Prior Influence in Bayesian Statistics

by

Michael Lavine
University of Minnesota
Technical Report No. 504
November, 1987

Research supported by National Institute of Health Grant No. GM 2527

PRIOR INFLUENCE IN BAYESIAN STATISTICS

I. INTRODUCTION

Objective

The work described here is motivated by a position that was well expressed by Berger (1986):

> "The robust Bayesian position can be roughly stated as follows: An answer to a statistical problem is a good answer only if ... the answer would approximately equal the posterior Bayes answer for any reasonable sampling model and prior distribution...."

But how can we tell whether "the answer would approximately equal the posterior Bayes answer for any reasonable sampling model and prior distribution: This report provides a guide by doing two things: describing classes of sampling models and prior distributions that are useful surrogates for the class of all "reasonable sampling model(s) and prior distribution(s)", and showing how to compute the resulting classes of posterior Bayes answers to particular statistical problems.

We will be calculating Bayes answers of the form $\int\phi(P)\text{posterior}(dP)$, or equivalently, $E_\pi[\phi(P)|\underset{\sim}{X}]$, where P is a possible sampling model for the data $\underset{\sim}{X}$, ϕ is a real-valued functional of P, posterior is the usual Bayes posterior measure and $E_\pi[.]$ means expectation using π as the prior. Four common examples are $\phi(P)=P(S)$ where S is a set, $\phi(P)=\int X dP$, $\phi(P)=1_B(P)$ where 1_B is the indicator function and $\phi(P)=L(P,a)$ where a is an action and L is a loss function. These make $\int\phi(P)\text{posterior}(dP)$ equal to the predictive probability that the next observation lies in the set S, the predictive mean, the posterior probability that P lies in the set B and the posterior expected loss of a, respectively. Berger (1987)

calls these quantities "ratio-linear" because they are the ratios $\int\phi(P)f(\underset{\sim}{X}|P)\pi(dP)/\int f(\underset{\sim}{X}|P)\pi(dP)$ of linear functionals of the prior π where $f(\underset{\sim}{X}|.)$ is the likelihood function.

Because we are conditioning on $\underset{\sim}{X}$ and treating it as known conclusions about the set of possible posterior answers apply only to the particular data set we are using. It can easily turn out that a single class of sampling models and priors will lead to either small or large sets of posterior answers depending on the data that were observed.

Section 2 introduces density-bounded classes of priors for parametric families and shows how to compute the resultant suprema and infema of ratior-linear posterior quantities. Section 3 gives a variation of density-bounded classes. Section 4 uses density bounds to define classes of periparametric priors and compute the ranges of ratio-linear answers. Section 5 discusses some issues raised in Section 4 while Section 6 shows how to extend the periparametric results to a regression setting. Finally, Section 7 presents a technical theorem relating periparametric classes to the Prohorov metric.

2. DENSITY BOUNDED CLASSES OF PRIORS

Introduction to the Class

This chapter defines Γ, a class of prior distributions for a parametric family, and shows how to compute psup = $\sup_{\pi\in\Gamma}\{E_\pi[\phi(\theta)|\underset{\sim}{X}]\}$ where θ is a parameter value indexing the sampling model P, ϕ is a real-valued function and $\underset{\sim}{X}$ is the observed data. Berger and Berliner (1986), Berger and Sellke (1987), Sivaganesan and Berger (1987), DeRobertis and Hartigan (1981), and others have done similar things for different classes of priors.

Let $\{P_\theta:\theta\in\Theta\}$ be a parametric family of distributions all having a density with respect to the same underlying measure and let $f(\underset{\sim}{X}|\theta)$ denote the joint density of the data $\underset{\sim}{X}$ in the usual fashion. For measures L and U on Θ, we say $L\le U$ if $L(B)\le U(B)$ for all measurable $B\subset\Theta$. Let $L\le U$ and $L(\Theta)<1<U(\Theta)<\infty$. Define Γ, a density-bounded class of probability measures by $\Gamma = \{\pi:L\le\pi\le U;\pi(\Theta)=1\}$.

Sometimes we will want to use an upper boundary U that has infinite mass. That will usually not pose any problem. DeRobertis and Hartigan (1981) use the related class of measures $\{\pi:L\le\pi\le U\}$ where $\pi(\Theta)$ need not be 1.

The class Γ is called density-bounded because it is often more natural to define Γ by bounds on densities. Without loss of generality, let L and U have densities ℓ and u with respect to some measure ν. Then $\Gamma=\{\pi:\ell(\theta)\le p(\theta)\le u(\theta)\ \nu\text{ a.s.};\ \int p(\theta)\nu(d\theta)=1\}$ where p is the density of π with respect to ν. In almost all applications we can take ν to be Lebesgue measure.

We will be using Γ to represent uncertainty in the prior distributions. In any particular problem we try to choose L and U so that Γ contains almost all of the plausible or reasonable priors and almost none of the implausible or unreasonable ones.

L, U and π are defined as measures on the parameter space Θ, but are equivalent to measures on the set of distributions $\{P_\theta:\theta\in\Theta\}$ where Θ is identified with $\{P_\theta:\theta\in\Theta\}$ and θ with P_θ. We use whichever notation seems convenient; i.e., we use $E[.|\theta]$ and $E[.|P_\theta]$ or $\phi(\theta)$ and $\phi(P_\theta)$ interchangeably.

One density-bounded class of priors is $\Gamma=\{\pi:\epsilon\pi_0\leq\pi\leq(1/\epsilon)\pi_0:\pi(\Theta)=1\}$ where π_0 is a fixed prior and $\epsilon\in[0,1]$ is a fixed scalar. However, this class only contains priors with the same type of tail behavior as π_0. We often want L to have smaller tails and U to have larger tails than π_0.

Density-bounded classes of priors are special cases of ϵ-contamination classes. An ϵ-contamination class is a set of priors $\{(1-\epsilon)\pi_0 + \epsilon q|q\in Q\}$ where π_0 is a fixed prior, $\epsilon\in[0,1]$ is a fixed scalar and Q is a class of allowable contaminations. A lower bound L is not a prior because $L(\Theta)<1$. It can be written as $(1-\epsilon)\cdot(L/L(\Theta))$ where $(L/L(\Theta))$ is a prior and $\epsilon=(1-L(\Theta))$. Any prior that falls between L and U can be written as $(1-\epsilon)\cdot(L/L(\Theta)) + \epsilon q$ where q is a prior from some allowable class that is determined by L and U. It is not true that every ϵ-contamination class is a density-bounded class.

Computing sup and inf of $E_\pi[\phi(\theta)|\underset{\sim}{X}]$

The set $\Gamma^* = \{E_\pi[\phi(\theta)|\underset{\sim}{X}] : \pi \in \Gamma\}$ of all possible posterior expectations of ϕ is an interval. To see this suppose that c_1 and c_2 are in Γ^* and $c_i = E_{\pi_i}[\phi(\theta)|\underset{\sim}{X}]$ where $\pi_i \in \Gamma$.

Define $\pi_\epsilon = (\epsilon\pi_1 + (1-\epsilon)\pi_2)$. Then $\pi_\epsilon \in \Gamma$ for any $\epsilon \in [0,1]$ and $E_{\pi_\epsilon}[\phi(\theta)|\underset{\sim}{X}]$ is a continuous function of ϵ. So $[c_1, c_2] \in \Gamma^*$.

We can characterize Γ^* by its endpoints. Define

$$\text{psup} = \sup_{\pi \in \Gamma} E_\pi[\phi(\theta)|\underset{\sim}{X}]$$

$$\text{and pinf} = \inf_{\pi \in \Gamma} E_\pi[\phi(\theta)|\underset{\sim}{X}].$$

We show how to compute psup; the technique for finding pinf is similar. Usually we do not find psup directly but employ an algorithm that estimates psup as accurately as desired.

The algorithm is based on being able to test, for any $q \in [0,1]$, whether psup is less than q. The test works by finding $\pi_q \in \Gamma$ such that

$$\text{psup} < q \quad \Leftrightarrow \quad E_{\pi_q}[\phi(\theta)|\underset{\sim}{X}] < q.$$

If we can find such a π_q then the following algorithm estimates psup with accuracy "tolerance":

```
lowbound = 0                /* bounds on psup */
highbound = 1

while ( highbound - lowbound    >   tolerance )
{
     q = ( highbound + lowbound ) / 2
     find π_q
     ptemp = E_{π_q} [φ(θ)|X]
                    /* ptemp is storing the expectation */

     lowbound = max ( lowbound, ptemp )
     if ( ptemp < q )   highbound = q
}

print ( highbound + lowbound ) / 2     /* final estimate of psup */
```

If this algorithm is implemented on a computer there will be unavoidable imprecision in computing ptemp. It may be worthwhile to estimate the error and change two statements of the algorithm to "lowbound = max (lowbound, ptemp-err)" and "if (ptemp+err < q) highbound=q."

Before showing how and why the algorithm works we give some motivation for it and discuss the ideas behind finding π_q with the right posterior expectation of ϕ.

For a fixed $q \in [0,1]$ we want to construct $\pi_q \in \Gamma$ such that ptemp $= E_{\pi_q}[\phi(\theta)|\underset{\sim}{x}]$ satisfies ptemp $< q \iff \text{psup} < q$.

By definition

$q \le pinf \Rightarrow ptemp \ge q$ and

$q > psup \Rightarrow ptemp < q$ regardless of how we choose π_q.

If we arrange that $q \in (pinf, psup] \Rightarrow ptemp \ge q$

then $psup < q \Leftrightarrow ptemp < q$.

Of course, we do not know whether $q \in (pinf, psup]$. However, if we assume $q \in (pinf, psup]$ and construct π_q accordingly then $psup < q \Leftrightarrow ptemp < q$ even if the assumption is wrong.

Ignoring the endpoint, assume $q \in (pinf, psup)$. There exists $\pi \in \Gamma$ such that $E_\pi[\phi(\theta)|\underset{\sim}{X}] = q$. The idea behind finding π_q is to start with this π and move mass around trying to increase the posterior expectation of ϕ. Because $q < psup$ we should be able to achieve that goal. When using the algorithm we find π_q directly without first finding π.

The posterior expectation of ϕ is the weighted average of $\phi(\theta)$ where each θ is given its posterior weight. It may seem obvious that to increase the posterior expectation we should move prior mass to θ's with large values of $\phi(\theta)$. However, this is not always true. A θ where $\phi(\theta)$ is large may have a small value of $f(\underset{\sim}{X}|\theta)$ and hence receive little posterior weight. Increasing the prior weight on that θ would not help much to increase the posterior expectation of ϕ.

Example 2.1:

Let X_1 and X_2 be Bernoulli random variables that are independent given the Bernoulli parameter θ. Let the prior π be defined by $\pi(.1) = .8$,

$\pi(.8)=\pi(.9)=.1$. What is $Pr_\pi[X_2=1|X_1=0]$? Begin by computing the posterior distribution.

$Pr_\pi[X_1=0] = (.9)(.8)+(.2)(.1)+(.1)(.1) = .75.$

$Pr_\pi[\theta=.1|X_1=0] = (.8)(.9)/.75 = .96.$

$Pr_\pi[\theta=.8|X_1=0] = (.1)(.2)/.75 \doteq .027.$

$Pr_\pi[\theta=.9|X_1=0] = (.1)(.1)/.75 \doteq .013.$

Therefore, $Pr_\pi[X_2=1|X_1=0] \doteq (.1)(.96) + (.8)(.027) + (.9)(.013) \doteq .1293.$

Create the prior π' by moving mass to the right. Say π' is defined explicitly by $\pi'(.1)=.8$, $\pi'(.9)=.2$. What is $Pr_{\pi'}[X_2=1|X_1=0]$? Again, start by computing the posterior.

$Pr_{\pi'}[X_1=0] = (.9)(.8)+(.1)(.2) = .74.$

$Pr_{\pi'}[\theta=.1|X_1=0] = (.8)(.9)/.74 \doteq .973.$

$Pr_{\pi'}[\theta=.9|X_1=0] = (.2)(.1)/.74 \doteq .027.$

Therefore, $Pr_{\pi'}[X_2=1|X_1=0] \doteq (.1)(.973) + (.9)(.027) \doteq .1216.$

In the example $\phi(\theta)=Pr[X_2=1|\theta]=\theta$. Moving weight to the right, from $\theta=.8$ to $\theta=.9$, (to large $\phi(\theta)$) decreased the predictive probability of a 1 on the next observation. The reason is that $\theta=.9$ has a small likelihood so the extra prior mass on $\theta=.9$ is heavily discounted in the posterior, thereby increasing the posterior mass on $\theta=.1$. The effect is to decrease the predictive probability that $X_2=1$.

We need a compromise between putting prior mass on θ's with large values of $\phi(\theta)$ and putting prior mass on θ's where $\phi(\theta)$ may be somewhat smaller but where the likelihood $f(\underset{\sim}{X}|\theta)$ is larger. The following theorem shows how to make the compromise.

Theorem 2.1:

Let $q \in (\text{pinf}, \text{psup})$ and let $h(\theta) = (\phi(\theta)-q) \cdot f(\underset{\sim}{X}|\theta)$. For a scalar z define $A_z=\{\theta:h(\theta)<z\}$, $B_z=\{\theta:h(\theta)=z\}$ and $C_z=\{\theta:h(\theta)>z\}$. If π_q is any prior in Γ satisfying $\pi_q(A_z)=L(A_z)$ and $\pi_q(C_z)=U(C_z)$ for some z then

$$E_{\pi_q}[\phi(\underset{\sim}{\theta})|X] \geq q.$$

Before proving the theorem we show that the conclusion is not vacuous, i.e., that there is some $\pi_q \in \Gamma$ that satisfies the conditions of the theorem for some z. Define the functions $\underline{g}(y) = L(A_y)+L(B_y)+U(C_y)$ and $\bar{g}(y) = L(A_y)+U(B_y)+U(C_y)$. Let $z=\inf\{y:\underline{g}(y)<1\}$. If $\underline{g}(z)\leq 1$ and $\bar{g}(z)\geq 1$ then it is clear that the required π_q exists, although it may not be unique.
$\underline{g}(z+1/k) = \underline{g}(z) - (U-L)\{\theta:h(\theta)\in(z,z+1/k]\}$. Take the limit as $k\to\infty$. $\lim \underline{g}(z+1/k) = \underline{g}(z)$ because (U-L) is continuous. But $\underline{g}(z+1/k)\leq 1$ by definition of z so $\underline{g}(z)\leq 1$. Similarly, $\underline{g}(z-1/k) = \bar{g}(z) + (U-L)\{\theta:h(\theta)\in(z-1/k,z)\}$ and $\lim \underline{g}(z-1/k) = \bar{g}(z)$ so that $\bar{g}(z)\geq 1$.

Proof of Theorem 2.1

Because $q \in (\text{pinf}, \text{psup})$ there exists a $\pi \in \Gamma$ such that $E_\pi[\phi(\theta)|\underset{\sim}{X}]=q$. Let p be the density of π and p_q be the density of π_q. Let $S=\{\theta:p(\theta)>p_q(\theta)\}$ and $T=\{\theta:p(\theta)<p_q(\theta)\}$. Because $\pi_q(A_z) = L(A_z)$ and $\pi_q(C_z) = U(C_z)$ SC_z and TA_z are both empty.

$$E_{\pi_q}[\phi(\theta)|X] \geq q$$

$$\Leftrightarrow \quad \int \phi(\theta) f(\underset{\sim}{X}|\theta)\pi_q(d\theta) \geq q\int f(\underset{\sim}{X}|\theta)\pi_q(d\theta)$$

$$\Leftrightarrow \quad \int \phi(\theta) f(\underset{\sim}{X}|\theta)\pi(d\theta) + \int \phi(\theta) f(\underset{\sim}{X}|\theta)(\pi_q-\pi)(d\theta)$$
$$\geq \quad q\int f(\underset{\sim}{X}|\theta)\pi(d\theta) + q\int f(\underset{\sim}{X}|\theta)(\pi_q-\pi)(d\theta)$$

$$\Leftrightarrow \quad \int \phi(\theta) f(\underset{\sim}{X}|\theta)(\pi_q-\pi)(d\theta) \geq q\int f(\underset{\sim}{X}|\theta)(\pi_q-\pi)(d\theta)$$

$$\Leftrightarrow \quad \int h(\theta)(\pi_q-\pi)(d\theta) \geq 0$$

$$\Leftrightarrow \quad \int_T h(\theta)(\pi_q - \pi)(d\theta)$$
$$\geq \quad \int_S h(\theta)(\pi - \pi_q)(d\theta).$$

But,

$$\int_T h(\theta)(\pi_q - \pi)(d\theta)$$
$$\geq \; z\int_T (\pi_q - \pi)(d\theta) \; = \; z\int_S (\pi - \pi_q)(d\theta)$$
$$\geq \; \int_S h(\theta)(\pi - \pi_q)(d\theta). \qquad \text{QED}$$

The third "$\Leftrightarrow$" follows because $\int\phi(\theta)f(\underset{\sim}{X}|\theta)\pi(d\theta) = q\int f(\underset{\sim}{X}|\theta)\pi(d\theta)$. The equality in the next to last line follows because π and π_q must both integrate to one so $\int_T(\pi - \pi_q) = \int_S(\pi_q - \pi)$.

Notice that the theorem is nothing more than the usual argument about balancing masses on a seesaw. Imagine the line as a seesaw balanced at the point q as in Figure 2.1. Each $\theta\in\Theta$ occupies a point on the seesaw corresponding to $\phi(\theta)$. The weight of each θ is its posterior weight so $q=\int\phi(\theta)\text{posterior}(d\theta)$. Let $\phi(\theta_1)$ and $\phi(\theta_2)$ both be greater than q, so θ_1 and θ_2 are on the right hand arm of the seesaw. Consider moving a small amount of prior mass from θ_1 to θ_2. Will this make the right hand side of the seesaw go up or down? Equivalently, will this decrease or increase the posterior expectation of ϕ? We know that the right hand side of the seesaw will go up if $(\phi(\theta_2)-q)(\Delta\text{posterior}(\theta_2))$ is less than $(\phi(\theta_1)-q)(\Delta\text{posterior}(\theta_1))$. Since $\Delta\text{posterior}(\theta_i)$ is approximately proportional to $(\Delta\text{prior}(\theta_i))(f(\underset{\sim}{X}|\theta_i))$ and Δprior must be the same for θ_1 and θ_2 we need only look at $h(\theta)$, as the theorem tells us.

Theorem 2.1 shows how to implement the step "find π_q" in the algorithm. Let q be given and define $h(\theta)$ as in the theorem. We can find z by a procedure such as

```
high_z  =  (some big number)    /* maximum possible value of h */
low_z   =  (some small number)  /* minimum possible value of h */
do
{
   z  =  (high_z + low_z)/2
   if  ( g(z) > 1 )    low_z  = z
   if  ( ḡ(z) < 1 )    high_z  = z
} until ( g(z)≤1 & ḡ(z)≥1 )
```

Then we define π_q as in the theorem. The conclusion is that $q \in (pinf, psup)$ implies $E_{\pi_q}[\phi(\theta)|\underset{\sim}{x}] \geq q$, and that in turn implies that $psup < q \Leftrightarrow E_{\pi_q}[\phi(\theta)|\underset{\sim}{x}] < q$. An example shows how the technique works in practice.

Example 2.2:

Let $X_1,\ldots,X_n,X_{n+1}$ be conditionally i.i.d. Bernoulli θ random variables. We observe $\underset{\sim}{X} = X_1,\ldots,X_n$ and want to compute the predictive probability that X_{n+1} is equal to 1, that is, $Pr[X_{n+1}=1|\underset{\sim}{X}]$. Let π_0 be the uniform prior and fix $\epsilon\in(0,1)$. Let $L=\epsilon\pi_0$ and $U=(1/\epsilon)\pi_0$.

Let s be the number of successes and f the number of failures in $\underset{\sim}{X}$. psup ($=\sup Pr[X_{n+1}=1|\underset{\sim}{X}]$) is a function of ϵ, s and f. For $\epsilon\in\{1,.9,.8,\ldots,.1\}$ and $s,f\in\{0,5,10\}$ the algorithm generated the results in Table 2.1. Each curve in Figure 2.2 is a plot of psup as a function of ϵ. The left hand set of curves is for s=0 the middle set is for s=5 successes and the right hand set is for s=10. Within each set the top curve is for f=0, the middle curve is for f=5 and the bottom curve is for s=10. Going across the page shows the effect of increasing s. Going down the page shows the effect of increasing f.

The results for ϵ=1 are exactly what would have been obtained by an ordinary Bayesian analysis with a uniform prior. As ϵ decreases the class of priors increases so psup increases.

For Example 2.2 h is a function that decreases from 0, reaches a minimum, increases through 0 to a positive maximum and decreases again to 0, as in Figure 2.3. Therefore π_q will have a special form. The unit interval will be partitioned into three sections. Either π_q will be equal to U on the outer sections and equal to L in the middle or else π_q will be equal to L on the outer sections and equal to U in the middle. Figure 2.3 illustrates this for p_q, the density of π_q.

One way to use the results of this section is to specify beforehand an L and U that capture our uncertainty about what prior to use. Then we compute the corresponding values of psup and pinf, which tell us something about our posterior state of ignorance concerning future observations.

But it may be difficult to decide in advance on unique satisfactory bounds on the prior measure. In that case we can look at the pair (pinf,psup) as a function of L and U. We may observe that for all reasonable choices of L and U the pair (psup,pinf) lies in a small region and that (psup-pinf) is small. Then we can be confident in stating our predictions for future values.

On the other hand (psup,pinf) may cover a large area or (psup-pinf) may be large for reasonable choices of L and U. Then we would know that our predictions can vary quite a bit over classes of reasonable priors.

Examples 2.3 and 2.4 indicate some problems that can be solved by the algorithm of this section.

Example 2.3:

Let $X_1,\ldots,X_{n+1}$ be conditionally iid $N(\theta,1)$ and let $\underset{\sim}{X} = (X_1, \ldots, X_n)$. Let π_0 be a prior for θ. Fix $\epsilon\in(0,1)$ and let $L=\epsilon\pi_0$ and $U=(1/\epsilon)\pi_0$. Find the sup and inf of $Pr_\pi[\theta\in B|\underset{\sim}{X}]$, $Pr_\pi[X_{n+1}\in S|\underset{\sim}{X}]$ and $E_\pi[X_{n+1}|\underset{\sim}{X}]$ over all priors π bounded by L and U.

Example 2.4:

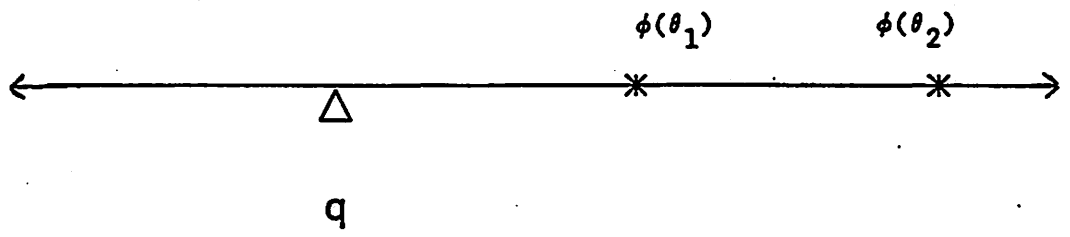
Take the previous example but let L be the 0 measure and U be proportional to Lebesgue measure.

It is easy to find sup $E_\pi[X_{n+1}|\underset{\sim}{X}]$ for this last example. For any real number k we can assign prior probability 1 to a set of θ's satisfying $E[X_{n+1}|\theta]>k$. The posterior will assign probability 1 to the same set of θ's so the predictive mean will be greater than k. Therefore sup $E_\pi[X_{n+1}|\underset{\sim}{X}] = \infty$.

TABLE 2.1

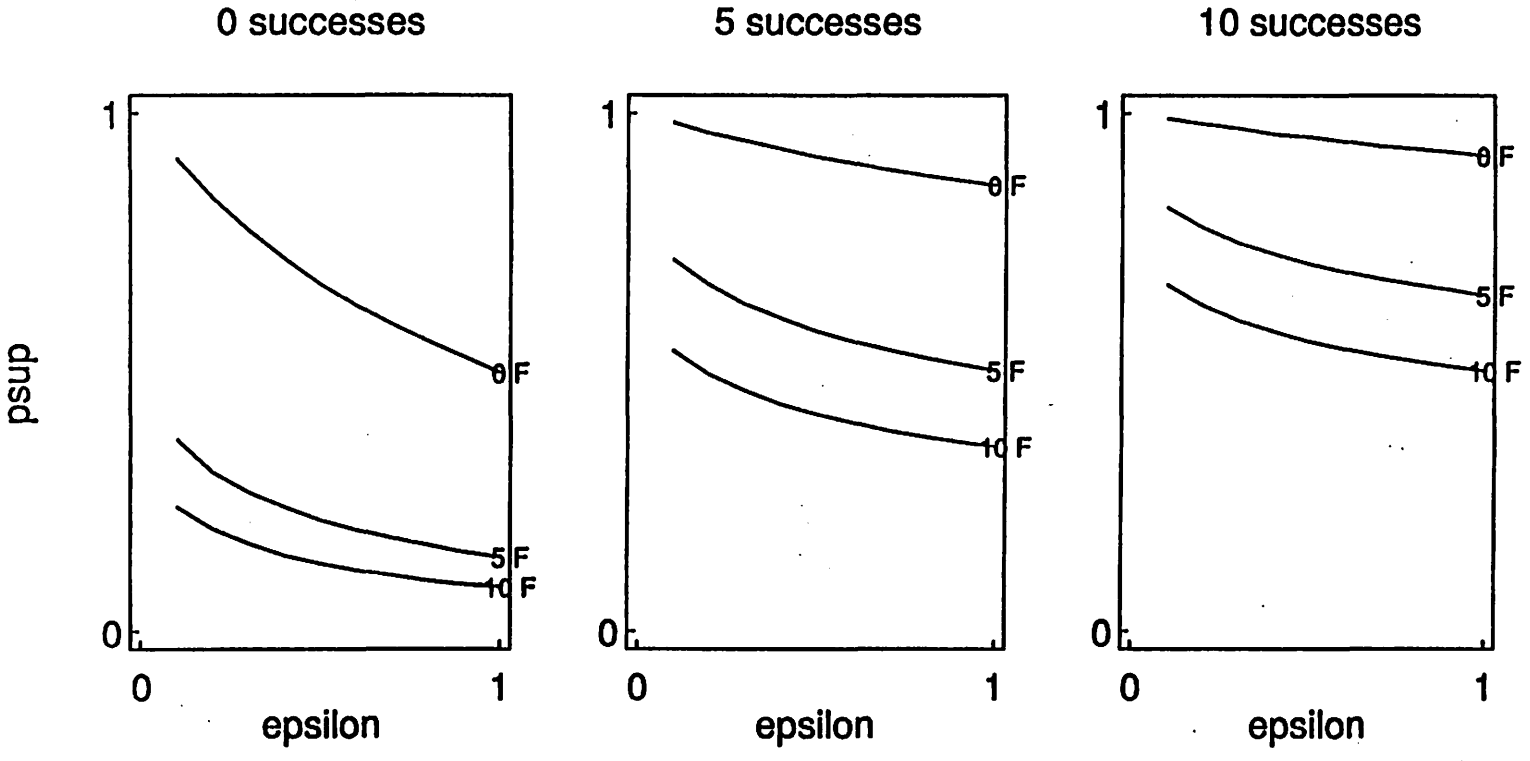
Psup as a function of the number of successes, the number of failures and the bounds on the prior determined by ϵ. See Example 2.2.

		0 successes	5 successes	10 successes
	ϵ= 1	.500	.862	.919
	.9	.530	.871	.927
	.8	.560	.881	.933
	.7	.593	.891	.938
0	.6	.628	.905	.947
	.5	.669	.917	.956
failures	.4	.719	.932	.960
	.3	.773	.948	.971
	.2	.836	.962	.980
	.1	.912	.982	.990
	ϵ= 1	.146	.504	.650
	.9	.156	.516	.661
	.8	.169	.529	.671
	.7	.182	.543	.682
5	.6	.198	.560	.695
	.5	.216	.580	.710
failures	.4	.241	.605	.730
	.3	.270	.632	.751
	.2	.309	.670	.780
	.1	.371	.718	.819
	ϵ= 1	.087	.358	.504
	.9	.093	.367	.513
	.8	.100	.377	.522
	.7	.110	.389	.533
10	.6	.120	.405	.546
	.5	.133	.421	.560
failures	.4	.149	.440	.580
	.3	.171	.466	.601
	.2	.200	.497	.631
	.1	.242	.542	.670



Pictorial representation of Theorem 2.1

FIGURE 2.1



psup as a function of the number of successes, the number of failures and epsilon. Within each set of three curves the top one is for 0 failures, the middle one for 5 and the bottom one for 10 failures.

FIGURE 2.2

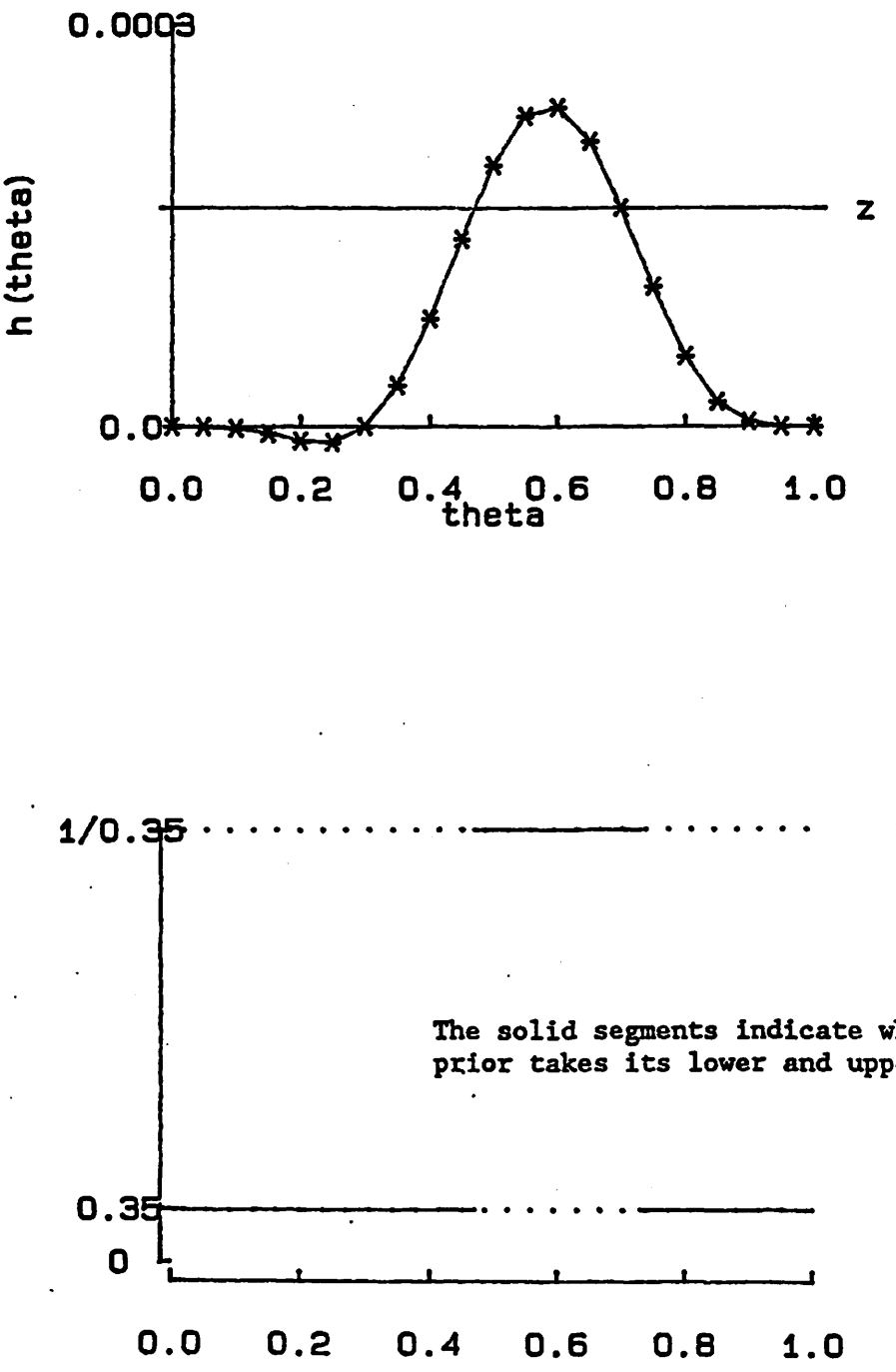


Illustration of the algorithm

FIGURE 2.3

3. A VARIATION OF DENSITY-BOUNDED CLASSES

Definitions and Examples

This section gives a variation of density-bounded classes of priors. Let θ be a multidimensional parameter. For ease of exposition take $\theta=(\theta_1,\theta_2)^t$ although the ideas work for arbitrary dimension. Let $\Theta=\Theta_1\times\Theta_2$ where $\theta_i\in\Theta_i$. Sometimes the prior distribution π has a natural decomposition into a marginal distribution π^m for θ_1 and a conditional distribution $\pi^c(\cdot|\theta_1)$ for θ_2 given θ_1. In such a case density-bounded classes for both π^m and π^c may be a natural way to represent uncertainty in the prior.

Let L and U be measures on Θ_1 such that $L\le U$ and $L(\Theta_1)<1<U(\Theta_1)<\infty$. For each $\theta_1\in\Theta_1$ let $L(\cdot|\theta_1)$ and $U(\cdot|\theta_1)$ be measures on Θ_2 such that $L(\cdot|\theta_1)\le U(\cdot|\theta_1)$ and $L(\Theta_2|\theta_1)<1<U(\Theta_2|\theta_1)<\infty$. Define the class of priors Γ by

$$\Gamma = \{\pi: L\le\pi^m\le U;\ L(\cdot|\theta_1)\le\pi^c(\cdot|\theta_1)\le U(\cdot|\theta_1)\ U\ \text{a.s.}\}$$

Example 3.1:
Given θ and σ let $X_1,\ldots,X_n$ be i.i.d. $N(\theta,\sigma^2)$. Let π_0 be a prior distribution for θ and σ defined by a marginal distribution π_0^m for σ and a conditional distribution π_0^c for θ given σ, say π_0^m=gamma(a,b) and $\pi_0^c=N(0,\sigma^2)$. Fix δ and ϵ in (0,1). Let Γ be the class of priors π satisfying $\delta\pi_0^m\le\pi^m\le(1/\delta)\pi_0^m$ and $\epsilon\pi_0^c(\cdot|\sigma)\le\pi^c(\cdot|\sigma)\le(1/\epsilon)\pi_0^c(\cdot|\sigma)$.

Example 3.2:
I know that in a recent campus election approximately 1000 votes were cast for my favorite candidate. I do not know n, the total number of votes cast, or θ, the fraction of the votes favoring my

range of priors for n and a range of conditional priors for θ given n, where the prior for θ given n would be centered near 1000/n.

Example 3.3:
Given θ and ϵ let $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ be independent Bernoulli random variables where each X_i has parameter $(\theta+\epsilon)$ and each Y_i has parameter $(\theta-\epsilon)$. A model like this might arise when there are two ways of administering a treatment. θ represents the average effect of the treatment and 2ϵ represents the difference in effect between the two ways of administering the treatment. We may want to use the class Γ determined by bounds on the marginal distribution of θ and the conditional distribution of ϵ given θ.

Non Example 3.1:
Another way of modeling the previous example leads to a class of priors not covered by the techniques of this chapter. Again let $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ be conditionally independent Bernoulli random variables. Let each X_i have parameter $(\theta+\epsilon_1)$ and each Y_i have parameter $(\theta+\epsilon_2)$. Let Γ^* be a class of priors determined by a density-bounded class for the marginal distribution of θ, density-bounded classes for the conditional distribution of ϵ_i given θ and in which ϵ_1 and e_2 are i.i.d. given θ. In some ways Γ^* is similar to the class described in Example 3.3. The crucial difference is that posterior expectations require integrating twice with respect to the conditional distribution of ϵ_i given θ. This report does not show how to compute psup in such a case.

The following two examples show that the classes of priors described in this section are neither special cases nor generalizations of density-bounded classes.

Example 3.4:
Let X and Y be Bernoulli random variables. Consider the class Γ of distributions given by the following set of marginal distributions for X and conditional distributions for Y given X.

$$\Pr[X=0] = 1-\Pr[X=1] \in [.25,.75].$$
$$\Pr[Y=0|X=0] = 1-\Pr[Y=1|X=0] \in [0,.25].$$
$$\Pr[Y=0|X=1] = 1-\Pr[Y=1|X=1] \in [.75,1].$$

For this class of distributions $0\le\Pr[0,0]\le 3/16$, $3/16\le\Pr[0,1]\le 3/4$, $3/16\le\Pr[1,0]\le 3/4$ and $0\le\Pr[1,1]\le 3/16$ and no tighter bounds are possible. However, not all distributions lying inside the bounds are members of the class Γ. The joint distribution $\Pr[0,0]=\Pr[0,1]=\Pr[1,1]=3/16$, $\Pr[1,0]=7/16$ lies within the bounds but is not in Γ, therefore Γ is not a density-bounded class.

Example 3.5:

Let X and Y be Bernoulli random variables. Consider the density-bounded class Γ of distributions that give no more than probability 1/2 to any of the four points (0,0), (0,1), (1,0) and (1,1). For this class the marginal distribution for X satisfies $0\le\Pr[X=0]\le 1$, the conditional distribution for Y given X satisfies $0\le\Pr[Y=0|X]\le 1$ and no tighter bounds are possible. But these are no restrictions at all. Consequently, the class Γ cannot be described by bounds on the marginal distribution for X and the conditional distribution for Y given X.

Computing psup

We now show how to compute psup for Γ's that are defined by bounds on π^m and π^c. The technique for pinf is similar. As in the previous section we test whether psup$\ge$q by finding $\pi_q\in\Gamma$ such that

$$\text{psup} \ge q \quad \Leftrightarrow \quad E_{\pi_q}[\phi(\theta)|\underset{\sim}{X}] \ge q.$$ And, as before, we only need show

$$q\in(\text{pinf},\text{psup}) \quad \Rightarrow \quad E_{\pi_q}[\phi(\theta)|\underset{\sim}{X}] \ge q.$$

We start defining π_q by defining $\pi_q^c(\cdot|\theta_1)$, the conditional distribution of θ_2 given θ_1. Let

$$h(\theta) = (\phi(\theta)-q) \cdot f(\underset{\sim}{X}|(\theta)).$$

For a fixed θ_1 h is a function of θ_2. Take $\pi_q^c(\cdot|\theta_1)$ to be the distribution that puts as much weight as possible on θ_2's where $h(\theta)$ is large, analogously to the definition of π_q in Theorem 2.1. For each θ_1 there will be a number $z(\theta_1)$ and sets

$A_{z(\theta_1)} = \{\theta_2 : h(\theta) < z(\theta_1)\}$

$B_{z(\theta_1)} = \{\theta_2 : h(\theta) = z(\theta_1)\}$

$C_{z(\theta_1)} = \{\theta_2 : h(\theta) > z(\theta_1)\}$ such that

$\pi_q^c(A_{z(\theta_1)}|\theta_1) = L(A_{z(\theta_1)}|\theta_1)$ and $\pi_q^c(C_{z(\theta_1)}|\theta_1) = U(C_{z(\theta_1)}|\theta_1)$.

Define in this way $\pi_q^c(\cdot|\theta_1)$ for every $\theta_1 \in \Theta_1$.
Let $h^m(\theta_1) = \int h(\theta)\pi_q^c(d\theta_2|\theta_1)$. Treat h^m exactly as h in Theorem 2.1. There will be a number z and sets $A_z=\{\theta_1 : h^m(\theta_1)<z\}$, $B_z=\{\theta_1 : h^m(\theta_1)=z\}$ and $C_z=\{\theta_1 : h^m(\theta_1)>z\}$ such that π_q^m will be a prior satisfying $\pi_q^m(A_z)=L(A_z)$ and $\pi_q^m(C_z)= U(C_z)$.

There is at least one such π_q determined by π_q^c and π_q^m in this way. The following theorem is the analog of Theorem 2.1 and shows that ϕ has the required conditional expectation under the prior π_q.

Theorem 3.1:
Let $\pi \in \Gamma$ satisfy $E_\pi[\phi(\theta)|\underset{\sim}{X}] = q$. Such a π exists by an argument similar to the one in Chapter 2.
Let π_q be defined as above.

Then $E_{\pi_q}[\phi(\theta)|X] \geq q$.

Proof:

The proof is in two parts. Define an intermediate prior μ by $\mu^m = \pi^m$ and $\mu^c(\cdot|\theta_1) = \pi_q^c(\cdot|\theta_1)$. We first prove $E_\mu[\phi(\theta)|\underset{\sim}{X}] \geq q$ and then show $E_{\pi_q}[\phi(\theta)|\underset{\sim}{X}] \geq q$.

Part 1.

$E_\mu[\phi(\theta)|\underset{\sim}{X}] \geq q$

$$\Leftrightarrow \quad \int [\, \int \phi(\theta) f(\underset{\sim}{X}|\theta) \pi_q^c(d\theta_2|\theta_1) \,] \, \pi^m(d\theta_1)$$

$$\geq \quad \int [\, q \int f(\underset{\sim}{X}|\theta) \pi_q^c(d\theta_2|\theta_1) \,] \, \pi^m(d\theta_1)$$

Theorem 2.1 says that the quantity in square brackets on the left-hand side is greater than or equal to the quantity in square brackets on the right-hand side for every value of θ_1. Therefore $E_\mu[\phi(\theta)|\underset{\sim}{X}] \geq q$.

Part 2.

$E_{\pi_q}[\phi(\theta)|\underset{\sim}{X}] \geq q$

$$\Leftrightarrow \quad \int \int \phi(\theta) f(\underset{\sim}{X}|\theta) \pi_q^c(d\theta_2|\theta_1) \, \pi_q^m(d\theta_1)$$

$$\geq \quad q \int \int f(\underset{\sim}{X}|\theta) \pi_q^c(d\theta_2|\theta_1) \, \pi_q^m(d\theta_1)$$

$$\Leftrightarrow \quad \int \int \phi(\theta) f(\underset{\sim}{X}|\theta) \pi_q^c(d\theta_2|\theta_1) \, \pi^m(d\theta_1)$$

$$+ \int \int \phi(\theta) f(\underset{\sim}{X}|\theta) \pi_q^c(d\theta_2|\theta_1) \, (\pi_q^m - \pi^m)(d\theta_1)$$

$$\geq \quad q \int \int f(\underset{\sim}{X}|\theta) \pi_q^c(d\theta_2|\theta_1) \, \pi^m(d\theta_1)$$

$$+ q \int \int f(\underset{\sim}{X}|\theta) \pi_q^c(d\theta_2|\theta_1) \, (\pi_q^m - \pi^m)(d\theta_1)$$

We know from Part 1 that the first term on the left is at least as large as the first term on the right. And the second term on the left is at least as large as the second term on the right by the argument in Theorem 2.1.

QED.

For some multidimensional parametric families it seems natural to specify the prior by giving a marginal prior for the first parameter and then a sequence of conditional priors for the rest of the parameters. In these cases it may be most natural to specify a class of priors by giving lower and upper bounds for the marginal and conditional priors.

In other situations it may be natural to specify a prior in which the parameters are independent. For these cases the marginal prior for θ_j is the same as the conditional prior for θ_j given $\theta_1,\ldots,\theta_{j-1}$. We could specify a class of priors either by giving bounds for the conditional priors or by giving bounds for the joint prior of all the parameters. We can use whichever method best captures our uncertainty about the prior and then compute psup and pinf using the techniques of this section or the previous one.

This concludes our discussion of parametric models. In reality we only believe parametric models when the data are multinomial but often use them when we believe the data follow a distribution that is close to some known parametric family. The next two sections show how to compute psup for models that include distributions that are close to, but not members of, a given parametric family.

4. PERIPARAMETRIC MODELS

"peri-prefix...1:all around:about:round...2:near...3a:enclosing or surrounding", Webster's Third New International Dictionary of the English Language Unabridged, G. and C. Merriam Company, Publishers, Springfield, MA, 1976.

Point of View

Let the real-valued random variables $X_1, X_2, \ldots, X_n = \underset{\sim}{X}$ be independent observations from the same sampling distribution. Let Ω be the set of all possible sampling distributions for X_1. A typical parametric Bayesian analysis identifies points in a parameter space Θ, usually a subset of some Euclidean space, say R^k, with points in a subset of Ω, say $\{P_\theta: \theta \in \Theta\}$. The prior is a probability measure on Θ and is equivalent to a probability measure on Ω that gives probability one to the subset.

This report takes the point of view that the distribution on Ω is the fundamental object, not the distribution on the parameter space. Henceforth, the terms "prior" and "posterior" refer to probability measures on Ω, not Θ. This point of view is mentioned explicitly by Lindley (1972) and is implicit in the work of Ferguson (1973). Because Θ is identified with a subset of Ω it may seem overly nice to call the measure on Ω more fundamental than the measure on Θ. But it is both correct and useful, as explained below.

Consider the parametric family of densities $f(x|\theta)=\theta\exp(-\theta x)dx$ for $\theta\in(0,\infty)$ and the prior density $p_1(\theta)=\exp(-\theta)d\theta$. Each θ, a real number, has been identified with P_θ, an element of Ω. In this case $\theta = 1/\int xP_\theta(dx)$. Another parameterization of the same set of densities is $f(x|\beta)=\beta^{-1}\exp(-x/\beta)d\beta$. Now the parameterization is $\beta = \int xP_\theta(dx)$.

The change of variables $\beta=1/\theta$ gives the prior density $p_2(\beta)=\beta^{-2}\exp(-1/\beta)d\beta$.

Most statisticians would agree that the first parameterization and prior are equivalent to the second parameterization and prior. $\Theta=(0,\infty)$ is the same in each case but the densities p_1 and p_2 are different. The two situations are equivalent because they describe the same distribution on Ω. The distribution on Ω is more fundamental than the distribution on Θ.

A typical Bayesian analysis might call for the computation of the posterior mean of the parameter. But the posterior mean has different interpretations in the two parametrizations. In one case it is $E\{1/\int xP(dx)\}$. In the other it is $E\{\int xP(dx)\}$. Whether either of these is useful in a real data problem depends on that problem and can only be determined by thinking about Ω.

The elements of Ω are distributions, so the term "posterior mean" should refer to the average of those distributions. It is another distribution, another element of Ω and need not correspond to any parameter value. This interpretation of the posterior mean is usually called the predictive distribution.

Periparametric Models

Priors on Ω that give probability 1 to a parametric subset are usually implausible. This section describes a class of priors that put their mass on a subset of Ω that is close to, surrounds and encloses a parametric family. We call both the subset of Ω and the class of priors "periparametric." Computing psup and pinf for this class accounts explicitly for deviations from the parametric family.

Generic elements of Ω will be denoted by capital letters like P and Q, possibly with subscripts. Θ will be a parameter space identified with $\{P_\theta : \theta \in \Theta\} \subset \Omega$. The notation P_θ and Q_θ means that these elements are associated with the parameter value θ. Usually P_θ will be the element of Ω identified with θ and Q_θ will be a point optimizing some function in a neighborhood of P_θ.

Periparametric classes of priors can be useful when we believe that the data are distributed approximately as some parametric distribution or when a parametric prior approximately describes our a priori beliefs.

Example 4.1:

Let X_1 and X_2 be random variables that are conditionally independent given their common distribution. Suppose we believe that the underlying distribution is close to exponential and also that the relative likelihood that the distribution is close to exponential(θ_i) rather than exponential(θ_j) is approximately $\exp(\theta_j - \theta_i)$. The standard parametric model given by the two formulae $f(x|\theta) = \theta e^{-\theta x} dx$ and $p(\theta) = e^{-\theta} d\theta$ approximately represents these a priori opinions. (It is a wonderful circumstance that our opinions are computationally convenient.) We know how to compute $\Pr[X_2 \in S | X_1]$ using the standard model. But we want to know how the result will change if we account for uncertainty about both $f(x|\theta)$ and $p(\theta)$.

Here Ω would be the set of all probability measures on the positive reals. The standard formulae are equivalent to a prior probability measure π_0 on Ω. The problem is to find Γ, a class of priors, that are all close to π_0 in some appropriate sense and to compute psup and pinf over the class Γ. Figure 4.1 shows Ω, Θ and a shaded region that is the set of all distributions that are approximately exponential or close to exponential, in some sense to be defined later. This section describes classes of priors that put all their mass on the shaded region. The

next section shows how to use priors that may put some mass outside of the shaded region.

Let P_θ be the probability measure with density $\theta e^{-\theta x}$ on the positive reals and $N(\theta) \subset \Omega$ be the set of probability measures that are close to P_θ, or for which P_θ is a good approximation, in this still undefined notion of closeness. To connote the idea of closeness $N(\theta)$ is called a neighborhood, even though it has no topological significance.

Let π be a prior created from π_0 by spreading each mass or density element $\pi_0(P_\theta)$ throughout $N(\theta)$. This means $\pi(N(\theta)) \geq \pi_0(P_\theta)$, which has a sensible interpretation even when both sides are zero:

$$\Pr_\pi[\ \bigcup_{\theta \in B} N(\theta)\ ] \geq \Pr_{\pi_0}[\ B\ ] \text{ for every measurable } B \subset \Omega.$$

If the neighborhoods are disjoint then the previous relationships become equalities. Often, our vague a priori notions do not distinguish very well between π_0 and π because P and P_θ are very similar for every $P \in N(\theta)$. Therefore we want to include such priors π in the class Γ of plausible priors.

Because π is a probability measure on Ω it tells us how to pick a random $P \in \Omega$. For priors that put all their mass in $\cup_\theta(N(\theta))$ we can think of picking P's as a two step process: first choose $P_\theta \in \Theta$ and then $P \in N(\theta)$. The distribution of such a process can be described by the marginal distribution of P_θ and the conditional distribution of P given P_θ. To allow for uncertainty in both parts of the prior we use a class of priors determined by a class of marginal distributions for P_θ and a class of conditional distributions for P given P_θ.

There are three parts to describing a periparametric class of priors Γ - defining $N(\theta)$ for each $\theta\in\Theta$, giving the class of marginal distributions for P_θ and giving the class of conditional distribution for P given P_θ. This section gives one way to define each part. The next section discusses modifications and alternatives to each of these parts that lead to different and sometimes more appropriate classes.

Density bounds provide one way to define the neighborhoods $N(\theta)$. Each P_θ is a probability measure on the sample space χ. Let L_θ and U_θ be two measures on χ satisfying$L_\theta\leq P_\theta\leq U_\theta$ and $L_\theta(\chi)<1<U_\theta(\chi)<\infty$. The top part of Figure 4.1 shows Ω, Θ, θ and $N(\theta)$. The bottom part shows the density $f(x|P_\theta)$ and the two curves ℓ_θ and u_θ that are the densities of L_θ and U_θ. We define $N(\theta)$ to be the set of all $P\in\Omega$ bounded between L_θ and U_θ.

Example 4.1 continued:
In the previous example we thought the distribution of the X's was close to exponential. Let Θ index the set of exponential distributions so that $f(x|P_\theta)=\theta e^{-\theta x}$. Fix $\epsilon\in(0,1)$ and let $\ell_\theta(x)=\epsilon f(x|P_\theta)$ and $u_\theta(x)=(1/\epsilon)f(x|P_\theta)$. Let $N(\theta)$ consist of all distributions on the positive reals with densities between ℓ_θ and u_θ. Figure 4.1 pictures such a neighborhood.

Of course this neighborhood contains some densities that may seem implausible, such as the discontinuous ones. A later section will discuss that problem.

The second part of describing Γ is defining a class of marginal distributions for P_θ. But sections 2 and 3 defined classes of distributions for parametric families. We can use those same classes here. Or we can use any class of priors over which we can maximize and

minimize $E[\phi(\theta)|\underset{\sim}{X}]$. For specificity let the class of π^m's be a density-bounded class determined by an L and a U.

The last part of describing Γ is defining a class of conditional distributions for P given P_θ. It is hard to think about distributions on $N(\theta)$ when there is no parametric representation to help us. So we adopt the solution of allowing any conditional distribution whatsoever satisfying $Pr[P\in N(\theta)|P_\theta]=1$ for almost all P_θ. If, in some applied situation, we can think clearly enough to specify a different set of conditional distributions then we should use that set. But it is often too difficult to think about distributions on such complicated sets as $N(\theta)$.

Example 4.1 continued:
Now we can completely describe a Γ for the previous example. Use the $N(\theta)$ neighborhoods described there. Fix $\delta\in(0,1)$. Let $f_0(\theta)=e^{-\theta}$. Let $\ell=\delta f_0$ and $u=(1/\delta)f_0$. Use the set of P_θ distributions having densities bounded between ℓ and u. Use the set of conditional distributions for P given P_θ such that $Pr[P\in N(\theta)|P_\theta]=1$. This completely describes Γ. Now the goal is to compute psup and pinf for this class.

A Γ described by the three parts above may contain some prior distributions that seem unreasonable. In particular, both the P_θ-distribution and P may have discontinuous densities. However, it may be difficult to specify a more reasonable class that is both large enough and tractable. We can proceed by computing psup and pinf and seeing whether the range of posterior answers is large or small. If it is small then it doesn't matter that Γ contained some unreasonable priors. If the range is large then we can try to see which priors in Γ

give answers that are close to psup and pinf. If those priors are reasonable then again we don't worry about the unreasonable ones. But if it is the unreasonable priors that cause the range to be large then we can try to make a smaller Γ and recompute psup and pinf.

It is difficult to make all the decisions necessary to describe Γ completely. We must supply L, U and L_θ and U_θ for each θ. One approach is to compute psup and pinf for several choices of L, U, L_θ and U_θ and see which choices lead to large ranges of posterior answers. Without deciding precisely which choices are reasonable we may be able to decide that no reasonable choices lead to large ranges of posterior answers. Then we needn't worry about which choice we make. Or, we may see that some reasonable choices do give large ranges. Then we must conclude that we really don't know much about $E[\phi(P)|\underset{\sim}{X}]$.

Computing psup

To find psup and pinf we use the same algorithm as before. We start by proving that every value in (pinf,psup) is attainable as a posterior answer.

Theorem 4.1:
Let $q \in (\text{pinf},\text{psup})$. Then there exists a prior measure $\pi\in\Gamma$ such that $E_\pi[\phi(\theta)|\underset{\sim}{X}]=q$.

Proof:
Since $q\in(\text{pinf},\text{psup})$ there exist priors π_1 and π_2, both in Γ, such that

$$E_{\pi_1}[\phi(\theta)|\underset{\sim}{X}] < q < E_{\pi_2}[\phi(\theta)|\underset{\sim}{X}].$$

Let π_1^m and π_2^m denote the corresponding P_θ-distributions and π_1^c and π_2^c denote the corresponding conditional distributions for P given P_θ. Any prior $\mu_{\alpha,\beta}$ with P_θ-distribution $(1-\alpha)\pi_1^m+\alpha\pi_2^m$ and conditional distribution for P $(1-\beta)\pi_1^c+\beta\pi_2^c$ for $\alpha,\beta\in[0,1]$ is also in Γ.

$E_{\mu_{\alpha,\beta}}[\phi(\theta)|\underset{\sim}{X}]$ is a continuous function of α and β. QED

Next we define π_q for every $q\in[0,1]$. As always π_q is supposed to put as much mass as possible on P's where $h(P)=(\phi(P)-q)f(\underset{\sim}{X}|P)$ is large. Since π_q is a prior in Γ it can be described by its P_θ-distribution π_q^m and its conditional distribution π_q^c for P given P_θ. It is easiest to give the conditional distribution first. Assume for the moment that within each $N(\theta)$ there is a Q_θ that maximizes $h(P)$. That is, $h(Q_\theta) = \sup\{ h(P) : P\in N(\theta) \}$. Then $\pi_q^c(\cdot|P_\theta)$ is the measure that puts all its mass on Q_θ, i.e.

$$Pr_{\pi_q}[P=Q_\theta|P_\theta] = 1.$$

If there is more than one point in $N(\theta)$ that maximizes h then it makes no difference whether the conditional distribution assigns all its mass to one of them or spreads the mass around among all of them. The next section discusses the existence of Q_θ and what to do if there is no maximizing point. For now we assume that there is a maximizer within each neighborhood.

The last step in defining π_q is to give the P_θ-distribution π_q^m. As the conditional distribution of P given P_θ is degenerate at Q_θ and because we want to put prior mass where h(P) is large we require the P_θ-distribution to put as much mass as possible where the function $h^m(P_\theta) = h(Q_\theta)$ is large. So we proceed as in Section 3. The marginal

density for P_θ takes its lower bound when $h^m(P_\theta)<z$ and takes its upper bound when $h^m(P_\theta)>z$, for some appropriate value z.

To summarize, π_q is defined in two parts, the marginal distribution of P_θ and the conditional distribution of P given P_θ. The construction is similar to that in Section 3. The conditional distribution of P is a point mass on Q_θ, which maximizes h over the neighborhood $N(\theta)$. We define $h^m(P_\theta)$ to be $h(Q_\theta)$. This is similar to Section 3 where $h^m(\theta_1)$ was defined to be the integral of $h(\theta_1,\theta_2)^t$. Finally we take the P_θ-distribution to agree with either L or U according to whether $h^m(P_\theta)$ is less than or greater than z, where z is chosen to make π_q a proper distribution.

The only thing left to do in verifing the algorithm is to prove psup > q if and only if $E_{\pi_q}[\phi(P)|X] > q$.
The theorem and proof are similar to those in Section 3. P plays the role of θ_2 and P_θ plays the role of θ_1.

Theorem 4.2:

Let $q\in(pinf,psup)$ and π_q be defined as above.

Then $E_{\pi_q}[\phi(p)|\underset{\sim}{X}] \geq q$.

Proof:

Same as Theorem 3.1.

Example 4.1 continued:

We used the algorithm to perform some numeric computations. We computed $psup=sup_{\pi\in\Gamma}\{Pr_\pi[X_2>k_2|X_1=k_1]\}$ in the context of Example 4.1. as a function of k_1,k_2,δ and ϵ where δ and ϵ determine the class of marginal priors for P_θ and the size of $N(\theta)$. We performed the calculations for $k_1,k_2\in\{.5,1,2\}$ and $\delta,\epsilon\in\{1,.9,.8,\ldots,.1\}$.

The results for the example are given in Table 4.1 and plotted in Figure 4.2. Each plot in Figure 4.2 is for a different combination of k_1 and k_2. Each plot shows the contours of psup as a function of δ and ϵ for those fixed values of k_1 and k_2. The value of psup at the upper right corner of each plot, where $\delta=\epsilon=1$, is the value that would have been attained by an ordinary Bayesian analysis without allowing for uncertainty in the prior. As δ and ϵ decrease the class Γ increases so psup increases. When either δ or ϵ is equal to 0 there is enough freedom in the P_θ-distribution or the distribution of P given P_θ to make psup equal to 1. psup increases monotonically from the upper right corner to the left and lower sides of the plot.

The top side of the plot is where δ is fixed at 1 and ϵ is free to vary, so the P_θ-distribution is fixed and the conditional distribution for P can vary. The right side is where ϵ is fixed and δ varies. In every one of the nine plots the contour lines are fairly evenly spaced along the top but are bunched near the bottom of the right hand side. That means that a small change in the conditional distribution of P has a greater effect on psup than a small change in the P_θ-distribution. Changing the P_θ-distribution without changing the conditional distribution of P is the same as doing a standard Bayesian analysis of sensitivity to the prior where the likelihood is kept fixed. The results here indicate that small uncertainty in the likelihood is more important than small uncertainty in the distribution of the parameters of that likelihood, at least for Example 4.1.

TABLE 4.1

psup = sup Pr [$X_2 > k_2$ | $X_1 = k_1$] as a function of k_1, k_2, δ, ϵ.

δ controls the class of marginal priors for θ.

ϵ controls $N(\theta)$.

See Example 4.1.

$k_1 = .5$ $k_2 = .5$

ϵ \ δ	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.996	.986	.980	.966	.949	.929	.907	.883	.858	.832
.2	.994	.985	.970	.950	.927	.900	.870	.839	.805	.771
.3	.994	.979	.961	.937	.909	.877	.842	.805	.767	.728
.4	.992	.976	.953	.926	.893	.857	.818	.777	.736	.693
.5	.991	.973	.947	.916	.880	.839	.797	.753	.709	.664
.6	.990	.969	.941	.907	.867	.824	.779	.733	.685	.640
.7	.988	.966	.935	.899	.856	.810	.763	.713	.665	.617
.8	.987	.962	.931	.891	.846	.797	.747	.696	.646	.597
.9	.985	.959	.926	.884	.836	.786	.733	.680	.629	.579
1.0	.984	.958	.921	.877	.827	.775	.721	.666	.613	.563

$k_1 = .5$ $k_2 = 1$

ϵ \ δ	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.991	.992	.985	.974	.964	.949	.934	.916	.897	.877
.2	.990	.988	.977	.963	.946	.927	.905	.881	.855	.828
.3	.995	.983	.970	.953	.932	.908	.882	.853	.822	.791
.4	.994	.979	.964	.943	.920	.892	.862	.830	.796	.761
.5	.993	.979	.959	.935	.909	.878	.845	.809	.773	.735
.6	.992	.976	.954	.929	.898	.865	.829	.791	.752	.712
.7	.991	.973	.950	.921	.889	.854	.815	.774	.733	.691
.8	.990	.971	.946	.916	.881	.843	.802	.759	.716	.672
.9	.989	.969	.943	.910	.873	.833	.790	.745	.700	.655
1.0	.988	.967	.939	.905	.866	.824	.779	.733	.686	.640

TABLE 4.1 continued

k_1=.5 k_2=2

ϵ= \ δ=	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.995	.995	.989	.982	.976	.968	.958	.947	.935	.922
.2	.994	.991	.981	.974	.962	.951	.936	.921	.905	.886
.3	.993	.988	.979	.966	.952	.937	.919	.900	.880	.858
.4	.992	.984	.974	.958	.944	.925	.904	.882	.859	.833
.5	.992	.982	.970	.954	.935	.914	.891	.866	.840	.812
.6	.992	.979	.966	.949	.928	.905	.879	.852	.823	.793
.7	.992	.977	.963	.944	.921	.896	.869	.839	.809	.777
.8	.991	.979	.961	.940	.916	.889	.859	.828	.795	.761
.9	.991	.978	.960	.937	.910	.882	.851	.817	.783	.747
1.0	.991	.977	.957	.933	.906	.875	.843	.808	.772	.735

k_1=1 k_2=.5

ϵ= \ δ=	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.992	.978	.959	.934	.904	.872	.836	.800	.763	.726
.2	.984	.968	.939	.903	.863	.819	.774	.727	.681	.635
.3	.986	.958	.922	.879	.831	.780	.727	.674	.622	.574
.4	.982	.951	.908	.858	.803	.746	.688	.631	.577	.527
.5	.979	.943	.895	.840	.780	.717	.655	.595	.540	.487
.6	.979	.936	.883	.823	.758	.691	.627	.564	.507	.454
.7	.977	.930	.873	.808	.739	.669	.601	.537	.478	.425
.8	.974	.925	.863	.793	.721	.648	.578	.513	.454	.401
.9	.972	.920	.854	.781	.704	.629	.557	.490	.431	.380
1.0	.970	.915	.845	.768	.689	.611	.536	.469	.410	.360

TABLE 4.1 continued

$k_1=1$ $k_2=1$

$\epsilon=$ \ $\delta=$	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.994	.985	.971	.953	.931	.907	.880	.852	.822	.792
.2	.992	.974	.954	.929	.899	.865	.829	.791	.752	.714
.3	.990	.969	.942	.909	.872	.832	.789	.745	.700	.657
.4	.988	.963	.930	.892	.850	.803	.756	.707	.658	.611
.5	.986	.957	.920	.877	.829	.779	.727	.674	.622	.573
.6	.983	.953	.912	.864	.812	.757	.701	.645	.591	.541
.7	.981	.949	.904	.852	.795	.737	.678	.619	.564	.512
.8	.982	.944	.896	.840	.780	.718	.656	.596	.540	.487
.9	.981	.941	.890	.830	.767	.701	.637	.575	.517	.465
1.0	.979	.938	.883	.820	.754	.686	.619	.556	.497	.444

$k_1=1$ $k_2=2$

$\epsilon=$ \ $\delta=$	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.990	.990	.979	.969	.954	.940	.922	.904	.883	.861
.2	.995	.983	.969	.951	.930	.908	.883	.857	.829	.800
.3	.993	.979	.959	.937	.911	.883	.852	.820	.786	.752
.4	.992	.975	.952	.925	.894	.861	.826	.788	.750	.712
.5	.990	.971	.945	.915	.880	.843	.803	.762	.720	.678
.6	.988	.967	.940	.905	.867	.826	.783	.738	.693	.649
.7	.987	.964	.934	.897	.856	.811	.765	.717	.670	.623
.8	.986	.961	.929	.890	.846	.798	.749	.699	.649	.601
.9	.985	.959	.925	.883	.836	.786	.734	.681	.630	.581
1.0	.984	.958	.921	.877	.827	.775	.721	.666	.613	.563

TABLE 4.1 continued

$k_1=2$ $k_2=.5$

ϵ= \ δ=	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.985	.958	.922	.881	.836	.788	.740	.691	.642	.598
.2	.976	.936	.886	.830	.769	.708	.648	.588	.531	.479
.3	.969	.919	.857	.789	.718	.649	.581	.516	.458	.404
.4	.962	.904	.833	.757	.679	.602	.528	.462	.400	.348
.5	.958	.890	.812	.728	.643	.561	.484	.414	.355	.305
.6	.953	.879	.794	.704	.611	.523	.442	.374	.318	.271
.7	.949	.869	.777	.680	.580	.487	.406	.341	.286	.244
.8	.946	.859	.762	.656	.549	.454	.375	.312	.261	.220
.9	.942	.850	.747	.632	.520	.424	.347	.286	.239	.200
1.0	.938	.842	.731	.607	.492	.397	.322	.264	.219	.184

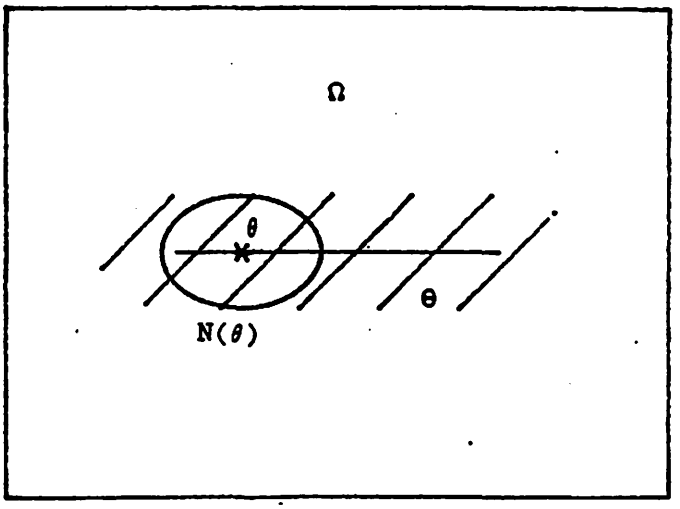
$k_1=2$ $k_2=1$

ϵ= \ δ=	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.983	.968	.942	.911	.876	.839	.799	.759	.718	.677
.2	.982	.950	.912	.868	.819	.769	.717	.666	.615	.567
.3	.977	.937	.887	.833	.776	.716	.656	.598	.543	.491
.4	.973	.924	.868	.805	.739	.672	.607	.545	.486	.433
.5	.968	.914	.850	.780	.707	.635	.565	.500	.440	.387
.6	.964	.905	.834	.757	.679	.602	.528	.460	.401	.349
.7	.960	.896	.819	.737	.654	.571	.494	.425	.367	.318
.8	.959	.888	.806	.719	.629	.541	.462	.395	.339	.292
.9	.956	.881	.793	.701	.605	.514	.435	.368	.314	.269
1.0	.953	.874	.782	.684	.582	.489	.410	.345	.293	.250

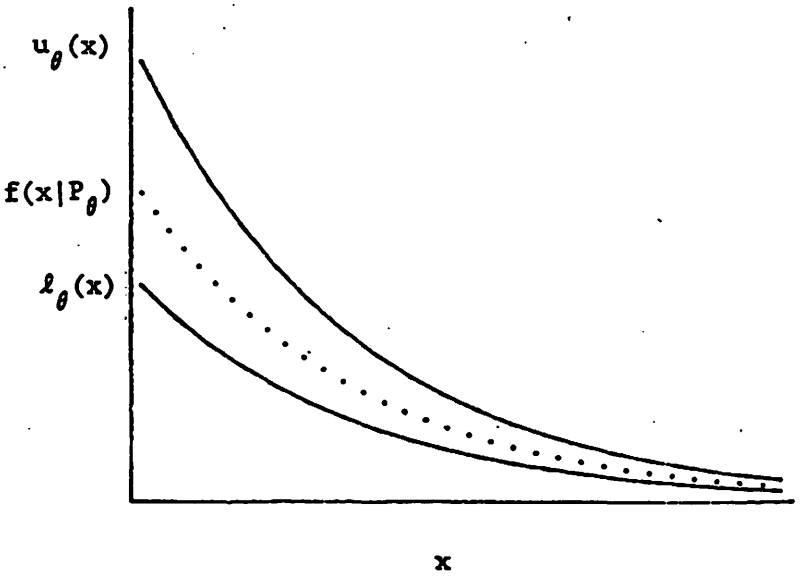
TABLE continued

k_1=2 k_2=2

ϵ= δ=	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.993	.978	.961	.939	.915	.889	.861	.832	.802	.770
.2	.989	.967	.939	.906	.871	.833	.794	.754	.713	.672
.3	.985	.956	.921	.880	.836	.790	.742	.694	.647	.601
.4	.981	.949	.906	.858	.807	.754	.700	.646	.595	.545
.5	.980	.942	.893	.840	.782	.723	.664	.607	.551	.500
.6	.978	.935	.882	.823	.760	.696	.633	.572	.515	.462
.7	.976	.930	.872	.807	.740	.672	.606	.542	.483	.431
.8	.974	.924	.862	.794	.722	.650	.581	.515	.456	.404
.9	.972	.920	.854	.780	.705	.630	.557	.491	.431	.380
1.0	.970	.915	.845	.768	.689	.611	.536	.469	.410	.360

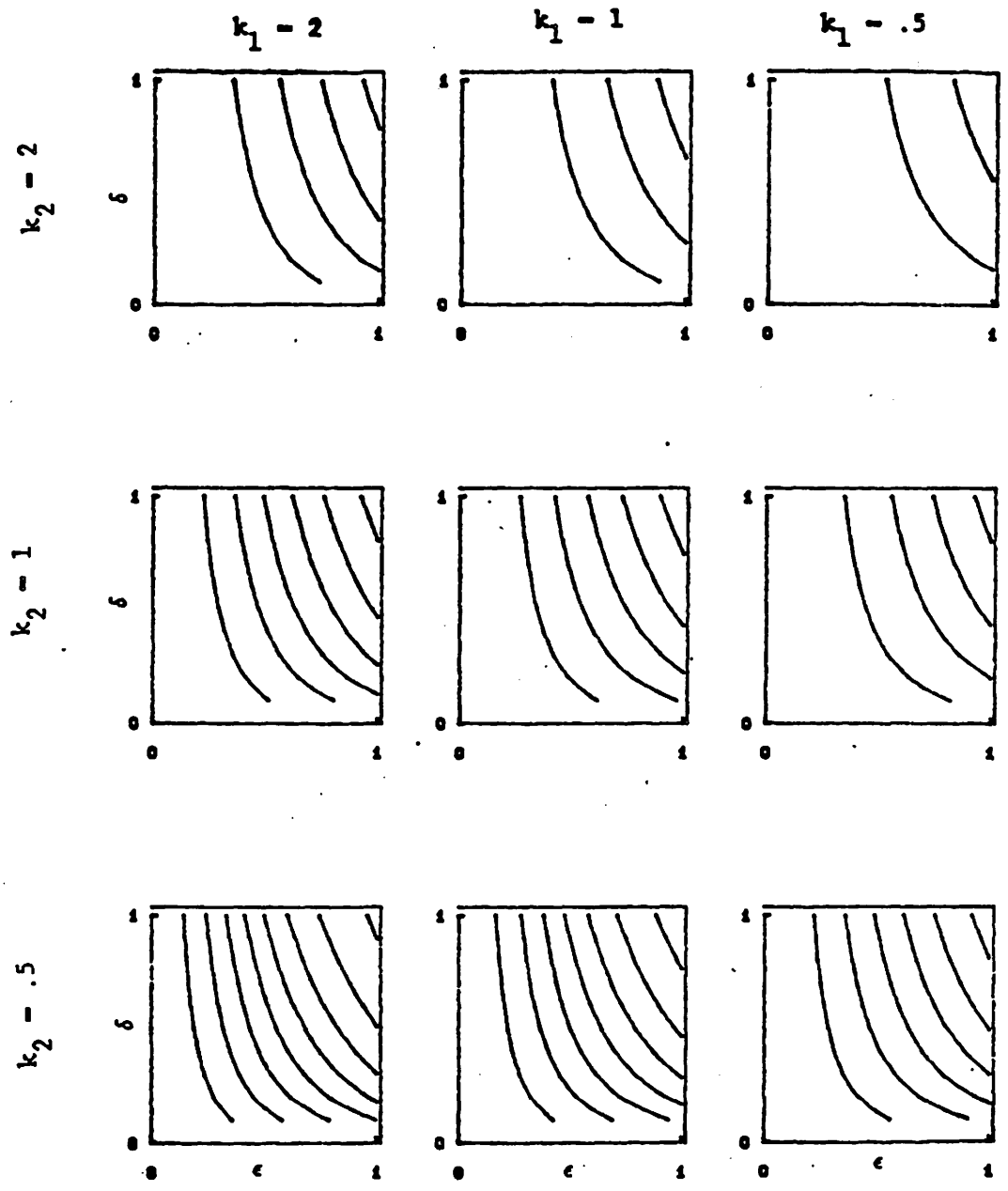


The set of all distributions



The neighborhood $N(\theta)$

FIGURE 4.1



Contour plots of psup = sup $Pr[X_2 > k_2 | X_1 = k_1]$ as a function of δ and ϵ where δ indexes the class of priors for P_θ and ϵ indexes the size of $N(\theta)$ See Example 4.1

FIGURE 4.2

5. MORE PERIPARAMETRIC MODELS

This section discusses some issues raised in the previous section. First is the question of Q_θ. Is there a Q_θ in each neighborhood that maximizes the function h? If so, what is it? If not, how does the algorithm work? Next we discuss alternatives for the three critical choices that determine M: defining $N(\theta)$, defining the set of conditional distributions for P given P_θ and defining the set of marginal distributions for P_θ. Lastly we talk about Prohorov metric neighborhoods as another way to get classes of priors.

Q_θ

Q_θ is supposed to maximize $h(P)=(\phi(P)-q)f(\underset{\sim}{X}|P)$ subject to the restriction that $L_\theta \le P \le U_\theta$. Apart from the restriction only two aspects of P matter, $\phi(P)$ and $f(\underset{\sim}{X}|P)$. It is almost true that we can choose Q_θ to optimize these two aspects of P independently of each other. For ease of exposition we assume every $P\in\Omega$ has a density with respect to Lebesgue measure and that $f(\underset{\sim}{X}|P)=f(X_1|P)$.

If we try to optimize the two parts separately then, because $f(\underset{\sim}{X}|P)$ is nonnegative, we should choose Q_θ to maximize $\phi(P)$. For most useful choices of ϕ this is possible. For example, when $\phi(P)=P(S)$, the P probability of the set S, then maximizing $\phi(P)$ means assigning as much probability as possible to the set S.

Let $\bar{S}$ be the complement of S. If $U_\theta(S)+L_\theta(\bar{S}) \le 1$ we take $Q_\theta(S)=U_\theta(S)$ and define Q_θ on the set $\bar{S}$ so that Q_θ is a probability measure. Otherwise, if $U_\theta(S)+L_\theta(\bar{S}) \ge 1$, we take $Q_\theta(\bar{S})=L_\theta(\bar{S})$

and define Q_θ on the set S so that Q_θ is a probability measure. In either case Q_θ maximizes P(S) subject to the restrictions.

After $Q_\theta(S)$ has been determined we can tell whether $(Q_\theta(S)-q)$ is positive or negative. Q_θ should either maximize or minimize $f(X_1|P)$ according to the sign of $(Q_\theta(S)-q)$. $f(X_1|P)$ is just the density of P evaluated at the observed value X_1. To maximize or minimize $f(X_1|P)$ Q_θ should have density either u_θ or ℓ_θ at X_1. That is $f(X_1|Q_\theta)=u_\theta(X_1)$ if $Q_\theta(S)>q$ and $f(X_1|Q_\theta)=\ell_\theta(X_1)$ if $Q_\theta<q$.

Once $Q_\theta(S)$ and $f(X_1|Q_\theta)$ have been determined any other features of Q_θ are irrelevant. We can extend the definition of Q_θ in any way at all that makes Q_θ a probability measure.

To summarize, Q_θ is determined by these three rules.

1) If $U_\theta(S)+L_\theta(\bar{S}) \leq 1$ then for every $T\subset S$ $Q_\theta(T) = U_\theta(T)$, else, for every $T\subset\bar{S}$, $Q_\theta(T) = L_\theta(T)$.
2) If $Q_\theta(S) \geq q$ then $f(X_1|Q_\theta) = u_\theta(X_1)$, else $f(X_1|Q_\theta) = \ell_\theta(X_1)$.
3) Extend rules 1 and 2 so that Q_θ is a probability measure and $Q_\theta\in\Gamma$.

There is a problem with defining Q_θ by these rules. Since densities are defined only up to sets of measure 0 rule 2 looks like it might be nonsense, or at least meaningless. In other words $f(X_1|P)$ is not well defined. There are two ways to resolve this. We can require that all the f(x|P) be continuous functions of x, or we can define f(x|P) to be the derivative of the c.d.f. and only consider those P for which the derivative exists.

Both of these solutions give meaning to rule 2 in a way that preserves the intuition about densities being infinitesimal probability masses. However, they can lead to another problem, namely that rule 2 can conflict with rule 1. For example, rule 1 might specify that Q_θ has density u_θ on S, but if $X_1 \in S$, $Q_\theta(S) < q$ and we use continuous densities then rule 2 might require $f(x|Q_\theta) \doteq \ell_\theta(X_1)$ in a neighborhood of X_1, contradicting rule 1.

In that case consider a sequence $\{Q_{\theta,n}\}$ in which $Q_{\theta,n}$ obeys rule 2 in a neighborhood about X_1 of size 1/n and otherwise obeys rule 1. Posterior and predictive probabilities of sets computed along this sequence will approach the posterior and predictive probabilities computed using the Q_θ with a density discontinuous at X_1, $f(X_1|Q_\theta) = \ell_\theta(X_1)$. Therefore we get the same psup regardless of whether the f(x|P) are required to be continuous. From now on we will use discontinuous densities and not worry about their uniqueness. We do require that the densities are bounded between ℓ_θ and u_θ for <u>all</u> x, not just for almost all x.

It might seem that the Q_θ's we use are, in some instances, unreasonable, should not be in the support of any reasonable prior, and therefore Γ is too big. It is not trivial to say exactly which distributions are reasonable. As we have mentioned, simply requiring continuity of f(x|P) does not change the value of psup. It may be that we want to bound the modulus of continuity, or make f(x|P) smoother in some other way. Such restrictions, while they may capture our sense of reasonableness better, can be harder to specify and work with. For the present we will continue to allow P's with discontinuous densities.

$N(\theta)$

Levy metric, L^r norm on either densities or cdf's and total variation norm are a few extensively studied metrics on Ω that could be used to define $N(\theta)$. Another possibility is to elicit quantiles from the expert and define $N(\theta)$ to be the set of probability measures with the specified quantiles. Why not use one of these methods to define neighborhoods?

One answer is that we can use these other definitions of $N(\theta)$. If we really think of closeness in Ω according to one of those definitions then computing the corresponding psup and pinf will tell us something useful. However, I contend, it is usually more natural, or at least useful, to think of closeness in Ω being determined by density bounds. An example will show some problems that arise when computing psup. For specificity, we take neighborhoods determined by Levy metric (Loeve (1977, p228)). The Levy distance between two cdf's is the size of the largest square that fits between them. This distance has the property that a sequence $\{F_n\}$ of cdf's converges to F in Levy metric if and only if $\{F_n\}$ converges weakly to F.

Example 4.1 continued:
Consider again Example 4.1 in which the X's have approximately an exponential distribution. Fix $\epsilon>0$ and define $N(\theta)$ to be the ϵ-neighborhood of P_θ in the Levy metric. Use the same class of marginal distributions for P_θ and conditional distributions for P given P_θ as before. Let $S=(1,\infty)$. Compute
$psup=sup_{\pi\in\Gamma}\{Pr_\pi[X_2\in S|X_1=x_1]\}$.

The solution to the example problem is psup=1 for every value of x_1. We will show that for every $k\in(0,1)$ there exists a prior $\pi\in\Gamma$ such that $Pr_\pi[X_2\in S|X_1=x_1] \geq k$. Within each $N(\theta)$ there exists a Q_θ such that $f(x_1|Q_\theta)=0$. Figure 5.1 shows P_θ and Q_θ. Take the conditional distribution of P given P_θ to be $P=P_\theta$ if $P_\theta(S)>k$ and $P=Q_\theta$ if $P_\theta(S)\leq k$. Because $f(X_1|P)>0$ only if $P(S)>k$ the posterior assigns probability 1 to distributions that put at least mass k on the set S. Therefore, the predictive probability of S is greater than k.

What has gone wrong here? One answer is that nothing is wrong. If we really think of closeness as being similar to weak convergence, or Levy metric, then we really don't know much about the predictive distribution of future observables. But this is an unsatisfactory answer; there is a problem with the definition of $N(\theta)$. P_θ and Q_θ have vastly different densities at some points, x_1 included. Q_θ can have density 0 on a set of positive P_θ measure. We get psup=1, and pinf=0, because $f(x_1|P_\theta)/f(x_1|Q_\theta)$ can be very different from 1.

One way of thinking about whether two densities are close is to ask whether you could tell them apart by observing data that was coming from one of them. In the case of P_θ and Q_θ it is easy to tell them apart when x_1 is observed. However, when $N(\theta)$ is defined by properly chosen density bounds the likelihood ratio cannot get too large or too small and it is much harder to distinguish between densities in $N(\theta)$. That is why we must at least consider neighborhoods determined by bounds on densities. We may want to have more restrictions as well, say by bounding the densities and the Levy metric. But density bounds must be considered.

Distribution of P given P_θ

After defining the neighborhoods $\{N(\theta):\theta\in\Theta\}$ the second critical point in describing Γ was to give a set of conditional distributions for P given P_θ. Section 4 allowed any conditional distribution satisfying $\Pr[P\in N(\theta)|P_\theta]=1$. We now consider other possibilities.

One reason to consider other distributions is that we may not be completely sure that the true sampling distribution is close to the parametric family. There may be a small probability, say α, that the sampling distribution lies far from the family. We can model this uncertainty by using a class of priors satisfying $\pi(\cup_{\theta\in\Theta}N(\theta))\geq 1-\alpha$. We give a brief description of how to find π_q for two sets of conditional distributions for P given P_θ that satisfy the previous inequality.

One model for the distribution of P is that after selecting P_θ we choose $P\in N(\theta)$ with probability at least $1-\alpha$. This means $\Pr[P\in N(\theta)|P_\theta]\geq 1-\alpha$ for all $\theta\in\Theta$. For this setup we find Q_θ as in section 4, and also find $Q\in\Omega$ maximizing h. That is, $h(Q)=\sup_{P\in\Omega}(h(P))$. Then the conditional distribution $\Pr[P=Q_\theta|P_\theta]=1-\alpha$ and $\Pr[P=Q|P_\theta]=\alpha$ puts as much weight as possible where h is large. If π_q has this conditional distribution for P given P_θ then a revised version of Theorem 4.2 holds and the algorithm works.

A second set of conditional distributions that give priors with $\pi(\cup_{\theta\in\Theta}N(\theta))\geq 1-\alpha$ is that in which for some θ P is in $N(\theta)$ with probability 1 but for other θ P is arbitrary. More formally, $\Pr[\Pr[P\in N(\theta)|P_\theta]=1]\geq 1-\alpha$. In this expression $\Pr[P\in N(\theta)|P_\theta]$ is random, it depends on P_θ. It is equal to 1 with probability at least $1-\alpha$.

For this set of conditional distributions we should take $\Pr[P=Q_\theta|P_\theta]=1$ if $h(Q_\theta)>z$ and $\Pr[P=Q|P_\theta]=1$ if $h(Q_\theta)<z$. This choice puts as much weight as possible where h is large. Again, a revised version of Theorem 4.2 obtains.

Another reason to consider different sets of conditional distributions is to control how mass is spread around the neighborhood of each P_θ. Instead of associating with each θ a single neighborhood $N(\theta)$ that gets conditional probability 1 we can construct an increasing sequence of neighborhoods $\{N_i(\theta)\}$ where the i-th neighborhood gets probability p_i. Then we would use the set of conditional probabilities satisfying $\Pr[P\in N_i(\theta)|P_\theta]=p_i$. To find π_q we would find $Q_{\theta,i}\in N_i(\theta)$ such that $h(Q_{\theta,i})=\sup\{h(P):P\in N_i(\theta)\}$ and let π_q have conditional distribution $\Pr[P=Q_{\theta,i}|P_\theta]=p_i-p_{i-1}$. Then we define $h^m(P_\theta)=\Sigma(p_i-p_{i-1})h(Q_{\theta,i})$ and take π_q to have marginal distribution for P_θ that puts as much weight as possible where h^m is large. This generalizes section 4 where $p_1=1$ and $h^m(P_\theta)=h(Q_\theta)$.

Distribution of P_θ

The third and final critical point in defining Γ was giving the set of marginal distributions for P_θ. Section 4 used density-bounded classes. One variation is to use the classes of distributions discussed in section 3. There θ was multidimensional, $\theta^t=(\theta_1,\ldots,\theta_k)$ say, and a class of priors was given by lower and upper bounds on all the conditional distributions $\pi^c(\theta_i|\theta_1,\ldots,\theta_{i-1})$. In this notation the θ_j's can be vector valued and have different dimensions. Section 3 explained how to choose π_q to optimize the posterior expectation of ϕ.

Another possibility is the DeRobertis-Hartigan (1981) class of priors, $\Gamma^m=\{\pi: L\leq\pi\leq U\}$ where π is any measure, not necessarily proper. If L has infinite mass then so will π. DeRobertis and Hartigan discuss an example in which both L and U are proportional to Lebesgue measure on the real line. They also solve the problem of maximizing and minimizing the posterior expectation of $g(\theta)$ over the class. If we define $g(\theta)=\phi(Q_\theta)$ then we can use Γ^m as the class of P_θ-distributions and find psup by the usual algorithm.

Several authors including Huber (1973), Sivaganesan (1986) and Sivaganesan and Berger (1987) give results on maximizing and minimizing the posterior expectation of $\phi(P)$ over ϵ-contamination classes. The results are for particular choices of the function ϕ and the class of allowable contaminations. The general rule is that whenever we can figure out what "putting as much prior mass as possible where h^m is large" means then we can apply the algorithm.

Example 5.1:
Consider the ϵ-contamination class $\Gamma^m=\{\pi = (1-\epsilon)\pi_0 + \epsilon\gamma : \gamma\in G\}$ where $\epsilon\in(0,1)$ is fixed, π_0 is a fixed prior and G is the class of all possible distributions on Θ. If there exists $\theta'\in\Theta$ such that $h^m(\theta')=\sup_{\theta\in\Theta}h(\theta)$ define γ' to be the distribution degenerate at θ'. Clearly $(1-\epsilon)\pi_0 + \epsilon\gamma'$ puts as much prior mass as possible where h^m is large. We can use the algorithm by proving a version of Theorem 4.2.

Example 5.2:
Berger (1987) mentions the quantile class of distributions $\{\tau : c_i\leq\int_{I(i)}\tau(d\theta)\leq d_i,\ i=1,\ldots,m\}$ where I(i) is the i-th element of a partition of Θ, and $c_i, d_i\in[0,1]$ are fixed bounds on the prior probability of I(i). We can find π_q as follows. Choose $\theta_i\in I_i$ to

maximize $h^m(\theta)$. Within I(i) assign mass c_i to θ_i. Then find j such that $h^m(\theta_j)=\max(h^m(\theta_i))$ and give θ_j all the remaining mass, but not more than d_j-c_j. Continue until all the mass is assigned. The resulting marginal prior for P_θ puts as much mass as possible where h^m is large, so we can prove the appropriate theorem and apply the algorithm.

Prohorov neighborhoods

We have described classes of priors on Ω by classes of P_θ-marginals and P-conditionals. A different class of priors is the set $\{\pi : d(\pi,\pi_0)\leq\epsilon\}$ where d is the Prohorov metric (Billingsley (1968)), π_0 is some fixed prior and ϵ is a fixed scalar in (0,1). $d(\pi,\pi_0)$ is defined to be the infimum of α's satisfying $\pi_0(B) \leq \pi(B^\alpha)+\alpha$ for all measurable $B\subset\Omega$ where B^α is the union of all open balls of radius α centered at a point in B. This definition requires a distance defined between members of Ω so that B^α is defined.

The Prohorov metric is appealing both for its interpretation of closeness of priors and because convergence in the Prohorov metric is equivalent to weak convergence (Billingsley (1968)). The interpretation for us is that $d(\pi,\pi_0)\leq\epsilon$ means that if π_0 puts mass c on the set B then π must put approximately the same amount of mass, at least $c-\epsilon$, on a nearby set, B^ϵ. There is a close relationship between Prohorov neighborhoods and the Γ's we have been studying.

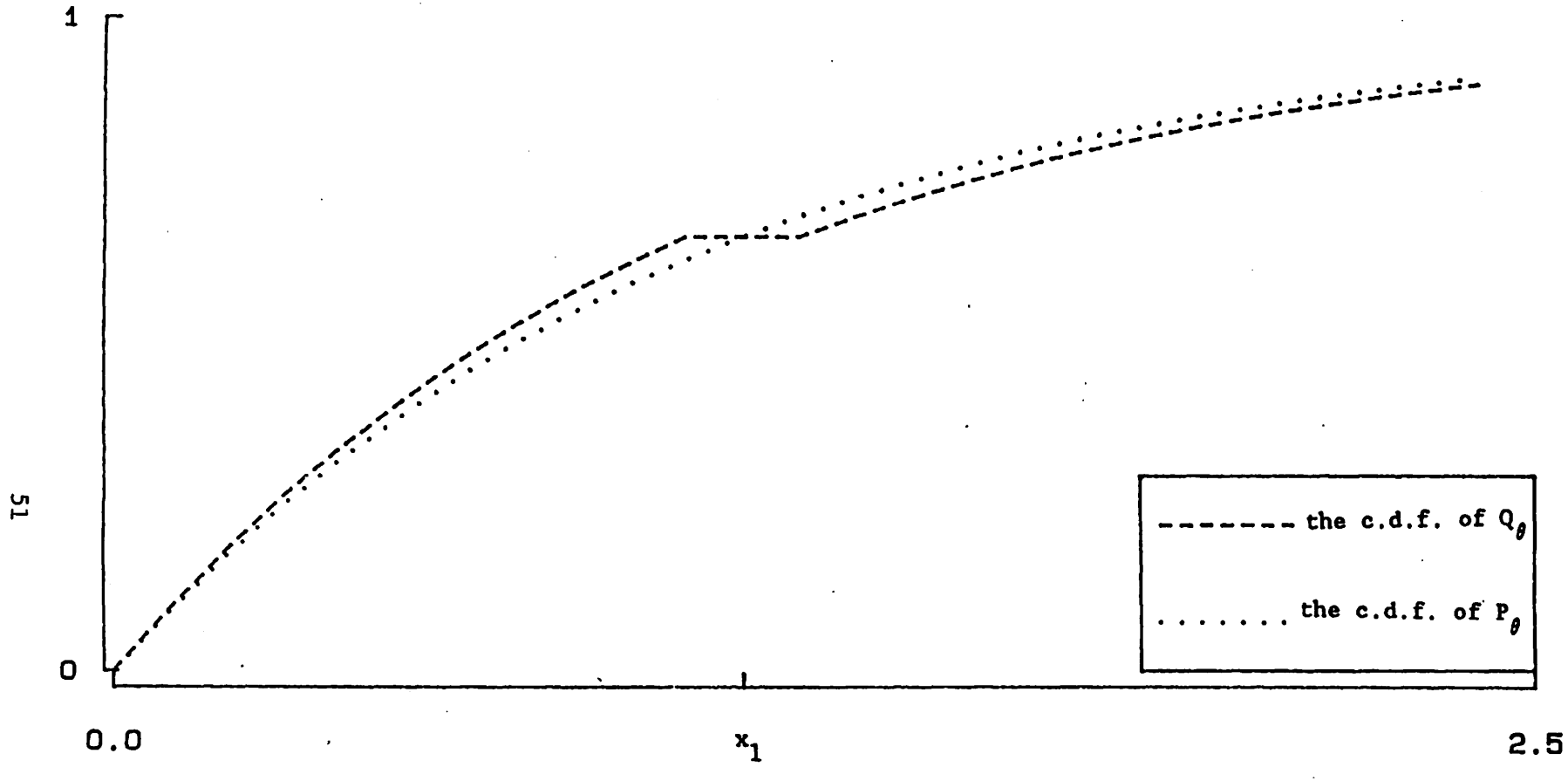
One aspect of the relationship is that $N(\theta)$ and P^α are both supposed to represent the set of points near a given point. We are required to use a metric to define P^α. We may, if we choose, use the same metric to define $N(\theta)$. We discussed previously what happens to psup when $N(\theta)$ is defined by a standard metric and when the conditional distribution for P

given P_θ satisfies either $Pr[P\in N(\theta)|P_\theta]\geq 1-\epsilon$ or $Pr[Pr[P\in N(\theta)|P_\theta]=1]\geq 1-\epsilon$. Those conditions both satisfy $E[Pr[P\in N(\theta)|P_\theta]]\geq 1-\epsilon$ so those results are relevant here.

Another aspect of the relationship is that the Γ's we use have interpretations similar to Prohorov neighborhoods. Suppose that $N(\theta)=P^\epsilon$ is defined by some metric and that the conditional distribution of P given P_θ satisfies $E[Pr[P\in N(\theta)|P_\theta]]\geq 1-\epsilon$. Then, for $B\subset\Omega$ and $\pi\in\Gamma$

$$
\begin{aligned}
1-\epsilon &\leq \int Pr[P\in N(\theta)|P_\theta] \\
&= \int I_B Pr[P\in N(\theta)|P_\theta] + \int (1-I_B) Pr[P\in N(\theta)|P_\theta] \\
&\leq \int I_B Pr[P\in N(\theta)|P_\theta] + 1 - \pi^m(B) \\
&\Rightarrow \int I_B Pr[P\in N(\theta)|P_\theta] \geq \pi^m(B)-\epsilon .
\end{aligned}
$$

Hence $\pi(B^\epsilon) \geq Pr[P_\theta\in B, P\in N(\theta)] = \int I_B Pr[P\in N(\theta)|P_\theta] \geq \pi^m(B)-\epsilon$. So $d(\pi,\pi_m)\leq\epsilon$ and $\Gamma\subset\{\pi : d(\pi,\pi_m)\leq\epsilon\}$. This leads naturally to the question "Is $\Gamma=\{\pi : d(\pi,\pi_m)\leq\epsilon\}$?" I don't know the complete answer but Appendix A proves a related result for the case where π_0 is discrete.



$Q_\theta \in N(\theta)$ See Example 4.1

FIGURE 5.1

6. REGRESSION

This section generalizes the previous discussion, which has been primarily about conditionally i.i.d. random variables, to the regression setting. In a typical regression setting the observation for the i-th case consists of the random variable Y_i and the covariate X_i which may be multidimensional. We assume that Y_i has a distribution in some parametric family indexed by θ, say $Y_i \sim P_{\theta_i}$ and that θ_i is a function of X_i and an unknown, possibly multidimensional, parameter β. The regression function is known, say $\theta_i = r(X_i,\beta)$.

The Y_i's are assumed to be independent given β and the X_i's. We also require that someone, usually the "expert," provides a prior opinion about β expressed by a probability measure. Of course the θ notation is superfluous. Instead of P_{θ_i} we could write $P_{r(X_i,\beta)}$. We use whichever notation is more convenient.

Example 6.1 Probit Regression:

Racine et al (1986) consider a probit regression in an acute toxicity test. They say "Typically such a test proceeds by administering various dose levels of the substance to batches of animals and subsequently observing their responses. The latter are most often characterized in terms of a simple dichotomy: for example, alive or dead." The Racine data are in the following table.

Dose (mg/ml)	Number of Animals	Number of Deaths
422	5	0
744	5	1
948	5	3
2069	5	5

In this example X_i is the i-th dose and Y_i is the number of deaths in the i-th group of animals. Y_i is taken to be binomially distributed with parameter $(\theta_i,5)^t$ where $\theta_i = \Phi(\beta_0+\beta_1 \ln(X_i))$ and Φ is the standard normal cdf. Some experts had a prior opinion about the chemical substance that some statisticians summarized as a bivariate normal prior distribution for $\beta=(\beta_0,\beta_1)^t$ with mean $\mu=(-17.31,\ 2.57)^t$ and covariance matrix

$$\Sigma = \begin{pmatrix} 1053.72 & -156.45 \\ -156.45 & 23.24 \end{pmatrix}.$$

Example 6.2 Linear Regression:
The usual linear regression setting assumes $(Y_i) \sim N(X_i^t\beta,\ \sigma^2)$. The regression function is $r(X_i,\beta,\sigma^2) = (X_i^t\beta,\ \sigma^2)$. Weighted regression, non-linear regression, non-normal regression and generalized linear models all fit into the framework we have described.

A choice of parametric family $\{P_\theta:\theta\in\Theta\}$, regression function r and distribution for β is equivalent to a prior distribution on $\Omega\times\Omega\times\ldots\times\Omega$ where Y_i has distribution P_i, an element of the i-th factor of Ω. The Y_i's are assumed to be independent given the P_i's. The number of factors can be large enough to accomodate future cases for which prediction is desired. Of course we usually can't specify a unique satisfactory prior distribution on $\Omega\times\Omega\times,\ldots,\times\Omega$ so we want to consider a class of plausible priors and see how much our inference varies over the class. We will define a class of priors using neighborhoods $N(\theta_i)\subset\Omega$ and lower and upper

bounds on both the regression function and the distribution of β. The new features are the regression function and the corresponding set of lower and upper bounds.

We are taking the X_i's to be fixed and known. If desired, errors in measurement could be modeled by the following scheme. Let χ_i be the set of possible "true values" for the covariate in the i-th case and define $N(\chi_i) = \cup\{N(r(X_i,\beta)):X_i\in\chi_i\}$. $N(\chi_i)$ contains all the points in Ω close to the parametric distributions corresponding to any possible "true value" of the covariate. To model the situation in which the X_i's are fixed but are measured with a small amount of error use $N(\chi_i)$ instead of $N(\theta_i)$ in the definition of the class of priors.

We can decompose the prior distribution on $\Omega\times\Omega\times\ldots\times\Omega$ into two parts, the marginal distribution of $\underset{\sim}{P}_\theta = (P_{\theta_1},P_{\theta_2},\ldots,P_{\theta_n})$ and the conditional distribution of $\underset{\sim}{P} = (P_1,P_2,\ldots,P_n)$ given $(P_{\theta_1},P_{\theta_2},\ldots,P_{\theta_n})$.

We define Γ, a class of priors, by a class of marginal distributions for $\underset{\sim}{P}_\theta$ and a class of conditional distributions for $\underset{\sim}{P}$ given $\underset{\sim}{P}_\theta$. For the class of conditional distributions we take the set of all distributions satisfying $\Pr[P_i\in N(\theta_i)|P_{\theta_i}] = 1$ for all $i\in\{1,\ldots,n\}$.

A marginal distribution for $P_{\theta_1},\ldots,P_{\theta_n}$ is determined by a regression function and a distribution on β. For the class of marginal distributions we take the set of all distributions determined by any regression function between a set of lower and upper bounds and any distribution for β between another set of lower and upper bounds. That is, if r is real valued, $\underset{\sim}{P}_\theta$ can have any distribution determined by r

and π where $\ell_r(X_i,\beta) \le r(X_i,\beta) \le u_r(X_i,\beta)$ and $\ell_\beta \le \pi \le u_\beta$. If r is vector valued the description of the bounds is different. When π is a prior let π^m denote the distribution of the β's and π^c denote the distribution of $\underset{\sim}{P}$ given $\underset{\sim}{P}_\theta$.

Before proceeding further we must resolve one more issue. Should $X_i=X_j$ imply $P_i=P_j$? We usually want $X_i=X_j$ to imply $\theta_i=\theta_j$ and hence $N(\theta_i)=N(\theta_j)$. But this does not mean $P_i=P_j$ or even that P_i and P_j have the same conditional distribution given $\underset{\sim}{P}_\theta$. Many possibilities would have their uses. In some circumstances we might require $P_i=P_j$. In others we might require P_i and P_j to be independent but have the same conditional distribution given $\underset{\sim}{P}_\theta$. In still others we might not require the conditional distributions to be the same. Each possibility says something different about our prior opinion. We will discuss how to find psup when we allow different conditional distributions. The modification for identical distributions is easy.

Now the question is how to compute psup for this type of Γ. In general we can find psup for any function $\phi(\underset{\sim}{P})$. As a specific example suppose we are interested in the predictive distribution of a future observable and take $\phi(\underset{\sim}{P})=P_n(S)$.

First we express the conditional distribution of Y_n given $X_1,\ldots,X_n,Y_1,\ldots,Y_{n-1}$ and some particular regression function r.

$$\Pr\left[\, Y_n \in S \mid X_1,\ldots,X_n,Y_1,\ldots,Y_{n-1} \,\right] =$$

$$\frac{\int P_n(S) \prod_{i=1}^{n-1} f(Y_i|P_i)\ \pi(d\underset{\sim}{P})}{\int \prod_{i=1}^{n-1} f(Y_i|P_i)\ \pi(d\underset{\sim}{P})} .$$

Using the same idea as before we see that this probability is greater than or equal to q if and only if

$$\int (P_n(S)-q) \prod_{i=1}^{n-1} f(Y_i|P_i)\ \pi(d\underset{\sim}{P}) \geq 0.$$

We look for choices for π and r that make the inequality true when $q\in(\text{pinf},\text{psup})$. As in Section 4 we first try to maximize $P_n(S)$ and then either maximize or minimize the product of the $f(Y_i|P_i)$ according to the sign of $(P_n(S)-q)$. The X_i's are fixed and we are constrained by $P_i\in N(\theta_i)=N(r(X_i,\beta))$. The development here is similar to that in previous sections.

We start by treating r and β as fixed and finding $\pi_q^c(\cdot|r,\beta)$. For each β we find $P^*_{n,r,\beta}\in N(r(X_n,\beta))$ that maximizes P(S) over all $P\in N(r(X_n,\beta))$. Then for each $i\in\{1,\ldots,n-1\}$ we find $P^*_{i,r,\beta}\in N(r(X_i,\beta))$ that maximizes or minimizes $f(Y_i|P)$, according to whether $(P^*_{n,r,\beta}(S)-q)$ is positive or negative. We take $\pi_q^c(\cdot|r,\beta)$ to be degenerate at $(P^*_{1,r,\beta},\ldots,P^*_{n,r,\beta})^t$.

That gives the conditional distributions of the P_i's for a fixed pair (r,β). Now we treat only β as fixed and find the best regression function for that β.

Since we know what the $P'_{i,r,\beta}$s are for each r and we are treating β as fixed we can think of the integrand $(P_n(S)-q)\Pi f(Y_i|P_i)$ as a function of r. We want to choose r^* to maximize the integral.

Now r is itself a function and the only aspect of r that matters is the set of values $r(X_1,\beta),\ldots,r(X_n,\beta)$. We start defining r^* by choosing the value $r^*(X_n,\beta)$. We take $r^*(X_n,\beta)$ to be that value in $[\ell_r(X_n,\beta),\ u_r(X_n,\beta)]$ that maximizes

$(P^*_{n,r,\beta}(S) - q)\ \Pi_{\{i:X_i=X_n\}} f(Y_i|P^*_{i,r,\beta})$. Then we choose the other values. We take, for example, $r^*(X_k,\beta)$ to be the value in $[\ell_r(X_k,\beta),\ u_r(X_k,\beta)]$ that either maximizes or minimizes $\Pi_{\{i:X_i=X_k\}} f(Y_i|P^*_{i,r,\beta})$ according to whether $(P^*_{n,r,\beta}(S)-q)$ is positive or negative.

Now we know what r and the P_i's should be for any fixed value of β. The only thing left to do is choose π^m_q to maximize the integral. But we have done this sort of thing before. We treat the integrand as a funciton of β, say $h(\beta)$. We take the density of π^m_q to be u_β whenever $h(\beta)>z$ and to be ℓ_β whenever $h(\beta)<z$. We choose z to make π^m_q be a probability measure. We know from previous results that

$$\Pr_{\pi_q, r^*}[\ Y_n \in S | X_1,\ldots,X_n,Y_1,\ldots,Y_{n-1}\ ] \geq q \quad \Leftrightarrow \quad \text{psup} \geq q.$$

Example 6.1 continued

Example 6.1 discusses a probit regression with a bivariate normal marginal prior π^m_0 for β and a regression function $r_0(X,\beta) = \Phi(\beta_0+\beta_1 \ln(X))$. Consider the following bounds on the regression function and the prior for β.

$$r_0(X/\sqrt{2},\beta) \leq r(X,\beta) \leq r_0(\sqrt{2}X,\beta) \text{ and}$$
$$\pi^m_0(\beta)/\sqrt{2} \leq \pi^m(\beta) \leq \sqrt{2}\pi^m_0(\beta)$$

Compute psup and pinf, the largest and smallest probabilities of success (death) for a dose x as r and π range over their permissible values.

Predictions were made for the following six dosages in the context of Example 6.1.

X=	245	403	665	1097	1808	2981
ln(X)=	5.5	6.0	6.5	7.0	7.5	8.0

pinf and psup were computed under eight conditions that form a 2^3 design. The three factors and their levels are

1)	use the original r_0	let r vary as described above
2)	use the original π_0^m	let π^m vary as described above
3)	make predictions before looking at the data (prior)	make predictions after looking at the data (posterior)

Computed values of pinf and psup are in Table 6.1.

When we use a range for the regression function then pinf and psup are closer together before performing the experiment than afterward. It is as though collecting data has increased our uncertainty or decreased our understanding of the drug's hazards. One possible explanation is that there are more variables we can use in the posterior case either to increase psup or to decrease pinf. Suppose we are trying to predict the probability of a success at dosage X=x. In the prior case, before observing data, the only part of the regression function that matters is $r(x,\beta)$. In the posterior case, after observing data, $r(422,\beta)$, $r(744,\beta)$, $r(948,\beta)$ and $r(2069,\beta)$ are also important. We can manipulate those parts of the regression function to lower pinf and raise psup.

The posterior probability of success is $\int r(x,\beta)\ \text{posterior}(d\beta)$. When we have data at $X_1,\ldots,X_4$ to work with we can choose $r(X_i,\beta)$ to give more posterior weight to β_s' where $r(x,\beta)$ is large and hence increase psup. However, we expect this effect to disappear as we collect even more data because any values for $r(X_i,\beta)$ that are far from the mle = (# of successes)/(# of trials) will be severely downweighted by the likelihood.

TABLE 6.1

(pinf, psup) for Example 6.1, the regression example

ln(dose)	fixed π fixed r	bounded π fixed r	fixed π bounded r	bounded π bounded r
Prior predictions				
5.5	(.042, .042)	(.029, .060)	(.032, .060)	(.023, .085)
6.0	(.072, .072)	(.051, .103)	(.048, .139)	(.034, .190)
6.5	(.215, .215)	(.161, .279)	(.092, .667)	(.065, .731)
7.0	(.804, .804)	(.743, .856)	(.376, .912)	(.311, .938)
7.5	(.929, .929)	(.901, .951)	(.870, .953)	(.821, .967)
8.0	(.958, .958)	(.941, .971)	(.942, .969)	(.918, .978)
Posterior predictions				
5.5	(.001, .001)	(.000, .003)	(.000, .243)	(.000, .282)
6.0	(.007, .007)	(.004, .013)	(.000, .512)	(.000, .543)
6.5	(.112, .112)	(.086, .141)	(.000, .980)	(.000, .986)
7.0	(.854, .854)	(.788, .866)	(.025,1.000)	(.019,1.000)
7.5	(.985, .985)	(.975, .991)	(.396,1.000)	(.364,1.000)
8.0	(.997, .997)	(.995, .999)	(.670,1.000)	(.723,1.000)

7. CLOSING REMARKS

Statisticians often perform analyses that begin by assuming a parametric distribution for the random variables. We hope that the assumption, made for computational convenience, doesn't lead us too far astray. The basis for the hope is usually a belief that the parametric model is approximately right or that the analysis is robust in some appropriate sense.

The research described in this paper was motivated by a dissatisfaction with that approach. We have seen how to do Bayesian analyses on classes of distributions that are not limited to parametric families. But neither are they completely general. They are periparametric and therefore tied to the underlying parametric family. Periparametric families are useful when there is a parametric family we believe to be approximately right.

Here is a list of key points in the technique for computing psup over periparametric classes of priors.

1) Use a class of priors rather than a single prior.

2) Priors are measures on Ω, not Θ.

3) Define a prior by its marginal on Θ and its conditional on Ω given θ.

4) For every $q \in [0,1]$ we can test whether psup is greater than q.

Section 4 uses those four points to define periparametric classes of priors and compute the corresponding set of posterior answers. Sections 5 and 6 discuss some issues raised in Section 4 and give some generalizations.

Density bounds arise naturally in defining the neighborhoods $\{N(\theta)\}$ that are used in constructing periparametric priors. As a bonus we can use them as in Sections 2 and 3 to define classes of priors on parametric families.

Box and Tiao (1962) discuss something that might be called sensitivity to the likelihood. But what are called different likelihoods in the usual terminology are different subsets of Ω in our terminology. Sensitivity to the likelihood deals with the effects of different priors on Ω and is really just another aspect of sensitivity to the prior.

This research was supported by the National Institute of General Medical Sciences GM 2527.

8. APPENDIX

Theorem:

Let μ be a measure on a measurable space χ. Suppose μ has finite support, say $\{x_1, x_2, \ldots, x_n\}$, and that $\mu(\chi)<\infty$. For each $i\in\{1,\ldots,n\}$ let there be an associated measurable set $N_i\subset\chi$. Suppose ν is another finite measure on χ satisfying

$$(*) \qquad \nu\left(\bigcup_{i\in I} N_i\right) \geq \mu\left(\bigcup_{i\in I} x_i\right)$$

for every $I\subset\{1,2,\ldots,n\}$. Then ν has a representation as

$\nu = \nu_1 + \nu_2 + \ldots + \nu_n + \alpha$ where

a) each ν_i is a measure satisfying $\nu_i(N_i) = \nu_i(\chi) = \mu(x_i)$ and

b) α is a measure.

The first equality in condition a) says that ν_i assigns mass only to the i-th set. The second equality says that the total mass of ν_i is the same as $\mu(x_i)$. It is as though the point mass from μ were spread throughout the set N_i to become the measure ν_i.

This theorem is related to Prohorov neighborhoods and the class Γ of Section 5. Let d be the Prohorov metric and B^ϵ be the ϵ-neighborhood of B. If $d(\nu,\mu)\leq\epsilon$, then $\nu(B^\epsilon)\geq\mu(B)-\epsilon$. Suppose μ is a discrete probability measure on Θ, a parametric subset of Ω. The theorem says that the collection of probability measures ν satisfying the stronger condition $\nu(B^\epsilon)\geq\mu(B)$ is the periparametric class Γ, of priors with P_θ-marginal μ and P-conditional satisfying $\Pr[P\in P_\theta^\epsilon | P_\theta]=1$. Clearly every $\nu\in\Gamma$ satisfies $\nu(B^\epsilon)\geq\mu(B)$. But if $\nu(B^\epsilon)\geq\mu(B)$ the theorem says $\nu=\Sigma\nu_i$ where

$\nu_i(P_\theta^\epsilon)=\nu_i(\Omega)$. ν is equal to the measure that has μ as its marginal and $\nu_i/\nu_i(P_\theta^\epsilon)$ as its conditional. So $\nu\in\Gamma$.

Proof of Theorem:

The proof is by induction on n, the number of points in the support of μ. The theorem is obviously true for n=1 and is not hard to show directly for n=2.

Assume the theorem holds when the support of μ is n-1 points. We want to show that it is also true when the support is n points. Let μ be a finite measure with support $\{x_1,\ldots,x_n\}$ and ν be a finite measure satisfying (*). When m is a measure and S is a measurable set let $(m|S)$ denote the measure m restricted to the set S, i.e. $(m|S)(B) = m(BS)$ for all measurable sets B.

Because ν satisfies (*) in relation to μ,ν also satisfies (*) in relation to $(\mu|\{x_1,\ldots,x_{n-1}\})$, a measure with n-1 support points. Therefore ν has a representation

$\nu = \nu_1 + \nu_2 + \ldots + \nu_{n-1} + \alpha$ where

a) each ν_i is a measure satisfying $\nu_i(N_i) = \nu_i(\chi) = \mu(x_i)$ and

b) α is a measure.

If $\alpha(N_n) \geq \mu(x_n)$ we could define

$$\nu_n = \frac{\mu(x_n)}{\alpha(N_n)} \cdot (\alpha|N_n) \text{ and}$$

$$\alpha^* = \alpha - \nu_n.$$

Then we would be done because $\nu_1,\ldots,\nu_n,\alpha^*$ would satisfy a) and b). So assume that $\alpha(N_n) < \mu(x_n)$. The idea behind the proof is to start with

$\nu_1,\ldots,\nu_{n-1},\alpha$ and modify them to create $\nu_1^*,\ldots,\nu_{n-1}^*,\alpha^*$ in such a way that $\nu_1^*,\ldots,\nu_{n-1}^*,\alpha^*$ still satisfy a) and b) and $\alpha^*(N_n)\geq\mu(x_n)$. If we can show that such modifications exist we will be done.

Partition the set $\{1,\ldots,n-1\}$ into two subsets I and $\bar{I}$ such that $i\in I$ if and only if there exists a sequence of indices $a_1,\ldots,a_k=i$ such that

1) $\nu_{a_1}(N_n) > 0$ and

2) $\nu_{a_j}(\bar{N}_n N_{a_{j-1}}) > 0$ for $j \in \{2,\ldots,k\}$.

Claim: I is not empty and $\alpha\left(\bar{N}_n \cup_{i\in I} N_i\right) > 0$.

Proof of Claim:

$$\text{If} \quad \sum_{1}^{n-1} \nu_i(N_n) = 0$$

then $\alpha(N_n) = \nu(N_n) \geq \mu(x_n)$ by (*). This is impossible because we are assuming $\alpha(N_n) < \mu(x_n)$. But

$$\text{if} \quad \sum_{1}^{n-1} \nu_i(N_n) > 0$$

then there exists some particular i' such that $\nu_{i'}(N_n) > 0$. I is not empty because $i'\in I$.

By the definition of I if $i\in I$ and $j\in\bar{I}$ then $\nu_j(N_i)=0$. Let

$A = N_n \cup \left(\cup_{i\in I} N_i\right)$. Again using (*)

$$\mu(x_n) + \sum_{i\in I} \mu(x_i) \leq \nu(A)$$

$$- \sum_{i\in I} \nu_i(A) + \alpha(A)$$

$$- \sum_{i\in I} \nu_i(A) + \alpha(N_n) + \alpha(\bar{N}_n \cup_{i\in I} N_i)$$

$$- \sum_{i\in I} \mu(x_i) + \alpha(N_n) + \alpha(\bar{N}_n \cup_{i\in I} N_i)$$

$$\Rightarrow \mu(x_n) - \alpha(N_n) \leq \alpha(\bar{N}_n \cup_{i\in I} N_i).$$

We are assuming that the left hand side is strictly positive, which proves the claim.

Any sequence of distinct indices $(a_1,\ldots,a_j : a_i \in I)$ will be called a chain of length j. If the quantities

$$\alpha(\bar{N}_n N_{a_j}),\ \nu_{a_j}(\bar{N}_n N_{a_{j-1}}),\ \nu_{a_{j-1}}(\bar{N}_n N_{a_{j-2}}),\ldots,\nu_{a_2}(\bar{N}_n N_{a_1}),\ \nu_{a_1}(N_n)$$

are all positive the chain is said to be available. Otherwise the chain is said to be broken. The claim implies that there is at least one available chain.

We can use an available chain to define new ν's and a new α. Suppose the chain $(a_1,\ldots,a_j)$ is available. Let r be the minimum of

$$\alpha(\bar{N}_n N_{a_j}),\ \nu_{a_j}(\bar{N}_n N_{a_{j-1}}),\ \nu_{a_{j-1}}(\bar{N}_n N_{a_{j-2}}),\ldots,\nu_{a_2}(\bar{N}_n N_{a_1}),\ \nu_{a_1}(N_n)$$

We can define

$$\alpha^* = \alpha - \frac{r}{\alpha(\bar{N}_n N_{a_j})}(\alpha|\bar{N}_n N_{a_j}) + \frac{r}{\nu_{a_1}(N_n)}(\nu_{a_1}|N_n)$$

$$\nu^*_{a_j} = \nu_{a_j} - \frac{r}{\nu_{a_j}(\bar{N}_n N_{a_{j-1}})}(\nu_{a_j}|\bar{N}_n N_{a_{j-1}}) + \frac{r}{\alpha(\bar{N}_n N_{a_j})}(\alpha|\bar{N}_n N_{a_j})$$

$$\nu^*_{a_{j-1}} = \nu_{a_{j-1}} - \frac{r}{\nu_{a_{j-1}}(\bar{N}_n N_{a_{j-2}})}(\nu_{a_{j-1}}|\bar{N}_n N_{a_{j-2}}) + \frac{r}{\nu_{a_j}(\bar{N}_n N_{a_{j-1}})}(\nu_{a_j}|\bar{N}_n N_{a_{j-1}})$$

$$\vdots$$

$$\nu^*_{a_2} = \nu_{a_2} - \frac{r}{\nu_{a_2}(\bar{N}_n N_{a_1})}(\nu_{a_2}|\bar{N}_n N_{a_1}) + \frac{r}{\nu_{a_3}(\bar{N}_n N_{a_2})}(\nu_{a_3}|\bar{N}_n N_{a_2})$$

$$\nu^*_{a_1} = \nu_{a_1} - \frac{r}{\nu_{a_1}(N_n)}(\nu_{a_1}|N_n) + \frac{r}{\nu_{a_2}(\bar{N}_n N_{a_1})}(\nu_{a_2}|\bar{N}_n N_{a_1})$$

For an index i not in $\{a_1,\ldots,a_j\}$ define $\nu^*_i = \nu_i$. Every term that appears with a plus sign also appears with a minus sign so it is still

true that $\nu = \nu_1^* + \ldots + \nu_{n-1}^* + \alpha$. Every term that was added or subtracted is a measure with total mass r. For every k the measure that was added to ν_k put all its mass on N_k. So $\nu_1^*,\ldots,\nu_{n-1}^*,\alpha$ satisfy conditions a) and b) above. Also $\alpha^*(N_n) = \alpha(N_n) + r$ so we are closer to the goal of $\alpha^*(N_n) \geq \mu(x_n)$.

We want to proceed in this fashion, using available chains to redefine the measures until we reach the goal. We will always use a superscript * to indicate the next redefinition of the process and a lack of superscript to indicate the current state. We know from the claim that as long as we have not reached the goal there are still available chains. The question is whether we can reach the goal in finitely many steps.

When a chain is used to redefine the measures that chain becomes broken. Since there are only finitely many chains it may appear that we must reach the goal in finitely many steps. However, it is possible for a broken chain to be fixed by a subsequent redefinition. If, in addition, r, the amount of mass by which we are able to increase $\alpha(N_n)$, is decreasing fast enough, we may never reach the goal. To conclude the proof of the theorem we show that if the chains are used in the proper order then no broken chain can be fixed and made available by any subsequent redefinition.

The ordering is simple. Order the chains by length and use the shortest chains first. For chains of the same length the ordering does not matter. We'll look first at what happens to chains of length 1 and then consider longer chains.

Suppose a chain of length 1 is available and we use it to redefine the measures. By renumbering we can let 1 be the only index in the

chain. So, $\alpha(\bar{N}_n N_1)>0$ and $\nu_1(N_n)>0$. After the redefinition either $\alpha^*(\bar{N}_n N_1)=0$ or $\nu_1^*(N_n)=0$. We see by the general form for redefining the measures that α increases on N_n and decreases on $\bar{N}_n$ so that $\alpha^*(\bar{N}_n N_1)$ can never become positive again once it has reached 0. Likewise, ν_1^* can only decrease on N_n. Therefore a chain of length 1 can never become available again once it has been broken.

For longer chains the argument is more complex. We will examine the conditions that must hold at the first time any chain is made available by a redefinition that comes from a chain of equal or greater length. We will see that such a situation is impossible and conclude that there is no such first time.

Say that Chain 1 has indices $a_1,\ldots,a_j$ and is now broken. Chain 2 with indices $b_1,\ldots,b_k$ is available and, if we redefine the measures using Chain 2, Chain 1 will become available. Assume $k\geq j$. Here are the quantities of interest for the two chains.

Chain 1	Chain 2
$\alpha(\bar{N}_n N_{a_j})$	$\alpha(\bar{N}_n N_{b_k})$
$\nu_{a_j}(\bar{N}_n N_{a_{j-1}})$	$\nu_{b_k}(\bar{N}_n N_{b_{k-1}})$
$\vdots$	$\vdots$
$\nu_{a_{i+1}}(\bar{N}_n N_{a_i})$	$\nu_{b_{\ell+1}}(\bar{N}_n N_{b_\ell})$
$\nu_{a_i}(\bar{N}_n N_{a_{i-1}})$	$\nu_{b_\ell}(\bar{N}_n N_{b_{\ell-1}})$
$\vdots$	$\vdots$
$\nu_{a_2}(\bar{N}_n N_{a_1})$	$\nu_{b_2}(\bar{N}_n N_{b_1})$
$\nu_{a_1}(N_n)$	$\nu_{b_1}(N_n)$

We know from the discussion on chains of length 1 that neither

$\alpha(\bar{N}_n N_{a_1})$ nor $\nu_{a_1}(N_n)$

will increase when we update the process. If the redefinition is to make Chain 1 available it must be by changing one of the other quantities. To be specific let's say that

$\nu_{a_i}(\bar{N}_n N_{a_{i-1}}) = 0$ but that $\nu^*_{a_i}(\bar{N}_n N_{a_{i-1}})$ will be positive.

Also we'll assume that the other quantities from Chain 1 are positive and leave the interested reader to deal with the case when more than one quantity becomes positive simultaneously. The redefinition will change

ν_{a_i} only if a_i appears in Chain 2. To be specific let's say $a_i = b_j$.

From the general form for redefinition we see that $\nu_{a_i} = \nu_{b_\ell}$ will increase on the set $\bar{N}_n N_{a_{i-1}}$ only if $\nu_{b_{\ell+1}}(\bar{N}_n N_{a_{i-1}}) > 0$ or $\alpha(\bar{N}_n N_{a_{i-1}}) > 0$.

But if $\alpha(\bar{N}_n N_{a_{i-1}}) > 0$ then the chain $a_1, a_2, \ldots, a_{i-1}$ is available and is shorter than Chain 2. Because we use short chains first $(a_1, \ldots, a_{i-1})$ must have been broken at some earlier point and then repaired. But repairing Chain 1 is supposed to be the first instance of repair. Therefore $\alpha(\bar{N}_n N_{a_{i-1}}) = 0$ and hence $\nu_{b_{\ell+1}}(\bar{N}_n N_{a_{i-1}}) > 0$.

Now consider two more chains.

Chain 3 = $(a_1, a_2, \ldots, a_{i-1}, b_{\ell+1}, b_{\ell+2}, \ldots, b_k)$ and

Chain 4 = $(b_1, b_2, \ldots, b_{\ell-1}, a_i, a_{i+1}, \ldots, a_j)$.

Because $\nu_{b_{\ell+1}}(\bar{N}_n N_{a_{i-1}}) > 0$ Chain 3 is available.

Also, $\nu_{a_i}(\bar{N}_n N_{b_{\ell-1}}) - \nu_{b_\ell}(\bar{N}_n N_{b_{\ell-1}}) > 0$, so that Chain 4 is available.

Because Chain 4 is available and we are contemplating using Chain 2, Chain 4 must be at least as long as Chain 2. Hence $j-i \geq k-\ell > k-(\ell+1)$ and Chain 3 is shorter than Chain 1. Therefore Chain 3 was used and broken earlier but is now available again. That contradicts the assumption that Chain 1 will be the first chain to be repaired by a subsequent redefinition.

Assuming that Chain 1 is the first chain to be repaired by a subsequent redefiniton leads to a contradiction. We conclude that there is no such first chain and hence that no chains are repaired by subsequent redefinitions. Therefore there can be only finitely many redefinitions. The proof is complete.