PROBLEMS IN BAYESIAN STATISTICS RELATING TO DISCONTINUOUS PHENONEMA, CATASTROPHE THEORY, AND FORECASTING

J. Q. SMITH

Ph. D.

University of Warwick

Department of Statistics

November, 1977.

UNIVERSITÄTSBIBLIOTHEK DER TECHNISCHEN UNIVERSITÄT HANNOVER
UND TECHNISCHE INFORMATIONSBIBLIOTHEK

The Library
University of Warwick

Coventry CV4 7A

Great Britain

Dissertations Acquisition
Universitätsbibliothek und TIB
Welfengarten 1B
D 3000 Hannover 1
(Ferderal Republic of Germany)

Your Ref.:	Our Ref.:	Telfon	Date
	1.1.4/le	(0511)762-3419	November 30, 1978

ATTENTION PLEASE ! This is a claim for the b.m.thesis, which has been requested from us at first on July 26, but till today we haven't received any copy or answer from. Can you please proove, what has happend with our request?

Dear Ladies and Gentlemen,

We kindly ask you to send us a 35mm microfilm/xerox/microfiche copy (copies) of the following thesis (theses) for purchase. The invoice please send to the address, as above.
The title of thesis (theses):

AUTHOR : Smith , J Q : Thesis 1977

TITLE : Problems in bayesian statistics relating to discontinuours phenomena, catastrophe theory and forecasting.

VERIFIED: - . - . -

Thanking you very much in advance for your kind attention to our letter, I remain

Your faithfully

i.A. Anette Luecke

(Miss) Anette Luecke
Library Assistant

TABLE OF CONTENTS

Page No.

Abbreviations

Unless otherwise stated $n(\theta,V)$ represented a normal distribution with mean θ and variance V. In Chapter 2 I use $\hat{P}(\theta)$ to denote the posterior distribution of θ when $P(\theta)$ is the prior. All other abbreviations are defined in the text.

List of Figures:

ACKNOWLEDGEMENTS

I wish to thank all those in the Department of Statistics at Warwick University whose help and enthusiasm contributed to this thesis. In particular I should like to thank Jeff Harrison for his bombardment with conceptual ideas and for checking the thesis before typing. I need also thank Keith Ord and Robin Reed for help in some of the proofs in Chapters 2 and 3, Tom Leonard for his general encouragement and references and Tony O'Hagan for directing my attention to the t-normal product given in Chapter 5; the first page of working here is his.

I should like to thank Pam for putting up with numerous neurotic storms during my writing up and Terri for making such a good job of the typing. Finally, I thank the referee's in advance for reading the thesis and for their comments.

DECLARATION

All work in this thesis is to the best of my knowledge original, except for the page of algebra on the t-normal product which as mentioned in the acknowledgements is some research of Dr. O'Hagan.

To Pamela

SUMMARY

The aim of this thesis is to generalise Bayesian Forecasting processes to models where normality assumptions are not appropriate. In particular I develop models that can change their minds and I utilise Catastrophe Theory in their description.

Under squared-error loss types of criteria the estimates will be smoothed out, so for model description and prediction I need to use bounded loss functions. Unfortunately the induced types of estimators have not been investigated very fully and so two chapters of the thesis represent an attempt to develop theory up to a necessar level to be used on Times Series models of the above kind.

An introduction to Catastrophe Theory is then given. Catastrophe Theory is basically a classification of C^{∞}-potential functions and since the expected loss function is in fact itself a potential function, I can use the classification on them. Chapters 6 and 7 relate the topologies of the posterior distribution and loss function to the topologies of the posterior expected loss hence a Bayes classification of posterior distributions is possible.

In Chapter 8, I relate these results to the forecasting of non-stationary time series obtaining models which are very much akin to the simple weighted moving average processes under which lies this firm mathematical foundation. From this I can generate pleasing models which adjust in a "Catastrophic" way to changes in the underlying process generating the data.

1). Introduction to Thesis

May I take the earliest opportunity to apologise for the style in which this thesis is written. I write in the first person singular for three quite inadequate reasons. The first is that I find the passive tense difficult to read, the second is that the word "we" is a trifle regal for me to feel comfortable using it and the third is that for a scientist who believes only in subjective reality it expresses his basic philosophy succinctly and frequently.

The reason I have included this note is that I feel it important to outline the layout of the rest of the thesis so that the reader knows which parts are the most important.

Chapters 6, 7 and 8 represent the core of results obtained and are the central chapters of the thesis. Chapters 6 and 7 deal with Bayes estimation problems relating the topology of the posterior distribution to the topology of the posterior expected loss. Chapter 8 gives a generalisation of the Steady Model defined in Harrison and Stevens (1) for general distributions.

Chapter 4 and 5 give an introduction to Catastrophe Theory and examples of its uses in getting to understand the topology of well known likelihoods, posterior distributions and expected loss functions. These two chapters therefore make much less intense reading than the rest of the thesis.

Chapter 2 is at the beginning of this work because its subject matter is extremely important *conceptually* to the mood of the rest of the thesis. I use a stability rather than an axiomatic approach to show that to act sensibly in a Bayesian framework I must work with bounded likelihoods and loss functions. Having come to the latter conclusion, which I must remark is not new, though it is a point never emphasised by theoretical Bayesians, I realised that most of the usual estimation procedures used by practical

Bayesians are bad. It was therefore necessary to develop a theory around bounded loss. This is what Chapter 3 is all about.

Thus Chapters 2 and 3 are reference chapters and only the *results* therein contained are of importance to the rest of the thesis.

Finally may I point out that there is a summary at the end of each chapter which can be used by the reader to get an idea of the main points covered in them.

2. A DISCUSSION OF BAYESIAN INFERENCE

2.1. Introduction

Before using any inferential procedure it is important to understand how and when that procedure can be used. In the following chapter the author explores when it is likely that a Bayesian analysis will give reasonable and coherent results (the word "coherent" being used in the non-technical sense). There are two ways in which such an exploration can begin.

(i) An axiomatic base

(ii) Identification of various "counterexamples" to the system.

The axiomatic basis for Bayesian statistics is well documented (See, for example, De Groot (1), De Finetti (1), Raiffa (1)). I must admit to feeling that although these axioms give a good feel for what one is doing when making inference about distributions, the implications of seemingly innocuous assumptions can often be much more than one would suppose. It follows that they may induce a type of structure that is clearly undesirable.

Rather than get lost in these obscurities (to which I will refer throughout), I would like to start with the conclusion of the axiomatic systems and criticise the building rather than the bricks. The concluding statement from the axiomatic systems is:

"I can express all my prior opinions about a particular parameter $\theta \in \mathbb{R}$ in terms of a probability distribution $P(\theta)$, my prior belief that θ lies in a Borel set A being represented by the number $P(A)$".

Now there is a gap here, an inferential step, that many seem to have missed. *How* am I to express my prior beliefs in terms of the measure P, that is, do I substitute what I conceive heuristically as my subjective probability distribution for θ or something else?

If there is to be any meaning to Bayes inference I obviously must do the former. For this link to be made I feel that it is at least necessary (but by no means sufficient) that the following two criteria are met.

Criteria

1). If 2 prior distributions $P_1(\theta)$, $P_2(\theta)$ are "close", then their posterior distributions $P_1(\theta)$, $P_2(\theta)$ are close

i.e. $M : P(\theta) \to \hat{P}(\theta)$ is continuous.

2). If 2 priors $P_1(\theta)$ $P_2(\theta)$ are "close" then their associated decisions (or estimates) d_1, d_2 respectively must be "close"

i.e. $M : P_k(\theta) \to d_k$ is a continuous map.

If these criteria did not hold, then prior probabilities would lose their intuitive meaning since *the exact* form of the prior would have to be known for any sense to come from the inference. But how can one be sure of the exact form? Obviously one could not since the intuitive, subjective idea of probability is necessarily fuzzy.

Of course, in the usual statistical tradition (see Wilkinson (1)) I have not specified what I mean precisely by the above two criteria. Firstly the word "close" must be defined. This is done by introducing a topology onto the distribution functions which is related to the class of probability measures in some direct sense.

2.2. A choice of topologies on F. (the class of distribution functio

Clearly there are many topologies that one could use to give this idea of closeness. Perhaps the weakest such topology is generated by the Lévy metric.

Definitions

Define the distance in *Lévy metric* $\rho_L(F,G)$ between two distribution functions F and G to be the infimum of all $h > 0$ such that

$$F(\theta-h) - h \leq G(\theta) \leq F(\theta+h) + h. \quad \text{for all } \theta \in \mathbb{R}.$$

The usefulness of this topology is that

$F_n \xrightarrow{w} F$ properly (i.e. properly in distribution) iff $\rho_L(F_n,F) \to 0$ (See Moran (1)).

Although this is a useful property for our purposes it should be noted that it linearly varies its measure of distance as the parameter θ is transformed linearly.

The *uniform metric* $\rho_o(F,G)$ is defined by

$$\rho_o(F,G) = \sup_{\theta \in \mathbb{R}} |F-G|$$

The *variation metric* $\rho_v(F,G)$ is defined by

$$\rho_v(F,G) = \sup_{\{u \in C_o : \|u\| = 1\}} \{ \| \int_{\mathbb{R}} U dF - \int_{\mathbb{R}} U dG \|' \}$$

Lemma 2.1.

(i) The Lévy metric is at least as weak as the Uniform metric

(ii) The Uniform metric is at least as weak as the Variation metric

Proof. (i) is an obvious consequence of the fact that the Lévy metric convergence is equivalent to convergence in distribution.

(ii) Let X, Y has distribution functions F and G respectively

Then $\rho_o(F,G) = \sup_{\theta \in \mathbb{R}} |\mathbb{E}(\chi_{(-\infty\theta]}(X)) - \mathbb{E}(\chi_{(-\infty\theta]}(Y))|$

where $\chi_{[a,b]}(X) = \begin{cases} 1 & X \in [a,b] \\ 0 & \text{otherwise} \end{cases}$

and $\rho_v(F,G) = \sup_{\{u \in C_o : \|u\| = 1\}} |\mathbb{E}\, U(X) - \mathbb{E}\, U(Y)|$

which since $C_o^{\prime}|F$ is dense in F under the usual metric

$$= \sup_{\{u \text{ measurable}: \|u\| = 1\}} |\mathbb{E}\, U(X) - \mathbb{E}\, U(Y)|$$

$$\geq \rho_o(F,G). \qquad \square$$

In fact this weakness is strict. Consider these two counterexamples.

Counterexample 1.

Let $F = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{otherwise} \end{cases}$ and $F_n = \begin{cases} 0 & \text{if } x \leq \frac{1}{n} \\ 1 & \text{otherwise} \end{cases}$

Then $\rho_L(F,F_n) \to 0$ as $n \to \infty$ yet $\rho_o(F,F_n) = 1$ for all n.

Counterexample 2.

It can be shown that if F and G are absolutely continuous

$$\rho_v(F,G) = \int_{-\infty}^{\infty} |f(x) - g(x)|\,dx \qquad \text{(See Feller (1))}.$$

Let F be distributed rectangularly on [0,1] and $G_m(x)$ defined by

$$G_m(x) = \frac{1}{m} \sum_{n=1}^{m} J_n^{(m)}(x)$$

where $J_n^{(m)} = \begin{cases} 1 & x \in \left(\frac{n-1}{m} + \frac{1}{m^2}, \infty\right) \\ 0 & x \in \left(-\infty, \frac{n-1}{m}\right] \\ \text{increasing elsewhere} \end{cases}$

is continuous and differentiable. (For the specific construction of such functions see Bröcker (1)).

Then G_m is an absolutely continuous distribution on [0,1] and

$$\rho_o(F,G_m) \leq m \quad \text{yet it is easily checked that}$$

$$\int_{\infty}^{\infty} |f(x) - g_m(x)|dx > \tfrac{1}{2} \qquad \text{for all } m > M$$

where M is a suitably large constant.

Hence $\rho_o(F,G_m) \to 0$ and $\inf_{m>M} \rho_v(F,G_m) \geq \tfrac{1}{2}$ as $m \to \infty$. □

Now I shall introduce a little notation.

Let τ_i '>' τ_j signify that ρ_i induces a weaker topology τ_i than ρ_j does τ_j.

τ_i '≡' τ_j signify that the induced topologies are equivalent.

Then the above comments can be summarised by

Weak convergence '≡' τ_L '>' τ_o '>' τ_v.

I shall now proceed as follows. If I can show an inferential procedure to satisfy Criteria 1 and 2 with respect to weak topologies (in particular the ones induced by ρ_L) then I will provisionally accept it. If, however, no matter how strong the topology which I put on the distribution fucntions Criteria 1 and 2 are still not satisfied I will reject the inference as fatuous if used in this general setting.

To show that any counterexamples are not extrodinary in some sense I will assume that to reject an inference a $C^n(\mathbb{R})$ counterexample (i.e. an n × differentiable distribution where n is an arbitrary integer).

2.3. Topologies on $F|C^n(-\infty\infty)$

Let $\tau^{(n)}$ be the topology induced by the basis $B^{(n)}(F,\varepsilon)$ where

$$B^{(n)}(F,\varepsilon) = \{G \in F|C^n(-\infty,\infty): \sup_{\theta\in\mathbb{R}}(|D^nF(\theta)-D^nG(\theta)| < \varepsilon\}$$

and τ_n be the topology induced by the basis $B_n(F,\varepsilon)$ where

$$B_n(F,\varepsilon) = \{G \in F|C^n_{(-\infty,\infty)}: \sup_{\theta\in\mathbb{R}} \{\max_{0\le i\le n} | D^i F(\theta)-D^iG(\theta)| < \varepsilon\}$$

So, for example $\tau^{(1)}$ close says that the p.d.f. of G are uniformly close to F

τ_3 close says that the distribution functions, p.d.f and its derivative, of G are uniformly close to F,

and so on.

Clearly as n becomes large, the topology τ_n becomes extremely strong and hence should satisfy anyone.

Using the notation of the previous section the following diagram is true in $\underline{F}|C^n_{(-\infty,\infty)}$

weak convergence '$\equiv$' τ_L '$\equiv$' τ_o '$>$' τ_v '$>$' $\tau^{(1)}$ '$\equiv$' τ_1 '$>$' τ_2 '$>$' ... '$>$' τ_n

To prove this, it is necessary to show

(i) Convergence in $\tau_o \Rightarrow$ weak convergence in $F|C^n_{(\infty,\infty)}$

(ii) Convergence in $\tau^{(1)} \Rightarrow$ Convergence in variation,

since all other "inequalities" have either been proved already or are trivial.

Lemma 2.2.

(i) Convergence in $\tau_o \Longrightarrow$ weak convergence in $F|C^n_{(-\infty,\infty)}$

(ii) Convergence in $\tau^{(1)} \Longrightarrow$ Convergence in variation.

Proof: (i) Note that since $F(\theta)$ is a continuous *distribution function* it is uniformly continuous on $\mathbb{R}$.

Hence $\sup_{\theta \in \mathbb{R}} |F(\theta) - F(\theta+\delta)| < \eta(\delta)$ where $\eta(\delta) \to 0$ as $|\delta| \to 0$

So $\sup_{\theta \in \mathbb{R}} \{F(\theta-\varepsilon) + \varepsilon < F_n(\theta) < F(\theta+\varepsilon) - \varepsilon\}$ can be arranged in the form

$= \sup_{\theta \in \mathbb{R}} \{\tau_1(\varepsilon) < F_n(\theta) - F(\theta) < \tau_2(\varepsilon)\}$: where $\tau_1(\varepsilon)\ \tau_2(\varepsilon) \to 0$ as $\varepsilon \to 0$

convergence weakly $\Longleftrightarrow$ convergence in Lévy metric, so from the definition of the Lévy metric

$$\rho_L(F_n(\theta),F(\theta)) \to 0 \Rightarrow \varepsilon \to 0$$
$$\Rightarrow \tau_1(\varepsilon),\ \tau_2(\varepsilon) \to 0$$
$$\Rightarrow \rho_o(\varepsilon)(F_n(\theta),F(\theta)) \to 0 \text{ as required}$$

(ii) Let $G \in B^{(1)}(F,\delta)$

Then $\int_{-\infty}^{\infty} |f(x)-g(x)|dx \le \int_{|x|<\delta^{-\frac{1}{2}}} |f(x)-g(x)|dx + \int_{|x|>\delta^{-\frac{1}{2}}} f(x)dx +$

$$\int_{|x|>\delta^{-\frac{1}{2}}} g(x)dx \qquad 2.3.1$$

Note that $\int_{|x|>\delta^{-\frac{1}{2}}} g(x)dx = \int_{|x|>\delta^{-\frac{1}{2}}} f(x)dx + \int_{|x|>\delta^{-\frac{1}{2}}} (f(x)-g(x))dx.$

Hence (2.3.1) becomes

$$\leq 2\left\{\int_{|x|<\delta^{-\frac{1}{2}}} |f(x) - g(x)|dx + 1 - F(\delta^{-\frac{1}{2}}) + F(-\delta^{-\frac{1}{2}})\right.$$

$$< 2\,\delta^{\frac{1}{2}} + \eta(\delta) \quad \text{where } \eta(\delta) = 1 - F(\delta^{-\frac{1}{2}}, + F(-\delta^{-\frac{1}{2}}) \to 0$$

as $\delta \to 0$.

The result follows. □

The reader will be pleased to hear that I can make some comments on Bayesian inference now.

2.4. Restrictions on prior to posterior analysis

Restriction 1. The Likelihood function $\ell(\theta|x)$ is measurable.

Comment. This is in fact an extremely mild condition on ℓ, in fact it is difficult to construct a problem when this is not the case. However it does emphasise one point, namely that the family of sample distributions that have been chosen to represent the experiment have to be "sensible" in some way.

Restriction 2. The likelihood function must be bounded.

Comment. This is a restriction that does not seem to be commonly realised and often "occurs" in practice. To demonstrate why this does not satisfy Criteria 1 suppose $\ell(\theta|x) \to \infty$ as $\theta \to 0$ and is continuous on $(0,k)$ where $k > 0$. I can without loss of generality consider

$$\ell_1(\theta|x) = \ell(\theta|x)|(0,k) \text{ since I can assume that it is}$$

appropriate to put prior measure zero on θ outside this range. For construction purposes transform the interval $(0,k)$ smoothly to $(-\infty,\infty)$.

I now have a likelihood $\ell(\theta|\ x)$ on $\mathbb{R}$ such that:

$$\ell(\theta|x) \to \infty \qquad \text{as } \theta \to -\infty$$

$$\ell(\theta|x) \text{ is defined and finite elsewhere.}$$

Consider the distribution function $F_1(\theta|a)$: $f_1(\theta|a) \propto n(a,1)$ $a \in \mathbb{R}$.

Then $\underset{a\in\mathbb{R}}{\text{Sup}}\ (\underset{\theta\in\mathbb{R}}{\text{Sup}}\ (\underset{1\le k\le n}{\max}\ |D^k F_1(\theta|a)|))$

$$= \underset{\theta\in\mathbb{R}}{\text{Sup}}\ (\underset{1\le k\le n}{\max}\ |D^k F_1(\theta|O)|) \qquad \le M_n \in \mathbb{R} \qquad 2.4.1.$$

since $D^k f(\theta|O) = P_k(\theta) \exp\{-\tfrac{1}{2}\theta^2\}$ where $P_k(\theta)$ is a polynomial.

Now suppose I have chosen a prior $F(\theta)$ and let $G(\theta)$ be defined by

$$G(\theta|\alpha,a) = (1-\alpha)F(\theta) + \alpha F_1(\theta|a)$$

Then it is easy to check, using the above comment (2.4.1) that for each $\varepsilon > 0$ there is an A such that if $\alpha < A$, $a \in \mathbb{R}$

$$G(\theta|\ \alpha,a) \in B_n(F,\varepsilon), \text{ provided } \underset{x\in\mathbb{R}}{\text{Sup}}\ (\underset{1\le k\le n}{\max}\ D^k(F(x))) \le M$$

However, if $\hat{G}$ and $\hat{F}$ represent the posterior distributions using respective priors G and F,

$$\hat{G}(\theta) \propto (1-\alpha)\ F^*(\theta) + \alpha\ F_1^*(\theta|a) \qquad 2.4.2.$$

$$\text{where } F^*(\theta|a) = \int_{\infty}^{\theta} \ell(\theta|x)dF(\theta) \qquad 2.4.3.$$

$$F_1^*(\theta|a) = \int_{\infty}^{\theta} \ell(\theta|x)dF_1(\theta|a) \qquad 2.4.4.$$

Hence

$$\hat{G}(\theta) = \frac{(1-\alpha)F^*(\infty)\hat{F}(\theta) + \alpha F_1^*(\infty|a)\hat{F}_1(\theta|a)}{(1-\alpha)F^*(\infty) + \alpha F_1^*(\infty|a)} \qquad 2.4.5.$$

$$= (1-\alpha^*)\ \hat{F}(\theta) + \alpha^*\ \hat{F}_1(\theta|a) \qquad 2.4.6.$$

$$\text{where } \alpha^* = [1 + (\alpha^{-1}-1)F^*(\infty)(F^*(\infty|a))^{-1}]^{-1} \qquad 2.4.7.$$

providing it exists. Since ℓ is unbounded at $-\infty$, as $a \to -\infty$ $F^*(\infty|a) \to \infty$. It follows that for each fixed α there is an $R(\alpha)$ such that for all $a < R(\alpha)$

$$\frac{(\alpha^{-1}-1)F^*(\infty)}{F^*(\infty|a)} < 1 \quad \text{i.e.} \quad \alpha^* > \frac{1}{2}. \qquad 2.4.8.$$

Finally note that $F(a) \to 0$ as $a \to -\infty$ whereas

$$F_1(a\ a) = \tfrac{1}{2} \text{ for all } a \in \mathbb{R} \qquad 2.4.9.$$

Hence there is an R_1 (depending on α and hence ε) such that if

$$a < R_1(\varepsilon)$$

$$|G(a|a) - F(a)| = |\alpha^* F_1(a|a) - (1-\alpha^*)F(a)| > \frac{1}{3}. \qquad 2.4.10.$$

Hence for all priors F and all $\varepsilon > 0$ there is a G such that $G \in B_n(F,\varepsilon)$ and yet

$$\hat{G} \notin B_o(\hat{F},\eta) \quad \text{if } \eta < \frac{1}{3}.$$

In words this means that any prior distribution I choose must be specified exactly, otherwise the posterior distribution and hence my consequent inference will be more or less arbitrary. Hence Criteria 1 is not met. □

It follows that in general I cannot make sensible inference in a Bayesian setting (or for that matter m.l.e. approach see Edwards (1)) using unbounded likelihoods. The difficulty is only an apparent one, however, for the following reasons. In any experiment (and here I echo Bartlett et al (1)) I can only take a measurement within a tolerance governed by my measuring instrument. So rather than take an observation x I take an observation $x \pm \varepsilon$. Hence I will, in general, observe a set of positive measure rather than a point. It will follow that the corrdsponding likelihood will then be bounded, so this criteria is always met in good modelling.

Since the form of the tolerance region $(x \pm \varepsilon)$ taken are often communicated by the scaling in which the observations are given, I feel that the invariance principle applied to the scaling in the sample distribution of the data is a bad Criterion for discrimination between inferences.

Restriction 3

My prior F must be such that $\int_{\mathbb{R}} \ell(\theta|x)dF(\theta) \neq 0$.

Comment

This means that apriori I must have the Prob $(\ell(\theta|x) > 0) > 0$ Even a Bayesian is going to be confused if every value of θ he thinks a priori is possible his data tells him is not and values of θ he diened impossible his data tells him have positive likelihood.

This highlights indirectly an important deficiency of Bayesian analysis. If one remembers that θ is just a label for a particular family $F(\theta)$ of sampling distributions then on putting a prior on θ we automatically put zero measure on all other sampling distributions. So I cannot see when my data (whilst not rejecting as above) seems to contradict the family.

For example suppose that I assume that the random variable X is normally distributed with mean μ and unit variance and I have prior on θ which is normal mean 10 variance 1. If I now take 1,999 observations of which

999 have value 10.0000

1000 have value -10.000.

Then my posterior distribution (which I quote "contains all the posterior information") says θ is normally distributed mean 0 variance 0.0005.

But what rubbish is this! I am almost positive that the points have come from a n(0,1) distribution! This is obviously ridiculous, but without sidestepping the formalism there is no way out of the problem. The more data points I have the more information there is contradicting my prior choice of distribution functions, but I cannot adapt to more likely ones since I have put measure zero on them.

Of course this problem arises from the fact that I am not putting a prior "distribution" across function space in a sensible way. At the moment I am researching into more sensible methods but it should be noted that the inferential procedures I have found so far, contradict De Groots 5^{th} axiom (1),that is, that I can compare the chance of each distribution function being right with a uniform distribution on [0,1]. For interesting analagous problems see Ferguson (1), Leonard (4).

Henceforth write $\hat{F}(\theta)$ as the posterior distribution of θ given data $\underset{\sim}{x}$.

Restriction 4

Let $T(\ell(\theta|x))$ be the set of discontinuities of ℓ. Then $T(\ell(\theta|x))$ must have measure zero with respect to the prior distribution F.

Comment

First I will give an example of the restriction.

Example

$$\text{Let } \ell(\theta|x) = \begin{cases} 0 \text{ on } \mathbb{Q} \cap [0,1) \\ 1 \text{ otherwise} \end{cases}$$

The prior $F(\theta)$ is a rectangular distribution on [0,1] and the prior $F_n(\theta)$ defined by

$$F_n(\theta) = \begin{cases} 0 & x < 0 \\ m/n & x \in [m, m+1) \quad 0 \le m \le n-1 \\ 1 & \text{otherwise} \end{cases}$$

Then $\rho_L(F, F_n) = \frac{1}{n} \to 0$ as $n \to \infty$

But $\hat{F}(\theta)$ is a rectangular distribution on [0,1]

$$\hat{F}_n(\theta) \text{ is } \begin{cases} 0 & x < 1 \\ 1 & x \ge 1 \end{cases} \quad \text{if } \rho_L(\hat{F}, \hat{F}_n) = 1 \text{ for all } n.$$

Hence if I have a family of sampling distributions labelled by θ, I must arrange the family in such a way that the priors are ordered in a natural continuous way with respect to their labels, i.e. the closeness of distribution functions is mimicked by closeness in θ. This is a topic that I will go into more detail about in a later section.

Having now discussed the restrictions I can show that under them Criteria 1 is satisfied.

Preservation under Lévy norm

Lemma 2.3. If (i) $\ell(\theta|x)$ is measurable

(ii) $\theta_n \xrightarrow{w} \theta$ and $P(T(\ell(\theta|x)) = 0$ (where T is defined above)

then $\ell(\theta_n) \xrightarrow{w} \ell(\theta)$

Proof [See Billingsey (13)] □

Lemma 2.4. If $\ell(\theta|x)$ is bounded above, then

$$\ell(\theta_n) \xrightarrow{\omega} \ell(\theta) \Rightarrow \mathbb{E}(\ell(\theta_n)) \to \mathbb{E}(\ell(\theta))$$

Proof [See Billingsey (1)] □

<u>Theorem</u> Let $P_n \to P$ in Lévy metric, and $\ell(\theta)$ be a likelihood,

then provided (i) $\ell(\theta)$ is bounded and measurable

(ii) $P(T(\ell)) = .0$

(iii) $\int_{-\infty}^{\infty} \ell(\theta) dP(\theta) > 0,$

then $\hat{P}_n \to \hat{P}$ in Lévy metric.

<u>Proof</u>. As a consequence of Lemmas 1 and 2 and assumption (iii)

$$\int_{\mathbb{R}} \ell(\theta)\, dP_n(\theta) \to \int_{\mathbb{R}} \ell(\theta)\, dP(\theta) = P^*(\theta) > 0 \qquad 2.4.11.$$

where convergence is in Levy metric

Also for each t

$$P_n^*(t) = \int_{\infty}^{\infty} \chi(-\infty,t]\ \ell(\theta) d\ell_n(\theta) \to \int_{\infty}^{\infty} \chi(-\infty,t] \ell(\theta) dP(\theta) = P^*(t) \quad 2.4.12.$$

as $n \to \infty$

at all continuity points of P, by replacing $\ell(\theta)$ by $\chi(-\infty,t)\ \ell(\theta)$

and using Lemmas 2.3 and 2.4.

But $\hat{P}_n(t) = \dfrac{P_n^*(t)}{P_n^*(\infty)}$, $\hat{P}(t) = \dfrac{P^*(t)}{P^*(\infty)}$, which by (2.4.11) and (2.4.12)

gives that $\hat{P}_n(t) \to \hat{P}(t)$ at all points of continuity of $\hat{P}(t)$

(Since $P^*(\infty) > 0$)

The result follows. □

To conclude this section, consider the following theorem.

<u>Theorem 2.5</u>. Suppose that (i) $\ell(\theta)$ is bounded above by M and is measurable

(ii) $\int_{\mathbb{R}} \ell(\theta) dF(\theta) > 0$

Then $F_n \to F$ in τ_v topology implies $\hat{F}_n \to \hat{F}$ in τ_v topology in the class absolutely continuous priors.

Proof Firstly, as in the previous theorem, by premise (ii) it is sufficient to prove that

$$F_n \to F \Rightarrow F_n^* \to F^* \quad \text{using the notation above.}$$

Well, since $\ell(\theta)$ is measurable $f^* = \ell f$ and $f_n^* = \ell f_n$ are Lebesgue measurable as a consequence of the D.C.T, it follows that $\hat{F}$ and $\hat{F}_n$ are absolutely continuous.

So $\rho_v(F^*F_n^*) = \int_{-\infty}^{\infty} |\ell(\theta)f(\theta) - \ell(\theta)\, f_n(\theta)|d\theta$ by a comment in the previous section

$$\leq M\int_{-\infty}^{\infty} |f(\infty)-f_n(\theta)|d\theta$$

$$= M\rho_v(F,F_n). \quad \text{The result follows} \qquad \square$$

So if I work in the class of all absolutely continuous distribution functions, in fact Restriction 4 is no longer needed provided the ρ_v metric is used. Perhaps the moral of this story is that when dealing with Bayesian inference, the Lévy topology is a little too weak and that it might be more sensible to restrict oneself to absolutely continuous priors when considering continuous phenomena. My personal feelings are that Restriction 4 is much more difficult to justify than the restrictions imposed above.

Note that there is still a need for a natural ordering of parametrisation of the family of sample distributions since 2 likelihoods ℓ_1 and ℓ_2 equal a.s. will give the same inference. The set of measure zero on which they are different may contain the "true" sample distribution so anomalies (which I will not at this stage go into) could arise.

2.5. Stability of Bayes Estimates

The reader will be familiar with the way a Bayes decision/estimate d is made. Without loss of generality I can assume that my utility is linear (by absorbing it into the loss function (see Chapter 3) so my Bayes decision corresponds to the infimum of the expected loss (with respect to prior F and loss function L) written E(L,F,d) (where L and F may be omitted if no confusion arises.)

Since my concern is estimation rather than general decision making I will assume I have a loss function of the form: $L(d-\theta)$.

Unbounded loss functions

Definition An unbounded loss function $L(d-\theta)$ has the property that

$$\sup_{\theta \in S} L(d-\theta) = \infty \qquad \text{for all } d \in \mathbb{R}.$$

where S is the extended support of the posterior distribution F (i.e. the smallest open interval containing all points such that $f(\theta) > 0$ (see Chapter 3))

Until now I think that it is safe to say that the bulk of Bayes estimation have corresponded to such loss functions. The fact that this is theoretically absurd is apparent when one sees that I am taking the infimum of a function E(F,d) which may or may not occur depending on the particular convention I employ to obtain F. I will elucidate this point.

Claim

The rate of convergence to zero of the tails of my likelihood should not make a significant difference to any inference I make.

Justification

Suppose that I have a family of sample distributions $\mathcal{G}$ parametrised by θ and have observed a measurable set

$$T = (x^* - \varepsilon,\ x^* + \varepsilon).$$

Following the same sort of argument as for Criteria 1, the family

$$G = \{G_\theta(x) : \theta \in \mathbb{R}\}$$

of sample distributions cannot be specified *precisely*, each must be considered as a representative of itself and distributions "close" to it in any practical situation. Again I must define "close" so I shall use the strong definition in the section preceding this, since I am looking for counterexamples.

Now consider an alternative family of sample distributions G^* defined by

$$G^* = \{(1-\alpha)G_\theta(x) + \alpha F_1(s\ a) : \theta, a \in \mathbb{R},\ 0 \le \alpha \le 1\}$$

where $F_1(x|a)$ is a normal distribution with mean a and variance 1. In an exactly analogous way to the example in Restriction 2 in the previous section, it can be shown that for all $\varepsilon > 0$ there is an A such that for all $0 \le \alpha \le A$ and $a, \theta \in \mathbb{R}$

$$G \in G^* \qquad G \in B_n(G_\theta(x), \varepsilon).$$

provided $\sup_{x \in S(G_\theta)} \{\max_{1 \le i \le n} D^i G_\theta(x)\} \le M$ for some $M \in \mathbb{R}$.

In particular putting $a = x^*$ where x^* is defined above, the likelihood induced by $G(\theta, a, \alpha)$, $\ell^*(\theta)$ is such that

$$\ell^*(\theta) \to \alpha \qquad \text{as } |\theta| \to \infty$$ provided that the original family with likelihood $\ell(\theta)$ has the property $\ell(\theta) \to 0$ as $|\theta| \to \infty$. I leave the reader to check that small perturbations of the likelihood do not affect the continuity arguments of the previous section.

The point of this claim is to show that perturbations of my prior distributions in the tail under τ_n topology can induce the same perturbations in the tails of the posterior distribution if the problem is changed an indissemable amount, since by wiggling my family of sample distriubtions a bit I can make the likelihood constant in the tails. Hence without loss of generality I can consider perturbations of the posterior distribution rather than the prior, when the perturbations are in the tail of the posterior.

I can now return to the original problem, that of constructing counterexamples to using unbounded loss functions.

Lemma 2.6.

Suppose that a function $L(\theta) \to \infty$ as $\theta \to \infty$ and finite elsewhere in $\mathbb{R}_{>0}$. Then there exists a distribution function F such that $F_2 \in F | C^n(-\infty,\infty)$ such that

$$\int_{-\infty}^{\infty} L(\theta)\, dF_2(\theta) = \infty \ , \quad \text{where} \quad \sup_{\theta \in \mathbb{R}} \{\max_{1 \le i \le n} D^i F_2(\theta)\} \le M_n \ ,$$

Proof

Since $L(\theta) \to \infty$ there exists points $t_1 \ldots t_n$ such that

$L(\theta) > 2$, $\theta \in A_1 = [t_1-1, t_1+1]$

$L(\theta) > 2^4$, $\theta \in A_2 = [t_2-1, t_2+1]$

$\vdots$

$L(\theta) > 2^j$, $\theta \in A_j = [t_j-1, t_j+1]$

where $A_i \cap A_j = \emptyset$ $i,j \in \mathbb{N}$.

Let $F_2(\theta) = \sum_{i=1}^{\infty} \frac{1}{2^i} J_i(\theta)$ where $J_i(\theta)$ is some chosen C^∞ distributio function with p.d.f. having support in (t_i-1, t_i+1) and such that $\sup_{\theta \in \mathbb{R}} \{\max_{1 \le i \le n} D\, J_j(\theta)\} \le M_n$ for at J.

Then (i) $F(\theta)$ is a C^∞ distribution function

(ii) $$\int_{\mathbb{R}} L(\theta)dF(\theta) > 2\,\frac{1}{2} + 2^2.\,\frac{1}{2^2} + \ldots = \infty$$

□

Now suppose I do a Bayesian analysis using a loss function $L(d-\theta)$ and a posterior distribution F, giving rise to a Bayes decision d^*, the minima of expected loss, its associated expected loss being $E(L,F,d^*)$. Perturbing F slightly in ρ_n topology by replacing it by

$G = (1-\alpha)F + \alpha F_2$ where F_2 is defined in the preceeding lemma,

I see that the associated expected loss no longer exists at d^* i.e. this is in fact a *worst* decision. So my posterior decision depends crucially on the (usually conjugate) form I have chosen to approximate it. This is obviously unacceptable.

Consider this even more stunning counterexample when I use convex loss functions (advocated by De Groot (2)) and including the squared error loss currently in vogue.

Let $(\theta-d) = S$ and suppose $L(s)$ is differentiable and $L'(s)$ is strictly monotone.

Lemma 2.7.

If L is defined as above and $E(L,F,d)$ exists for all $d \in \mathbb{R}$, then it has exactly one minima, provided $F'(\theta) \neq 0$, $\theta \in \mathbb{R}$.

Proof

It is obviously sufficient to prove that $E(L,F,d)$ has exactly one stationary point since $E(L,F,d) \to \infty$ as $d \to \infty$, so this stationary point must be a minima.

$$\text{Well } E'(F,L,d) - E'(F,L,0) = \int_{\mathbb{R}} L'(s)dF(s+d) - \int_{\mathbb{R}} L'(s)dF(s)$$

$$= \int_{\mathbb{R}} (L'(s)-L'(s-d))dF(s)$$

which is strictly monotone in d since L'(s) is strictly monotone in its argument (by definition). Hence E'(F,L,d) is strictly monotone. It follows that it will cut the axis d = 0 at most once. □

Theorem 2.8.

Let $L(\theta-d)$ be a convex loss function with its derivative $L'(s)$ continuous and my original posterior distribution function F is such that

$$\sup_{\theta\in\mathbb{R}} \{ \max_{0\le k\le r} |D^* F(\theta)| \} \le M \quad \text{where } f(\theta) \neq 0 \text{ on } \mathbb{R}.$$

Then for all ε(small) > 0 A^*(large) > 0 and $n \in \mathbb{N}$ there is a posterior distribution function G such that

$$\rho_n(F,G) < \varepsilon \quad \text{and} \quad |d(F) - d(G)| > A^*$$

where d(F) and d(G) are the Bayes decisions corresponding to F and G respectively.

Proof. Define $F_1(a,\theta)$ as $n(a,1)$ and

$G(\alpha,a,\theta) = (1-\alpha)F(\theta) + \alpha F(a,\theta)$. Then it has previously been shown that for $\varepsilon > 0$ there is an $\alpha^* > 0$ such that for all α $\alpha < \alpha^*$ and $a \in \mathbb{R}$

$$\rho_n(G(\alpha,a),F) < \varepsilon.$$

Let E(d,G) represent the expected loss with respect to decision d and distribution function L. Then d(G) such that E'(d(G),G) = 0 gives the (unique by Lemma 2.7) Bayes decision.

Well, $E'(d,G(\alpha,a)) = \int_{-\infty}^{\infty} L'(s)\,((1-\alpha)\,f(s+d)+\alpha f_1(a;s+d))ds = 0$

implies $E'(d,F) = \frac{\alpha}{1-\alpha}\,E'(d,F_1(a))$.

Fix an arbitary $d \in \mathbb{R}$. Then it is easily checked that

$$E'(d,F_1(a)) \to \infty \quad \text{as } a \to \infty$$

$$E'(d,F_1(a) \to -\infty \quad \text{as } a \to -\infty \text{ since } L' \text{ is}$$

increasing and unbounded.

It follows that for all $d \in \mathbb{R}$ there is an $a(d) \in \mathbb{R}$ such that

$E'(d,F) = \frac{\alpha}{1-\alpha}\,E'(d,F(a))$, i.e. such that d is the unique Bayes decision with respect to $G(\alpha,a(d),\theta)$.

The result is now clear. □

I hope that the reader is now satisfied that such contortions of a proper Bayesian analysis are just not on. The question remains "Is Criteria 2 satisfied by a proper Bayesian analysis?" (i.e. one where bounded loss functions are used). The answer is almost. First a definition.

Definition

A decision $d(\hat{F})$ is said to be *stable within J* with respect to topology induced by the metric ρ if for all $\eta > 0$ there is an $\varepsilon > 0$ such that

$$\rho(\hat{F},\hat{G}) < \varepsilon \implies |\, C(\hat{G}) - d(\hat{G}) \,| < \eta$$

where C is some point in J.

Call a decision simply *stable* if $J = \{d(\hat{F})\}$

Theorem 2.9.

Suppose (i) $L(s) = L^+(s) + L^-(s)$

where $L^+(s) = \begin{cases} L(s) & s > 0 \\ 0 & \text{otherwise} \end{cases}$ $L^-(s) = \begin{cases} L(s) & s < 0 \\ 0 & \text{otherwise} \end{cases}$

and such that $\left.\begin{matrix} L^+ \\ L^- \end{matrix}\right\}$ are $\left\{\begin{matrix} \text{right} \\ \text{left} \end{matrix}\right.$ $\left.\begin{matrix} \text{continuous increasing} \\ \text{decreasing} \end{matrix}\right\}$ and bounded by $\begin{matrix} M_1 \\ M_2 \end{matrix}$

(ii) $\int_{\mathbb{R}} L^+(\theta-d)dF(\theta)$ and $\int_{\mathbb{R}} L^-(\theta-d)dF(\theta)$ are continuous.

then $E(L,F_n,d) \to E(L,F,d)$ uniformly as $\rho_L(F_n,F) \to 0$

Proof

$$E(L,G,d) = \int_{\mathbb{R}} L^+(\theta-d)dG(\theta) + \int_{\mathbb{R}} L^-(\theta-d)dG(\theta)$$

$$= M_1^{-1} \int_{\mathbb{R}} H^+(d-\theta)dG(\theta) - M_2^{-1} \int_{\mathbb{R}} H^-(d-\theta)dG(\theta)$$

where H^+ and H^- are distribution functions. Each of the above integrals is therefore a convolution. Hence

$$E(L,G,d) = M_1^{-1} U_G^{(1)}(d) + M_2^{-1} U_G^{(2)}(d)$$

where $U_G^{(1)}$ and $U_G^{(2)}$ are the convolutions mentioned above.

It is a well known result that (See Feller (1))

$$\rho_L(F,F_n) \to 0 \implies \rho_L(U_G^{(i)}, U_{G_n}^{(i)}) \to 0 \quad i = 1,2.$$

so provided $U_G^{(1)}$ and $U_G^{(2)}$ are continuous by Lemma 2.2.

$$\rho_L(F,F_n) \to 0 \implies \rho_o(U_G^{(i)}, U_{G_n}^{(i)}) \to 0 \quad i = 1,2.$$

So in particular

$E(L,F_n,d) \to E(L,F,d)$ uniformly as $\rho_L(F_n,F) \to 0$. □

Comments

The conditions on this theorem need some discussion. Firstly if F is continuous (see Feller (1) p.147) condition (ii) is automatically satisfied and even if L^+ and L^- are not respectively right and left continuous the proof holds. So inthis case the theorem is true for all bounded loss functions of the form L(s). Secondly if I allow F to be discontinuous and L continuous in S, the conditions of the theorem are also met. So unless my loss function and distribution function are very discontinuous the theorem holds.

Finally notethat if $\rho_L(L_n,L) \to 0$, $E(L_n,F,d) \to E(L,F,d)$ by symmetry of the convolution operator. Putting these two results together gives the following Corollary.

Corollary 2.9.1.

If the conditions of Theorem 2.9 are met for all L_n and

$\max \{\rho_L(L_n,L), \rho_L(F_n,F)\} \to 0$, then

$$E(L_n,F_n,d) \to E(L,F,d) \text{ uniformly.} \qquad \square$$

Thus if Loss functions and distributions are close, so is the expected loss function. This is the most important result for Bayesians. However the Bayes estimate is one step away from this, since I am interested in the infimum of such functions.

Theorem 2.10

Let $\mathcal{D}$ be the set of minima of expected loss with respect to the originally chosen posterior distribution E(L,F,d). Suppose there is no sequence $d_j \in \mathcal{D}$ such that

$$\lim_{j\to\infty} E(L,F,d) \to E(L,F,d(F))$$

and $\lim_{j\to\infty} d_j \neq d(F)$ where d(F) is a Bayes decision.

Then if $E(L,F_n,d) \to E(L,F,d)$ uniformly $d(F)$ is stable within J where

J = set of original Bayes decisions.

Proof

Suppose to the contrary there exists a sequence of distributions such that

$E(L,F_n,d) \to E(L,F,d)$ and yet

$\lim_{n\to\infty} |d(F_n) - d(F)| > \varepsilon$ where $d(F) \in J$ and $d(F_n)$ is a Bayes decision w.r.t. $E(L,F_n,d)$.

Since $E(L,F_n,d) \to E(L,F,d)$ uniformly

$\inf_{d\in\mathbb{R}} E(L,F_n,d) \to \inf_{d\in\mathbb{R}} E(L,F,d)$ (since $E \geq 0$ by definition)

Hence $E(L,F_n,d(F_n)) \to E(L,F,d)$ as $n \to \infty$ contradicting the hypothesis. □

Since this condition is extremely weak (since it is on the posterior I end up with) I have in fact proved that condition 2 is satisfied provided I use a bounded loss function in the Bayesian setting.

A related piece of work has subsequently emerged from Kadane & Chuang (1) dealing in a weaker sort of way with general loss functions of the form $L(d,\theta)$. They seem however to completely miss the point that the infimum of an expected loss function has no meaning if that expected loss function does not exist.

2.6. A Preview of the Kernel of the Thesis

The reader may be wondering what this has to do with Catastrophe Theory and Time Series. The answer is that Catastrophe Theory is a classification theorem about families of potential functions

THE EXPECTED LOSS FUNCTION IS A POTENTIAL FUNCTION

since the behaviour (or decisions) are governed by its minima. Thus theorems classifying minima of potentials are applicable.

Now if convex loss functions are used Lemma 2.7 shows that only one minima can arise on the expected loss function. Since Catastrophe Theory is a classification in terms of the number of minima, it would be redundant. But I have shown that such abortions of Bayesian analysis admit ridiculous and unacceptable results. I *must* work with bounded loss functions so Catastrophe Theory is applicable.

Now consider the last theorem, but this time, rather than interpret $\hat{F}$ as the original posterior distribution let it represent the "best" representation of my posterior beliefs. Suppose the corresponding expected loss function E(L,F,d) is smooth and has 2 minima and 1 maxima where the 2 minima m_1, m_2 are each Bayes decision with respect to $\hat{F}$.

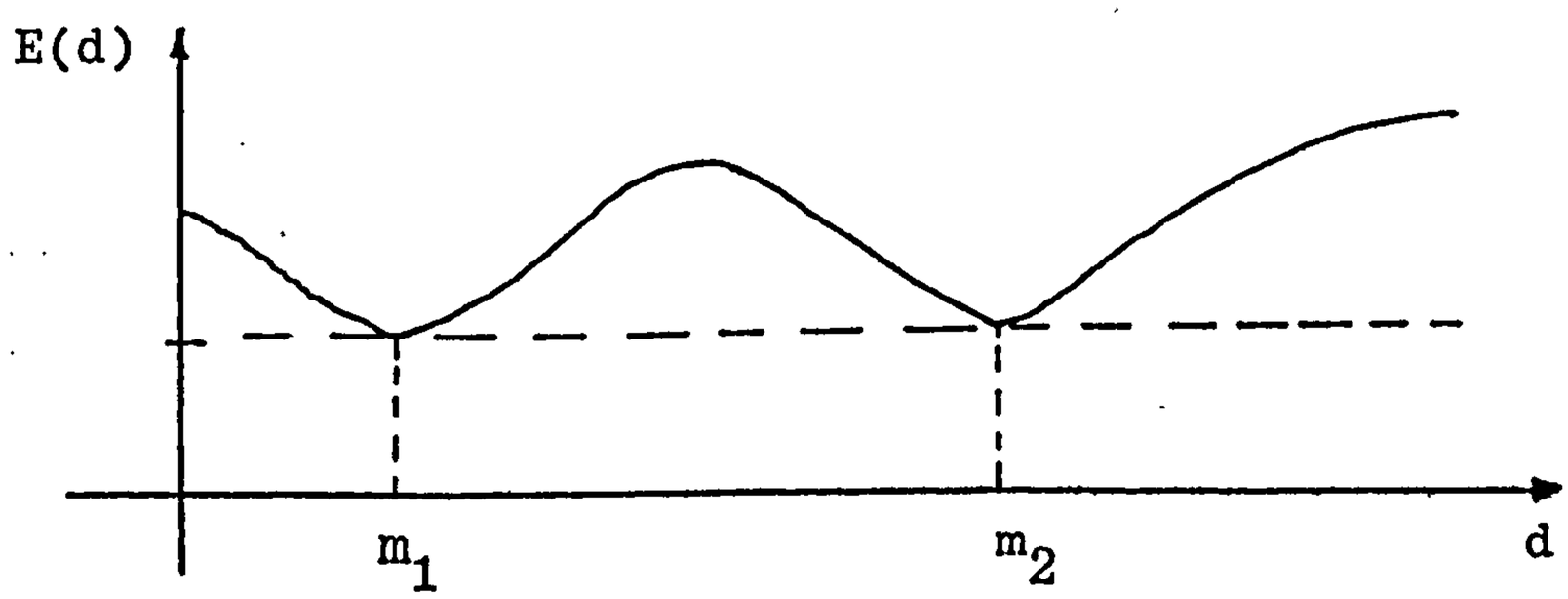


Fig. 2.1.

If F_1 and F_2 are 2 perturbations of $\hat{F}$, i.e. 2 approximations of the "best" representation of my posterior beliefs, then no matter how good my approximations, the Bayes decision $d(F_1)$, $d(F_2)$ of F_1 and F_2 respectively could be very different and unique (e.g. $d(F_1)$ near m_1, $d(F_2)$ near m_2) Hence a bifurcation (classified in Catastrophe Theory) is observed.

Certainly in a one-off situation such as $\hat{F}$ (with two identical minima) would be an extreme oddity. (see for example Morse Theory). But suppose $\{\hat{F}\}$ is in fact a parametrised family $\{\hat{F}(t) : t \in \mathbb{R}\}$ where t might for example represent time. Then one often finds that the family $\{E(L,\hat{F}(t),d) : t \in \mathbb{R}\}$ goes through evolution (1) → (2) → (3) pictured below, as t increases.

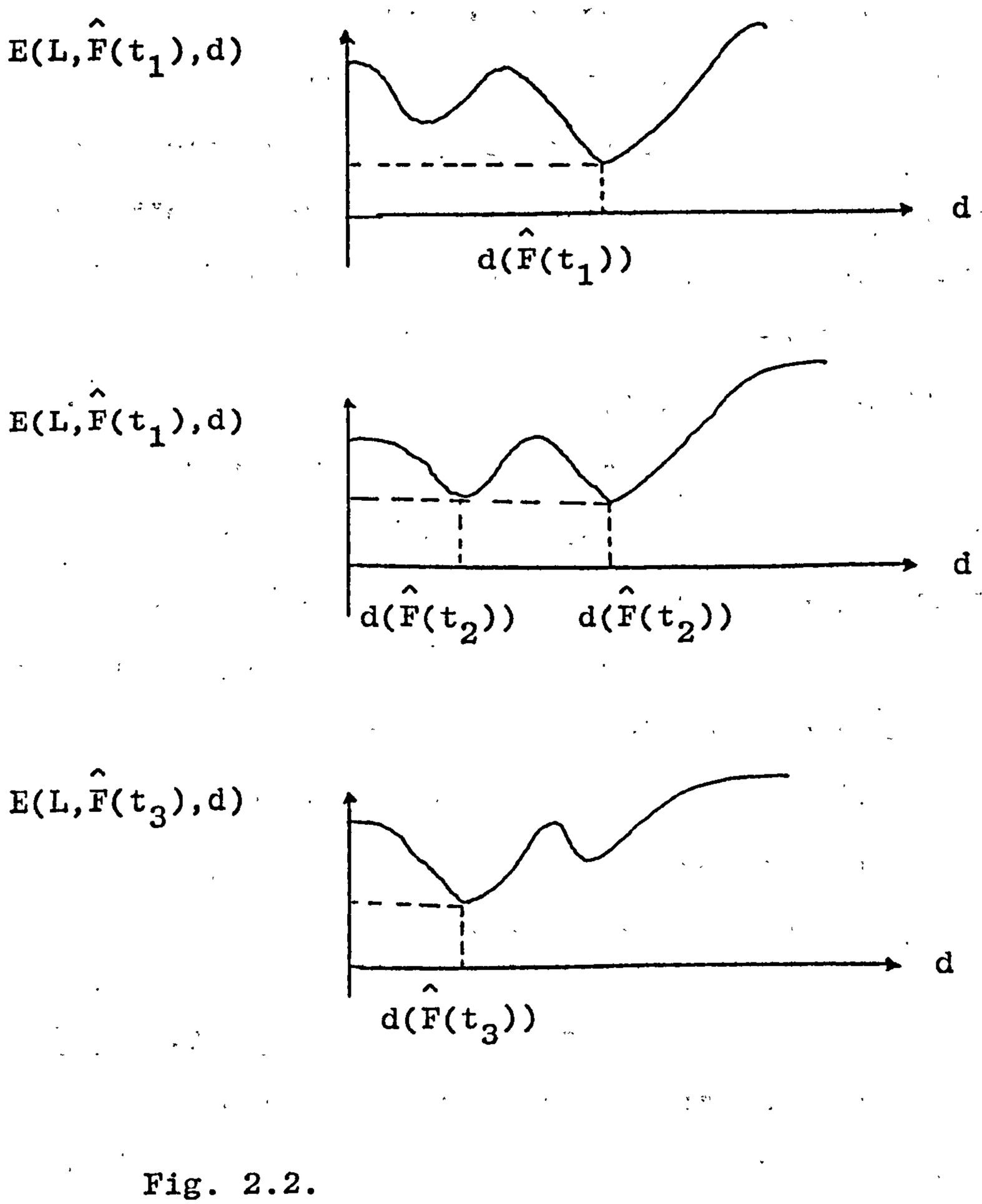


Fig. 2.2.

The Bayes decision thus experiences a jump from a point in a neighbourhood of $d(F(t_1))$ to a point near $d(F(t_3))$. Hence Catastrophes occur (in a Bayesian sense) readily in:

Time Series

Sequential Analysis

Removal of nuisance parameters and many other fields. The first of these is the most striking and so I shall concentrate on it in this thesis.

n × Differentiability of the expected loss follows if either:

(i) L is n × differentiable w.r.t. d and bounded in some sense (see Burrill (1) for details)

(ii) the distribution function used is sufficiently smooth.

Since such loss functions/distributions functions will be dense (under ρ_L metric) in the space of all bounded loss functions/F, the last theorem and the previous Corollary mean that without loss of generality I can make assumptions of smoothness of E(d).

Since no-one has yet developed Bayesian inference with bounded loss functions far enough I must devote a chapter to these problems (Chapter 3) before going on to using Catastrophe Theory on them. For this I need the concept of natural parametrisation.

2.7. Natural Parametrisations

It does not seem to be widely realised that Bayesian inference using a loss function L is invariant under transformations of the parameter, even though the posterior moments/mode are not, since

$E(d) = \int_{\mathbb{R}} L(\theta,d)dF(\theta)$ does not depend on the parametrisation of θ.

The transformation is just absorbed into the loss function. Therefore the most sensible way in which to define a "natural"parametrisation is by specifying the loss structure and then changing the parametrisatio of θ and d such that L is in a simple form. In most problems of an inferential nature it is possible to convert the original loss function into the form $L(\theta-d)$ by an appropriate choice of parametrisation of θ.

Note that this implies that before I make a Bayes estimate I must say what it is I am looking for. (i.e. fix $L(\theta-d)$. This seems sensible but in an inferential setting the appropriate loss function may at first seem difficult to find: many statisticians (including Bayesians) don't *like* to state explicitly what they are looking for, for fear that their results may seem subjective. The following suggestions are for those at a loss choosing their right parametrisation - i.e. when the loss structure is not immediately obvious to them.

Definition

Call a parametrisation θ of a family F_o of sample distributions *ρ-continuous* if

$$\rho(F_{\theta'}(x),\ F_{\theta+\eta}(x)) < k(\eta) \qquad \text{where } k(\eta) > 0 \to 0 \text{ as } \eta \to 0.$$

How often there exists a ρ-continuous parametrisation of such a family of distribution functions is a moot point and requires further research. Certainly if this condition is met, then problems outlined in Restriction 4 etc, are bypassed. I will take this one step further.

Definition

Call a parametrisation θ of a family of sample distributions F_θ ρ-*natural* if

$$\rho(F_\theta(x),\ F_{\theta+\eta}(x)) = k(\eta) \qquad \text{where } k(\eta) > 0,$$

$$k(\eta) \to 0 \text{ as } |\eta| \to 0.$$

and k is locally about 0 1-1.

If such a parametrisation exists, then it is easily seen that it is natural with respect to the metric ρ in the sense that it measures a function of the ρ-distance between distribution functions of the same family. Hence the loss function depends only on the distance (i.e. ρ-distance) on t between the chosen sample distribution and the actual sample distribution. The last condition of the definition ensures that for positive η , k will separate the distributions that are close to the chosen F_θ.

A question now arises: How many "natural" parametrisations are there for a particular family of distributions?

Theorem 2.11

A ρ-natural parametrisation is unique up to linear transformations of the said parametrisation.

Proof

Clearly from the definition, if θ is a natural parametrisation then aθ + b will be.

Conversely suppose both θ and J(θ) are ρ-natural paramterisation i.e.

$$\rho(F_{\theta_1}(x),\ F_{\theta_2}(x)) = k_1(\theta_1 - \theta_2) \qquad 2.7.1.$$

$$\rho(F_{\theta_1}(x),\ F_{\theta_2}(x)) = k_2(J(\theta_1) - J(\theta_2)) \qquad 2.7.1.$$

and assume without loss of generality. J(θ) = 0.

By definition, on some neighbourhood $(0,\varepsilon)$, (say) k_2 will have an inverse, so putting $k_3 = k_2^{-1}\, k_1$ (1) and (2) can be written

$$k_3(2\lambda) = J(\lambda+\mu) - J(\mu-\lambda) \quad \text{where } \lambda = \tfrac{1}{2}(\theta_1-\theta_2) \qquad 2.7.3.$$
$$\mu = \tfrac{1}{2}(\theta_1+\theta_2)$$

$$\text{or } k_3(2\lambda) = J(y+x) - J(y) \quad \text{where } x = 2\lambda \qquad 2.7.4.$$
$$y = \mu-\lambda.$$

for sufficiently small λ (wx)

Putting $y = 0$ in (4)

$$k_3(2\lambda) = J(x) - J(0) = J(x) \text{ by assumption.}$$

Hence $J(y+x) = J(x) + J(y)$, so that J is a linear function of θ . □

So I have at least found that such a definition gives me a unique parametrisation (up to linear transformation). It should be noted that a function k defined above has properties induced by the metric namely

(i) k is symmetric about 0

(ii) k is concave downwards

(i.e. $k(x) + k(y) \leq k(x+y)$) $\quad x,y \in \mathbb{R}$

so it looks something like Fig. 2.3.

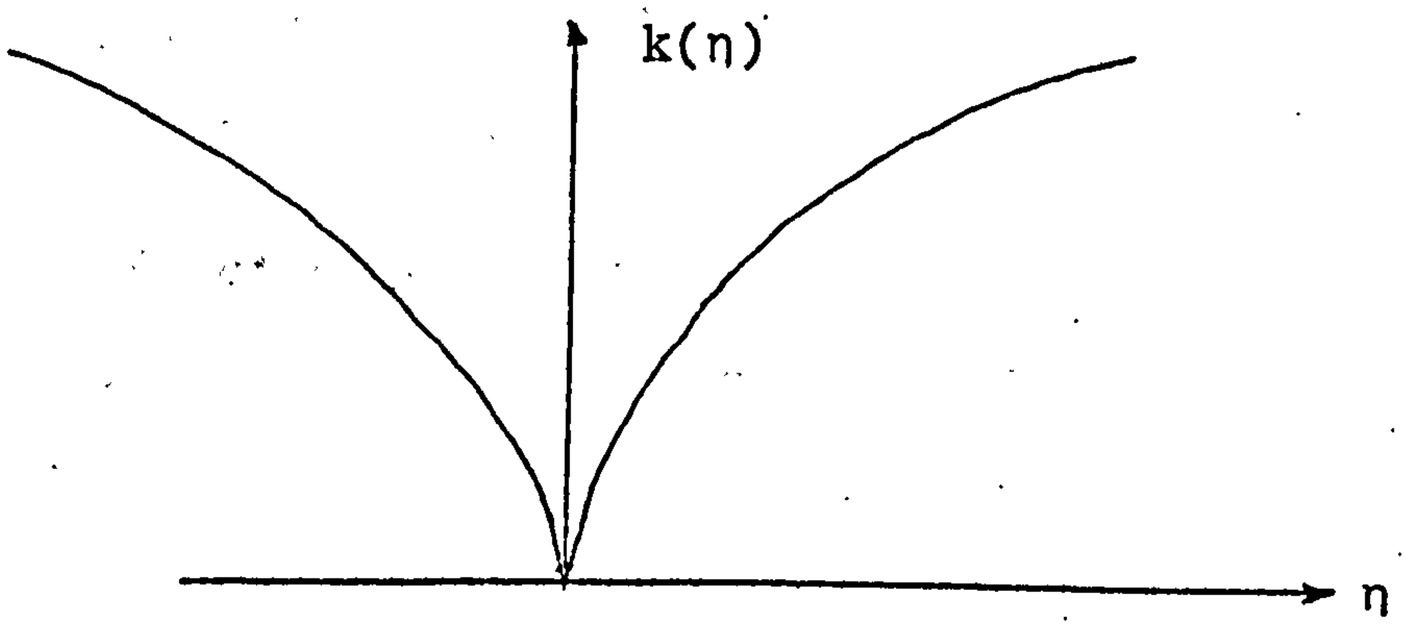


Fig. 2.3. The function $k(\eta)$.

Before I can proceed any further I must choose a particular metric. This, as I have said before, will depend very much on the situation I am parametrising (i.e. my loss structure). The simplest metric to consider is the ρ_o-topology for which the following theorem gives many results for ρ_o-natural parametrisation.

Theorem 2.12

Let $F_x(.|\theta)$ be a family of continuous distribution functions of $X \in I_x \leq \mathbb{R}$ with parameter $\theta \in I_\theta \leq \mathbb{R}$. Suppose:

(i) There exists a transformation T: $X \to Y$ such that

$$\rho_o(F_y(.|\theta_1), F_y(.|\theta_2)) = r[(\rho_o(F_x(.|\theta_1), F_x(.|\theta_2))]$$

for each θ_1 $\theta_2 \in I_\theta$ where r is increasing in some $[0,\varepsilon)$ neighbourhood of 0 with $r(0) = 0$.

(ii) That $F_y(y|\theta+h) = F_y(y-h|\theta)$ for each $\theta \in I_\theta$

(iii)That for each $\theta \in I_\theta$ $F_y(y|\theta)$ is strictly increasing on an open interval (possibly infinite) I_y and constant on I_y^c.

Then θ is itself a ρ_o-natural parametrisation

Proof

$$\rho_o(F_x(.|\theta+h), F_x(.|\theta)$$

$$= r[\rho_o(F_x(.|\ \theta+h), F_y(.|\theta))] \text{ by (i)}$$

$$= r[\rho_o(F_y(y|\theta), F_y(y-h|\theta))] \text{ by (ii)}$$

which is { a function of h only
{
{ increasing in h for small $h > 0$ by (iii) and the
{
{ definition or r

So θ is a ρ_o-natural parametrisation. □

Notes

1). If T is a monotone function of x, then condition (i) of the theorem is satisfied fatuously from the definition of the a distribution function.

2) If each $F(.|\theta)$ $\theta \in I$ has a symmetric p.d.f. about a common point (w.l.o.g. O) then if T is some even function then again condition (i) of the theorem is satisfied since

$$\rho_o(F(.|\theta), (F(.|\theta_2) = \sup_{x \in I_x} |F(x|\theta_1) - F(x|\theta_2)| = \sup_{x>O \in I_x} |F(x|\theta_1) - F(x|\theta_2)$$

So since T is 1-1 on $x > O$ the argument above applies.

Examples

1). If the family is of the form $F(\theta-x)$, then θ is already in its ρ_o-natural parametrisation (e.g. normal mean, t-distribution mode etc.)

2). If the family is symmetrical and of the form $F(\frac{\mu-x}{\theta})$ put $T(x) = \ell n|\mu-x|$. It is then easily seen, using note 2, that $\ell n\ \theta$ satisfies condition (ii) of the theorem and so (modulo condition (iii)) gives a ρ_o-natural parametrisation. (e.g. normal variance V has a ρ_o-natural parametrisation $\ell n\ V$ or a $\ell n\ v + b$ for any constants a, and b).

Although the search for such parametrisations is obviously very interesting it is a bit off the track of the thesis so I will leave most further classification for my further research.

Summary

I have shown that to use Bayesian inference successfully I must have:

i) A reasonable set of sample distributions to distinguish between (Restriction 1)

ii) One of these sample distributions is the "right" one (Restriction 2)

iii) The likelihood function must be bounded (Restriction 2)

iv) The sample distribution must be ordered in a sensible way (Restriction 4)

v) I must use *bounded* loss functions to get my estimates.

I have noted that with (v) I must admit the possibilities of Catastrophes. In the last section I have suggested a way in which a sensible ordering of sample distributions (iv) can be found.

3. SOME PROPERTIES OF ESTIMATES MADE UNDER BOUNDED LOSS

3.1. Introduction

In this section I will give a classification of Bayes decisions made when using bounded loss functions, the classification being slanted towards decisions which are estimates of a parameter. The idea is ultimately to use this on Time Series data to make sequential decisions.

Although many theorists (e.g. De Groot (1) have emphasised that the axiamatic systems forming the basis of Bayesian statistics imply the use of bounded loss, it has not been until recently that any serious work has gone into looking at the properties of the induced estimates (see Lindley (1) and (2)). On the other hand estimates using, for example, quadratic loss functions have almost exhaustively been looked at (e.g. Chao (1) and De Groot and Rao (2)).

In the last chapter it was pointed out the sort of pitfalls around when unbounded loss structures are used. In any real life situation resourses always have an upper bound anyway.

Of course, there are some difficulties which arise from thinking of estimates as decisions. Perhaps the most poignient is the fact that the *personal* utility function of the decision maker must be specified before an optimality criteria is well defined.

Utility Functions

Suppose that I have found my posterior distribution $P(\theta)$ of θ and that I have a loss function $L(\theta,d)$ with an associated decision $d \in D$ the decision space. The loss function will represent my rational assessment of losses incurred from certain decisions d when the true value of the parameter is θ. (Hence my loss function correspond to - De Groot's gain function (De Groot (1)).

The posterior $P(\theta)$ then induces a posterior distribution for the random variable $L(\theta,d)$ where d is fixed, for each $d \in D$. Call the distribution associated with $L(\theta,d)$ $P_d(\ell')$.

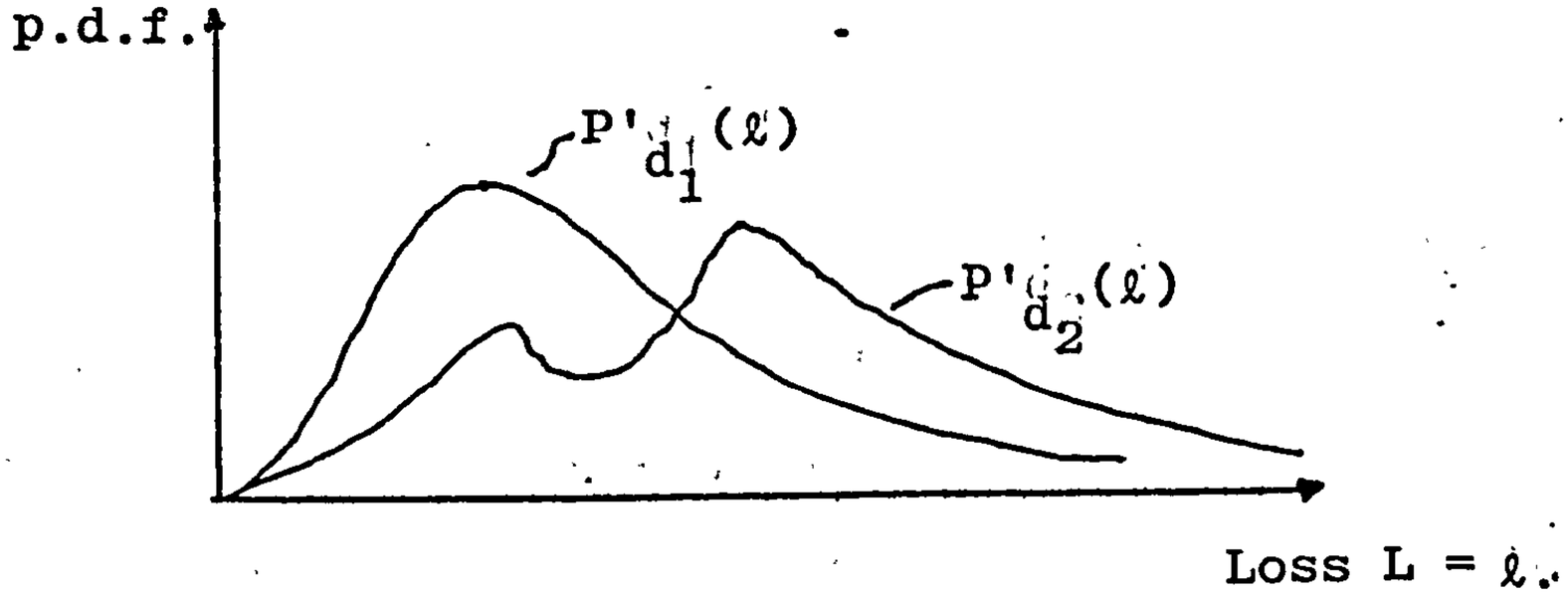


Fig. 3.1.

The axiomatic system then states that to make a "sensible" choice for $d \in D$. I must pick a $d^* \in D$ such that with respect to some strictly decreasing function of L, $U(\ell),d^*$ maximises

$$\{ \int_{\mathbb{R}>0} U(\ell)\, dP_d(\ell) : d \in D\}.$$

Equivalently I must pick a d^* such that there is a strictly increasing reparamentisation of L, A(L) such that

$$E(A(L),d) = \{ \int_{\mathbb{R}>0} A(L)dP_d) : d \in D\} \text{ is minimised.}$$

U is called the *Utility function* and A = -U I will call the *Anxiety function* (Note the invariance of optimal decisions under linear increasing transformations of A (or U), so without loss of generality assume A(O) = O).

The sceptical critic may assert that this statement about a best choice says very little. For example if $L(\theta,d)$ is of the form $L_1(\theta-d)$ where L_1 is symmetric about $(\theta-d)$ and increasing on $\mathbb{R}_{>0}$, I can find for any other fixed $L_2(\theta-d)$, symetric and increasing on $\mathbb{R} > 0$ an anxiety function A such that

$$L_2(\theta-d) = A(L_1(\theta-d))$$

Hence unless I can give A (or U) a *concrete* meaning, then the class of all loss functions symmetric in $\theta-d$ and increasing in $\mathbb{R} > 0$ give the same class of optimal decisions, so it does not matter what the form of the loss function is.

Now unlike the loss function L, A is a far more subjective quantity to L. Whereas L represents the rational assessment of the situation as perceived by the decision maker, A augments the situation to fit the optimism, pessimism, fears, expectations in fact the total emotional state of mind of this person.

There are 2 major misconceptions about the nature of Utility/Anxiety functions in the literature that firstly need to be removed before I can proceed.

(i) The "sensible" - non emotional utility/anxiety function is the linear one.

The consequence of this wrongly inferred statement is that it is commonly assumed that to be rational, I must choose a decision which minimises my expected loss C rather than my expected anxiety). A study of the build up of the axiamatic system C in e.g. De Groot (1)) should satisfy the reader that this is in fact a completely erroneous deduction. This hoped-for correspondence is just a case of wishful thinking. Anyway all Bayes decision (including those made under linear utility functions) contain an emotional element about them.

(ii) The utility/anxiety function is independent of the amount of information I think I have.

Suppose that I have a data set {X} and I am making inference about a parameter θ using a loss function L. The estimate I make will correspond to the Bayes decision obtained from using some anxiety functions A which the reader will remember is supposed to summarise, in part, the optimism and expectations of the decision maker. If I then increase the data set to include another much larger set of data {Y} is it reasonable to assume that the optimism of the decision maker will be the same as it was before? I think that it is extremely hopeful to expect this and so feel that the statement above is usually wrong - a point perhaps missed by many Bayesions when criticizing the Classical approach to inference.

With those difficulties in mind, it seems a good starting point in the classification of Bayes estimates to look at loss functions whose corresponding Bayes estimates are independent of the choice of Anxiety function.

Theorem 3.1.

If a loss function can only take two values for all θ and d, then the associated Bayes decision in invariant under choice of Anxiety function.

Proof: Any increasing transformation of 2 points is equal to an increasing linear transformation of those two points, and the Bayes decision is invariant under increasing linear transformations of the Anxiety function. □

3.2. The Step Loss function and its properties

If I am estimating and I have found a "natural" parametrisation (see §2) I can restrict my class of loss function to those which satisfy :

:

$$(L1) \quad \sup_{d\in D\theta\in H} L(\theta.d) = 1 \qquad \inf_{d\in D,\theta\in H} L(\theta,d) = 0$$

(Since L is bounded the above assumption loses no generality)

$$(L2) \quad L(\theta,d) = L(\theta-d) = L(d-\theta)$$

(L3) L is increasing in $|\theta-d|$

Note that for any increasing A continuous in [0,1] $A(L(\theta-d))$ also satisfies (L2) and (L3) and can be normalised so that (L1) can hold.

One Anxiety function invariant class of loss functions is the step loss functions.

Definition The *step loss function* with guage b is defined:

$$S_b(\theta-d) = \begin{cases} 0 & |\theta-d| \leq b \\ 1 & \text{otherwise} \end{cases}$$

Hence a misestimate by an amount less than b incurs no loss (i.e. the estimate is adequate) but misestimation by more than an amount b incurs maximum loss (i.e. the estimate is then inadequate). The reader may note the closeness between this loss function and confidence limits.

In what follows I shall assume that the posterior distribution for θ, $F(\theta)$ is twice differentiable, though most results in this section will carry through. It will be shown in the next section that these step-loss functions alone determine the class of estimates obtained from all loss functions satisfying (L1) (L2) (L3). First some properties of S_b.

Notation Let $E_b(d) = \int_{\mathbb{R}} S_b(\theta-d)\, dF(\theta) = F(d-b) + 1 - F(d+b)$.

The turning points d* of $E_b(d)$ all satisfy

$$f(d^*-b) = f(d^*+b) \qquad (3.2.1)$$

So in particular, all minima including the Bayes decision, will satisfy this equation. The next lemma will show that usually equation (3.2.1) has a unique solution which is the Bayes decision. So the Bayes estimate under any step loss function can be obtained in closed form for most of the standard distributions. A table giving some of these solutions is given in Section 3.5 of this chapter.

Definitions

Let the *support* of the continuous p.d.f. $f(\theta)$ of a distribution function $F(\theta)$ be the closure of the set of points θ such that $f(\theta) > 0$.

Let the *extended support* S(F) be defined by the interval (k_1,k_2) where

k_1 = inf {Support of F}

k_2 = sup {Support of F}

Call a distribution function $F(\theta)$ on $\mathbb{R}$ *properly unimodal* (or just unimodal if no confusion can arise) if $f'(\theta) = 0$ has exactly one solution in S(F) and $f(\theta)$ is continuous in $\mathbb{R}$.

$\phi(b)$, the *generalised location of* θ given posterior distribution $F(\theta)$ of θ is defined to be the set of points in S(F) satisfying the equation

$$f(\phi(b_1-b) = f(\phi(b) + b) \quad b \le \tfrac{1}{2}(k_2-k_1)$$

where k_1 and k_2 are defined above. □

In general, for each b, $\phi(b)$ is the set of points which contains (but not necessarily properly) all sensible estimates of θ under step loss guage b. It can be thought of a mapping

$$\mathbb{R} > 0 \quad \to S(F)$$

$$b \quad \to \phi(b)$$

It will be seen that this map is central to the whole discussion.

First a look at some of the properties of $\phi(b)$ when $F(\theta)$ is unimodal (properly).

Lemma 3.2.1.

If $G(\theta) = F(P(\theta+m))$, $P > 0$, $m \in \mathbb{R}$, where F is a 2 × differentiable function, then for every $d_1 \in \phi_G(b)$ there exists a $d_2 \in \phi_F(Pb)$ such that

$$d_1 = P^{-1} d_2 + m$$

Proof: If $d_1 \in \phi_G(b)$ then

$$g(d_1-b) = g(d_1+b)$$

i.e. $$f(P(d_1-b+m)) = f(P(d_1+b+m))$$

Hence $P(m+d_1) = d_2$ for some $d_2 \in \phi_F(Pb)$ □

It follows that for the next Theorem I can assume without loss of generality that the mode of a properly unimodal distribution function $F(\theta)$ is at zero, since all the stated properties will carry over by linear transformation of $F(\theta)$.

Theorem 3.2.

If $F(\theta)$, the posterior distribution of θ is 2 × differentiable on $S(F)$ and unimodal with zero mode then:

(i) $\phi(b)$ is a differentiable function on $(0, \frac{1}{2}(k_2-k_1))$

[Hence in particular the set $\phi(b)$ is a single point for each b]

(ii) $\phi(b) \in (-b,b)$

[Hence in particular $\lim_{b\to 0} \phi(b) = 0$ = mode of $F(\theta)$]

(iii) $|\phi'(b)| < 1$ $\quad b \in (0, \frac{1}{2}(k_2-k_1))$

(iv) If $f(\theta)$ is symmetric about 0 then

$$\phi(b) = 0 \text{ for all } b \in (0, \tfrac{1}{2}(k_2-k_1))$$

(v) If $F(\theta)$ has support $[k_1,\infty)$ $\phi(b) > b \in k_1$ for all $b \in \mathbb{R} > 0$

[Hence $\phi(b) \to \infty$ as $b \to \infty$].

Proof

(i) Write $F = (\text{range of } f(\theta) : \theta \in (k_1k_2))^0$

Since f is unimodal, $f'(\theta) > 0$ for $\theta \in (k_1,0)$ and $f'(\theta) < 0$ for $\theta \in (0,k_2)$ I can therefore define function

$$g_1 : F \to (k_1,0) \text{ by } g_1(.) = [f(.)|(k_1,0)]^{-1} \qquad (3.2.2)$$

$$g_2 : F \to (0,k_2) \qquad g_2(.) = [f(.)|(0,k_2)]^{-1} \qquad (3.2.3)$$

where g_1 and g_2 are differentiable with

$$g_1'(y) > 0 \text{ and } g_2'(y) < 0 \qquad y \in F \qquad (3.2.4)$$

Let

$$g_3 : F \to (0,\tfrac{1}{2}(k_2-k_1) \text{ be defined by } g_3(.) = \tfrac{1}{2}(g_2(.)-g_1(.)) \qquad (3.2.5)$$

$$g_4 : F \to S(F) \qquad g_4(.) = \tfrac{1}{2}(g_1(.)+g_2(.)) \qquad (3.2.6)$$

Then $g_3.g_4$ are also both differentiable functions and

$g_3'(y) < 0$ $y \in F$ by (3.2.4) and (3.2.5)

So $g_3^{-1}: (0,\tfrac{1}{2}(k_2-k_1) \to F$ is again a differentiable function (3.2.7)

The result now follows from checking that

$$\phi(b) = g_4(g_3^{-1}(b)) \qquad \square$$

(ii)

Since f is strictly decreasing on $(0,k_2)$ $d-b \geq 0 \to f(d-b) > f(d+b)$

f is strictly increasing on $(k_1,0)$ so $d+b \leq 0 \to f(d+b_1 > f(d-b)$ for $b \in (0,\tfrac{1}{2}(k_2-k_1))$.

The result follows.

(iii) Since g_3^{-1} is strictly decreasing.

$$f(\phi(b_1) - b_1) > f(\phi(b_2) - b_2) \quad b_1 < b_2 \quad b_1, b_2 \in (0, \tfrac{1}{2}(k_2-k_1))$$

$$f(\phi(b_1) + b_1) > f(\phi(b_2) + b_2)$$

But $\phi(b) - b < 0$ for all $b \in (0, \tfrac{1}{2}(k_2-k_1))$

$$\phi(b) + b > 0$$

and $f'(\theta) > 0 \qquad \theta \in (k_1, 0)$

$$f'(\theta) < 0 \qquad \theta \in (0, k_2)$$

and so $\phi(b_1) - b_1 > \phi(b_2) - b_2 \rightarrow \phi(b_2) - \phi(b_1) < b_2 - b_1$

$$\phi(b_1) + b_1 < \phi(b_2) + b_2 \rightarrow \phi(b_2) - \phi(b_1) > -(b_2-b_1)$$

The result now follows from the fact that ϕ is differentiable □

(iv) If $f(\theta)$ is symmetrical about 0, then $f(-b) = f(b)$ for all $b \in (0, \tfrac{1}{2}(k_2-k_1))$. This together with property (i) gives the result.

(v) Because $f(\theta) = 0 \quad \theta < k_1$ and positive elsewhere, to satisfy

$$f(\phi(b) - b) = f(\phi(b) + b), \ \phi(b) > b+k_1 \qquad \square$$

3.3. A Classification of Bayes Decisions using Step loss functions

In this section I will assume my loss function satisfies L1, L2 and L3 of Section 3.2 and that the anxiety function has been absorbed into $L(\theta,d)$ (so that I can assume A linear). The Bayes decision will then be the absolute minima of expected loss. A powerful theorem can now be obtained relating the Bayes decisions with respect to $L(\theta,d)$ to Bayes decisions under step loss, hence underlying the importance of the $\phi(b)$ map.

Lemma 3.3.1

Let $L(\theta-d)$ satisfy L1, L2 and L3 of Section 3.2 and suppose the posterior distribution $F(\theta)$ of the parameter θ is absolutely continuous.

Then $E_L(d) = \int_{\mathbb{R}>0} E_b(d)\,dG(b)$

where G is a probability measure on $\mathbb{R} > 0$ and $E_L(d)$ is the expected loss with respect to $F(\theta)$ and $L(\theta-d)$.

Proof

Let Y have distribution function $G(s) = \begin{cases} L^*(s) & s > 0 \\ 0 & \text{otherwise} \end{cases}$

where $L^*(s)$ is the right continuous version of $L(s)$. Then $-Y$ has associated distribution function $H(s) = 1 - G(-s)$. 3.3.1.

Thus $\int_{\mathbb{R}} L(d-\theta)dF(\theta)$

$= \int_{\mathbb{R}} (1-H(d-\theta))dF(\theta) + \int_{\mathbb{R}} G(d-\theta)dF(\theta)$ since $F(\theta)$ is absolutely continuous. By reversing these two convolution formulae

$$= 1 - \int_{\mathbb{R}} F(d+b)dH(-b) + \int_{\mathbb{R}} F(d-b)\, dG(b),$$

which on resubstituting via (3.3.1) and rearranging

$$= \int_{\mathbb{R}_{>0}} (1 - F(d+b) + F(d-b))\, dG(b)$$

$$= \int_{\mathbb{R}_{>0}} E_b(d)\, dG(b) \quad \text{as required} \qquad \square$$

Lemma 2. Let $F(\theta)$ be the differentiable posterior distribution of θ. Then

(i)	$E_b'(d) < 0$	$k_1 < d < \inf \phi(b)$	$b < \frac{1}{2}(k_2-k_1)$
(ii)	$E_b'(d) > 0$	$\sup \phi(b) < d < k_2$	$b < \frac{1}{2}(k_2-k_1)$
(iii)	$E_b'(d) \leq 0$	$k_1 \leq d \leq \frac{1}{2}(k_1+k_2)$	$b \geq \frac{1}{2}(k_2-k_1)$
(iv)	$E_b'(d) \geq 0$	$\frac{1}{2}(k_1+k_2) \leq d \leq k_2$	$b \geq (k_2-k_1)$

Proof (i) and (ii) are a consequence of the fact that $E_b(d) \to 1$ as $|d| \to \infty$ (ii) and (iv) follow directly from noting that for $b \geq \frac{1}{2}(k_2-k_1)$,

$$E_b(d) = \begin{cases} 1 - F(d+b) & d < k_2 - b \\ 0 & |d-\frac{1}{2}(k_2+k_1)| \leq b - \frac{1}{2}(k_2-k_1) \\ F(d-b) & d > b + k_1 \end{cases} \qquad \square$$

Theorem 3.3. If loss function L(s) satisfying L1, L2, L3 is chosen with the additional property that

$$L(s) = \begin{cases} 0 & s \le b_1 \text{ where } b_1 < \tfrac{1}{2}(k_2-k_1) \\ a(s) & b_1 < s < b_2 \qquad 0 < a(s) < 1 \\ 1 & s \ge b_2 \end{cases}$$

where $s = \theta - d$, then all the stationary points d^* of $E_L(d)$ (which of course include all the Bayes decisions) with respect to posterior distribution $F(\theta)$ (with $S(F) = (k_1,k_2)$ in $S(F)$ satisfy

$$d^* \in [d_1,d_2]$$

where $d_1 = \inf\{\phi((b_1,b_2^*))\}$

$d_2 = \sup\{\phi((b_1,b_2^*))\}$

$b_2^* = \inf\{\tfrac{1}{2}(k_2-k_1),b_2\}$.

Proof

$$E_2(d) = \int_{\mathbb{R}_{>0}} E_b(d)\, dG(b) \qquad \text{where G is defined in Lemma 1.} \qquad 3.2.2.$$

$$= \int_b^{b_2} E_b(d)\, dG(b) \qquad \text{b, the restriction on L(s) above} \qquad 3.3.3.$$

$$= E_1(d) + E_2(d) \qquad 3.3.4.$$

where $E_1(d) = \int_{b_1}^{b_2^*} E_b(d)\, dG(b)$ and $E_2(d) = \begin{cases} 0 & \text{if } b_2^* = b_2 \\ \int_{\frac{1}{2}(k_2-k_1)}^{b_2} E_b(d)dG(b) & \text{otherwise} \end{cases}$ 3.3.5.

By (i) and (ii) of Lemma 2,

$$E_b'(d) < 0 \qquad d < d_1 \qquad b \in (b_1,b_2^*) \qquad 3.3.6.$$

$$E_b'(d) > 0 \qquad d > d_2 \qquad b \in (b_1,b_2^*) \qquad 3.3.7.$$

Also by the definition of G and the properties of L given above, G(b) ascribes positive weight to (b_1,b_2^*)

Hence $E_1'(d)$ $\quad O < k_1 < d < d_1$ } by commuting the integration
$E_2'(d)$ $\quad O > k_2 > d > d_2$ } and differentiation operations.3.3.9.

Note that if $\frac{1}{2}(k_2-k_1) = b^*$ then $\frac{1}{2}(k_2+k_1) \in [d_1,d_2]$ since

$$\lim_{b \to \frac{1}{2}(k_2-k_1)} \phi(b) = \tfrac{1}{2}(k_1+k_2) \qquad 3.3.10$$

So by (iii) and (iv) of Lemma 2, by interchanging differentiation and integration I have that

$$E_2'(d) \leq O \quad k_1 < d < d_1 \qquad 3.3.11$$
$$E_2'(d) \geq O \quad k_2 > d > d_2$$

Hence combining (3.3.8), (3.3.9), (3.3.10), (3.3.12) with (3.3.4) I have that

$$E_2'(d) \leq O \quad k_1 < d < d_1 \qquad 3.3.13$$
$$E_2'(d) > O \quad k_2 > d < d_2$$

The result follows. □

Corollary 3.3.2

Any Bayes decision with respect to a loss function L satisfying conditions L1, L2, L3, and such that there is an $\varepsilon > O$ such that $L(\frac{1}{2}(k_2-k_1)-\varepsilon) > O$ must lie in an interval $[d_1,d_2]$

where $\quad d_1 = \inf \{\phi((O,\frac{1}{2}(k_2-k_1)))\}$ □
$\quad\quad d_2 = \sup \{\phi((O,\frac{1}{2}(k_2-k_1)))\}$

Suppose then that I know that nothing is lost if I misestimate by a quantity less than $b_1 \geq O$ and that my estimates are as bad as one another if they are out by more than a quantity b_2. Then under linear anxiety function, any sensible estimate must lie in the interval

$[d_1(b_1,b_2^*),\ d_2(b_1b_2^*)]$ defined in the theorem.

Notes

1). The above set is in fact *invariant* under changes of Anxiety function A (assuming A is *strictly* increasing in L) since

$$L \text{ constant on } (0,b_1) \text{ and } (b_2,\infty)$$
$$\Leftrightarrow A(L) \text{constant on } (0,b_1) \text{ and } (b_2,\infty)$$

Therefore the above set does not depend on the emotional state of the decision maker.

2). A partial converse is true if I assume $F(\theta)$ is unimodal. In this case it can easily be shown that for all $d^* \in [d_1,d_2]$ and $\varepsilon > 0$ there exists an Anxiety function A such that

$$|d_1^* - d^*| < \varepsilon \quad \text{where } d_1^* \text{ is a Bayes decision with respect to anxiety } A(L)$$

So the interval $[d_1(b_1,b_2^*),\ d_2(b_1,b_2^*)]$ is a very natural one to look at. It would be difficult to shrink the interval further whilst still keeping the invariance property above.

3). The term "generalised location map" for $\phi(b)$ relates to the fact that the location interval $[d_1,d_2]$ depends on $\phi(b)$ only.

4). If $\phi(b)$ is strictly increasing (decreasing on $b \in (0,\infty)$, the argument above can be refined to allow only estimates in the *open* interval

$$(d_1,d_2)$$

3.4. A new representation of a posterior distribution function

Suppose that my known anxiety function and loss function are combined in the usual way to get the function $L(\theta-d)$ satisfying conditions L1, L2 and L3. If, in addition, I add the mild constraint that $1-L(s)$ is integrable and a random variable X has p.d.f. $f_x(x)$ satisfying

$$f_X(s) \propto 1 - L(s), \text{ then } 1 - E_2(d) \propto f_Z(d)$$

where $f_Z(z)$ is the p.d.f. of the random variable $Z = X + \theta$ where parameter θ has associated posterior distribution $F(\theta)$. Using, for example, characteristic functions, it is trivial to show that with $f_Z(z)$ and $f_X(x)$ I can retrieve $F(\theta)$. So, by the above, if I communicate the functions $E_2(d)$ and $L(s)$ I can also get back to $F(\theta)$. Hence by retaining the loss function and anxiety function satisfying in combination L1, L2, L3 and the integrability condition, together with the expected anxiety, means that I keep all the posterior information that I have summarised in the posterior distribution of my parameter.

It is amusing to note that this is not the case with the unbounded squared error loss, since then E(d), the associated expected loss satisfies

$$E(d) = Var(\theta) + (d - E(\theta))^2$$

i.e. E(d) only communicates the posterior mean and variance, so loses most of the information in $F(\theta)$.

It would now be more satisfactory if it were possible to remove the inherent dependence on anxiety function in the above, to obtain a 'depersonalised' representation of my posterior information It was shown in the previous section that $\phi(b)$ synthesises all anxiety function invariant information about the posterior *location* of the parameter θ. In an exactly analogous way

$$\psi(b) = \{E_b(r(b)) : r(b) \in \phi(b)\}$$

where $E_b(d)$ is defined in Section 3.2 can be thought of as an invariant spread map. The next theorem shows that communicating the pair $(\phi(b),\psi(b))$ rather than the posterior mean and variance retains *all* the posterior information in $F(\theta)$. Hence by breaking up the posterior

distribution of θ in this way I isolate the factors associated with location from those associated with spread whilst retaining the whole of the information.

Lemma 3.4.1.

Let $F(\theta)$ be a distribution function $2\times$ differentiable on $\mathbb{R}$ with extended support $S(F)$ and $\bar{F}$ = range of $f(\theta)$. Then

(i) if $A = \{y \in \bar{F}$: there exists an $x \in S(F)$: $f(x) = y$ and $f'(x) = 0\}$ then A^c is dense in $\bar{F}$.

(ii) for all $c \in S(F)$ such that $\begin{cases} f'(c) > 0 \\ f'(c) < 0 \end{cases}$ there is another point $\begin{cases} c^* > c \\ c^* < c \end{cases}$

such that $\begin{cases} f'(c) \leq 0 \\ f'(c) \geq 0 \end{cases}$ and $f(c) = f(c^*)$.

Proof (i)

Let $D = \bigcup_{i\in I} D_i \subset S(F)$ be the set of all points $x \in S(F)$: $f'(x) = 0$

where D_i's $i \in I$ represent the disjoint non-empty closed intervals comprising the set D 3.4.2.

Note that $f(x) = f(y)$ $x,y \in D_i$ since f is differentiable, so let $f_i \in \bar{F}$ be defined by

$$f_i = \{f(x) : x \in D_i\} \quad 3.4.3.$$

Let $C = C^c = \bigcup_{i\in I^*} C_i \subset S(F)$ 3.4.4.

where C_i $i \in I^*$ represent the disjoint non-empty open intervals comprising the set C

Obviously I^* countably infinite $\Rightarrow$ I countably infinite. 3.4.5.

But every non-empty open set contains a rational. Hence I^* is countably infinite by (3.4.3) and (3.4.5), A consists of a countably infinite number of points. But $\bar{F}$ is an interval of non-zero length

Hence A^c is dense in $\bar{F}$ □

Proof (ii)

Suppose $f'(c) > 0$. Then in particular $f(c) > 0$. Let $y > 0$, then there exists an $\varepsilon > 0$ such that $|y| < \varepsilon$ implies

$$f(c+y) > f(c) \qquad 3.4.6.$$

if $f(c+y) > f(c)$ for all $y \in \mathbb{R}_{>0}$ f is not a p.d.f. Hence by Rolles Theorem there is a $y \in \mathbb{R}_{>0}$ such that

$$f(c+y) = f(c).$$

Finally let $y^* = \inf \{y \in \mathbb{R}_{>0} : f(c+y) = f(c)\}$. If $f'(c+y) > 0$ then there is an ε^* such that

$$f(c+y^* - \varepsilon^*) < f(c). \qquad 3.4.7.$$

(3.4.6) and (3.4.7) together with Rolles theorem implies that there is a z $c < z < c+y$ such that

$f(z) = f(c)$ contradicting the definition of y^*

Hence putting $c^* = c+y^*$ gives the result.

An analogous argument proves the result for $f'(c) < 0$ □

Theorem 3.4.

Let $(\phi(b), \psi(b))$ consist of pairs $(r(b), E_b(r(b)))$ where $r(b) \in \phi(b)$. Then $(\phi(b),\psi(b))$ determine their distribution function $F(\theta)$, which is $2 \times$ differentiable on $\mathbb{R}$, uniquely.

Proof

Choose an $r(B) \in \phi(B)$ such that

(i) $f'(r(B)-B) > 0$ for some fixed $B > 0$. 3.4.8.

(ii) $f'(r(B)+B) < 0$

Then since f' is continuous, there is some $\varepsilon > 0$ such that

$f'(x) > 0 \quad x \in A_1$ where $A_1 = (r(B)-B)-\varepsilon, r(B)-B+\varepsilon)$ 3.4.9.

$f'(x) < 0 \quad x \in A_2$ $A_2 = (r(B)+B-\varepsilon, r(b)+B+\varepsilon)$

Now choose a properly unimodal p.d.f. h(x) such that

$$h(x) \propto f(x) \qquad x \in A_1 \cup A_2 \qquad 3.4.10.$$

For h(x), Theorem 3-2(i) states that $\phi_h(B)$ is a *function* of B with derivative $|\phi_h'(B)| < 1$. It follows that there is a $\delta > 0$ such that for any η, $|\eta| < \delta$

$$\phi_h(B+\eta) - (B+\eta) \in A_1 \qquad 3.4.11.$$
$$\phi_h(B+\eta) + (B+\eta) \in A_2$$

So for all $b \in (B-\delta, B+\delta)$ there is an $r(b) \in \phi(b)$ such that $r(b) = \phi_h(b)$.

Hence at B I can define the derivative of r(B) by $\phi_h'(B)$. So I can now differentiate $E_B(r(B))$.

$$-\tfrac{1}{2} E'_B(r(B)) = \tfrac{1}{2}[f(r(B)+B)+f(r(B)-B)+r'(B)(f(r(B)+B)-f(r(B)-B))]$$
$$= f(r(B)+B) = f(r(B)-B). \qquad 3.4.12.$$

since f(r(B)+B) = f(r(B)-B) by definition.

Hence all point $c \in S(F)$ which can be written

$$c = r(b) + b \text{ where } f(c) < 0 \quad f(c-2b) > 0$$
$$\text{or} \quad c = r(b) - b \text{ where } f(c) > 0 \quad f(c+2b) < 0$$

for some $b \in \mathbb{R}_{>0}$ and $r(b) \in \phi(b)$ have f(c) defined by (3.4.12). By combining Lemmas (i) and (ii) it is easily seen that elements of a set of points c S(F) whose range is dense in F has this property. Thus since the p.d.f. $f(\theta)$ is continuous it can be reconstructed using limiting arguments on this dense set.

This completes the proof. □

Note: If $F(\theta)$ is properly unimodal, the only point whose image is not accessible by equation (3.4.12) is the mode.

Hence no information is lost by summarising the posterior distribution $F(\theta)$ by the pair $(\phi(b), \psi(b))$. In fact if there is more than one stationary point of $E_b(d)$ for some value of b, there is duplication of information in this pair. (It is fairly obvious that not *all* the local stationary points of $\phi(b)$ are needed for the construction (3.4.12) in the last Theorem. I feel it may be possible to characterise $F(\theta)$ in terms of

(i) all possible *Bayes* decisions with respect to step loss,

(ii) their associated loss,

but this is for further research.

A partial converse of the above result is possible for properly unimodal distribution functions. By this I mean that from a pair of functions (r(b), q(b)) obeying certain conditions, I can construct a properly unimodal distribution function $F(\theta)$ with associated pair $(\phi(b),\psi'(b) = (r(b),q(b))$

Conditions on r(b)

Clearly by Theorem 1(i) I need that

(1) $|r'(b)| < 1 \quad b \in (0,B)$ where B is possibly infinite

(This is in fact all I need)

Conditions on q(b)

From equation (3.4.12) it was shown that $q(b) = 2f(d+b)$ Hence I certainly need the following conditions.

(2) $q(b) < 0 \quad b \in (0,B)$

(3) $q'(b) > 0 \quad b \in (0,B)$

(4) $\lim_{b\to B} q(b) = 0$ (for the continuity of $f(\theta)$ on $\mathbb{R}$)

(5) $\lim_{b\to 0} q'(b) = 0$ (since $F(\theta)$ must be *properly* unimodal)

(6) $\int_0^B q(b)db = 0$ (by the derivation of (3.4.12)).

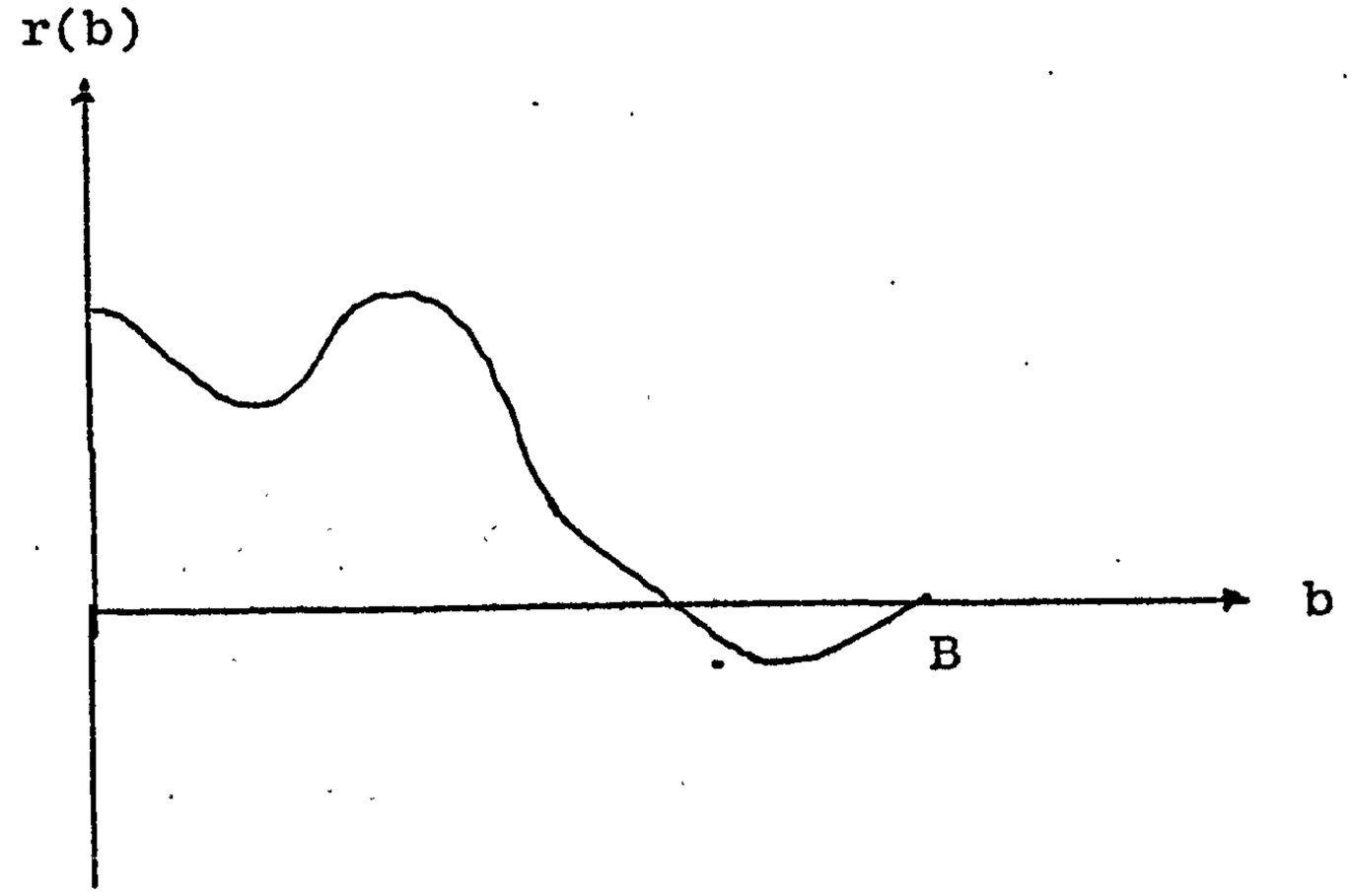
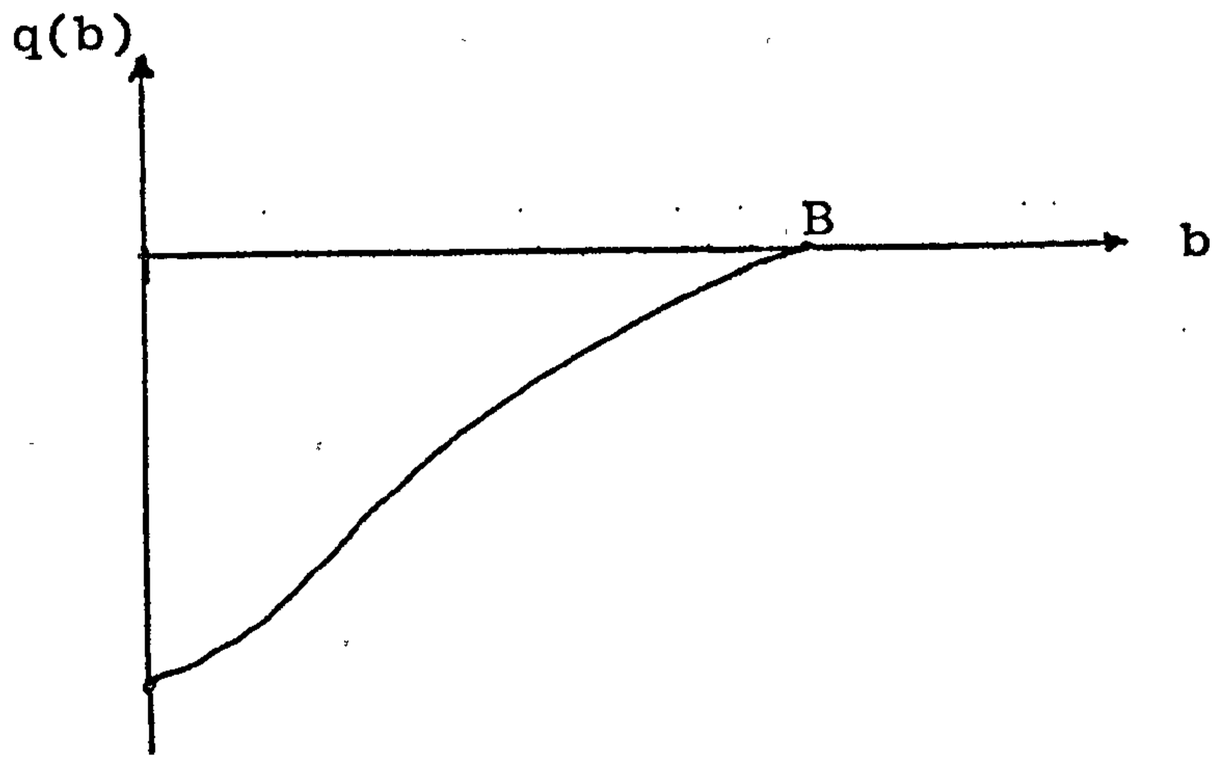


Fig. 3.2.



Theorem 3.5

Suppose the pair r(b) and q(b) satisfy conditions C1 ... C6. Then there exists a unique properly unimodal (distribution function F(θ) 2 × differentiable on $\mathbb{R}$ such that

$$f(r(b)+b) = f(r(b)-b) \quad \begin{cases} -\frac{1}{2}q(b) & b \in [0,B) \\ 0 & \text{otherwise} \end{cases}$$

Proof. The function r(b)+b is differentiable and strictly increasing on (O,B) by C1 and so invertible.

Let $t_1(x) : (r(O), r(B)+B) \to (O,B)$

be defined by $b_1^{-1}(x) = r(b)+b$.

Then $t_1'(x) > O$. $x \in (r(O), r(B)+B)$. 3.4.13.

Let $f(x) = \frac{1}{2}q(t_1(x))$. $x \in (r(O), r(B)+B)$

Then $f(x) > O$ by C2

$\lim_{x \to r(B)+B} f(x) = O$ by C4.

Also $f'(x) = -\frac{1}{2}q'(t_1(x))\, t_1'(x)$

So $\lim_{x \to O} f'(x) = O$ by C5.

$f'(x) < O$ $x \in (r(O), r(B)+B)$ by C4 and 3.4.13.

Similarly, since $r(b)-b$ is strictly decreasing on (O,B), and differentiable by C1, it is also invertible.

Let $t_2: (r(B)-B, r(O)) \to (O,B)$ be defined by

$t_2^{-1}(x) = r(b)+b$.

Then $t_2'(x) < O$. 3.4.14.

Let $f(x) = -\frac{1}{2}q(t_2(x))$ $r(B)-B \leq x \leq O$

Then $f(x) > O$ $x \in (r(B)-B, O)$ by C2

$\lim_{x \to r(B)-B} f(x) = O$ by C4

Also $f'(x) = -\frac{1}{2}q'(t_2(x)\, t_2'(x)$

So $\lim_{x \to O} f'(x) = O$ by C5

$f'(x) > O$ $x \in (r(B)-B, O)$ by C4 and 3.4.14.

Let $f(r(O)) = \lim_{x \downarrow O} f(x)$

$f(x) = O \quad x \in (r(B)-B, r(B)+B)^c$

Thus if f integrates to 1 it will be a differentiable properly unimodal p.d.f. with mode $r(O)$ and $(\phi_f(b), \psi_f(b)) = (r(b), q'(b))$ with extended support $(M_1 M_2)$ where $M_1 = r(B)-B$

$M_2 = r(B)+B$.

Well,

$$-2\int_{M_1}^{M_2} f(x)dx = \int_{M_1}^{O} f(x)dx + \int_{O}^{M_2} f(x)\,dx$$

$$= \int_{M_1}^{O} q(t_1(x)dx + \int_{O}^{M_2} q(t_2(x))dx$$

$$= \int_{O}^{B} q(b)\,(r(b)+b)'\,db - \int_{O}^{B} q(b)(r(b)-b)'db$$

$$= \int_{O}^{B} q(b)\,(r'(b)+1)db - \int_{O}^{B} q(b)(r'(b)-1)db.$$

$$= 2\int_{O}^{B} q(b)db.$$

$$= \pm 2 \quad \text{by C6.}$$

The result follows. □

Thus it is possible to construct a prior for a particular decision maker by getting him to specify his action and its associated loss under these step loss functions which are, as mentioned before, utility invariant. From this approach it is also possible to construct personal utilities. For related work see (Becker et al (1), Davidson et al (1)). However there is now good evidence to suggest that people in fact behave incoherently (Festinger (1) and (2)). Some analytic reasons for this phenomena will be made clearer in a later section.

Of more importance in the statistical setting is the natural way I can use the preceding results in the analysis of Time Series. Recently Bayesian Forecasting techniques (Harrison & Stevens (1)) have been very useful to generate models and formalise processes occurring in Time Series. However, most Time Series models either assume normality of observations or use the moment type of approach of Kalman Filtering (Box-Jenkins (1) Kalman (1)). Working for some time on the problem of generalising the *Bayesian* Forecasting approach to processes other than normal, I found that it is very easy to specify the model in terms of the decision space. This was then shown to be equivalent to a generalisation of the usual sample space model - the equivalence catalysing very interesting and simple results. In particular mixtures of processes (causing "jumps") can easily be formalised in this way. The subject is developed in depth in Chapter 8. The subsequent part of this chapter contains many results useful for such an analysis whist at the same time generating interesting results for time independent processes.

3.5. Examples of Standard $\phi(b)$ functions

The vast proportion of standard distributions are symmetrical and unimodal and the only possible choice of location for these, under some symmetric criteria, would appear to be their mode. The proof of Theorem 1(iv) confirms this conjecture i.e. $\phi(b) = \mu$ for all $b \in \mathbb{R}_{>0}$

where μ is the mode.

Also if a p.d.f. $f(\theta)$ is strictly decreasing on its extended support $S(F)$ it is trivial to show (by working directly from $E_b(d)$) that

$$\phi(b) = \mu + b \quad \text{where } \mu = \inf S(F)$$

This means, for example, that

$$\phi(b) = b \qquad \text{for the exponential distribution}$$

$$f(\theta) \propto \exp(-\Lambda\theta)$$

and $\phi(b) = \theta_o + b$ for the Paneto distribution where the Paneto is given in the form $f(\theta) \propto \begin{cases} \theta^{-(\alpha+1)} & \theta > \theta_o \\ 0 & \text{otherwise} \end{cases}$

where $\theta_o > 0$.

Table 1 lists some of the more common properly unimodal distribution functions under common parametrisations, and their corresponding $\phi(b)$ functions. (The working for these examples is omitted since it is like finding the mode of a p.d.f., i.e. elementary manipulation).

Sometimes explicit solutions are not possible, but this is not very much of a disadvantage since numerical solutions can easily be worked out using a pocket calculator, and the corresponding intervals found.

Table 1. $\phi(b)$ for some standard distributions.

Name of Distribution	Form of associated p.d.f. f(x)	$\phi(b)$ function	mode = m
Gamma Family			
1) Gamma	$f(x) \propto \begin{cases} x^{\alpha-1}\exp-\beta x, & x \in \mathbb{R}_{>0} \\ 0 & \text{otherwise} \end{cases}$ $\alpha>1, \beta>0$	$b \operatorname{Coth}(b/m)$	$\frac{\alpha-1}{\beta}$
2) Inverted Gamma	$f(x) \propto \begin{cases} x^{\alpha+1}\exp-\beta x^{-1}, & x \in \mathbb{R}_{>0} \\ & \text{otherwise} \end{cases}$ $\alpha, \beta>0$	$b+R^{-1}(2m/b)$ where $R(t)=\ell n(1+2t^{-1})t(2+t)$	$\frac{\beta}{\alpha+1}$
3) Log-Gamma	$f(x) \begin{cases} \exp(-\beta \exp x+\alpha x) & x \in \mathbb{R} \\ & \alpha, \beta>0 \end{cases}$	$m + \ell n \left(\frac{b}{\operatorname{Sinh} b}\right)$	$\ell n(\alpha/\beta)$
Beta Family			
4) Beta	$f(x) \propto \begin{cases} x^{\alpha-1}(1-x)^{\beta-1} & 0<x<1 \\ 0 & \text{otherwise} \end{cases}$ $\alpha, \beta>1.$	$\phi(b)$ satisfies $\ell n\left(\frac{\phi(b)+b}{\phi(b)-b}\right) = (m^{-1}-1)\ell n\left(\frac{1-\phi(b)-b}{1-\phi(b)+b}\right)$	$(1+\frac{\beta-1}{\alpha-1})^{-1}$
5) Logistic Beta	$f(x) \propto \exp \alpha x(1+\exp x)^{\alpha+\beta}$ $\alpha, \beta>0$	$\ell n\left[\frac{\exp 2rb - 1}{\exp b-\exp(2r-1)b}\right]$ where $r=\frac{\exp m}{1+\exp m}$	$\ell n(\alpha/\beta)$
Others			
6) Lognormal	$f(x) \propto \begin{cases} x^{-1}\exp -\frac{1}{2}(\gamma+\delta \ell n x)^2 & x \in \mathbb{R}_{>0} \\ 0 & \text{otherwise} \end{cases}$	$(m^2 + b^2)^{\frac{1}{2}}$	$\exp[-(\gamma\delta^{-1}+\delta^{-2})$
7) Log-F.	$f(x) \propto \left(\frac{\exp\{(1+\beta/\alpha)^{-1}x\}}{1+\beta/\alpha \exp x}\right)^{\alpha/2}(1+\beta/\alpha)$ $x \in \mathbb{R}$ $\alpha, \beta>0.$	$\frac{m}{2} - \ell n[\operatorname{Sinh} b)(\exp\{2(1+\exp -m/2)b-1\})^{-1} - \exp -b]$	$2 \ell n(\alpha/\beta)$

Variance Estimation for a normal distribution

Squared-error admissibility was in vogue a while ago before James and Stein (1) showed some rather paradoxical results using the concept as a criteria for judgement. To add wood to the fire, it has long been known that the M.V.U.E. S^2 of variance V of a normal distribution is squared-error inadmissible. But:

(i) The sensible parametrisation for V is ℓn V (as shown in Chapter 2) and with a "vague" prior of conjugate form the decision region for all symmetric loss functions is

$$V \in (\frac{n-1}{n} S^2, \infty) \text{ (transform from Table 1)}$$

which certainly contains the M.V.U.E.

(ii) If I have a symmetric loss function on the variance itself, the decision region is

$$V \in (\frac{n-1}{n+1} S^2, \infty)$$

which again contains the M.V.U.E.

In fact it could be argued that under this Bayes criteria that S^2 is a better estimate than the M.L.E. which lies on the boundary of region (i) and certainly better for estimates of the form

$\frac{n-1}{n+k} S^2$ where $1 < k < 4$, which have been proposed as alternatives to S^2. Of course prior information usually will not be vague so these results must be taken with a pinch of salt, but I think that the example illustrates how misconceived the concept of inadmissibility is, and how it can discard out of hand estimations which are quite acceptable.

Comparison of lognormal and gamma distributions

It has been well-known that under certain situations, gamma distributions can be approximated by the lognormal. Bartlett et al (1) and this has been used widely in Bayesian statistics (e.g.Leonard(3)). It is a moot point however how this approximation should be made.

In Fig 3.3, I have graphs of the functions $\phi(b)$ of the Gamma and lognormal when they have the same mode. Differences in estimates under particular loss structures can then be compared. The estimates will be most notably different when $b \simeq 2.13m$ where m is the mode

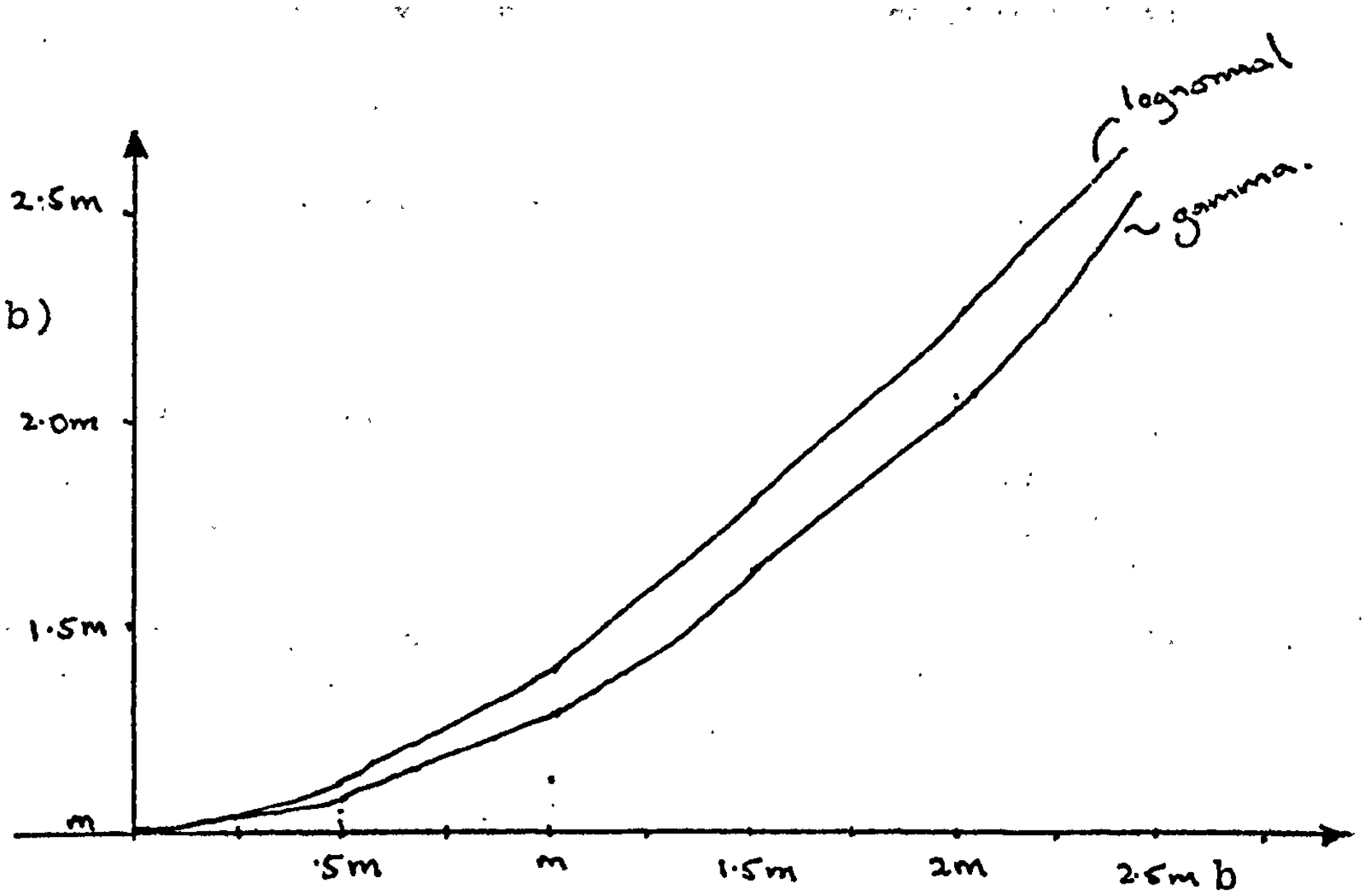


Fig. 3.3.

The above comparison suggests that in a Bayesian estimation situation, an approximation minimising the distance between these two $\phi(b)$ functions could be utilised.

3.6. Expanding families of distribution functions

The first thing that an observant reader will see is that in many cases the mode is intrinsic to the generalised location map for many of the standard families. The following theorem gives the reason for this.

Theorem 3.6.

Suppose $f_1(\theta), f_2(\theta)$ are p.d.f.s of θ. Then

$$\phi_{f_1}(b) = \phi_{f_2}(b) \quad \text{for all } b \in (0, \tfrac{1}{2}(k_2-k_1))$$

if and only if there exists a 1-1 transformation T such that

$$f_2(\theta) = T(f_1(\theta)).$$

where notation is borrowed from Chapter 8

Proof

The sufficiency is obvious since it is then clear that

$$f_1(d+b) = f_1(d-b) \iff f_2(d+b) = f_2(d-b).$$

Conversely if T is not monotonic then there exist points S_1, S_2 such that

$$f_1(S_1) \neq f_1(S_2) \qquad \text{and } S_1 < S_2 \text{ (say).}$$
$$f_2(S_1) = f_2(S_2)$$

or visa-versa. Putting $b = \tfrac{1}{2}(S_2-S_1)$ and $d = \tfrac{1}{2}(S_1 + S_2)$, this becomes

$$f_1(d-b) \neq f_1(d+b)$$
$$f_2(d-b) = f_2(d+b)$$
or $$\phi_{f_1}(b) \neq \phi_{f_2}(b)$$

Hence the theorem is proved. □

Notes

1) For all symmetric unimodal pairs of p.d.f.s $(f_1(\theta), f_2(\theta))$ there is always such a 1-1 transformation T linking them.

2) Typical examples of the equivalences are the following:

(i) $f_1(\theta) \propto \exp(-h(\theta))$

$f_2(\theta) \propto (1 + ch(\theta))^{-\nu}$

where $h(\theta) > 0$ for all $\theta \in S(F_1)$ and $\nu, c > 0$

Since $f_2 = (1 - c\ \ell n f_1)^{-\nu}$ which is a 1-1 transformation

(ii) $f_2(\theta) \propto f_1^{\nu}(\theta)$ where $\nu > 0$.

The above theorem provokes the following definitions.

Definitions

Call $F(\underset{\sim}{\alpha}) = \{f(\theta|\underset{\sim}{\alpha}): \underset{\sim}{\alpha} \in A\}$ an *expanding family* of p.d.f.s if for every $f^* \in F(\underset{\sim}{\alpha})$ there exists a subset

$F_{f^*} \subset F$ such that:

(i) For every $f \in F_{f^*}$,

$\phi_f(b) = \phi_{f^*}(b)$ for all $b \in (0, \frac{1}{2}(k_2 - k_1))$.

(ii) There exists a reparametrisation R of the family $F(\alpha)$

$R: A \to B = (B_1, B_2, \ldots B_n)$

$\underset{\sim}{\alpha} \to \underset{\sim}{\beta} = (\beta_1, \ldots\ldots \beta_n)$ such that

$F_{f^*} \supseteq \{f(\theta|\underset{\sim}{\beta}): \beta_1 \in B_1\}$.

Call a family $F(\underset{\sim}{\alpha})$ a *linear expanding family* if for every

$f^* \in F(\underset{\sim}{\alpha})$ $\quad F_{f^*} \supseteq P_{f^*}$

where $P_{f^*} = \{f(\theta): f(\theta) \propto (f^*(\theta))^r,\ r \in \mathbb{R}_{>0}\}$.

Examples. The following examples can be checked from Table 1:

Linear expanding p.d.f.s

Normal

t-distribution with unknown degrees of freedom

Gamma

Inverted-Gamma

Beta.

Expanding families

All unimodal symmetric families with one "spread" variable

Log-normal

Log-F

Inverse-Logistic transformation of the Beta.

Obviously not all families are expanding, for example the usual parametrisation of the F distribution is not.

The definition of expanding families is used in the expression of increased uncertainty with time which is explained in Chapter 8.

3.7. A link with the prior likelihood approach.

Suppose instead of a Bayesian approach in which I end up with posterior p.d.f. $f(\theta)$, I use a prior likelihood approach (Edwards (1)). Then the prior likelihood (assumed of the same form as the prior p.d.f) and sample likelihood are combined in exactly the same way to give a posterior likelihood function of θ

$$\ell(\theta) \propto f(\theta).$$

Then, because $\phi_f(b)$ does not use the measure underlying the Bayes posterior (i.e. I am not integrating $f(\theta)$) it could equally be considered as a map of $\ell(\theta)$. Consider $\phi_f(b)$ in this light for a moment.

Firstly the intervals obtained from $\phi_f(b)$ are measure invariant. Also Theorem 3.6 can be restated in the following way.

If for every pair $(\theta_1,\theta_2) \in S(F)$ the pair of posterior likelihoods (ℓ_1,ℓ_2) has the property

$$\ell_1(\theta_1) \leq \ell_1(\theta_2) \iff \ell_2(\theta_1) \leq \ell_2(\theta_2).$$

then ℓ_1 and ℓ_2 have the same "location" or $\phi_f(b)$ function. In fact I could choose to define $\phi(b)$ functions in terms of these equivalence classes. In a sense this definition would be more appealing because of the invariance to the transformation of the parameter. Certainly it emphasises the strong similarity between proper Bayesian analysis and likelihood analysis which just does not come over if I immediately turn to the conventional loss functions.

I can summarise the sentiment in the following:

(i) If I only have information about the ordering induced by the posterior likelihood then the only invariate "location" variables I can communicate are the stationary points of the likelihood (which of course include the M.L.E).

(ii) If in addition I have a "natural" (up to linear transformations) scaling on θ (induced for example by some loss structure), then I have all the information I need for making location estimates on θ in terms of the $\phi(b)$ function.

(iii) It is not until I need to guage how "good" my actual estimates are ($\psi(b)$) that I need to integrate i.e. use a measure. This is when I need the actual *numerical* values of the posterior likelihood/distribution (up to linear transformations) rather than just the ordering induced by them.

I shall now return to the main theme.

3.8. On to Bimodal distributions - The symmetric product

Bimodal distributions are hardly touched upon in most statistical inference, often because they tend to cause acute embarrassment for many criteria of judgement (e.g. m.l.e. squared-error loss, inadmissibility) being used today. Hence there is a tendency to assume that they do not exist, sufficiency and other concepts are created to cloud the issue.

The following lemma and theorem give the reason for the importance of a sample mean and a good interpretation of it.

Lemma 3.7.1.

Suppose $X_1 \ldots X_n$ are independent random variables with differentiable, unimodal, symmetric sampling p.d.f.'s $f_i(x|\theta)$ where θ is a location parameter, and $1 \leq i \leq n$.

Then if a uniform prior on $[m_1-k, m_2+k]$ is used for θ

where $m_1 = \min_{1\leq i\leq n} x_i$ $\qquad$ $m_2 = \max_{1\leq i\leq n} x_i$

any local minima of expected loss with respect to step loss S_b lies in $[m_1,m_2]$, $b \leq \frac{1}{2}(m_2-m_1+2k)$

Proof Any local minima d has to satisfy the equation

$$\sum_{i=1}^{n} (\ell n f_i(d-x_i-b) - \ell n f_i(d-x_i+b)) = 0 \quad b < \tfrac{1}{2}(m_2-m_1+2k)$$

Suppose $d > m_2$. Then for $1 \leq i \leq n$ $\quad d-x_i > 0$ so

$$\ell n f_i(d-x_i-b) - \ell n f_i(d-x_i+b) > 0 \quad 1 \leq i \leq n$$

since each f_i is symmetric and unimodal.

Hence adding the equations

$$\sum_{i=1}^{n} \ell n\ f_i(d-x_i-b) - \ell n\ f_i(d-x_i+b) > 0$$

Hence d cannot be a local minima of expected loss. A similar argument proves the case for $d < m_1$. □

This tees up to the following theorem.

Theorem 3.7. Suppose $X_1 \ldots X_n$ are independent random variables where X_i has sampling p.d.f. $f_i(x|\theta)$ on $\mathbb{R}$ $1 \le i \le n$, $f_i(x|\theta)$ is symmetric and unimodal $1 \le i \le n$, and where θ is a location parameter for each of the X_i's.

If (i) $I_i(\theta-x) = \ell n\, f_i(x|\theta)$ is $1 \times$ differentiable with respect to θ everywhere on $\mathbb{R}$

(ii) For all $M_1 < M_2 \in \mathbb{R}$

$$r_i(x) = \frac{I_i'(x+k_i|\theta)}{\sum_{j=1}^{n} I_j'(x+k_j|\theta)} \to R_i \text{ as } x \to \infty \text{ for all } k_1, k_2 \in [M_1, M_2]$$

(iii) a uniform prior on $[-2b,2b]$ on θ is used, then $\phi(b)$ satisfies

$$\phi(b) \to \sum_{i=1}^{n} R_i x_i \text{ as } b \to \infty$$

Note: From the definition $\sum_{i=1}^{n} R_i = 1$.].

Proof $d = \phi(b)$ must satisfy, by the symmetry of I,

$$\sum_{i=1}^{n} I(b-y_i) - \sum_{i=M}^{n} I(b+y_i) = 0 \qquad 3.8.1.$$

where $y_i = d-x_i$. By Lemma 7.1 it was shown that

$$\min_{1\le i\le n} x_i \le d \le \max_{1\le i\le n} x_i \quad \text{so}$$

$$|y_i| \le m \quad \text{where } m = \max_{1\le i\le n} x_i - \min_{1\le i\le n} x_i \qquad 3.8.2.$$

By the first mean value theorem

$$\sum_{i=1}^{n} I_i(b-y_i) - \sum_{i=1}^{n} I_i(b+y_i) = \sum_{i=1}^{n} 2y_i\, I_i'(b+\theta_i y_i) \qquad 3.8.3.$$

where $|\theta_i| < 1$ $1\le i\le n$

Note that $\sum_{i=1}^{n} I_i'(b+\theta_i y_i) \quad 0 > b \quad m$ since each f_i is properly unimodal 3.8.4.

So dividing (3.8.3) by (3.8.4) and implementing equation (3.8.1) d must satisfy.

$$\sum_{i=1}^{n} 2y_i \, I_i'(b+\theta_i y_i) \left(\sum_{i=1}^{n} I_j'(b+\theta_j y_j) \right)^{-1} = 0$$

But by using condition (ii) and statement (3.8.2) this becomes

$$\sum_{i=1}^{n} (x_i - d) \, R_i + t(b) = 0 \text{ where } t(b) \to 0 \text{ as } b \to \infty .$$

Hence $d = \phi(b) \to \sum_{i=1}^{n} R_i x_i$ as $b \to \infty$ □

The importance of the result of course hinges on how often condition (ii) is met and the value of the limit in particular cases. The following corollories give two important examples.

Corollary 3.7.1. If $f_i(x|\theta) = f(x|\theta) \quad i = 1...n$ then if

(i) I is a polynomial or

(ii) f is an inverse polynomial

then $\phi(b) \to \bar{x}$ as $b \to \infty$.

Proof It is easy to check that condition (ii) of Theorem 3.7 is satisfied by showing

$$\frac{I'(x+k_1|\theta)}{I'(x+k_2|\theta)} \to 1 \text{ as } x \to \infty .$$

[In both cases this is a ratio of 2 polynomials the highest order top and bottom terms dominating]

It follows that $R_i = 1/n \; 1 \leq i \leq n$ in both (i) and (ii) and hence that

$$\phi(b) \to \bar{x}$$

□

Corollary 3.7.2.

If $f_i(x|\theta) \propto f(H_i(x-\theta))$ $\quad i = 1...n$, then

(i) If $I = \ell n\, f(x|\theta)$ is a polynomial of degree $m, R_i = H_i^m \left(\sum_{j=1}^{n} H_j^m\right)^{-1}$

$1 \le i \le n$

(ii) If f is an inverse polynomial of degree $m, R_i = 1/n$ $\quad 1 \le i \le n$.

Proof

Using the result in Corollary 7.1 it is sufficient to prove in case

(i) $\dfrac{HI'(Hx)}{I'(x)} \to H^m$ as $x \to \infty$ $\quad$ where $I = \sum_{i=0}^{m} I_i x_i$

in case (ii)

$$\frac{HI'(Hx)}{I'(x)} = \frac{Hg'(Hx)}{g(Hx)} \cdot \frac{g(x)}{g'(x)} = 1 \quad \text{as } x \to \infty$$

where $f^{-1}(x) = g(x) = \sum_{i=0}^{m} g_i x_i$.

This is again easily seen since highest terms in each expansion dominate

[In (i) $H \dfrac{I'(Hx)}{I'(x)} \to \dfrac{H\ Im\ H^{m-1}\ x^{m-1}}{m.\ x^{m-1}} = H^m$

In (ii) $H\dfrac{g'(Hx)}{g'(x)} \quad \dfrac{g(x)}{g(Hx)} \to \dfrac{Hm\ g_m(Hx)^{m-1}\ g_m x^m}{m\ x^{m-1}\ g_m\ g_m(Hx)^m} = 1$] □

Corollary 3.7.3. If $f_1(x|\theta)$ (defined above) has a tail steeper in degree than $f_i(x|\theta)$ $\quad z \le i \le n$, in cases (i) and (ii) of Corollary 3.7.2.

$$\phi(b) \to x_1 \text{ as } b \to \infty .$$

Proof. Use the dominance of highest terms as in Corollary 3.7.1, but on $r_i(x)$ (defined in Theorem 3.7) directly. □

It can now be seen how the sample mean can be an important quantity even when bounded loss functions are used. When $X_1....X_n$ are independent identically distributed random variables each with symmetric unimodal sampling p.d.f.s and a vague prior is used, the holy $\phi(b)$ function tends to $\bar{x}$ as long as the tails of the sampling p.d.f.'s are not too steep. (See Corollary 7.1). Note, however, that $\bar{x}$ is more of a landmark than a decision.

Often it is found that $\phi(b)$ is an increasing fucntion of b. In this case (and many others) all Bayes decisions under symmetric bounded loss must lie in $(m,\bar{x})$ where m is the highest posterior mode. So $\bar{x}$ need not be a bounded-Bayes decision at all, it often marks an end point of an interval outside which estimates are poor, unless of course it is sensible to use a strong prior distribution. Note $\bar{x}$ itself is outside this interval.

The effects on posterior decisions of different forms of prior distribution is for further research. However in the final Corollary 7.3 it is seen that the value of the random variable X which has the steepest tail dominates the decision when a vague prior is used so that

$$\phi(b) \to x \qquad \text{as } b \to \infty \ .$$

In for example David [1], he shows that the analysis of the posterior distribution of a location parameter θ using a prior p.d.f. $f_o(\theta-\mu)$ with location μ is equivalent to the analysis using a uniform prior on θ and assuming an extra random variable X_o with p.d.f. $f_o(x-\theta)$ has taken an observed value μ. So by the above comment, if the tail of f_o is too steep,

$$\phi(b) \to \mu \qquad \text{as } b \to \infty \ .$$

This may be sensible or it may not, depending on the strength of particular pieces of information. Conversely if the tail of the prior is less steep than that of one of the sampling distributions of the random variables being observed, the prior location gets ignored as $b \to \infty$. This in fact, is a special case of the analysis in the above reference.

Notice that using the notation above, if $f_o = f$ and conditions of Corollary 7.1 hold, then

$$\phi(b) \to \frac{n\bar{x}+\mu}{n+1} \qquad \text{as } b \to \infty \ .$$

Corollary 7.2 emphasises the difference in asympototic behaviour between distributions in the exponential family and those distributions like the t with inverse polynomial tails (note in particular the difference between the tail behaviour of $\phi(b)$ for the normal distribution and for the t-distribution with high degrees of freedom).

The author feels that much of what is done using asymptotic approximations, characteristic functions etcetera, uses properties of the sampling distribution in their *tails*. In most cases it is unethical to assume that these are known precisely. Clearly $\bar{x}$ often crops up as a limit because of the above intimate link with the tails as shown above.

3.9. Convex Utilities (Assume L1, L2 and L3 hold)

It is natural to ask (following comments in 3.1) whether fuller use of a particular loss structure can be used to find where its corresponding Bayes estimates might lie. To answer this question one first has to consider the types of utilities that might be employed by a decision maker making an estimate. One that springs to mind might be the convex (inward) utility function which is pictured in Fig. 3.4.

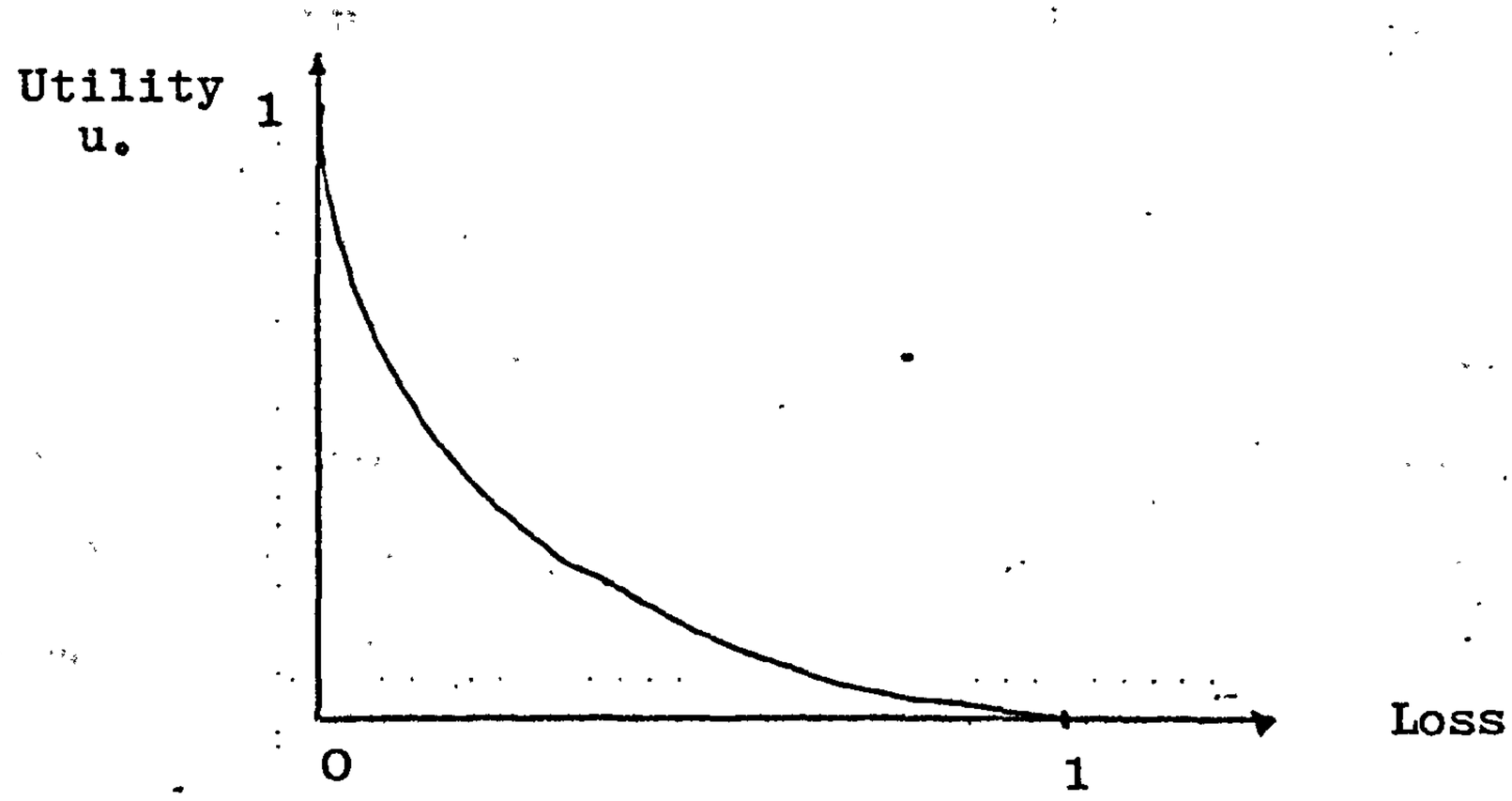


Fig. 3.4.

i.e. those utilities u such that

(i) u is twice differentiable

(ii) $u''(L) \geq 0$ for all $\ell \in [0,1]$.

This is obviously an optimistic class of utilities and constrains the decision maker never to "cut his losses" for example. Each utility in the class takes at least as much note of small errors in estimation as it does large.

Assuming this type of utility therefore, a more restrictive class of loss functions can be considered, the "convex inward" systems loss function shown in Fig. 3.5.

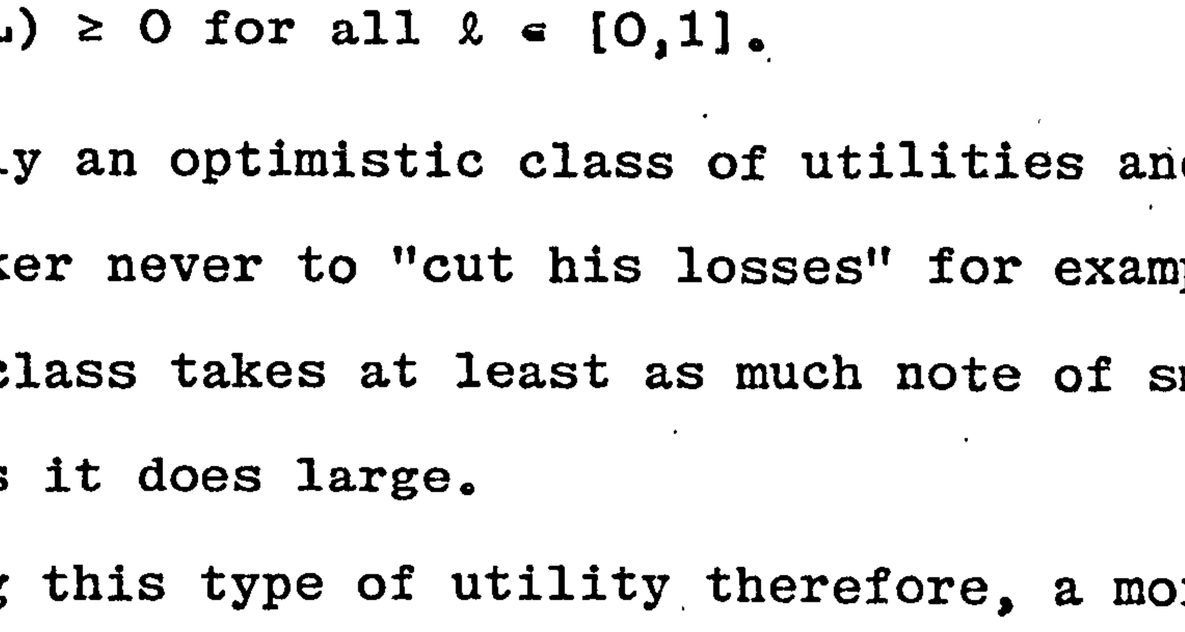


Fig. 3.5.

A typical example of such a loss function that the author has seen used is

$$1 - \exp-|\, d - \theta|.$$

As before the class of corresponding expected utilities for a fixed posterior distribution under these two restrictions can be summarised in terms of a linear utility u, and "convex inward" loss function L*, by defining

$$L^*(s) = (1 - u(L(s))$$

where U and L are the original utility and loss function respectively. So I lose nothing in the analysis by forgetting the utility function and just working with expected loss.

As in the first part of the Chapter a "basis" for this class can now be constructed. This is done with what are called ramp loss functions.

Definition

The *Generalised Ramp loss function* $R_{b,c}(s)$ is defined by

$$R_{b,c}(s) = \begin{cases} 0 & |s| < b \\ K(|s| - b) & b \le |s| \le c \\ 1 & c < |s| \end{cases}$$

where $s = d-\theta$

$k = (c-b)^{-1}$

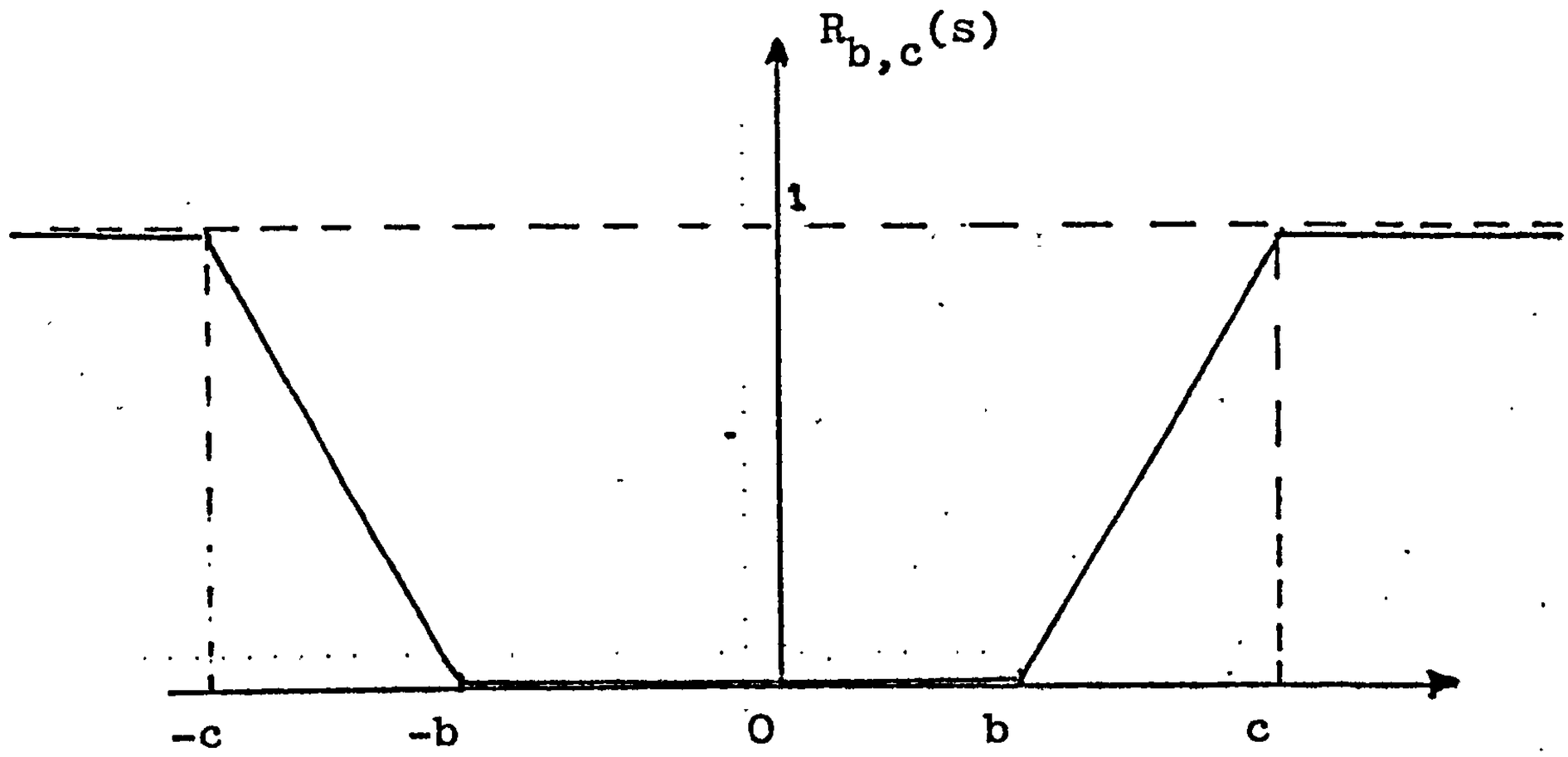


Fig. 3.6.

Theorem 3.8.

The expected loss function $E_R(d)$ with respect to loss function $R_{b,c}(\theta-d)$ and differentiable posterior distribution $F(\theta)$, has minima all satisfying the equation

$$F(d+c)-F(d+b) = F(d-b) - F(d-c).$$

Proof

$$E_R(d) = 1-F(d+c) + F(d-c) + k\left(\int_{-c}^{-b} -(s+b)f(s+d)ds + \int_{b}^{c}(s-b)f(s+d)ds\right)$$

which on rearranging and changing the arguments of the integrals becomes

$$= 1 - F(d+c) + F(d-c) + k\int_{0}^{c-b} s(f(d+s+b)-f(d-s-b))ds$$

Integrating by parts and rearranging this becomes:

$$= 1 - k\int_{0}^{c-b}(F(d+s+b) + F(d-s-b))ds.$$

Hence the stationary points are given by

$$\int_{b}^{c}(f(d+s) + f(d-s))ds = 0$$

The result follows. □

Definition

Define the *simple ramp* loss function (or simply the *ramp* loss function) to be the generalised ramp loss function with b = 0.

The minima of expected loss then satisfy the simpler equation

$$F(d^*+c) - F(d^*) = F(d^*) - F(d^*-c) \qquad 3.9.1.$$

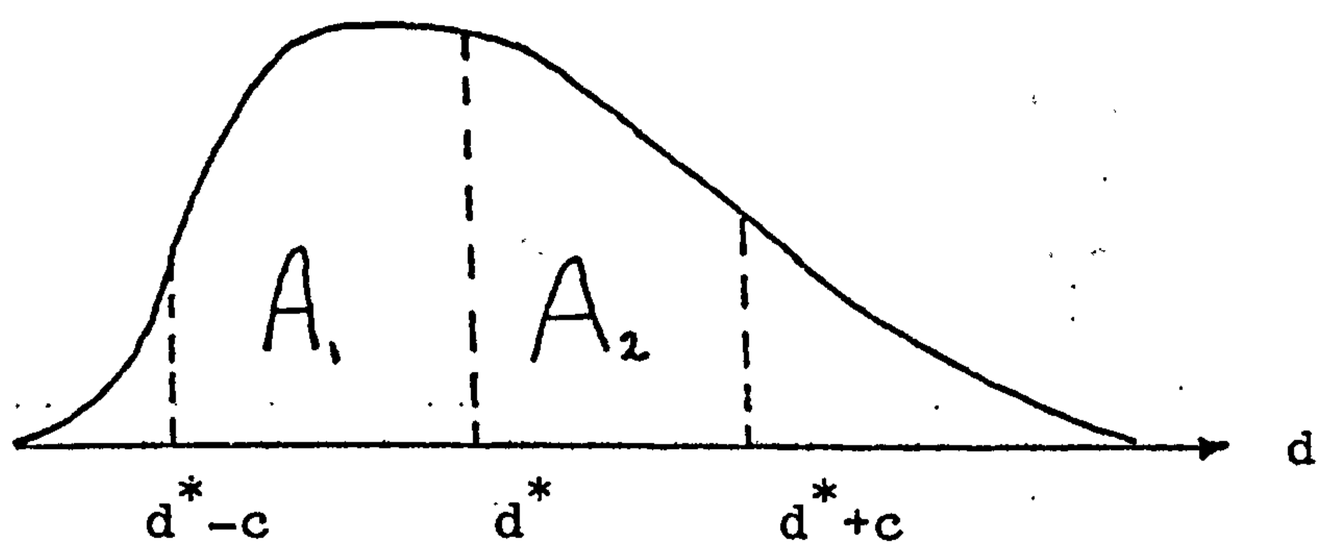


Fig. 3.7. Note that the local stationary points of $E_R(d)$ are given when Area A_1 = Area A_2.

Now the basis Lemma. Let $E_{R(b)}(d)$ denote the expected loss with respect to the simple ramp loss function $R_b(\theta-d)$

Lemma 3.9.1.

Suppose $L''(s) \leq 0$ a.s. Then $E_L(d)$ can be expressed in the form

$$E_L(d) = \int_{\mathbb{R}_{>0}} E_{R(b)}(d)\ \alpha(b)db \qquad \text{where } \alpha(b) \text{ is a p.d.f.}$$

Proof. It has been shown in Lemma 3.3.1. that in general

$$E_L(d) = \int_{\mathbb{R}_{>0}} E_b(d)\ L'(b)db \quad \text{where } L'(b) \text{ is a p.d.f.} \qquad 3.9.2.$$

Also $\int_0^b E_c(d)dc = \int_0^b [1-F(d+c) - F(d-c)]dc$ which on integrating by parts gives

$$= bE_b(d) + \int_{-b}^{b} |c| dF(d+c) \qquad 3.9.3.$$

$$= bE_{R(b)}(d).$$

Hence integrating (3.9.2) by parts and using (3.9.3) I get

$$E_L(d) = b\, E_{R(b)}(d)\, L'(b)]_{0-}^{\infty} \int_{\mathbb{R}_{>0}} E_{R(b)}(d) bL''(b)db.$$

Since $L'(b)$ is a p.d.f. $bL'(b) \to 0$ when $b \to 0$ or $b \to \infty$, so

$$E_L(d) = \int_{\mathbb{R}_{>0}} E_{R(b)}(d)\, \alpha(b)db$$

where $\alpha(b) = -b\, L''(b)$.

It is easy to check that $(b) \geq 0$ and $\int_{\mathbb{R}_{>0}} \alpha(b)db = 1$.

The result now follows. □

Let $\phi_R(b)$ = set of stationary points of $E_{R(b)}(d)$ for $d \in S$, where S is defined in the previous section.

The following Theorem ensues.

Theorem 3.9.

Suppose L(s) is bounded by 1 symmetric and 2 × differentiable a's such that

(i) $L'(s)$ is decreasing $s > 0$

(ii) $L(b) = 1$ $b > k$ (where k could be ∞)

Then all Bayes decisions with respect to L lie in the interval $[d_1, d_2]$

where $d_1 = \inf \{\phi_R(0,k)\}$

$d_2 = \sup \{\phi_R(0,k)\}$

Proof. Using Lemma 3.9.1. I can substitute $E_R(b)$ for E(b) in Theorem 3.3 and the argument can be reproduced exactly (Note that I have no end point difficulties this time). □

The problem now is that the $\phi_R(b)$ functions are a little bit more difficult to find explicitly, since they are functions of the distribution function (see Lemma (3.9.1) rather than the p.d.f (as $\phi(b)$ was). Most commonly used distributions are represented in p.d.f. form, so the set $\phi_R(0,k]$ may be difficult to obtain explicitly.

This next theorem partially removes these difficulties.

Theorem 3.10

Let $F(\theta)$ be a properly unimodal distribution with $\theta_f(b)$ a non-decreasing function of b. Let

$\mathcal{L} = \{L_b(s) : b \in \mathbb{R}_{>0}\}$ be some family of bounded symmetric loss functions such that

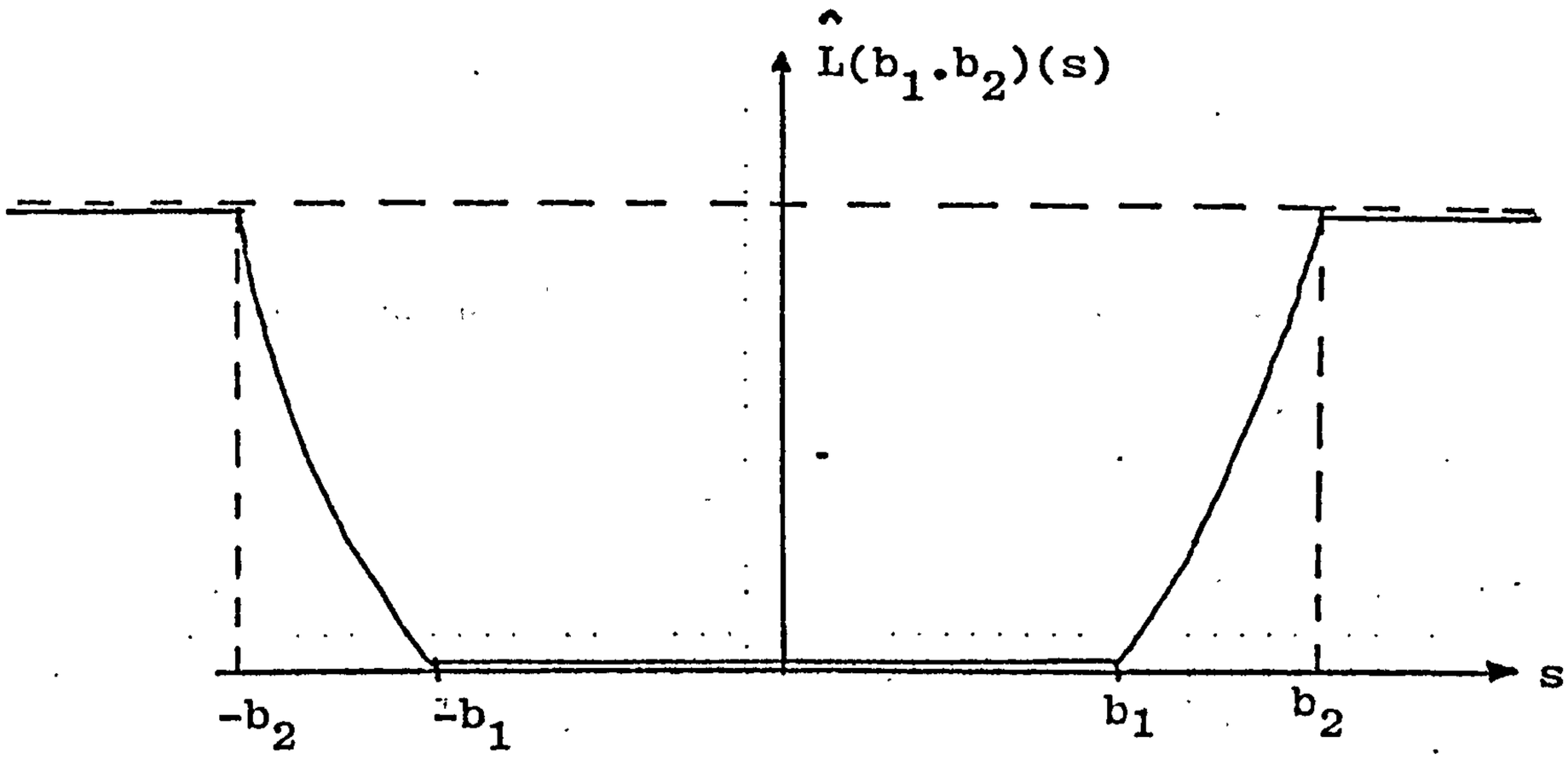
$$L_b(s) = \begin{cases} k(b)\, L^*(s) & L(s) < c(b) \\ c(b) & \text{otherwise} \end{cases}$$

where c is strictly increasing in b, k(b) a function of b only and $L^*(s)$ some symmetric loss function (not necessarily bounded).

Then $\inf_{d \in I} [E(L_b, F, d)]$

where I is any closed interval containing a local minima s increasing in b.

Proof. First note that without loss of generality I can assume that k(b) = 1 since k(b) will not affect the positions of the minima of $E(L_b, F, d)$.



Suppose $0 < b_1 < b_2$. Then

$$L_{b_2}(s) = L_{b_1}(s) + \hat{L}(b_1,b_2)(s)$$

where $\hat{L}(b_1,b_2)(s) = L_{b_2}(s) - L_{b_1}(s)$ is itself a bounded loss function pictured above.

$$\text{So } E(L_{b_2},F,d) = E(L_{b_1},F,d) + E(\hat{L}(b_1,b_2),F,d)$$

By Theorem 3.9 and using the fact that $\phi_f(b)$ is non-decreasing, $E(L_{b_1},F,d)$ has all its stationary points in S in

$$(\phi_f(0),\ \phi_f(b_1))$$

and $E(\hat{L}(b_1,b_2),F,d)$ has all its stationary points in S in

$$(\phi_f(b_1),\ \phi_f(b_2))$$

So in particular $E(\hat{L}(b_1,b_2),F,d)$ will be decreasing on $(\phi_f(0),\phi_f(b))$.

The result follows □

Note also that the same argument can be used to prove that all local maxima are decreasing in b

Corollary

Under the conditions of the theorem, the Bayes decision is increasing in b □

Note

Just because f is unimodal does not mean that E(L,F,d) is going to have one minima for all bounded symmetric loss functions L(s). Exact conditions for when this is the case are given in Chapter 6.

The trajectory of the Bayes decision could very well look like Fig. 3.8.

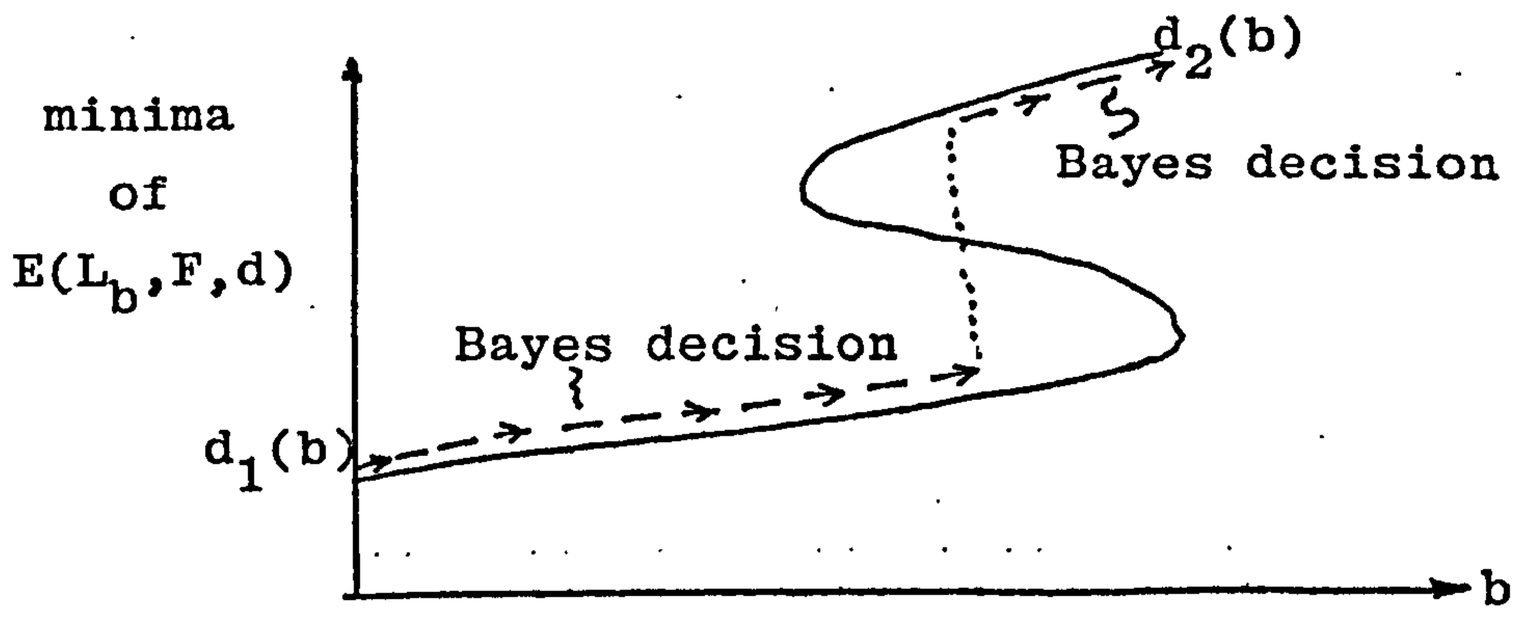


Fig. 3.8.

The reason I have included the previous theorem so soon is that it gives rise to the next corollary for which I need the following Theorem.

Theorem 3.11.

Suppose M is the median and in the mode of a properly unimodal distribution function F. Then

(i) $\max \{\phi_R(b)\} \to M$ as $b \to \infty$

(ii) $\min \{\phi_R(b)\} \to m$ as $b \to 0$

Proof Without loss of generality suppose m = 0.

(i) $\phi_R(b)$ is the set of points given by the equation

$$F(\phi_R(b)+b) - F(\phi_R(b)-b) - 2F(\phi_R(b)) = 0 \qquad 3.9.4.$$

which can be rewritten as

$$1 - 2F(\phi_R(b)) = R(b) \quad \text{where} \qquad 3.9.5.$$

$$R(b) = 1-F(\phi_R(b)+b)-F(\phi_R(b)-b).$$

Suppose there exists a sequence $(\phi_R(b))$ in $\{\phi_R(b)\}$ such that

$$\phi_R(b) \to \infty \quad \text{as } b \to \infty$$

Then $F(\phi_R(b)+b) - F(\phi_R(b)) \to 0$ as $b \to \infty$ which by (3.9.4) gives

$$F(\phi_R(b)-b) - F(\phi_R(b) \to 0$$

ie. $F(\phi_R(b)-b) \to 1$

Hence for all $b > B$(say) $\phi_R(b)-b > 0$ $\Rightarrow \Leftarrow$

(since then $F(\phi_R(b)+b)-F(\phi_R(b)) < F(\phi_R(b)) - F(\phi_R(b)-b)$.

An analogous argument shows that there is no sequence in $\{\phi_e(b)\}$ such that $\phi_R(b) \to -\infty$ as $b \to \infty$.

So for all b $|\{\phi_R(b)\}| \leq M$. Hence $R(b) \to 0$ as $b \to \infty$

So by (3.9.5)

$$1 - 2F(\phi(b)) \to 0 \qquad \text{ie.}$$

$F(\phi(b)) \to \frac{1}{2}$ as $b \to \infty$, so $\lim_{b\to\infty} \phi_R(b)$ is the unique median.

(ii) $\{\phi_R(b)) \leq [-b,b]$ because if there is a $\phi_R(b) \in \{\phi_R(b)\}$

such that $\phi_R(b) > b$ then

$$F(\phi_R(b)+b) - F(\phi_R(b)) < F(\phi_R(b)) - F(\phi_R(b) - b)$$

and if $\phi_R(b) < -b$

$$F(\phi_R(b)+b) - F(\phi_R(b)) > F(\phi_R(b)) - F(\phi_R(b)-b).$$

The result follows. □

Corollary 3.11.1

If the loss function L used in convex downwards and F has a non-decreasing $\phi(b)$ function, then any Bayes decision with respect to L and F lies in the interval

[m,M] where m = mode of F

M = median of F.

Proof

Since the ramp loss functions $\{R_b(s),\ b \in \mathbb{R}_{>0}\}$ define a family of loss functions satisfying the conditions of Theorem 3.10, where L^* is the absolute loss function, $\min\{\phi_R(b)\}$ and $\max\{\phi_R(b))\}$ are non-decreasing in b.

It follows by Theorem 3.11, that the Bayes decision lies in $[\lim_{b\to 0} \min\{\phi_R(b)\} \quad \lim_{b\to\infty} \max\{\phi_R(b)\}]$. By the previous Lemma this implies any Bayes decision lies in

[m,M] as required □

Finally it should be noted that related work on some aspects of decision theory of the same flavour but of a different emphasis was given by Wald (1) and Blackwell and Girshick (1). The only related work I can find in recent statistical literature pertinent to this chapter is a paper by Baron (1) who investigates the effects on Bayes decisions of convex utilities when absolute or squared error loss functions are used.

Another good reason for using Step loss functions is given by Savage (1) where he shows that heuristic optimal decisions do not correspond with theoretical ones when the loss space has more than two elements.

Summary

Some general properties of estimates made using bounded loss functions and distributions that are possibly multimodal have been listed and a characterisation of general posterior distributions carried out. A few examples for standard distributions are presented. Finally I give some refinements of the aforementioned results under a convex utility hypothesis.

4. CATASTROPHE THEORY IN STATISTICS

4.1. Introduction

In the first part of this chapter I will give a very brief introduction to the parts of Catastrophe Theory pertinent to statistics illustrating the principles with a couple of examples in the second half.

I do not pretend that these first few pages give the reader more than a glimpse of what Catastrophe Theory is all about, so I shall give a set of references so that a fuller appreciation by the reader might be possible.

The best layman's introduction I have seen is a book by Poston and Stewart (1) and the non technical paper Isnard and Zeeman (1) could be usefully looked at, though many of the applications of the theory could at best be described as hopeful. A full proof of the Classification Theorem was first given by Mather (1). Other proofs are given by Trotman and Zeeman (1) and a less general but easier proof is in Bröcker (1). The classic book on the whole subject is written by Thom (1). It was he who conjectured the Classification Theorem in the first place. Unfortunately this is not the most readable of expositions.

There have been numerous applications of Catastrophe Theory made in the last few years; sadly many of them are less than adequate and others are blatently wrong. The best examples are in either the field of engineering or biology. For examples of the latter see Zeeman (2) and (3). On the sociological side the only papers I have seen that I have found at all convincing are Zeeman (4) and (5). It is from the latter that I take the statement of the Classification Theorem.

The Classification Theorem

Catastrophe Theory is a classification of potential functions, the classification being given by the following theorem.

Let C and X be manifolds, dim C $\leq$ 5 and $V \in C^\infty(C \times X)$. Suppose that V is generic in the sense that the related map

$$C \to C^\infty(X).$$

is transverse to the orbits of the group

$$\text{Diff}(X) \times \text{Diff}(\mathbb{R}) \text{ acting on } C^\infty(X).$$

("Most" C^∞ functions V are generic, genericity being open dense in Witney C^∞-topology).

Let $M \subset C \times X$ be given by

$$D_x V = 0$$

and let $\chi : M \to C$ be induced by the projection $C \times X \to C$.

Thom's Classification Theorem

(a) M is a manifold of the same dimension as C.

(b) Any singularity of χ is equivalent to an elementary catastrophe

(c) χ is stable under small perturbations of V.

The number of elementary catastrophes depends upon the dimension of C (and not on X)

Dim C	1	2	3	4	5	6
Elementary Catastrophes	1	2	5	7	11	(∞)

□

Hence provided dim C $\leq$ 5 I can classify what a Potential function V will look like locally, provided V is "nice" (i.e. C^∞ and generic). The proof of the above theorem is extremely arduous taking up some 60 sides of writing.

Following the notation of the theorem, X is commonly called the *behaviour space* and C the *control space* C the latter is a misnomer for applications I have in mind).

In statistics the above classification can be used in (at least) two very distinct ways which I shall call Type I and Type II classifications.

Type I Classification - Classification of expected loss functions

Type II Classification - Classification of processes acting on the Sample Space.

Type I Models

These are the easiest models to deal with. It has already been mentioned that expected loss functions are potential functions, so I can use the Classification Theorem directly on them. In this case:

X is the decision space.

C will depend on (a) The model being used [and hence the *form* of the prior distribution, likelihood and the loss functions]

(b) All the parameters of the model [including the loss function]

(c) The data set.

The effect in X on the decision d under local changes in particular $\xi \in C$ will be classified provided the expected loss function is C^∞ and generic (the former of these conditions being satisfied for example if the loss function $L(\theta-d)$ is C^∞.) This means in particular that if ξ moves smoothly with time, the different types of sharp changes or catastrophes in E(d) that occur will be catalogued by the theorem.

It cannot be strongly enough emphasised that the classification above is local and not global. In many cases however E(d) will in fact experience these local properties globally. This I shall look at in more detail in Chapters 6 and 7.

Type II Models

These models are a little more delicate to describe. Argue thus:

. Perhaps one of the most used methods in Statistical Analysis is Linear Regression, and the reason for its usefulness is clear. Many functional relationships are linear at least locally since they have a Taylor series expansion which converges and in many cases this local approximation is an adequate global description.

The question then is: How do I generalise this methodology? A possibility is polynomial regression but it is often found that a piecewise linear model is at least as adequate and I am back where I started. However, add the postulate that the relationship is induced by a potential function and Catastrophe Theory gives a wider class of "shapes" of relationships within which to expand the number of possible models. Just as in the linear case I will project the local Taylor series approximation into the global and hope for the best. What is interesting is that these relationships cannot be approximated linearly so I begin to move away from the much loved linear regression model.

I shall usually work with a random variable taking values on a *Canonical Catastrophe manifold* (I will define "canonical"later). In addition it will be noted that

X is the range of the random variable in question

C will depend on (a) The form of the model [prior and likelihood]

(b) The parameters of the model

(c) The data points.

Because of their greater simplicity I will spend most of the thesis analysing Type I models. However Chapter 7 shows that there are big links between the two types in fact. Type II models will be discussed in some of the remainder of this chapter and briefly in the last chapter.

In both the applications of Catastrophe Theory cited above it would seem sensible to limit the classification to those models where X, the Behaviour space, is the one dimensional manifold $\mathbb{R}$ as a first step (In fact I get no further than this in this exposition). Furthermore for Type I models I always choose an absolute minimumof expected loss and so canonical forms which evolve no new absolute minima are of less interest (but see §4.3 later). On these two counts I will not make any mention of Umbilics that might appear in Statistics in this thesis. A list of the classified Catastrophes is given below for $\mathcal{C} \leq 4$.

Codim	Name	Canonical Form
1	Fold	$V(x) = x^3/3 - ax$
2	Cusp (Dual Cusp)	$V(x) = x^4/4 - bx^2/2 - ax$ $(V(x) = -(x^4/4 - bx^2/2 - ax))$
3	Swallowtail	$V(x) = x^5/5 - cx^3/3 - bx^2/2 - ax$
3	Hyperbolic Umbilic	$V(x) = x^3 + y^3 + cxy - bx - ay$
3	Elliptic Umbilic	$V(x) = x^3 - xy^2 + c(x^2+y^2) - bx - ay$
4	Butterfly (Dual Butterfly)	$V(x) = x^6/6 - dx^4/4 - cx^3/3 - bx^2/2 - ax$ $(V(x) = -(x^6/6 - dx^4/4 - cx^3/3 - bx^2/2 - ax))$
4	Parabolic Umbilic	$V(x) = x^2y + y^4 + dx^2 + cy^2 - bx - ay$

The Umbilics also have "duals" that I have not included. Note that the duals are not geometrically different but there is an interchange of maxima and minima.

The Fold, Cusp, Swallowtail and Butterfly are the main Catastrophes of interest so I will briefly describe these in turn below. It must be remembered however that these manifolds can be pulled and stretched (see the statement of the theorem) and still remain equivalent in the terms of the theorem. The *Canonical Form* is the simplest element in each of the equivalence classes. Heuristically the Canonical forms can be thought of as a truncated Taylor series of each member in the equivalence class the truncation taken at a point where it does not affect the geometrical form. For an exact exposition see Thom (1) or Poston and Stewart (1).

The Fold Catastrophe

$X = \mathbb{R}$ $\quad C = \mathbb{R}$ $\quad V(x,a) = \frac{x^3}{3} - ax$

$c = a$

M_V is given by $x^2 - a = 0$ (See Fig. 4.1).

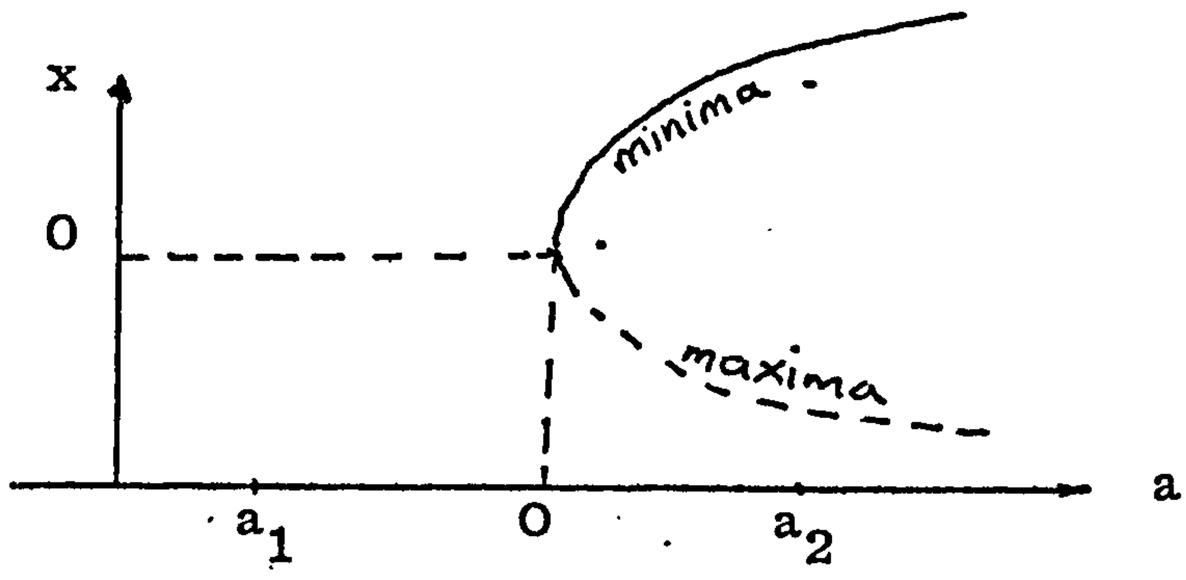


Fig. 4.1.

$V(x,a_3)$ $\qquad$ $V(x,0)$ $\qquad$ $V(x,a_1)$

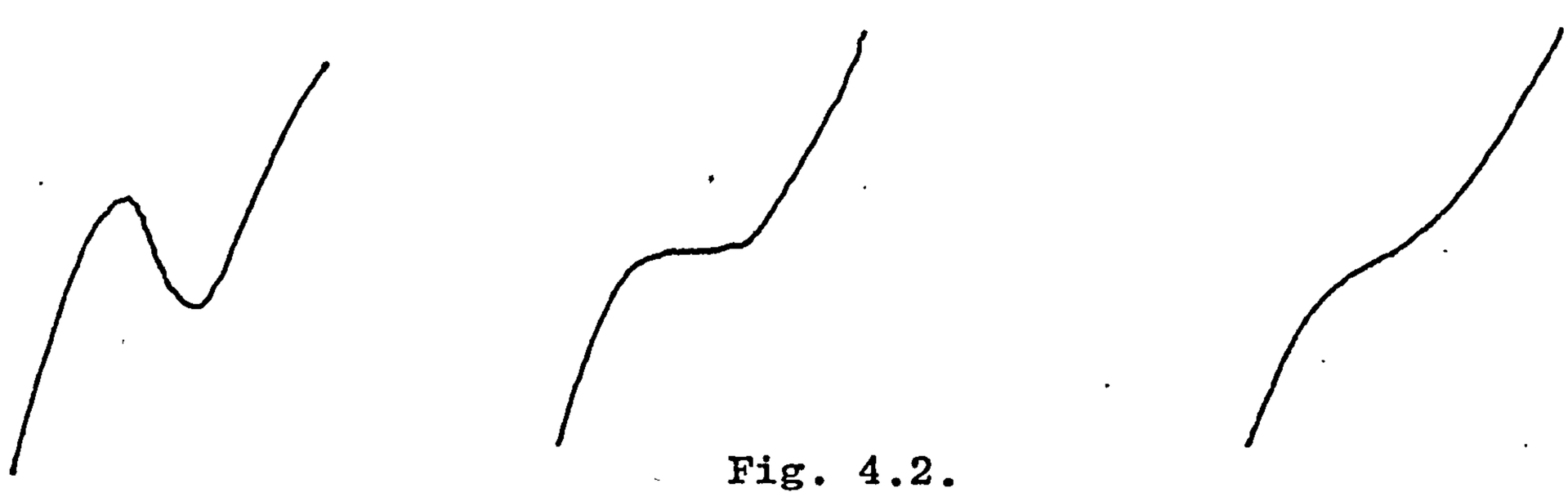


Fig. 4.2.

The position of the maxima and minima of V(x) are given in Fig. 4.1 and some illustrative potential functions given in Fig. 4.2. Notice that no minima (except $-\infty$) exists for $a \leq 0$, and hence no point of equilibrium.

The Cusp Catastrophe

$X = \mathbb{R}$ $\quad \underset{\sim}{c} = \mathbb{R} \times \mathbb{R}$ $\quad \underset{\sim}{c} = (a,b)$ $\quad V(x,\underset{\sim}{c}) = x^4/4 - bx^2/2 - ax$

M_V is given by $x^3 - bx - a = 0$.

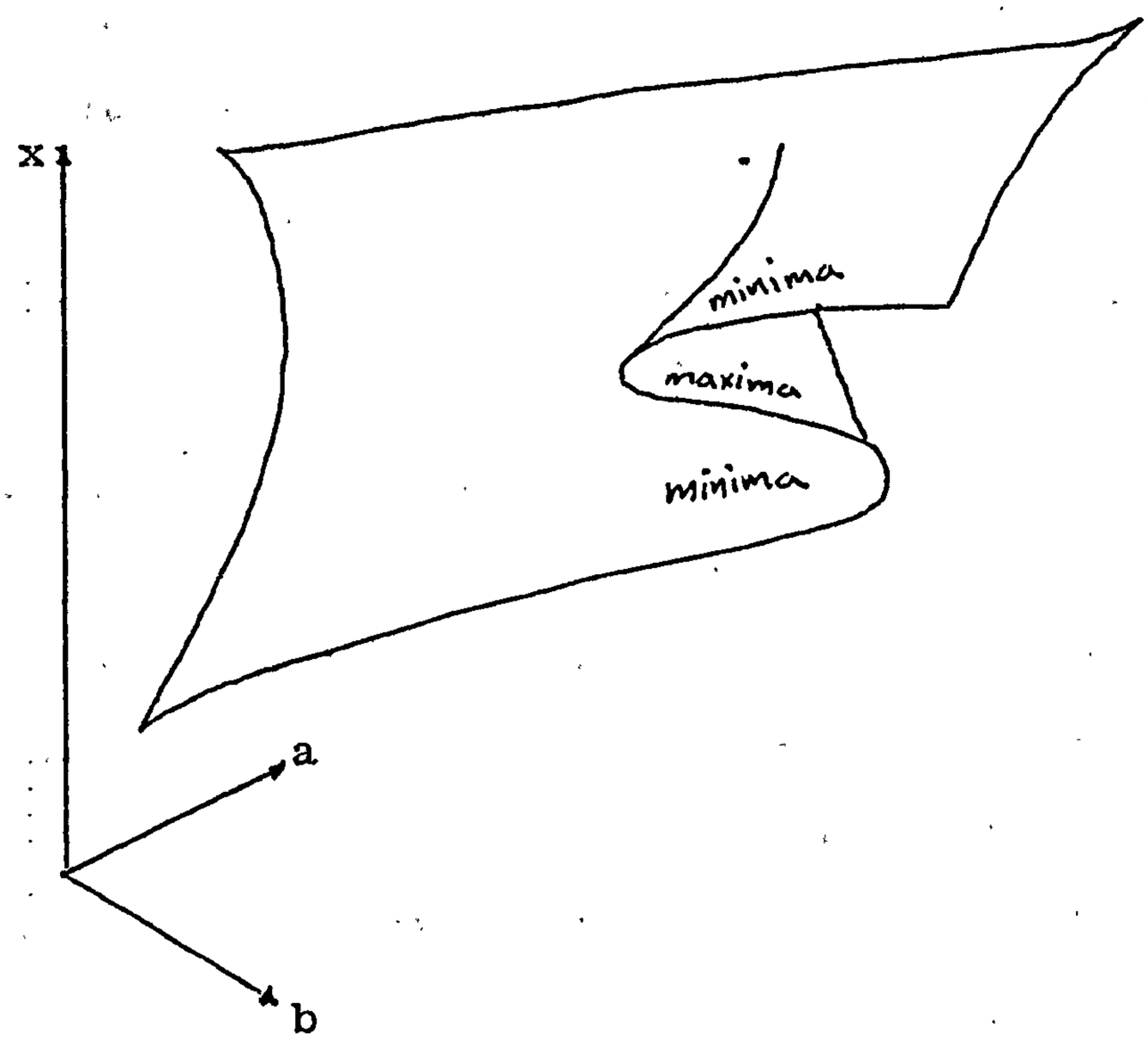


Fig. 4.3. The Cusp Manifold

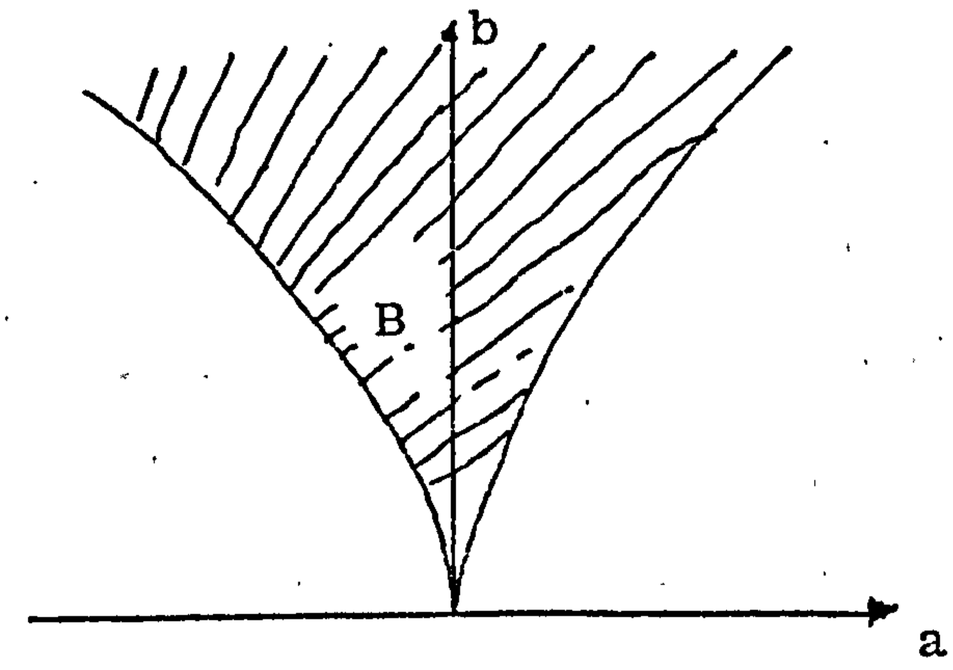


Fig. 4.4. The Bifurcation Set

The Cusp Catastrophe is the most well known and useful of all the Catastrophe, M_V given by the graph in Fig. 4.3. Of main interest is the number of minima (and maxima) of V(x) for particular values of the pair (a,b). If I solve

$$V'(x) = 0 \qquad \text{with}$$

$$V''(x) = 0 \qquad \text{(differentiation being with respect to x).}$$

I obtain the *Fold points*, which as projection on to the Control Space gives the cusp graph in Fig. 4.4. Notice that in the shaded region B (called the *Bifurcation set*) that is bounded by the projection of the fold lines there exist 2 minima (and one maxima) of V(x) and outside B in $(B^c)^o$ there exists exactly one stationary point, a minimum. The point x = a = b = 0, the solution of

$$V'(x) = V''(x) = V'''(x) = 0$$

is called the *Cusp point*

Note that along the fold lines there exists 1 turning point and one minimum except at the cusp point where there is just 1 minimum V(x) is often used to represent two conflicting regimes, the regimes being modelled by the different "parts" of the manifold.

For the *Dual Cusp* put V*(x) = -V(x), V(x) defined above. This just makes all the maxima minima and vice versa. B then contains 1 local minima and $(B^c)^o$ none. However since this has no conflict of regimes it tends to be of less interest. For both cusps

a is called the normal factor

b is called the splitting factor.

The difficulty in analysing the "higher" catastrophes is one of trying to picture the geometry (I do not have enough dimensions to draw them). I circumvent the problem temporarily for the case of the Swallowtail Catastrophe by just drawing the projection of the fold points onto the Control space (now 3-dimensional) as I did for the Cusp and hence get a representation of regions containing certain classes of maxima and minima.

The Swallowtail Catastrophe

$X = \mathbb{R}$ $\quad c = \mathbb{R}^3$ $\quad V(x),\underline{c}) = \frac{1}{5}x^5 - \frac{1}{3}x^3 - \frac{1}{2}x^2 - ax$

$\underline{c} = (a,b,c)$

M_V is given by $x^4 - cx^2 - bx - a = 0$

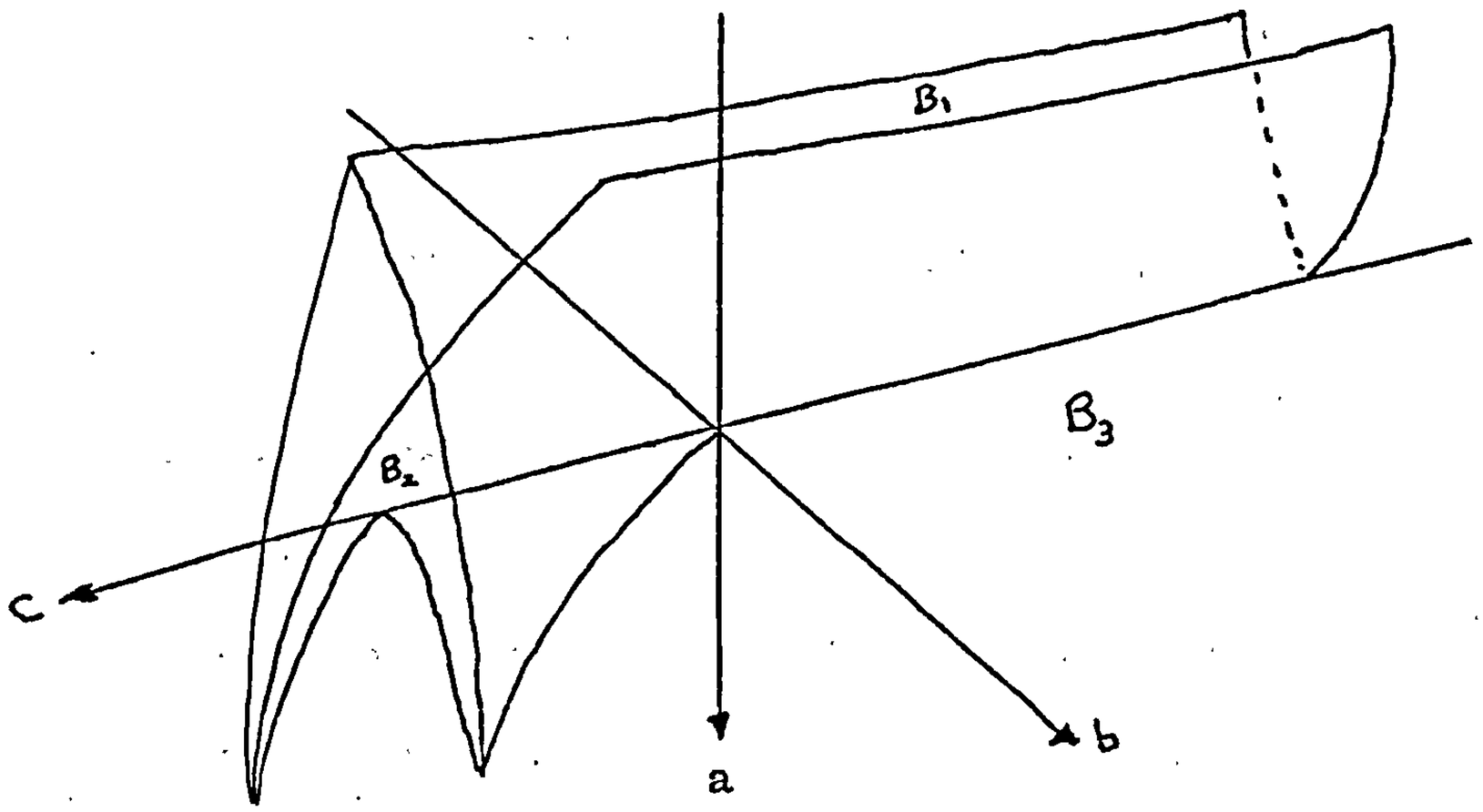


Fig. 4.5. The Swallowtail Control Space

Very roughly speaking the Swallowtail Catastrophe is a Fold Catastrophe with a "kink" in it. The projection of the Fold surfaces on to the control space partition it into the 3 regions marked on Fig. 4.5.

$V(x,\xi) \in B_1^o$ the potential function has no turning points (as in the Fold with a < 0)

$V(x,\xi) \in B_2^o$ the potential function has 2 local maxima and 2 local minima

$V(x,\xi) \in B_3^o$ the potential function has 1 local maxima and 1 local minima.

Hence for c < 0, this behaves like a Fold Catastrophe. For c > 0 and a, b small in absolute size. $V(x,\xi)$ will look like Fig. 4.6.

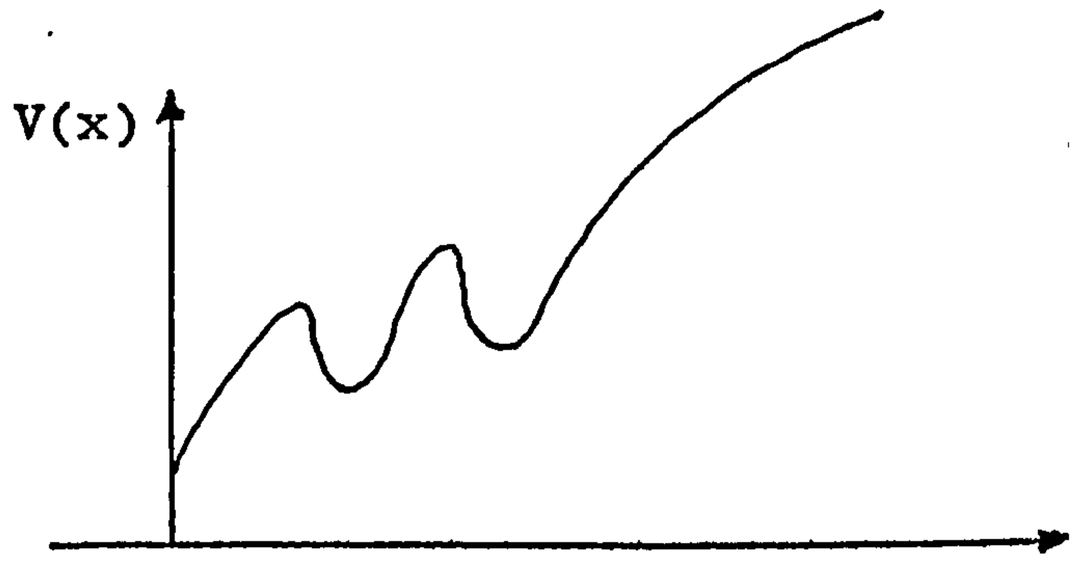


Fig. 4.6. A Swallowtail potential

The Butterfly Catastrophe

Perhaps the most interesting of the Catastrophes is the Butterfly Catastrophe. The difficulty of seeing this geometrically is now acute, however, since even the Control space is 4-dimensional.

$X = \mathbb{R}$ $C = \mathbb{R}^4$ $V(x,\xi) = x^6/6 - dx^4/4 - cx^3/3 - bx^2/2 - ax$

$\xi = (a,b,c,d)$

which has manifold M_v of stationary points given by

$$x^5 - dx^3 - cx^2 - bx - a = 0.$$

In this case (a) is called the *normal factor*

(b) is called the *splitting factor*

(c) is called the *bias factor*

(d) is called the *butterfly factor*.

I represent the control space in stochastic form to make up for the lack of dimensions hence.

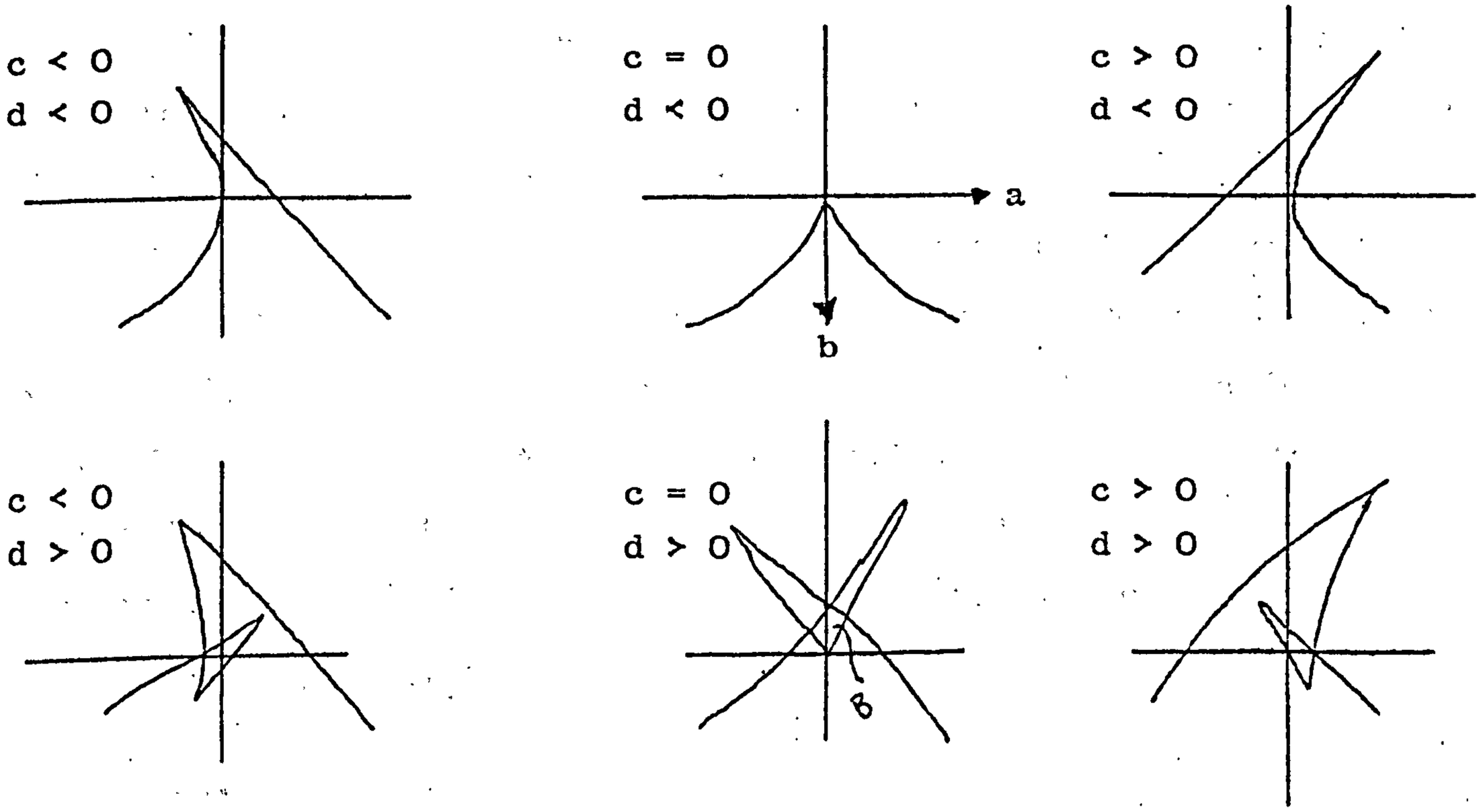


Fig. 4.7. The Butterfly Catastrophe's projected Fold points

Intuitively it is fairly easily seen how these shapes evolve into one another for varying c and d. The most interesting potential arises from c = 0, d > 0 (see Fig. 4.7) where 5 distinct regions are defined, the central one mark B containing 3 minima and 2 maxima and hence a minima which could be described as a Compromise minima being between two other regimes.

The Dual Butterfly is of less interest, the potential function being the wrong way up for most models.

For a much fuller exposition of the above geometry and others read Woodcock and Poston (1) and Thom (1).

Catastrophes with Boundaries

In addition to these Catastrophes there are some others of great importance that have not been publicised though as yet nothing has been published on them; namely Catastrophes with Boundaries. This is a cause of great embarrassment to me since I want to use the concepts therein contained without giving a thorough exposition of the subject. Basically the idea is not to restrict the Behaviour space X to be open. I restrict the Behaviour space to a compact interval, (the astute reader will recognize I would like such intervals for sake of integrability). For example let

$$X = [q,r].$$

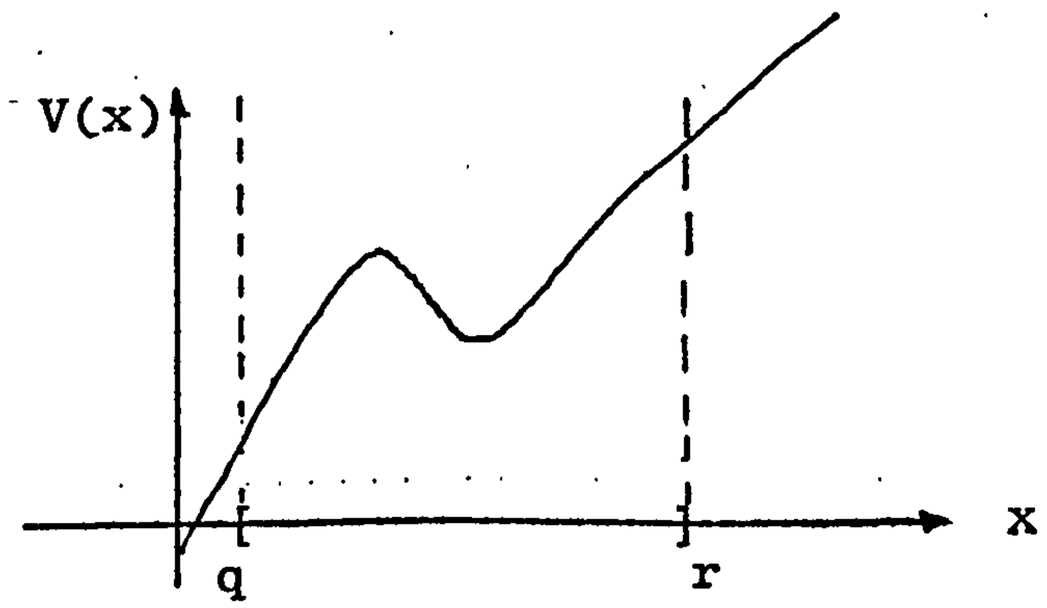


Fig. 4.8. The Restricted Behaviour Space

In Fig. 4.8 V(x) has a local minima at q. This must therefore be included in any analysis of behaviour that subsequently occurs. V(x) in Fig. 4.8 has 2 minima and 1 maxima and so will behave more like a Cusp than a Fold in its evaluation. Obviously the Dual Catastrophes regain their importance too. Most examples of Fold and Swallowtail that I give will, in fact, be these catastrophes with *boundaries*.

4.2. Rules in Catastrophe Theory

Because Catastrophes have been given a local classification, the discipline of Catastrophe Theory is best placed in a dynamic setting where local perturbations are natural phenomena. Hence I assume that the Controls go through a (mostly) smooth path with time inducing a corresponding movement in the Behaviour space

i.e. $V_t(x,\underset{\sim}{c}) = V(x,c(t))$.

The theory is still not complete enough for any application however, since I have not yet specified which minimum of V(x) to choose for my Behaviour point when more than one minimum exists. A method of making such a choice is called a *Rule*.

The Delay Rule

Choose that $\begin{cases}\text{minima} \\ \text{turning point}\end{cases}$ $x^*(t)$ of $V_t(x,\underset{\sim}{c})$ such that

(i) $x^*(t)$ is continuous in t if possible

(ii) If $t = t^*$ is a point where it is not possible for $x^*(t)$, to be continuous then $x^*(t^*)$ will be a turning point of $V_t(x,\underset{\sim}{c})$.

In this case define $x^*(t)$ to be left-continuous and

$$x(t^*) = \lim_{t \to t^*, t > t^*} x(t)$$

to be the minima adjacent to $x^*(t^*)$.

The Delay Rule, for example, describes the trajectory of a marble on a smooth surface under the influence of gravity. It is the main Rule used so far in modelling Catastrophes. Because of (i) in the definition given only the Controls ξ at time t it may not be possible to determine the appropriate behaviour $x^*(t)$. For example, in the Cusp Catastrophe, if the Controls ξ lie in B (see Fig. 4.4) then depending on the trajectory through C beforehand $x^*(t)$ will either lie on the top or the bottom sheet of the manifold M_V.

Hence this particular rule is necessarily stochastic having a "memory". Thus it generates many interesting models and would seem pertinent to the study of Type II models described above.

Its usefulness lies in the fact that it depends solely on M_V and not the potential function V and so I can concentrate solely on M_V in any analysis. Note that it is also a "local" rule and so fits in with the local classification of Thom's theorem. I will illustrate the Delay Rule with an example.

<u>The Type II model on the Cusp Catastrophe with Delay Rule</u>

Let θ be the Behaviour variable, and a the normal factors with $b > 0$ the splitting factor and suppose for simplicity that θ is deterministically linked with a and b

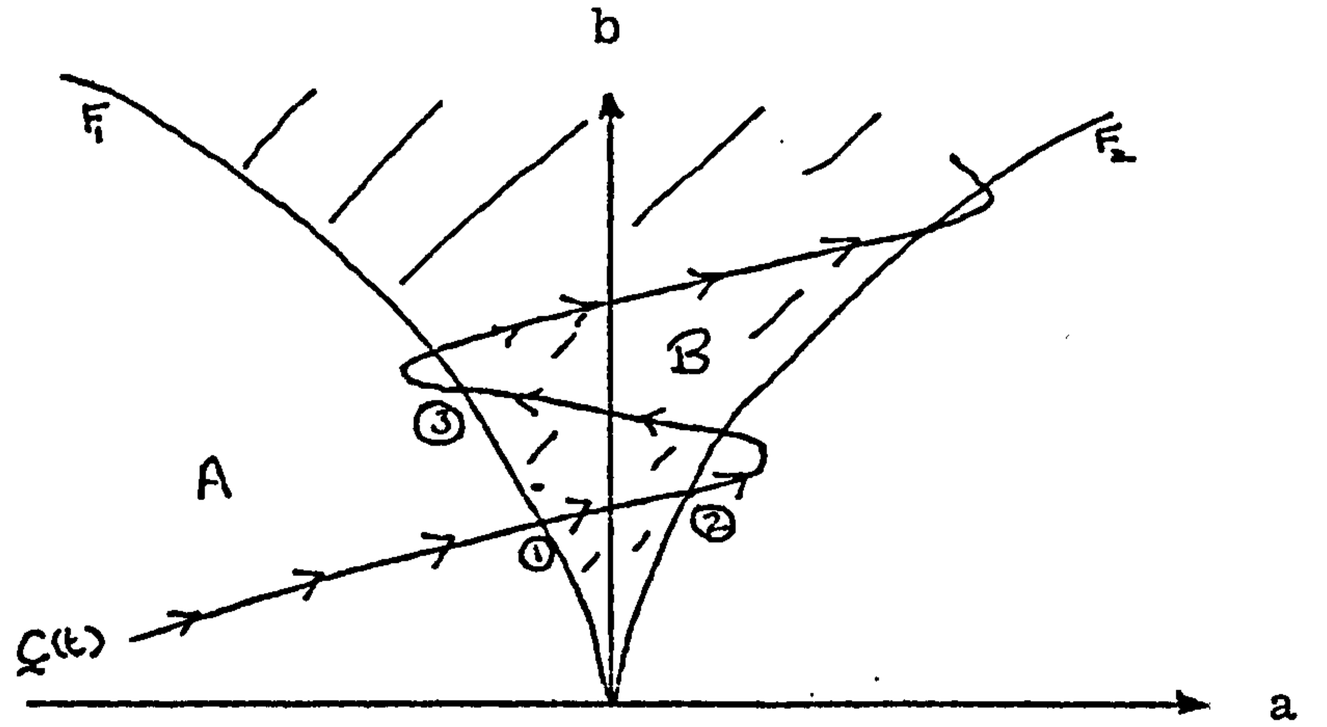


Fig. 4. 9. A path on the control space of the Cusp Catastrophe

Suppose the control factors $\underset{\sim}{\zeta} = (a,b)$ undergo the smooth evolution $\underset{\sim}{\zeta}(t)$. Under the Delay Rule θ will make a smooth trajectory

$$\theta = M_1(a,b) \qquad \text{(say)}$$

until I reach the point labelled ② in Fig. 4.9 where suddenly θ will follow another smooth model

$$\theta = M_2(a,b) \qquad \text{(say)}$$

having gone through a large rise in value. θ continues then to be governed by M_2 until it reaches a point ③ shown above where it drops down onto model M_1 again. Everytime it meets and crosses the projected fold line F_1 from the top model M_2 or the fold line F_2 from the bottom smooth model $M_1(a,b)$. θ changes models.

Formally this can be described by the equation

$$\theta_t = \chi_t(a,b)\, M_1(a,b) + (1 - \chi_t)\, M_2(a,b)$$

where $\chi_t = \begin{cases} 1 & \text{if } \chi_{t-} = 1 \text{ and } (a,b) \in A \cup B \\ & \text{or if } \chi_{t-1} = 0 \text{ and } (a,b) \in B \\ 0 & \text{otherwise} \end{cases}$

where $\chi_{t_1-} = \lim_{t>t_1,\, t\to t_1} \chi_t$ and A and B are marked on Fig. 4.9.

Now since M_1 and M_2 are smooth they can be approximated piecewise linearly. This provides the following illustration.

Example

Let M_1 be given by $$\psi_t = \psi_{t-1} + \varepsilon_t^{(1)} \qquad 4.2.1.$$

Let M_2 be given by $$\phi_t = \phi_{t-1} + \varepsilon_t^{(2)}$$

where $\varepsilon_t^{(1)}$, $\varepsilon_t^{(2)}$ are identically distributed, independant random variables.

Note then that $\theta_s = \psi_s$ if and only if $\chi_{s+1} = 1$ and similarly

$$\theta_s = \phi_s \text{ if and only if } \chi_{s+1} = 0.$$

Without loss of generality assume $\chi_t = 1$ and

$$\chi_{t-r}(a,b) = 0$$

$$\chi_{t-r+j}(a,b) = 1 \quad \text{for all } j \in \mathbb{N},\ 1 < j \leq r.$$

Then 4.2.1 gives that

$$\theta_t = \begin{cases} \theta_{t-r} + \sum_{t=0}^{r-1} \varepsilon_{t-r}^{(2)} & \text{if } \zeta(t) \in (A \cup B)^c \\ \theta_{t-1} + \varepsilon_t^{(1)} & \text{otherwise} \end{cases}$$

Suppose I assume now that in addition the error terms are normally distributed, the resultant series will then look like Fig. 4.10. Typically I could let θ_t represent the "level" of the process in a Bayesian Forecasting Setting (Harrison and Stevens (1)). Priory would then be put on ψ_o and ϕ_o and I would do the usual Bayes update. I have here something akin to the Kalman Filter Multiprocess Model but with more structure on the mixing parameters. The continuous time analogue can be phrased in the same way.

Hence I have with the Delay Rule acting on the different Catastrophes a host of Type II models which are very different to the normal Time Series models at present employed. It is quite clear that I could spend the rest of the thesis analysing estimation problems and the like arising from these models and do some case studies. I have preferred however in this exposition to consider theoretical aspects of Type I models so there is not space to include such a study. This is something I shall return to at a later date, though I touch upon these models again in the last chapter.

Finally two words of warning for anyone wanting to dabble in these models. The first is that such Time Series models get very complicated very quickly. I hope that Statistical techniques will be used only on sensibly constructed and explainable models and that the data is not just "fitted". (As in polynomial regression, the number of possible models is so large that something will be a "good fit"!).

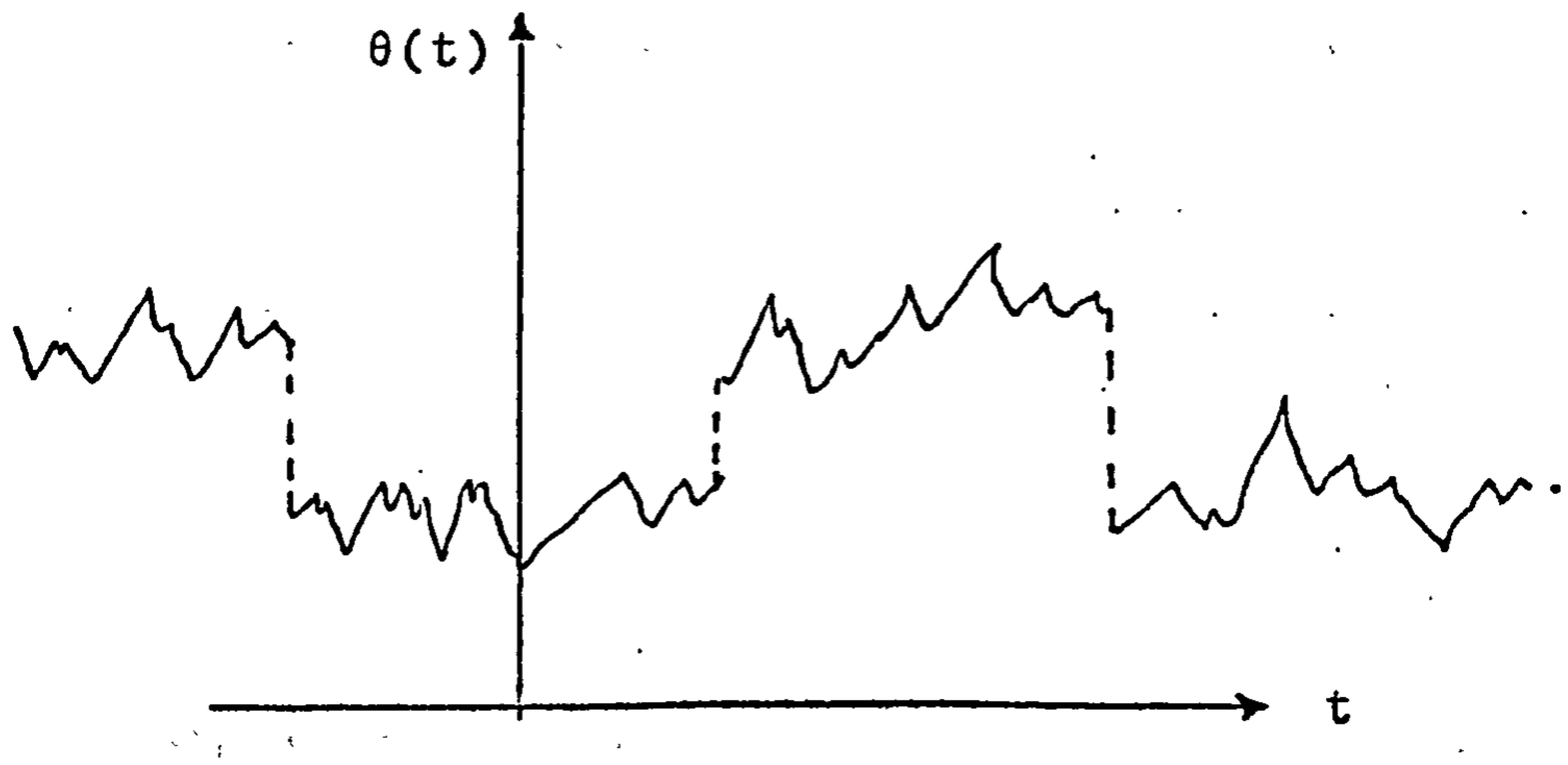


Fig. 4.10. Typical Times Series generated by Cusp

Secondly traditional methods of testing between connections between variables are not adequate for these models. For example suppose I have a Hysteresis Loop (i.e. the path on the controls marked out in Fig. 4.9. but with b remaining constant). Then even if I have a deterministic relationship between the behaviour variable θ and the normal factor a, the points will be distributed as in Fig. 4.11.

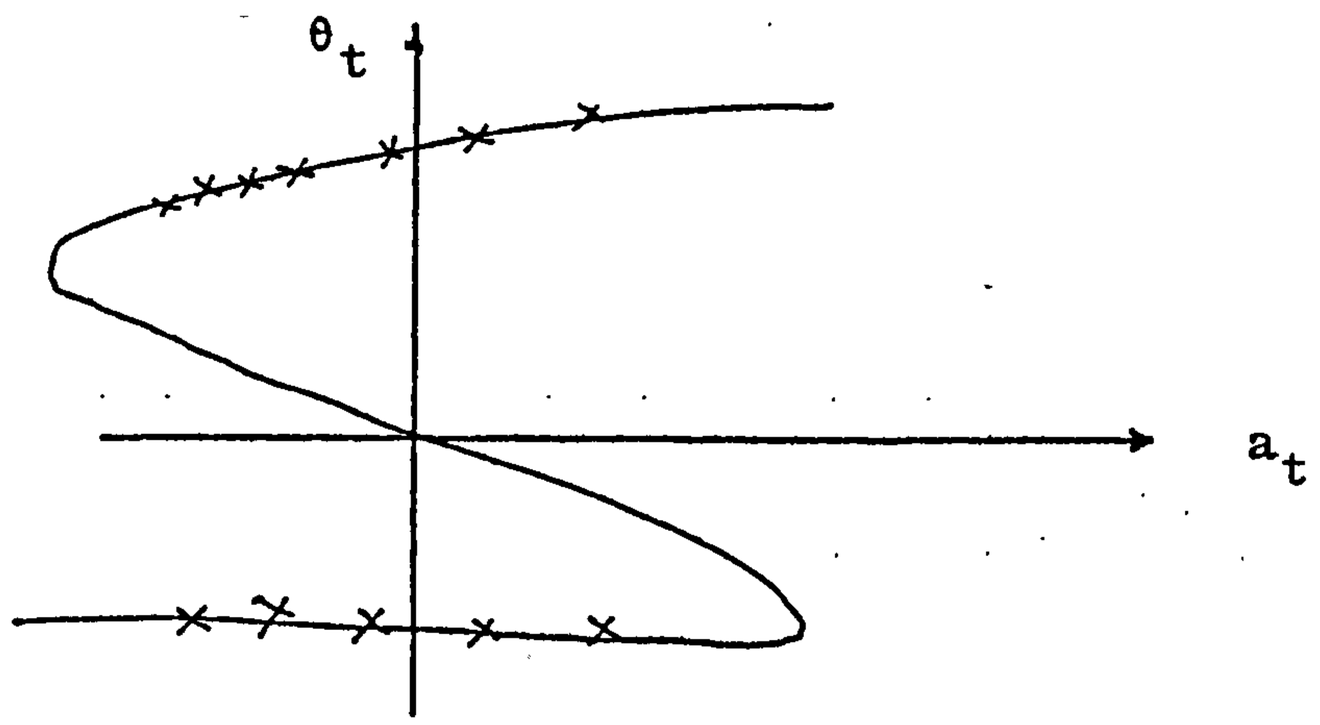


Fig. 4.11. The Hysteresis Loop

To argue that because there is little correlation between θ and a and thus no connection (as someone in fact has) is obviously misguided.

Maxwell's Rule

Another rule that has been used is called Maxwell's Rule. Here simply I have that

Behaviour $x^*(t) = \inf V_t(x,c)$.

This is of course the Bayes decision rule for Type I models above where the Potential function V(x) is the expected loss. Unlike the Delay Rule it is independent of the sequence at decisions leading up to it and therefore basically non sequential.

It is also easily seen that it is dependent not only on the manifold M_V but also the actual Potential functions V generating it.

It is basically a global rule in the sense that I need to know the whole of the Potential function rather than just its local form (as in the Delay Rule) before I can choose the correct Behaviour point. Note however that for local linear models Maxwell's Rule and the Delay Rule are the same, since only one stable behaviour point exists at any one time. For non-linear models they give in a sense two extremes of behaviour and are important as limiting cases for other rules.

4.3. Bayes Statistical Rules

Other rules can come into play for Type I models which make them stochastic in nature. The simplest way of incorporating some form of time dependence into the Bayesian Inference framework is to put a cost on the changing of a decision.

Cost of Change Rules

"Change" must be well defined and the most simple way of doing it is in the following way

Let the *Step Cost of Change Function* $C(d_t - d^*_{t-\eta})$ be defined by

$$C(d_t - d^*_{t-\eta}) = \begin{cases} 0 & | d_t - d^*_{t-\eta}| < b\eta \\ k & \text{otherwise} \end{cases}$$

where $\eta > 0$ and $d^*_{t-\eta}$ was the chosen Bayes decision at time $t-\eta$.

Hence if I make decisions on a set whose points are η apart then changing a decision by less than $b\eta$ incurs no loss, whereas changing the decision by more than $b\eta$ incurs a loss of k units.

Let $E_t(d)$ represent the expected loss at time t if there was no cost in changing a decision and let $G_t(d)$ be the total expected loss associated with the model, then

$$G_t(d) = \begin{cases} E_t(d) & d\varepsilon\ (d^*_{t-\eta} - b\eta,\ d^*_{t-\eta} + b\eta) \\ k + E_t(d) & \text{otherwise} \end{cases}$$

Hence it adapts $E_t(d)$ as shown in Fig. 4.12.

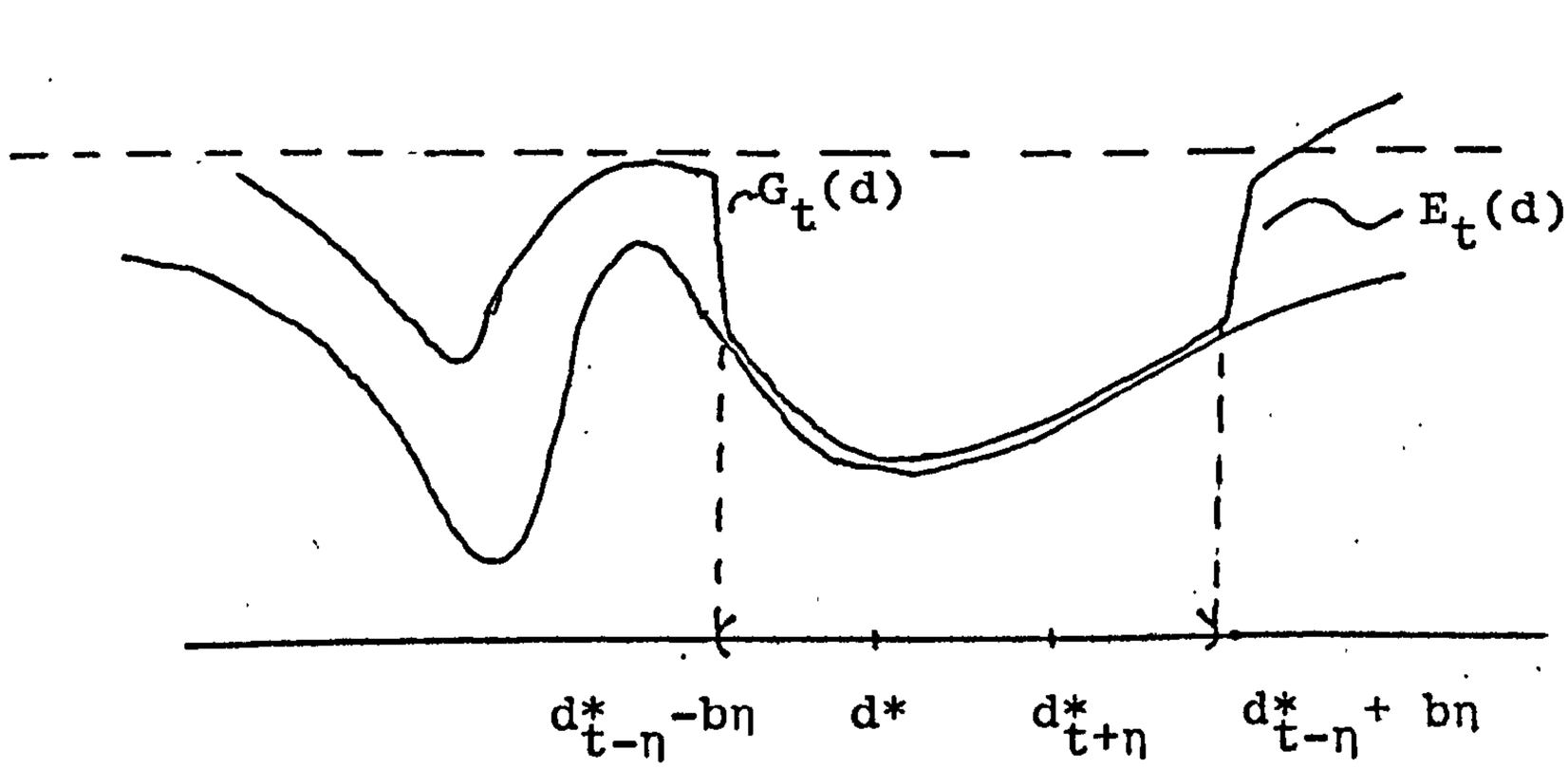


Fig. 4.12. The Expected Loss function with Cost on Change

Suppose I use the convention in Chapter 2 that

$$0 \leq E_t(d) \leq 1$$

and fix $\{E_t(d) : t \in \mathbb{R} \geq 0\}$. Then as $k \to 1$ and $\begin{cases} b \to \infty \\ b\eta \to 0 \end{cases}$ it can easily be seen that if $E_t(d)$ moves smoothly with time $d^*(t)$ will follow the trajectory of the local minima $d^*(0)$ whilst this minima exists. If it disappears, then since b is very large it will "almost" instantaneously latch on to the adjacent minima. Hence I have that

when $k \to 1$ Step Cost of Change Rule → Delay Rule

$b \to \infty$, $\eta b \to 0$

$k \to 0$ Step Cost of Change Rule → Maxwell's Rule.

$b\eta \to \infty$

Thus Delay Rules and Maxwell's Rule can be seen as limiting cases of Cost of Change Rules. For moderate values of k and $b\eta$ I obtain a Delay Rule in the neighbourhood of the Cusp point, in the case of the Cusp Catastrophe, gradually becoming a Maxwell Rule as I get further away from this Cusp point. (See Fig. 4.13).

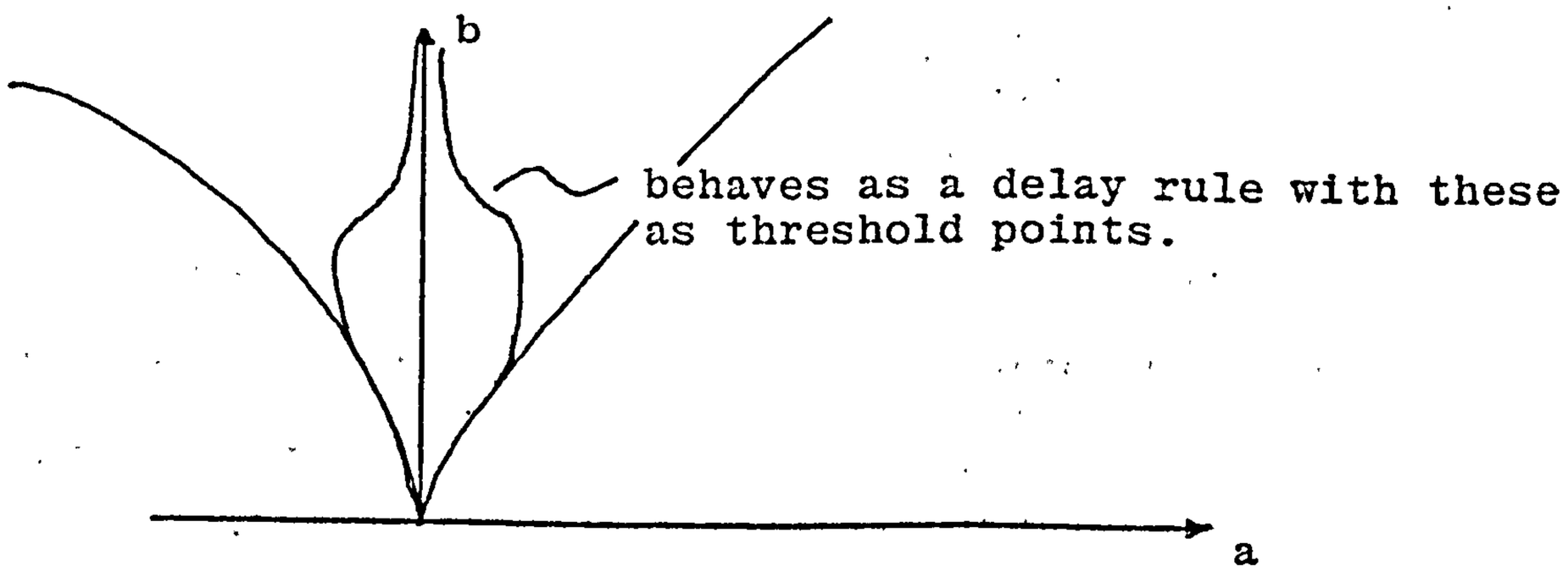


Fig.4.13. Cost of Change effect on Control space of Cusp

It should in passing be noticed that in many cases the Cost of Change function would not be Markov in the sense that the cost of change would have to be offset by future long term benefits of change. This would have the effect of averaging weighted future expected loss functions. I will not go into this here.

An Illustration: Judgement Under Stress

Here I will follow the content of a paper by Zeeman (6). In an experiment by Drew, Colquham and Long (1), subjects were given a small dose of alcohol and asked to drive at what they thought

would be normal speed on a simulator. The effect of the alcohol on their speed was plotted against an introversion/extroversion scale and the graph on Fig. 4.13a was obtained.

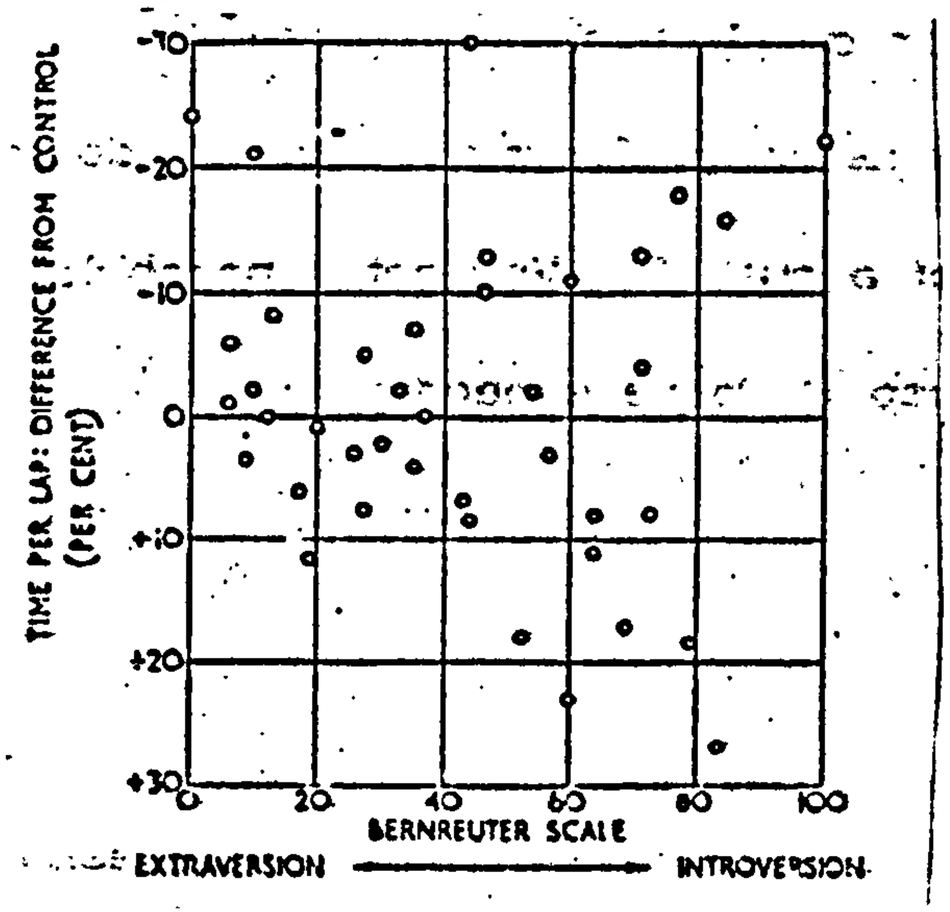


Fig. 4.13a.

Zeeman argues that one takes cues as one drives to help estimate the speed of the car, but if one's integrative capacity decreases, for example because of alcohol intake, the middle cues are missed out. Hence instead of obtaining a unimodal distribution representing one's understanding of speed one tends to a bimodal mixture of an overestimation distribution and an underestimation distribution centred at $S_n + B$ and $S_n - B$ respectively (say) where S_n represents the actual speed, and $B > 0$.

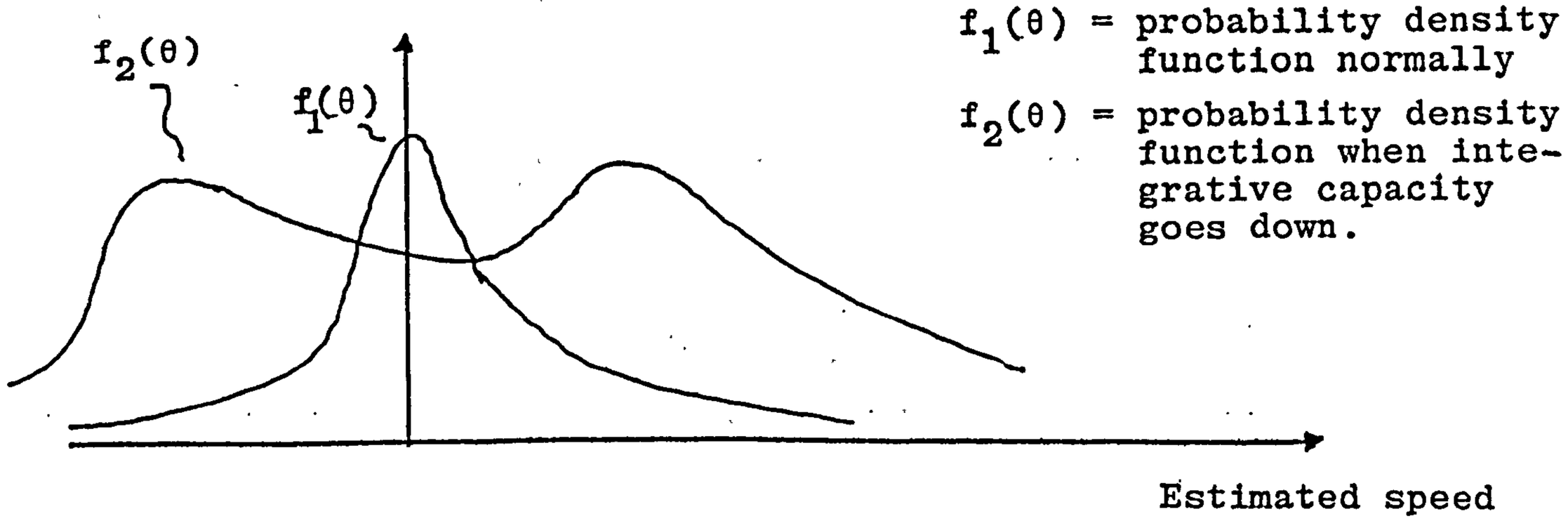


Fig. 4.14. Estimated Speed

Increasing ones speed towards the second mode and away from the true speed, the proportion of overestimation cues to underestimation cues increases and vice-versa decreasing one's speed, the proportion goes down.

For the sake of simplicity I am now going to assume that the probability density $f_2(\theta)$ is a mixture of two normal distributions with equal variances, and the loss function being used is the conjugate one. (see Lindley (1). These assumptions are in no way crucial to the argument given below (see Chapter (6), but they do make the analysis easier. In addition I will assume that the distribution for each individual is the same and each of the subjects acts as if he is a Bayesian, i.e. minimises his expected loss. The difference in decision making between the introvert and extrovert can now be studied.

It could be said that the extrovert makes estimates using a large value of his k parameter in his loss function, taking an expansive view of things, hence he tends to compromise between the modes in Fig. 4.14 and choose an estimate in the region of the time speed. The introvert on the other hand will tend to work with a loss function with a small k value. In this case the expected loss will go bimodal. (See Figs.4.15 and 4.16). (For precise result see Chapter 7, Theorem 7.1).

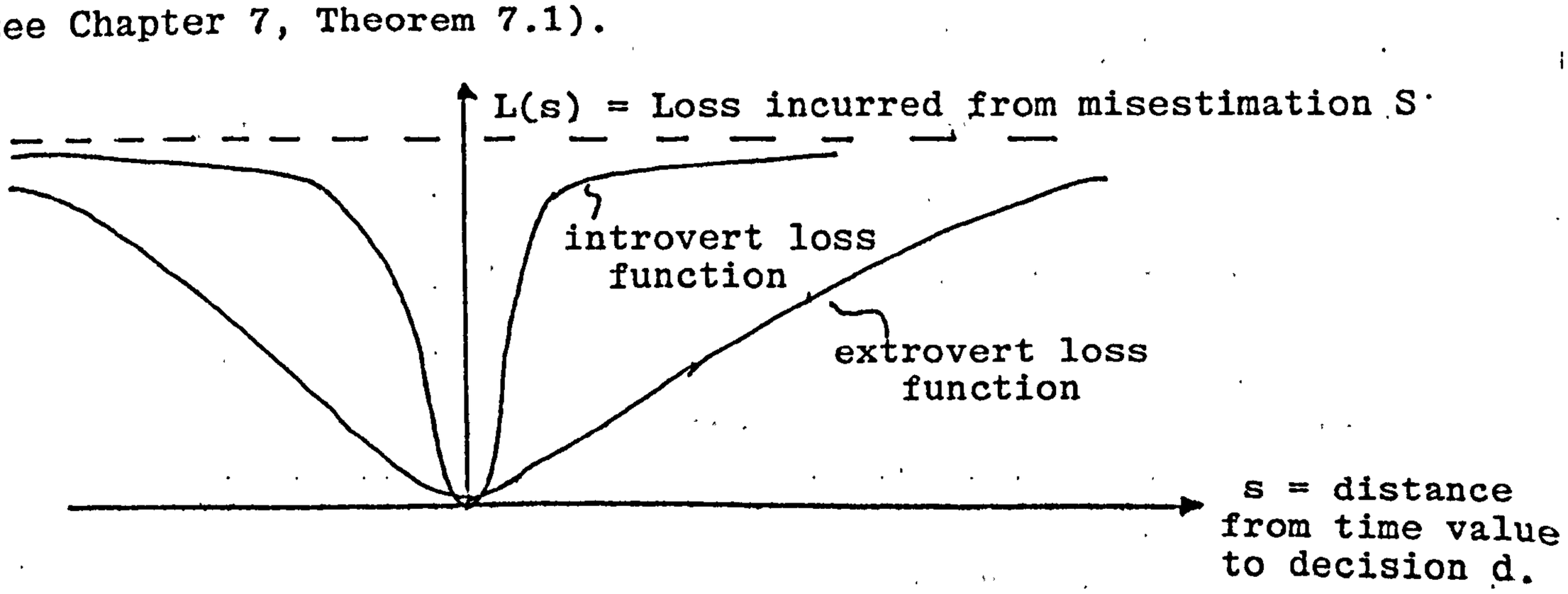


Fig. 4.15. Loss functions of Subjects

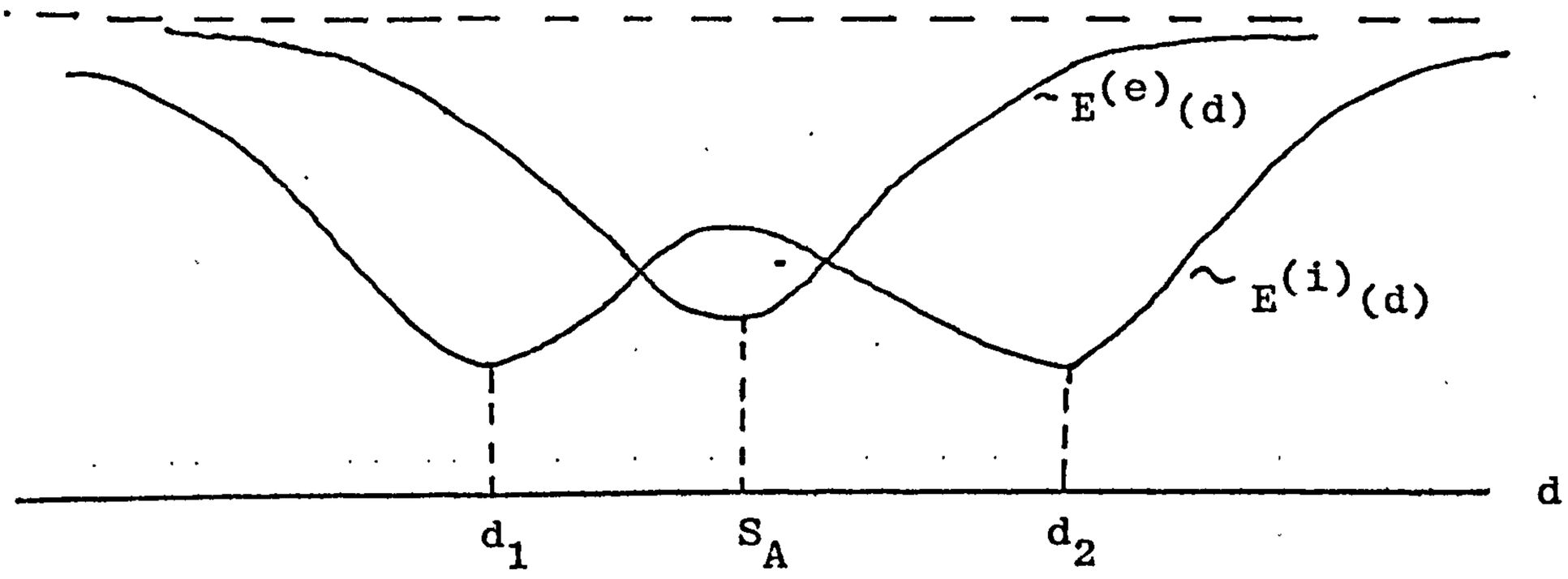


Fig. 4.16. Expected loss functions of Subjects

where $E^{(e)}_{(d)}$ = expected loss for extrovert

$E^{(i)}_{(d)}$ = expected loss for introvert

S_A = actual speed.

If I now refer to Fig. 4.16 and look at the case when the driver is travelling at the true speed it is seen that the extrovert estimates correctly. The introvert, however has two equally favourable decisions as to the speed he is travelling at d_1 and d_2 (both wrong). He will therefore accelerate to d_2 or decelerate to d_1. Then the proportion of low speed cues, in the former case, will go up. However, there is a cost to change in this model, he cannot reach d_1 immediately and will lose face if he changes his mind suddenly. Suppose he imposes the Step Cost of Change and takes a very short term view of things. If the Step Cost of Change has a form close to that tending to a Delay Rule, then the introvert will follow the decision d_2 as he accelerates.

Of course this will have the effect of reducing the number of overestimation cues and increasing the number of underestimation cues until the ratio of the mixture variables is such that the local minima

at d_2 disappears. The Delay Catastrophe then occurs and the subject immediately breaks towards the only remaining minima of expected loss d_1.

This illustrates how Bayesian decision theory can logically model Zeeman's geometry into an application in Psychology without handwaving arguments that usually are banded about.

Note that the step cost of change model is just a simple example of cost of change models in general which seem well worth developing into a theory, though I have not yet done this. The obvious next extension is to allow $C(d_t - d^*_{t-\eta})$ to be any function increasing in $|d - d^*_{t-\eta}|$. In this case $C'(x) \neq 0$ would imply that the behaviour would take discrete jumps, even though the phenomena considered was in all senses a continuous one.

Model Breakdowns and the Delay Rule

Nearly always models are only local descriptions of global phenomena. Typically a random variable $\theta(\underset{\sim}{C}_t)$ depending on factors $\underset{\sim}{C}_t$ will be well modelled only if $\underset{\sim}{C}_t$ are fairly constant through time. If on the other hand the $\underset{\sim}{C}_t$ jump about the relationship imposed by my approximating model relating θ_t to $\underset{\sim}{C}_t$ will no longer be valid.

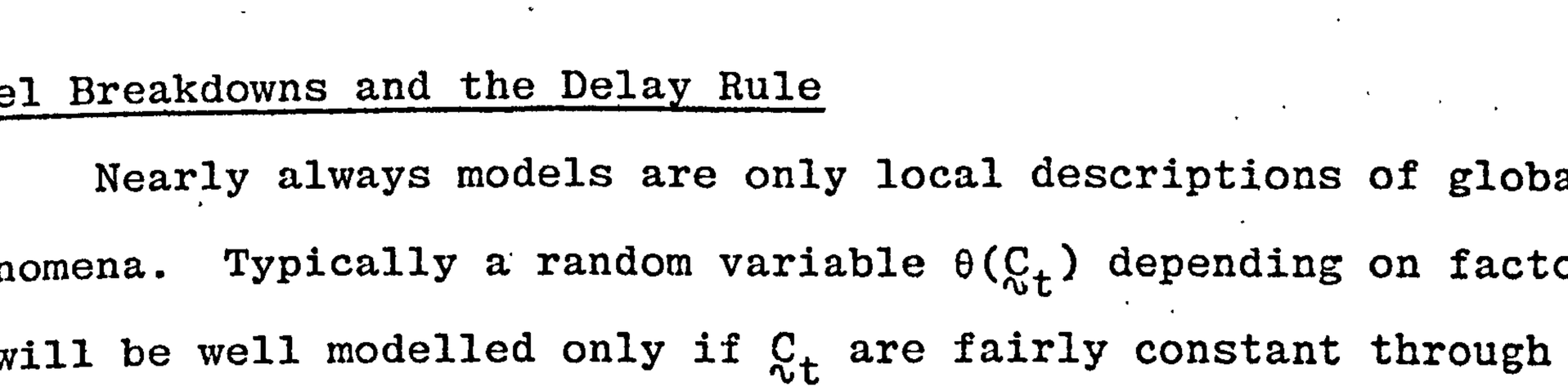


Fig. 4.17.

Suppose θ_t is now a random variable on $\mathbb{R}$.

Suppose then that $F(\theta_t)$ is my posterior distribution arising from the locally approximating model given above. It follows that replacing $F(\theta_t)$ by $G(\theta_t)$ is likely to give a better summary of my posterior beliefs where $G(\theta_t)$ is defined by

$$g(\theta_t) \propto \begin{cases} f(d^*_{t-\eta-a}) & -C \le \theta \quad d^*_{t-\eta} - a \\ f(\theta) & d^*_{t-\eta} - a \le \theta \le d^*_{t-\eta} + a \\ f(d^*_{t-\eta}+a) & d^*_{t-\eta} + a < \theta \le C \\ 0 & \text{otherwise} \end{cases}$$

where C is a very large positive number

$d^*_{t-\eta}$ is the previous decision made.

Thus if $\theta \in A = [d^*_{t-\eta} - a, d^*_{t-\eta}+a]$ I am sufficiently sure of my approximating model whereas if θ lies outside A I feel completely ignorant about θ_t. Note that a similar type of approximation is given in a different context by Leonard (2).

Now let $H(\theta_t)$ be the *true* distribution of θ_t at time t. (See Fig. 4.17). If I am right about my assessment of when my approximating model is reasonable then

$$h(\theta_t) \propto g(\theta_t) \quad \text{when } \theta_t \in A.$$

but different (in general) otherwise. For simplicity suppose I use the simple Step loss function $S_b(\theta-d)$ where $b < a$. It can then be seen that provided C is large enough

$$E_b(G,d) \propto \begin{cases} E_b(H,d) & d \in B \\ T(d-d^*_{t-\eta}) & d \in B^c \end{cases}$$

where T(r) is increasing in $|r|$ and

$$B = (d^*_{t-\eta} - (a-b),\ d^*_{t-\eta} + (a-b))$$

The result is seen in Fig. 4.18.

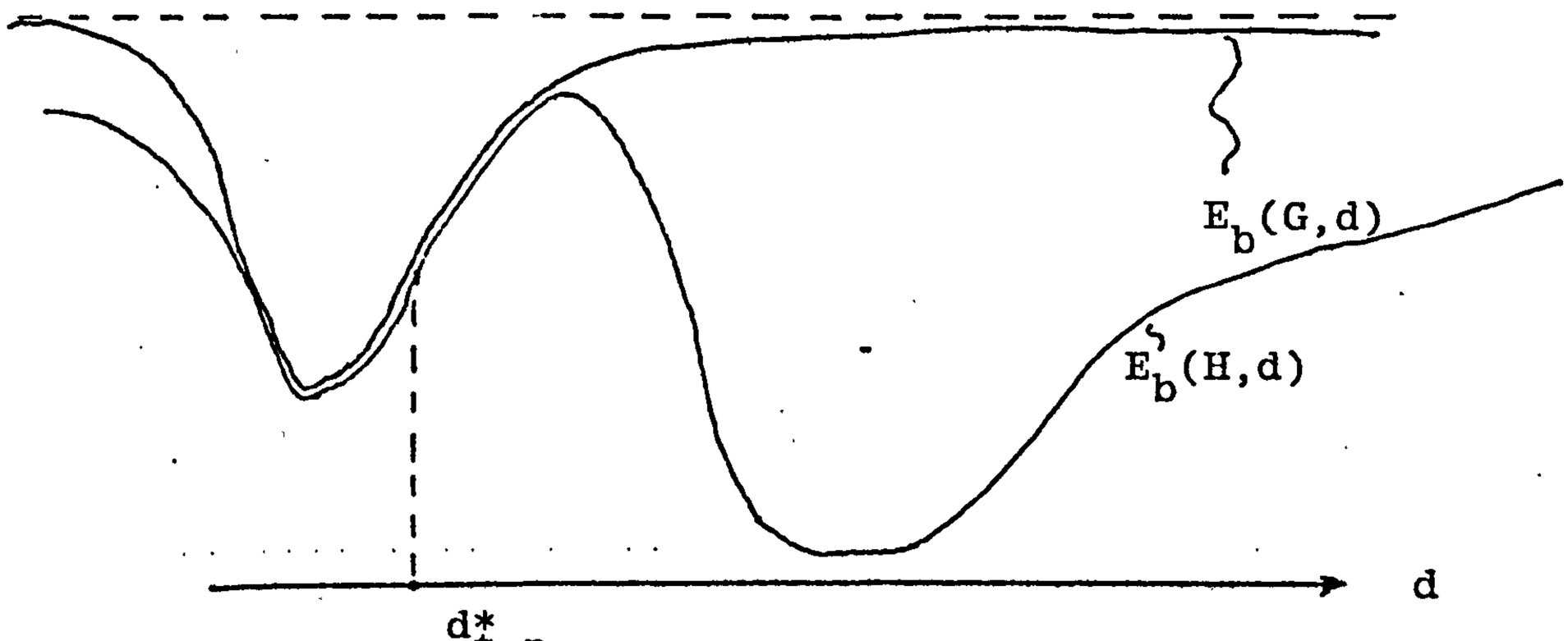


Fig. 4.18. The true expected loss and an approximation

It is quite clear that here again I have an analogous example to the cost of change rule. As max(a,b) → 0 so I will tend to a delay rule. I find it fascinating that there is this kind of link between cost of change and model uncertainty models. Obviously the topic needs more research, but even with a more complicated model, for uncertainty and other loss functions I would expect similar phenomena.

I note in passing that I can make similar arguments for the case when the *loss function* is only partially known, but I do not include this for want of space.

Summary

In this chapter I have outlined some of the principles of Catastrophe Theory. I then showed two very distinct ways of incorporating the theory into statistics giving two examples of these types of approach. Catastrophe Rules are explained and new ones introduced for my own purposes.

5. SOME COMMON CATASTROPHES OCCURRING IN STATISTICS

In the following chapter I will introduce the reader to a Catastrophe Theory approach to analysis of problems he will be familiar with in the field of Statistical Inference. This list is by no means exhaustive and is meant to emphasise the fact that examining the geometric rather than algebraic aspects of likelihood, posterior distributions, and posterior expected loss functions can elucidate problems that otherwise may seem very obscure.

The first catastrophes I want to examine are those induced by common likelihood functions, (i.e. occur due to the family of sample distributions that are chosen)

5.1. Bivariate Normal (means and variances known)

I can without loss of generality, assume that random variables X,Y each have means equal to zero and unit variances, but with unknown covariance ρ.

Then the joint sample distribution f(x,y) is given by

$$f(x,y) \propto \frac{1}{(1-\rho^2)^{\frac{1}{2}}} \exp -\tfrac{1}{2}\{(1-\rho^2)^{-1}(x^2-2\rho xy+y^2)\} \qquad 5.1.1.$$

which has log likelihood kernel $\ell(\rho)$ given by:

$$\frac{1}{n}\ell(\rho) = \tfrac{1}{2}[-\ell n(1-\rho^2) - (1-\rho^2)^2(S_x^2-2\rho S_{xy}+S_y^2) \qquad 5.1.2.$$

where $S_x^2 = \sum_{i=1}^{n} \frac{x_i^2}{n}$ $S_y^2 = \sum_{i=1}^{n} \frac{y_i^2}{n}$ $S_{xy} = \sum_{i=1}^{n} \frac{x_i y_i}{n}$.

Differentiating and rearranging (5.1.2) it is easily shown that the stationary points of $\ell(\rho)$ in (-1,1) are given by the equation

$$\rho(1-\rho^2) + (1+\rho^2)\, S_{xy} - \rho(S_x^2 + S_y^2) = 0. \qquad 5.1.3.$$

Putting $r = \rho - \frac{1}{3} S_{xy}$ I can write this in the form:

$$r^3 - br - a = 0 \qquad 5.1.4.$$

where $b = 1 + \frac{1}{3}(S_{xy})^2 - S_x^2 - S_y^2$ 5.1.5.

$a = \frac{1}{27} S_{xy} (-2(S_{xy})^2 - 9 (2 + S_x^2 + S_y^2))$ 5.1.6.

which is of course a cusp catastrophe with

a = normal factor

b = splitting factor.

Suppose now taking $\left.\begin{matrix} x_1 \ldots x_n \\ y_1 \ldots y_n \end{matrix}\right\}$ it is found that $S_{xy} = 0$

so that the analysis is simply that $a = 0$ 5.1.7.

$b = 1 - S_x^2 - S_y^2$ 5.1.8.

$r = \rho$. 5.1.9.

Case 1

If $S_x^2 + S_y^2 < 1$ then $b > 0$ and equation (5.1.4) is now

$\rho^3 - (1 - (S_x^2 + S_y^2) \rho = 0$ which has solutions

at 0 and $\pm (1 - (S_x^2 + S_y^2)^{\frac{1}{2}}$. It is easily checked that the point $\rho = 0$ is then a local minima of $\ell(\rho)$ hence a "worst" estimate of the log likelihood (See Fig. 5.1).

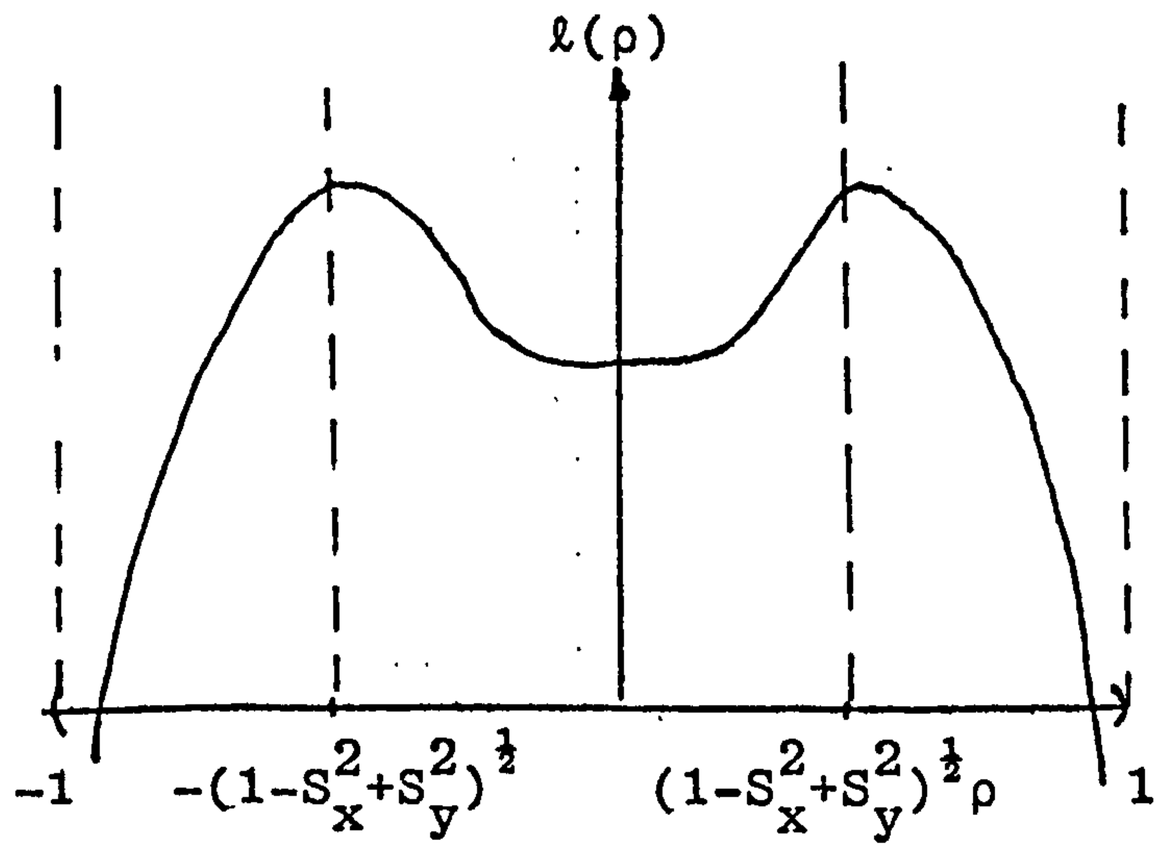


Fig. 5.1.

As $S_x^2 + S_y^2$ become larger, the maxima of $\ell(\rho)$ tend to zero from both sides, until at the point $S_x^2 + S_y^2 = 1$, they merge to give a unique maxima at O.

Case 2

$S_x^2 + S_y^2 > 1$, this maxima remains fixed at zero. Note that as $n \to \infty$ $S_x^2 + S_y^2 \to 2$ so in the limit the value $\rho = O$ will become the unique M.L.E.

Thus there is a cusp at $(S_x^2 + S_y^2, S_{xy}) = (1,O)$ and the description above can be summarised by a section of the cusp catastrophe where holding a = O (See Fig. 5.2).

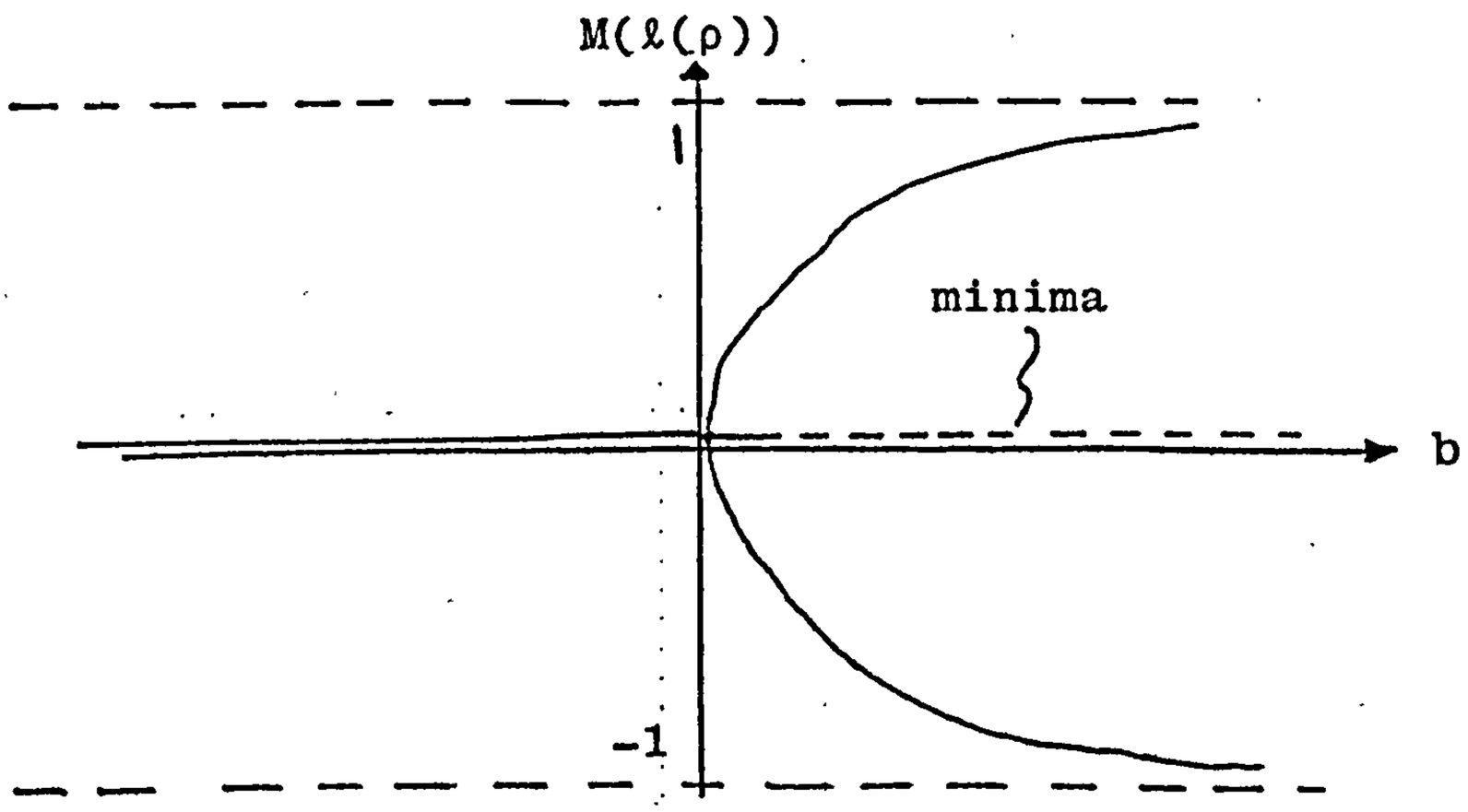


Fig. 5.2. Where $M(\ell(\rho))$ gives the stationary points of $\ell(\rho)$

$$b = 1 - (S_x^2 + S_y^2)$$

Notice that by local approximations at the cusp point I can use the new controls:

$$a^* = - S_{xy}$$
$$b^* = 1-(S_x^2 + S_y^2)$$

Thus, and this approximation is a fairly good one, on the control space I can represent the likelihood of ρ by the summarised form of Fig. 5.3.

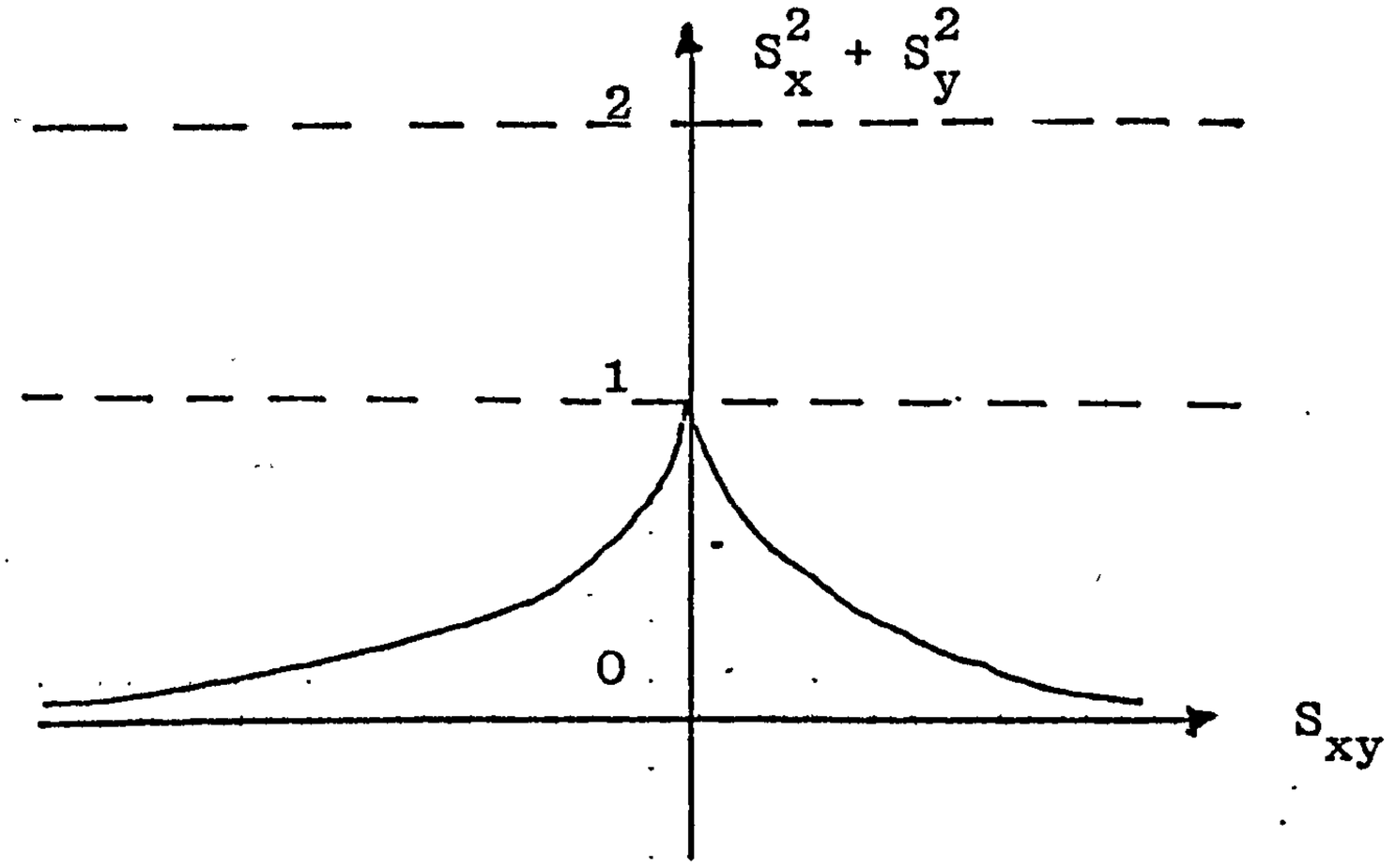


Fig. 5.3.

It follows that one would expect similar kinds of behaviour for all S_{xy} in a fairly large neighbourhood of O, the only difference being that the symmetry is lost.

The bimodality is not surprising since for if

$$\begin{cases} x_1 \dots x_n \\ \\ y_1 \dots y_n \end{cases}$$

have very close values, (all near zero) one would expect some dependence between X and Y, but the *form* of the dependence will not be clear.

The problem is not so obscure as it looks in the sense that there are many situations where the variance of X and Y are known fairly precisely, but their covariance structure is unknown. Hence in a Bayesian analysis with sharp priors on the variance one would expect the same sort of problems to arise.

This is the first simple example where control variables have been expressed in terms of the sufficient statistics, and characterisations of the graphs of minima of the likelihood have been used to summarise the data.

5.2. Simple Heirachical Normal Model

Let $\begin{cases} X_1 \ldots X_n \\ \\ \theta_1 \ldots \theta_n \end{cases}$ be random variables such that

$$X_i|\theta_i \sim n(\theta_i,1) \qquad 1 \leq i \leq n$$

$$\theta_i|V \sim n(O,V) \qquad 1 \leq i \leq n$$

Putting $S_\theta^2 = \sum_{i=1}^{n} \frac{\theta_i^2}{n}$ $S_x^2 = \sum_{i=1}^{n} \frac{x_i^2}{n}$, the log-likelihood kernel of

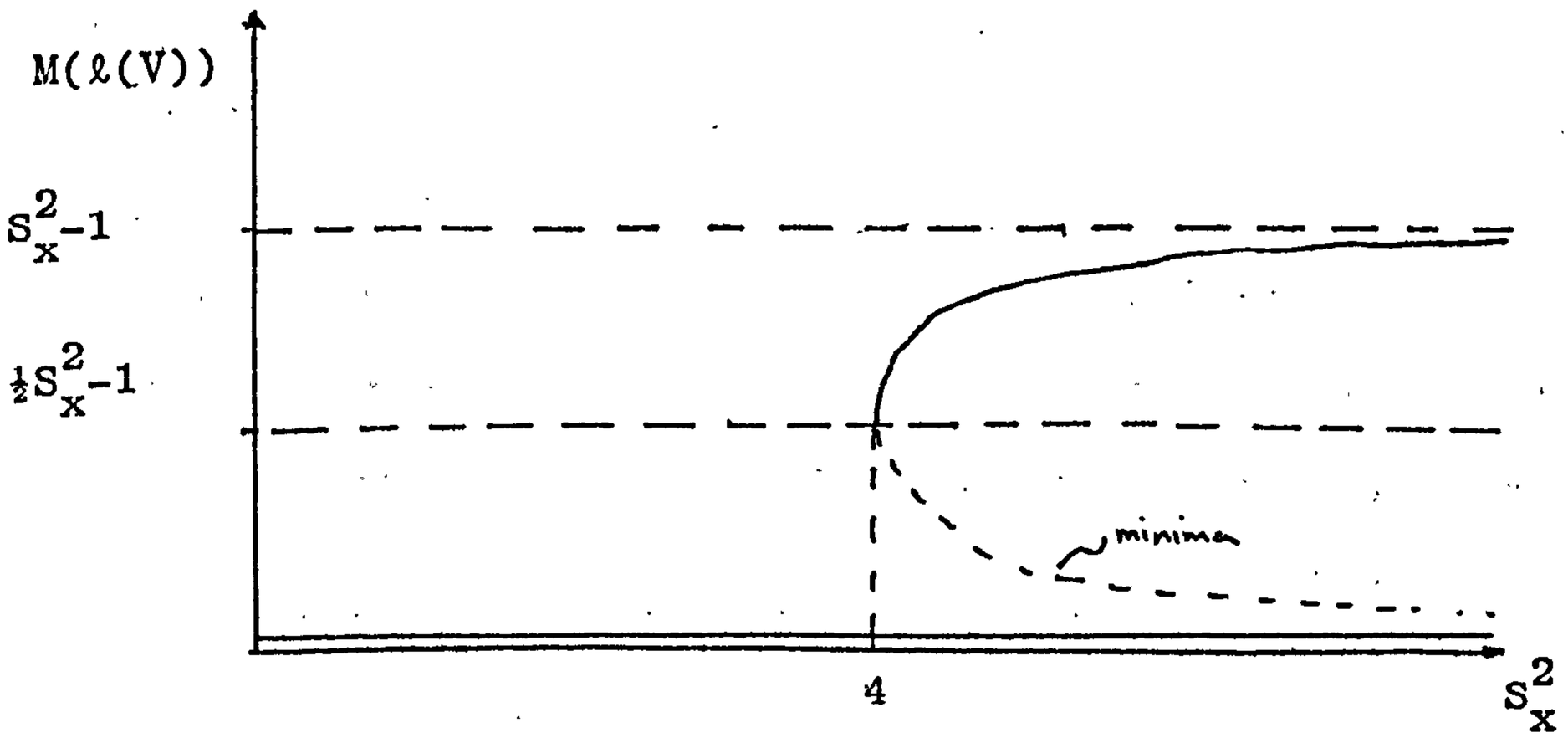
θ_i and V $1 \leq i \leq n$ given X_i $1 \leq i \leq n$ $\ell(\underset{\sim}{\theta},V)$ is given by:

$$-2\ell(\underset{\sim}{\theta},V) = n\left[S_\theta^2(1+V^{-1}) - 2\sum_{i=1}^{n} \frac{x_i\theta_i}{n} + \ell n\ V\right]$$

which has stationary points $V^* = \begin{cases} O & S_x^2 \leq 4 \\ \\ \{O \text{ and } \frac{(S_x^2 - 2)\pm(S_x^2 - 2)^2-4)^{\frac{1}{2}}}{2} & S_x^2 > 4 \end{cases}$

$$\theta_i^* = (1+V^{*-1})^{-1} x_i$$

The one sufficient statistic for V, S_x^2 thus gives a fold catastrophe



Although this type of model is in vogue at the moment, there are many criticisms of identifiability to consider. Leaving these aside it should be noted that the likelihood function is also unbounded and hence I cannot do a sensible Bayes analysis on it (See Chapter 2).

This is easily rectified by putting a lower bound on V. So let V be restricted to the range $[\varepsilon,\infty]$ where $\varepsilon > 0$. $M(\ell(V))$ then looks like Fig. 5.4.

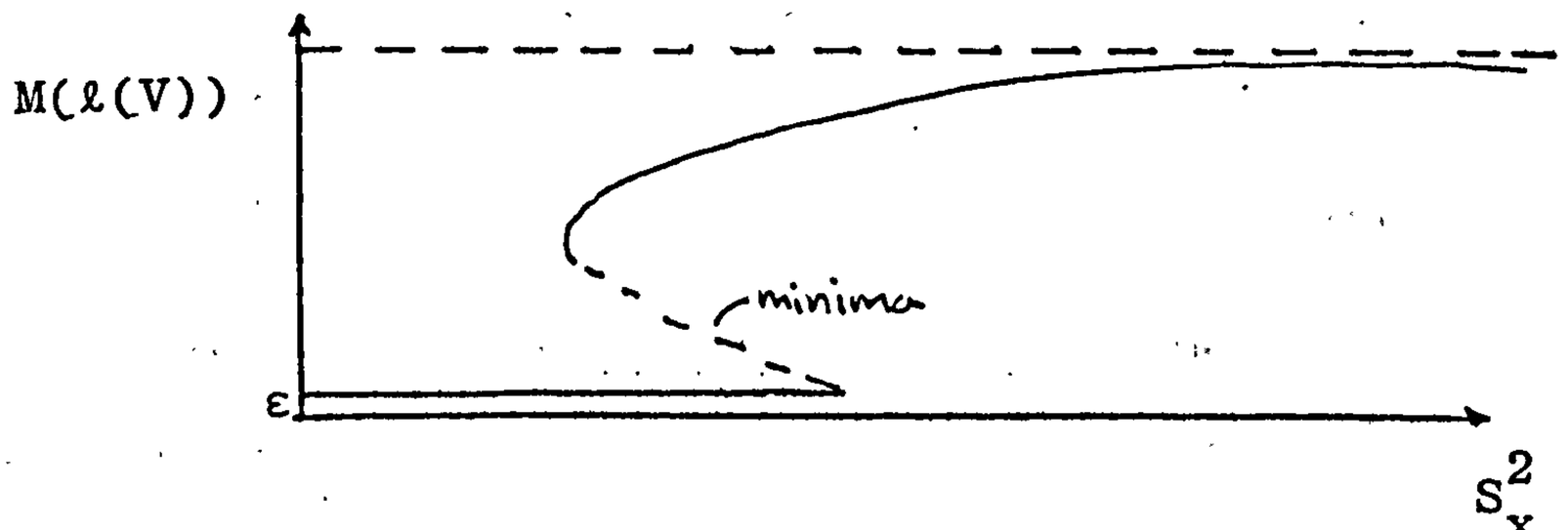


Fig. 5.4.

This is a fold catastrophe with boundary, and behaves very much like the cusp catastrophe.

In a Bayesian analysis therefore one msut be prepared to get at least a 2 modal distribution across the n+1 dimensional joint posterior distribution. The odd topology can of course be dodged by integrating out the θ_i's but usually the θ_i's are the quantities of interest, so I want to estimate the n+1-tuple $(\underset{\sim}{\theta},V)$, so marginalisati in a loss function approach is not really justified.

It happens that Catastrophes occur in a very large proportion of non-trivial likelihood functions on the advent of particular forms of data, topical examples are those arising in Time Series which are often grotesque. However I purposely leave these out of this exposition since their meaning is often not as clear as the archetypes above.

For more interesting Catastrophes arise in a Bayesian setting where the likelihoods concerned look very simple, but in which the posterior distribution goes bimodal. This occurs typically in 2 distinct types of situation:

(i) Specifying alternatives (Mixed models)

(ii) Contradiction of prior information by data (Product models).

Many models of the form (i) have been studied in detail by Dickey (1) and (2) and general theorems for this case will be left to the next chapter. Models of type (ii) occur when prior/likelihood has flat tails and have been studied by David (1). Elucidating the latter category is the following example.

5.3. The Normal Sample distribution-Student t prior distribution model

I can assume that without loss of generality the t-prior is normalised and that only one observation is taken. Hence let

$$X \sim n(\theta, W)$$

where θ has p.d.f. $f(\theta) \propto (\nu+\theta^2)^{-\frac{1}{2}(\nu+1)}$ 5.3.1.

Then the log-posterior kernel $\ell(\theta)$ is given by

$$\ell(\theta) = -\tfrac{1}{2}(\nu+1)\ \ell n(\nu+\theta^2) - \tfrac{1}{2}\ w^{-1}(\theta-x)^2 \qquad 5.3.2.$$

Differentiating and rearranging the formula for the stationary points is given by:

$$\psi^3 - b\psi - a = 0 \qquad 5.3.3.$$

where $\psi = \theta - \frac{1}{3}x$

$a = -\frac{1}{3}(k-2)x + 2/27\ x^3$

$b = \frac{1}{3}x^2 - (k+1)v$

$k = W(1+v^{-1})$

This is thus a Canonical Cusp Catastrophe with

a = normal factor

b = splitting factor

Note, in passing, that a represents the assymetry in the situation, whilst b is a function of the distance between the observation and prior estimate and hence the "split" of the information, as expected.

The Bifurcation set is given by

$$27\ a^2 \leq 4b^3$$

which on rearranging reduces to

$$t(z) = 4z^2 + (k^2-20k-8)z + 4(k+1)^2 \leq 0 \qquad 5.3.4.$$

where $z = \frac{x^2}{v} > 0$

Since t(z) is a nose down parabola (4) will have no solutions if t(z) has no roots. For t(z) to have real roots I must have that

$$(k^2 - 20k - 8)^2 \geq 4^3(k+1)^3 \qquad 5.3.5.$$

which on rearranging gives

$$k(k - 8)^3 \geq 0 \qquad 5.3.6.$$

i.e. $k \geq 8$ 5.3.7.

If this is the case then $k^2 - 20k - 8 < 0$ so $t(z) = 0$ has 2 solutions in $[0,\infty)$.

Hence this means that, translating into the original notation, bimodality will occur for some values of x iff

$$W \geq \frac{8v}{v+1}$$

In this case bimodality of the posterior of θ will occur for values of x lying in two open intervals symmetric about the zero point (i.e. the prior estimate of θ). In fact moving x on $\mathbb{R}$ I can induce a flow on the control space which is given in Fig. 5.5. Note that as $|x| \to \infty$ control values tend to the bifurcation lines.

The same phenomena of course is true from t-sample distribution and normal prior or for "vague prior" and likelihood made up of 2 observations, are from a t-distribution and are from a normal distribution. Again I will refer the reader to David (1) for a fuller discussion of this.

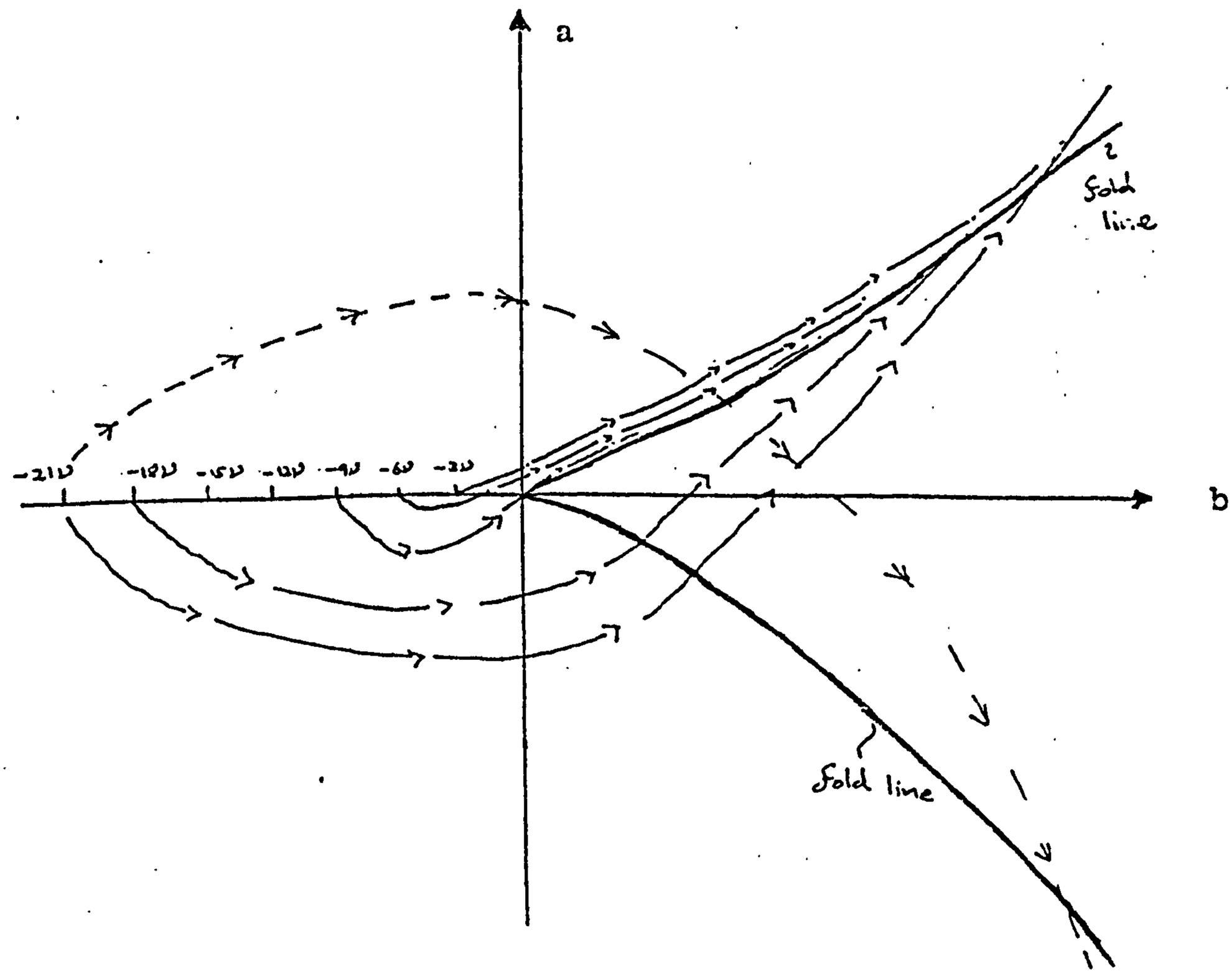


Fig. 5.5. Solid lines shows path of controls as $x > 0$ increasing

Dotted line shows path of controls as $x < 0$ decreases

5.4. The Student -t prior and sample distribution

To start with assume that I take one observation x, from a random variable X which has Student-t p.d.f.

$$f_2(x|\theta) \propto (\nu_2 + W_2^{-1}(\theta-x_1)^2)^{-\frac{1}{2}(\nu_2+1)} \qquad 5.4.1.$$

Let the prior p.d.f. of θ, $f_1(\theta)$ also be t-distributed

$$f_1(\theta) \propto (\nu_1 + W_1^{-1}(\theta-x_o)^2)^{-\frac{1}{2}(\nu_1+1)} \qquad 5.4.2.$$

The log-posterior kernel is then:

$$-\tfrac{1}{2}(\nu_1+1)\ \ell n(\nu_1+W_1^{-1}(\theta-x_o)^2) - \tfrac{1}{2}(\nu_2+1)\ \ell n(\nu_2+W_2(\theta-x_1)^2) \qquad 5.4.3.$$

which has stationary points defined by the equation:

$$(\nu_1+1)W_1^{-1}(\theta-x_o)(\nu_2+W_2^{-1}(\theta-x_1)^2)+(\nu_2+1)W_2^{-1}(\theta-x_1)(\nu_1+W_1^{-1}(\theta-x_o)^2)$$

$$= 0 \qquad 5.4.4.$$

So again this gives the *Canonical Cusp Catastrophe* since this is a cubic. If for simplicity I assume $\nu = \nu_1 = \nu_2$, then I can rearrange (5.4.4) in the form

$$\psi^3 - b\psi - a = 0 \qquad 5.4.5.$$

where $\psi = \theta - \bar{x}$

$$a = \frac{1}{4}\nu(W_2-W_1)(x_o-x_1)$$

$$b = S^2 - \bar{W}\nu$$

$$\bar{x} = \tfrac{1}{2}(x_o+x_1)$$

$$\bar{W} = \tfrac{1}{2}(W_1+W_2)$$

$$S^2 = \tfrac{1}{2} \sum_{i=0,1} (x_o-\bar{x})^2.$$

With the extra symmetry of this model over the previous one it is very easy to see that the splitting factor b represents the symmetrical "split" in the model whilst the normal factor a incapsulates all assymetrical components of the model.

Notice the difference between the normal product and the t-product above. Whilst the former never experiences a split the latter always has some observation which will make the posterior distribution go bimodal. Notice also that for $W_1 = W_2$ and $\nu > 1$ the posterior mean of $\theta, \bar{x}$ goes from being a unique posterior mode to the unique *antimode* as the distance between observation and prior estimate increases.

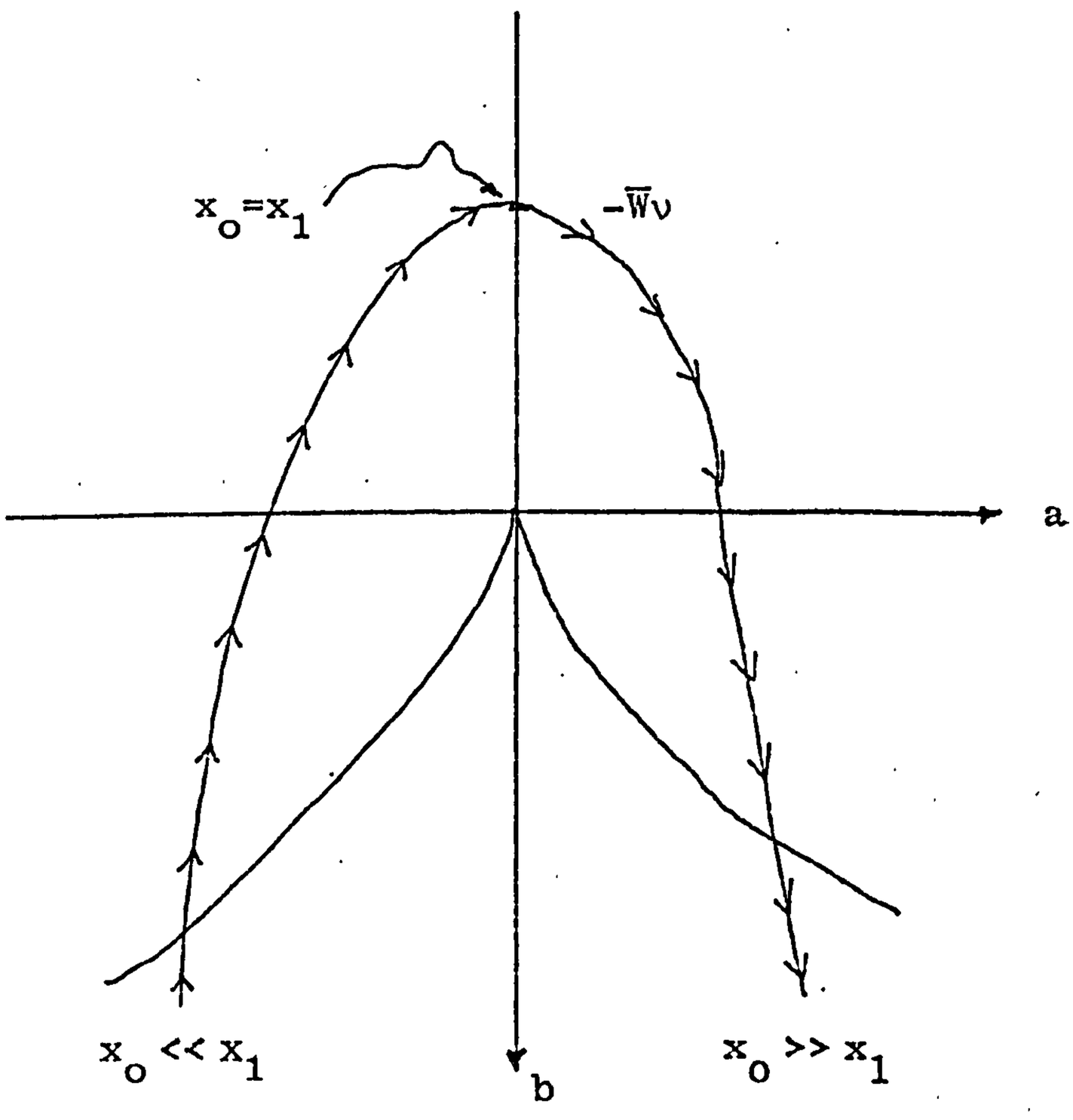


Fig. 5.6.

Trajectory across control space as $x_o - x_1$ increases

Higher order catastrophes occur when I take more than one observation. For simplicity assume $W_1 = W_2$. If I observe a new value x_2 of the random variable X, then log-posterior kernel then becomes

$$-\tfrac{1}{2}(\nu+1)(\ell n(r+(\theta-x_o)^2)+\ell n(r+(\theta-x_1)^2)+\ell n(r+(\theta-x_2)^2) \qquad 5.4.6.$$

where $r = \nu W$

which has stationary values on the manifold given by the equation:

$$(\theta-x_o)(r+(\theta-x_1)^2)(r+(\theta-x_2)^2)+(\theta-x_1)(r+(\theta-x_o)^2)(r+(\theta-x_2)^2)$$

$$+(\theta-x_2)(r+(\theta-x_o)^2)(r+(\theta-x_1)^2) = 0 \qquad 5.4.7.$$

This quintic is of course an example of a *Canonical Butterfly Catastrophe*.

(The Butterfly Catastrophe is also obtained with assumptions

$\{ w_1 \neq w_2 \neq w_3$ but the algebra is more messy)

$\{ \nu_1 \neq \nu_2 \neq \nu_3$

With some rearrangement this can be reduced to.

$$\psi^5 - d\psi^3 - c\psi^2 - b - a = 0.$$

where $\psi = \theta - \bar{x}$

$$\bar{x} = \frac{1}{3}\sum_{i=0}^{2} x_i$$

$$d = 2(S^2-\nu W)$$

$$c = 3u$$

$$b = -(\nu^2W^2+T) < 0$$

$$a = -u(1+\tfrac{1}{2}S^2).$$

$$S^2 = \frac{1}{3}\sum_{i=0}^{2} (x_i-\bar{x})^2$$

$$T = \frac{1}{3}([\bar{x}-x_o)(\bar{x}-x_1]^2+[(\bar{x}-x_o)(\bar{x}-x_2)]^2+[(\bar{x}-x_1)(\bar{x}-x_2)]^2)$$

$$u = (x_1-\bar{x})(x_2-\bar{x})(x_3-\bar{x}).$$

Hence the Butterfly Factor d varies with S^2 the bias and normal factors c and a respectively with a statistic u defined above, which synthesises all the assymmetry in the model, and the splitting factor b is a function of the rather odd statistic T. It is instructive to check how I move over the control space for variance changes in x_1x_2 and x_3.

It should be noted that with symmetry (i.e. u = O), the statistic T will become a fourth sample moment and S^2 is a second sample moment, of course. So now order moments seem to be useful summary statistics in this situation.

Of course, the most interesting classification is in terms of the posterior expected loss, since as mentioned in Chapter 4 the use of Catastrophe Theory is in classifying potential functions. I will go into this classification generally in Chapters 6 and 7 but for the purposes of these examples, examination of properties with respect to step loss functions seems the most simple choice.

Definition

The *asymmetric step loss function* L(B,A,(-d)) is defined by

$$L(B,A,(-d)) = \begin{cases} O & |\theta-d| \leq B \\ 1 & (\theta-d) > B \\ A & (\theta-d) < -B \end{cases} \quad \text{where } A \geq 1,\ B > O.$$

I will call B the *guage* of the loss function and A the *asymetry constant*.

If $F(\theta)$ is the posterior distribution it is easily seen that

$$E(F,B,A,d) = A\,F(d-B) - (1 - F(d+B)) \qquad 5.5.1.$$

where $E(F,B,A,d)$ is the expected loss with respect to $L(B,A,(\theta-d))$ and $F(\theta)$. Hence the stationary points d^* of (5.5.1) satisfy

$$A\,f(d^*-B) = f(d^*+B). \qquad 5.5.2.$$

5.5. The t-distribution under asymmetric step loss

The t-distribution occurs widely as a posterior distribution in Bayesian analysis. The most common occurrence the unknown mean and variance/normal sample distribution conjugate analysis (De Groot,(1) for which the posterior marginal distribution for the mean is a Student-t.

Suppose $f(\theta) \propto (r + \theta^2)^{-\frac{(r+1)}{2}}$ where $r > 0$.

Then using (5.5.2) and rearranging the equation, the stationary points under $L(B,A,(\theta-d))$ are given by:

$$\psi^2 - a = 0 \qquad 5.5.3.$$

where $\psi = d^* + c$

$$a = c^2 + r + B^2$$

$$c = B\left[\frac{A^{2(r+1)^{-1}}+1}{A^{2(r+1)^{-1}}-1}\right]$$

Hence again I have the *Canonical Fold Catastrophe*.

Fig. 5.7. gives the evolution of E(d) for different values of a.

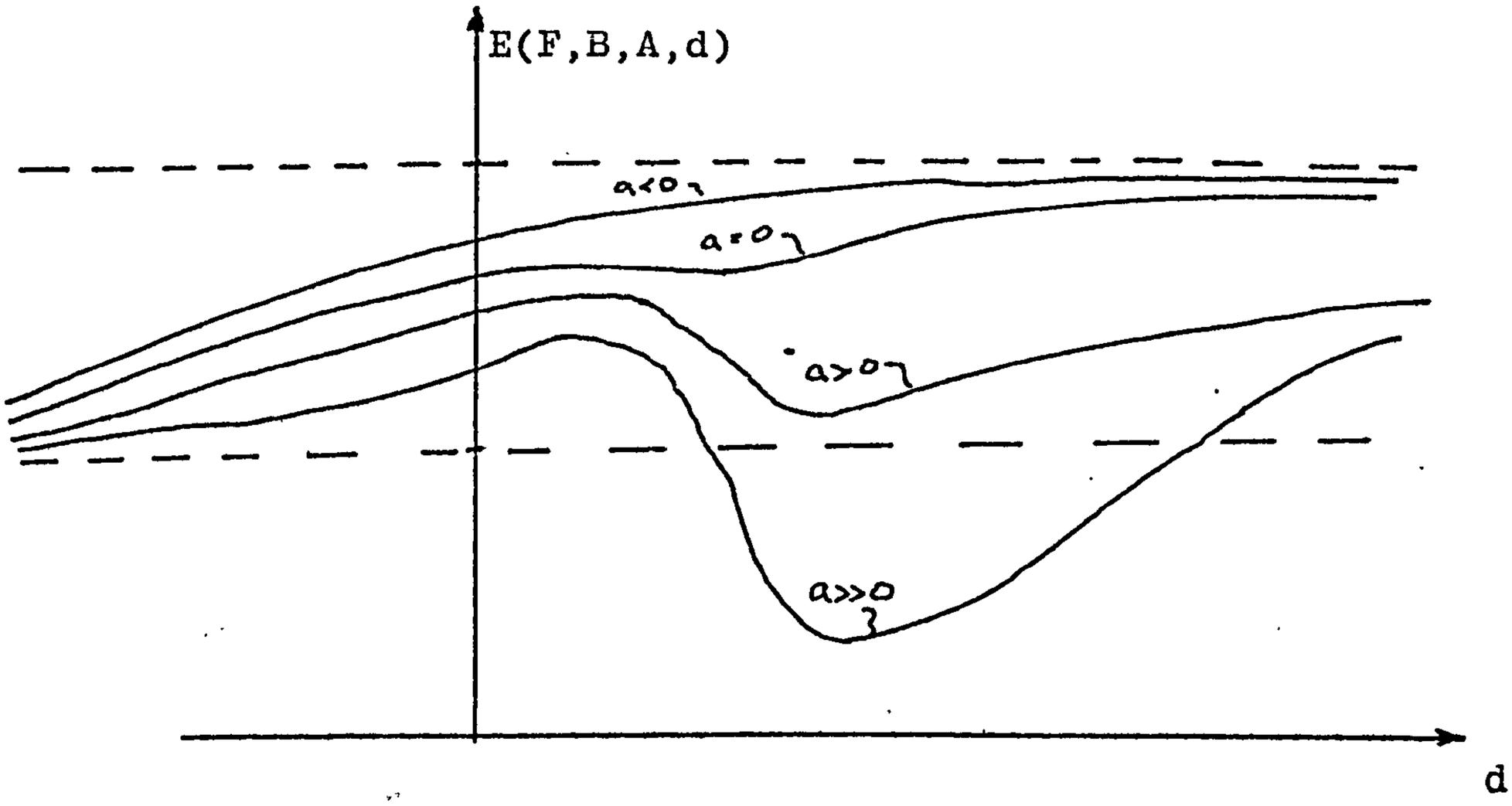


Fig. 5.7.

For a < 0, the Bayes decision is at $-\infty$ indicating that with a large amount of uncertainty r the optimal decision is to cut losses and go for a loss of 1 unit. As a increases to become positive there will be a point when the Bayes decision maker suddenly decides that it is worth while making a guess and hence a Catastrophe occurs. Notice that this sort of phenomena is not a property of this particular loss function and smooth ones will give exactly the same sort of qualitative behaviour provided they are asymmetric.

This illustrates another point about the partitioning of distribution functions by their $\phi(b)$ functions as described in Chapter 3. Once I allow for asymmetrical loss functions, qualitative behaviour inside each member of the partition will change radically from one another. For example the Student-t and usual distribution have the same $\phi(b)$ function, but whereas the t-distribution gives a fold catastrophe with asymmetric step loss, the normal gives a linear function for its stationary points.

It is amusing to note that the Bayes decision does not exist for a < 0, hence the necessity for the compactness of the decision space in Chapter 2. If I insist on a lower bound for my decision I of course obtain the *Fold Catastrophe with Boundary* which as mentioned before, behaves very much like the *Cusp Catastrophe*.

Referring back to equation (5.3.3) it is now obvious that the *possibility* of a Fold Catastrophe depends solely on the degrees of freedom for any specific step loss function, and as these increase so the decision will stabilise. Hence if $F(\theta)$ represents the marginal posterior distribution of the mean of a normal sampling distribution, the *possibility* of a Fold Catastrophe depends on the number of observations I have taken. If these are small in number, and their sample variance is large, then I am likely to choose the 'cut-loss' decision at $-\infty$. If the sample variance is small and the number of observations large then I am likely to be prepared to make a 'proper' decision. Equation (5.5.3) thus sums up the way in which the topology of $E(d)$ behaves for changes in the sufficient statistics of this specific problem.

5.6. The t-product under step loss

Suppose prior on mean θ has p.d.f. $f_o(O)$ given by

$$f_o(\theta) \propto (\nu W_o + (\theta - x_o)^2)^{-\frac{1}{2}(\nu+1)}$$

and an observation x_1 is taken from a random variable X having p.d.f. $f_1(\theta)$ given by

$$f_1(\theta) \propto (\nu W_1 + (\theta - x_1)^2)^{-\frac{1}{2}(\nu+1)}$$

Using equation (5.5.2), the stationary points of expected asymmetric loss are given by the manifold.

$$(\nu W_o+(\theta-B-x_o)^2)(\nu W_1+(\theta-B-x_1)^2)=A^*(\nu W_o+(\theta-B-x_o)^2)(\nu W_1+(\theta+B-x_1)^2)$$

$$\text{where } A^* = A^{\frac{2}{\nu+1}}$$

Notice for A > 1 this is a quantic in θ and hence an example of the *Swallowtail Catastrophe.*

If A = 1 however, then it is easily seen that the quartic term vanishes which after some rearrangement can be written

$$\psi^3 - b\psi - a = 0$$

$$\text{where} \quad a = \tfrac{1}{4}\nu\,(W_1-W_o)(x_o-x_1)$$

$$b = S^2 - \overline{W}\nu - B^2.$$

$$\psi = \theta - \overline{x}$$

$$\overline{x} = \tfrac{1}{2}(x_o+x_1)$$

$$S^2 = \tfrac{1}{2} \sum_{i=0,1} (x_o-\overline{x})^2$$

$$\overline{W} = \tfrac{1}{2}(W_o+W_1)$$

which is exactly the same form as equation (5.4.5) except that a quantity B^2 has been subtracted from the splitting factor b. Hence the topology of this potential function is very similar to the topology of the posterior distribution.

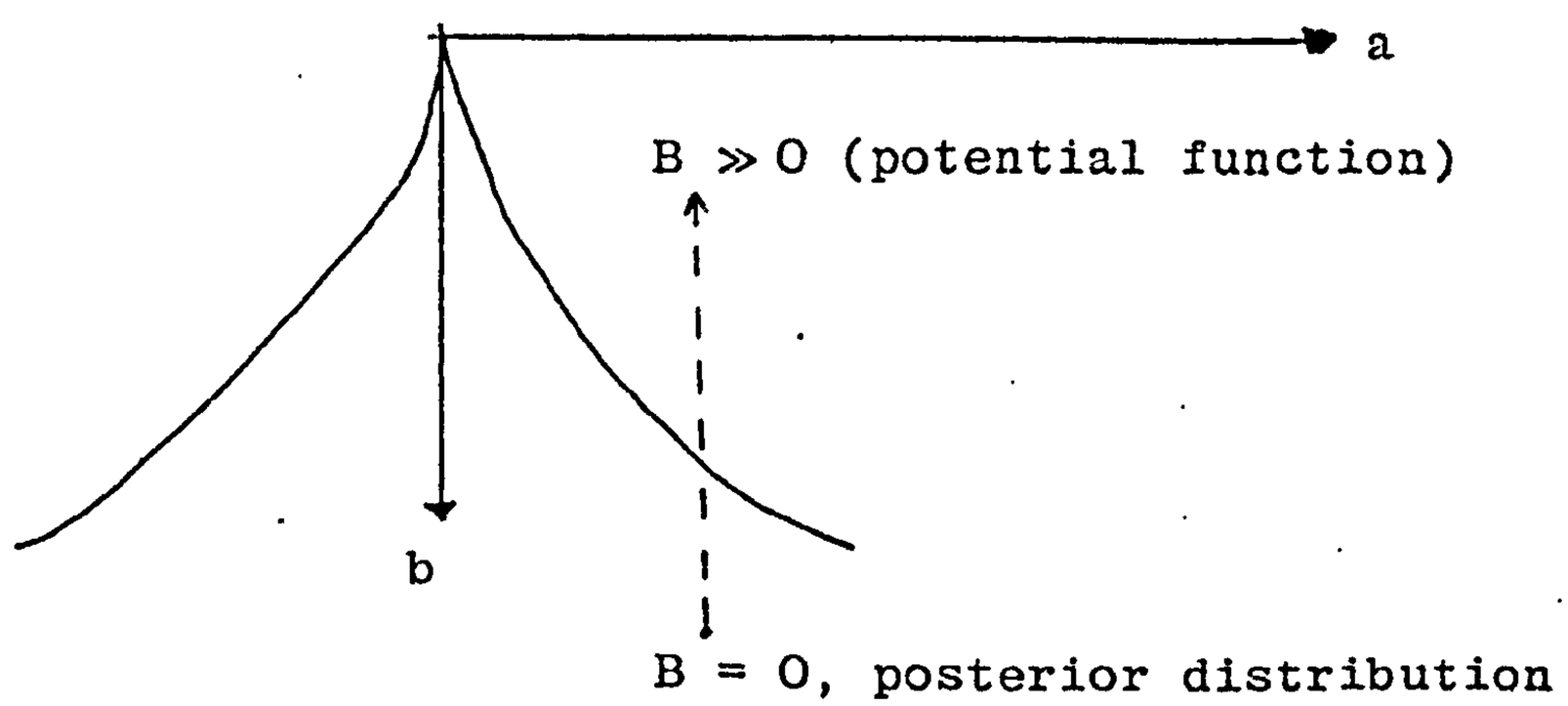


Fig. 5.8.

Summary

Examples of some canonical forms of Catastrophes occurring in Statistics have been given. It has been shown that Catastrophes representing conflict between prior, sample information and requirements (i.e. loss structure) can simply illucidated using the t-distribution.

6. A CLASSIFICATION OF EXPECTED LOSS ARISING FROM GENERAL DISTRIBUTION FUNCTIONS

Introduction

In this chapter I succeed at least partially in classifying the sorts of forms of Catastrophes in the topology of expected loss arising from certain combinations of a general distribution function with bounded symmetric loss.

The first step is obviously to look at the topology of expected loss arising from unimodal distributions. The closest associated work I can find was done by Ibraginov (1) way back in 1956, where here he was concerned with the problem of when the sum of two independent unimodal random variables was again unimodal. I however cannot proceed by the analogy outlined in Chapter 3 because I have an extra symmetry condition on $L(\theta-d)$, but the work in Chapter 3 now comes in very useful and Lemma 3.3.1 makes the results quite simple to prove.

In the second part of the chapter I tentatively start a classification of multimodal distributions under step loss functions.

6.1. Topology of Bayes decisions arising from unimodal distributions

It would be tempting to speculate following Theorem 3.2(i) that unimodal distributions with bounded symmetric loss functions gave rise to expected loss functions with only one minima. This however is surprisingly not the case.

Unless otherwise stated I will use the following notation:

All distributions $F(\theta)$ will be assumed properly unimodal, twice differentiable with mode at zero. Parameter θ and decision d will be in the same space $\mathbb{R}$.

The loss function L will be bounded above by 1, symmetric in $(\theta-d)$ and satisfying the condition in Theorem 3.3, (i.e. with respect to $F(\theta)$ there is no decision d_1 such that the associated expected loss $E(d_1) = 0$).

Let $E_b(d) = 1 - F(d+b)^{-} + F(d-b)$

(the expected loss with respect to posterior distribution $F(\theta)$ and $S_b(\theta-d)$, the step loss with guage b defined in Chapter 3)

$S(F)$ be the extended support of F (defined in Chapter 3)

Let $[d_1(f),d_2(f)] \leq S(F)$ be the interval obtained in Theorem containing all turning point of expected loss in $S(F)$ with respect to posterior distribution $F(\theta)$ and loss functions of the form described above.

Let $E_L(d)$ = expected loss with respect to $(\theta-d)$ and $F(\theta)$

(L will be omitted if no confusion is likely to arise).

$\tau(\theta) = f'(\theta)\,[f(\theta)]^{-1}$ (Fisher's Score).

This first theorem gives sufficient conditions on $F(\theta)$ for one minima of expected loss only to appear when a loss function of the above form is used.

Theorem 6.1.

If $\tau(\theta)$ is strictly decreasing on $[d_1(F),d_2(F)]$ and

$\tau(\theta) > \tau(d_1)$ $\quad \theta < d_1(F)$ for all $\theta \in (S(F))^{o}$, then

$\tau(\theta) < \tau(d_2)$ $\quad \theta > d_2(F)$

$E_L(d)$ has exactly one minima regardless of the loss function of the above form which is chosen.

Proof

By Lemma 3.3.1 I can write

$$E(d) = \int_{\mathbb{R}_{>0}} E_b(d)\, dG(b)$$

where G(b) is the distribution function on $\mathbb{R}_{>0}$ defined in the above Lemma. It is now sufficient to prove that for any L(θ-d) of the above form E(d) has no maxima or turning point on S(F)

Well, suppose d^* is a maxima or turning point then it certainly must satisfy

$$E'(d) = \int_{\mathbb{R}_{>0}} (f(d^*-b) - f(d^*+b))dG(b) = 0 \text{ and}$$

$$E''(d^*) = \int_{\mathbb{R}_{>0}} (f'd^*-b) - f'(d^*+b))dG(b)$$

$$= \int_{\mathbb{R}_{>0}} (\tau(d^*-b)f(d-b) - \tau(d^*+b)f(d+b))dG(b) \leq 0.$$

So in particular

$$E''(d^*) - \tau(d^*)E'(d^*)$$

$$= \int_{\mathbb{R}_{>0}} ([\tau(d^*-b) - \tau(d^*)]f(d^*-b) - [\tau(d^*+b) - \tau(d^*)]f(d^*+b))dG(b) \leq 0$$

But under the conditions of the theorem and the constraints put on L(θ-d), the integrand is always non negative and positive for some values of b of measure > 0 with respect to G(b).

Hence $E_L(d)$ has exactly one turning point, a minima, as S(F). □

It is interesting to see that the conditions of the theorem depend on the derivative of Edward's support function (Edwards (1)), being monotonic decreasing with θ.

Examples of distributions satisfying the conditions of Theorem

1) All Symmetric distributions that are unimodal

This is because $d_1 = d_2$ = point of symmetry and $\tau(\theta) > 0$ for all $\theta < \mu$ -the point of symmetry and $\tau(\theta) < 0$ $\theta > \mu$

2) Gamma distribution

The Gamma probability density function

$f(\theta) \propto \theta^{\alpha-1} \exp(-\beta\theta)$ $\theta > 0$ $\beta > 0, \alpha \geq 1$

has $\tau(\theta)$ given by

$\tau(\theta) = (\alpha-1)\theta^{-1} - \beta$ which satisfies the conditions of the theorem.

3) Beta distribution

The Beta probability density function

$f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ $\alpha,\beta \geq 1$ $0 < \theta < 1$

has $\tau(\theta)$ given by

$\tau(\theta) = (\alpha-1)\theta^{-1} - (\beta-1)(1-\theta^{-1})$ which again satisfies the conditions of the theorem.

Note that in the proof of the above theorem I have used the fact that a sufficient condition to ensure E(d) has exactly one minima in $S(F)$ is that there exists a $K \in \mathbb{R}$ such that for all $b \in \mathbb{R}_{>0}$ and $d \in [d_1(F), d_2(F)]$

$$E_b''(d) + kE_b'(d) > 0. \qquad 6.1.1.$$

Obviously the smaller the interval $[d_1(F), d_2(F)]$ (heuristically the more symmetric $f(\theta)$ is) the easier it is to satisfy the above equation. Sometimes the following Corollary is easier to prove for specific cases and utilises (6.1.1) when $k = 0$. Remember I have assumed $F(\theta)$ have mode at zero.

Corollary 6.1.1.

Let $d_1 = 0$ and δ_1 be the first postive root of $f''(\theta) = 0$

If (i) $d_2 < \delta_1$

(ii) $f'(d_2-b) > f'(d_2+b) \qquad 0 < b < d_2$

(iii) $f''(\theta) = 0$ has at most two solutions in $(d_2, 2d_2)$

then E(d) has exactly one minima on S(F)

Proof.

(ii) implies that $f'(d_2-b)-f'(d_2+b) > 0$ for all $b \in \mathbb{R}_{>0}$ since

$f'(\theta) > 0 \qquad \theta < 0$

$f'(\theta) > 0 \qquad \theta > 0$ since $f(\theta)$ is unimodal with mode at 0.

By condition (iii) $f'(d-b)-f'(d+b)$ will take its minimum value $d \in [0,d_2]$ when $d = d_2$.

Hence $E''(d) > 0$ for all $b \in \mathbb{R}_{>0}$ $d \in [0,d_2(f)]$ □

Now I begin a search for catastrophes on properly unimodal distributions $f(\theta)$. A necessary condition is that I obtain 2 minima of expected loss and hence at least are local maxima on S(F). With the differentiability assumption this means that I need to find a $d^* \in S(F)$ with the property

$$E'(d^*) = 0$$

$$E''(d^*) < 0$$

This provokes the following theorem.

Theorem 6.2.

Suppose there exists $b_1, b_2 \in \mathbb{R}_{>0}$ and a $d^* \in S(F)$ with the following properties:

(i) $E'_{b_1}(d^*) > 0$

(ii) $E'_{b}(d^*) < 0$

and there exists a $k \in \mathbb{R}$ such that

(iii) $E''_{b_1}(d^*) + kE'_{b_1}(d^*) < 0$

(iv) $E''_{b_2}(d^*) + kE'_{b_2}(d^*) \leq 0.$

Then there exists a loss function $L(\theta-d)$

$$L(\theta-d) = \alpha S_{b_1}(\theta-d) + (1-\alpha)S_{b_2}(\theta-d) \qquad 0 < \alpha < 1$$

such that d^* is a local maxima of expected loss.

Proof

By (i) and (ii) I can choose an α, $0 < \alpha < 1$ such that

$$E'_L(d^*) = \alpha E'_{b_1}(d^*) + (1-\alpha)E'_{b_2}(d^*) = 0 \qquad 6.1.2.$$

Also

$$E''_L(d^*) = \alpha E''_{b_1}(d^*) + (1-\alpha)E''_{b_2}(d^*) < 0$$

by (iii), (iv) and (6.1.2), □

which has the following Corollary.

Corollary 6.2.1.

Suppose $f(\theta)$ has extended support $[R\ \infty)$ where $R < 0$ and

$\tau(\theta) \to 0$ as $\theta \to \infty$

Then $F(\theta)$ has a maxima of expected loss with respect to some loss functions of the form used in the previous theorem on $S(F)$.

Proof

All I need to do is to show that the four conditions of the previous theorem are satisfied.

Since $f(\theta)$ has extended support $[R,\infty)$ $[d_1(F),d_2(F)] < [O\infty)$.

Let $d^* \leq (O,\infty)$, then if $O < b_1 < d^*$,

$$E'_{b_1}(d^*) = f(d^*-b_1) - f(d^*+b_1) > O$$

since $f'(d^*) < O$, so condition (i) of Theorem 6.2. is satisfied.

$$E''_{b_1}(d^*) + kE'_{b_1}(d^*) = -2b_1 \left\{ \frac{f'(d^*+b_2)-f'(d^*-b_1)}{2b_1} + k\frac{f(d^*+b_1)-f(d^*-b_1)}{2b_1} \right\} \qquad 6.1.3.$$

So if there is a $k=k^*$ such that

$f''(d^*)+k^*f'(d^*) > O$, then equation (6.1.3) implies that for small b_1 condition (iii) of Theorem (6.2) is satisfied.

Well, let d^* be any number between the first and the second (if it exists, otherwise ∞) solution of

$$f''(d^*) = O.$$

Then $f''(d^*) > O.$

It follows that there exists a small positive $k = k^*$ such that

$f''(d^*)+k^*f'(d^*) > O$ and so condition (iii) of Theorem (6.2) holds.

Obviously if b_2 is chosen such that

$$b_2 > d^* - R$$

then

$$E'_{b_2}(d^*) = -f(d^*+b_2) < O$$

so (ii) of Theorem (6.2) is satisfied.

Finally

$$E''_{b_2}(d^*) + k^* E'_{b_2}(d^*) = -(f'(d^*+b_2) - k^* f(d+b_2))$$

$$= -f(d^*+b_2)\ (\tau(d^*+b_2) + k^*)$$

< 0 for large enough b_2, since $\tau(x) \to 0$ and $k^* > 0$. Hence (iv) of Theorem (6.2) is satisfied. Hence the Corollary is proved. □

Transform $F(\theta)$ linearly so that it lies on the range $[0,\infty)$ and the following standard distributions on $\mathbb{R}_{>0}$ provide examples of this corollary holding.

Examples of distributions satisfying the conditions of Theorem (6.2)

1). All distributions with inverse polynomial tails properly unimodal with support $[0,\infty)$

So for example the F-distribution and Inverted Gamma are prime examples.

2). Lognormal distribution. □

It is also clear that the above Corollary generalises to distributions on $\mathbb{R}$. For example.

Corollary 6.2.2.

Suppose $F(\theta)$ is properly unimodal mode 0 with extended support $S(F) = \mathbb{R}$ and

(i) The right hand tail of $f(\theta)$ is an inverse polynomial

(ii) The left hand tail of $f(\theta)$ is an inverse polynomial of higher order or exponential of any order.

or vice-versa. Then there exists a loss function $L(\theta-d)$ of the prescribed form such that $E_L(d)$ has at least one local maxima.

Proof

(i) and (ii) imply that $[d_1(F),d_2(F)] \supset (0,\infty)$ so the arguments to prove Conditions (i) and (iii) are proved in exactly the same way as in Corollary (6.2.1), bychoosing b_1 small.

Clearly $E'_{b_2}(d^*) = f(d-b_2)-f(d+b_2) > 0$ for large enough b_2 so (ii) of Theorem (6.2) is satisfied.

Finally

$$E''_{b_2}(d^*) + k^*E'_{b_2}(d^*) = -f(d^*+b_2)\ [k + \tau(d^*+b_2) - \psi(b_2)]$$

where

$$\psi(b_2) = \frac{f(d^*-b_2)}{f(d^*+b_2)}\ [\tau(d^*-b_2) + k]$$

Clearly under conditions (i) and (ii) above $\psi(b_2) \to 0$ as $b_2 \to \infty$, hence for large enough b_2

$$E''_{b_1}(d^*) + k^*E'_{b_2}(d^*) < 0.$$

Hence condition (iv) of Theorem (6.2) is again satisfied.

The result follows. □

So I have found in many common distributions a loss function such that the expected loss has at least 2 minima. A question remains: Is the class of all such loss fucntions for any fixed distriubtion function pathological in some sense? The evidence strongly suggests that they are not. For example rather than the loss function in Theorem (6.2) I could instead use a loss function of the form

$$L(\theta-d) = \alpha \int_{\mathbb{R}_{>0}} S_b(\theta-d)dG(b)+(1-\alpha) \int_{\mathbb{R}_{>0}} S_b(\theta-d)dG_2(b)$$

where $0 < \alpha < 1$ and $G_1(b)$, $G_2(b)$ are C^∞ distribution functions with non zero weight respectively in

$b \in (0,\varepsilon)$ where ε is small

$b \in (M,\infty)$ where M is large

It is easy to check again that Theorem (6.2) would still hold. In fact it is not difficult to see heuristically that I could extend this class much further by allowing positive loss in the region (ε,M) which is small compared with the weight on $(0,\varepsilon) \cup (M,\infty)$. Hence for fixed distribution fucntions satisfying the conditions of Corollary (6.2.1) all loss functions looking something like Fig. 6.1. (not unreasonable?) are likely to cause non uniqueness of minima problems.

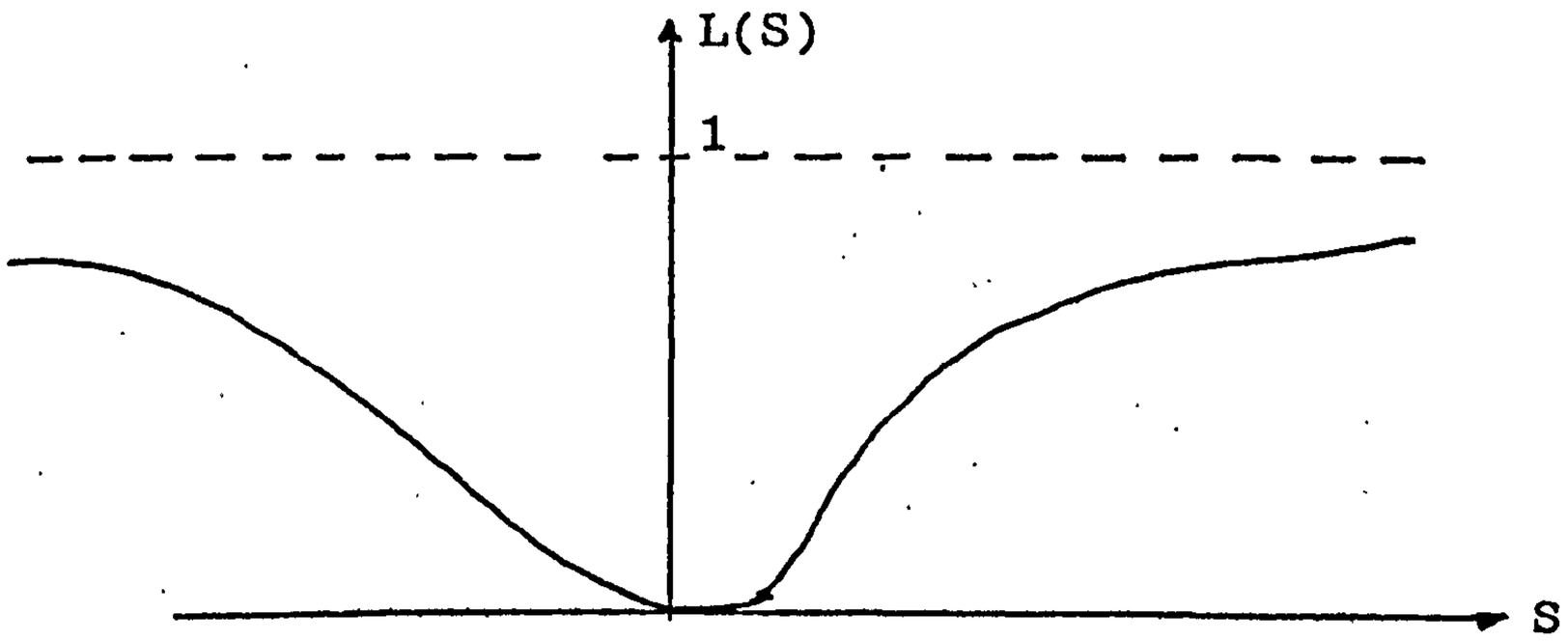


Fig. 6.1. A loss function causing bifurcation

Another approach to tackling this problem of classification would be to turn the problem on its head and try to isolate those loss functions $L(\theta-d)$ which have the property that $E_L(d)$ has only one minima for *any* unimodal distribution function $F(\theta)$ I care to choose.

Unfortunately this approach gives even less joy. I can generalise Ibragimov's (1) work (I do not include this for want of space) to show that only loss functions $L(\theta-d)$ satisfying the conditions of the section plus the additional condition

$$\frac{\partial^2}{\partial S^2}\,(\ell n(1-L(S))) < 0 \quad \text{for all } S \in \mathbb{R}$$

belong to this class. It is quite obvious that this is not a natural restriction to make on a loss function especially under the observation in Chapter 3 that in fact I should be minimising $E_L(d)$ where $L(\theta-d) = A(L^*(\theta-d))$ where A is some Anxiety function which is personal to the decision maker and not objective.

Perhaps the best solution is to restrict attention to loss functions under the constraint

$$L(S) = 1 \qquad |S| \geq S_o(F)$$

where S_o is chosen dependent on distribution function $F(\theta)$ so that the corresponding interval for Bayes decisions (see Theorem 3.11) is small enough so that a theorem similar to Corollary (6.1.1) can be invoked. This way conceptual problems arising from choice of Anxiety function are also avoided. Obviously there are theorems here to be proved at a later date but it should be noted that they will depend heavily on the particular choice of fixed distribution function being considered.

Returning to the original problem now, it should be noted that I have not yet proved anything outstanding, there could in fact be a reprieve from the last theorem if the extra minima appearing in $E_L(d)$ by the construction of the theorem never become global minima. Then they would never effect the topology of the Bayes decision. But in fact the infimum of E(d) jumps across these local minima.

Theorem 6.3.

Suppose $F(\theta)$ satisfies the condition of Corollary (6.2.1).

Write $d(\alpha) = \{d: E_{L(\alpha)}(d(\alpha)) = \inf_{d \in S(F)} E_{L(\alpha)}(d)\}$

where $L(\alpha)(\theta-d) = \alpha\, S_{b_1}(\theta-d) + (1-\alpha)\, S_{b_2}(\theta-d)$

where $0 \leq \alpha \leq 1$ and $0 < b_1 < b_2$ are chosen as in the construction of Corollary (6.2.1).

Then $d(\alpha)$ is not a continuous function of α.

Proof

Without loss of generality assume that $d_1 < d_2$ where

$$d_1 = \{d: E_{b_1}(d_1) = \inf_{d \in S(F)} E_{b_1}(d)\} \quad (d_1 \text{ unique by Theorem } 3.2)$$

$$d_2 = \{d: E_{b_2}(d) = \inf_{d \in S(F)} E_{b_2}(d)\} \quad (d_2 \text{ unique by Theorem } 3.2)$$

Then by Theorem (3.3.) $d(\alpha) \in [d_1, d_2]$ $0 \leq \alpha \leq 1$ and

$d(0) = d_1$ $\qquad$ $d(1) = d_2$. $\qquad$ 6.1.4.

Write $d^{(1)}(\alpha)$ the least minima of $E_{L(d)}(d)$

$d^{(2)}(\alpha)$ the least maxima greater than $d^{(1)}(\alpha)$ (if it exists)

$d^{(3)}(\alpha)$ the least minima greater than $d^{(1)}(\alpha)$ (if it exists)

Suppose $d(\alpha)$ were continuous. Then by (6.1.4).

$$d(\alpha) = d^{(1)}(\alpha) \quad \text{for all} \quad \alpha \in [0,1]. \qquad 6.1.5.$$

By Theorem (3.10) any new minima of expected loss must be greater than $d^{(1)}(\alpha)$

Let $\alpha^* = \{\text{Inf } \alpha \in [0,1]: E_{L(\alpha)}(d)$ has at least 2 minima

Then for α in some right neighbourhood of α^* $(\alpha^*, \alpha^* + \varepsilon_1)$ $\varepsilon_1 > 0$ (say)

$d^{(1)}(\alpha)$, $d^{(2)}(\alpha)$ $d^{(3)}(\alpha)$ exist and satisfy

$$d''(\alpha) < d^{(2)}(\alpha) < d^{(3)}(\alpha)$$

and $E_{L(\alpha)}(d^{(1)}(\alpha)) < E_{L(\alpha)}(d^{(2)}(\alpha))$ 6.1.6.

Again by Theorem (3.10) $d^{(2)}(\alpha)$ is decreasing and $d^{(1)}(\alpha)$ and $d^{(3)}(\alpha)$ increasing in α. In a left neighbourhood of 1

$$\alpha \in (1-\varepsilon_2, 1] \qquad \varepsilon_2 > 0 \qquad \text{(say)}$$

$E_{L(\alpha)}(d)$, has only 1 minima. Hence for some $0 < \alpha' < 1$, $d^{(2)}(\alpha')$ must disappear. By the above component, the only way it can disappear is by merging with $d^{(1)}(\alpha)$ into a point of inflexion. It follows that $d^{(1)}(\alpha')$ is not then an absolute minima of $E_{L(\alpha)}(d)$

The result follows. □

One consequence of the above theorem is that I can find an α^* $0 < \alpha^* < 1$, a corresponding loss function $L(\theta-d)$ of the form above and a distribution function $F(\theta)$ satisfying Corollary (6.2.1) such that $E_{L(\alpha^*)}(d)$ has at least two Bayes decisions d_1 and d_2 (say) where d_1 and d_2 one isolated (i.e. there are no Bayes decision in the interval (d_1, d_2). Any slight perturbation of $F(\theta)$ using this fixed loss function is likely to cause a flip from d_1 to d_2 or vice-versa. In particular if $F_t(\theta)$ is a distribution evolving smoothly with time t, one must be prepared for a sudden change in optimal decision at some point even if $F_t(\theta)$ is unimodal. This point will be touched upon later.

Returning to the theorem under the class of mixed step loss defined above and distribution $F(\theta)$ satisfying Corollary (6.2.1), since $d(\alpha)$ is increasing in α, the possible Bayes decisions will not cover the interval $[d_1, d_2]$, but there will be at least one region $(r_1, r_2) \in [d_1, d_2]$ where no Bayes decision can possibly lie.

Example (All notation as in last theorem)

Suppose a collection of Bayes decision makers are assembled and told that the posterior distribution of a certain parameter θ is $F(\theta)$ (which for the sake of argument I will assume is an Inverted Gamma giving just one region (r_1,r_2) defined above). A pseudo observation $\theta = x$ from $F(\theta)$ is worked out by usual computer techniques, but hidden from the decision makers.

Each decision maker is then forced to make an estimate $\hat{\theta}$ in the knowledge that

if	$-\lvert\hat{\theta} - x\rvert < b_1$	he wins £50
$b_1 \leq$	$\lvert\hat{\theta} - x\rvert < b_2$	he wins nothing $\quad 0 < b_1 < b_2$.
$b_2 <$	$\lvert\hat{\theta} - x\rvert$	he loses £100

Depending on his Anxiety function, the decision maker will either make a decision in $[d_1,r_1]$ or $[r_2,d_2]$ and therefore could be classified respectively as optimistic or pessimistic. For example those Bayesians who normally go for posterior modes could be expected to choose a decision in $[d_1,r_1]$ and those who use posterior means plum for a decision in $[r_2,d_2]$. Hence I have here a simple example of a population bifurcating when forced to make a decision about a seemingly smooth phenomenon. □

Finally note that generically, the only type of Catastrophe I can expect to isolate using this familty of loss functions is a Fold Catastrophe (though there may be many). Allowing a two parameter family of loss functions, I could then expect Cusp catastrophes and so on. What the normal factors and splitting factors are likely to be in these types of situation I must leave for research at a later date. They will depend both on the loss function and distribution function concerned.

6.2. Topology of Expected loss for Multimodal distributions

It was shown in the last section that one could not guarantee getting an expected loss function with only one minima even if the distribution concerned was properly unimodal. However, it was found that on restricting the loss functions concerned to the (utility invariant) step loss functions the topology of the probability density function carried over to the expected loss (see Theorem 3.2(1)). As a first attempt to classify multimodal distributions therefore it might be hoped that the nice property above might be preserved in some way. The next theorem shows that this is not the case unless some restriction is put on the multimodal distributions being considered. First some definitions.

Notation

Let $F(n)$ denote the class of all distribution functions $F(\theta)$ continuous on $\mathbb{R}$ and twice differentiable on $(S(F))^{o}$ such that its corresponding probability density function $f(\theta)$ has exactly n stationary points on $(S(F))^{o}$ where $S(F)$ is the extended support of $F(\theta)$.

Let $S(n)$ denote the class of all distribution functions $F(\theta)$ continuous on $\mathbb{R}$ and twice differentiable on $(S(F))^{o}$ for which the expected loss $E_b(d)$ with respect to any step loss function has a maximum of n stationary points on the part of the range of $E_b(d)$ defined by $E_b(d) \in (0,1)$.

Note that if n is even then $F(n)$ and $S(n) = \emptyset$ unless there are points of infection and also that if $F(\theta) \in F(n)$, $F(\theta) \in S(m)$ where $m \geq n$ (just let $b \to 0$ on the step loss). For notational convenience I will allow $m,n = \infty$.

Theorem 6.4.

If $n \geq 3$ for all $m \in \mathbb{N}$ there exists a distribution function $F(\theta) \in \mathcal{F}(n)$ such that

$$F(\theta) \in S(m^*) \quad \text{for some } m^* \geq m.$$

Proof. It is sufficient to prove the theorem for odd m. I first prove the case for n = 3.

Fix $m \in \mathbb{N}$, m odd.

Clearly I can choose a distribution function $F^*(\theta)$ in $\mathcal{F}(1)$ with the following two properties.

(i) $F^*(\theta)$ has mode at O and $S(F^*) = [-A,A]$ $A \in \mathbb{R}_{>O}$.

(ii) $f^*(y) = f^*(-y)$ has exactly $\frac{1}{2}(m-3)$ solutions for $y \in (O,A)$.

Define the probability density function $f(\theta)$ of a distribution function $F(\theta)$ by the equation

$$f(\theta) = \tfrac{1}{2}[f^*(\theta) + f^*(2A-\theta)]$$

For stationary points d of expected loss using loss function $S_b(\theta-d)$ the equation

$$f(d-b) = f(d+b) \quad \text{is satisfied.}$$

Putting $b \in A$ this has the same number of solutions as

$$f(y) = f(y+A).$$

This by definition of $f(\theta)$ (which is of course symmetric) has the same number of solutions as

$2 \times$ {no.of solutions of $f^*(y) = f^*(-y)$ $y \in (O,A)$} + solutions $\{f_tA) = f(A),\ f(O) = f(2A),\ f(A)=f(3A)\}$

$$= 2\left(\frac{m-3}{2}\right) + 3$$

$$= m.$$

Hence for this particular $F(\theta)$

$$\sup_{b \in \mathbb{R}_{>0}} \{\text{no of turning point of } E_b(d)\} \geq m.$$

so $F(\theta) \in S(m^*)$ for some $m^* \geq m$.

The theorem is now proven for $n = 3$.

For the case $n > 3$ consider a distribution function $F_n(\theta) \in F(n)$ such that

$f_n(\theta)|[-A,3A] \propto f(\theta)$ and the proof carries through. The result follows. □

An obvious way of disallowing the construction given in the last theorem is by the following definition.

If $F(\theta) \in F(n)$ write $S(F) = [M_o(F), M_{n+1}(F)]$ (M_o, M_{n+1} possibly infinite) and

$M_1 < M_2 < \ldots < M_n$ be the stationary points of $F(\theta)$.

Write $A_i(F) = (M_{i-1}, M_i)$ $\quad 1 \leq i \leq n+1$.

In passing note that now $f_i(\theta) = f(\theta)|A_i$ $\quad 1 \leq i \leq n+1$ is strictly monotonic. 6.2.1.

Definition

Call a distribution function $F(\theta)$ *ordinary* if for each pair (A_i, A_j) $1 \leq i < j \leq n+1$, and $b \in \mathbb{R}_{>0}$ there is at most one $d^*(b,i,j)$ such that

(i) $f(d^* - b) = f(d^* + b)$

(ii) $d^* - b \in A_i$

$d^* + b \in A_j$

In this case call $d^*(b,i,j)$ the *$(i,j)^{th}$ stationary point* with respect to b.

The following theorem insues.

Theorem 6.5.

If $F(\theta)$ is an ordinary distribution function, then $F(\theta) \in F(n)$ implies

$$F(\theta) \in S(m) \qquad \text{where } n \leq m \leq 2n-1.$$

Proof

Choose an arbitary but fixed $b \in \mathbb{R}_{>0}$. Suppose there exist

$$i \neq i^* \qquad j \neq j^* \qquad i,j,i^*,j^* \in \mathbb{N}$$

such that d_1 and d_2 are respectively $(i,j)^{th}$ and $(i^*,j^*)^{th}$ stationary points and

$$i + j = i^* + j^*.$$

without loss of generality assume $i < i^*$, so that $j > j^*$. But then

$$(d+b) - (d_1-b) \qquad (d_2+b) - (d_2-b)$$

Hence since $F(\theta)$ is ordinary for all $b \in \mathbb{R}_{>0}$ and $k \in \mathbb{N}$ $2 < k \leq 2n+1$ there exists at most one (i,j)th stationary point of $E_b(d)$ with $i+j = k$.

Hence $F(\theta) \in S(m)$ where $m \leq (2n+1) - 2 = 2m-1$.

The other inequality comes from the comment preceding the last theorem. □

Note that for general ordinary distribution $F(\theta)$ the upper bound on m cannot be lowered. To see this choose $b \in \mathbb{R}_{>0}$ large enough so that a $(1,n+1)$ stationary point exists.

Then by suitable choice of $F(\theta)$ I can make sure that all

$(1,j)^{th}$ turning points $\qquad 2 \leq j \leq n$

and $(i,n+1)$ turning points $\qquad 1 \leq i \leq n$

(for example let $f(\theta) = 0$ at the ends of each interval $A_i(F)$).

Hence $F(\theta) \in S(2n-1)$.

Definition

If distribution function $F(\theta)$ is ordinary, call the *k-stationary point* of $E_b(d)$ the unique $(i,j)^{th}$-stationary point such that

$$i + j = k.$$

So at least for a certain class of distribution functions, the associated step expected loss is well behaved in some sense. However the reader may see that to isolate ordinary distributions analytically directly from their definition is difficult. The following theorem makes it easier to detect whether a distribution function is ordinary or not.

Theorem 6.6

$F(\theta) \in \bar{F}(n)$ is ordinary if and only if for all pairs (i,j) such that

$$j = i+2p \qquad 1 \leq p \leq \frac{n-i+1}{2} \qquad 1 \leq i \leq n$$

$g_{ij} : B_{ij} \to \mathbb{R}$ defined by

$g_{ij}(b) = [f_i^{-1} - f_j^{-1}](t)$ is strictly monotonic on B_{ij}

where B_{ij} is the intersection of the ranges of f_i and f_j and

$f_i(\theta) = f(\theta)|A_i(F) \qquad 1 \leq i \leq n+1.$

Proof

Suppose $F(\theta)$ is not ordinary. I can then rewrite the conditions of the definition of ordinary distributions to say that there exist pairs (x_1,y_1), $(x_2y_2) \in \mathbb{R}^2$ with

$x_1 \neq x_2 \in A_i$ and $y_1 \neq y_2 \in A_j$ such that

(i) $x_1 - y_1 = x_2 - y_2$

(ii) $f(x_1)-f(y_1) = f(x_2) - f(y_2) = 0$

for some pair (i,y) $1 \leq i < j \leq n+1$

Let $f(x_1) = t_1$, $f(x_2) = t_2$. Then $t_1 \neq t_2$ since $f(\theta)$ is invertible on A_i. Hence by (i) $g_{ij}(t_1) = g_{ij}(t_2)$ and so g_{ij} is not strictly monotonic.

Conversely if there exists a pair (i,j) $1 \leq i < j \leq n+1$ such that g_{ij} is not strictly monotonic, i.e. for which there exist a $t_1 \neq t_2 \in B_{ij}$ such that

$$g_{ij}(t_1) = g_{ij}(t_2).$$

Let
$$x_1 = f_i^{-1}(t_1) \qquad y_1 = f_j^{-1}(t_1)$$
$$x_2 = f_i^{-1}(t_2) \qquad y_2 = f_j^{-1}(t_2)$$

Then (i) and (ii) are satisfied and hence $F(\theta)$ is not ordinary. □

From the above equivalence it may be apparent that "ordinary" distributions are not as typical as they should be for a general analysis multimodal distributions. For example suppose $F(\theta) \in F(3)$ with antimode a_1 and modes m_1, m_2 such that

$$m_1 < a_1 < m_2$$

$$S(F) = (-\infty, \infty)$$

Then as $t \downarrow f(a_1)$ it is clear that

$$\frac{\partial g_{13}(t)}{\partial t} \to -\infty \quad \text{and} \quad \frac{\partial g_{24}(t)}{\partial t} \to \infty .$$

So for $F(\theta)$ to be ordinary $\{g_{13}(t)$ must be increasing. This in turn implies $\{g_{24}(t)$ must be decreasing

$f(m_1) = f(m_2)$, for if $f(m_1) > f(m_2)$ (say) then as $t \uparrow f(m_2)$

$$\frac{\partial g_{24}(t)}{\partial t} \to \infty$$

There are several possibilities for generalising the definition of ordinary distributions that I am working on at present and will present at a later date. The next theorem is, however, of some interest.

Theorem 6.7.

Let ordinary distribution function $F(\theta) \in \bar{F}(3)$ have probability density function $f(\theta)$ with stationary points

$$m_1 < a_1 < m_2 \quad \in (S(F))^o$$

where m_1, m_2 are modes and a_1 is an antimode. Then $F(\theta)$ will have 3 minima of $E_b(d)$ for some $b \in \mathbb{R}_{>0}$ if and only if both $g_{13}(t)$ is strictly decreasing and $g_{24}(t)$ is strictly increasing. Otherwise $F(\theta)$ will have at most two minima of $E_b(d)$ for all $b \in \mathbb{R}_{>0}$.

Proof

Since $F(\theta)$ is ordinary, by Theorem (6.5) $F(\theta) \in S(j)$ $3 \leq j \leq 5$. For 3 minima of $E_b(d)$ to exist it is necessary and sufficient that

(i) $F(\theta) \in S(5)$ (i.e. all 5 k-stationary points exist for some $b^* \in \mathbb{R}_{>0}$)

(ii) Every one of these 5 stationary points must be either a maxima or a minima for b^*.

(ii) in turn implies

k-stationary point is a $\begin{cases}\text{minima if k is } \begin{cases}\text{odd}\\ \text{even}\end{cases}\\ \text{maxima}\end{cases}$ 6.2.2.

For necessity suppose first that $0 < b^* < \frac{1}{2}(m_2 - m_1)$, then the 5-stationary point is the $(2,3)^{th}$-stationary point and hence if d^* is such that

$$E'(d^*) = f(d^* - b^*) - f(d^* + b^*) = 0$$

then $$E''(d^*) = f'(d^* - b) - f'(d^* + b) < 0.$$

Thus the 5-stationary point is a maxima by (6.2.2). So for b^* to satisfy (i) and (ii)

$$b^* > \tfrac{1}{2}(m_2-m_1) \qquad 6.2.3.$$

Thus if (i) and (ii) are to be satisfied b^* must be chosen so that the 5-stationary point is the $(1,4)^{th}$ stationary point.

If (d^*,b^*) has an $(1,4)^{th}$ stationary point in particular

$$f(m_1) - f(m_1+2b^*) > 0 \qquad 6.2.4.$$

Also if $g_{24}(t)$ is strictly decreasing using (6.2.2)

$$f(x) - f(x+2b^*) > 0 \quad \text{for all } x \in A_2.$$

Hence $E_b(d)$ will have no 6-turning point $b^* > \tfrac{1}{2}(m_2-m_1)$. Using an exactly analogous argument, if $g_{13}(t)$ is strictly increasing and b^* is chosen so that there is a $(1,4)^{th}$ stationary point then $E_{b^*}(d)$ will have no 4-stationary point. Hence necessity is proved.

For sufficiency all I need do is check that the 4-stationary point and 6-stationary point are maxima for some $b^* \in \mathbb{R}_{>0}$, for then by the nature of $E_b(e)$, $E_{b^*}(d)$ must have three minima as well. Since $g_{13}(t)$ is strictly increasing

$$f(x) - f(x+2b) \text{ is strictly decreasing } x \in A_3$$

taking a minimum value at $x = a_1$

and since $g_{14}(t)$ is strictly increasing

$$f(x-2b)-f(x) \text{ is strictly increasing } x \in A_3$$

taking maximum value at $x=a_1$.

Let $\eta=\min\{|(f(m_1)-f(m_1+2b^*))-(f(a_1)-fa_1+2b^*)|, |(f(m_2)-f(m_2-2b^*))-(f(a_1)-f(a_1-2b^*)|\}$

where b^* is chosen such that b^* $\tfrac{1}{2}(m_2-m_1)$ and close enough to $\tfrac{1}{2}(m_2-m_1)$ such that

$$\max\{|f(m_1) - f(m_1+2b^*)|, |f(m_2) - f(m_2-2b^*)|\} < \eta/2.$$

Then there exist $x_1 \in A_2$ and $x_2 \in A_3$ such that:

$$f(x_1) - f(x_1-2b^*) = 0 , \qquad f'(x_1) - f'(x_1-2b^*) > 0 \qquad 6.2.5.$$

$$f(x_2+2b^*) - f(x_2) = 0 , \qquad f'(x_2+2b^*) - f'(x_2) > 0 \qquad 6.2.6.$$

Putting $d_i = x_i - b^*$ $1 \leq c \leq 2$ (6.2.5) (6.2.6) imply that

$$E'_{b^*}(d_i) = 0 \qquad E''_{b^*}(d_i) > 0 \qquad 1 \leq i \leq 2.$$

Hence for b^* the 4-stationary point and 6-stationary point are both maxima of expected loss.

The result follows. □

Let $G(\theta)$ be given by

$$g(\theta) = \tfrac{1}{2} [f(\theta-\mu) + f(\theta+\mu)]$$

where $f(\theta)$ is the probability density function of a symmetric unimodal distribution $F(\theta)$ with $S(F) = (-\infty,\infty)$. The comment preceding the theorem together with the theorem itself imply that I can expect at most 2 minima of expected loss $E_b(d)$ with respect to step loss and distribution $G(\theta)$, provided $G(\theta)$ is ordinary. Thus the topology of $E_b(d)$ will not be "worse" than that of $g(\theta)$ in this sense. Reassuringly it seems that most mixtures $G(\theta)$ of the form above where $F(\theta)$ satisfies some loose regularity conditions will in fact be ordinary so at least Theorem 3.2.(1), can be generalised to a certain class of bimodal distributions.

Summary

Properly unimodal distributions that always have one minima of expected loss with respect to bounded symmetric loss functions have been classified together with those which can have more than one minima. An example of an application of this phenomena is given. A tentative attempt to classify multimodal distributions under step loss starts another rich area for research.

7. CATASTROPHES ARISING FROM MIXTURES

In the last chapter some examples of Catastrophes in Statistics were given. However it would be more valuable if some general results were possible so that in any particular situation the analyst would know whether or not he is likely to come across certain types of these singularities. Obviously I would like a classification in terms of the Expected loss potential function.

The easiest of such potential functions to classify are those arising from symmetric loss and distributions which are (discrete) mixtures of a particular family of distributions, because I can then interchange the order of integration so that I get a (discrete) mixture of a potential family. In this chapter, therefore, I attempt a classification of such mixtures. Because of the symmetry, the most interesting Catastrophes are the symmetric ones (Cusp and Butterfly). The following theorem concerns the former.

Theorem 7.1.

Suppose a posterior distribution of the form

$$g(\theta) = \alpha f(\theta+\mu) + (1-\alpha)\, f(\theta-\mu) \qquad \mu > 0$$

where f is symmetric, unimodal and generic, is obtained.

Let $E(d) = \int_{-\infty}^{\infty} L(d-\theta)\, f(\theta)d\theta$

$$E^*(d) = \int_{\infty}^{\infty} L(d-\theta)\, g(\theta)d\theta$$

$$R(\theta) = E^{(''')}(\theta)(E'(\theta))^{-1}$$

and E(d) be C^∞.

(Note $E^*(d) = \alpha E(d+\mu) + (1-\alpha)\, E(d-\mu)$.)

Then if (i) L is symmetric and bounded by 1.

(ii) $E''(x) = 0$ has one solution in $(0,\infty)$ namely $x = \eta$

(iii) $E'''(x) = 0$ has one solution in $(0,\infty)$ namely $x = \lambda$

(iv) $R((0,\eta)) \cap R((\eta,\lambda)) = \phi$

the potential $E^*(d)$ exhibits one unique catastrophe its coordinates given by

$$(d,\alpha,\mu) = (0,\tfrac{1}{2},\eta)$$

In this case (a) the normal factor, is a function of α only

(b) the splitting factor, is an increasing function of μ only.

Proof

$E(d)$ is symmetric since f and L are.

For a Catastrophe to occur at $d = D$ for $E^*(d)$ I need the first 3 terms of the Taylor expansion of $(E^*)'(d)$ to vanish about D, i.e.

$$AE'(\mu+D) + E'(D-\mu) = 0 \qquad 7.1.1.$$

$$AE''(\mu+D) + E''(D-\mu) = 0 \qquad 7.1.2.$$

$$AE'''(\mu+D)+E'''(D-\mu) = 0 \qquad 7.1.3.$$

where $A = \alpha(1-\alpha)^{-1}, \mu > 0$.

Since E is symmetric about zero these may be written:

$$AE'(\mu+D) = E'(\mu-D) \qquad 7.1.4.$$

$$AE''(\mu+D) = -E''(\mu-D) \qquad 7.1.5.$$

$$AE'''(\mu+D)= E'''(\mu-D) \qquad 7.1.6.$$

In the region $\{D: |D| > \mu\}$ $E(d)$ has no stationary points hence I need only search for Catastrophes in the region

$$\{D : |D| \leq \mu\} .$$

Conditions (ii) and (iii) imply $\eta < \lambda$.

From Condition (ii) and Equation (7.1.5)

$$\mu - D \leq \eta \quad \text{and} \quad \mu + D \geq \eta \qquad 7.1.7.$$

and Condition (iii) with Equation (7.1.6) implies

$$\mu - D < \lambda \qquad 7.1.8.$$

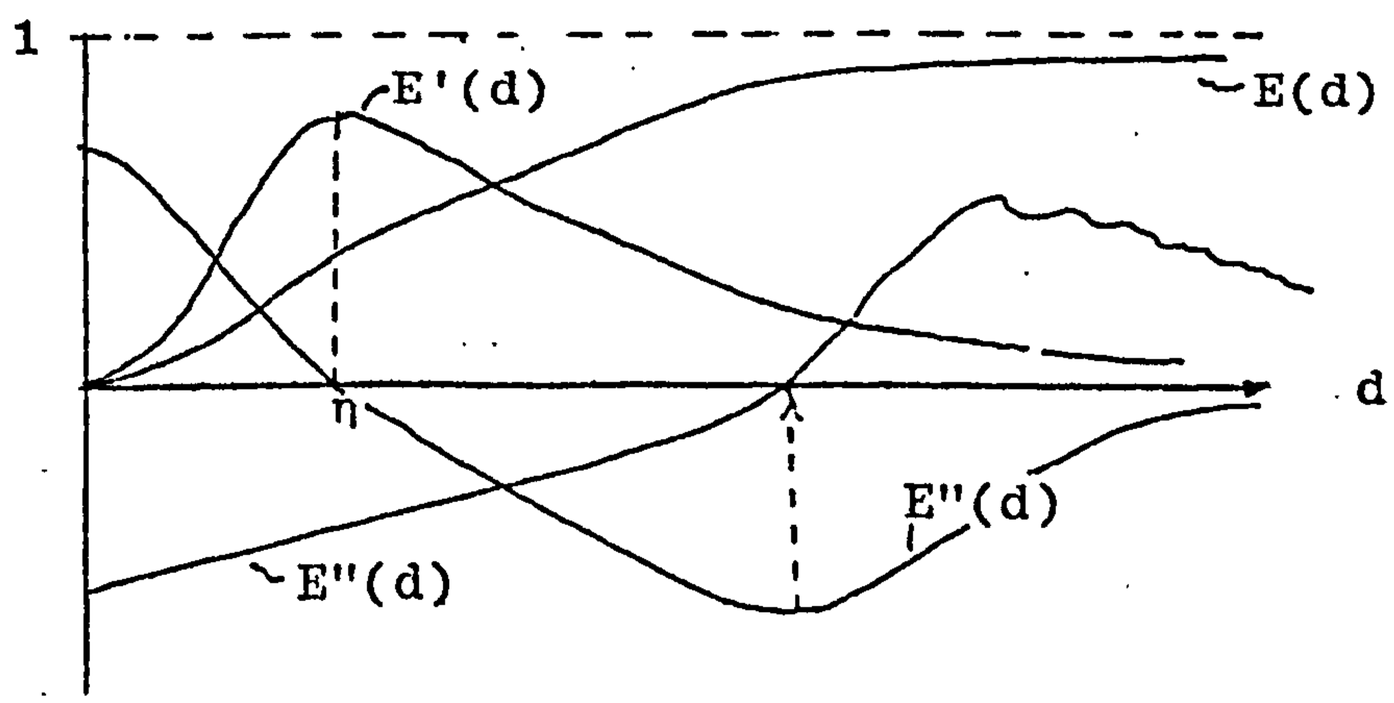


Fig. 7.1. Graph of E(d) and some of its derivatives.

Hence

$$(\mu-D) \in (0,\eta] \text{ and } (\mu+D) \in [\eta,\lambda) \qquad 7.1.9.$$

Dividing (7.1.6) by (7.1.4) the equation

$$\frac{E'''(\mu-D)}{E'(\mu-D)} = \frac{E'''(\mu+D)}{E'(\mu+D)} \quad \text{must be satisfied} \qquad 7.1.10.$$

Hence Condition (iv) and Equation (7.1.9) give

$$\mu-D = \mu+D = \eta$$

so D = 0 is a necessary condition for a cusp in E(d).

In this case (7.1.4), (7.1.5) and (7.1.6) become

$$(A-1)E'(\mu) = 0 \qquad 7.1.11.$$

$$(A+1)E''(\mu) = 0 \qquad 7.1.12.$$

$$(A-1)E'''(\mu) = 0 \qquad 7.1.13.$$

Since $E'(\mu) > 0$ (11) implies

$$A = 1 \qquad \alpha = \tfrac{1}{2}$$

Since $A+1 > 0$ (12) implies

$$E''(\mu) = 0 \qquad \mu = \eta.$$

and (7.1.13) is automatically satisfied. Thus there is a unique cusp at

$$(d,\alpha,\mu) = (0, \tfrac{1}{2}, \eta).$$

Finally, expanding the Taylor series abouf this point, truncating after the 3rd term gives the manifold

$$d^3 + a_1 d^2 + a_2 d + a_3 = 0$$

where
$$a_1 = 6\frac{E'''(\bar{\mu}+\eta)\bar{\alpha}}{E^{iv}(\mu+\eta)}$$

$$a_2 = 6\frac{E''(\eta+\bar{\mu})}{E^{iv}(\eta+\bar{\mu})}$$

$$a_3 = 12\bar{\alpha}\frac{E'(\eta+\bar{\mu})}{E^{iv}(\eta+\bar{\mu})}$$

$$\bar{\alpha} = \alpha - \tfrac{1}{2}$$

$$\bar{\mu} = \mu - \eta$$

which on truncating the Taylor series for coefficients in terms of $\bar{\alpha}$ and $\bar{\mu}$ to the lowest power, can be rewritten as:

$$\psi^3 - b\psi - a = 0$$

where
$$\psi = d + 2r(\eta)\bar{\alpha}$$

$$b = -6r(\eta)\bar{\mu}$$

$$a = -12\frac{E'(\eta)\bar{\alpha}}{E^{iv}(\eta)}$$

and $$r(\eta) = \frac{E'''(\eta)}{E^{iv}(\eta)}.$$

This completes the proof. □

The really crucial condition of the theorem is Condition (ii). Suppose for example that instead $E''(x) = 0$ had 3 solutions on $(0,\infty)$. When $\alpha = \frac{1}{2}$ equations (7.1.11) (7.1.12) and (7.1.13) have then 3 solutions when $D = 0$, hence there are 3 cusps along this line. The induced map of stationary points of $E(x)$ across the section $\alpha = \frac{1}{2}$ would then look something like Fig. 7.2. corresponding to the evolution of expected loss under increasing μ given by Fig. 7.3.

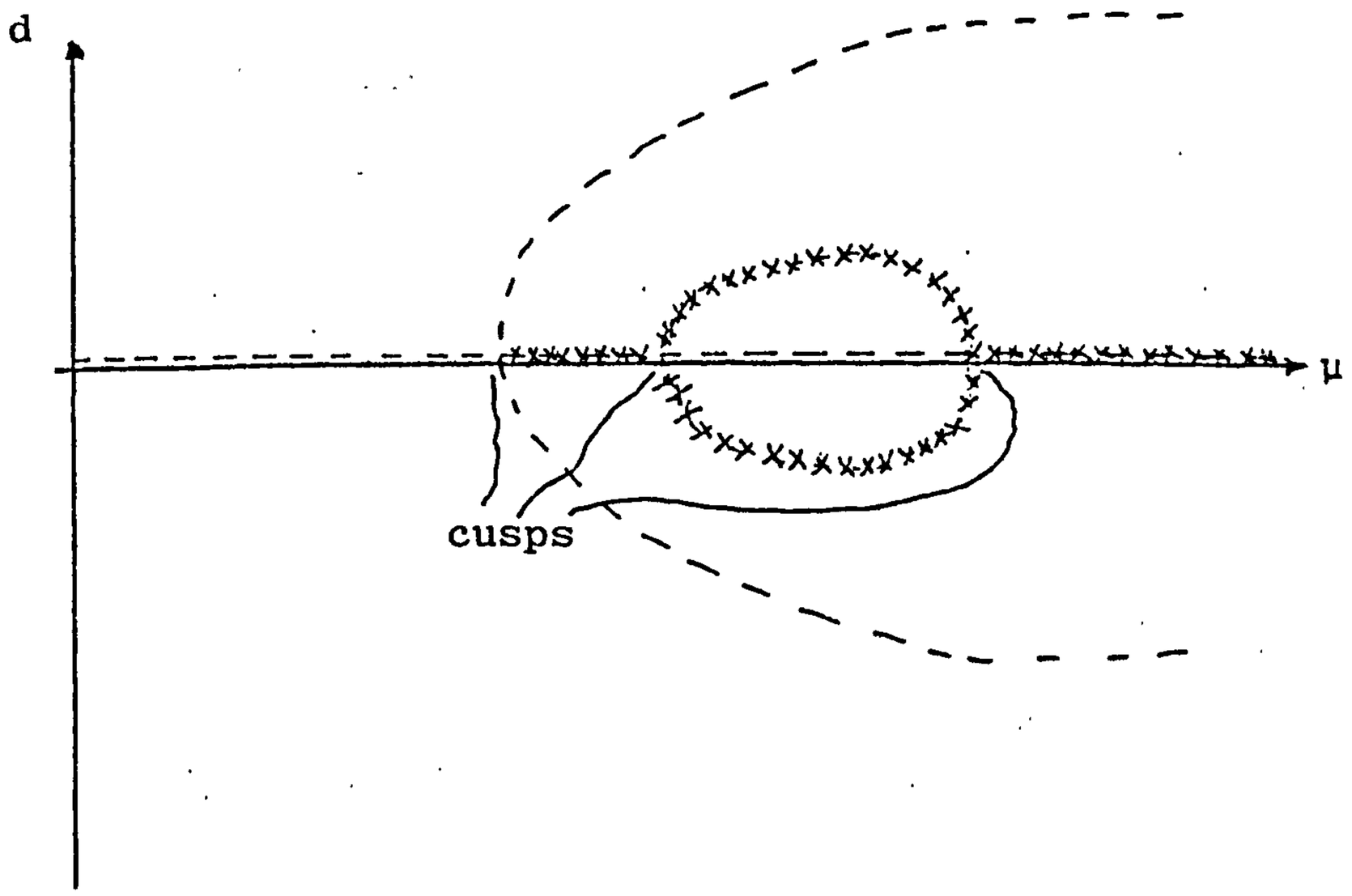


Fig. 7.2. Section of Manifold when $\alpha = \frac{1}{2}$.

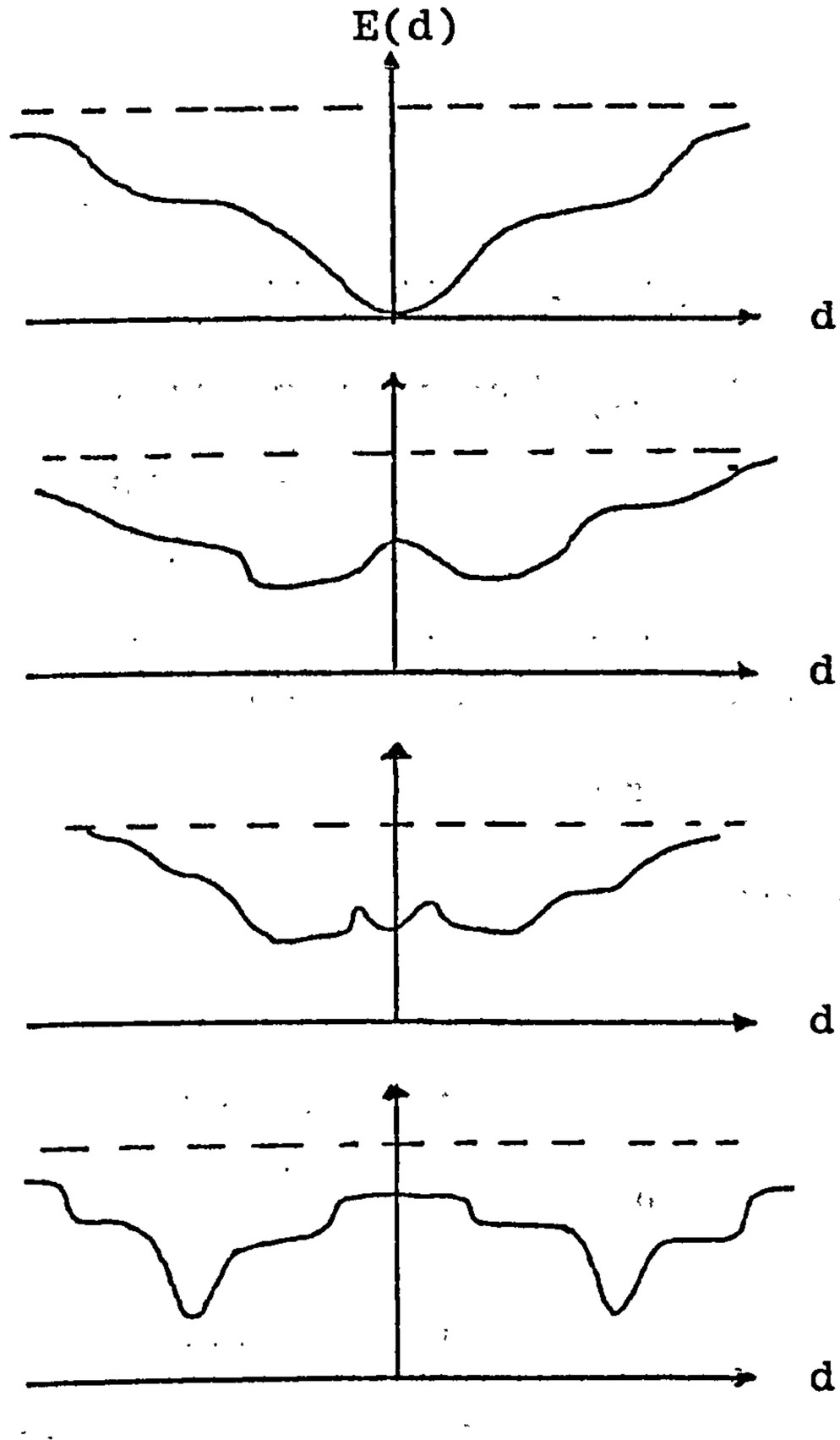


Fig. 7.3.

Note however that provided E(d) is generic, locally, any stationary points can be at worst folds or cusps and in particular if η is the smallest solution of E"(x) = 0 a cusp point will again be observed.

A rather crude theorem giving sufficient conditions for E"(x) = 0 to have exactly one solution is given below.

Theorem 7.2.

Let $L(\theta-d)$ be any 3 × differentiable symmetric loss function strictly increasing on $\mathbb{R}_{>0}$ with upper bound 1 and $f(x)$ be a continuous 2 × differentiable (a.e) p.d.f. symmetric about 0 such that

$$\delta(x) = \frac{f''(x)}{f'(x)} \text{ is increasing and strictly negative on } \mathbb{R}_{>0}.$$

Then $E''(d) = 0$ has only one solution in $\mathbb{R}_{>0}$.

Proof

By Lemma 3.3.1 of Chapter 3 I can write

$$E(d) = \int_0^\infty E_b(d)dG(b) \qquad 7.2.1.$$

where G is a probability distribution which since L is strictly increasing G' has support $[0,\infty)$

Since by the above conditions $E(d)$ is 3 × differentiable it is sufficient to show that whenever:

$$E''(d) = 0 \text{ then } E'''(d) > 0 \qquad d \in \mathbb{R}_{>0} \qquad 7.2.2$$

since then $E'(d)$ can have at most one stationary point a minima. Using (7.2.1) above

$$E''(d) = \int_{\mathbb{R}_{>0}} (f'(d-b)-f'(d+b))dG(b) \qquad 7.2.3.$$

$$E'''(d) = \int_{\mathbb{R}_{>0}} (f''(d-b)-f''(d+b))dG(b) \qquad 7.2.4.$$

Let $k_1(d,b) = \delta(d)-\delta(d-b)$

$k_2(d,b) = \delta(d+b)-\delta(d)$

then by the above conditions on δ

$$k_1(d,b) \geq 0 \quad \text{when} \quad 0 < b \leq d \qquad 7.2.5.$$

$$k_1(d,b) < 0 \quad \text{when} \quad 0 < d < b \qquad 7.2.6.$$

$$k_2(d,b) \geq 0 \quad \text{for all} \quad d,b \in \mathbb{R}_{>0} \qquad 7.2.7.$$

Hence

$$T(d) = \int_{\mathbb{R}_{>0}} (t_1(d,b)f'(d-b)+k_2(d,b)f'(d+b))dG(b) < 0 \qquad 7.2.8.$$

since as mentioned before G' has support $\mathbb{R}_{\geq 0}$

But it is easily checked that

$$E'''(d) = -(T(d) + \delta(d)E''(d)) \qquad 7.2.9.$$

and so whenever $E''(d) = 0$ then $E'''(d) > 0 \quad d \in \mathbb{R}_{>0}$.

The result follows □

Example

If f(x) is a Double-exponential p.d.f., then $\delta(x)$ is constant on $\mathbb{R}_{>0}$ and hence for any loss function of the form above any mixture of 2 Double exponential distributions can have at most 1 cusp in its expected loss at the point of symmetry between them

An Illustration: The mixture of 2 normal distributions

In multiprocess modelling (Harrison and Stevens (1)) and many other situations, mixtures of normal distributions arise as the posterior distribution $F(\theta)$ of a parameter θ i.e.

$$f(\theta) = \sum_{i=1}^{n} \alpha_i \, n(\mu_i, W_i)$$

where $n(\mu,W)$ is the normal p.d.f. with argument θ, mean μ and variance W and

$$\sum_{i=1}^{n} \alpha_i = 1 \quad \text{with} \quad \alpha_i > 0 \qquad 1 \leq i \leq n.$$

Lindley's conjugate loss function to the normal distribution (1) can be written

$$L(\theta-d) = 1 - \exp\{-\tfrac{1}{2} k^{-1}(\theta-d)^2\} \qquad 7.2.10$$

where k is some positive constant.

Notice that as $k \to 0$ the Bayes estimate will tend to the highest mode of the posterior distribution and as $k \to \infty$ the corresponding Bayes estimate tends to the posterior mean so that the family of loss functions is fairly wide. Using this loss function the posterior expected loss can be easily computed to be

$$E(d) = \sum_{i=1}^{n} \alpha_i[1-(2\pi k)^{\frac{1}{2}} n(\mu_i, w_i + k)] \qquad 7.2.11.$$

Corollary 7.1.1.

An expected loss of the form given by (7.2.11) with n = 2 and $w_1 = w_2$ exhibits a single cusp catastrophe at the point

$$\alpha = \tfrac{1}{2}$$
$$(\mu_1-\mu_2)^2 = 4(w + k).$$

Proof. Check the conditions of Theorem 1 are satisfied and solve the equation

$f''(x) = 0$ where

$f(x) = \exp[-\frac{1}{2}(w+k)^{-1}(x-\mu)^2]$ to obtain the value for $\mu_1 - \mu_2$ at the cusp □

The graph of the control space on this particular catastrophe is given in Fig. 7.4.

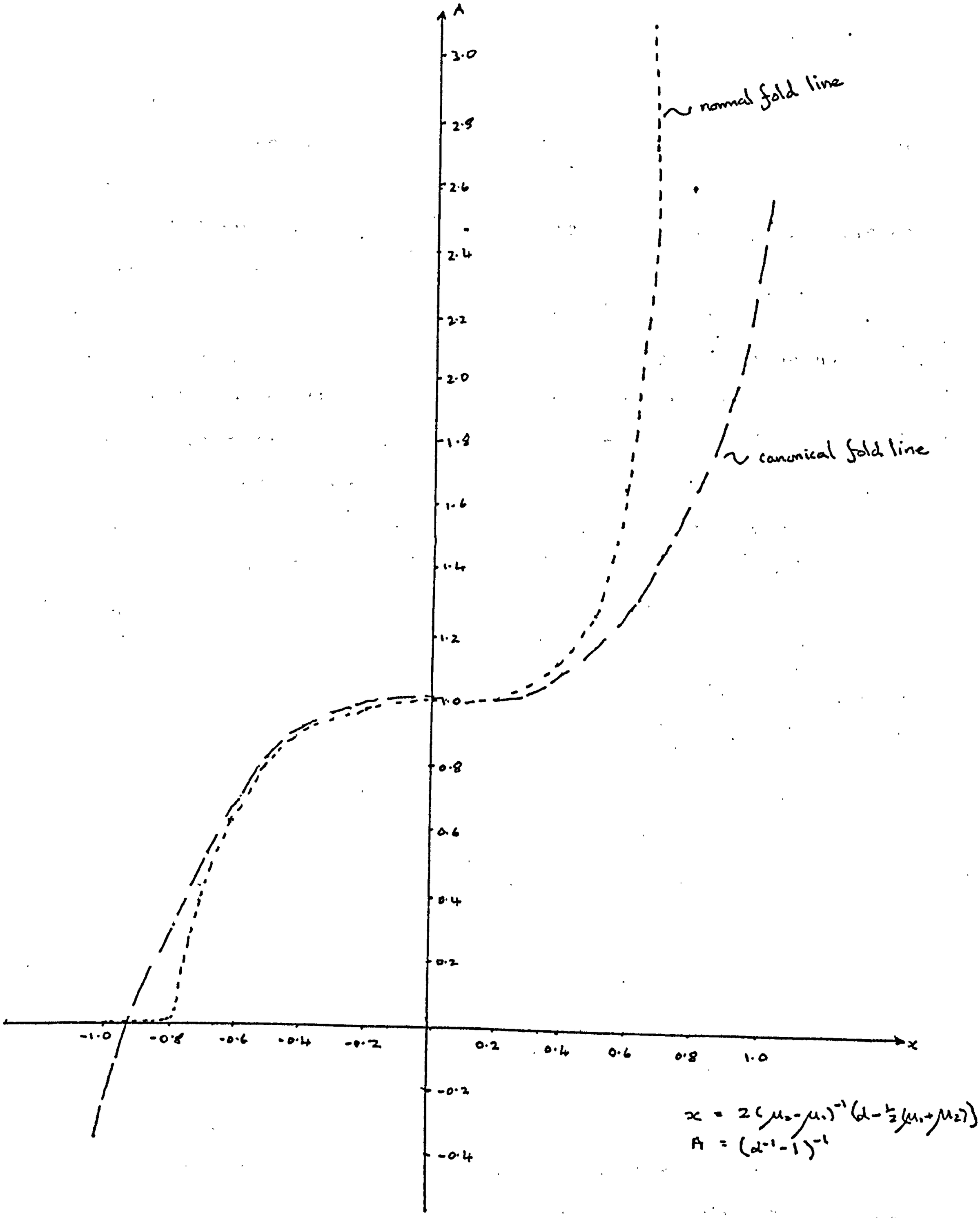


Fig. 7.4. Fold lines in (x, a) space

An aside

The above result emphasises an important fallacy, sometimes this sort of argument is heard,

"A loss function $L(\theta-d)$ which is symmetric and analytic can be expanded in a Taylor Series as

$$a_o - a_2(\theta-d)^2 + \text{other higher even terms.}$$

Ignoring these higher terms, this will give the posterior mean as Bayes estimate and thus the posterior mean can be seen as a second order approximation to the true Bayes estimate".

Well, the normal conjugate loss given above is certainly analytic. Suppose I consider a posterior distribution of the form given in the Corollary above and suppose $\alpha_1 = \alpha_2 = \frac{1}{2}$, then the graph of stationary points of E(d) is given by Fig. 7.5.

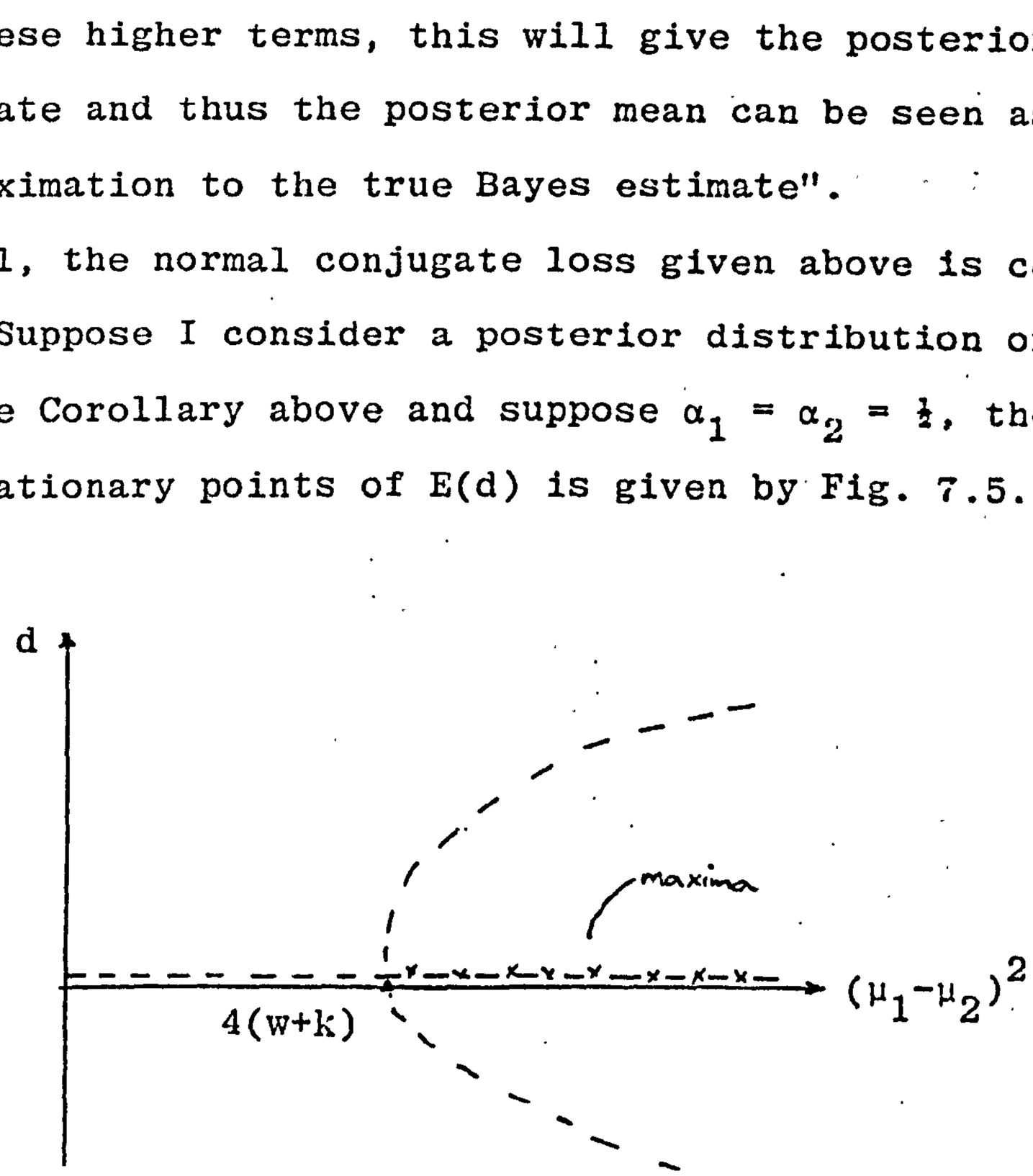


Fig. 7.5.

Notice that as $|\mu_1-\mu_2|$ becomes large the posterior mean $\frac{1}{2}(\mu_1 + \mu_2)$ becomes a *maxima* of expected loss and hence a worst decision.

The last Corollary can be generalised to the case when $W_1 \neq W_2$ as the following Theorem shows.

Theorem 7.3.

$$\text{The expected loss } E(d) = \sum_{i=1}^{2} \alpha_i(1-(2\pi k)^{\frac{1}{2}}n(\mu_i,v_i))$$

where $\sum_{i=1}^{2} \alpha_i=1$ $\alpha_i > 0$ $1 \leq i \leq 2$ and $n(\mu,V)$ is the normal probability density function with argument d, mean μ and variance V, has no swallowtail points if $V_1 \neq V_2$.

Proof

Without loss of generality I can assume that $V_1 > V_2$ and $\mu_1 \neq \mu_2$. The fold points of E(d) are then given by the points satisfying the two equations

$$E'(d) = 0 \quad E''(d) = 0.$$

which after rearrangement give the respective equations

$$(y-a) \exp -\{\tfrac{1}{2}cy^2\} = B(y-b) \qquad 7.3.1.$$

$$(1+acy-cy^2) \exp -\{\tfrac{1}{2}cy^2\} = B \qquad 7.3.2.$$

where
$$y = d - (V_1\mu_1-V_2\mu_2)(V_1-V_2)^{-1}$$

$$B = (1-\alpha^{-1})V_1^{\frac{1}{2}} V_2^{-\frac{1}{2}} \exp \tfrac{1}{2}\{(V_1-V_2)^{-1}(\mu_1-\mu_2)^2\}$$

$$c = (V_2-V_1)(V_1V_2)^{-1}$$

$$b = V_2(\mu_1-\mu_2)(V_1-V_2)^{-1}$$

$$a = V_1(\mu_1-\mu_2)(V_1-V_2)^{-1}$$

Eliminating the exponential term I find that the fold points must satisfy the cubic

$$(1 + acy - cy^2)(y-b) - (y-a) = 0 \qquad 7.3.3.$$

which can be written

$$z^3 - a_1 z - a_o = 0 \qquad 7.3.4.$$

where $z = y - \frac{1}{3}(a+b)$

$$a_1 = \frac{1}{3}(a+b)^2 - ab$$

$$a_o = c^{-1}(a-b) + \frac{2}{27}(a+b)^3 - \frac{ab}{3}(a+b)$$

Since $a_1 \neq 0$ for any real values of a and b the cusp of this cubic in z can never be achieved. But a swallowtail point is exactly a cusp point of fold points.

This completes the proof. □

An Application in Normal Hypothesis Testing

Some important conceptual work has been done (e.g. Dickey (1) and (2)) relating Hypothesis Testing to Bayesian statistics by considering mixtures of Alternative and Null hypotheses as priors and updating in the usual way. Another interesting link is given in the following

Suppose $X_1 \ldots X_n$ are independent identically distributed random variables, having normal distribution with mean μ and variance V, the null hypothesis being

$$H_o: \quad \mu = 0$$

against the alternative

$$H_1: \quad \mu \neq 0.$$

Using the usual type of arguments in a Bayes setting this implies that I have a prior $\hat{p}$.d.f. $\hat{p}(\mu)$ of of a form

$$\hat{p}(\mu) = (1-\beta)\chi_o(\mu) + \beta(2\pi w)^{-\frac{1}{2}} \exp -\tfrac{1}{2}(w^{-1}\mu^2)$$

where $\chi_o(\mu)$ is the indicator function for μ located at 0

β represents my prior belief that the null hypothesis is wrong

and W is large (and be tended to ∞ in due course)

On observing $x_1 \ldots x_n$ using the usual Bayes arguments I obtain a posterior probability density function $p(\mu)$ of μ of the form

$$p(\mu)=(1-\beta_1(\bar{x}))\chi_o(\mu) + \frac{\beta_1(\bar{x})}{\sqrt{2\pi}} (W^{-1}+nV^{-1})^{-\frac{1}{2}} \exp-\tfrac{1}{2}\{(W^{-1}+nV^{-1})(\mu-\frac{W\bar{x}}{W+n^{-1}V})^2)\}$$

where $\beta_1(\bar{x}) = (1 + \frac{\gamma(\bar{x})}{\delta(\bar{x})})^{-1}$

$$\delta(\bar{x}) = \frac{\beta}{\sqrt{2\pi(W+n^{-1}V)}} \exp -\tfrac{1}{2}\{(W+n^{-1}V)^{-1}\bar{x}^2\}$$

$$\gamma(x) = \frac{(1-\beta)}{\sqrt{2\pi\ n^{-1}V}} \exp -\tfrac{1}{2}\ \{nV^{-1}\bar{x}^2\}.$$

Now ideally I would like to let $W \to \infty$ so that I have a flat prior over the alternative but this is impossible since as $w \to \infty$ the alternative part of the prior tends to an improper prior distribution (which of course gives zero probability to any finite interval) so that posterior weighting on the alternative $\beta_1(\bar{x})$ tends to zero regardless of $\bar{x}$. The problem is simply overcome, however, by allowing β to be a function of W.

The only function β, W giving a non trivial limit (i.e. such that

$\lim_{w\to\infty} \beta_1(\bar{x}, W) \to 0$ or 1 for all values $\bar{x}$)

can, from the equation for $\beta_1(\bar{x})$ above, be seen to be a function

$$\beta(W) = (1 + f(W))^{-1}$$

such that

$$\lim_{w\to\infty} \left(\frac{W^{\frac{1}{2}}}{f(W)}\right) = B \quad \text{where } B \in \mathbb{R}_{>0}$$

otherwise $\gamma/\delta \to 1$ or 0

Since I am going to take limits anyway without loss of generality write $\beta(W)$

$$\beta(W) = (1 + BW^{-\frac{1}{2}})^{-1}$$

in which case as $w \to \infty$

$$\frac{\gamma}{\delta} \to B \exp -\tfrac{1}{2}(\bar{x}^2\, nV^{-1})$$

It is easily checked that this now gives sensible results. For example as $|\bar{x}| \to \infty$, $\gamma/\delta \to 0$ and hence.

$$\beta_1(\bar{x}) \to 1$$

so that all my posterior weight goes on the alternative. Similarly if $\bar{x} = 0$ the posterior distribution is then symmetric about 0 so that any Bayes decision under symmetric loss will give the null hypothesis $\mu = 0$.

Hence in representing this Hypothesis Testing procedure in this way as a limit of a proper Bayesian procedure it is seen that the corresponding posterior distribution for μ is given by

$$(1 - \beta_1(\bar{x}))\, \chi_0(\mu) + \beta_1(\bar{x}) \frac{1}{\sqrt{2\pi n^{-1}V}} \exp -\tfrac{1}{2}(nV^{-1}(\bar{x}-\mu)^2)$$

where $\beta_1(\bar{x}) = (1 + B \exp -\tfrac{1}{2}(nV^{-1}\bar{x}^2))^{-1}$

To make a Bayes decision about μ I now use a loss function (for convenience's sake the conjugate loss function of the form in equation (7.2.10) to obtain a posterior expected loss of the form

$$E(d) = \sum_{i=1}^{2} \alpha_i (1 - (2\pi k)^{\frac{1}{2}} n(\mu_i, U_i))$$

where again $n(.,.)$ represents the normal probability density function with argument d and

$$\mu_1 = 0 \qquad U_1 = k \qquad \alpha_1 = 1 - \beta(\bar{x})$$

$$\mu_2 = \bar{x} \qquad U_2 = n^{-1}V+k \qquad \alpha_2 = \beta(\bar{x}).$$

A diagram of this expected loss is given in Fig. 7.6. Note in passing that I have brought only two extra constants B and k into the model, B being linked to my prior beliefs and k being linked to my criteria of judgement via the loss function.

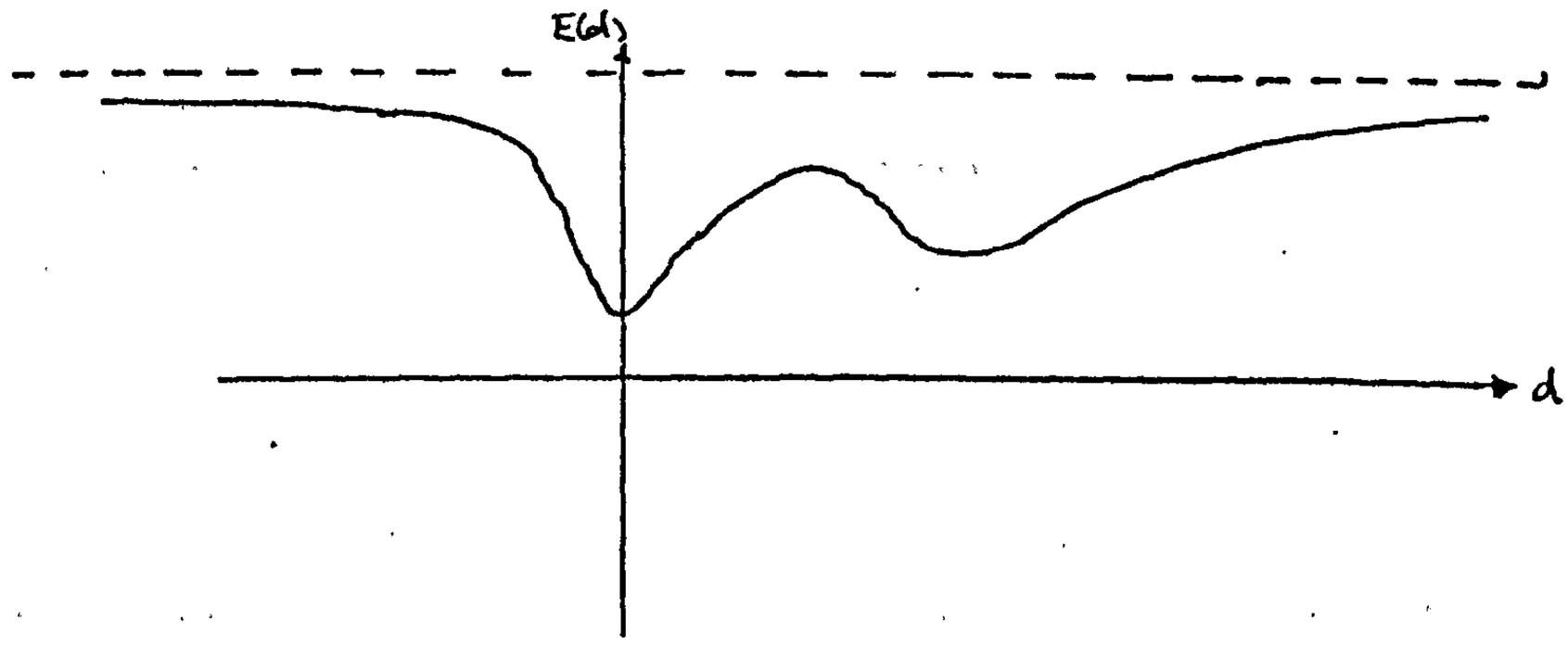


Fig. 7.6.

Now suppose n is large so that $U_1 \simeq U_2 = k$. Then this almost satisfies the conditions of Corollary (7.1.1). If $U_1 = U_2 = k$ then there would be a cusp point at

$$\beta_1(x) = \tfrac{1}{2} \qquad 7.3.5.$$

and $$\bar{x}^2 = 4k \qquad 7.3.6.$$

where (7.3.5) can be written in the form

$$\left| \frac{\bar{x}}{(n^{-1}V)^{\frac{1}{2}}} \right| = (2 \ln B)^{\frac{1}{2}}$$

Suppose that k is chosen such that $k < \frac{x^2}{4}$.

Then the decision space has been split into two sheets by the cusp one sheet corresponding to minima of E(d) near O (the null hypothesis sheet) and minima of E(d) near $\bar{x}$ (the alternative hypothesis sheet). See Fig (7.7)

Since I have symmetry (n large) to a first approximation, the Bayes decision (i.e. the lowest minima of expected loss) will be approximately determined by the line

$$\beta_1 = \tfrac{1}{2}$$

Modulo this approximation therefore it follows that the null hypothesis sheet holds the Bayes decision if

$$\left|\frac{\bar{x}}{(n^{-1}V)^{\frac{1}{2}}}\right| < (2\ \ell nB)^{\frac{1}{2}}$$

and the Bayes decision is on the alternative sheet if

$$\left|\frac{\bar{x}}{(n^{-1}V)^{\frac{1}{2}}}\right| > (2\ \ell nB)^{\frac{1}{2}}$$

These of course are the standard acceptance and rejection regions obtained from classical arguments.

Hence Classical hypothesis testing procedures can be seen as a particular form of a limiting case of Bayes decision making under bounded loss criteria. Unlike other Bayes analogues however it represents hypothesis testing in a *qualitative* way. When testing the hypothesis $\mu = 0$ I do not really believe that μ could possibly be exactly O, just that the value of μ is near O. The above correspondence presents this point very clearly.

When $k \geq \frac{\bar{x}^2}{4}$ the posterior distribution will be unimodal and hence I get the typical $k \to \infty$ I obtain the mean which always compromises (and in so doing loses the topology of the situation)

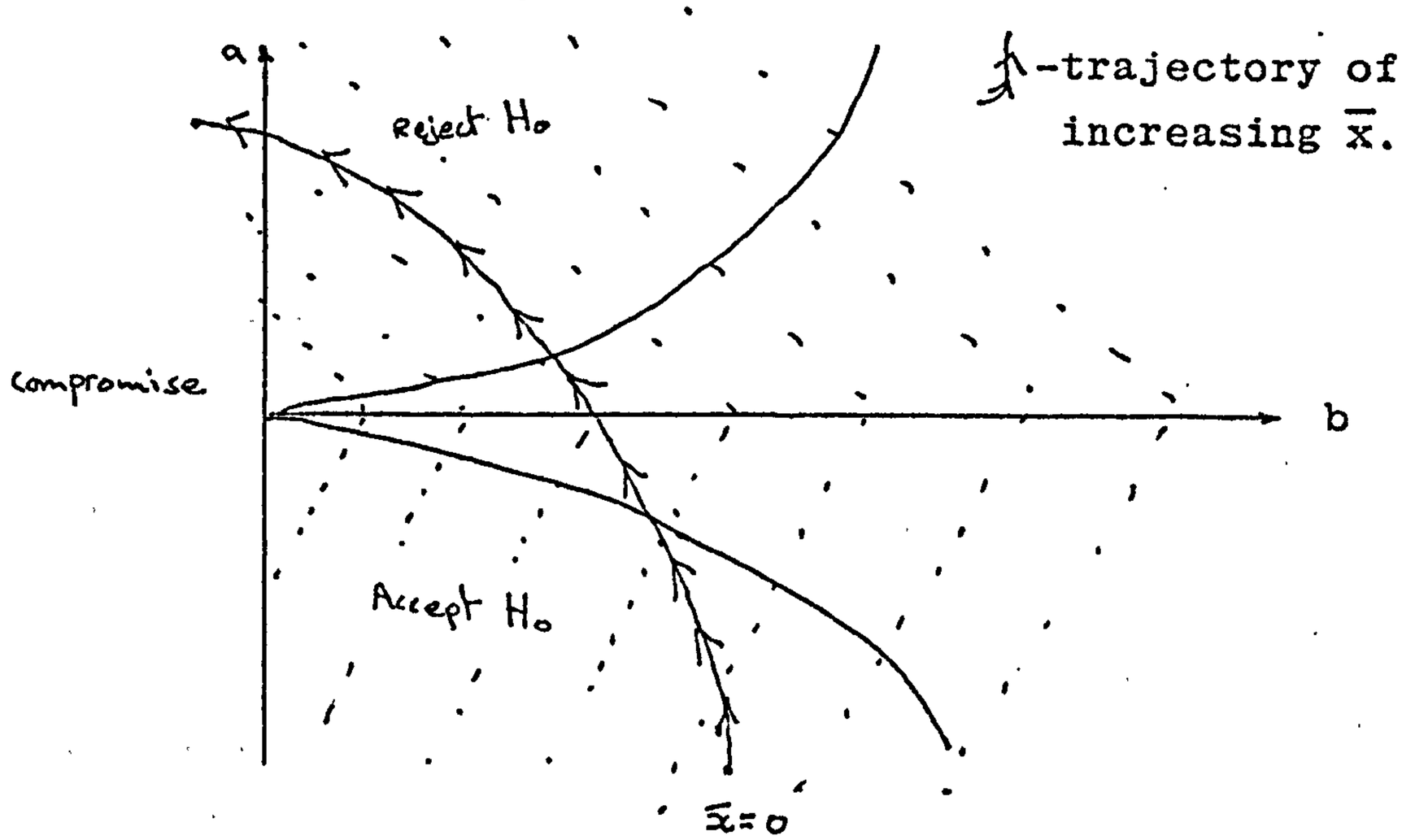


Fig. 7.7.

It would be very interesting to see (for small values of a) how the Bayes decision moves around the 4 dimensional control space, but this is some research that is yet to be completed.

The Butterfly Mixture

At the beginning of this chapter I mentioned that the other main Catastrophe to appear from mixtures of symmetric distributions is the Butterfly. This time I need a 3-mixture to illustrate the point.

Theorem 7.4.

Suppose a posterior probability density function is obtained which is of the form

$$g(\theta) = \alpha_1 f(\theta-\mu_1) + \alpha_2 f(\theta) + \alpha_3 f(\theta+\mu_3)$$

where $f(\theta)$ is a symmetric unimodal probability density function and

$$\sum_{i=1}^{3} \alpha_i = 1 \qquad \text{where } \alpha_i > 0 \qquad 1 \leq i \leq 3.$$

$$\text{Let } E(d) = \int_{-\infty}^{\infty} L(d-\theta)f(\theta)d\theta$$

$$E^*(d) = \int_{-\infty}^{\infty} L(d-\theta)g(\theta)d\theta$$

$$R(d) = E'''(d)(E'(d))^{-1}$$

and $E(d)$ be generic and C^{∞}

Suppose (i) $R(d)$ is monotonic on $\mathbb{R}_{>0}$

(ii) $E^{(iv)}(d)\ (E^{(ii)}(d))^{-1}$ is monotonic on (η,∞)

where as before η is the unique solution in $(0,\infty)$ of $E''(x) = 0$. Then there exists a unique Butterfly Catastrophe along the plane $d = 0$ whose coordinates in the Control Space are given by

$$(\alpha_1\alpha_2\alpha_3, \mu_1\mu_3) = (\alpha,\ 1-2\alpha,\ \alpha,\ \mu, \mu)$$

where μ is the unique solution in (η,∞) of

$$\frac{E^{(iv)}(d)}{E^{(ii)}(d)} = \frac{E^{(iv)}(0)}{E^{(ii)}(0)} \qquad 7.4.1.$$

and α is given by the equation

$$\alpha = \tfrac{1}{2}\left(1 - \frac{E^{(ii)}(\mu)}{E^{(ii)}(0)}\right)^{-1}$$

Furthermore such a Butterfly Catastrophe will have

a and c as linear combinations of $\mu_1 - \mu_3$ and $\alpha_1 - \alpha_3$

and b and d as linear combinations of $\mu_1 + \mu_3$ and $\alpha_1 + \alpha_2$

where a = normal factor b = splitting factor

c = bias factor d = butterfly factor.

Proof

Any Butterfly point at d = O must satisfy

$$\alpha_1 E^{(i)}(d-\mu_1) + \alpha_2 E^{(i)}(d) + {}_3E^{(i)}(d+\mu_3)\Big|_{d=O} = O$$

$$1 \leq i \leq 5$$

which on using the symmetry of E gives

$$-\alpha_1 E^{(i)}(\mu_1) + \alpha_2 E^{(i)}(O) + \alpha_3 E^{(i)}(\mu_3) = O \qquad 7.4.2.$$

$$\alpha_1 E^{(ii)}(\mu_1) + \alpha_2 E^{(ii)}(O) + \alpha_3 E^{(ii)}(\mu_3) = O \qquad 7.4.3.$$

$$-\alpha_1 E^{(iii)}(\mu_1) + \alpha_2 E^{(iii)}(O) + \alpha_3 E^{(iii)}(\mu_3) = O \qquad 7.4.4.$$

$$\alpha_1 E^{(iv)}(\mu_1) + \alpha_2 E^{(iv)}(O) + \alpha_3 E^{(iv)}(\mu_3) = O \qquad 7.4.5.$$

$$-\alpha_1 E^{(v)}(\mu_1) + \alpha_2 E^{(v)}(O) + \alpha_3 E^{(v)}(\mu_3) = O \qquad 7.4.6.$$

Dividing (7.4.4) by (7.4.2) and using that $E^{(i)}(O) = E^{(iii)}(O) = E^{(v)}(O) = O$ implies that

$$\frac{E^{(iii)}(\mu_1)}{E^{(i)}(\mu_1)} = \frac{E^{(iii)}(\mu_3)}{E^{(i)}(\mu_3)} \qquad 7.4.7.$$

which on using (ii) gives

$$\mu_1 = \mu_3 = \mu \quad \text{(say)} \qquad 7.4.8.$$

Resubstituting into (7.4.2) gives

$$\alpha_1 = \alpha_3 = \alpha \quad \text{(say)} \qquad 7.4.9.$$

Dividing (7.4.5) by (7.4.3) and using (7.4.8) and (7.4.9) above gives

$$\frac{E^{(iv)}{}_{(\mu)}}{E^{(ii)}{}_{(\mu)}} = \frac{E^{(iv)}{}_{(O)}}{E^{(ii)}{}_{(O)}}$$

Also $\mu \in (\eta,\infty)$ since $E^{(ii)}{}_{(\mu)} > O$ is necessary if (7.4.3) is to hold. Hence by (ii) μ is uniquely given by the above and α satisfies

$$\alpha = \left[2\left(1 - \frac{E^{(ii)}{}_{(\mu)}}{E^{(ii)}{}_{(O)}}\right)\right]^{-1}$$

Conversely if $(\alpha_1,\alpha_2,\alpha_3,\mu_1,\mu_2)$ satisfy the above requirements then each of the first 5 equations hold.

The first part of the theorem is now proven.

For the second part of the theorem consider singularities arising from general small perturbations of the butterfly point of the form

$$G\rho(d) = (1+\varepsilon_1)\,E(d-\mu-\lambda_1) + (1-(\varepsilon_1+\varepsilon_2))AE(d)+(1+\varepsilon_2)E(d+\mu+\lambda_2)$$

where $A = \alpha^{-1} - 2$

The k^{th} coefficient of the Taylor series expansion of $G\rho(d)$ is then given by

$$(1+\varepsilon_1)E^{(k)}(-\mu-\lambda_1) + A(1 - (\varepsilon_1+\varepsilon_2))E^{(k)}(O)+(1+\varepsilon_2)E^{(k)}(\mu+\lambda_2) \quad 7.4.10.$$

If k is odd. $E^{(k)}(-\mu-\lambda_1) = -E^{(k)}(\mu+\lambda_1)$

Hence (7.4.10) becomes

$$E^{(k)}(\mu+\lambda_2) - E^{(k)}(\mu+\lambda_1) - \varepsilon_1 E^{(k)}(\mu+\lambda_1) + \varepsilon_2 E^{(k)}(\mu+\lambda_2)$$

which on taking the first 2 terms of the Taylor expansion with respect to $(\varepsilon_1,\varepsilon_2,\lambda_1,\lambda_2)$

$$= (\lambda_2 - \lambda_1)\,E^{(k+1)}(\mu) + (\varepsilon_2 - \varepsilon_1)E^{(k)}(\mu) \quad 7.4.11.$$

If k is even $E^{(k)}(-\mu-\lambda_1) = E^{(k)}(\mu+\lambda_1)$

Hence expanding (7.4.10) to its first 2 terms in $\varepsilon_1,\varepsilon_2,\lambda_1,\lambda_2$ again gives

$$2E^{(k)}(\mu) + (\lambda_1+\lambda_2)E^{(k+1)}(\mu) \pm AE^{(k)}(O) + (\varepsilon_1+\varepsilon_2)\,[E^{(k)}(O)+E^{(k)}(\mu)]$$

and by definition of μ and A

$$= (\lambda_1+\lambda_2)\,E^{(k+1)}(\mu) + (\varepsilon_1+\varepsilon_2)\,[E^{(k)}(O) + E^{(k)}(\mu)].$$

So if $a_n(n!)^{-1}$ is the coefficient of d^n in the Taylor expansion of $G\rho(d)$ then

$$\begin{aligned}
a_1 &= (\lambda_2-\lambda_1)\,E^{(ii)}(\mu) + (\varepsilon_2-\varepsilon_1)E^{(i)}(\mu)\\
a_2 &= (\lambda_2+\lambda_1)\,E^{(iii)}(\mu) + (\varepsilon_1+\varepsilon_2)E^{(ii)}(\mu) + E^{(ii)}(O))\\
a_2 &= (\lambda_2-\lambda_1)E^{(iv)}(\mu) + (\varepsilon_2-\varepsilon_1)E^{(iii)}(\mu)\\
a_4 &= (\lambda_2+\lambda_1)E^{(v)}(\mu) + (\varepsilon_1+\varepsilon_2)E^{(iv)}(\mu) + E^{(iv)}(O))\\
a_5 &= (\lambda_2-\lambda_1)E^{(vi)}(\mu) + (\varepsilon_2-\varepsilon_1)E^{(v)}(\mu).
\end{aligned}$$

Since a_5 has no O^{th} order part with respect to $(\varepsilon_1,\varepsilon_2,\lambda_1,\lambda_2)$ it will have no effect on the first term expansion of the controls.

Also $E^{(vi)}(O) \neq O$ if E is generic so a_6 has nonvanishing first term in its Taylor expansion and so will only effect the controls proportionalitywise.

It follows that

$$\begin{aligned}
a &\propto (\lambda_2-\lambda_1)\,E^{(ii)}(\mu) + (\varepsilon_2-\varepsilon_1)E^{(i)}(\mu) \qquad 7.4.12.\\
b &\propto (\lambda_2+\lambda_1)\,E^{(iii)}(\mu) + (\varepsilon_1+\varepsilon_2)(E^{(ii)}(\mu) + E^{(ii)}(O))\\
c &\propto (\lambda_2-\lambda_1)\,E^{(iv)}(\mu) + (\varepsilon_2-\varepsilon_1)E^{(iii)}(\mu)\\
d &\propto (\lambda_2+\lambda_1)\,E^{(v)}(\mu) + (\varepsilon_1+\varepsilon_2)(E^{(iv)}(\mu)+E^{(iv)}(O))
\end{aligned}$$

Rewriting $G\rho(d)$ in its original variables $(\alpha_1,\alpha_2,\alpha_3,\mu_1,\mu_2)$ now gives the result. □

Comments

1). It is easily checked that a mixture of 3 normal probability density functions

$$f(\theta) = \alpha_1\, n(\mu_1,w) + \alpha_2\, n(0,w) + \alpha_3\, n(\mu_3,w)$$

where $n(\mu,w)$ is a normal probability density function with argument θ, mean μ and variance w and

$$\sum_{i=1}^{3} \alpha_i = 1 \quad \text{with } \alpha_i > 0 \qquad 1 \le i \le 3$$

under conjugate loss of the form

$$L(\theta-d) = \exp\{-\tfrac{1}{2}k^{-1}(d-\theta)^2\}$$

gives an expected loss satisfying the conditions of the theorem. In fact the Butterfly point is given by

$$d = 0$$

$$\mu_1 = \mu_3 = \sqrt{3(w+k)}$$

$$\alpha_1 = \alpha_3 = \tfrac{1}{2}\,(1+2\exp(-\tfrac{3}{2}))^{-1}$$

$$\alpha_2 = 1 - 2\alpha_1$$

2). In practice it has been found difficult to isolate Normal factors from Bias factors and Splitting factors from Butterfly factors because locally they tend to have similar effects on the topology. Here again the theorem above emphasises this point since the exact linear combination in (7.4.12) of each of the factors depends on the exact topological form of E(d). However it is easily seen that "symmetric" perturbations around the Butterfly point are encapsulated in the Splitting and Butterfly factors and the "antisymmetric" is the Normal and Bias factors.

Summary

The last Chapter has shown how one can use Catastrophe theory to classify some of the more simple forms of expected loss arising from mixtures of processes. With Theorem 7.1 and Theorem 7.4, I now have the apparatus to redo rigorously in terms of potential theory many of the examples of Catastrophes in the Social Sciences (Zeeman (3) and (4)) which Zeeman has heuristically rooted out.

An example using Theorem 7.1 is given which gives a novel link between Hypothesis Testing and Bayesian inference.

8. GENERALISED BAYES FORECASTING

8.1. Introduction

To study non-linear temporal processes successfully sooner or later a complete rewrite and reformulation of existing time series methodology must be attempted. The reason for this is that the standard formulation relies heavily on the linearity of the model - an assumption which for example is contrary to Catastrophe models.

In this chapter I shall begin such a reformulation. Obviously such a project is too large for comprehensive coverage in one chapter of a thesis so I have concentrated most of my attention on the steady (or first difference) model giving a careful formulation of the problems and many examples. I will then briefly indicate how the procedure carries through to other models.

Firstly I shall give a discussion of existing methodology.

8.2. Normal Bayes Forecasting

A very useful and robust method of forecasting has been the Bayes Forecasting approach introduced by Harrison and Stevens (1). For a full exposition I refer the reader to the above reference. Briefly the model is specified in two stages.

Stage 1 (Observation Eqn)

$$\underset{\sim}{Y}_t = F_t \underset{\sim}{\theta}_t + \underset{\sim}{V}_t \qquad V_t \sim n(\underset{\sim}{0},V)$$

Stage 2 (System Eqn)

$$\underset{\sim}{\theta}_t = G \underset{\sim}{\theta}_{t-1} + \underset{\sim}{W}_t \qquad W_t \sim n(\underset{\sim}{0},W).$$

where F_t, G are known matrices

$\underset{\sim}{V}_t, \underset{\sim}{W}_t$ are "error" vectors of random variables

$\underset{\sim}{Y}_t$ is the observational vector of random variables

$\underset{\sim}{\theta}_t$ is the system vector of random variables.

The point of splitting the usual sort of Box-Jenkins (1) type at ARIMA model into two stages is that the model is much more easy to construct and interfere with in any specific situation. This is due to the fact that each θ_t represents the "level" of one of the things I am estimating and has an interpretative value to it. For example one θ_t may represent the "true" level of demand for a particular product and Y_t represent the sales for that item during that time stage. Hence theoretically the connection between demand between two particular products at some time stage may be specified together with the connection between the actual sales in a natural way. Also if something outside unusual happens it is possible to predict the effect on the levels or observations so the model can be quickly readjusted to meet the new situation.

Because of the nature of the normal distribution the distribution of $\underset{\sim}{\theta}_t$ conditional on collections of the $\underset{\sim}{Y}_t$'s will be normal. Using the short hand

$$\underset{\sim}{\theta}_t|\underset{\sim}{Y}_t^t = \underset{\sim}{\theta}_t|Y_t, Y_{t-1} \cdots$$

the distribution of $\underset{\sim}{\theta}_t|\underset{\sim}{Y}^t$ has mean which is a weighted average of combination of the observative (the weights and combinations depending on F_t,G,V and W) and variance tending to a constant after the edge effect of the prior wears off.

These weighted averages will smooth over the Y_t's, the more weight put on past observations the "larger" the matrix V is than W. In passing it should be noted that such results are not recent, for example Muth (1) and Whittle (1) give some special cases of the update formulae.

The purpose of this chapter is to generalise the approach to include models where the Y_t's are not necessarily normally distributed. In particular I can the model Type II Catastrophes (see Chapter 4).

8.3. Some Difficulties in Formalising a Generalisation

Consider first the special case of when the observations Y_t are symmetric. A very obvious first try is to keep the same form of the equations "Stage 1" and "Stage 2" with error terms V_t, W_t symmetric distributions. This approach however encounters two problems.

Problem 1 Tractability. Because a convenient error term in the second stage cannot be found in general, the distribution of $\theta_t|Y^t$ gets out of hand and incomputable as t increases.

Problem 2 No "natural" or conjugate distributional choices for V_t and W_t exist in general and different distributions will give different results.

Obviously without symmetry these problems multiply.

In the literature I have seen three possible solutions proposed all having major disadvantages in a forecasting context. The most common of these is the Control Theory approach (see Snyder (1), Kalman (1), Maachi et al (1), and more recently Cleverson and Zidek (1) where the process is considered on the first two moments only, rather than on the distribution of the second stage random variable. In non-normal situations however, such moment approaches are notorious for giving misleading results (for example see Chapter 2)

The second method proposed by Leonard (1) is to transform the second stage parameters to normality and evolve the transform random variables in the usual way. This, I hope the reader will appreciate immediately removes two main advantages of using the two stage model over other techniques, namely:

(i) The transformed θ_t's and their evolution have lost their heuristic significance.

(ii) The approach is inelegant and messy.

As a third solution Harrison and Stevens ([1]) tentatively suggest that at each time stage a normal approximation to the second stage random variable should be made equating means and variances. Apart from completely losing the form of the random variable θ_t this again suffers from bad stability properties mentioned in Chapter 2.

So none of these solutions looks very promising. Let us return to the original normal model and check these statements about it given below.

(i) $\underset{\sim}{\theta}_t$ is in a "natural" parametrisation, firstly because it has meaning, as stated before, and therefore can be ascribed to some loss function which implies a natural (up to linear transformations) parametrisation. Secondly for $\underset{\sim}{W}_t$ to be specified as marginally symmetrical in its components, θ_t must be in a particular parametrisation. Symmetry of W_t is only preserved under linear transformations.

(ii) In Stage 2 $\underset{\sim}{W}_t$ is only chosen to be normally distributed because it is convenient to do this to get a nice updating relation.

(iii) In Stage 2 the adding of an error term is only a convenient way of specifying the joint distributions of the random variables θ_t in such a way that our "uncertainty" about $\underset{\sim}{\theta}_t|\underset{\sim}{Y}^{t-1}$ is greater than the uncertainty of $\underset{\sim}{\theta}_{t-1}\ \underset{\sim}{Y}^{t-1}$

With these concepts in mind I will start my own generalisation of the Bayesian forecasting approach. The most important Bayesian forecasting model is the Steady Model, so I will start with this and expand up.

Before I begin, note that in the normal case Stage 1 can be rewritten

$$\underset{\sim}{Y}_t|\underset{\sim}{\theta}_t \sim n(0,\underset{\sim}{V})$$

This equivalent representation is usually more useful (since it removes the need for the additivity of error on the first stage) when dealing with general distributions.

8.4. The Steady Model

The steady model for the Dynamic Linear Model (See Harrison and Stevens (1) is given by the equations

Stage 1 $Y_t = \theta_t + V_t \qquad V_t \sim n(O,V)$ 8.4.1.

Stage 2 $\theta_t = \theta_{t-1} + W_t \qquad W_t \sim n(O,W)$ 8.4.2.

where all error components i.e. $\{V_t, W_t \; t \in \mathbb{N}\}$ are independent.

The first part of this chapter discusses how to analyse such processes for other distributions.

Write $y^t = y_1 \ldots y_t$, then the normal steady model gives

$$\theta_t | y^t \sim n(m_t, C_t) \qquad 8.4.3.$$

where $m_t = m_{t-1} + A_t(y_t - m_{t-1})$ 8.4.4.

$$A_t = (C_{t-1} + W)\,(C_{t-1} + W + V)^{-1} \qquad 8.4.5.$$

$$C_t = C_{t-1} + W - A_t^2\,(C_{t-1} + W + V) \qquad 8.4.6.$$

It is easy to see that C_t does not depend on the observations y^t but just on the value of t, and as $t \to \infty$ $C_t \to C$ where

$$C = \tfrac{1}{2}W\,[(1 + 4r)^{\frac{1}{2}} - 1] \text{ where } r = V/W \qquad 8.4.7.$$

In fact if I just started at t = O with $C_o = C$, then C_t will not depend on t at all. So the convergence with t is due to the effects of the prior distribution. Forgetting the effects of the prior I obtain the "steady" state of the steady model.

$$m_t = m_{t-1} + A(y_t - m_{t-1}) \qquad 8.4.8.$$

$$C_t = C \qquad 8.4.9.$$

where A is the constant derived by substituting C for C_{t-1} in (8.4.5). It can be checked that this gives m_t as an exponentially weighted average of the observations y^t as $t \to \infty$.

It is now possible to generalise the process. Write

$f_t(\theta)$ = probability density function of θ_t given y^t 8.4.10.

$f_{t+1}^{(1)}(\theta)$ = probability density function of θ_{t+1} given y^t 8.4.11

Then (8.4.8) and (8.4.9) can be written

$$f_{t+1}^{(1)}(\theta) \propto (f_t(\theta))^k \qquad 8.4.12.$$

where $k = (1 + WC^{-1})^{-\frac{1}{2}}$

$$= \{1 + 2([1 + 4r]^{-\frac{1}{2}} - 1)^{-1}\}^{-\frac{1}{2}} < 1 \qquad 8.4.13.$$

an expression just depending on r, the variance ratio. This suggests the following extension to non-normal processes.

8.5. The General Steady Model

I shall begin this section with a few mysterious definitions and then given an explanatory theorem.

Suppose for all $t \in \mathbb{N}$ parameters θ_t are a-priori all contained in a bounded interval I, and let $F(I)$ denote the set of all absolutely continuous distribution functions with extended support contained in this interval.

Definitions

Choose a set of continuous functions

$$T_t : \mathbb{R}_{>0} \to \mathbb{R}_{>0} \qquad t \in \mathbb{N} \text{ such that}$$

(i) for each $t \in \mathbb{N}$ T_t is concave downwards

i.e. for all $x,y \in \mathbb{R}_{>0}$ such that $0 < y < x$

$$T(x) - T(x-y) > T(x+y) - T(x) > 0.$$

(ii) T_t converges uniformly to T (say) as $t \to \infty$ where T satisfies (i).

Then the *steady map* S_t *induced by* T_t is the map

$$S_t : F(I) \to F(I)$$

$$F(\theta) \to G_t(\theta)$$

where $g_t(\theta)$, the probability density function of $G_t(\theta)$ has the property

$$g_t(\theta) \propto T_t(f(\theta)).$$

where $f(\theta)$ is the corresponding density function of $F(\theta)$.

It is easily checked that S_t is well defined since I is compact.

Call the *Simple Steady Map* a steady map such that $T_t = T$ for all $t \in \mathbb{N}$.

Call a *Steady Model* (S.M) with respect to T_t, $t \in \mathbb{N}$ a model for which there exist maps S_t, $t \in \mathbb{N}$ such that

$$F^{(1)}_{t+1}(\theta) = S_t(F_t(\theta)) \qquad t \in \mathbb{N}$$

where $F_t(\theta)$ is the distribution of parameter θ_t given observations $y_1 \ldots y_t$

$F^{(1)}_{t+1}(\theta)$ is the distribution of parameter θ_{t+1} given observations $y_1 \ldots y_t$

and S_t is the steady map induced by T_t satisfying conditions (i) and (ii) above.

Finally call a *Simple Steady Model* (S.S.M) with respect to T the S.M. with $T_t = T$ for all $t \in \mathbb{N}$.

Now the clarifying theorem as promised.

Theorem 8.1.

Let $F_t(\theta)$ and $F^{(1)}_{t+1}(\theta)$ be defined as above and $F_t(\theta)$ be differentiable on $\mathbb{R}$. Then any (S.M) satisfies

(i) $\phi_t(b) = \phi_{t+1}(b)$

where $\phi_t(b)$ is the generalised location map associated with $F_t(\theta)$

$\phi_{t+1}(b)$ is the generalised location map associated with $F^{(1)}_{t+1}(\theta)$

(ii) If $b \in \mathbb{R}_{>0}$ is chosen such that

$$f_t(\theta) > f_t(d_b - b) \qquad \theta \in B_1$$

$$f_t(\theta) < f_t(d_b - b) \qquad \theta \in B_2$$

where $B_1 = (d_b - b, d_b + b)$

$B_2 = [d_b - b, d_b + b]^c$

and $d_b \in \phi(b)$

then $E_b(F_t, d_b) < E_b(F^{(1)}_{t+1}, d_b)$

where $E_b(G,d)$ denotes the expected loss with respect to step loss $S_b(\theta - d)$ distribution G and decision d.

Proof

(i) is a direct consequence of Theorem 3.6.

(ii) Since T is concave downwards for all $c \in \mathbb{R}_{>0}$ there exists an R(c) such that

$$R(c)\ T(x) > x \qquad 0 < x < c$$

$$R(c)\ T(x) < x \qquad c < x.$$

so by the above condition in particular

$$R(d_b-b)\ T_t(f_t(\theta)) < f_t(\theta) \qquad \theta \in B_1$$

$$R(d_b-b)\ T_t(f_t(\theta)) > f_t(\theta) \qquad \theta \in B_2$$

Hence in particular

$$\frac{\int_{\theta\in B_1} f_t(\theta)\ d\theta}{\int_{\theta\in B_2} f_t(\theta)\ d\theta} > \frac{\int_{\theta\in B_1} R(d_b-b)\ T(f_t(\theta))\ d\theta}{\int_{\theta\in B_2} R(d_b-b)\ T(f_t(\theta))\ d\theta}$$

which on rearranging gives

$$\frac{1}{\int_{\theta\in B_2} f_t(\theta)\ d\theta} > \frac{\int_{\mathbb{R}} T(f_t(\theta))\ d\theta}{\int_{\theta\in B_2} T(f_t(\theta))\ d\theta}$$

Hence $E_b(F_t,d_b) < E_b(F_{t+1}^{(1)},d_b)$

by the definition of $F_{t+1}^{(1)}$. This concludes the proof. □

Corollary

If $F_t(\theta)$ is properly unimodal, then

$E_b(F_t,d_b) < E_b(F_{t+1}^{(1)},d_b)$ for all $b \in \mathbb{R}_{>0}$. □

A discussion of the definitions

(1) Condition (ii)

The reason I must consider a set of transformations T_t rather than a single transformation T for the definition of a S.M. is purely (as in the normal case) to allow for the effects of the original prior density $p(\theta_o)$ comprising of information not gathered from the data. When considering a Steady Model by its very name I am primarily interested in its behaviour as $t \to \infty$ and the model "steadies down". Thus the edge effects introduced by the prior are of little consequence theoretically and in later examples I will restrict myself only to S.S.M's.

(2) Theorem (8.1) (i)

For the Steady Model, the Generalised Location Map $\phi(b)$ will be the same for the distribution of $\theta_{t+1}|y^t$ as it was for $\theta_t|y^t$. Hence the set of all unemotional (or Anxiety function invariant) local minima of expected loss will remain constant over this time period. This can be interpreted as the "location" of the process being fixed. Note that I have mimicked the normal steady model where the modes of the distributions of $\theta_t|y^t$ and $\theta_{t+1}|y^t$ are the same, as are the $\phi(b)$ functions associated with these two distributions (by the symmetry of the probability density functions $f(\theta_t|y^t)$ and $f(\theta_{t+1}|y^t)$.

(3) Theorem (8.1) (ii)

Clearly the final condition required is that the "spread" of the distribution of $\theta_{t+1}|y^t$ is greater than that of $\theta_t|y^t$. This is expressed succinctly in terms of Bayesian inference by using Theorem (8.1) (ii) which says that the expected step loss associated with the Bayes decision increases provided $f(\theta_t|y^t)$ is well behaved in the sense of the restriction of the theorem, for example if it is unimodal.

(The reason this restriction needs to be imposed is that one would expect a general flattening of $f(\theta_t|y^t)$ for $f(\theta_{t+1}|y^t)$. If d_b is in a neighbourhood of a deep antimode, the antimode will grow less deep under the steady model and hence $E_b(F(\theta_t|y_t),d_b)$ may actually decrease. (see Fig. 8.1).)

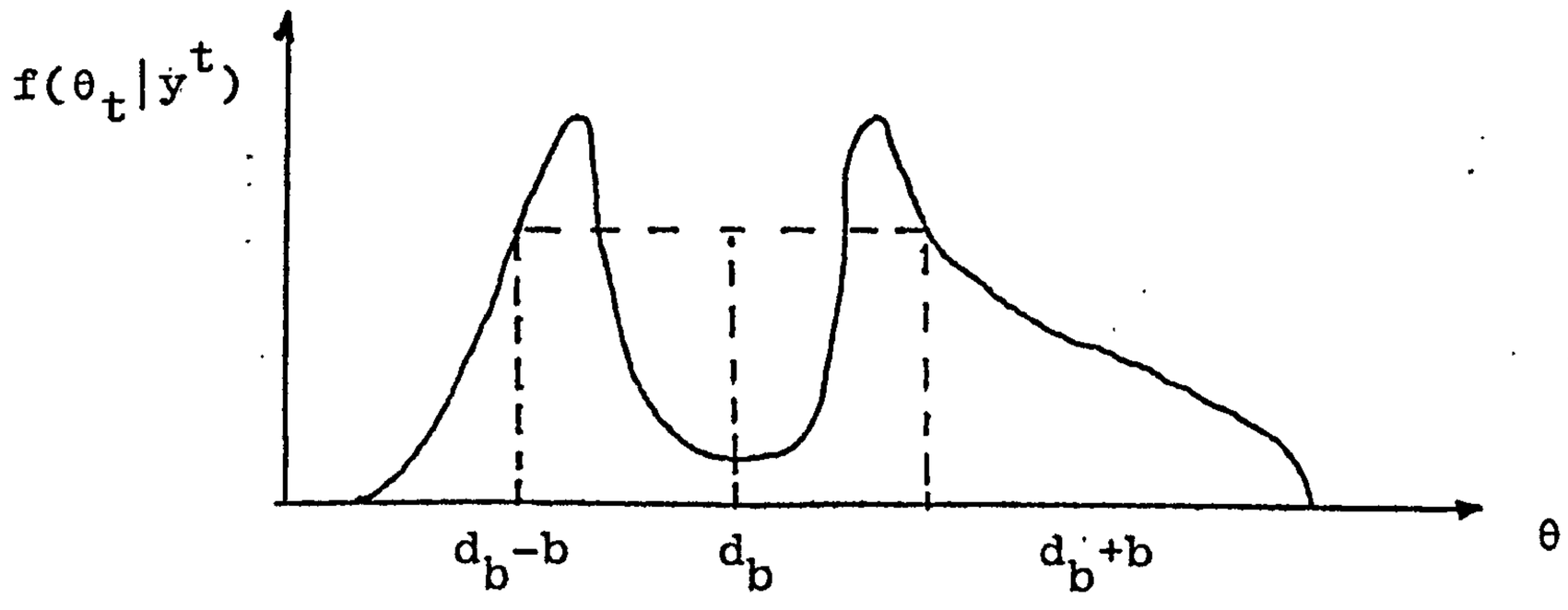


Fig. 8.1.

8.6. The Power Steady Model

The definitions above do contain one snag; the parameter space for θ_t $t \in \mathbb{N}$ must be bounded. Although this might be a reasonable restriction in many practical situations it is a theoretic embarrassment since nearly all tractable density functions have support in either $\mathbb{R}$ or $\mathbb{R}_{>0}$. This means that the induced density $f(\theta_{t+1}|y^t)$ may not be integrable and so S defined above not well defined. I can surmount the difficulty in a very interesting way however by imposing one more restriction.

Let B bea compact set and $A \subset B$. Write $B\backslash A = B \cap A^c$.

Let $f(\theta)$ be a probability density function on B

$f_A(\theta)$ be the probability density $f(\theta|\theta \in A)$

$f_{B\backslash A}(\theta)$ be the probability density $f(\theta|\theta \in B\backslash A)$

Then $f_A(\theta) = \begin{cases} p_1^{-1} f(\theta) & \theta \in A \\ 0 & \text{otherwise} \end{cases}$

$f_{B\setminus A}(\theta) = \begin{cases} p_2^{-1} f(\theta) & \theta \in B\setminus A \\ 0 & \text{otherwise.} \end{cases}$

for constants p_1 and p_2.

Condition (iii)

Choose functions T_t satisfying (i) and (ii) in the above definitions with the additional condition that for all probability density functions with support contained in B

$$T_t(f(\theta)) \propto a_1 T_t(f_A(\theta)) + a_2 T_t(f_{B\setminus A}(\theta))$$

where a_1 and $a_2 \in \mathbb{R}_{>0}$.

Definition

Call a S.M. (S.S.M) with $T_t(T)$ satisfying condition (iii) a *Power Steady Model* (P.S.M) (*Simple Power Steady Model* (S.P.S.M))

Note that this implies that

$$S_t(f(\theta)) = (p_1^*)S_t(f_A(\theta)) + (p_2^*)S_t(f_{B\setminus A}(\theta))$$

where $p_1^*, p_2^* > 0$ and $p_1^* + p_2^* = 1$.

Interpreting p_1^* as the probability that $\theta_{t+1}|y^t \in A$ and p_2^* as the probability $\theta_{t+1}|y^t \in B\setminus A$ it can now be seen that a P.S.M. is a model for which the restriction of θ to lie in a particular interval does not effect the evolution of its distribution function. In particular if I apriori restrict $\theta_i \in B$ to lie in the interval $A \subset B$ $i \in \mathbb{N}$ the distribution I obtain for $\theta_t|y_t$ is the same as the one I obtain by restricting θ_i to the interval A only at times $i = t-r, t-r+1 \ldots \infty)$ $r > t$.

This of course a property of the non time dependent Bayes analysis and a little thought should convince the reader that this is an intuitive condition to impose.

Without loss of generality drop the subscript on T.

Theorem (8.2)

T satisfies Condition (iii) if and only if

$$T(x) = \psi x^t \quad \text{for some} \quad \begin{cases} k \in (0,1) \\ \\ \psi \in \mathbb{R}_{>0} \end{cases}$$

Proof

$$T(f) \equiv T(p_1^{-1} f_A + p_2^{-1} f_{B\setminus A})$$

$$\equiv T(p_1^{-1} f_A) + T(p_2^{-1}(f_{B\setminus A}) \quad \text{since A and B\A are disjoint}$$

$$= a_1 T(f_A) + b_2 T(f_{B\setminus A}) \quad \text{if and only if}$$

$$T(cx) \propto T(x) \quad \text{for all } k,\ x \in \mathbb{R}_{>0}$$

which is equivalent to the condition

$$W(y+c) - W(y) = J(c) \quad \text{for all } y,\ c \in \mathbb{R}$$

where $W(y) = \ell n(T(\exp y))$

But this means that W is a linear function of y.

Hence (iii) is satisfied if

$$T(x) = \psi x^k \qquad k,\ \psi \in \mathbb{R}_0.$$

But Condition (i) will be satisfied in this case if and only if

$$k \in (0,1).$$

This completes the proof. □

It should be noted that without loss of generality I can drop the constant ψ since proportional T induce the same transformation $S : F(I) \to F(I)$.

It is obvious that the normal steady model is in fact a P.S.M. and so I have achieved the generalisation I require. It is also obvious that I could have proceeded by using a prior likelihood

approach (Edwards (1)) The P.S.M. model can then be formulated as

$$\ell n \ell(\theta_{t+1}|y^t) = k\, \ell n \ell(\theta_t|y^t) + \text{constant}$$

where ψ is the likelihood function and $0 < k < 1$.

This would have the advantage of removing the integrability condition on $T(\ell(\theta))$ and the disadvantage in the difficulty to find a spread concept.

I said at the beginning of this section that the P.S.M. got round the necessity for the condition of compactness of support of . The reason for this is as follows. A consequence of Bayes rule is that if W_t Z_t and U_t are random variables with

$$W_t = Z_t|A \qquad \text{then}$$

$$f(X_t|Y_t = y_t,\ X_t \in A) = f(W_t|Y_t = y_t) \qquad 8.6.1.$$

Suppose random variables θ_t with support contained in $\mathbb{R}$ are such that

$$\psi_t = \theta_t|A \qquad \text{is a P.S.M.} \quad \text{for all compact subsets of } \mathbb{R}, A. \qquad 8.6.2.$$

Then clearly by the definition of a P.S.M. and (8.6.2)

$$f(\psi_t|y_t) = f_A(\theta_t|y^t) \qquad 8.6.3.$$

where $f_A(\theta_t|y_t) = \begin{cases} p^{-1} f(\theta_t|y^t) & \theta_t \in A \\ 0 & \text{otherwise} \end{cases}$

and p is a constant to make $f_A(\theta_t|y^t)$ integrate to unity.

Thus (8.6.3) is satisfied if and only if the θ_t's satisfy

$$f(\theta_{t+1}|y^t) \propto f^{k_t}(\theta_t|y^t)$$

where $0 < k_t < 1$.

Hence provided I remember to restrict $\theta_t|y^t$ *a postiori* to a compact set I always have a well defined distribution which agrees with the

formulation given in Theorem 8.2. In this way the posterior (possibly improper) distribution on θ_t can thus just be seen as a limiting distribution of a proper P.S.M.

All the examples I shall give later (except one) have $f(\theta_t|y^t)$ integrable for all times t so usually no such problems arise anyway. The exception is the t-product.

The P.S.M. has the following pleasing properties

Let $$I_t = \int \ell_n f_t(\theta)\, dF_t(\theta)$$

$$I^{(1)}_{t+1} = \int \ell_n f^{(1)}_{t+1}(\theta)\, dF_t(\theta)$$

Then I_t and $I^{(1)}_{t+1}$ represent Shannons negative entropy at time t for θ_t and θ_{t+1} respectively. This can be used as an alternative measure of speed.

Theorem 8.3.

Let θ_t be a P.S.M. with associated transformations T_t given by

$$T_t(x) = x^{k_t} \qquad k_t \in (0,1)$$

$$t \in \mathbb{N}.$$

Then (i) $T(0) = 0$

(ii) $I^{(1)}_{t+1} = k_t\, I_t$

(iii) $f(\theta_{t+1}|y^t) \to R(\theta_{t+1}|y^t)$ pointwise as $k_t \to 0$.

where R is a rectangular distribution on the support B of $f(\theta_t|y^t)$.

Proof

(i) and (ii) follow directly from the definitions. For (iii) let

$$A_n = \{\theta \in B: \frac{1}{n} < f(\theta_t|y^t) < n\}$$

Then $B^{\sigma} = \bigcup_{n=1}^{\infty} A_n$. 8.6.4.

For each fixed $n \in \mathbb{N}$.

$$\theta_t \in A_n \implies \left(\frac{1}{n}\right)^k < (p_n(k_t))^{-1} f(\theta_{t+1}|y^t) < n^k$$

where $p_n(k_t) = \int_{A_n} f^{k_t}(\theta_t|y^t)\, d\theta_t$.

So $f_{A_n}(\theta_{t+1}|y^t) \to p_n^{-1}$ uniformly as $k \to 0$

where $p_n = \lim_{k_t \to 0} \int_{A_n} f^{k_t}(\theta_t|y^t) d\theta_t$.

and $f_{A_n}(\theta_{t+1}|y^t) = \begin{cases} p_n^{-1} f(\theta_{t+1}|y^t) & \theta_{t+1} \in A_n \\ 0 & \text{otherwise.} \end{cases}$

Let $p_n \to p$. By (8.6.4) and the definition of a P.S.M. it follows

$$f_{B^o}(\theta_{t+1}|y^t) \propto \lim_{n\to\infty} f_{A_n}(\theta_{t+1}|y^t) \to p^{-1} \text{ pointwise.}$$

The result follows. □

Let $f^{(r)}_{t+r}(\theta) = f(\theta_{t+r}|y^t)$ where $r \in \mathbb{N}$

$f_t(\theta) = f(\theta_t|y^t)$ as before.

It is then easily seen that

$$f^{(r)}_{t+r}(\theta) = \underbrace{S(S(S(S(S}_{r \text{ times}} f_t(\theta)) \qquad r \in \mathbb{N}$$

This is often complicated to work out but in the case of the S.P.S.M. with coefficient $k \in (0,1)$ it is easily seen that

$$f^{(r)}_{t+r}(\theta) \propto (f_t(\theta))^{k^r} \qquad r \in \mathbb{N}.$$

Hence predictions r steps ahead are easily found. In particular Theorem 8.3 (iii) gives that

$$f^{(r)}_{t+r}(\theta) \to R(\theta) \text{ pointwise as } r \to \infty$$

where $R(\theta)$ is the rectangular distribution on B the support of $f_t(\theta)$. Hence the distribution $F^{(r)}_{t+r}(\theta)$ expresses complete ignorance about θ_{t+r} as $r \to \infty$ apart from the fact that θ must lie in the set B. This I think the reader will agree is a useful property.

The S.P.S.M. can also easily be extended for times $t \in \mathbb{R}_{>0}$ instead of just $t \in \mathbb{N}$ and thus deal with situations where the times of observations are irregular. Just put

$$f^{(r)}_{t+r}(\theta) \propto (f_t(\theta))^{k^r} \quad \text{where } r,\ t \in \mathbb{R}_{>0}.$$

More than one observation at any time point causes no difficulty I just update the prior with the likelihood in the usual way. Due to lack of space I will not develop these generalisations in this thesis but leave them to a later date. One more theorem.

Theorem 8.4.

Let θ_t denote a P.S.M. and $\{F_t : t \in \mathbb{N}\}$ be the set of properly unimodal distributions described as before. Let the mode of F_t be m_t. Then if m_{t-1} is the most likely value of θ_t given $\underline{y}^t$

$$m_t = m_{t-1}$$

Proof

$$f_t(\theta) \propto \ell(\theta, y_t)\ (f_{t-1}(\theta))^k \quad \text{by Bayes Rule.}$$

and $$\frac{\partial}{\partial\theta} f_t(\theta)\Big|_{m_{t-1}} = 0 \quad \text{since}$$

$$\frac{\partial}{\partial\theta} \ell(\theta, y_t)\Big|_{m_{t-1}} = 0 \quad \text{and} \quad \frac{\partial}{\partial\theta} f_{t-1}(\theta)\Big|_{m_{t-1}} = 0$$

The result follows since f_t is unimodal. □

This shows the fundamental difference between the P.S.M. and other formulations of the steady model. Whereas moment approaches typically adjust observations towards a posterior mean the P.S.M. adjusts towards the posterior mode of the underlying "level" distributions.

I am now at a stage to give some explicit examples of P.S.M's.

8.7. Examples of the Simple Power Steady Model

In this section I will give three examples of distributions on the level parameter θ which give S.P.S.M's. I choose the usual conjugate forms in the first two examples simply for the sake of tractability and simplicity, I could if I wished work outside the usual conjugate forms. In the same way all distributions for θ I will choose are in fact Linear Expanding Families (See 3). Again the reason for this choice is just to give recognisable forms to the posterior distributions.

1). Beta - (Negative) Binomial Distribution

Write the Beta distribution $B(\delta,\gamma)$ as the distribution with corresponding probability density function satisfying

$$f(\theta) \propto \begin{cases} \theta^{\delta}(1-\theta)^{\gamma} & \theta \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

where $\delta, \gamma \geq 0$.

If $\theta_t|y^t$ has distribution $B(\delta_t,\gamma_t)$, then since the Beta distribution is a Linear expanding family, under the S.P.S.M. with associated coefficient k

$$\theta_{t+1}|y^t \text{ has distribution } B(\delta_{t+1}^{(1)}, \gamma_{t+1}^{(1)})$$

where $\delta_{t+1}^{(1)} = k\,\delta_t$

$$\gamma_{t+1}^{(1)} = k\,\gamma_t \qquad 8.7.1.$$

1a). Binomial distribution of first stage

Let Y_t $t \in \mathbb{N}$ have binomial distributions given by

$$P(Y_t = y_t) = \begin{cases} \binom{n}{y_t} \theta^{y_t}(1-\theta)^{n-y_t} & y_t \in Z_{\geq 0} \\ 0 & \text{otherwise} \end{cases}$$

Then if $\theta_t|y^t$ has distribution $B(\delta_t,\gamma_t)$ under the S.P.S.M. using (8.7.1) and the usual conjugate analysis,

$\theta_{t+1}|y^{t+1}$ has distribution $B(\delta_{t+1}, \gamma_{t+1})$ where

$$\boxed{\begin{aligned} \delta_{t+1} &= k\,\delta_t + y_{t+1} \\ \gamma_{t+1} &= k\,\gamma_t + n - y_{t+1} \end{aligned}}$$

It follows that if θ has a Beta distribution in particular

$$\delta_T + \gamma_T \to n(1-k)^{-1} \qquad \text{as } T \to \infty$$

$$\delta_T \to \sum_{t=0}^{\infty} k^t\, y_{T-t} \qquad \text{as } T \to \infty$$

Thus the posterior mode m_T of ${}_T y^T$ satisfies

$$m_T = (1-k) \sum_{t=0}^{\infty} k^t\, p_{T-t}$$

where $p_t = \frac{y_t}{n}$ is the proportion of success at time t.

Hence I obtain in the limit under the S.P.S.M. a distribution $\theta_T|y^T$ which has mode an *Exponentially Weighted Moving Average* (E.W.M.A) of the proportion of successes up to that time.

It will soon be realised by the reader that the E.W.M.A. is intrinsic to the location of posterior distributions arising from S.P.S.M's. Anyone who has worked in practice will know how useful this average is to forecasting. This time however I have given theoretical reasons why this type of estimator should be expected to be good. Of even more importance is that since in this Bayesian approach information is expressed in terms of a distribution, ideas of the accuracy of estimators are easily worked out.

The estimator of θ_{t+1} will be, of course, the Bayes decision obtained as a minima of the expected loss with respect to the

distribution of $\theta_{t+1}|y^t$ and its spread the associated expected loss.

1b). <u>The Negative Binomial Distribution on first stage</u>

Let Y_t have negative binomial distributions given by

$$P(Y_t=y_r) = \begin{cases} \binom{r+y_{t-1}}{y_t} \theta_t^r \ (1-\theta_t)^{y_t} & y_t = 0,1,2.... \\ 0 & \text{otherwise} \end{cases}$$

Borrowing the notation above, the corresponding recurrent relationship for the S.P.S.M. is then

$$\delta_{t+1} = k\,\delta_t + r$$

$$\gamma_{t+1} = k\,\gamma_t + y_t$$

Again, with Beta prior on θ_1

$$\delta_T \to r(1-k)^{-1} \qquad T \to \infty$$

$$\delta_T + \gamma_T \to \sum_{t=0}^{\infty} k^t(r + y_{T-t}) \qquad T \to \infty$$

Thus in particular the posterior mode m_T satisfies

$$m_T^{-1} = (1-k)^{-1} \sum_{t=0}^{\infty} k^t (P_{T-t})^{-1}$$

where $P_{T-k} = \dfrac{r}{r+y_{T-k}}$, the proportion on succession

$r + y_{T-t}$ trials. Again this is a E.W.M.A.

2. <u>Gamma-Poisson/Exponential Distribution</u>

Write the Gamma distribution $G(\gamma,\beta)$ as the distribution with corresponding probability density function $f(\theta)$ satisfying

$$f(\theta) \propto \begin{cases} \theta^{\gamma} e^{-\beta\theta} & \theta > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\gamma,\beta \in \mathbb{R}_{>0}$.

If $\theta_t|y^t$ has distribution $G(\gamma_t,\beta_t)$ then since the Gamma distribution is a linear expanding family, under the S.P.S.M. with associated coefficient k,

$$\theta_{t+1}|y^t \text{ has distribution } G(\gamma_{t+1}^{(1)},\beta_{t+1}^{(1)})$$

where $$\gamma_{t+1}^{(1)} = k\,\gamma_t$$

$$\beta_{t+1}^{(1)} = k\,\beta_t. \qquad 8.7.2.$$

2a). Poisson observations on the first stage.

Let Y_t $t \in \mathbb{N}$ have Poisson distribution given by

$$P(Y_t = y_t) = \begin{cases} \dfrac{(\exp - \theta_t)\theta_t y_t}{(y_t)!} & \text{for } y_t \in Z_{>O} \\ \\ O & \text{otherwise} \end{cases}$$

Then if $\theta_t|y^t$ has distribution $G(\gamma_t,\beta_t)$ under the S.P.S.M. using (8.7.2) and the usual conjugate analysis

$$\theta_{t+1}|y^{t+1} \text{ has distribution } G(\gamma_{t+1},\beta_{t+1}) \text{ where}$$

$$\boxed{\begin{aligned} \gamma_{t+1} &= k\,\gamma_t + y_{t+1} \\ \beta_{t+1} &= k\,\beta_t + 1 \end{aligned}}$$

It follows that if θ_1 has a Gamma distribution, in particular

$$\gamma_t \to \sum_{t=O}^{\infty} k^t\, y_{T-t} \qquad \text{as } T \to \infty$$

$$\beta_t \to (1-k)^{-1} \qquad \text{as } T \to \infty .$$

Thus the posterior mode of $\theta_T|y^T$, m_T satisfies

$$m_T = (1-k) \sum_{t=O}^{\infty} k^t\, y_{T-t}.$$

the E.W.M.A.

Note that since the $\phi(b)$ function for the Gamma distribution is strictly increasing, any Bayes decisions made under symmetric loss will be strictly greater than this value (See §3).

2b). Exponential observations on the first stage.

(Here for simplicity I will consider a S.P.S.M. on θ whereas strictly by argument in Chapter 2, a S.P.S.M. on $\ln \theta$ might be more sensible).

Let Y_t have exponential distribution given by the probability density function $f(\theta_t)$ where

$$f(\theta_t) \propto \begin{cases} e^{-t^{y_t}} & y_t \in \mathbb{R}_{>0} \\ 0 & \text{otherwise} \end{cases}$$

Borrowing the notation from above again, the corresponding recurrent relationship for the S.P.S.M. is then

$$\boxed{\begin{aligned} \gamma_{t+1} &= k\,\gamma_t + 1 \\ \beta_{t+1} &= k\,\beta_t + y_t \end{aligned}}$$

Again with Gamma prior on θ_1

$$\gamma_T \to (1-K)^{-1} \qquad \text{as} \qquad T \to \infty$$

$$\beta_T \to \sum_{t=0}^{\infty} k^t\, y_{T-t} \qquad \text{as} \qquad T \to \infty$$

Thus in particular the posterior mode m_T satisfies

$$m_T^{-1} = (1-k) \sum_{r=0}^{\infty} k^t\, y_{T-t}$$

A Bayes decision with respect to symmetric loss will always be strictly greater than m_T given above for the "rate of decay" of this process.

3) Student t sample distribution Steady Model

Suppose the sample distribution for observations in the S.P.S.M is not normal, but perhaps more sensibly assumed to have a student t probability density function

$$\text{i.e.} \quad p(Y_t|\theta_t) \propto (1 + V^{-1}(\dot{y}_t - \theta_t)^2)^{-\alpha} \qquad \alpha > \tfrac{1}{2} \qquad 8.7.3.$$

Either using limiting arguments or using a prior on θ_1 of the form

$$p(\theta_1) \propto (1 + V^{-1}(y_o - \theta_o)^2)^{-\alpha} \qquad 8.7.4.$$

using the S.P.S.M. at time t = T I have the t-product for $\theta_T|y^T$

$$p(\theta_T|y^T) \propto \prod_{t=0}^{T} (1 + V^{-1}(y_t-\theta_T)^2)^{-\alpha k^{(T-t)}} \qquad 8.7.5.$$

Obviously this does not have the same neat posterior form as the normal but has great advantages in practice because it is a more realistic model. I shall examine the posterior distribution more closely.

Slowly varying observations

The posterior mode(s) m_T of $\theta_T|y^T$ is obtained by taking logs and differentiating once giving m_T as the solution of the equation

$$\sum_{t=0}^{T} k^{(T-t)} \frac{(m_T - y_t)}{(V + (m_T-y_t)^2)} = 0 \qquad 8.7.6.$$

This in general may have many solutions m for m_T but if the observations y_T are *close together*

$$V + (m_T-y_t)^2 \simeq V$$

so (8.7.6) reduces to

$$m_T \simeq \left(\sum_{r=0}^{T} k^{T-t}\right)^{-1} \sum_{r=0}^{T} k^{T-t} y_t$$

which as $T \to \infty$ gives

$$m_T \simeq (1-k) \sum_{t=0}^{T} k^{T-t} y_t$$

the E.W.M.A. (just as in the normal case!).

For Bayes estimates under symmetric bounded loss I obviously require the types of conditions on $\phi(b)$ discussed in Chapter 3. By Theorem 3.7.

$$\phi(b) \to y_T \text{ as } b \to \infty$$

so roughly speaking as long as observations are not flying about too much, Bayes estimates will lie approximately an interval around

(i) the exponentially weighted moving average up to y_T

(ii) the observation y_T.

Outlying Observations

Suppose now that y_T is a long way away from the body of the other data y^{T-1}. Then since for a Student t-product $p(\theta)$.

$$p'(\theta) \to 0 \qquad \text{as} \qquad |\theta| \to \infty$$

the posterior distribution $p(\theta_T|y^T)$ will be at least bimodal. The bimodality will express the difficulty in discerning whether there has been a change of "level" at time T or whether y_T was in fact a rogue observation. If k is very close to 1, the highest mode will tend to be in the region of $\mathbb{R}$ around y^{T-1}, and in fact if these early data points were fairly close as above, the highest posterior mode of $\theta_T|y^T$ will be approximately an E.W.M.A. of the y^{T-1} data. Conversely if k is small so that information from time stage to time stage is weak, the highest mode will be the one in a neighbourhood of y_T.

Usually, it will be appropriate to choose k large. In this outlier situation I then have that Bayes estimates will lie approximately in an interval around.

(i) the E.W.M.A. up to y_{T-1}

(ii) the observation y_T.

If then y_{T+1} y_{T+2} ... are observed in a region of y_T, the highest mode will flip from the E.W.M.A. of y^{T-1} to the mode near the new observations.

Thus it can be seen that such a model has many practical advantages over the normal model usually used. Firstly it retains the possibility of a pragmatic choice of estimate all important in modelling. Secondly it registers "jumps" in level and adjusts to them far more quickly than does the normal approach. Hence model "break-downs" are unlikely to occur (it is in this sense a far better local first order approximation model than the normal).

Truncation of the Parameter Space

As hinted on previously in this chapter, another big advantage of this method is that the analysis does not depend on whether I truncate the parameter space. For example in the Poisson-Gamma process given in Example 2, I could put a truncate Gamma rather than just a Gamma on the second stage. The effect would be to give a Gamma distribution with the same parameters as posterior distribution but this time truncated over the original trancation interval. In many applications it is far more sensible to put a truncated conjugate prior that the actual conjugate distribution to represent prior beliefs. For example, in the normal model there is usually a constrain on the upper and lower band of the level.

Another advantage in the P.S.M. is that I do not need conjugacy for it to work or in fact be tractable. However it can clearly be seen that in all the cases cited above I will tend towards the conjugat posterior distribution anyway, regardless of the original prior I use. This is however due to the fact that I am working with expanding families of distributions.

8.8. Multivariate Simple Power Steady Models

In the same way as for the univariate case, I need to find an up-date relationship of the distribution of $\underset{\sim}{\theta}_t|y^t$ to the distribution of $\underset{\sim}{\theta}_{t+1}|y^t$ where this time $\underset{\sim}{\theta}_t|y^t$ and $\underset{\sim}{\theta}_{t+1}|y^t$ are random *vectors*

Writing $f_t(\theta)$ as the probability density function of $\underset{\sim}{\theta}_t|y^t$ and

$f^{(1)}_{t+1}(\theta)$ as the probability density function of $\underset{\sim}{\theta}_{t+1}|y^t$

an obvious first candidate would be to let

$$f^{(1)}_{t+1}(\theta) \propto (f_t(\theta))^k \quad \text{where } k \in (0,1). \qquad 8.8.1.$$

It is fairly clear that any generalisation of the univariate S.P.S.M. must be expected to at least contain this type of model. It is not quite general enough unless there is some sort of symmetry about the parameters $\underset{\sim}{\theta}$ since information about each component of $\underset{\sim}{\theta}$ dissipates at the same rate, but it does allow for the following two examples.

4. Dirichlet/Multinomial distribution

Suppose Y_y $1 \leq t \leq T$ are random variables taking values

$Y_t \in K_r$ with probability $\theta_r(t)$ $1 \leq r \leq n$, where

$1 \leq t \leq T$

for each t $1 \leq t \leq 1$ $0 < \theta_r(t)$, $\sum_{r=1}^{n} \theta_r(t) = 1$. Then assuming I have a model governed by (8.8.1) and that $\theta_r(t)$ are distributed Dirichlet i.e.

$$f_t(\theta) \propto \prod_{r=1}^{n} \theta_r{}^{S_r(t)} \qquad 0 < S_r(t) \text{ for } 1 \leq r \leq n$$

$1 \leq t \leq T.$

then this steady model generates the recurrence relationships

$$S_r(t) = \begin{cases} k\, S_r(t-1) + 1 & y_t \in K_r \\ k\, S_r(t) & \text{otherwise} \end{cases}$$

Define the random variable

$$X_i(r) = \begin{cases} 1 & \text{if } X_i \in k_r \\ 0 & \text{otherwise.} \end{cases}$$

Then in the limit

$$S_r(T) \rightarrow \sum_{i=0}^{\infty} k^i x_{T-i}(r) \qquad \text{as } T \rightarrow \infty .$$

Of course by the definiton of $X_i(r)$ this implies that

$$\sum_{r=1}^{n} S_r(T) \rightarrow (1-k)^{-1} .$$

Note therefore that in the limit $\theta_r(T)$, the probability $Y_T \in K_t$ has marginal distribution with mode

$$(1-k) \sum_{i=0}^{\infty} k^i x_{T-i}(r) \qquad 8.8.3.$$

again an E.W.M.A. of the observations.

Hence instead of using a parametric evaluation of my distributic for the random variable Y_t I can, in fact, use this type of non-parametric histogram approximation. Thus I have a non-parametric time series. Of course I would obviously need to have fairly frequent observations for this to work better than parametric forms of the same problem, but it is still rather exciting.

I can even go one further than this by using a paper by Fergusor (1). I shall not go into any detail, but in the limit his estimate for the distribution of Y_t is given by

$$F_T(y|Y^t) = 1/T \sum_{i=1}^{T} X_i(-\infty, y)$$

where $X_i(A) = \begin{cases} 1 \text{ if } Y_i \in A \\ 0 \text{ otherwise} \end{cases}$

Well, in the Times series steady model case I obtain the result

$$F_T(y|Y_o \ldots Y_T) = (1-k) \sum_{i=0}^{T} k^i X_{T-i}(-\infty, y)$$

for the posterior distribution of the distribution of Y_T at time T. Again, by similar arguments to those of Ferguson, the posterior expectation of Y_T, μ_T is given by

$$\mu_T = (1-k) \sum_{i=0}^{\infty} k^i y_{T-i}$$

and posterior variance V_T given in the limit as

$$V_T = (1-k) \sum_{i=0}^{\infty} k^i (y_{T-i}-\mu_T)^2,$$

an avalanche of E.W.M.A.'s

5. Normal-Variance Unknown Steady Model

Often in the normal steady model the variance is unknown and needs to be estimated. This can be done very simply by using euqation (8.8.1). Throughout I will use the usual conjugate analysis advocated by De Groot (1) and follow his notation as much as possible.

I will assume that I am primarily interested in the logarthmic transform of the variance V. This is the 'hatural parametrisation" of V advocated in Chapter 2. However, if the reader preferred working directly with either the precision or variance similar limiting result, though slightly different would arise.

Let $n(\mu,V)$ denote the normal distribution mean μ variance V

$G(\alpha,\beta)$ denote the Gamma distribution i.e.

$$G(\alpha,\beta, x) \propto \begin{cases} x^{\alpha-1} \exp - \beta x & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha,\beta > 0$.

Since I have chosen the conjugate forms and the family is Linearly

Expanding it follows that if data Y_t, $t \in \mathbb{N}$

$$Y_t \sim n(\theta_t, r_t^{-1}) \text{ then}$$

$$\theta_t|y^t, r_t \sim n(\mu_t, \tau_t r_t^{-1})$$

$$r_t|y^t \sim G(\alpha_t, \beta_t)$$

For some parameters μ_t, τ_t, α_t, β_t. If I use the log transform on r

$$p(\theta_t, \ell n\ r_t|y^t) \propto r_t^{\alpha_t+\frac{1}{2}} \exp(-\tfrac{1}{2} r_t \{\tau_t(\theta_t-\mu_t)^2 + 2\beta_t\})$$

Using the S.P.S.M. I have the following recurrence relationship.

Time $t \mid y^t$	Time $t+1 \mid y^t$	Time $t+1 \mid y^{t+1}$
μ_t	μ_t	$\mu_{t+1} = \dfrac{k\tau_t \mu_t + y_{t+1}}{k\ \tau_{t+1}}$
τ_t	$k\ \tau_t$	$\tau_{t+1} = k\ \tau_t + 1$
$\alpha_t + \frac{1}{2}$	$k(\alpha_t + \frac{1}{2})$	$\alpha_{t+1} = k(\alpha_{t-1} + \frac{1}{2})$
β_t	$k\ \beta_t$	$\beta_{t+1} = k\ \beta_t + \dfrac{k\ \tau_2(y_{t+1} - \mu_t)^2}{2(k\ \tau_t + 1)}$

(I have used the standard update formulas (see De Groot (1) Page 169).

I can now find limiting forms for all these parameters. It is easily checked that

$$\tau_T \to \tau = (1-k)^{-1} \qquad 8.8.4.$$

$$\alpha_T \to \alpha = \tfrac{1}{2}\ ((1-k)^{-1} - 1) \qquad 8.8.5.$$

$$\text{as } T \to \infty$$

By the update relationships of μ_t and β_t with the limiting forms given in (8.8.4) and (8.8.5) gives that

$$\mu_{t+1} = k\ \mu_t + (1-k)\ y_{t+1} \qquad 8.8.6$$

$$\beta_{t+1} = k\ (\beta_t + \tfrac{1}{2}\ S^2_{t+1}) \qquad 8.8.7.$$

where $S_{t+1} = y_{t+1} - \mu_t$.

Hence $\mu_T \to (1-k) \sum_{t=0_T}^{T} k^t\ y_{T-t}$ 8.8.8.

and $\beta_T \to \tfrac{1}{2}\ k \sum_{r=-0}^{} k^t\ S^2_{T-t}$ 8.8.9.

As might be expected the posterior mean of Y_t is the same as in the variance unknown case. The posterior distribution of the variance however is very interesting. The mode of the log transform of the precision (the lower bound of the estimate for the variance (see Chapter 3)) m_T is given by

$$m_T = \frac{\beta_T}{\alpha_T}$$

which by (8.8.8) and (8.8.9) has limiting form

$$m_T = (1-k) \sum_{r=-0}^{T} k^t S_t^2 \qquad 8.8.10.$$

the E.W.M.A. the reader has grown to expect. Note that the corresponding variance of the level θ_t is given by (1-k) m_T. This suprisingly simple and intuitive result I find especially pleasing.

6. <u>Mean vector of observations Multivariate Normally distributed, (Covariance Matrix Known)</u>

Suppose $\underset{\sim}{Y}_t$ is an n vector $Y_t^{(1)} \ldots Y_t^{(n)}$ of observations at time t. Let

$$\underset{\sim}{A}\underset{\sim}{Y} \quad n(\underset{\sim}{\theta}_t, \underset{\sim}{V}_t)$$

where $n(\underset{\sim}{\mu},\underset{\sim}{V})$ represents the normal distribution with mean vector $\underset{\sim}{\mu}$ and covariance matrix $\underset{\sim}{V}$. Then assuming that it is appropriate to use the evaluation given by equation (8.8.1) on the vector $\underset{\sim}{\theta}_t$ (for example one could usually make this assumption if the $\underset{\sim}{\theta}_t$ were apriori exchangeable) I have the following evaluation.

If $\theta_t|y^t \sim n(\underset{\sim}{\mu}_t,\underset{\sim}{V}_t)$

then $\theta_{t+1}|y^t \sim n(\underset{\sim}{\mu}_{t+1}^{(1)}, \underset{\sim}{V}_{t+1}^{(1)})$ where

$$\boxed{\begin{aligned} \underset{\sim}{\mu}_{t+1}^{(1)} &= \underset{\sim}{\mu}_t \\ V_{t+1}^{(1)} &= k^{-1} V_t \end{aligned}} \qquad \text{where } k \in (0,1]$$

Notice that for time t to time t+1, adding no more information, the means will remain the same. *So will the correlation vector.* This is the first derivation from the Bayesian forecasting in Harrison and Stevens (1) where they would usually put independent errors on the second stage so that the correlation over this interval would automatically change. Note too that in my model the marginal variance of each component of the mean vector $\theta_t^{(i)}$ will increase, as would be expected.

It would now be quite simple to use the exchangeability model proposed by Lindley and Smith (1) in a time series setting.

The Principle of Stacking

Although the method of updating given by equation (8.8.1) is useful it is a bit too restrictive for many cases because, as mentioned before it assumes that information about each component of the random vector decays at the same rate. In many applications this is not the case, some parts of the model must of necessity be assumed far more "stable" with time than others because their likelihood component caves in much flatter so that otherwise information would be decayed away faster than it came in.

The simplest way I have found to deal with this problem is by "stacking" the respective distributions. Let $\theta_1 \ldots \theta_n$ be parameters ordered in such a way that "information" about θ_i decays at least as fast as information about θ_{i+1} $1 \leq i \leq n-1$. Use the notation of the previous section.

Definition *A Stacked Simple Steady Model* S.S.S.M. is one for which

$$f_{t+1}^{(1)}(\theta_i|\theta_{i+1} \ldots \theta_n) \propto (f_t(\theta_i|\theta_{i+1} \ldots \theta_n))^{k_i} \qquad 1 \leq i \leq n-1$$

$$f_{t+1}^{(1)}(\theta_n) \propto (f_t(\theta_n))^{k_n}$$

where $0 < k_i \leq k_{i+1} \leq 1 \qquad 1 \leq i \leq n-1.$

Notes

1). Equation (8.8.1) gives a special case of a S.S.S.M. with $k_i = k \quad 0 < k < 1$.

2). If $\theta_{1,t} \ldots \theta_{n,t}$ are individually S.S.M's and are independent then trivially $\underset{\sim}{\theta}$ is a S.S.S.M.

3). If $\theta_1 \ldots \theta_n$ is a S.S.S.M. and I keep parameters $\alpha_1 \ldots \alpha_m$ fixed then $(\theta_1 \ldots \theta_n, \alpha_1 \ldots \alpha_m)$ is a S.S.S.M. Hence I can replace $\alpha_1 \ldots \alpha_m$ by these estimates without losing the form of the model.

4). In general a S.S.S.M. marginalised across one of its parameters is *not* and S.P.S.M.

5a). Normal Variance unknown

Here I consider case 5) but instead of the model imposed by equation (8.8.1) replace it by the more general S.S.S.M. putting $\theta_1 = \theta$ and $\theta_2 = \ell n\ V$ it is easily checked that the update of μ_t and τ_t remain the same with k_1 written for k and the update of α_t and β_t are given by the equation.

Time $t \mid y_{t+1}$	Time $t+1 \mid y_{t+1}$
$\alpha_t + \frac{1}{2}$	$\alpha_{t+1} = k_2(\alpha_t + \frac{1}{2}) + \frac{1}{2}$
β_t	$\beta_{t+1} = k_2\beta_t + \dfrac{k_1 \tau_t (y_{t+1} - \mu_t)^2}{2(k\ \tau_t + 1)}$

The limiting forms of τ_t and α_t are the same with k replaced by k_1 and k_2 respectively as in equations (8.8.4) and (8.8.5). Also using the limiting form for τ

$$\beta_{t+1} = k_2(\beta_t + \tfrac{1}{2} S_{t+1}^2). \qquad \text{where } S_{t+1} = y_{t+1} - \mu_t$$

The mode of the log transform of the variance m_T is now given by

$$m_T = (1-k_2) \sum_{t=0}^{T} k_2^t S_t^2$$

Hence I obtain the same weighted average forms, but this time with a longer "memory" than the weight average form for the posterior mode of θ_t which is given by

$$m_T^* = (1-k_1) \sum_{t=0}^{T} K_1^t \, y_{T-t}.$$

6a). Mean vectors for Bivariate normal

For simplicity take the model given in 6) with n = 2. The new update system now gives the recurrence relations (using well known multivariate normal theory)

$$\mu_{t+1}^{(1)} = \mu_t$$

$$V_{t+1}^{(1)} = \begin{bmatrix} V_{11}^{(1)} & V_{12}^{(1)} \\ V_{12}^{(1)} & V_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} k_1^{-1} V_{11} - (k_1^{-1}-k_2^{-1})V_{12}, k_2^{-1}V_{12} \\ k_2^{-1} V_{12} \, , \quad k_2^{-1} V_{22} \end{bmatrix}$$

$$\text{where } V_t = \begin{bmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{bmatrix}$$

Obviously this update can be generalised for general sizes of vectors but the equations are a bit tedious to write down.

7). The Mixed Model

This is almost a trivial consequence of the definition of a S.S.S.M. Suppose I have no models giving different evaluations for a parameter θ_t with time each of these being S.P.S.M's. Then by note 3 I have a Stacked Simple Steady Model if I incorporate the weighting on the n models as the distribution of a lower stage random variable $\theta_{2,t}$.

7a). The Normal Mixture S.S.S.M.

Let $\theta_{1,t}|\theta_{2t} \sim n(\mu_1 + \theta_{2t}, V)$ 8.8.11

and $\theta_{2t} \sim P(\theta_{2t} = 0) = \alpha$

$P(\theta_{2t} = \Delta) = 1-\alpha$ 8.8.12

where $0 < \alpha < 1$. Then I have a mixture of 2 normals with equal variances. Imposing a S.S.S.M. with for example

$$k_2 = 1. \quad k_1 = k.$$

I have a simple update form for the mixture, each normal component updating as before and α being adjusted as in the static case.

Catastrophe Models in Time Series

I now refer to Chapter 4. On the S.S.M. models I can expect Catastrophes on the decision space (Type I model) if the distribution satisfies the condition in Chapter 6. For example, putting a loss function on the variance component in Example 5 above, between observations I have smooth movement and possible catastrophe's since the Inverted Gamma has a polynomial tail. Again Example 7a, gives an evolution of a mixture which is steady and allows for the sort of Catastrophes discussed in Chapter 7. So at last, in these late stages of the thesis I can start modelling general forms of Catastrophes in Statistics. For Type II models I need to get away from the steady model. I hope that the reader has been convinced that this model gives remarkably single and intuitive results for a large class of examples I must now generalise away from the Steady Model and try and formulate a theory for general time series.

8.9. A Discussion of some of the Drawbacks of Box-Jenkins Modelling

Usually in the analysis of time series the unit of the stationary series is used (See Box-Jenkins (1), Anderson (1)) the rest of the model being represented as combinations of these stationary series. This can be done by:

(i) Isolating the Trend and Seasonal from the model
(by least squares?)

(ii) Differencing until a stationary series is arrived at.

Procedure (i) is rarely criticised but is to my opinion very difficult to justify. In most practical time series Trend and Seasonal account for most of the variation in the data, but these are immediately "taken out" in an ad hoc way to leave an ARIMA model for which there is a reasonable amount of theory. Polynomial regression in particular is very often misused in the context of removing trends. I feel that one should take the attitude that there is always a model underlying a particular time series and the modeller should always be prepared to make an assumption of the dynamic governing the model so give him a sensible family of trend curves. Other statisticians seem to have misgivings about some of these points (e.g: see Whittle (1)).

Having followed procedure (i) the Box-Jenkins modeller then uses procedure (ii) and possibly differences once or twice to obtain a time series he can deal with. Usually he will difference once. I can illustrate why one can expect this to be the case by the following example,

Let us assume that the Trend component is known and additive and that there is no seasonal component

Then the original series $Y_t = h(t) + Z_t$

where h(t) represents the trend. If Z_t were stationary(see Fig. 8.23) then I would expect flow lines pulling Y_t on to h(t). If, however, Z_t is a first difference stationary model then h(t) represents a *vector field* and Z_t is jostled across its flow lines by random variation (see Fig. 8.3). When working in disciplines like Economics one often talks of an upward drift of an observable. The latter interpretation is then more realistic.

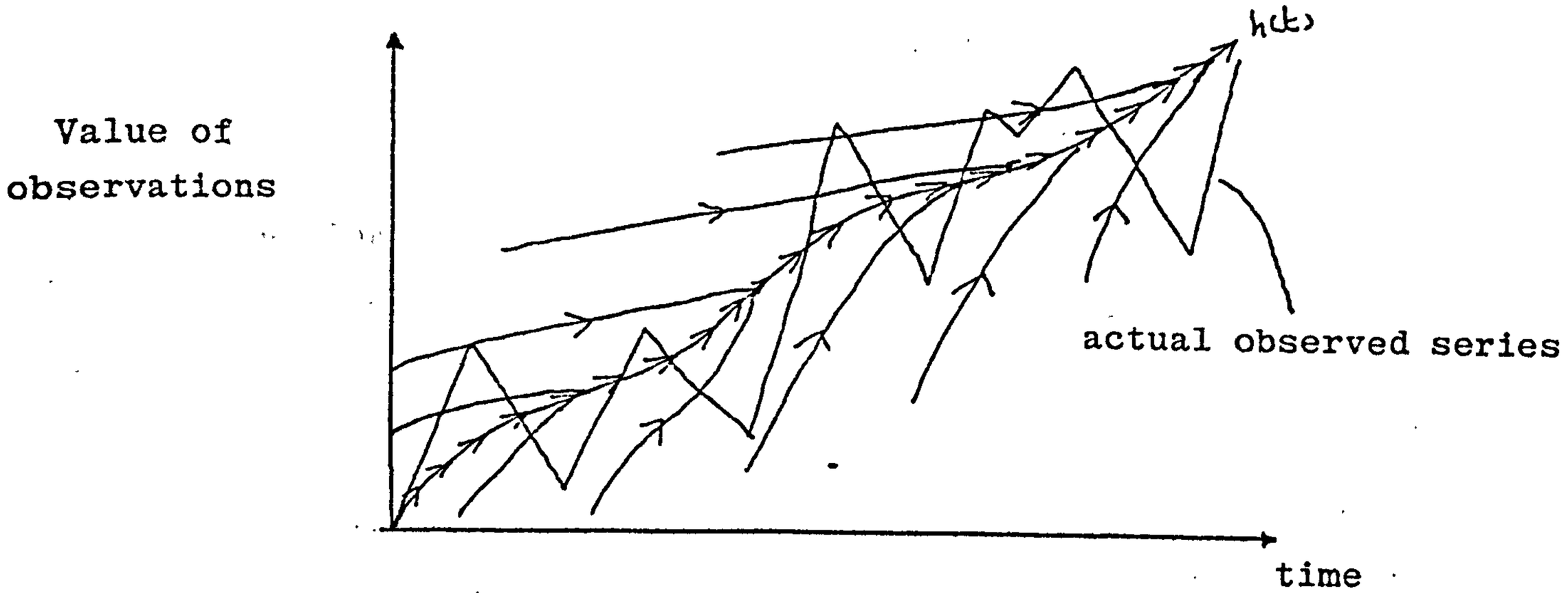


Fig. 8.2.

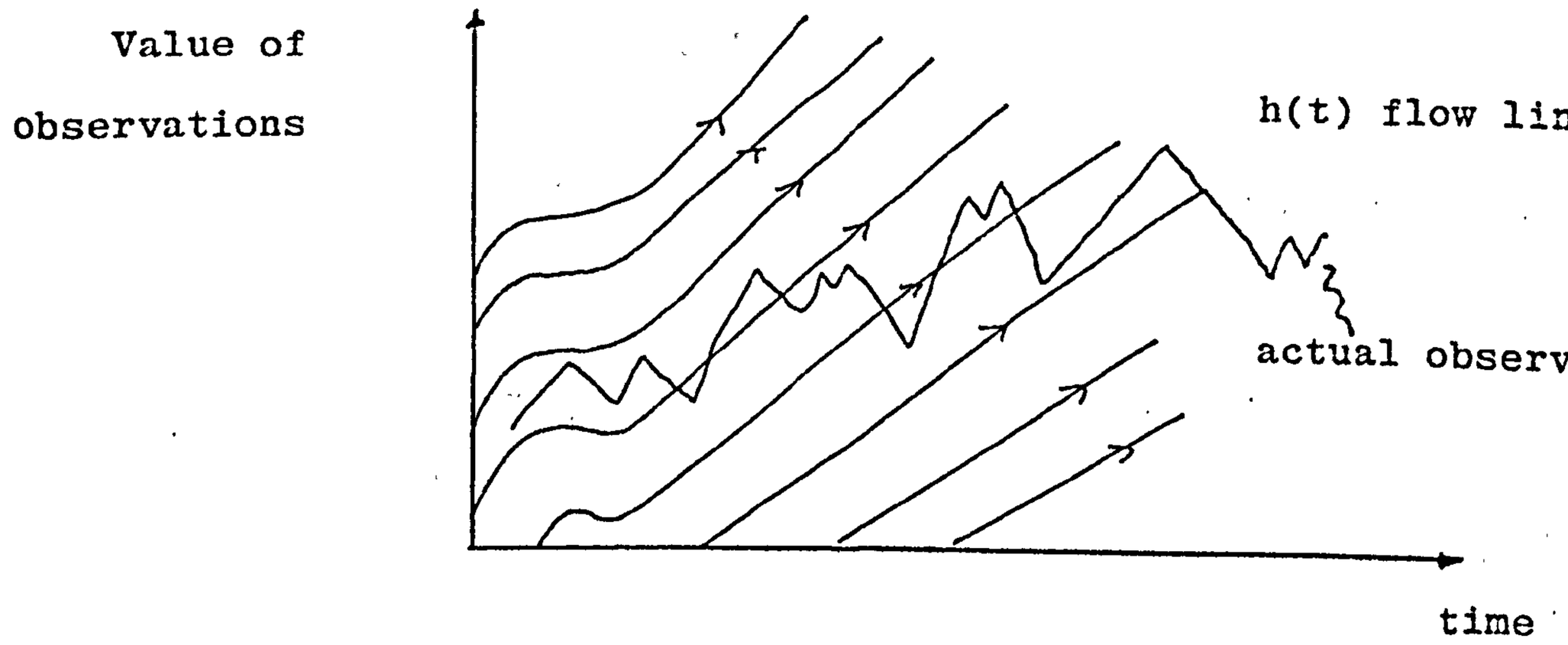


Fig. 8.3.

But why work with stationary processes as the "unit" in the first place? Would it not be better to work in units of Steady models instead? This provokes the following definition.

Definition

Suppose random variables Y_t $t \in \mathbb{N}$ are such that

$Y_t|\underset{\sim}{\theta}_t \sim g(y_t \mid \underset{\sim}{\tau}_t(\underset{\sim}{\theta}_t))$ where τ_t is a known function

$f(\underset{\sim}{\theta}_t^{(1)}) \propto \rho_t(f(\underset{\sim}{\theta}_{t-1}))$ where ρ_t is a known function

and where the notation is as before. Then call Y_t a Bayes Time Series (B.T.S).

Example - Trend model

This can simply be given by putting

$\tau_t(\theta) = \theta_t + h(t)$ where h(t) is some known function of t

and θ_t is a S.P.S.M. i.e. ρ_t is the power transformation

$f(\theta_t) \propto (f(\theta_{t-1}))^k \qquad 0 < k < 1.$

Example - Periodic functions model (Normal Case)

Here I follow Harrison and Stevens (1) using the same analogy as before. So put ρ_t to be the linear function.

$\rho_t(\underset{\sim}{\theta}_t) = \sum_{i=1}^{n} \rho_{it}(\theta_{it})$ where ρ_{it} are periodic function of t

and

$f(\underset{\sim}{\theta}_t) \propto (f(\underset{\sim}{\theta}_{t-1}))^k \qquad 0 < k < 1.$

Most typically ρ_{it} will be sine and cosine functions. To update now just use the usual Bayes formulae.

Example - Type II Catastrophe Models

Referring back to Chapter 4, Section 1 the model proposed here is in fact a B.T.S. with T_t the identity map and the evolution of θ_t defined by

$$f(\theta_t^{(1)}) \propto \begin{cases} (f(\ _{t-r}))^{k^r} & \text{if } \xi(t) \in (A \cup P)^c \\ (f(\theta_{t-1}))^{k} & \text{otherwise.} \end{cases}$$

Some Further Ideas and a Final Generalisation

It will have been seen that all the conjugate models proposed in this Chapter have been 2 stage heirarchical models. What may have been missed is that each one can be expressed equivalently as a 3 stage model of a *deterministic* character.

For example the Steady Gamma-Poisson process of Example 2, given at the beginning of the previous section could be written in the form

$$\begin{aligned} Y_t|\theta_t &\sim \text{Poisson } (\theta_t) \\ \theta_t|\gamma_t^{(1)},\beta_t^{(1)} &\sim \text{Gamma } (\gamma_t^{(1)},\beta_t^{(1)}) \\ \gamma_t^{(1)} &= k\,\gamma_{t-1} \\ \beta_t^{(1)} &= k\,\beta_{t-1} \end{aligned}$$

where again I have used the notation of the example.

Again the standard Growth Model (Harrison and Stevens (1) can be specified in terms of deterministic up-dates on the second stage mean and variance over the time interval [t,t+1).

Synthesising these pieces of information it follows that I am dealing with fundamentally *deterministic evolutions* on the hyper parameters. Of course such deterministic evolutions must be specified for well chosen reasons together with reasons for remaining in conjugate form (as in the above examples of the Steady Model). However, I hope that the reader will appreciate the scope of models this opens up to be moulded by the practioner for his various uses. All he needs do is specify how he feels the distribution of

$$\theta_t|y^t \text{ evolves in the distribution of } \theta_{t+1}|y^t$$

and then use Bayes rule. The true Bayesian should not flinch at such a proposal.

The myth of additive error models is widespread but there is no reason for the Bayesian to restrict himself to just the considerations of these. Many difficulties arise from such models which are easily side stepped. For example in the continuous time case the additive model will restrict the user to stable distributions for his underlying level unless he works in specifications just using the first two movements which are notorious (for example see Chapter 2) for giving misleading estimates in non-normal situations.

A methodology must be *simple* and *flexible* enough to give models to the practicioner he can use on situations he is commonly confronted with. For this reason the Bayesian (or for that matter prior likelihood) framework has a real advantage over the conventional approach. I will continue my research into the many specific types of model arising from these ideas on the completion of this thesis.

<u>Summary</u>

A generalisation of the Normal Steady model has been given across all distributions in a well argued way. Many examples were presented. A brief statement of how to generalise the procedure to non-steady models was then given.

LIST OF REFERENCES

ANDERSON T.W. (1) "The Statistical Analysis of Time Series", Wiley 1971

BARON D.P. (1) "Point Estimation and Risk Preferences" J.A.S.A. Vol. 68, 1973

BARNARD G.A. (1) "The use of the likelihood function in statistical practice" Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol 1, 1967, p. 27-40.

BARTLETT M.S., and Kendall D.G. (1) "Statistical analysis of variance, heterogenerity and the logarithmic transform" J.R.S.S.B. Vol 8, 1948 p.128-38.

BECKER G.M., De GROOT M.H., MARSHAK, J. (1) "Stochastic models in choice behaviour" Behavioural Sci. 8 1963, p.41-55.

BILLINGSLEY P. (1) "Convergence of Probability Measure" McGraw-Hill 1972.

BLACKWELL D, and GIRSHICK M.A. (1) "Theory of Games and Statistical Decisions", Wiley, 1954.

BOX G.E.P. and JENKINS G.M. (1) "Time Series Analysis, Forecasting and Control" Holden-Day 1970

BRÖCKER, Th (1) "Differential Games and Catastrophes" Cambridge University Press, 1975.

BURILL C.W. (1) "Measure, Integration and Probability" McGraw-Hill, 1972.

CHAO M.T. (1) "The assymptotic behaviour of Bayes Estimators" Annals of Mathematical Statistics 1976 p.601-608.

CLEVENSON M.L. and ZIDEK J.V. (1) "Bayes Linear Estimations of the intensity function of the non-stationary Poisson Process" J.A.S.A. Vol 72 1977

DAVIDSON D, SUPPES P. and SIEGEL S (1) "Decision Making: An Experimental Approach" Stanford University Press 1957.

DAVID, A.P. (1) "Posterior means for large observations" Biometrika Vol 60, 1973 p. 664-6.

DE FINETTI, B. (1) "Theory of Probability Volume 1" Wiley 1974.

DE GROOT M.H. (1) "Optimal Statistical Decisions" McGraw-Hill 1970.

_____ and RAO M.M. (2) "Bayes estimation with convex loss" Annals of Mathematical Statistics 1963 p. 839-843.

DICKEY J.M. (1) "The weighted likelihood ratio, linear hypotheses on Normal location parameters" Annals of Mathematical Statistics Vol 42 p. 204-223.

_____ (2) "Scientific reporting and personal probabilities Students hypothesis" J.R.S.S.B. Vol 35 p. 285-305.

DREW, G.C., COLQUHOUN W.P. and LONG, H.A. (1) "Effects of small doses of alcohol and skill resembling driving" Medical Research Council Memo Vol 38 1959

EDWARDS A.W.F. (1) "Likelihood" Cambridge University Press, 1972

FELLER W. (1) "An introduction to Probability Theory and Its Applications, Volume 2" 2nd Edition Wiley 1971

FERGUSON T.S. (1) "A Bayesian Analysis of some non-parametric problems" Annals of Mathematical Statistics Volume 1, 1973 p.1-32.

FESTINGER, L. (1) "A theory of cognitive dissonance" Stanford University Press 1962.

_____ (2) "Conflict, decision and dissance" Stanford University Press, 1964.

HARRISON P.J., and STEVENS C.F. (1) "Bayesian Forecasting (with Discussion)" J.R.S.S.B. Vol. 38, 1976.

IBRAGIMOV I.A. (1) "On the composition of Unimodal Distributions" Theory of Probability and Its Applications Vol. 1, 1956.

ISNARD C.A. and ZEEMAN E.C. (1) "Some models from catastrophe theory in the Social Sciences" Edinburgh Conference 1972 Warwick University Work Paper.

JAMES W. and STEIN C. (1) "Estimation with Quadratic loss" Proceedings of the 4th Berkeley Symposium Volume 1, p.361-79.

KADANE J.B. and CHUANG D.T. (1) "Stable Decision Problems" Technical Report Carnegie-Melon University 1977.

KALMAN R.E. (1) "New methods in Wiener filtering theory" Proceedings of the 1st Symposium of Engineering Applications of Random Function Theory and Probability. Wiley 1963.

LEONARD,T., and HARRISON P.J. (1) "Bayesian updating for two-stage dynamic models" 1978 (Submitted to Technometrics)

LEONARD, T. (2) "A modification of the Bayes estimate for the mean of a normal distribution" Biometrika Vol.61 1974 p.627-8.

_____ (3) "A Bayesian approach to Linear models with unequal variances" Technometrics Vol 17, 1975

_____ (4) "Density Estimation, Stochastic Processes and Prior Information (with discussion) J.R.S.S.B. 1978 (to appear)

LINDLEY,D.V. (1) "A class of utility functions" Annals of Statistics Vol 4, p. 1-10, 1976.

LINDLEY, D.V.(2) "The use of prior probability distributions in statistical inference and decisions" Proceedings of the 4th Berkeley Symposium in Mathematical Statistics and Probability Vol 1, p.453-468.

LINDLEY, D.V. and SMITH,A.M.F. (1) "Bayes estimates for the linear model" J.R.S.S.B. Vol 34, 1972, p.1-41.

MAACHI, O. and PICINBONE B.C. (1) "Estimation and Detection of Weak Optical Signals" Institute of Electrical and Electronic Engineers Transactions on Information Theory IT-18 1975 p.562-73½

MATHER J.N. (1) "Stability of C^{∞} mappings I: The division theorem" Annals of Mathematics Vol 87, 1968 p.89 - 104.

MORAN P.A.P. (1) "An Introduction to Probability Theory" Claredon Press, Oxford, 1968.

MUTH J.F. (1) "Optimal Properties of Exponentially weighted Forecasts" Journal of the American Statistical Society Vol 55, 1960, p.299-306.

POSTON T. and STEWART I.N. (1) "Taylor expansions and Catastrophes" Pitman, 1976.

RAIFFA, H. (1) "Risk, ambiguity and the Savage axioms;Canwent" Quarterly Journal of Economics 1975, p.690-694.

SAVAGE L.J. (1) "The foundations of Statistics" Wiley, 1954.

SNYDER D.L. (1) "Random Point Processes" Wiley, 1976.

TROTMAN D.J.A. and ZEEMAN E.C. (1) "The classification of elementary catastrophes of codimension ≤5". Warwick University Working Paper, 1974.

THOM R. (1) "Structural Stability and Morphogenesis" Benjamin, 1975.

WALL, A. (1) "Statistical Decision Functions" Wiley 1950.

WHITTLE, P. (1) "Prediction and Regulation by Linear Least-Squares Methods" English Universities Press 1963.

WILKINSON, G.N. (1) "On resolving the controversy in Statistical Inference" J.R.S.S.B. Vol 39 1977, p. 119-72.

WOODCOCK A.E.R. and POSTON T. (1) "A Geometrical Study of Elementary Catastrophes" Springer-Verlag, 1974.

ZEEMAN E.C. (1) "Differential equations for the heartbeat and nerve impulse" Towards a theoretical biology 4" Ed. C.H. Waddington English Unviersities Press 1971 p.8-67.

(2) "Primary and Secondary waves in developmental biology" - Lectures in Maths in the Life Sciences 7. American Mathematical Society, Providence 1974.

(3) "On the unstable behaviour of stock exchanges" Journal of Mathematical Economics Vol. 1, 1974. 39-

(4) "Levels of Structure in Catastrophe Theory". International Congress of Mathematicians, Vancouver 1974.

(5) "A mathematical model for conflicting judgement caus by stress, applied to possible misestimation of speed caused by alcohol" Warwick University Working Paper, 1975.