DISSERTATION

THE POOLING OF PRIOR DISTRIBUTIONS VIA LOGARITHMIC AND SUPRA-BAYESIAN METHODS WITH APPLICATION TO BAYESIAN INFERENCE IN DETERMINISTIC SIMULATION MODELS

Submitted by

Paul J. Roback

Department of Statistics

In partial fulfillment of the requirements

for the degree of Doctorate of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 1998

COLORADO STATE UNIVERSITY

June 22, 1998

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY PAUL J. ROBACK ENTITLED THE POOLING OF PRIOR DISTRIBUTIONS VIA LOGARITHMIC AND SUPRA-BAYESIAN METHODS WITH APPLICATION TO BAYESIAN INFERENCE IN DETERMINISTIC SIMULATION MODELS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTORATE OF PHILOSOPHY .

Committee on Graduate Work

Adviser

Department Head

ABSTRACT OF DISSERTATION

THE POOLING OF PRIOR DISTRIBUTIONS VIA LOGARITHMIC AND SUPRA-BAYESIAN METHODS WITH APPLICATION TO BAYESIAN INFERENCE IN DETERMINISTIC SIMULATION MODELS

We consider Bayesian inference when priors and likelihoods are both available for inputs and outputs of a deterministic simulation model. Deterministic simulation models are used frequently by scientists to describe natural systems, and the Bayesian framework provides a natural vehicle for incorporating uncertainty in a deterministic model. The problem of making inference about parameters in deterministic simulation models is fundamentally related to the issue of aggregating (i.e. pooling) expert opinion. Alternative strategies for aggregation are surveyed and four approaches are discussed in detail—logarithmic pooling, linear pooling, French-Lindley supra-Bayesian pooling, and Lindley-Winkler supra-Bayesian pooling. The four pooling approaches are compared with respect to three suitability factors—theoretical properties, performance in examples, and the selection and sensitivity of hyperparameters or weightings incorporated in each method–and the logarithmic pool is found to be the most appropriate pooling approach when combining expert opinions in the context of deterministic simulation models.

We develop an adaptive algorithm for estimating log pooled priors for parameters in deterministic simulation models. Our adaptive estimation approach relies on importance sampling methods, density estimation techniques for which we numerically approximate the Jacobian, and nearest neighbor approximations in cases in which the model is noninvertible. This adaptive approach is compared

to a nonadaptive approach over several examples ranging from a relatively simple $\Re^1 \rightarrow \Re^1$ example with normally distributed priors and a linear deterministic model, to a relatively complex $\Re^2 \rightarrow \Re^2$ example based on the bowhead whale population model. In each case, our adaptive approach leads to better and more efficient estimates of the log pooled prior than the nonadaptive estimation algorithm. Finally, we extend our inferential ideas to a higher-dimensional, realistic model for AIDS transmission.

Several unique contributions to the statistical discipline are contained in this dissertation, including:

1. the application of logarithmic pooling to inference in deterministic simulation models;

2. the algorithm for estimating log pooled priors using an adaptive strategy;

3. the Jacobian-based approach to density estimation in this context, especially in higher dimensions;

4. the extension of the French-Lindley supra-Bayesian methodology to continuous parameters;

5. the extension of the Lindley-Winkler supra-Bayesian methodology to multivariate parameters; and,

6. the proofs and illustrations of the failure of Relative Propensity Consistency under the French-Lindley supra-Bayesian approach.

Paul J. Roback
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Summer 1998

ACKNOWLEDGEMENTS

This research was supported by the North Slope Borough (Alaska), the State of Alaska (through the Alaska Department of Community and Regional Affairs), and the National Oceanic and Atmospheric Administration (through the National Marine Mammal Laboratory to the Alaska Eskimo Whaling Commission), through North Slope Borough contracts 97-192 and 98-116; and by the National Oceanic and Atmospheric Administration (through the National Marine Mammal Laboratory) through contract 40ABNF601311. I am grateful to these organizations for their generous assistance.

I would like to thank my advisor, Geof Givens, for the direction and encouragement he provided me. I feel extremely fortunate to have worked with someone who is so willing to share his time and his considerable expertise, who is demanding yet realistic and pragmatic. The loyalty, confidence and friendship he extended were greatly appreciated.

Thank you to my other committee members—Jennifer Hoeting, Adele Howe, and Richard Tweedie—for their support and suggestions, and an extra thanks to Jennifer for helping me as a teacher through her example and encouragement. Thanks also to other professors at Colorado State (especially Hari Iyer) who helped broaden and deepen my statistical horizons.

Thanks to my parents, Jim and Pat Roback, for instilling in me an appreciation for a good education and a determined effort. Their love and influence are felt more than they know.

Thanks to my sisters Jill and Joni and my brother Peter. A few miles aren't nearly enough to break the close bonds we share, and I'm grateful for their support and inspiration and love during my years at CSU and before.

Thanks to my grandparents, especially Grandma Steenson for her enthusiasm and encouragement.

Thanks to my many friends scattered across this country for sharing their time and laughter and their hopes and experiences to help make this journey enjoyable. I'm constantly touched by the genuineness and permanence of the friendships I've made. In particular, thanks to Nels and Billy, the Boe guys (and devoted groupies), the Swat guys, the Indy people, the OFBL gang, and my fellow CSU stat-ers.

Finally, thank you to my wife Karen and my daughter Samantha. Samantha, in just a year, has brought feelings of joy and love along with a new perspective into my life that I never thought possible. And without Karen this never would have happened. Her unwavering support, unspoken selflessness, and unconditional love helped keep me going and made the experience rich. This is our PhD, and I'm excited by the possibilities which lie ahead now for us.

CONTENTS

Chapter 1

Introduction

1.1 Simulation Models

A *simulation model* is a set of assumptions about how a particular system works, often expressed in the form of mathematical equations and logical relationships. Simulation models often provide simplified yet reasonably accurate descriptions of how systems work, and they can be used to gain a deeper understanding of systems or to make predictions regarding future output values. In simulation modeling, one often studies the underlying system by numerically determining the outputs produced by a particular set of inputs.

1.1.1 Classifications

It is often helpful to classify simulation models along four dimensions as described below (the first three have been suggested by Law and Kelton (1991)):

Static vs. Dynamic. A static simulation model represents a system at a particular time point, or a system in which time plays no role. A dynamic simulation model represents a system as it evolves over time.

Continuous vs. Discrete. In discrete simulation, the system is assumed to change at only a countable number of points in time. Under continuous simulation, a system can change continuously with respect to time, so that continuous simulation models often involve differential equations that give relationships for the rates of change with time.

Deterministic vs. Stochastic. If a simulation model contains no random (probabilistic) components, then it is called deterministic; otherwise, it is called stochastic. Therefore, a model is deterministic if, given the same inputs, the output is always the same. For instance, a complicated and analytically intractable system of differential equations describing a chemical reaction might be a deterministic simulation model, while an inventory system in which demand is a random variable following a particular probability distribution might be stochastic.

Mechanistic vs. Empirical. Mechanistic models attempt to incorporate the causal framework of a system, while empirical models are based solely on observation. Most statistical models are empirical, in that they attempt to describe a set of data with a model which is as small and simple as possible, without necessarily providing a detailed description of causal relationships.

In this dissertation, we will focus on simulation models which are often dynamic and discrete, and always deterministic and mechanistic.

1.1.2 Examples of Simulation Models

In each of the three examples below, a mathematical model is used to simulate an actual system. Assumptions are made about how a system works, and the system is expressed in terms of mathematical formulas and logical relationships. Computers are employed to generate results because of either the complexity of the model (AIDS and whales) or the need for random number generation (bank tellers). Results from the simulation models are then used to guide decision-making and enhance understanding of the actual system. Each of our three introductory examples is dynamic and discrete, and the AIDS transmission and whale population models are deterministic and primarily mechanistic. The bank teller example is stochastic and primarily empirical, and it is presented to give a flavor of the potential range of simulation models.

Multi-teller bank with jockeying (Law and Kelton, 1991)

A bank with five tellers opens its doors at 9 AM. Each teller has a separate queue, and an arriving customer joins the shortest queue, selecting the leftmost shortest queue in the case of ties. In addition, customers tend to jockey among lines. Thus, if n_i is the total number of customers in front of teller i (in service plus waiting), then the customer from the tail of queue j will jockey to the tail of queue i whenever the completion of a customer's service at teller i causes $n_j > n_i + 1$.

The bank's management would like to determine the optimal number of tellers, considering both the quality of their customer service and the costs of operation. For various potential numbers of tellers, the managers would like to obtain estimates of the expected time-average total number of customers in queue, the expected average delay in queue, and the expected maximum delay in queue.

A day at the bank can be simulated after making certain assumptions. For instance, the interarrival times of customers may be assumed to be independent and identically distributed (i.i.d.) exponential random variables with mean 1 minute, and the service times of customers may be assumed to be i.i.d. exponential random variables with mean 4.5 minutes. The managers' desired response variables can be obtained by simulating many days with a fixed number of tellers. After generating response variables for various numbers of tellers, the managers may decide to vary distributional assumptions and repeat the process.

AIDS in San Francisco (Hethcote et al. 1991)

Hethcote *et al.* (1991b; 1991a) developed an epidemiological model of the incidence and prevalence of HIV and AIDS among homosexual men in San Francisco. Using population structure, migration patterns, and interaction patterns, Hethcote *et al.* formed a model which involved the infection of susceptible individuals and the progression of HIV-infected people through stages leading to AIDS

and eventually death. The AIDS model has a compartmental structure in which individuals move through the disease stages, interact with other men, alter their level of sexual activity, and infect partners with changing probabilities. In this way, the authors hoped to reconstruct the HIV epidemic. In fact, the model was designed to "incorporate all features considered essential for modeling the sexual spread of HIV infections in homosexual men in San Francisco and yet . . . be simple enough that estimates of the values of the parameters are possible."

Because of the abundance of data from sexual behavior studies in San Francisco, specific values of important model parameters could be set with a good degree of confidence in many cases. Important model parameters included fraction of population who were sexually very active, natural mortality rate, migration rate, transfer rate from sexually very active to sexually active, number of infectious stages, rate of progression from stage k of AIDS to stage $k+1$, relative infectivity of stage k men, starting date of the epidemic, starting date for reduction in average number of partners per month, and several others. Reasonable values of input parameters were found by fitting model outputs to estimated HIV and AIDS incidence rates, adjusting the most unreliable inputs to produce the best fitting outputs.

Such an AIDS model can serve several roles. In general, a simulation model for AIDS "serves as a framework for organizing and coalescing a wide variety of data." It describes the mechanisms which bring about disease progression, providing an answer to the question "How?" rather than just the question "How many?". It allows prediction of future incidences among the homosexual male population in San Francisco. It provides a vehicle for evaluating the potential effect of therapies and the optimal administration strategies. The model also allows researchers to test specific hypotheses such as: patients remain in the asymptomatic stage for at least 100 days on average; sexually very active males

have 5 times as many partners per month as other homosexual males; a susceptible male has a 25 percent chance of being infected during an encounter with an asymptomatic partner. Even more elaborate hypotheses can be evaluated, such as whether the unusual reduction in HIV incidence observed in 1982 was caused by oversaturation in the very sexually active group, or by a general reduction in sexual activity.

An application of the Bayesian inferential methods developed here to the AIDS model of Hethcote *et al.* is contained in Chapter 8.

Population Model for Bowhead Whales (Raftery, Givens, and Zeh, 1995)

The bowhead whale *Balaena mysticetus* was reduced to near extinction by excessive commercial harvesting in the second half of the nineteenth century. Today, the largest remaining stock can be found in the Western Arctic (Bering-Chukchi-Beaufort Seas), and these whales are protected from commercial harvesting by the International Convention for the Regulation of Whaling. The Convention does, however, allow limited aboriginal subsistence whaling because of the cultural and subsistence needs of Eskimo peoples in the area. To protect the maintenance of the bowhead stock and prevent a return to near extinction levels, the International Whaling Commission (IWC) sets a quota for aboriginal subsistence whaling. To determine this quota, the IWC relies on a population model for bowhead whales.

Raftery *et al.* (1995) discuss methods of inference for a population dynamics model for bowhead whales developed by Breiwick *et al.* (1984). This model combines three types of relevant information: recent surveys, historical whaling records, and biological information. Recent surveys contain yearly estimated counts from visual surveyance and hydrophone use during migration periods, and historical whaling records contain validated yearly kill counts. Biological information includes birth and death rates as estimated from photoidentification, exami-

nation of dead animals, and other sources. Additional inputs into the population dynamics model of Breiwick *et al.* include age-specific natural mortality and fertility that may be density-dependent. Outputs include yearly populations, broken down by age and sex.

IWC policy regarding aboriginal subsistence whaling quotas focuses on one particular model output—replacement yield (RY). RY is the greatest number of whales that could be harvested in a given year without decreasing the current population. Because they want to set conservative quota levels, the IWC is interested in not only a point estimate of replacement yield, but also in an expression of uncertainty regarding the estimate of RY.

In addition to helping the IWC form policy, a population dynamics model for bowheads can have several other valuable purposes. Such a model provides a description of a natural system rather than merely a means of making predictions. Because of its descriptive nature, a bowhead whale population model can be used to evaluate hypotheses (e.g. Do female whales reach maturity at age 20?), to simulate effects of proposed actions (e.g. What would the long-term effects be if no harvesting were allowed next year?), and to trace potential effects of system disruptions (e.g. What if there were a major oil spill in the bowheads' migration path?).

1.1.3 Other examples

The three detailed examples above illustrate that simulation models can be a very effective tool for answering questions and gaining knowledge in a variety of areas. In addition, simulation models are used quite heavily in engineering and manufacturing applications. One reason is that it is often simpler, more cost-effective, and less disruptive to experiment with a model of a system than the system itself. Such cases include an automobile manufacturing plant, where it

would be very costly to experiment with the actual assembly line, or the bank teller example, where experimentation with the number of tellers could disrupt customer service. Another reason for the proliferation of simulation models in the engineering and manufacturing arenas is the advantages in terms of time and resources which simulation models have over building physical models. Physical models, such as wind tunnels for automobiles or models of the surface of Mars for space discovery vehicles, are sometimes extremely useful, but often not worth the time and expense. Mathematical models can often describe the system as accurately and lead to inferences as valid as those from physical models, while avoiding construction issues.

A multitude of applications of simulation models to engineering and manufacturing problems have been described in the literature. For example, Law and McComas (1988) used simulation modeling to study a manufacturing facility which produced metal parts requiring three distinct subassemblies. They wanted to determine optimal numbers of containers, forklift trucks, output queue positions for the loaders, and required shifts for the assembler. Pritsker (1986) considers the unloading of oil tankers and the supplying of oil to a refinery. Optimal unloading patterns are sought to avoid frequent startups and shutdowns of the refinery, which requires a minimum level of oil in storage tanks.

Deterministic simulation models are also frequently used by scientists in diverse areas to describe natural systems. In addition to the bowhead whale population and AIDS transmission models previously described, we offer a small sampling of examples here. George and Grant (1983) describe a stochastic simulation model of brown shrimp population dynamics in Galveston Bay, Texas. This model is then used to evaluate the effects of management alternatives and changing environmental conditions on shrimp dynamics. Steinhorst *et al.* (1985) create a stochastic-deterministic simulation model of canopy cover development of several

shrub species. In particular, they focus modeling efforts on the first 15 years following clearcutting and burning in moist coniferous forests of northern Idaho. King and Johnson (1993) model five generations of tree breeding improvement using Monte Carlo simulation, where model details are based on the New Zealand radiata pine improvement program. They address the effects of different mating schemes on both per generation gain and genetic diversity. Gelman *et al.* (1996) estimate parameters in physiological pharmacokinetic models, illustrating their approach with an application to the metabolism of tetrachloroethylene (a potential occupational hazard in dry cleaners). Green and Strawderman (1996) develop a Bayesian version of an existing growth and yield model for slash pine plantations in the West Gulf region of the United States. Bogstad *et al.* (1992) create a model called MULTSPEC which describes the biology of the Barents Sea, modeling processes such as spawning, maturation, feeding, predation, migration, fishing, and mortality in populations from harp seal to plankton. Finally, many scientists (Cubasch and Cess, 1990; Oglesby *et al.*, 1989; Henderson-Sellers and Robinson, 1986; Simmons and Bengtsson, 1984) use nonlinear partial differential equations to create general circulation models for Earth's climate.

All of the models described here differ with respect to certain aspects: the complexity of the model, which can be extreme; the number of input and output parameters, which can climb quickly into the hundreds in some applications; and the level of sophistication in which current analysis techniques evaluate uncertainty in parameter inferences. Yet, they are all examples of the class from simulation models.

1.2 Simulation Models and Statistics

Despite the widespread use of simulation models among scientists and engineers, statisticians have been slow to investigate associated methods of inference.

Although some work has been done in important areas such as model validation (Rawlings, 1988) and Monte Carlo simulations (Hammersley and Hanscomb, 1964), statisticians have been slower to embrace the idea of inference for deterministic simulation models. The resistance of statisticians may be strongly related to their innate suspicion of models which involve no random variability. Scientists, however, often find it easier to think and express themselves in terms of deterministic models, and these models are often very reasonable approximations to the truth. The Bayesian framework, under which uncertainty about model inputs and outputs can be expressed in terms of prior distributions, provides a natural vehicle for incorporating uncertainty in a deterministic model. In addition, the Bayesian framework can provide improved methods for building simulation models and evaluating the sensitivity of these models to changes in parameter values.

1.3 Introduction to Bayesian Methods

A frequentist evaluates procedures based on repeated samplings (or hypothetical repeated samplings) for fixed values of unknown parameters. A Bayesian, on the other hand, considers unknown parameters random, and he bases inference on the distribution of these unknown parameters conditional on observed data. This conditional distribution is called the posterior distribution.

Under the Bayes approach, we seek an expression which summarizes what we know about the model parameter θ *after* taking the sample, assuming that we know the prior distribution $\pi(\theta)$ which summarizes what we know about θ *before* the sample. Thus, we want the posterior distribution of θ given the data x, a conditional distribution denoted by $p(\theta \mid x)$. Inference concerning θ will be based on this posterior distribution. Often, a vector η of hyperparameters is introduced, so that the prior distribution becomes $\pi(\theta \mid \eta)$ and the posterior distribution becomes $p(\theta \mid x, \eta)$. Then, Bayes' Theorem is used to obtain the

posterior (notation follows that of Carlin and Louis, (1996)):

$$p(\theta \mid x, \eta) = \frac{p(\theta, x \mid \eta)}{p(x, \eta)} = \frac{f(x \mid \theta)\pi(\theta \mid \eta)}{\int f(x \mid u)\pi(u \mid \eta)du} = \frac{f(x \mid \theta)\pi(\theta \mid \eta)}{m(x \mid \eta)}$$

where $m(x \mid \eta)$ is the marginal distribution of the data x given the value of the hyperparameters η, and $f(x \mid \theta)$ is the likelihood function for the data x. Specifically, with n independent observations, we have $f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$. If η is known, it can be suppressed in the notation. Through Bayes' Theorem, we see that the posterior density is proportional to the likelihood function times the prior density; we can think of the posterior as the prior after it has been updated to reflect the data.

Bayes' Theorem may also be used sequentially. Suppose we have two independently collected samples of data x_1 and x_2. Then (suppressing η),

$$\begin{aligned} p(\theta \mid x_1, x_2) &\propto f(x_1, x_2 \mid \theta)\pi(\theta) \\ &= f_2(x_2 \mid \theta) f_1(x_1 \mid \theta)\pi(\theta) \\ &\propto f_2(x_2 \mid \theta) p(\theta \mid x_1) \end{aligned}$$

Thus, the posterior for the full data set (x_1, x_2) is obtained by first finding $p(\theta \mid x_1)$, the posterior for θ given x_1, and then treating $p(\theta \mid x_1)$ as the prior for finding the posterior $p(\theta \mid x_1, x_2)$. With this relationship, updating the posterior when data arrives sequentially over time is a trivial matter, although this is an important issue in the consideration of "external Bayesianity" in Chapter 2.

1.3.1 Determining Prior Distributions

A crucial aspect of Bayesian analysis is the ability (and willingness) to determine a prior distribution for θ. Feelings about typical values of θ must be quantified in probability distribution form before looking at the data x, often relying on previous studies or expert opinion. Several approaches have been em-

ployed for determining $\pi(\theta)$; three common approaches as identified by Carlin and Louis (1996) are described here.

Elicited priors. Under this approach, a distribution is obtained which represents a subject-area expert's opinion about θ. Common elicitation methods include assigning probabilities to possible values of θ and defining parametric values for an assumed parametric form. If θ is discrete, assigning probabilities to possible values of θ is a natural process. If, however, θ is continuous, probabilities can be assigned to intervals on the θ-axis. Although discretizing a continuous distribution may seem less than ideal, it often provides adequate results, especially when computation of the posterior requires numerical integration. In the second elicitation method, a parametric distribution form is chosen, and the expert must merely define a few parameters or quantiles of this distribution.

Conjugate priors. This is an approach of analytical convenience. It is sometimes possible to select a prior which is conjugate to the likelihood, which means that the resulting posterior distribution belongs to the same distributional family as the prior. Morris (1983) showed that exponential families, a typical source of likelihoods, always have conjugate priors.

Noninformative priors. This is an approach used if no reliable prior information about θ exists, or if inference based solely on the data is desired. A noninformative prior distribution $\pi(\theta)$ contains no information about θ in some formal sense—a common sense of noninformativeness leads to the Jeffreys prior $\pi(\theta) = [I(\theta)]^{1/2}$, where $I(\theta)$ is the expected Fisher information such that $I(\theta) = -E_{\underline{x}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f(\underline{x} \mid \theta)\right]$ (Jeffreys, 1961). Berger (1985) reviews methods of finding noninformative priors.

1.3.2 Elicited Priors from Multiple Experts

Often, a decision maker preparing to use Bayesian analysis has sufficient knowledge about the parameter of interest, θ, to form a subjective prior distribution for θ. In some cases the decision-maker realizes his own knowledge of θ is limited, so she consults an expert about θ to obtain that expert's subjective prior. For example, an individual considering an investment in a particular company might consider the expert opinion of a financial analyst regarding that company's projected stock price in six months. In certain cases when the problem is deemed especially important, the decision maker may obtain opinions from more than one expert. For instance, if the investor above is responsible for a portion of a college's endowment, she would likely consult with multiple financial analysts. The investor now has several subjective prior distributions for θ, but she needs a single "consensus distribution" to combine with sample evidence in the form of a likelihood function to make Bayesian inference (or to use as a basis for decision-making in the absence of sample evidence). This problem of deriving a consensus distribution is discussed fully in Chapter 2.

1.3.3 Bayesian Methods and Deterministic Simulation Models

A common criticism of Bayesian analysis is the difficulty in obtaining reliable, well-justified prior distributions. Several studies have illustrated the statistical and psychological difficulties individuals encounter when attempting to assess their own probability distributions, even for parameters with which they are well familiar (Tversky and Kahneman, 1974; Savage, 1971). For deterministic simulation models, however, reliable priors often exist for many model parameters. In both the AIDS and the bowhead whales examples, data describing most model inputs exists from a wealth of previous studies.

In particular, consider the AIDS model. A generalized gamma distribution is used as the prior distribution for the AIDS incubation period according to work

by Longini *et al.* (1989; 1990). The size of the homosexual male population in San Francisco has been investigated by both Pickering *et al.* (1986) and Lemp *et al.* (1990). Contact rates between subpopulations can be described by the preferred mixing model of Jacquez *et al.* (1988). Grant *et al.* (1987) studied the probability of transmission per partner. Other model parameters have been the subjects of similar independent studies. As a result, the often sticky issue of prior elicitation brings about less cause for concern with many of the deterministic simulation models we will consider.

Bayesian inference of deterministic simulation models involves an additional step following the elicitation of prior distributions for model inputs and outputs. Because a set of inputs completely defines a corresponding set of outputs, a deterministic model actually contains multiple priors in input space—those prior distributions explicitly defined for the inputs, and those implicitly defined by prior distributions for the outputs and the deterministic equation mapping inputs and outputs. Similarly, output space also contains multiple priors. The priors in one parameter space, for example, must be pooled into a single prior before Bayesian inference can be made. It is this problem, especially in the face of complex deterministic models, which motivates the work in this dissertation.

1.3.4 An Introductory Example

To provide a background framework for the remaining chapters, consider the following inference problem involving a deterministic simulation model. Let $M(\theta) = (\theta - 3)/4 = \phi$ be a simple model linking an input θ to an output ϕ. For instance, M could be a population model in which ϕ is a growth rate parameter and θ is a fertility parameter. The ultimate goal is to make inference about ϕ. Suppose an expert on θ (Expert 1) and an expert on ϕ (Expert 2) are independently consulted, and they each offer a prior distribution which describes their beliefs

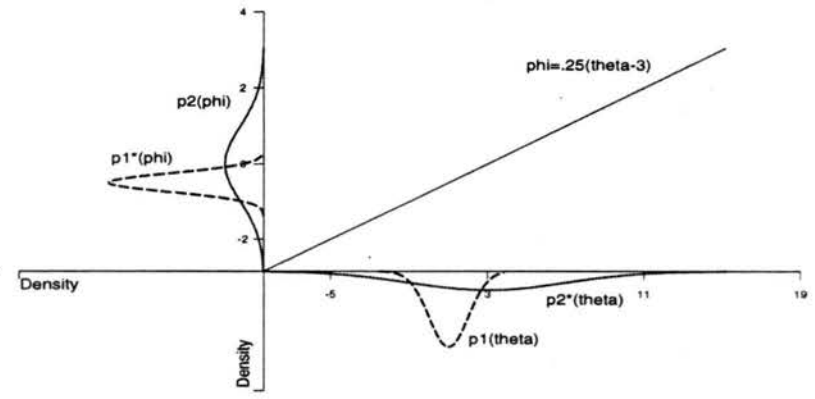

Figure 1.1: Priors and implied priors for θ and ϕ, for an introductory example.

about "reasonable values" of their parameter of expertise and corresponding uncertainty. Suppose the solicited prior distributions were: $\theta \sim p_1(\theta) = N(\theta; 1, 1)$ and $\phi \sim p_2(\phi) = N(\phi; 0, 1)$, where $N(x; \mu, \sigma^2)$ represents a normal density with mean μ and variance σ^2. This example is sketched in Figure 1.1.

Given M, Expert 2 implicitly professes an opinion about reasonable values of θ. Similarly, Expert 1 implicitly professes an opinion about ϕ. This occurs even if each expert has no knowledge of the other's field. Specifically, the implicit prior induced on θ by $p_2(\phi)$ is $p_2^*(\theta) = N(\theta; 3, 16)$ and the implicit prior induced on ϕ by $p_1(\theta)$ is $p_1^*(\phi) = N(\phi; -\frac{1}{2}, \frac{1}{16})$. Clearly, p_1 and p_2 are incoherent, because when both are expressed in the same space, they do not agree.

Suppose now some data D_ϕ are observed that provide direct evidence about ϕ in the form of a likelihood $L(D_\phi \mid \phi)$. This permits Bayesian inference about ϕ if the likelihood is combined with a single prior for ϕ. If no evidence is to be lost, the two available priors for ϕ, $p_2(\phi)$ and $p_1^*(\phi)$, must be pooled to form a single prior, $p(\phi)$. How should the pooled prior $p(\phi)$ be obtained? This is the question that drives this dissertation. Once $p(\phi)$ is formed, we can quantify the coherized

posterior for ϕ, $\pi(\phi) \propto p(\phi)L(D_\phi \mid \phi)$. Inference about ϕ, our primary goal, can be based on a sample from the posterior $\pi(\phi)$.

1.4 Scope of Work

In Chapter 2, we review the literature about aggregating expert opinion and forming group consensus. The approaches we discuss in detail include linear pooling, logarithmic pooling, supra-Bayesian methods, and group interaction methods. Chapters 3 and 4 focus on the log pooling approach. In Chapter 3, we describe a general algorithm for obtaining log pooled priors for deterministic simulation models. In particular, we detail an adaptive algorithm designed to drastically improve sampling efficiency. We illustrate in Chapter 4 how our adaptive importance sampling approach leads to less biased and more precise estimates of the pooled prior in several cases when compared to a simpler estimation algorithm. Five examples are used to illustrate the performance of the log pool and compare the two versions of the general algorithm in inference problems where priors linked by a deterministic simulation model must be pooled. These examples range from a relatively simple $\Re^1 \rightarrow \Re^1$ example with normally distributed priors and a linear deterministic model, to a relatively complex $\Re^2 \rightarrow \Re^2$ example based on the bowhead whale population model. In Chapters 5 and 6, two specific methods under the "supra-Bayesian" approach (a potential alternative to log pooling) are detailed and discussed. These methods, denoted the French-Lindley method and the Lindley-Winkler method, are applied to the same illustrative examples as used in Chapter 4 to examine their properties. Then, in Chapter 7, four approaches for pooling expert opinion—logarithmic pooling, linear pooling, the French-Lindley supra-Bayesian method, and the Lindley-Winkler supra-Bayesian method—are compared and contrasted. Finally, in Chapter 8 we extend our inferential ideas to

a larger (6 input and 6 output parameters), realistic model for AIDS transmission, and in Chapter 9 we summarize the important results of this research.

Chapter 2

Aggregating Expert Opinions

2.1 Simulation Models, Priors, and Likelihoods

Consider the Bayesian use of a deterministic simulation model, $\boldsymbol{M}(\boldsymbol{\theta})$: $\boldsymbol{\theta} \rightarrow \boldsymbol{\phi}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \Re^m$ and $\boldsymbol{\phi} \in \boldsymbol{\Phi} \subseteq \Re^p$. The deterministic simulation models we consider are often carefully crafted to explicitly model the mechanism of interest. These are unlike statistical models, which aim to empirically estimate the relationships between variables. Suppose that likelihoods $L(D_{\boldsymbol{\theta}}|\boldsymbol{\theta})$ and $L(D_{\boldsymbol{\phi}}|\boldsymbol{\phi})$ are available based on independent data $D_{\boldsymbol{\theta}}$ and $D_{\boldsymbol{\phi}}$. Further, suppose that priors $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\phi})$ are elicited independently on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, but that the functional relationship $\boldsymbol{M}$ is ignored during the development of these priors. Inference is desired about $\boldsymbol{\phi}$ and/or $\boldsymbol{\theta}$.

Why would one ignore $\boldsymbol{M}$ during prior elicitation? Particular variables that are inputs or outputs of a simulation model $\boldsymbol{M}$ are often the subject of basic scientific research in diverse disciplines. The model itself need not have been developed when such research was conducted, and the priors implied by the evidence gathered may not reflect a thorough understanding of $\boldsymbol{M}$ or even the knowledge that $\boldsymbol{M}$ exists. Even when the model is available, no one may have sufficient widespread expertise to synthesize knowledge from the diverse disciplines. Further, scientists and statisticians may be prevented from using $\boldsymbol{M}$ to synthesize the evidence because of the extreme complexity or noninvertibility of $\boldsymbol{M}$. Finally, any model misspecification may make it difficult or impossible to synthesize the

evidence in a manner wholly compatible with $\boldsymbol{M}$, and virtually all models are misspecified to some extent.

For example, in the whale population model example to which we apply our methods in Section 4.6, model inputs include a maximum potential productivity parameter, which would be the object of study by biologists relying mainly on examination of specimens. Outputs include current population abundance, which is estimated by ice-based visual and acoustical census which requires no biological expertise. It is not clear how to reconcile these two sources of evidence, even if one knows a good population model.

Even in the case where $m = p$ and $\boldsymbol{M}$ is a smooth, continuous, invertible mapping, the solution to this Bayesian inferential problem is nontrivial. For example, consider inference about $\boldsymbol{\phi}$. Likelihoods are invariant to parameter transformation, so we may re-express the joint likelihood as $L(D_{\boldsymbol{\phi}}, D_{\boldsymbol{\theta}}|\boldsymbol{\phi})$. We may also convert the prior $p_1(\boldsymbol{\theta})$ to $\boldsymbol{\phi}$-space using the Jacobian, obtaining $p_1^*(\boldsymbol{\phi}) = p_1(\boldsymbol{\theta})|\boldsymbol{J}(\boldsymbol{\phi})|$, where $\boldsymbol{\phi} = \boldsymbol{M}(\boldsymbol{\theta})$ and $\boldsymbol{J}(\boldsymbol{\phi})$ is the Jacobian (Bickel and Doksum, 1977). However, at this point there are two priors available for $\boldsymbol{\phi}$: $p_2(\boldsymbol{\phi})$ and $p_1^*(\boldsymbol{\phi})$. These two priors are unlikely to be equivalent because $\boldsymbol{M}$ was not taken into account during their elicitation, and because they represent two independent beliefs based on possibly different evidence. We will call $p_2(\boldsymbol{\phi})$ and $p_1^*(\boldsymbol{\phi})$ *incoherent* if they do not correspond to equivalent probability measures. The goal is to obtain a coherized prior for $\boldsymbol{\phi}$ which accounts for both p_2 and p_1^*. After such coherization has reduced $p_2(\boldsymbol{\phi})$ and $p_1^*(\boldsymbol{\phi})$ to a single prior, say $p(\boldsymbol{\phi})$, then standard Bayesian inference follows through the combination of $p(\boldsymbol{\phi})$ and $L(D_{\boldsymbol{\phi}}, D_{\boldsymbol{\theta}}|\boldsymbol{\phi})$.

2.2 An Introduction to Aggregating Expert Opinions

The problem of coherizing $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\phi})$ can be considered to fall under the framework of aggregating expert opinions, an area reviewed by French (1985) and

Genest and Zidek (1986). In general, a collection of experts, possibly separated by time or space, express opinions (prior distributions) about the same thing. The problem, then, is to pool opinions in a rational way that yields a single probability distribution from which inference can be made. How this might be done may depend on several issues:

- Is there a leader or meta-expert whose task is to complete the pooling and draw inference, and who may be able to assess the relative expertise, calibration, and correlation of the experts?
- Can the experts interact?
- Is there a specific inference problem to be solved? Or is pooling desired so that any (unspecified) arbitrary inference problem may later be solved? Or is no inference planned in the future, so pooling is seen only as a method of summary?

Lindley (1985) argues that pooling requires the concept of a meta-expert or "supra-Bayesian"; see Section 2.2.3. Unlike the sorts of problems Lindley considered, the experts with knowledge relevant to the inputs and outputs of a simulation model may work in diverse disciplines. It is much harder to assess the expertise and correlation of, say, one oceanographer, one physicist, one biologist, and one atmospheric scientist, than it is to assess a group of four scientists all from the same narrow field. For this reason, meta-experts may be hard to find for the sorts of simulation efforts we describe.

When simulation modeling results are used by a group of policymakers, it would be common for the scientific experts to interact only if they are members of that group. Often the experts which would be consulted about the parameters of a simulation model are from diverse disciplines in scattered geographic locations, and therefore they are never assembled in one place where interaction may occur.

In addition, expert opinions are sometimes gathered through literature reviews rather than direct conversations, creating another situation in which interaction is extremely unlikely.

Most simulation modeling exercises are motivated by a specific question, yet analysts are rarely inclined to limit their work to a particular inferential problem. More often, multiple questions are to be answered by a simulation effort, and many questions may not be fully formulated before analysis begins. Thus we focus here on pooling as a method to enable arbitrary, unspecified inferences.

2.2.1 Logarithmic Pooling

At this point, we must determine a sensible method for synthesizing individual probability distributions into a single group probability distribution. Genest (1984) contends that the experts should summarize "...their diverse beliefs using an externally Bayesian prescription. In this way, they would ensure that once new, objective information becomes available, a potential user could update this summary opinion with the same effect as if the experts themselves had observed the data jointly." In other words, Genest maintains that the pooling process should produce the same result from combining all expert priors into a single aggregate prior and then updating with a likelihood as from updating each expert's prior and then merging the resulting individual posterior distributions into a single group posterior. Thus,

$$T\left[\frac{L \cdot p_1(\theta)}{\int L \cdot p_1(\theta)d\theta}, \cdots, \frac{L \cdot p_k(\theta)}{\int L \cdot p_k(\theta)d\theta}\right] = \frac{L \cdot T(p_1, \ldots, p_k)(\theta)}{\int L \cdot T(p_1, \ldots, p_k)(\theta)d\theta}$$

where L is a likelihood function, and T is a pooling operator which produces to a single, coherized opinion. Therefore, $T(p_1, \ldots, p_k) : \Theta \rightarrow [0, 1]$ for each vector of opinions $(p_1, \ldots, p_k)$, where $p_i, i = 1, \ldots, k$, is a probability distribution reflecting expert i's beliefs about θ, and Θ is the set of all possible values of the parameter θ. This condition is termed *external Bayesianity* by Madansky (1978),

"prior to posterior coherency" by Weerahandi and Zidek (1978), and "the data independence property" by McConway (1978).

Genest (1984) proves that any externally Bayesian pooling operator T must be of the form

$$T(p_1(\theta), \ldots, p_k(\theta)) = \frac{\prod_{i=1}^{k} [p_i(\theta)]^{w_i}}{\int \prod_{i=1}^{k} [p_i(\theta)]^{w_i} d\theta}, \tag{2.1}$$

where $w_1, \ldots, w_k \geq 0, \sum_{i=1}^{k} w_i = 1$, and $p_i(\theta)$ is the i^{th} individual prior for some parameter θ. His proof requires just one minor condition: for any $\epsilon > 0$, there exists $A_\epsilon \subset \Theta$ with $0 < \mu(A_\epsilon) < \epsilon$, where Θ is the set of all possible values of the parameter θ and μ is some dominating measure on Θ (e.g. Lebesgue measure). In other words, the group probability distribution at θ must not depend upon densities of values which might have been obtained but were not. In the scenario described in Section 2.1, where two priors $p_2(\phi)$ and $p_1^*(\phi)$ are reduced to a single coherized prior $p(\phi)$, we can express the logarithmic opinion pool as

$$p(\phi) \propto p_2(\phi)^{\alpha} p_1^*(\phi)^{1-\alpha}, \tag{2.2}$$

where $0 < \alpha < 1$.

Another axiomatic approach can yield the logarithmic opinion pool. Genest (1984) shows that pooling operators T which have "relative propensity consistency" (RPC) must be logarithmic pools. Before defining the property of RPC, several quantities must first be defined. Let $\underline{b} = (b_1, \ldots, b_n)$ be a vector of opinions, $b_i : \Theta \to (0, \infty), i = 1, \ldots, n$. Opinions may expressed in many ways, from likelihood functions to significance levels to the method on which we will focus, probability densities. Θ, then, is the collection of items on which opinions are made; the nature of Θ will vary according to the way opinions have been quantified. For instance, Θ might consist of subsets of parameter values, pairs of odds, or discrete events. Let μ and ν be arbitrary elements of Θ, and let $\underline{d} = (\underline{b}, \mu, \nu)$.

Define two functions, RP and H, by $RP(\underline{d}) = \left(\frac{b_1(\mu)}{b_1(\nu)}, \ldots, \frac{b_n(\mu)}{b_n(\nu)}\right)$ and $H(\underline{d}) = \frac{T(\underline{b})(\mu)}{T(\underline{b})(\nu)}$. $RP(\underline{d})$ is the vector of ratios of each expert's opinions about μ and ν (where μ and ν are commonly parameter values or events involving parameter values), and $H(\underline{d})$ is the analogous ratio for the pooled opinion of the experts. Then, the pooling operator T has RPC if and only if

$$RP(\underline{d}_1) \geq RP(\underline{d}_2) \Rightarrow H(\underline{d}_1) \geq H(\underline{d}_2),$$

where $RP(\underline{d}_1) \geq RP(\underline{d}_2)$ refers to componentwise comparisons—i.e., $RP(\underline{d}_1) \geq RP(\underline{d}_2)$ if and only if $\frac{b_1(\mu_1)}{b_1(\nu_1)} > \frac{b_1(\mu_2)}{b_1(\nu_2)}, \ldots, \frac{b_n(\mu_1)}{b_n(\nu_1)} > \frac{b_n(\mu_2)}{b_n(\nu_2)}$. Genest (1984) shows that RPC pooling operators are always of the form in (2.1) as long as Θ contains at least three distinct points.

We focus attention on a special case of relative propensity consistency. If we let $\underline{d}_1 = (\underline{b}, \mu, \nu)$ and $\underline{d}_2 = (\underline{b}, \mu, \mu)$ in the definition above, we obtain the following result: T has RPC if and only if $\frac{b_1(\mu)}{b_1(\nu)} > 1, \ldots, \frac{b_n(\mu)}{b_n(\nu)} > 1$ implies that $\frac{T(\underline{b})(\mu)}{T(\underline{b})(\nu)} > 1$. That is, $T(\underline{b})(\mu) \geq T(\underline{b})(\nu)$ must hold whenever $b_i(\mu) \geq b_i(\nu)$ for $i = 1, \ldots, n$, or if all experts favor an event or set of parameter values μ over a corresponding ν, then the pooled opinion must also favor μ over ν. It is this special case, also called the dominance consistency property or Pareto optimality (Genest *et al.*, 1984), which we will refer to as relative propensity consistency (RPC) throughout the remainder of this dissertation.

As a method for creating a group probability distribution from a collection of individual expert opinions, the logarithmic opinion pool has several additional appealing features. For instance, the logarithmic pool is typically unimodal and less dispersed than other options, it is invariant under rescaling of individual opinions (Genest and Zidek, 1986), and it has a natural-conjugate interpretation (Winkler, 1968).

Still, some authors have found cause to criticize log pooling. French (1985) argues that the selection of weights is arbitrary. In our examples, we take the log pooling parameter to be $\alpha = 0.5$ so that the pooled prior is the geometric mean of the two original priors. In the case of simulation modeling, Raftery *et al.* (1996) point out that results are invariant to relabeling (redefining inputs as outputs, or vice versa) for any α when M is linear, and for $\alpha = 0.5$ when M is nonlinear. In other settings, α might correspond to the relative levels of 'faith' the log pooler has in $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\phi})$.

French (1985) also notes that the honesty of experts can not be verified, and calibration—accounting for an expert's abilities in assessing probability—is ignored. In addition, the logarithmic opinion pool has the characteristic that if any expert believes that a particular parameter value is impossible, then the group probability distribution must reflect a probability of zero for that parameter value. Some feel that this characteristic—the strong version of the zero preservation property—is too strong to be desirable (Genest and Zidek, 1986).

Lindley (1985) contends that external Bayesianity is not desirable in the supra-Bayesian (meta-expert) context (see Section 2.2.3). Lindley claims that a meta-expert's opinion of an expert may change if the meta-expert knows that the expert has acquired an additional piece of knowledge, making external Bayesianity unnecessary. While theoretically correct, Lindley's contention is often not of practical concern. For instance, in the deterministic simulation model setting, the decision maker collects prior opinions from experts on input and output variables, but because she is not an expert herself on any particular parameter, her evaluation of an expert will not change if she learns that the expert has learned another parameter-specific piece of information. If we can make the assumption (which seems reasonable in many settings including the pooling of priors linked by simulation models) that the decision maker's opinion about an expert will not

change if the decision maker learns that the expert has acquired another piece of knowledge, then external Bayesianity is indeed a sensible and desirable property.

2.2.2 Linear Pooling

A simplistic alternative to the logarithmic opinion pool is the linear opinion pool, under which:

$$T(p_1, \ldots, p_k) = \sum_{i=1}^{k} w_i p_i$$

where $w_1, \ldots, w_k \geq 0$ and $\sum_{i=1}^{k} w_i = 1$. Stone (1961) first proposed this pooling operator, although Bacharach (1979) attributes it to Laplace. McConway (1978) shows that "under mild regularity conditions, if the process of finding the consensus distribution is to commute with any possible marginalization of the distributions involved, then a linear opinion pool must be used." Specifically, $T(p_1, \ldots, p_k)$ must be of the form $\sum_{i=1}^{k} w_i p_i$ if

$$T(p_1, \ldots, p_k)(A) = G[p_1(A), \ldots, p_k(A)]$$

for some arbitrary function $G : [0,1]^k \rightarrow [0,1]$, where A is an event on Θ.

Equivalently, French (1985) describes the marginalization property as follows: "Suppose that interest is focused on a sub-σ-field of that over which the group members gave their opinions. It seems intuitively appealing that the same consensus distribution be obtained whether (i) their opinions are first combined into a consensus distribution over the complete σ-field and then a marginal distribution taken or (ii) the group members each give their marginal distributions over the sub-σ-field and a consensus distribution formed from these." However, Lindley (1985) points out that more information is contained under (i) than (ii), so that the equivalence of (i) and (ii) is not necessarily desirable. In addition, the linear pool does not allow one to highly weight an expert with particular expertise over a specific sub-σ-field of interest. Other problems with the linear opinion pool in-

clude the typically multi-modal shape to the group distribution and the absence of RPC or external Bayesianity.

2.2.3 Supra-Bayesian Pooling

Advocates of supra-Bayesian models take the following viewpoint: to a decision maker, the subjective probability assessments solicited from experts about the parameter of interest are random realizations (data) observed by the decision maker. Under this framework, a supra-Bayesian decision maker must be identified. This can be the individual who is soliciting expert opinions to address questions about a certain quantity θ, or it can be a fictitious construct which represents the group's synthetic personality. The supra-Bayesian decision maker is charged with the task of evaluating each of the experts in terms of the anticipated accuracy and precision of his or her projections, in addition to providing her own (possibly noninformative) prior distribution for θ. Then, the supra-Bayesian model produces a posterior distribution for θ in which the decision maker's prior is updated by a likelihood function which includes the experts' prior distributions for θ, the decision maker's evaluations of the experts, and any applicable data collected to provide more information about θ.

Expert Opinion as Data

Cooke (1991) points to a growing body of evidence which describes how expert opinion can be a very useful source of *data*. As a source of data, expert opinion is cheap and plentiful, and ideally it provides a synthesis of many different data sets and knowledge bases. Cooke is careful, however, to draw the distinction between "expert opinion" and "expert knowledge". Techniques which treat expert opinion as data must recognize the inherent uncertainty in assessments provided by experts and account for this uncertainty accordingly. Too often, expert opinion can be regarded as absolute knowledge—as black and white fact—and decisions made on

the basis of expert opinion under this assumption do not reflect the underlying uncertainty which exists.

Supra-Bayesian methods have been developed to treat expert opinion as data providing information which is relevant to a particular decision or unknown quantity, while concurrently recognizing the inherent uncertainty in the expert opinions and accounting for this uncertainty throughout the analysis. In general, these supra-Bayesian models claim to offer solutions to issues like calibration, honesty, correlation, and relative expertise. Those who believe the pooling problem is "solved" generally regard supra-Bayesian approaches as the solution.

However, these approaches are not without difficulty. Supra-Bayesian approaches can be difficult to implement. Most are derived for consideration of one single event, or for a discrete probability space; thus, they are most easily applied in problems with a specific, pre-determined inferential goal. In addition, extensions to continuous variables are usually not immediately clear. Supra-Bayesian approaches are not externally Bayesian, although by treating the expert priors as data rather than probabilities, supra-Bayesian reasoning leads to the conclusion that external Bayesianity is not necessarily a desirable property (see Section 2.2.1). Under supra-Bayesian approaches, one is required to establish hyperpriors describing the calibration and correlation of expert opinions, perhaps expressed on a log odds scale. However, even a rough guess for these hyperpriors can be difficult to provide, particularly when expertise is contributed from diverse fields. While mathematically attractive, the use of hyperpriors to assess expert opinions requires a level and range of knowledge rarely possessed or even available. Finally, these approaches blur the distinction between data and prior.

Variations on the Basic Framework

Many variations on the basic supra-Bayesian framework have been proposed. For example, Winkler (1968) proposed probably the first supra-Bayesian model in which he assumed that the experts' judgments represented sample information which belonged to the natural-conjugate family of distributions. Later, Morris (1977) added a calibration function which provides an assessment of each expert's predictive performance based on variables besides the parameter of interest. Cooke (1990) obtains a set of optimal weights based on calibration and entropy.

Lindley (1985) and French (1985) both propose supra-Bayesian solutions for the case in which θ is finite. Basically, they addresses the question: How should a decision maker update her belief about an event A after learning the opinions of several experts? Each expert is required to give an opinion about A in terms of log-odds. To implement the French-Lindley approach, the decision maker must evaluate the experts' abilities in terms of the log-odds of event A occurring or not occurring.

Winkler (1981) and Lindley (1983) consider continuous θ. They address the question: How should a decision maker update her belief about a parameter θ after learning the opinions of several experts? For Lindley's version, each expert is required to give the first two moments describing his or her belief about θ, although a particular distributional form is not required. Then, implementation of the Lindley-Winkler model requires that the decision maker evaluate the experts' abilities using a set of hyperparameters, including the covariances between experts.

In the remainder of this section, we survey the early supra-Bayesian approaches of Winkler (1968) and Morris (1977), and then in Chapters 5 and 6 the models of French-Lindley and of Lindley-Winkler are explored in much greater detail.

Winkler's Natural Conjugate Approach

The supra-Bayesian approach has roots in Winkler (1968). Winkler assumed that the experts' judgments represented sample evidence which could be described by the natural-conjugate family of distributions, so that they could easily be combined with the decision-maker's prior distribution using successive applications of Bayes' Theorem. As Winkler states, the experts' opinions ". . . can be thought of as being equivalent to sample information from the data-generating process of interest," a concept he illustrates using a Bernoulli data-generating process and a natural-conjugate family of beta distributions. While creating Bayesian probabilities which are more mathematically convenient, this method is valuable only if the experts' judgments can be well-expressed in terms of a probability distribution from the natural-conjugate family. In addition, the decision maker who solicited the expert opinions must be willing to assign weights which reflect the strength of each expert and the degrees of dependencies between the experts.

Morris' Calibration Approach

Morris (1977) summarizes his basic result as follows: "If the expert is calibrated so that he is an accurate probability assessor, then the decision maker's posterior equals the normalized product of the expert's prior and the decision maker's own prior." Thus:

$$p(\theta \mid x, d) = k \cdot h(\theta) \cdot \pi(\theta \mid d),$$

where d is the decision maker's prior state of information, $p(\theta \mid x, d)$ represents the posterior distribution for the decision maker, $\pi(\theta \mid d)$ is the decision maker's prior, and $h(\theta)$ is the composite prior for the experts, which involves the product of expert priors times a correction factor for the degree of dependence and the lack of calibration among experts. This model requires two assumptions, which

are later relaxed. The first is an invariance to scale, so that an expert's self-evaluated confidence in his predictive ability provides no information about θ. Second, an invariance to shift is assumed, so that the decision maker's assessment of an expert's bias does not depend on the true value of θ.

Core to Morris' model is the ability of the decision maker to calibrate the experts' priors by adjusting experts' probability assessments to reflect measured assessment ability. Morris recommends using a Calibration Function, which can be determined subjectively, or, even better, from an empirical summary of an expert's past probability assessment record. Thus, good assessment of calibration, according to Morris, requires that all experts provide probability assessments on the same set of variables for which the true values are or will be known. A weakness of this method is its application to multiple, dependent experts because of the difficulties in accurately assessing expert dependence.

Besides providing a general inferential framework for supra-Bayesian models, Morris' approach leads to a couple of interesting results. First, he proves the existence of a composite prior which measures the joint information contained in the subjective probability distributions of a collection of experts. In fact, the composite prior is a sufficient statistic for the information contained in the experts' probability assessments. Second, with independent and calibrated experts, Morris' approach reduces to a weighting scheme similar to that found in Winkler's natural conjugate approach when the experts and the decision maker all have priors from the same natural conjugate family.

2.2.4 Group Interaction Pooling

Under this behavioral model, the experts interact with each other, either face-to-face or through a distributed set of summary information, and they are allowed to revise their individual probabilities based on feedback. For example,

Winkler (1968) compares two such approaches. In the feedback and reassessment technique, each expert is presented with probability distributions from other experts as feedback, although no direct contact is allowed. In the group reassessment technique, the experts are gathered together and allowed to discuss the problem. The group can then form a single group probability distribution either while still gathered together or after individually updating their beliefs following the group discussion. Cox (1991) advocates group reassessment models over aggregation formulae because "communication that allows agents to exchange *explanations* for their opinions, rather than only numerical *strengths* of opinions (or of supports for beliefs, etc.), can improve both the efficiency of the consensus-building process and the quality of the common-knowledge base finally achieved."

One difficulty with direct interaction is that it can be challenging for experts from different fields to translate their knowledge to others, and to revise their beliefs based on evidence that is foreign to them. However, such efforts can also be very enlightening.

Another difficulty arises when the model is too complex to permit immediate, comprehensible feedback to assist those who want to modify opinions through interaction to achieve consensus. However, when such interactive prior transformations are successfully carried out, it can enable the experts to critique the model itself. In this way, prior coherization is related to issues of model validation and improvement.

Unfortunately, when pooling priors linked by a deterministic simulation model, group interaction models are rarely feasible. When Bayesian inference is desired about a parameter in a determinstic simulation model, often a single person (or group) with knowledge of the model will solicit expert opinion on each input and output parameter, obtaining not merely point estimates, but complete probability distributions which also reflect the experts' uncertainty. The expert priors

must then be pooled into a single prior about the parameter of interest. The first problem with using a group interaction method for this pooling is that it is rarely possible to gather the experts, scattered across scientific fields and geographic locations, together at once. Moreover, even if the experts were to meet, it would be difficult to focus the discussion on the important parameter. For instance, an "input expert" may not even be aware of the model which links his opinion to the outputs, and without good understanding of the model would contribute little to a discussion about outputs. In addition, many deterministic simulation models which are used to model real systems are large and highly complex, so that immediate feedback, such as in the case when an "output expert" changes his opinion slightly and wants to know the impact on the inputs, is an impossibility.

2.3 Summary and Preview

In Section 2.2, several theoretical properties were mentioned as a means for evaluating and comparing strategies for pooling expert opinion. Here we summarize those properties which we feel are important to consider when evaluating potential approaches for pooling priors linked by deterministic simulation models.

External Bayesianity demands that a pooling process produce the same consensus distribution regardless of the order of pooling and updating upon the receipt of new information. Even though Lindley (1985) argued from a supra-Bayesian perspective that external Bayesianity is not always desirable, we believe that under many pooling scenarios, including priors linked by deterministic simulation models, external Bayesianity is indeed intuitively sensible and desirable (see Section 2.2.1).

Relative propensity consistency , in the special case we are considering, demands that the pooled distribution favor ϕ_1 over ϕ_2 whenever each expert

favors ϕ_1 over ϕ_2. We believe that any reasonable pooling strategy should exhibit RPC, otherwise it would be difficult to place faith in the pooled results.

The zero preservation property has two forms—strong and weak. Strong ZPP implies that $p(\phi) = 0$ whenever $p_i(\phi) = 0$ for *any* expert i, while weak ZPP requires that $p(\phi) = 0$ whenever $p_i(\phi) = 0$ for *every* expert i. We feel that strong ZPP can be overly restrictive, especially in deterministic simulation model settings when some experts do not even express their beliefs in the parameter space of interest. If an "input expert", for instance, knew that his prior beliefs, when transformed by the model into output space, expressed no probability for certain reasonable output values, then the input expert would probably revisit his beliefs or the model itself. Weak ZPP, on the other hand, is intuitively reasonable and desirable, because if all experts feel that a particular parameter value is impossible, then the pooled prior has no basis for assigning non-negative probability at that point.

Based on the theoretical properties of the potential pooling approaches discussed in this chapter, we favor creating consensus priors via logarithmic pooling. Chapter 3 will outline a detailed algorithm for obtaining log pooled priors in the deterministic simulation model setting. A theoretically and mathematically attractive method, while possessing a nice foundation, still must perform well in realistic situations to be a truly useful method to the practitioner. In Chapters 4-7, therefore, we evaluate and compare logarithmic pooling and alternative pooling methods (linear pooling and two supra-Bayesian methods) over several examples. In each example, priors are linked by a deterministic simulation model, and the examples cover a range of complexity levels.

One point yet to be discussed is the strategy of pooling information which conflicts, given a deterministic model. Is it reasonable to average information which is not in agreement when expressed in the same parameter space? All of the mathematical pooling formulae discussed here are essentially averages. We feel that routine application of pooling formulae to obtain inference from simulation models is risky unless it is part of a larger effort that includes model validation and revision, and interaction between experts. Priors which are severely incoherent should not be coherized merely because they can be. Such cases can only be resolved when the experts can revise the model and their own opinions. When only mild incoherence is present, the averaging of opinion through pooling is more reasonable. Such philosophical issues, plus the computational challenges to implementing a pooling approach, point to the considerable potential for future development in this area. We expect that deterministic models will continue to be used in diverse areas of scientific inquiry, and techniques like those described here can encourage disciplined statistical thinking and uncertainty analysis in deterministic modeling.

Chapter 3

Log Pooling via Adaptive Importance Sampling

3.1 A General Algorithm

Consider a deterministic simulation model $\boldsymbol{M}(\boldsymbol{\theta})$: $\boldsymbol{\theta} \rightarrow \boldsymbol{\phi}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \Re^m$ and $\boldsymbol{\phi} \in \boldsymbol{\Phi} \subseteq \Re^p$. Prior distributions $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\phi})$ are elicited independently on the inputs and outputs, and we wish to make inference about certain model parameters, say $\boldsymbol{\phi}$. The basic approach is to convert $p_1(\boldsymbol{\theta})$ to $\boldsymbol{\phi}$–space, to coherize the priors with log pooling, then to make inference about $\boldsymbol{\phi}$.

When $\boldsymbol{M}$ is too complex to invert or to solve for the implicit priors, only $p_1(\boldsymbol{\theta})$, $p_1^*(\boldsymbol{\phi})$ and $p_2(\boldsymbol{\phi})$ may be sampled, and only $p_1(\boldsymbol{\theta})$, $p_2(\boldsymbol{\phi})$ and the likelihoods may be evaluated. Note that $p_1^*(\boldsymbol{\phi})$ may be sampled by sampling from $p_1(\boldsymbol{\theta})$ and applying $\boldsymbol{M}$ to the sample.

We estimate the posterior distribution for $\boldsymbol{\phi}$, $\pi(\boldsymbol{\phi} | D) \propto p(\boldsymbol{\phi}) L(D_{\boldsymbol{\phi}}, D_{\boldsymbol{\theta}} | \boldsymbol{\phi})$, using the steps described below. Obtaining the coherized prior $p(\boldsymbol{\phi})$ relies upon (possibly adaptive) selection of a suitable importance sampling function in $\boldsymbol{\theta}$-space whose mapping covers the support of both $p_2(\boldsymbol{\phi})$ and $p_1^*(\boldsymbol{\phi})$.

Assume $\boldsymbol{M}$ is a smooth, continuous mapping. A general algorithm for estimating the log pooled prior distribution $p(\boldsymbol{\phi})$ (and subsequently the posterior distribution $\pi(\boldsymbol{\phi} | D)$) is as follows:

1. Let $e(\boldsymbol{\theta})$ be a density function that induces a density function in $\boldsymbol{\phi}$-space that is a suitable importance sampling function for the log pooled prior. In

some cases, $e(\boldsymbol{\theta})=p_1(\boldsymbol{\theta})$ will suffice. We discuss strategies for choosing $e(\boldsymbol{\theta})$ in Section 3.2.

2. Sample $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n \sim$ i.i.d. $e(\boldsymbol{\theta})$.

3. Calculate $\boldsymbol{\phi}_i = \boldsymbol{M}(\boldsymbol{\theta}_i)$ for $i = 1,\ldots,n$.

4. Find a density estimate for $p_1^*(\boldsymbol{\phi})$ using $\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_n$. Although many methods work adequately in some situations, histogram methods (see Westergaard, 1968) are inherently non-smooth and unwieldy in higher dimensions, and kernel density methods (Parzen, 1962; Rosenblatt, 1956) can lead to severely biased importance weights for densities with bounded support. We have had the most success by numerically approximating the Jacobian, so $\widehat{p_1^*(\boldsymbol{\phi}_i)} = p_1(\boldsymbol{\theta}_i) \cdot \left|\widehat{\boldsymbol{J}(\boldsymbol{\phi}_i)}\right|$, where $\boldsymbol{\phi}_i = \boldsymbol{M}(\boldsymbol{\theta}_i)$. See Sections 3.3 and 3.4 for methodological details and consideration of noninvertible $\boldsymbol{M}$.

5. Find $\widehat{e^*(\boldsymbol{\phi}_i)} = e(\boldsymbol{\theta}_i) \cdot \left|\widehat{\boldsymbol{J}(\boldsymbol{\phi}_i)}\right|$ using the same methods as in step 4.

6. Obtain a representative sample from the coherized prior $p(\boldsymbol{\phi}) \propto p_2(\boldsymbol{\phi})^{\alpha} p_1^*(\boldsymbol{\phi})^{1-\alpha}$. Since this distribution is usually too difficult to sample from directly, we use importance weighting (Hammersley and Hanscomb, 1964) to represent $p(\boldsymbol{\phi})$. The estimated importance weights are $w_j \propto p_2(\boldsymbol{\phi}_j)^{\alpha} \widehat{p_1^*(\boldsymbol{\phi}_j)}^{1-\alpha} \Big/ \widehat{e^*(\boldsymbol{\phi}_j)}$ for $j = 1,\ldots,n$. Then, the $\boldsymbol{\phi}_j$ and their corresponding w_j represent $p(\boldsymbol{\phi})$. For further details on importance sampling, see Section 3.2.1.

7. Using additional importance weighting, combine $p(\boldsymbol{\phi})$ with the likelihood function $L(D_{\boldsymbol{\phi}}, D_{\boldsymbol{\theta}}|\boldsymbol{\phi})$ to obtain the posterior $\pi(\boldsymbol{\phi} \mid D) \propto p(\boldsymbol{\phi})L(D_{\boldsymbol{\phi}}, D_{\boldsymbol{\theta}}|\boldsymbol{\phi})$. This also permits inference regarding other quantities, say $\boldsymbol{\psi} = F(\boldsymbol{\theta},\boldsymbol{\phi})$ for some function F.

A similar algorithm could be followed to make inference about $\boldsymbol{\theta}$ using the posterior distribution $\pi(\boldsymbol{\theta} \mid D) \propto p_1(\boldsymbol{\theta})^{\alpha} \widehat{p_2^*(\boldsymbol{\theta})}^{1-\alpha} L(D_{\boldsymbol{\theta}}, D_{\boldsymbol{\phi}} | \boldsymbol{\theta})$. When $\boldsymbol{M}$ is invertible, we can proceed using $\widehat{p_2^*(\boldsymbol{\theta}_i)} = p_2\left(\boldsymbol{M}(\boldsymbol{\theta}_i)\right)\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|$, where the estimation of the Jacobian is performed according to the methods described in Section 3.3. The log pooled prior and posterior can be estimated using importance sampling, as in steps 6 and 7. When $\boldsymbol{M}$ is noninvertible, $p_2^*(\boldsymbol{\theta})$, and hence the log pooled prior, is not uniquely defined. Described in Section 3.4 is an approach for proceeding in this case; essentially, each possible inversion of $p_2(\boldsymbol{\phi})$ is equally weighted to reflect the $\boldsymbol{\phi}$-expert's ignorance about $\boldsymbol{\theta}$.

The effectiveness of this algorithm depends strongly on the quality of the importance sampling function, $e(\boldsymbol{\theta})$. In Section 3.2.3 we describe the details of a cumulative, adaptive mixture strategy for constructing a suitable $e(\boldsymbol{\theta})$. A simple, nonadaptive alternative would be to set $e(\boldsymbol{\theta})$=$p_1(\boldsymbol{\theta})$. Section 3.2.3 also mentions higher dimensional problems and cases when $m \neq p$.

3.2 Importance Sampling Function, $e(\boldsymbol{\theta})$

3.2.1 Importance Sampling

The SIR (Sampling/Importance Resampling) algorithm is, according to Rubin (1988), "an ubiquitously applicable noniterative algorithm for obtaining draws from an awkward distribution." The SIR algorithm was first proposed by Rubin in 1983 and more fully described in 1987 and 1988 articles by Rubin . In order to obtain a sample of size m from an unwieldy distribution $p(\phi)$, the SIR algorithm follows these steps:

1. obtain a sampling envelope $h(\phi)$ which (to some extent) approximates $p(\phi)$, and for which $h(\phi) > 0$ for all possible ϕ.

2. let $\phi_1, \ldots, \phi_M$ be a random sample from $h(\phi)$, where M is large relative to m.

3. let $w_j \propto p(\phi_j)/h(\phi_j)$ be the importance weight for $\phi_j, i = 1, \ldots, M$.

4. let $\phi'_1, \ldots, \phi'_m$ be a random sample from $\phi_1, \ldots, \phi_M$ with probabilities proportional to w_j.

Rubin (1988) shows that as $\frac{M}{m} \to \infty$, the points $\phi'_1, \ldots, \phi'_m$ are independent selections from $p(\phi)$ as desired. The choice of a practical ratio $\frac{M}{m}$ to make the approximation adequate depends on the adequacy of the sampling envelope $h(\phi)$.

3.2.2 A Nonadaptive Approach

The easiest approach is to let $e(\boldsymbol{\theta}) = p_1(\boldsymbol{\theta})$. Note that, in this case, $w_i \propto p_2(\boldsymbol{\phi}_i)^{\alpha}/\widehat{p_1^*(\boldsymbol{\phi}_i)}^{\alpha}$. If the support of $p_1(\boldsymbol{\theta})$ covers the support of $p_2^*(\boldsymbol{\theta})$, then the weighted empirical distribution of the sample converges to the log pooled distribution as the number of samples $n \to \infty$ (Hammersley and Hanscomb, 1964). However, even when this is true, there is a question of efficiency. If $p_1^*(\boldsymbol{\phi})$ and $p_2(\boldsymbol{\phi})$ are not very similar, sampling from $p_1^*(\boldsymbol{\phi})$ is not an efficient way to obtain points which represent $p_2(\boldsymbol{\phi})$, or, hence, $p(\boldsymbol{\phi})$. Efficiency is an important concern when the complexity of $\boldsymbol{M}$ or its Jacobian limits n. We therefore introduce an adaptive approach which periodically improves $e(\boldsymbol{\theta})$ while accumulating an effective and efficient importance sample. The increased complexity of the adaptive approach can sometimes be offset by even greater gains in sampling efficiency and hence in estimation performance when n is limited.

3.2.3 An Adaptive Approach

The adaptive envelope strategy attempts to determine a sample of points which is representative of the prior opinions of both experts. This is done by starting with a sample of points from $p_1(\boldsymbol{\theta})$ in the first iteration and then searching

for points which would map into $p_2(\boldsymbol{\phi})$ in the remaining iterations. Therefore, we adopt the strategy of aiming for an importance sampling function in $\boldsymbol{\theta}$-space that, when transformed into $\boldsymbol{\phi}$-space, can be used for importance sampling of $p_2(\boldsymbol{\phi})$. The rationale for this goal is that if one component of the mixture is $p_1(\boldsymbol{\theta})$, then such a function should also be adequate for sampling the log pooled prior. At each iteration, the current mixture is tested for adequacy, and additional mixture components are added in the most problematic regions until a suitable $e(\boldsymbol{\theta})$ is obtained. An importance weighted sample of $(\boldsymbol{\theta},\boldsymbol{\phi})$ pairs is cumulatively gathered and reweighted at each stage, so that the results are immediate when the final importance sampling function is identified.

Case 1: 1 input and 1 output (m = p = 1)

Begin with $e_1(\theta) = p_1(\theta)$. Take a random sample of size m_1 from $e_1(\theta)$. Use M to transform these points into ϕ-space, where we calculate importance weights

$$u_i = \frac{p_2(\phi_i)/\widehat{e_1^*(\phi_i)}}{\sum_{i=1}^{m_1} p_2(\phi_i)/\widehat{e_1^*(\phi_i)}},$$

for $i = 1, \ldots, m_1$. As described in Sections 3.3 and 3.4, $\widehat{e_1^*(\phi_i)}$ is based on a Jacobian estimate.

At this point, the true cumulative distribution function, F_{p_2}, corresponding to $p_2(\phi)$ is compared with the weighted empirical estimate, $\widehat{F}_{p_2}(\phi_{j:m_1}) = \sum_{i=1}^{j} u_i$, where $\phi_{j:m_1}$ represents the j^{th} order statistic from the collection of ϕ values, and the u_i are sorted in order of the ϕ_i. We test whether $\widehat{F}_{p_2}$ satisfies a goodness of fit criterion relative to F_{p_2} with the Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1939).

Instead of computing the p-value with the actual sample size, we use an 'effective sample size' which is more appropriate because it corrects for importance sampling inefficiency. The effective sample size does not count ϕ_i that have negligible u_i because they effectively contribute no information to the estimate $\widehat{F}_{p_2}$.

The effective sample size is the minimum of the actual sample size and an adjusted sample size, calculated as the number of ϕ_i in a 95% $p_2(\phi)$ probability interval, divided by .95. With this estimation of the number of ϕ_i that have non-negligible $p_2(\phi)$ density, those ϕ_i that contribute little to $\widehat{F}_{p_2}$ are ignored.

If the p-value from the K–S test is greater than .05, we stop and conclude that the importance sampling function $e_1(\theta) = p_1(\theta)$ is suitable because $\widehat{F_{p_2}}(\phi)$ is very similar to $F_{p_2}(\phi)$. However, if the p-value is below .05, we take an additional sample from θ-space. These additional points are stochastically selected from regions in θ-space where the difference between $\widehat{F_{p_2}}(\phi)$ and $F_{p_2}(\phi)$ is greatest. Specifically, the i^{th} observation in the additional sample taken at the 2^{nd} iteration is distributed according to

$$\theta_{2,i} \sim e_2(\theta) = \sum_{r=1}^{m_1} b_{1,r} N(\theta_{1,r}, v_1)$$

for $i = 1, \ldots, m_2$, where

$$b_{1,r} = p_2(\phi_{1,r}) \cdot \left| \widehat{F}_{p_2}(\phi_{1,r}) - F_{p_2}(\phi_{1,r}) \right| \Big/ \sum_{r=1}^{m_1} p_2(\phi_{1,r}) \cdot \left| \widehat{F}_{p_2}(\phi_{1,r}) - F_{p_2}(\phi_{1,r}) \right|,$$

and v_1 is a weighted estimate of the variance of $p_2^*(\theta)$ after the first iteration.

Thus, $e_2(\theta)$ is a mixture distribution of m_1 normal kernels centered at the points $\theta_{1,1}, \ldots, \theta_{1,m_1}$, weighted by how inadequately $p_2(\phi)$ is represented at $\phi_{1,r} = M(\theta_{1,r})$. The overall $e(\theta)$ becomes the weighted mixture of the two densities $e_1(\theta)$ and $e_2(\theta)$, since the points from all previous iterations are mixed with the points from the current iteration to form a combined current sample.

Subsequent $e_j(\theta)$ are defined analogously, using all information available at iterations $1, \ldots, (j-1)$. In general, at the j^{th} iteration, $e(\theta)$ is the cumulative mixture distribution defined by $e(\theta) = \left(\sum_{i=1}^{j} m_i e_i(\theta) \right) \Big/ \sum_{i=1}^{j} m_i$.

The algorithm continues until a mixture importance sampling function is found that produces a weighted empirical distribution function that is not rejected

by the goodness-of-fit test when compared to F_{p_2}. At this point, we conclude that the current mixture distribution for θ is a suitable $e(\theta)$. Therefore, the sample from $e(\theta)$, transformed to ϕ-space and paired with the corresponding importance weights $w_i \propto p_2(\phi_i)^{\alpha} \widehat{p_1^*(\phi_i)}^{1-\alpha} \Big/ \widehat{e^*(\phi_i)}$ should produce a reasonable estimate for $p(\phi)$.

Case 2: Higher Dimensions

Suppose $m = p > 1$. Denote by $F_{p_2,i}(\boldsymbol{\phi}_j), i = 1, \ldots, p$, the marginal cumulative distribution functions corresponding to the p dimensions of $\boldsymbol{\phi}$, evaluated at $\boldsymbol{\phi}_j$.

We treat the multivariate case analogously to Case 1. An initial sample is taken from $p_1(\boldsymbol{\theta})$, and importance weights u_i are again calculated as in Case 1. In this case, $\widehat{e^*(\boldsymbol{\phi})}$ is obtained using multivariate Jacobian estimation methods as described in Sections 3.3 and 3.4. Then p univariate marginal Kolmogorov-Smirnov tests are performed to check whether the weighted empirical distribution functions, $\widehat{F}_{p_2,i}(\boldsymbol{\phi}_{j:m_1})$, resemble $F_{p_2,i}(\boldsymbol{\phi}), i = 1, \ldots, p$. As before, the goodness-of-fit test relies on effective sample size, which is based on the number of ϕ's within the corresponding multivariate 95% probability interval. If any of these tests reject goodness of fit at a .05 significance level, then an additional sample is added to the mixture. Although multiple significance tests are being used, no adjustment is made because these p-values are being used as summary statistics, providing information on how close our empirical estimates of $\widehat{F}_{p_2,i}$ is to $F_{p_2,i}$.

A potential alternative to the series of marginal goodness-of-fit tests is the application of projection pursuit methods. Under projection pursuit density estimation (Friedman *et al.*, 1984), multivariate functions are estimated by combinations of univariate functions of carefully selected linear combinations of the variables. Projection pursuit density estimates of $p_2(\boldsymbol{\phi})$ are of the form: $p_{2M}(\boldsymbol{\phi}) =$

$p_{20}(\boldsymbol{\phi}) \prod_{m=1}^{M} f_m(\alpha_m \cdot \boldsymbol{\phi})$, where M is the number of iterations of the procedure, $p_{20}(\boldsymbol{\phi})$ is an initial estimate for the multivariate density, α_m is a vector specifying a direction in $\Re^p$ at which relative goodness of fit (compared to the model at the previous iteration) is maximized, $\alpha_m \cdot \boldsymbol{\phi} = \sum_{i=1}^{p} \alpha_{mi} \cdot \phi_i$ is a linear combination of the original coordinates, and f_m is a univariate function. The idea of projection pursuit is worthy of future study for its utility in finding adaptive importance samples in higher dimensions.

The new mixture component for the second iteration is itself a mixture distribution of m_1 multivariate normal kernels centered at the points $\boldsymbol{\theta}_{1,1}, \ldots, \boldsymbol{\theta}_{1,m_1}$ and weighted by how inadequately $p_2(\boldsymbol{\phi})$ is represented at $\boldsymbol{\phi}_{1,r} = \boldsymbol{M}(\boldsymbol{\theta}_{1,r})$. If we denote by $\widehat{F_{p_2}}$ and F_{p_2} the multivariate empirical and true cdfs, then the mixture weights $b_{1,r}$ can be expressed as

$$b_{1,r} = p_2(\boldsymbol{\phi}_{1,r}) \cdot \left| \widehat{F}_{p_2}(\boldsymbol{\phi}_{1,r}) - F_{p_2}(\boldsymbol{\phi}_{1,r}) \right| \Big/ \sum_{k=1}^{m_1} p_2(\boldsymbol{\phi}_{1,k}) \cdot \left| \widehat{F}_{p_2}(\boldsymbol{\phi}_{1,k}) - F_{p_2}(\boldsymbol{\phi}_{1,k}) \right| .$$

$\widehat{F_{p_2}}$ is calculated by summing the appropriate $u_i \propto p_2(\boldsymbol{\phi}_i)/e^*(\widehat{\boldsymbol{\phi}_i})$. Each multivariate normal kernel has covariance matrix equal to the weighted sample covariance matrix of the $\boldsymbol{\theta}_i$ with weights proportional to $p_2^*(\widehat{\boldsymbol{\theta}_i})/e(\boldsymbol{\theta}_i)$. Additional components are accumulated until no marginal empirical distribution function $\widehat{F_{p_2,i}}, i = 1, \ldots, p$, differs significantly from the truth, at which point $e(\boldsymbol{\theta})$ is taken to be the resulting mixture distribution. A sample from $e(\boldsymbol{\theta})$, transformed to $\boldsymbol{\phi}$-space and paired with the corresponding importance weights $w_j \propto p_2(\boldsymbol{\phi}_j)^{\alpha} p_1^*(\widehat{\boldsymbol{\phi}_j})^{\,1-\alpha} \Big/ e^*(\widehat{\boldsymbol{\phi}_j})$ should produce a reasonable estimate for $p(\boldsymbol{\phi})$.

Case 3: Higher, Unequal Dimensions

It should be possible to reduce such problems to Case 2 by the introduction of additional nuisance parameters. Raftery and Poole (1998) present a generalization for invertible models with $m \neq p$ which has attractive invariance properties. They describe a variation of log pooling which is suitable for such models and implement

it using a nonadaptive importance sampling strategy. We do not consider this case further here.

3.3 Jacobians $J(\theta_i)$ and $J(\phi_i)$ with Invertible M

Let $|\boldsymbol{J}(\boldsymbol{\theta}_i)|$ denote the Jacobian of $\boldsymbol{\phi}$ with respect to $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}_i$. Similarly, let $|\boldsymbol{J}(\boldsymbol{\phi}_i)|$ denote the Jacobian of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\phi}$, evaluated at $\boldsymbol{\phi}_i$. Then, for invertible $\boldsymbol{M} : \Re^1 \rightarrow \Re^1$, the estimated Jacobian $\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|$ is obtained using a discrete difference approach iterated to a prespecified relative convergence criterion. Under this approach $|\boldsymbol{J}(\boldsymbol{\theta}_i)|$ is estimated with $\left|\widehat{\frac{d\phi}{d\theta_i}}\right| \approx \frac{M(\theta_i+\epsilon_{ij})-M(\theta_i)}{\epsilon_{ij}}$. An iterative selection procedure is used to select ϵ_{ij}, where as the iteration number j increases, smaller and smaller ϵ_{ij} are used until $\left|\frac{\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|_j - \left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|_{j-1}}{\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|_{j-1}}\right| < \tau$ for some predefined convergence limit τ, where $\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|_j$ is the estimate for $|\boldsymbol{J}(\boldsymbol{\theta}_i)|$ based on ϵ_{ij}. If the convergence limit is not reached within a specified number of iterations, then we choose ϵ_{ij} such that $\left|\frac{\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|_j - \left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|_{j-1}}{\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|_{j-1}}\right|$ is minimized. For higher-dimensional problems, $\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|$ can also be calculated using discrete differences.

With invertible $\boldsymbol{M} : \Re^1 \rightarrow \Re^1$, $\left|\widehat{\boldsymbol{J}(\boldsymbol{\phi}_i)}\right|$ is merely the inverse of $\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|$. In higher dimensions, estimation of $|\boldsymbol{J}(\boldsymbol{\phi}_i)|$ is more complex because the coordinate-wise discrete perturbations can not be implemented in $\boldsymbol{\phi}$-space.

Instead, directional derivatives can be employed. For instance, in two dimensions, we can introduce discrete perturbations in $\boldsymbol{\theta}$-space, then solve for $\widehat{\frac{d\theta_1}{d\phi_1}}$ and $\widehat{\frac{d\theta_1}{d\phi_2}}$ according to:

$$a_{11}(\theta_1) = a_{12} \cdot \widehat{\frac{d\theta_1}{d\phi_1}} + a_{13} \cdot \widehat{\frac{d\theta_1}{d\phi_2}}$$

$$a_{21}(\theta_1) = a_{22} \cdot \widehat{\frac{d\theta_1}{d\phi_1}} + a_{23} \cdot \widehat{\frac{d\theta_1}{d\phi_2}}$$

where $a_{11}(\theta_1)$ is the directional derivative of θ_1 at $\boldsymbol{\phi}=\boldsymbol{M}(\boldsymbol{\theta})$ in the direction $a_{12}\boldsymbol{i}+a_{13}\boldsymbol{j}$, and a_{12} and a_{13} are the changes on the ϕ_1- and ϕ_2-axes, respectively, for a

1-unit change along the vector between $\boldsymbol{M}(\theta_1 + \epsilon_1, \theta_2)$ and $\boldsymbol{M}(\theta_1, \theta_2)$. Thus,

$$a_{11}(\theta_1) = \frac{\theta_1 + \epsilon_1 - \theta_1}{||\boldsymbol{M}(\theta_1 + \epsilon_1, \theta_2) - \boldsymbol{M}(\theta_1, \theta_2)||},$$

if $||\cdot||$ denotes Euclidean distance between points. a_{21}, a_{22}, and a_{23} are analogous for a perturbation of θ_2 rather than θ_1. $\widehat{\frac{d\theta_2}{d\phi_1}}$ and $\widehat{\frac{d\theta_2}{d\phi_2}}$ can be found using a similar set of equations. Finally, the Jacobian can be calculated using these estimates of the partial derivatives. A relative convergence criterion is used to choose the magnitude of discrete perturbations like ϵ_1.

3.4 Estimation under Noninvertible $\boldsymbol{M}$

3.4.1 Mapping $\boldsymbol{\phi} \rightarrow \boldsymbol{\theta}$

When $\boldsymbol{M}$ is noninvertible, we assume that there exists a partition of Θ such that $\Theta = \Theta_1 \cup \ldots \cup \Theta_r$ and $\Theta_i \cap \Theta_j = \emptyset, \forall i \neq j$, where $\boldsymbol{M}$ is invertible on $\Theta_i, \forall i = 1, \ldots, r$. In this case, $p_2^*(\boldsymbol{\theta})$ is not uniquely defined. If $p_2^{*(i)}(\boldsymbol{\theta})$ is the transformation of $p_2(\boldsymbol{\phi})$ onto Θ_i, let $p_2^*(\boldsymbol{\theta}_j) \equiv p_2^{*(i)}(\boldsymbol{\theta}_j)/r$. With this definition, we can estimate $p_2^*(\boldsymbol{\theta})$ using $\widehat{p_2^*(\boldsymbol{\theta}_i)} = p_2\left(\boldsymbol{M}(\boldsymbol{\theta}_i)\right)\left|\widehat{\boldsymbol{J}(\boldsymbol{\theta}_i)}\right|/r$. This definition for $p_2^*(\boldsymbol{\theta})$ with noninvertible $\boldsymbol{M}$, while somewhat arbitrary, is logically sensible. A certain $\boldsymbol{\phi}$-value may have been produced by several different $\boldsymbol{\theta}$-values, yet the $\boldsymbol{\phi}$-expert likely has no information about which $\boldsymbol{\theta}$ produced a particular $\boldsymbol{\phi}$. Therefore, this definition gives each potential $\boldsymbol{\theta}$-value equal weight.

3.4.2 Mapping $\boldsymbol{\theta} \rightarrow \boldsymbol{\phi}$

Suppose, as before, there exists a partition of Θ such that $\Theta = \Theta_1 \cup \ldots \cup \Theta_r$ and $\Theta_i \cap \Theta_j = \emptyset, \forall i \neq j$, where $\boldsymbol{M}$ is invertible on $\Theta_i, \forall i = 1, \ldots, r$. Then we can uniquely define $p_1^*(\boldsymbol{\phi}) = \sum_{i=1}^r p_1(\boldsymbol{\theta}_i) \cdot |\boldsymbol{J}(\boldsymbol{\phi}_i)|$, where $\boldsymbol{\phi} = \boldsymbol{M}(\boldsymbol{\theta}_i)$ for $i = 1, \ldots, r$. However, although we may know that $\boldsymbol{M}$ is a many-to-one mapping, we may not easily be able to identify $\Theta_1, \Theta_2, \ldots, \Theta_r$. One way to estimate $p_1^*(\boldsymbol{\phi}_i)$ in

this instance is to take a weighted sum of all estimates of $p_1^*(\phi)$ for ϕ's in a neighborhood of ϕ_i, where the weights are inversely proportional to the probability mass near θ. Thus, our estimate is:

$$\widehat{p_1^*(\phi_i)} \propto \frac{\sum_{j \in J} p_1(\theta_j) \left| \widehat{\boldsymbol{J}(\phi_j)} \right| \widehat{w(\theta_j)}}{\sum_{j \in J} \widehat{w(\theta_j)}},$$

where, when $m = p = 1$, $\phi_j = M(\theta_j)$, J = the set of subscripts identifying the k nearest neighbors of ϕ_i, θ_j = θ-value which produced ϕ_j, and $\widehat{w(\theta_j)} = 1/(e(\theta_j)\,|\theta_j - \theta_{j'}|)$, where j' indexes the nearest neighbor to θ_j. In higher dimensions, analogous estimates are used. For instance, if $\boldsymbol{M}$: $\Re^2 \to \Re^2$, then $|\theta_j - \theta_{j'}|$ is replaced by the area of the rectangle whose vertices are $(\theta_{j,1}, \theta_{j',2})$, $(\theta_{j',1}, \theta_{j',2})$, $(\theta_{j',1}, \theta_{j,2})$, and $(\theta_{j,1}, \theta_{j,2})$. If r is small and the number of neighbors used for the calculation of $p_1^*(\phi)$ is sufficiently large so that the sampled proportions of each inverse image of the ϕ-neighborhood approximate the $p_1(\theta)$ mass corresponding to each inverse image, then $\widehat{p_1^*(\phi)}$ will be a good estimate of $p_1^*(\phi)$.

3.5 Convergence

Suppose the importance weights w_j were exact, rather than estimated as in step 6 of Section 3.1. Then, for a suitable $e(\theta)$, importance weighting ensures that the weighted empirical distribution of the sample converges to the log pooled prior as the number of importance samples $n \to \infty$ (Hammersley and Hanscomb, 1964). Thus, by sampling the weighted sample with replacement with probabilities proportional to the importance weights, the resulting secondary sample (of size m) would converge in distribution to the log pooled prior as $\frac{n}{m} \to \infty$—this is the SIR algorithm (Rubin, 1987, 1988). A suitable $e(\theta)$ is one that has nonzero support wherever $p_2^*(\theta)$ and $p_1(\theta)$ have nonzero support; it is sufficient that $e(\theta)$ has nonzero support on Θ.

In step 6 of Section 3.1, however, the w_j are estimated. For invertible $\boldsymbol{M}$, the degree of approximation in this estimate is not a function of sample size, but

rather depends on the quality of the discrete difference approximations of the derivatives. The distributional convergence of importance sampling methods is not affected by this approximation except that the limiting distribution will be approximate if the derivatives are approximate.

In Section 3.2.3 we described an adaptive approach for selection of the importance sampling function $e(\boldsymbol{\theta})$. Although this algorithm is also iterative, convergence of the algorithm itself is not necessary for convergence of the coherized prior. Rather, the adaptive algorithm is designed to drastically improve importance sampling efficiency in a few steps. Thus, the algorithm could even be stopped before it "converges", as is done in the examples of Section 4.5 and 4.6. Regardless of convergence of the adaptive algorithm, the final results from it are guaranteed to be correct in a limiting final step, due to the importance weighting employed therein.

Chapter 4

Log Pooling: Examples

4.1 Description

In this chapter, two versions of the general algorithm in Chapter 3 for obtaining the log pooled prior $p(\boldsymbol{\phi})$ are compared. The nonadaptive version features $p_1^*(\boldsymbol{\phi})$ as the importance sampling envelope, while the adaptive version features the adaptively chosen sampling envelope $e(\boldsymbol{\phi})$ as described in Section 3.2.3.

We begin with simple, low-dimension models $\boldsymbol{M}$ to enable a degree of visualization of the pooling and results which is difficult with more complex $\boldsymbol{M}$. We also examine performance on a challenging multi-dimensional $\boldsymbol{M}$, which is relevant to a real biological population management problem.

Hereafter, we focus exclusively on the priors and the estimation of the log pooled $p(\boldsymbol{\phi})$ and $p(\boldsymbol{\theta})$, because what follows after obtaining $p(\boldsymbol{\phi})$ or $p(\boldsymbol{\theta})$ is standard Bayesian inference. In addition, we chose to focus primarily on pooled priors for the output space for these examples; we chose $\boldsymbol{\phi}$-space since inference is commonly desired on the output parameters $\boldsymbol{\phi}$.

4.1.1 Method of Evaluation

Let the integrated squared error be ISE $= \int(\widehat{F_p(\boldsymbol{\phi})} - F_p(\boldsymbol{\phi}))^2 d\boldsymbol{\phi}$, which is the squared area between the estimated and true cdfs of $p(\boldsymbol{\phi})$. ISE in $\boldsymbol{\theta}$-space is similarly calculated. Izenman (1991) states that ISE is a common choice for comparing density estimates; in fact, he recommends it as a measure of goodness

of fit over other commonly used measures. For most of the following examples, we have the luxury of knowing $F_p(\phi)$ and $F_p(\theta)$ so that we can evaluate the relative performances of the nonadaptive and adaptive algorithms discussed in Chapter 3.

Let relative integrated squared error be RISE = ISE(adaptive)/ ISE(nonadaptive), so that RISE < 1 indicates that the adaptive algorithm produced a better fit for log pooled $p(\phi)$ or $p(\theta)$. For each example, 20 replicates of a pooling analysis (either adaptive or nonadaptive) are run and the RISE is calculated. The distribution of RISEs can be used to compare the adaptive and nonadaptive methods for log pooling; a mean and confidence interval based on the 20 replicates are calculated, where the confidence interval is based on an assumption of normality whenever strong evidence of nonnormality does not exist. In addition, a 90% confidence interval is used because the 95% upper confidence bound is of primary interest. Within each replication, equal numbers of simulation runs are used to compare the adaptive and nonadaptive methods and to compute RISEs; for instance, 4 iterations of 200 model simulations each in the adaptive algorithm will be compared to 800 simulations in the nonadaptive algorithm. For problems involving higher dimensions, we also evaluate ISEs along each margin. Plots from a representative replication are shown to illustrate differences between the two algorithms for obtaining log pooled priors.

4.2 Introductory Example

Consider the introductory example described in Section 1.3.4. In this case, the support for the implicit $p_1^*(\phi)$ does not adequately cover the support for $p_2(\phi)$. Therefore, even with appropriate importance weights w_i to ensure the correct limiting distribution for the sample, there is a terrible problem of efficiency; for essentially any implementable n, a poor estimate of $p(\phi)$ results.

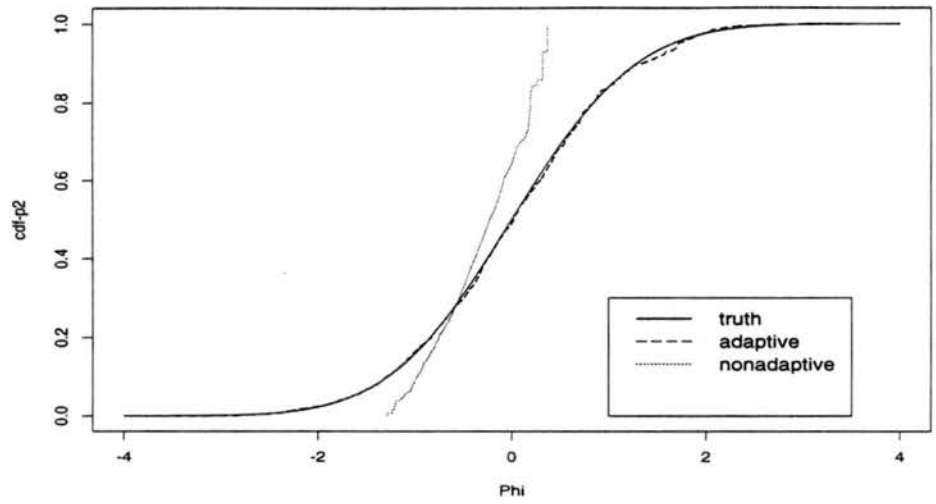

Figure 4.1: Introductory example: estimated and true cumulative distribution functions for $p_2(\phi)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods, and are shown for the seventh iteration, at which point the adaptive method converged.

Figure 4.1 shows the true (solid) and estimated (dashed) cumulative distribution functions for $p_2(\phi)$ at the final iteration required to find a suitable $e(\theta)$ by the adaptive mixture envelope strategy (see Section 3.2). In this particular replication, seven iterations of 200 simulation runs were required to achieve convergence. To review, the adaptive envelope strategy attempts to determine a sample of points which is representative of the prior opinions of both experts. This is done by starting with a sample of points from $p_1(\theta)$ in the first iteration and then searching for points which would map into $p_2(\phi)$ in the remaining iterations. As a result, the comparison of the true cdf of $p_2(\phi)$ to estimates of it is a key step in the process of estimating $p(\phi)$, and poor performance here will lead to poor overall results. For instance, points selected under the nonadaptive method produce to a poor estimate (dotted) of $F_{p_2}(\phi)$ in Figure 4.1, which will lead to poorer estimates of $F_p(\phi)$. For the adaptive method, the region of ϕ-space explored changes quickly from what is preferred by $p_1(\theta)$ to what is compatible with both priors. Figure 4.2 shows $e^*(\phi)$ at the final iteration; after starting with

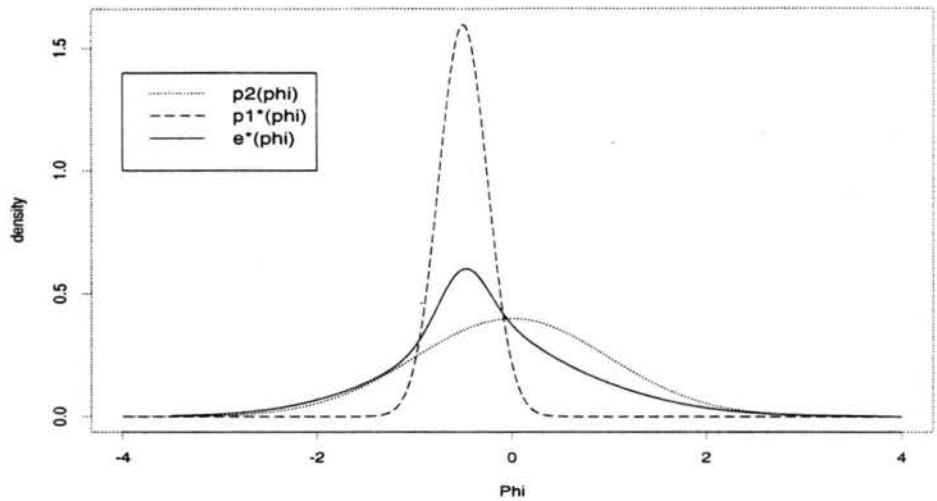

Figure 4.2: Introductory example: $e^*(\phi)$ at the final iteration of the adaptive sampling method.

a sample from $p_1^*(\phi)$ in the first iteration, $e^*(\phi)$ has become a density suitable for importance sampling for both p_1^* and p_2, and hence for the coherized prior $p(\phi)$.

Both methods actually perform well in this easy example (see Figure 4.3); however, the adaptive method provided estimates of $F_p(\phi)$ and $F_p(\theta)$ which were over twice as good as those from the nonadaptive method (average RISEs of .39 on the ϕ-axis and .39 on the θ-axis with 95% upper confidence bounds of .54 and .52, respectively).

The performances of the adaptive and nonadaptive algorithms by iteration are illustrated in Figure 4.4. Each point in the adaptive (dashed) line, for example, represents the average ISE at a particular iteration for the estimation of $F_p(\phi)$ over 20 replications of the adaptive method. Each replication was allowed to complete 8 iterations of 200 simulations runs regardless of whether the convergence criterion was met earlier. From this plot, we see that the adaptive method outperforms the nonadaptive method in this example in terms of both efficiency and precision. $\widehat{F_p}$ from the adaptive method improves drastically in Iteration 2—the first iteration which contains points other than from $p_1(\theta)$—after which the rate of improvement slows considerably. The degree of improvement in $\widehat{F_p}$ estimates

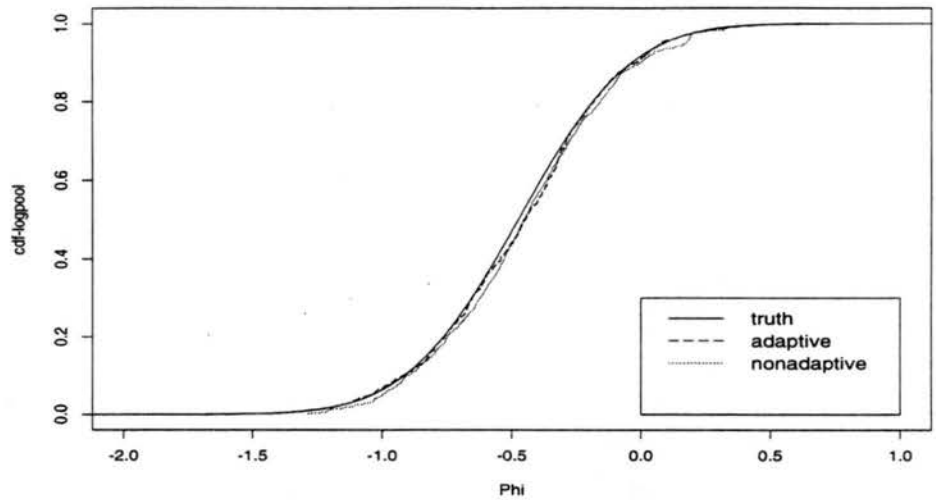

Figure 4.3: Introductory example: estimated and true cumulative distribution functions for $p(\phi)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods, and are shown for the seventh iteration, at which point the adaptive method converged.

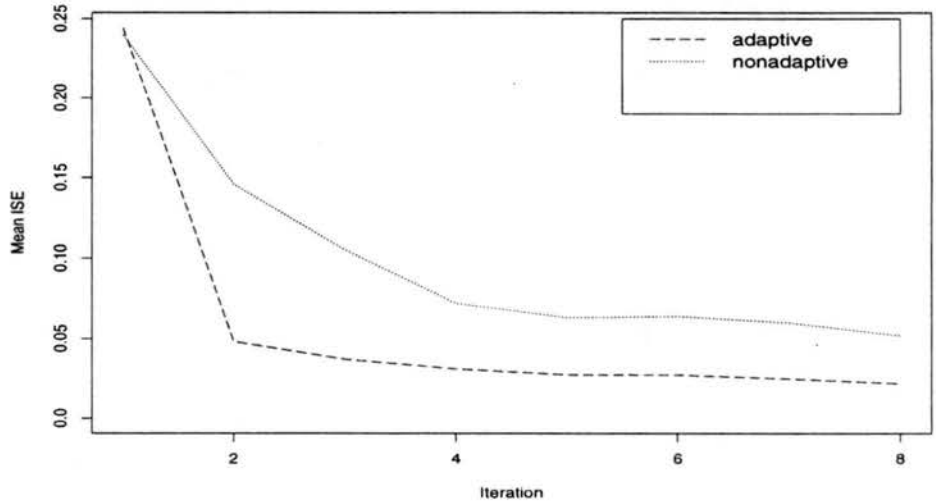

Figure 4.4: Introductory example: mean ISE by iteration for the estimation of $F_p(\phi)$ over 20 replications of the adaptive (dashed) and nonadaptive (dotted) methods.

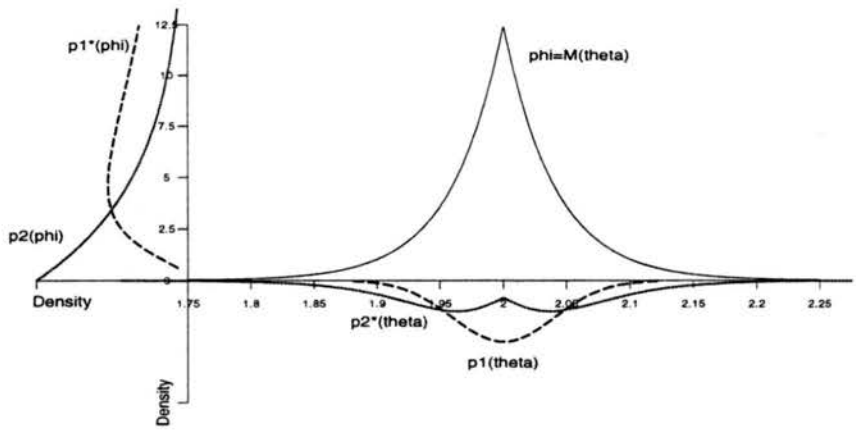

Figure 4.5: Cusp example: priors and implied priors for θ and ϕ, where θ is on the horizontal axis. The dashed densities correspond to p_1, and the dotted to p_2.

from the nonadaptive method also decreases with each iteration. At Iteration 2, the adaptive method has achieved a mean ISE which the nonadaptive method fails to achieve even after 8 iterations of 200 simulation runs each.

4.3 Cusp Example

Consider the noninvertible model $\phi = 50 \exp(-25\,|\theta - 2|)$, with priors $p_1(\theta) = N(\theta; 2, .04^2)$ and $p_2(\phi) = Exp(\phi; 2.5)$ where $Exp(x; \lambda)$ is the exponential distribution with mean λ. This example is sketched in Figure 4.5; it provides a first examination of how implementation proceeds for noninvertible models, using the approach detailed in Chapter 3. For instance, an estimate of $p_1^*(\phi)$ under noninvertible M is handled using a weighted average of estimates of $p_1^*(\phi)$ for a neighborhood of points around ϕ, where weights are inversely proportional to the probability of selecting the θ corresponding to ϕ. In this way, we need not explicitly know the partition of Θ-space which leads to invertibility within each partition in order to determine $p_1^*(\phi)$.

In this case, the implicit $p_1^*(\phi)$ conflicts with $p_2(\phi)$. Although the importance weights w_i for the coherized prior $p(\phi)$ are calculated appropriately, many

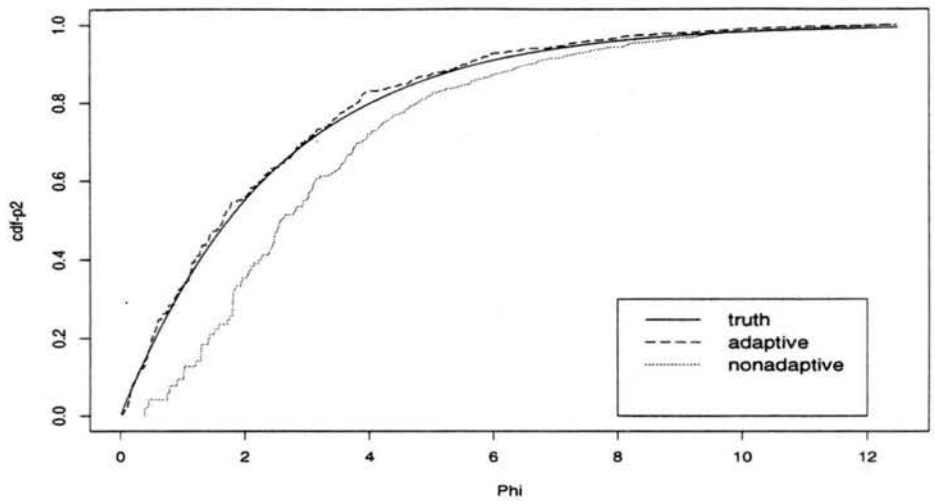

Figure 4.6: Cusp example: estimated and true cumulative distribution functions for $p_2(\phi)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods, and are shown for the second iteration, at which point the adaptive method converged.

θ_i sampled from $p_1(\theta)$ under the nonadaptive method are wasted because the corresponding ϕ_i receive negligible support by $p_2(\phi)$, and therefore by the coherized prior too. This can lead to a very biased estimate of the coherized prior, and hence a potentially biased posterior. In Section 2.3, we address the wisdom of pooling in a case where serious prior incoherence exists.

For a representative replication from this example, the true (solid) and estimated (dotted) cumulative distribution functions for $p_2(\phi)$ are shown in Figure 4.6, where estimation is performed using the nonadaptive algorithm ($n = 400$ model simulations). The bias of the nonadaptive method is due to severe sampling inefficiency. Figure 4.7 shows the true (solid) and estimated (dotted) coherized prior distribution functions. The bias persists.

Shifting to the adaptively chosen $e(\theta)$ fixes the problem. The dashed lines in Figures 4.6 and 4.7 show the results using the adaptive mixture envelope strategy. The adaptive strategy required two iterations in this replication (each with 200 simulations) to find a suitable sample for coherization which covered both possible

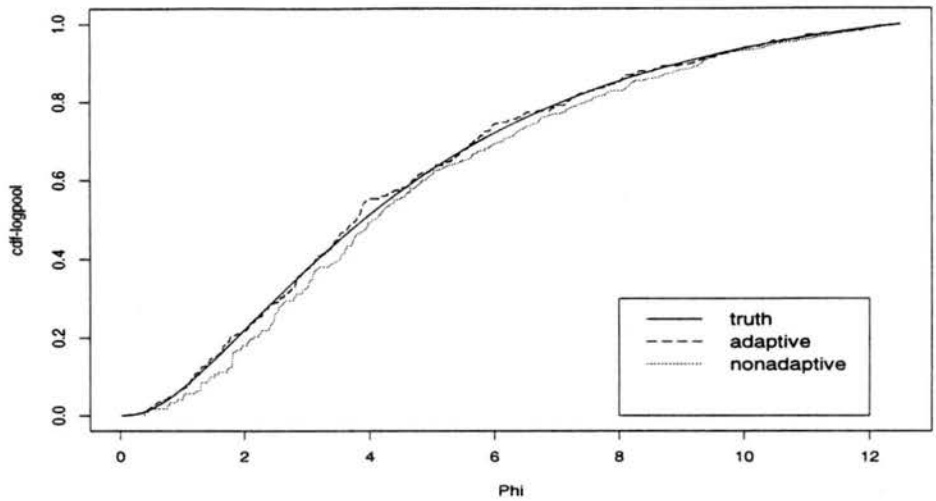

Figure 4.7: Cusp example: estimated and true cumulative distribution functions for $p(\phi)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods, and are shown for the second iteration, at which point the adaptive method converged.

inverse mappings of ϕ-space. The adaptive method provided a better estimate of the log pooled prior: the average RISE was .51 on the ϕ-axis and .50 on the θ-axis with 95% upper confidence bounds of .87 and .70 respectively.

The performances of the adaptive and nonadaptive algorithms by iteration are illustrated in Figure 4.8. Each point in the adaptive (dashed) line, for example, represents the average ISE at a particular iteration for the estimation of $F_p(\phi)$ over 20 replications of the adaptive method. Each replication was allowed to complete 8 iterations of 200 simulations runs regardless of whether the convergence criterion was met earlier. From this plot, we see that the adaptive method outperforms the nonadaptive method in this example in terms of both efficiency and precision. $\widehat{F_p}$ from the adaptive method improves drastically in Iteration 2—the first iteration which contains points other than from $p_1(\theta)$—after which the rate of improvement slows considerably. The degree of improvement in $\widehat{F_p}$ estimates from the nonadaptive method also decreases with each iteration. At Iteration 2,

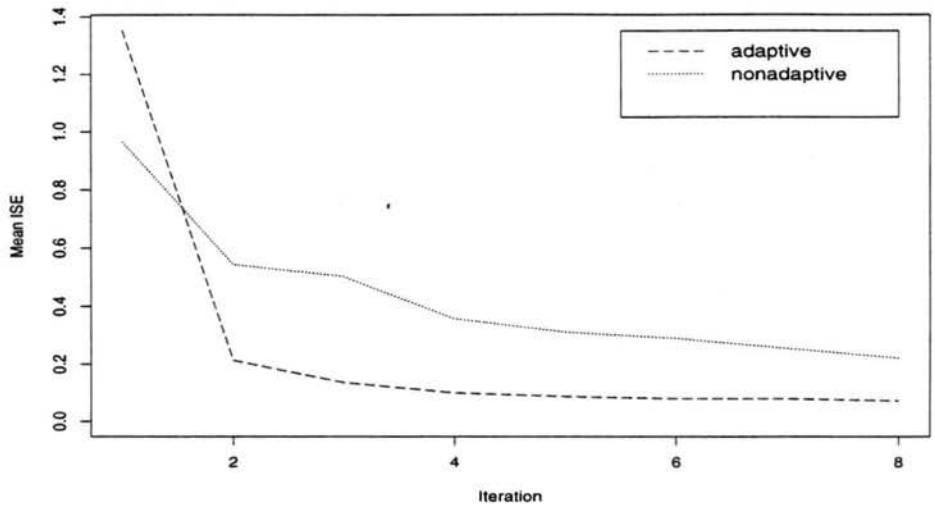

Figure 4.8: Cusp example: mean ISE by iteration for the estimation of $F_p(\phi)$ over 20 replications of the adaptive (dashed) and nonadaptive (dotted) methods.

the adaptive method has achieved a mean ISE which the nonadaptive method fails to achieve even after 8 iterations of 200 simulation runs each.

4.4 Linear $\Re^2 \to \Re^2$ Example

Having illustrated the value of our adaptive envelope selection strategy in one-dimensional problems, we set out to evaluate its performance in multi-dimensional problems. To that end, we considered a linear 2-dimensional case with

$$p_1(\boldsymbol{\theta}) = N_2\left[\begin{pmatrix} -1 \\ -3 \end{pmatrix}, \begin{pmatrix} 1 & .8 \\ .8 & 4 \end{pmatrix}\right],$$

$$p_2(\boldsymbol{\phi}) = N_2\left[\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -.1 \\ -.1 & .5 \end{pmatrix}\right],$$

$\phi_1 = \frac{1}{2}\theta_1 - \frac{1}{2}\theta_2 + 2$, and $\phi_2 = \frac{1}{4}\theta_1 + \frac{1}{8}\theta_2 - 1$.

These priors and corresponding implicit priors through $\boldsymbol{M}$ are sketched in Figure 4.9. Note, for example, that a region of agreement between $p_2(\boldsymbol{\phi})$ and $p_1^*(\boldsymbol{\phi})$ exists where the log-pooled prior $p(\boldsymbol{\phi})$ is probably large, but that a sample from $p_1^*(\boldsymbol{\phi})$ would not cover the entire support of $p_2(\boldsymbol{\phi})$.

In the pictured replication, 8 iterations of the adaptive algorithm (each with 200 model simulations) led to convergence along both $\boldsymbol{\phi}$-margins, as shown by the

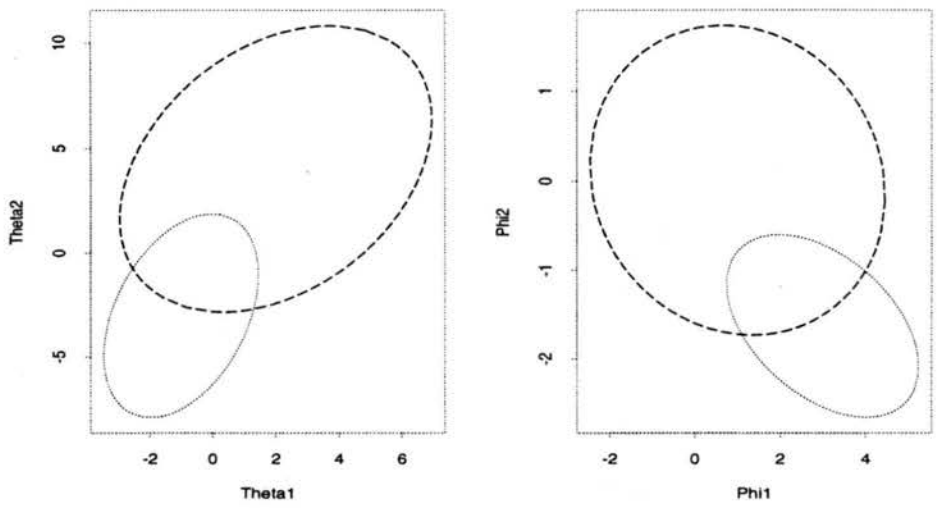

Figure 4.9: Linear $\Re^2 \rightarrow \Re^2$ Example: priors and implied priors for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The dashed 95% ellipsoids represent the $\boldsymbol{\theta}$-expert, while the dotted 95% ellipsoids represent the $\boldsymbol{\phi}$-expert.

marginal plots of $\widehat{F}_{p_2}$ versus F_{p_2} in Figures 4.10 and 4.11. As in the $\Re^1 \rightarrow \Re^1$ examples, estimates of $F_{p_2}(\boldsymbol{\phi})$ and $F_p(\boldsymbol{\phi})$ were obtained using equal numbers of simulation runs from both the adaptive and the nonadaptive methods (n=1600). Bias due to sampling inefficiency under the nonadaptive approach is easily observable by examining the dotted estimated cumulative distribution functions for p_2 in Figures 4.10 and 4.11.

Figures 4.12 and 4.13 then show the true (solid) marginal cumulative distribution functions for $p(\boldsymbol{\phi})$ compared to the estimates from the adaptive (dashed) and nonadaptive (dotted) methods. For the adaptive method, the lack of bias in estimating $p_2(\boldsymbol{\phi})$ led to greater efficiency in estimating $p(\boldsymbol{\phi})$ when compared to the nonadaptive approach. Numerical summary measures support conclusions suggested by Figures 4.12 and 4.13: the average RISE was .25 along the ϕ_1-axis, .38 along the ϕ_2-axis, and the average bivariate RISE in $\boldsymbol{\phi}$-space was .35 (with respective 95% upper confidence bounds of .41, .63, and .52). Thus, on average, the

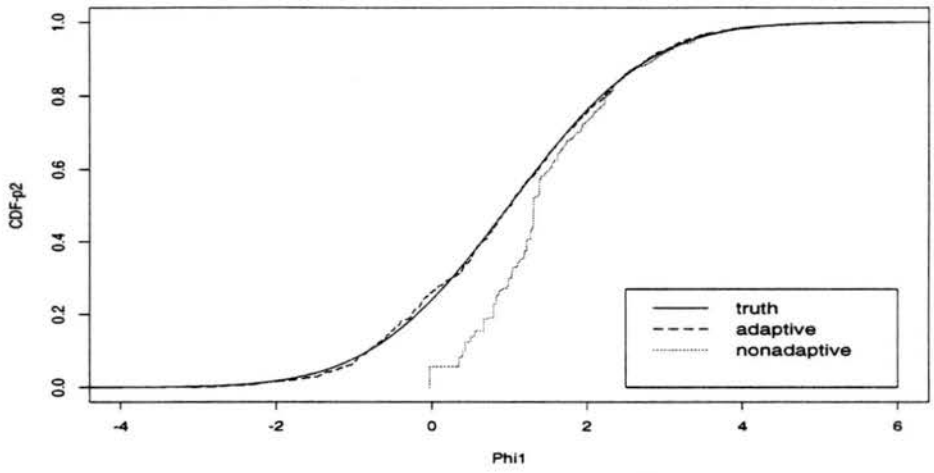

Figure 4.10: Linear $\Re^2 \rightarrow \Re^2$ Example: estimated and true cumulative distribution functions for $p_2(\phi_1)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods, and are shown for the eighth iteration, at which point the adaptive method converged.

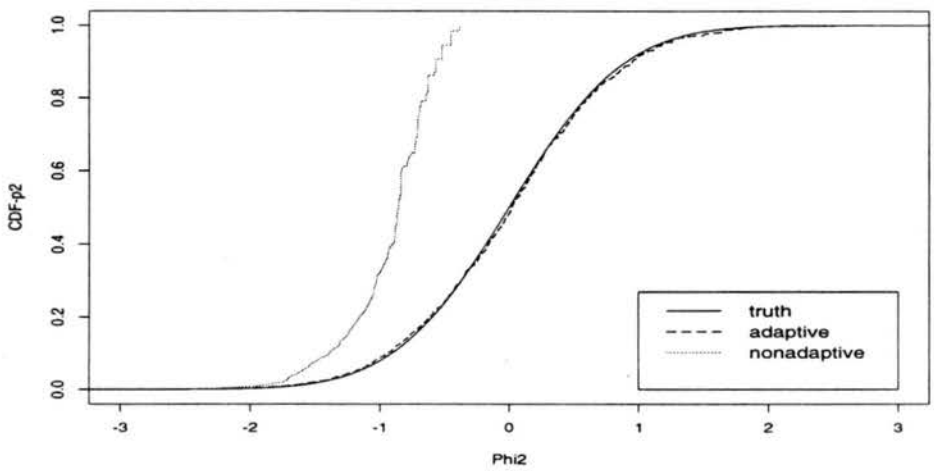

Figure 4.11: Linear $\Re^2 \rightarrow \Re^2$ Example: estimated and true cumulative distribution functions for $p_2(\phi_2)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods, and are shown for the eighth iteration, at which point the adaptive method converged.

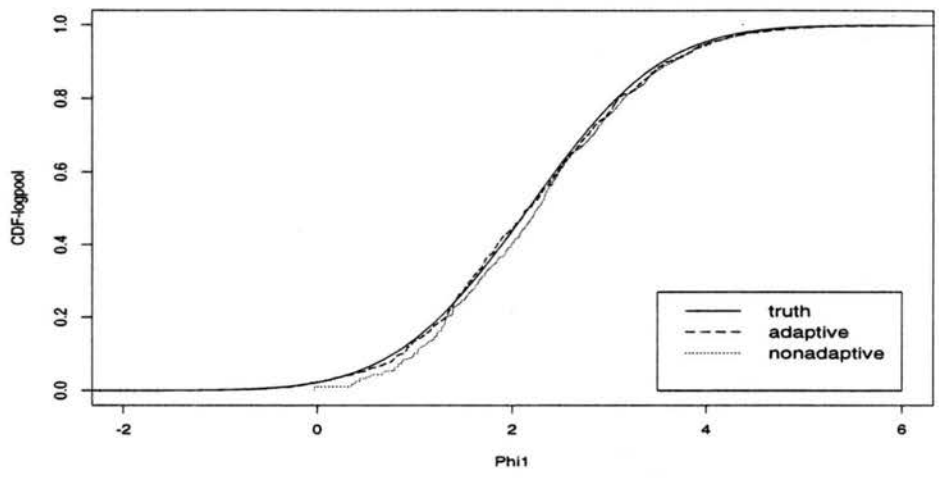

Figure 4.12: Linear $\Re^2 \rightarrow \Re^2$ Example: estimated and true cumulative distribution functions for $p(\phi_1)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods, and are shown for the eighth iteration, at which point the adaptive method converged.

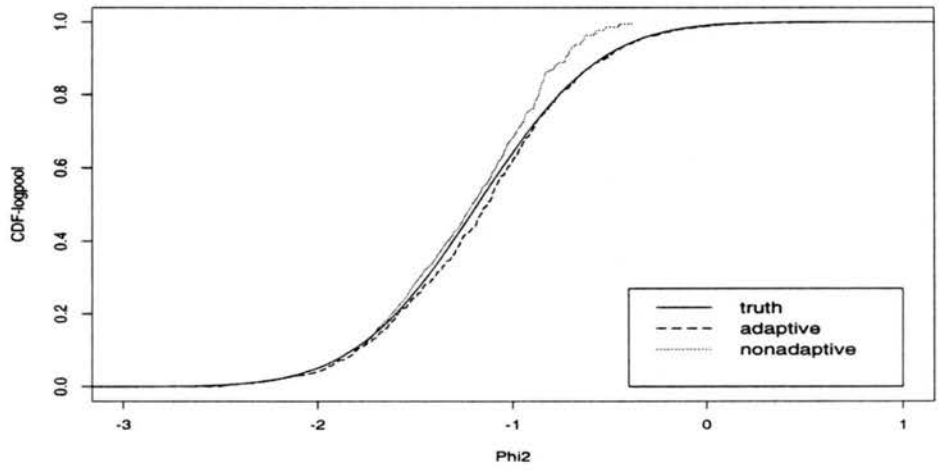

Figure 4.13: Linear $\Re^2 \rightarrow \Re^2$ Example: estimated and true cumulative distribution functions for $p(\phi_2)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods, and are shown for the eighth iteration, at which point the adaptive method converged.

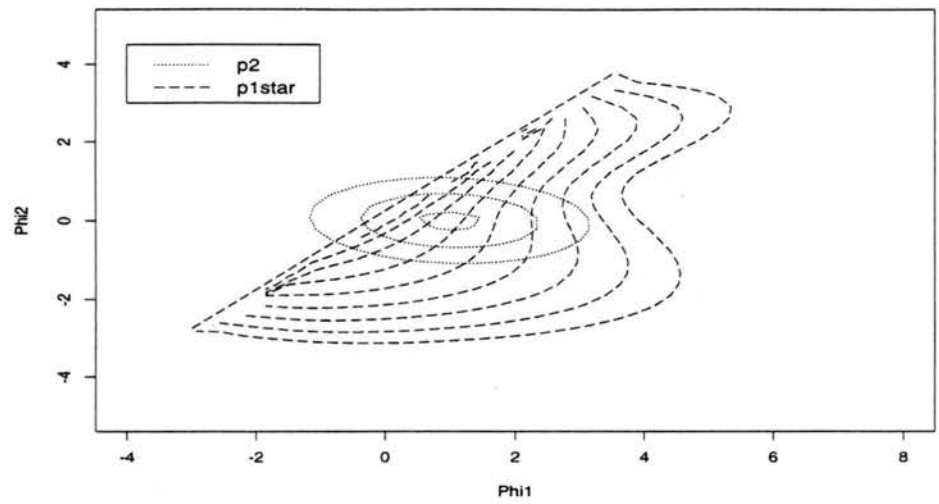

Figure 4.14: Noninvertible $\Re^2 \to \Re^2$ example: contour plots of $p_2(\boldsymbol{\phi})$ (dotted) and $p_1^*(\boldsymbol{\phi})$ (dashed).

adaptive algorithm produced an estimate of the log-pooled prior in $\boldsymbol{\phi}$-space which was nearly three times as good the analogous estimate using the nonadaptive algorithm and the same numbers of simulation runs. Even though both estimates were based on 1600 simulated points, the adaptive estimates were better able to approximate the truth due to more sensible selection of the 1600 points. Similar results were observed in $\boldsymbol{\theta}$-space (average RISEs of .36 along the θ_1-axis and .15 along the θ_2-axis, with 95% upper confidence bounds of .52 and .22).

4.5 Noninvertible $\Re^2 \to \Re^2$ Example

Consider the noninvertible model $\boldsymbol{M}$:$\Re^2 \to \Re^2$ in which $\phi_1 = \theta_1^2 + \theta_2$ and $\phi_2 = \theta_1 + \theta_2$, with bivariate normal priors:

$$p_1(\boldsymbol{\theta}) = N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right],$$

and

$$p_2(\boldsymbol{\phi}) = N_2\left[\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -.1 \\ -.1 & .5 \end{pmatrix}\right].$$

Contour plots of prior distributions are shown in Figures 4.14 and 4.15; note

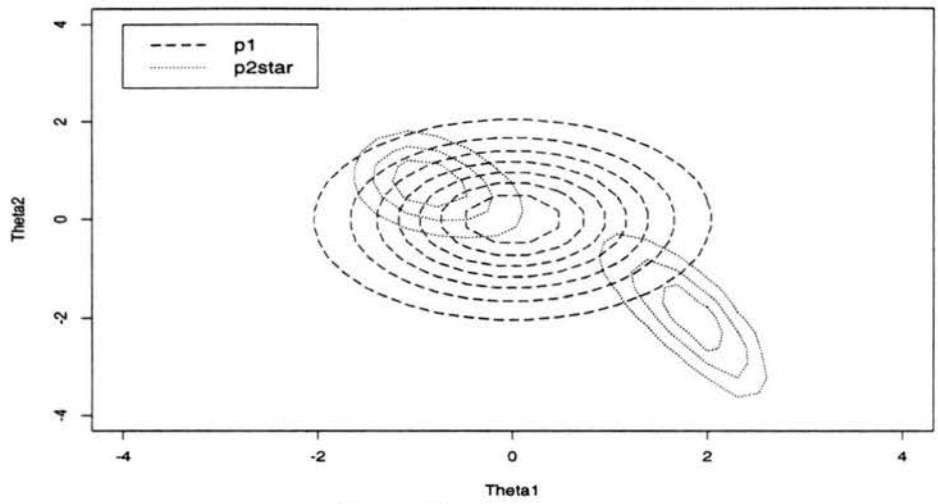

Figure 4.15: Noninvertible $\Re^2 \rightarrow \Re^2$ Example: contour plots of $p_1(\boldsymbol{\theta})$ (dashed) and $p_2^*(\boldsymbol{\theta})$ (dotted).

that $p_1^*(\boldsymbol{\phi}) = 0$ for $\boldsymbol{\phi}$ such that $\phi_1 < \phi_2 - \frac{1}{4}$, since there exists no $\boldsymbol{\theta}$ that produce outputs in that $\boldsymbol{\phi}$ half-plane.

In each replication, at most four iterations of the adaptive algorithm, with 400 model simulations per iteration, were used, while the nonadaptive estimates were obtained from at most 1600 simulations sampled from $p_1(\boldsymbol{\theta})$. To attempt to account for the noninvertibility, the method described in Section 3.4 was employed, using 5 nearest neighbors in $\boldsymbol{\phi}$-space to calculate estimates of $p_1^*(\boldsymbol{\phi})$ and $e^*(\boldsymbol{\phi})$.

The adaptive algorithm was stopped manually after 4 iterations because the mapping $\boldsymbol{M}$ prevents any $e(\boldsymbol{\theta})$ from inducing an $e^*(\boldsymbol{\phi})$ which covers the entire support of $p_2(\boldsymbol{\phi})$. As noted in Section 3.5, however, estimation performance is not dependent on convergence of the adaptation mechanism. The 4 iterations were often sufficient to achieve a improved importance sampling function $e(\boldsymbol{\theta})$. Initial evidence of an improved $e(\boldsymbol{\theta})$ in one representative replication can be found in Figures 4.16 and 4.17, which show estimated and true cumulative distribution functions for $p_2(\boldsymbol{\phi})$ along the ϕ_1- and ϕ_2-margins. Marginal estimates of the cdfs of $p(\boldsymbol{\phi})$ for this replication are presented in Figures 4.18 and 4.19; the good performance of the adaptive algorithm in this case was due to the improved importance

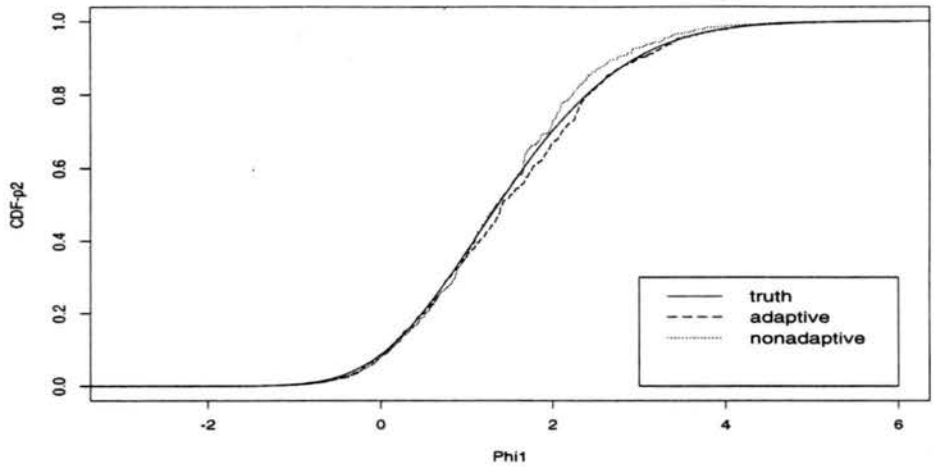

Figure 4.16: Noninvertible $\Re^2 \rightarrow \Re^2$ Example: estimated and true cumulative distribution functions for $p_2(\phi_1)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods.

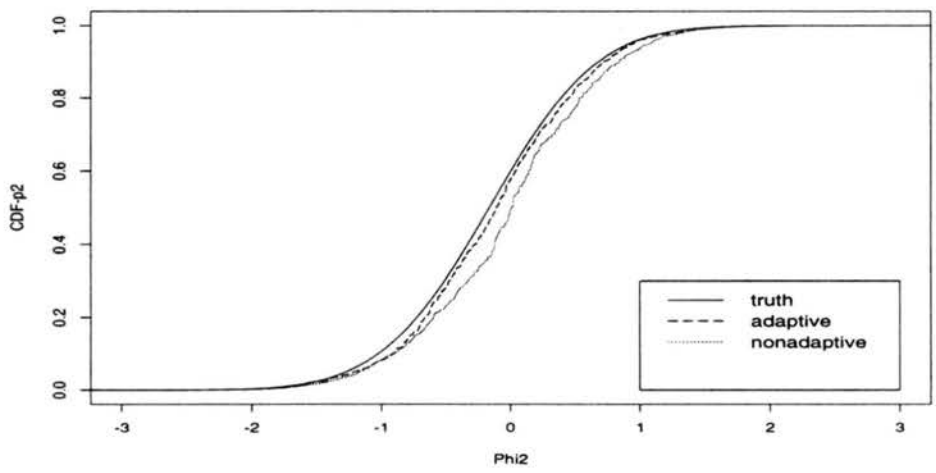

Figure 4.17: Noninvertible $\Re^2 \rightarrow \Re^2$ Example: estimated and true cumulative distribution functions for $p_2(\phi_2)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods.

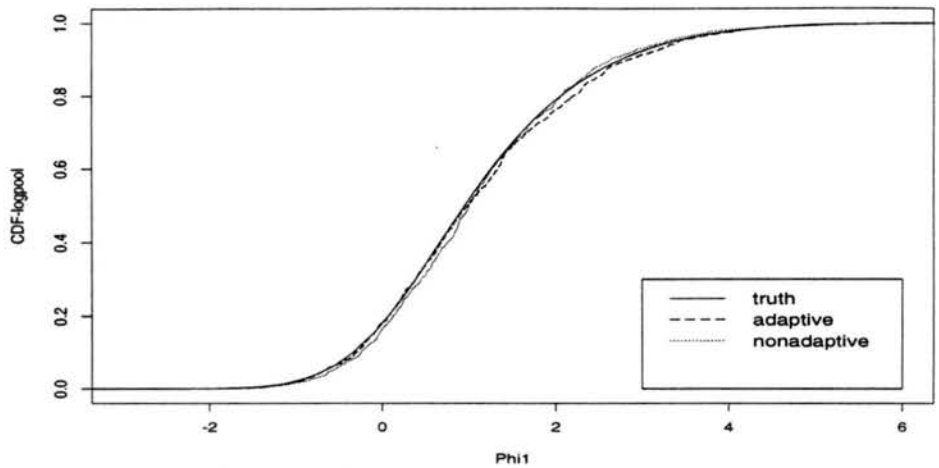

Figure 4.18: Noninvertible $\Re^2 \rightarrow \Re^2$ example: estimated and true cumulative distribution functions for $p(\phi_1)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods.

sampling function. A comparison of ISEs between the adaptive and nonadaptive methods for this example reveals a significant advantage for the adaptive method: the average RISE was .79 along the ϕ_1-axis, .59 along the ϕ_2-axis, and .60 in bivariate $\boldsymbol{\phi}$-space, with 95% upper confidence bounds of 1.14, .83, and .84 respectively. Results in $\boldsymbol{\theta}$-space were also encouraging (average RISEs of .45 along the θ_1-axis and .74 along the θ_2-axis with 95% upper confidence bounds of .79 and 1.02).

The performances of the adaptive and nonadaptive algorithms by iteration are illustrated in Figure 4.20. Each point in the adaptive (dashed) line, for example, represents the average ISE at a particular iteration for the estimation of $F_p(\boldsymbol{\phi})$ over 20 replications of the adaptive method. Each replication was allowed to complete 4 iterations of 200 simulations runs. From this plot, we see that the adaptive method outperforms the nonadaptive method in this example in terms of both efficiency and precision. $\widehat{F_p}$ from the adaptive method improves noticeably between Iterations 2 and 4, while the corresponding estimate from the nonadaptive method shows almost no improvement in the 4 iterations.

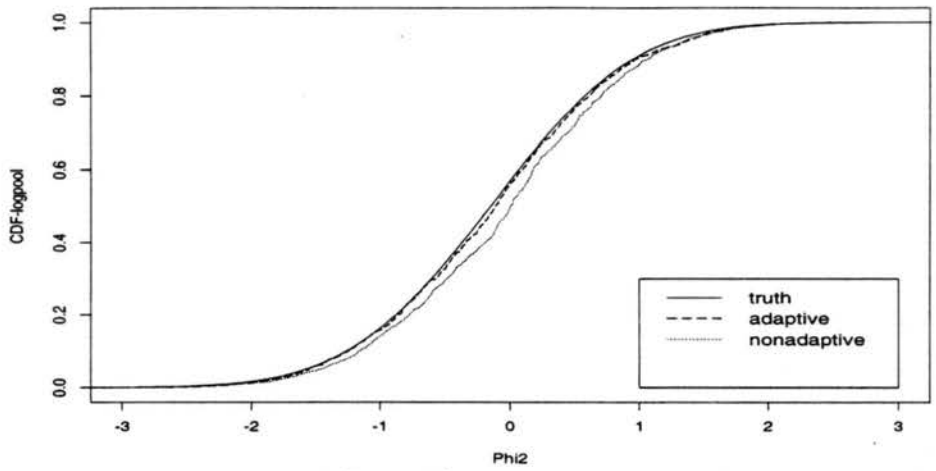

Figure 4.19: Noninvertible $\Re^2 \rightarrow \Re^2$ example: estimated and true cumulative distribution functions for $p(\phi_2)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods.

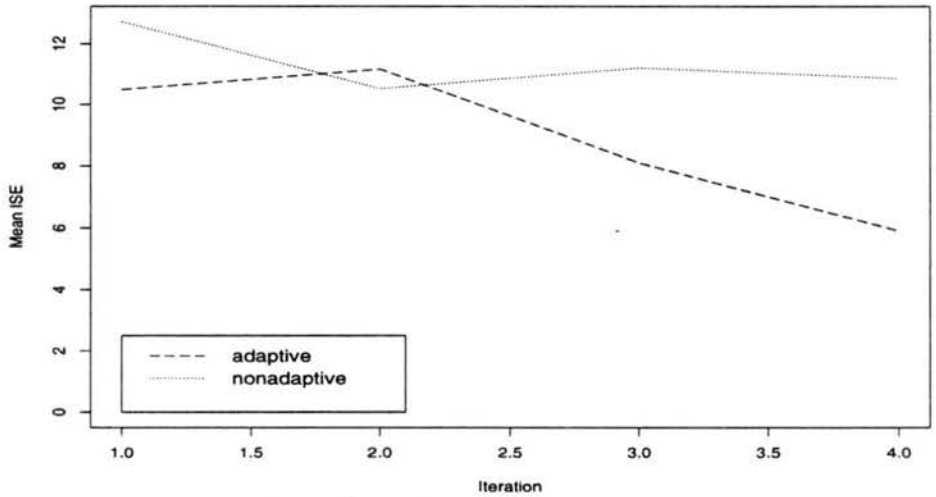

Figure 4.20: Noninvertible $\Re^2 \rightarrow \Re^2$ example: mean ISE by iteration for the estimation of $F_p(\phi)$ over 20 replications of the adaptive (dashed) and nonadaptive (dotted) methods.

4.6 Whale Population Model

We consider here a density-dependent population dynamics model which is applied to data on the Bering-Chukchi-Beaufort Seas stock of bowhead whales, *Balaena mysticetus*. Let the two inputs be $\theta_1 = K$ (the unexploited population size or carrying capacity) and $\theta_2 = \mu$ (a productivity parameter), and let the two outputs be $\phi_1 = P_{1993}$ (population size in 1993) and $\phi_2 = RY_{1993}$ (replacement yield in 1993). Replacement yield is the largest number of whales that can be hunted in a year without causing the population to decline. Inputs and outputs are linked by a recursive discrete-time model under which productivity depends on recent depletion (i.e. "density-dependent productivity"), and

$$P_{t+1} = P_t - C_t + \mu P_t \left(1 - (P_t/K)^2\right)$$

where C_t are fixed constants representing the number of catches (removals) at time t, and P_t is the population at time t.

Priors for the inputs and outputs may be established from various sources of available data (Raftery *et al.*, 1996; Givens *et al.*, 1995). We take $\theta_1 - 6400 \sim Gamma(2.8085, .0002886)$, where $Gamma(s, \lambda)$ is the gamma distribution with mean s/λ. The prior for $\theta_2 \mid \theta_1$ is derived from the depleted stock yield relationship $\theta_2 = (100 \cdot r + .94)(\theta_1 - 6081)$, where r is a productivity parameter for which data may in principle be observed and for which a prior is available (Raftery *et al.*, 1996)—i.e., $r \sim Gamma(8.2, 372.7)$. Schweder and Ianelli (1998) show that failure to account for severe nonlinearity in r can lead to numerically unreliable or intractable models. Figure 4.21 shows the relationship between r and K that is supported by the priors on outputs, given the model; the near dimension-reduction suggests that the model parameterized by μ should lead to more reliable inference. We adopt independent priors for ϕ_1 and ϕ_2 with $\phi_1 \sim N(7800, 1300^2)$. Although no direct data are available for ϕ_2, we use $\phi_2 \sim N(200, 75^2)$ because this covers

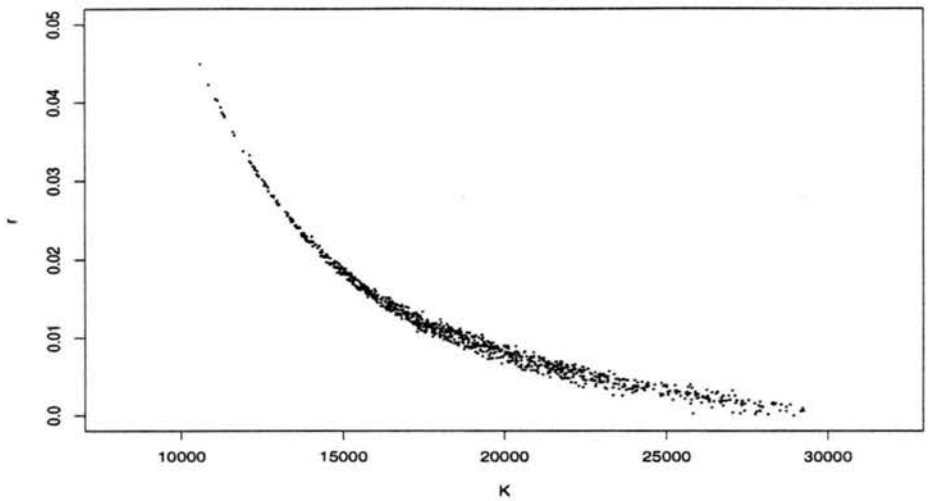

Figure 4.21: Whale population model: the relationship between r and K that is supported by the priors on outputs, given the model.

a range of catch levels that may be supported by this and similar whale stocks. (θ_1, θ_2) points which produce $\phi_1 \leq 0$ are excluded, since the population is known not to have been extinct in 1993.

Several features make this an extremely challenging model from which to obtain coherized priors. First, the model still produces a near-dimension reduction in some regions of parameter space. Figure 4.22 shows that $p_1^*(\boldsymbol{\phi})$ is concentrated on several thin manifolds in $\boldsymbol{\phi}$-space; a similar problem with the support of $p_2^*(\boldsymbol{\theta})$ is seen in Figure 4.23. One result of this problem is that small changes in $\boldsymbol{\theta}$ in certain regions of $\boldsymbol{\theta}$-space can bring about extremely large changes in $\boldsymbol{\phi}$. Second, a large degree of incoherence exists between the two priors when they are transformed to the same space, as is also evident in Figures 4.22 and 4.23.

Unless early convergence was attained, 4 iterations of the adaptive algorithm, with 400 model simulations per iteration, were used, while the nonadaptive estimates were obtained from 1600 simulations sampled from $p_1(\boldsymbol{\theta})$. For evaluation purposes, the true marginal and bivariate cdfs of $p_2(\boldsymbol{\phi})$ were obtained using a sample which was derived from a density specifically tailored, post-hoc, to this application. As with the noninvertible $\Re^2 \rightarrow \Re^2$ example of Section 4.5, the

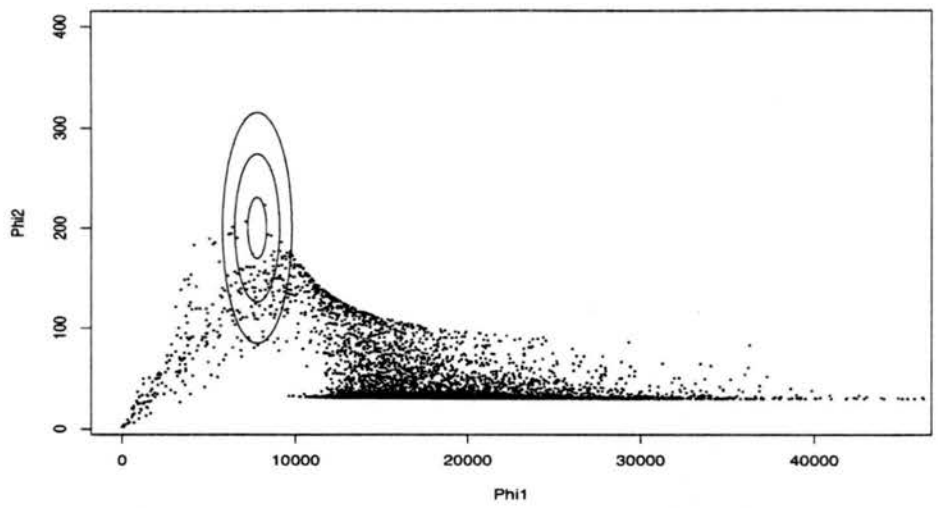

Figure 4.22: Whale population model: contour plot of $p_2(\boldsymbol{\phi})$ with points sampled from $p_1^*(\boldsymbol{\phi})$.

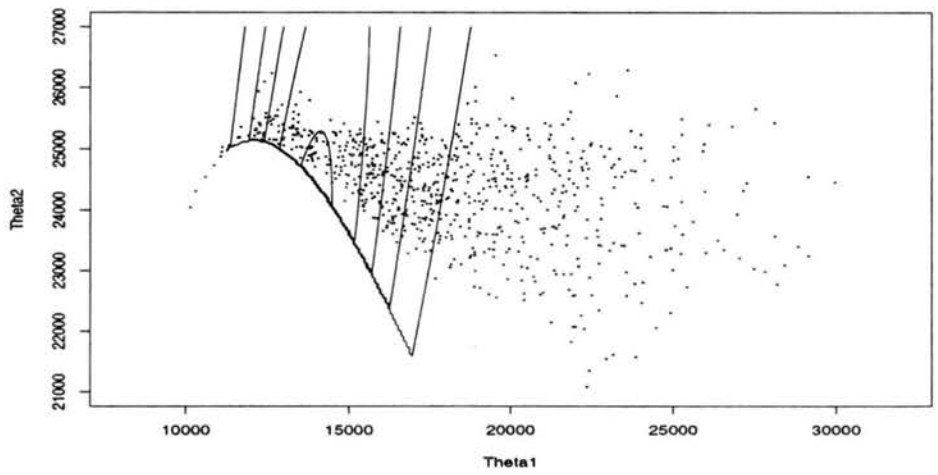

Figure 4.23: Whale population model: contour plot of $p_1(\boldsymbol{\theta})$ with points sampled from $p_2^*(\boldsymbol{\theta})$.

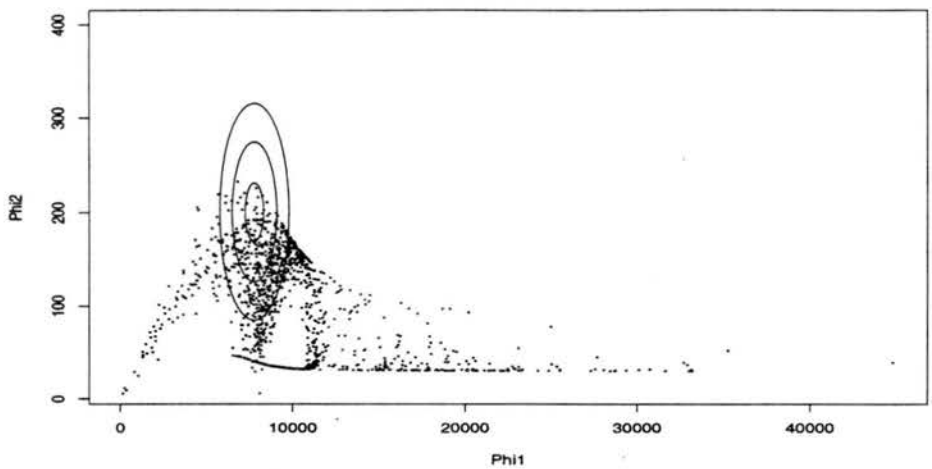

Figure 4.24: Whale population model: contour plot of $p_2(\boldsymbol{\phi})$ with points sampled from $e^*(\boldsymbol{\phi})$.

adaptive algorithm was stopped manually after 4 iterations—the 4 iterations were sufficient in most cases to achieve an improved importance sampling function $e(\boldsymbol{\theta})$ and better estimates of the log pooled prior. Figure 4.24 shows, in $\boldsymbol{\phi}$-space, the sample of points we obtained in this instance using the adaptive sampling strategy; a comparison to observations sampled from p_1 shown in Figure 4.22 confirms that the adaptive sampling envelope provides better coverage of $p_2(\boldsymbol{\phi})$. As a result, over 20 replications comparing the adaptive and nonadaptive methods in this whale population model, the average bivariate RISE in $\boldsymbol{\phi}$-space was .66 with a 95% upper confidence bound of .88. Average marginal RISEs were 1.21 on the ϕ_1-axis and .42 on the ϕ_2-axis, with respective 95% upper confidence bounds of 1.63 and .67. A point estimate above 1 on the ϕ_1-axis is not especially unexpected or discouraging. The goal of the adaptive approach is to produce better and more efficient estimates of the bivariate pooled prior, and marginal comparisons may not reflect the more complete picture of the improvements which have occurred. Marginal estimates of the cdf of $p(\boldsymbol{\phi})$ for a representative set of simulation runs are presented in Figures 4.25 and 4.26.

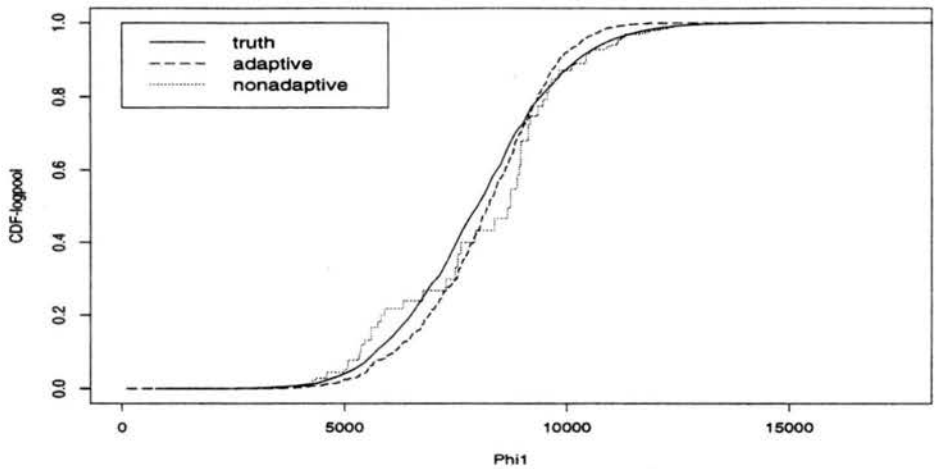

Figure 4.25: Whale population model: estimated and true cumulative distribution functions for $p(\phi_1)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods.

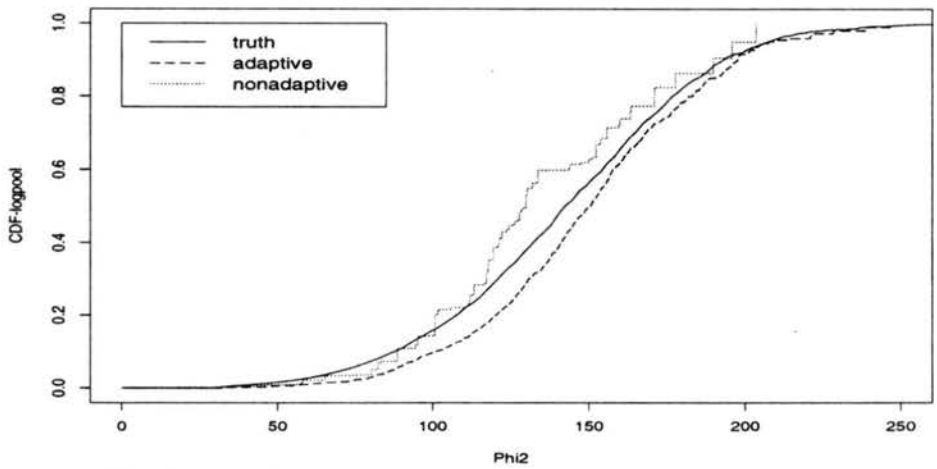

Figure 4.26: Whale population model: estimated and true cumulative distribution functions for $p(\phi_2)$. Estimates are based on adaptive (dashed) and nonadaptive (dotted) sampling methods.

This whale population model shows that the log pooling exercise did result in modified pooled beliefs about ϕ_2, replacement yield, compared to either $p_1^*(\phi_2)$ or $p_2(\phi_2)$. For example, the prior median replacement yield after log pooling is 142, compared to 200 for the initial prior on $\boldsymbol{\phi}$ and 31 for the initial prior on $\boldsymbol{\theta}$. Results from an analysis like this can enable policy makers to formulate whale management advice on the basis of all the evidence: $p_1(\boldsymbol{\theta})$, $p_2(\boldsymbol{\phi})$, and the likelihoods. Abandoning the pooling approach would mean ignoring either $p_1(\boldsymbol{\theta})$ or $p_2(\boldsymbol{\phi})$, which are based on important biological or survey research. Although dropping one prior simplifies the problem, such an analysis ignores evidence and may lead to inferior estimates and therefore to suboptimal resource management, which in this case might endanger the whale stock.

4.7 Summary

This chapter compared nonadaptive and adaptive algorithms for obtaining the log pooled prior over five examples in which priors were linked by a deterministic simulation model. In each of the examples, the adaptive method outperformed the nonadaptive method at a significant level. A numerical summary of the examples is provided in Table 4.1, which shows the relative efficiency of the two methods as measured by relative integrated squared error. An RISE below 1 indicates that the adaptive method approximated the true cumulative distribution function more closely than the nonadaptive method when compared over the same number of simulation runs. For the $\Re^1 \rightarrow \Re^1$ examples, RISEs are presented in both ϕ-space and θ-space; for the $\Re^2 \rightarrow \Re^2$ examples, RISEs are presented for marginal distributions in both $\boldsymbol{\phi}$-space and $\boldsymbol{\theta}$-space and for the bivariate distribution in $\boldsymbol{\phi}$-space. The exception is the whale example, where RISEs were not obtained in $\boldsymbol{\theta}$-space because of the difficulties in obtaining the true $p(\boldsymbol{\theta})$. For each

	RISE Evaluation Space				
			joint		
Example	ϕ_1	ϕ_2	(ϕ_1, ϕ_2)	θ_1	θ_2
linear $\Re^1$	.39 (.30,.52)			.39 (.28,.54)	
cusp $\Re^1$	.51 (.30,.87)			.50 (.35,.70)	
linear $\Re^2$	.25 (.16,.41)	.38 (.23,.63)	.35 (.23,.52)	.36 (.25,.52)	.15 (.10,.22)
noninvertible $\Re^2$	.79 (.55,1.14)	.59 (.41,.83)	.60 (.43,.84)	.45 (.25,.79)	.74 (.53,1.02)
whale $\Re^2$	1.21 (.90,1.63)	.42 (.26,.67)	.66 (.49,.88)		

Table 4.1: Logarithmic pooling: average RISEs (adaptive/nonadaptive) and corresponding 90% confidence bounds for 5 examples in Chapter 4.

example, average RISEs and corresponding 90% confidence bounds based on 20 replications are presented.

The primary strength of the adaptive technique is in instances when one is limited to few iterations and/or small numbers of model simulations. The nonadaptive alternative will often be effective (although still less efficient than the adaptive algorithm) when the problem allows one to process a sufficiently large sample. However, many attractive applications of these inference methods, such as climate models, disease transmission models, and biological population models can involve dozens of inputs and outputs and highly complex functional relationships between modeled quantities. In these cases, concerns about computing time and costs may require inference to be based on a fairly small number of model simulations, and an efficient sampling envelope like that defined by our adaptive

techniques will be a necessity. For these reasons, the results in this chapter were particularly encouraging.

Chapter 5

The French-Lindley Supra-Bayesian Method

The French-Lindley method used in the examples in Section 5.3 follows the results contained in French (1981). However, the groundwork for the approach was laid in an article by Lindley, Tversky, and Brown (1979, hereafter, LTB). Thus, the LTB method will be described first, and then French's extensions (1980,1981) will be detailed.

5.1 The Lindley, Tversky, and Brown (LTB) Method

The general problem of reconciling incoherent probability assessments can be illustrated with the following example. Let A be the event that a major earthquake hits southern California within the next decade, and let p_1 and p_2 be two different assessments of the probability of A. The probability assessments p_1 and p_2 could represent opinions of two experts or the results of two studies using different approaches, but in either case p_1 and p_2 will not generally be equal. The LTB method strives to reconcile p_1 and p_2 and produce a single estimate of the probability of A in light of p_1 and p_2. This reconciled probability is denoted $p(A \mid p_1, p_2)$.

LTB offer two approaches to the reconciliation problem, both of which rely on a framework in which both a decision maker N and at least one expert E exist. The decision maker N is a coherent person—she precisely and unbiasedly knows her own subjective probabilities about a particular event A. Incoherent persons may evaluate situations in a biased manner or be unable to precisely state their

beliefs in terms of probabilities. As an example, N could be an investor trying to decide whether to invest in Corporation X, where A is the event that Corporation X's stock price ends the year above 70. Initially, N has a subjective probability $p_0(A)$ for the occurrence of A. However, N also obtains an opinion from a financial analyst E, who states that his probability for A is $\lambda(A)$. The question LTB address is then: How should N change her belief about A in light of E's opinion?

Under an approach LTB label as *internal*, one assumes that an expert E has a set of "true" probabilities which are distorted in the elicitation process. These "true" probabilities, in addition to accurately reflecting E's beliefs, follow all basic rules of probabilities under summing, conditioning, and other operations. The goal is to estimate the expert's underlying "true" probabilities using observed probabilities elicited from the expert. As a result, the internal method is often used when an expert E offers several probability assessments about related events which are found to be incoherent in light of laws of probabilities.

A second approach, which LTB label *external*, is not concerned with "true" probabilities, but attempts to derive a set of probabilities which follow all basic probabilistic laws even when the original set of assessments do not. This method is often used for cases in which a decision maker N consults a group of experts, each of whom offers his or her probability for event A. These expert opinions will then form the basis for N's decision. Because we are primarily interested in this scenario, we focus our attention on the external approach.

Assume that E is an incoherent expert and N is a coherent decision maker. Both individuals consider a single event A. These ideas can later be extended to a sequence of events $A = (A_1, A_2, \ldots, A_m)$. Let $\pi(A)$ be E's true probability distribution for A, and let $\lambda(A)$ be E's stated probability assessment for A, where $\pi(A)$ and $\lambda(A)$ are often not equal because of E's difficulties in articulating probabilities. The external LTB method involves three key probabilities:

1. $p_0(A) = N$'s opinion about A

2. $p_0(\pi \mid A) = N$'s opinion of a coherent E's knowledge of A

3. $p_0(\lambda \mid \pi) = N$'s opinion of E as a probability assessor

An assumption is made that $p_0(\lambda \mid \pi, A) = p_0(\lambda \mid \pi)$—i.e., λ and A are independent, given the "true" probability π.

LTB advocate using log-odds rather than probabilities for λ and π for a couple of reasons. First, the assumption of constant variance for errors (such as in N's assessments of E) becomes more reasonable. Second, it reflects the greater accuracy with which people can evaluate high and low as opposed to mid-range probabilities. Third, the infinite ranges of log-odds allow one to model expert assessments and prior probabilities using normal distributions, which makes calculations easier. In terms of log-odds then, LTB make the following assumptions about their three key probabilities:

1. $p_0(A) = \alpha$

2. $p_0(\pi \mid A) \sim N(\mu_1, \tau^2)$, $p_0(\pi \mid \overline{A}) \sim N(\mu_2, \tau^2)$

3. $p_0(\lambda \mid \pi) \sim N(\pi, \sigma^2)$

With assumption 3, N views E's assessment of log-odds as unbiased with constant variance. With assumption 2, we quantify N's expectations about E's log-odds assessments in the cases in which A does or does not occur. In this way, $p_0(\pi \mid A)$ is related to Type I and Type II errors from statistics or false positives and false negatives in medical diagnostics. For a good probability assessor, $\mu_2 < 0 < \mu_1$, so that his or her average probability for events A which subsequently occur is above 50% and below 50% for events $\overline{A}$ which subsequently do not occur.

The LTB external method focuses on estimation of $p_0(A \mid \lambda)$—N's opinion about A after accounting for E's stated assessment of A. For the investment

example, the investor N will use the financial analyst E's opinion about A (the event that Corporation X's stock price ends the year above 70) to update her own opinion $p_0(A)$. As demanded by assumption 2, N must determine, before approaching E, what she thinks E will say in the case when A subsequently occurs (and similarly in the case when A subsequently does not occur).

We will now consider the external approach in greater detail. An application of Bayes' Theorem shows that $p_0(A \mid \lambda) \propto p_0(\lambda \mid A)p_0(A)$. $p_0(\lambda \mid A)$ can be assessed directly as what N predicts E will say he believes for each event A, or it can be derived using $p_0(\lambda \mid A) = \sum_\pi p_0(\lambda \mid \pi)p_0(\pi \mid A)$.

With the assumptions above, $p_0(\lambda \mid A) \sim N(\mu_1, \sigma^2 + \tau^2)$ and $p_0(\lambda \mid \overline{A}) \sim N(\mu_2, \sigma^2 + \tau^2)$. N's log-odds for A after accounting for E's stated assessment of the log-odds of A are given by

$$\begin{aligned}
l_0(A \mid \lambda) &= \ln \frac{p_0(A \mid \lambda)}{p_0(\overline{A} \mid \lambda)} \\
&\propto \log \frac{p_0(\lambda \mid A)p_0(A)}{p_0(\lambda \mid \overline{A})p_0(\overline{A})} \\
&= \log \left[\frac{p_0(A)}{p_0(\overline{A})}\right] + \log \left[\frac{\frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\left\{\frac{-(\lambda-\mu_1)^2}{2(\sigma^2+\tau^2)}\right\}}{\frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\left\{\frac{-(\lambda-\mu_2)^2}{2(\sigma^2+\tau^2)}\right\}}\right] \\
&= l_0(A) + \frac{1}{2(\sigma^2+\tau^2)}\left[\lambda^2 - 2\lambda\mu_2 + \mu_2^2 - \lambda^2 + 2\lambda\mu_1 - \mu_1^2\right] \\
&= l_0(A) + \frac{\lambda(\mu_1-\mu_2)}{\sigma^2+\tau^2} - \frac{\mu_1^2-\mu_2^2}{2(\sigma^2+\tau^2)}.
\end{aligned}$$

Thus, $l_0(A \mid \lambda) = l_0(A) + \frac{\lambda(\mu_1-\mu_2)}{\sigma^2+\tau^2} - \frac{\mu_1^2-\mu_2^2}{2(\sigma^2+\tau^2)}$.

5.2 The French Method

5.2.1 The Basic Approach: Single Decision Maker

French (1980) points out a flaw in the LTB approach. He believes that N's assessment of E should depend on N's own views about A—i.e. $p_0(\pi \mid A)$ depends

on $p_0(A)$. French justifies this extra conditioning in the following manner (1980, p. 45), with our notation inserted:

> When I come to state my subjective probabilities $p_0(\lambda \mid A)$ and $p_0(\lambda \mid \overline{A})$, I have already made explicit my beliefs about A. In other words, when I wonder what E will think about A, I already know what I think about A. Furthermore, I am aware why I hold my beliefs about A. Now this familiarity with judgments concerning A may well be relevant information to my judgments of λ.

For example, if N and E have tended to agree in the past in similar situations, then $p_0(\pi \mid A)$ might reasonably peak around $\pi = p_0(A)$. French's method (1980,1981), therefore, adds this extra bit of conditioning to the LTB approach.

Assume a decision maker N is interested in the occurrence of a particular event A. At this point, we allow N to solicit opinions from multiple experts. Returning to the investment example, N has preconceived notions about the probability of A occurring, but she realizes her limited knowledge regarding Corporation X and therefore solicits opinions from one or several financial analysts $(E_1, \ldots, E_k)$. Each analyst offers an opinion about the probability of event A, so the question becomes: What should N believe about the probability that the stock price of Corporation X reaches \$70 per share now? In other words, how should N update her beliefs on learning the opinions of the experts?

Let $p_0(A)$ be N's opinions for the probability of A, and let $p_i(A)$ be Expert i's opinion for the probability of $A, i = 1, \ldots, k$. Additionally, let $l_0(A) = \log\left(\frac{p_0(A)}{1-p_0(A)}\right)$ be the log-odds for N, and let $\lambda_i(A) = \log\left(\frac{p_i(A)}{1-p_i(A)}\right)$ be the log-odds for Expert $i, i = 1, \ldots, k$. The vector of log-odds $[l_0(A), \lambda_1(A), \ldots, \lambda_k(A)]'$ will be denoted by $\underline{l}(A)$, and the vector $[\lambda_1(A), \ldots, \lambda_k(A)]'$ will be denoted by $\underline{\lambda}(A)$. If the context is clear, the A will be dropped from $l_0(A), \lambda_1(A), \ldots, \lambda_k(A)$. The k

expert opinions, $\lambda_1(A), \ldots, \lambda_k(A)$, are considered not as probability statements, but as data which should be used to update N's prior opinion $l_0(A)$, forming a Bayesian posterior distribution for N's opinion of A. Therefore, we want to find an expression for $l_0(A \mid \underline{\lambda})$.

Similar to the LTB method, French makes the assumption that $\underline{l} \mid A \sim N(\underline{m}, V)$ and $\underline{l} \mid \overline{A} \sim N(\overline{\underline{m}}, V)$. For instance, $\underline{l} \mid A \sim N_{k+1}\left(\begin{bmatrix} m_0 \\ \underline{m}_1 \end{bmatrix}, \begin{bmatrix} V_{00} & V_{01} \\ V_{10} & V_{11} \end{bmatrix}\right)$, where m_0 is the expected log-odds which N would state in favor of A when A is subsequently found to be true, and $\underline{m}_1 = [m_1, \ldots, m_k]'$ is the vector of log-odds which N expects $E_1, \ldots, E_k$ to state in favor of A when A is subsequently found to be true. Also, V_{00} is the variance N expects in her own statements, $V_{01} = V_{10}' = [\sigma_{01}, \sigma_{02}, \ldots, \sigma_{0k}]$ is a vector of expected covariance terms between the statements of N and the individual experts $E_1, \ldots, E_k$, and

$$V_{11} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2k} \\ \vdots & \vdots & & \vdots \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_{kk} \end{bmatrix}$$

is a variance-covariance matrix for experts $E_1, \ldots, E_k$. $\overline{\underline{m}}$ is analogous to $\underline{m}$, but when A is subsequently found to be false. Note that a constant variance structure in the two cases is assumed.

Given these assumptions, we can use Bayes' Theorem and properties of the multivariate normal distribution to derive the following results (French, 1981).

Result 5.1 $\underline{\lambda} \mid A, l_0 \sim N_k(\underline{\mu}(l_0), W)$ *where* $\underline{\mu}(l_0) = \underline{m}_1 + V_{10}V_{00}^{-1}(l_0 - m_0)$ *and* $W = V_{11} - V_{10}V_{00}^{-1}V_{01}$.

Proof. First, note that $\underline{l} - \underline{m} \mid A \sim N_{k+1}(\underline{0}, V)$. Then, let

$$B = \begin{bmatrix} I_{1\times 1} & \underline{0}'_{1\times k} \\ -(V_{10}V_{00}^{-1})_{k\times 1} & I_{k\times k} \end{bmatrix}$$

so that

$$B(\underline{l} - \underline{m}) = B\begin{bmatrix} l_0 - m_0 \\ \underline{\lambda} - \underline{m}_1 \end{bmatrix} = \begin{bmatrix} l_0 - m_0 \\ \underline{\lambda} - \underline{m}_1 - V_{10}V_{00}^{-1}(l_0 - m_0) \end{bmatrix}$$

is (when conditioned on A) jointly normal with the $(k+1) \times 1$ mean vector $\underline{0}$ and covariance matrix given by

$$\begin{aligned} BVB' &= \begin{bmatrix} I_{1\times 1} & \underline{0}'_{1\times k} \\ -(V_{10}V_{00}^{-1})_{k\times 1} & I_{k\times k} \end{bmatrix} \begin{bmatrix} V_{00} & V_{01} \\ V_{10} & V_{11} \end{bmatrix} \begin{bmatrix} I_{1\times 1} & (-V_{10}V_{00}^{-1})'_{1\times k} \\ \underline{0}_{k\times 1} & I_{k\times k} \end{bmatrix} \\ &= \begin{bmatrix} V_{00} & V_{01} \\ \underline{0}_{k\times 1} & V_{11} - V_{10}V_{00}^{-1}V_{01} \end{bmatrix} \begin{bmatrix} I_{1\times 1} & (-V_{10}V_{00}^{-1})'_{1\times k} \\ \underline{0}_{k\times 1} & I_{k\times k} \end{bmatrix} \\ &= \begin{bmatrix} V_{00} & \underline{0}'_{1\times k} \\ \underline{0}_{k\times 1} & V_{11} - V_{10}V_{00}^{-1}V_{01} \end{bmatrix}. \end{aligned}$$

Because $\underline{\lambda} - \underline{m}_1 - V_{10}V_{00}^{-1}(l_0 - m_0)$ and $l_0 - m_0$ are independent (given A), the conditional distribution of $\underline{\lambda} - \underline{m}_1 - V_{10}V_{00}^{-1}(l_0 - m_0)$ given l_0 is the same as the unconditional distribution. Therefore,

$$\underline{\lambda} - \underline{m}_1 - V_{10}V_{00}^{-1}(l_0 - m_0) \mid A, l_0 \sim N_k(\underline{0}_k, V_{11} - V_{10}V_{00}^{-1}V_{01})$$

and

$$\underline{\lambda} \mid A, l_0 \sim N_k(\underline{m}_1 + V_{10}V_{00}^{-1}(l_0 - m_0), V_{11} - V_{10}V_{00}^{-1}V_{01}) \ \square$$

Result 5.2 $\underline{\lambda} \mid \overline{A}, l_0 \sim N(\underline{\overline{\mu}}(l_0), W)$ *where* $\underline{\overline{\mu}}(l_0) = \underline{\overline{m}}_1 + V_{10}V_{00}^{-1}(l_0 - \overline{m}_0)$ *and* $W = V_{11} - V_{10}V_{00}^{-1}V_{01}$.

Proof. Analogous to proof of Result 5.1.

Result 5.3

$$l_0(A \mid \underline{\lambda}) = \left[\underline{\mu}(l_0) - \underline{\overline{\mu}}(l_0)\right] W^{-1} \left[\underline{\lambda} - \frac{1}{2}(\underline{\mu}(l_0) + \underline{\overline{\mu}}(l_0))\right] + l_0(A)$$

Proof.

$$\begin{aligned} l_0(A \mid \underline{\lambda}) &= \log \frac{p_0(A \mid \underline{\lambda})}{p_0(\overline{A} \mid \underline{\lambda})} \\ &= \log \frac{p_0(\underline{\lambda} \mid A, l_0)p_0(A)}{p_0(\underline{\lambda} \mid \overline{A}, l_0)p_0(\overline{A})} \\ &= \log \frac{p_0(\underline{\lambda} \mid A, l_0)}{p_0(\underline{\lambda} \mid \overline{A}, l_0)} + \log \frac{p_0(A)}{p_0(\overline{A})} \\ &= l_0(\underline{\lambda} \mid A, l_0) + l_0(A) \end{aligned}$$

$$
\begin{aligned}
&= \log\left[\frac{\frac{1}{(2\pi)^{k/2}|W|^{1/2}}\exp\left\{-\frac{1}{2}(\underline{\lambda}-\underline{\mu}(l_0))'W^{-1}(\underline{\lambda}-\underline{\mu}(l_0))\right\}}{\frac{1}{(2\pi)^{k/2}|W|^{1/2}}\exp\left\{-\frac{1}{2}(\underline{\lambda}-\overline{\underline{\mu}}(l_0))'W^{-1}(\underline{\lambda}-\overline{\underline{\mu}}(l_0))\right\}}\right] + l_0(A) \\
&= -\frac{1}{2}(\underline{\lambda}-\underline{\mu}(l_0))'W^{-1}(\underline{\lambda}-\underline{\mu}(l_0)) + \frac{1}{2}(\underline{\lambda}-\overline{\underline{\mu}}(l_0))'W^{-1}(\underline{\lambda}-\overline{\underline{\mu}}(l_0)) \\
&\quad + l_0(A) \\
&= \underline{\mu}(l_0)'W^{-1}\underline{\lambda} - \overline{\underline{\mu}}(l_0)'W^{-1}\underline{\lambda} - \frac{1}{2}\underline{\mu}(l_0)'W^{-1}\underline{\mu}(l_0) \\
&\quad + \frac{1}{2}\overline{\underline{\mu}}(l_0)'W^{-1}\overline{\underline{\mu}}(l_0) + l_0(A) \\
&= \left[\underline{\mu}(l_0) - \overline{\underline{\mu}}(l_0)\right]W^{-1}\left[\underline{\lambda} - \frac{1}{2}(\underline{\mu}(l_0) + \overline{\underline{\mu}}(l_0))\right] + l_0(A) \quad \square
\end{aligned}
$$

Thus, Result 5.3 provides a prescription which allows the decision maker N to update her log-odds for the event A in light of the opinions of k experts (who expressed their assessments of the probability of A in the vector of log-odds $\underline{\lambda}$). With this approach, $l_0(A \mid \underline{\lambda})$ is essentially the posterior log-odds for the decision maker, where the vector of expert log-odds $\underline{\lambda}$ is treated as data for the decision maker, allowing usual Bayesian methods to be applied. By assuming a specific parametric form for $\underline{\lambda} \mid A$ and $\underline{\lambda} \mid \overline{A}$, the decision maker is able to account for possible miscalibration, dishonesty, and nonindependence among the experts.

French (1980, p.47) claims that the prescription for $l_0(A \mid \underline{\lambda})$ in Result 5.3 has a very natural interpretation. $\underline{\mu}(l_0) - \overline{\underline{\mu}}(l_0)$ is positive if E tends to assign positive log-odds to events A which subsequently occur and negative log-odds otherwise. $\underline{\lambda} - \frac{1}{2}(\underline{\mu}(l_0) + \overline{\underline{\mu}}(l_0))$ is positive if E's statement provides more evidence in favor of A occurring than the decision maker expects and negative otherwise. W^{-1} scales the decision maker's belief according to the expected variation in E's statement.

5.2.2 Combining Beliefs from a Group of Individuals

In his 1981 article, French was especially interested in how to combine a group of individuals' beliefs into a consensus probability distribution. He offers

two approaches based on his Bayesian updating scheme for a single decision maker presented above.

First, French suggests that everyone in the group update his or her initial opinion as if he or she is the group decision maker, soliciting opinions from everyone else. This results in a series of log-odds $l_0(A \mid \underline{\lambda})$, $l_1(A \mid l_0, \lambda_2, \ldots, \lambda_k)$, $\ldots$, $l_k(A \mid l_0, \lambda_1, \ldots, \lambda_{k-1})$. Ideally, at this point the group will have achieved a unanimous opinion, so that

$$l_0(A \mid \underline{\lambda}) = l_1(A \mid l_0, \lambda_2, \ldots, \lambda_k) = \ldots = l_k(A \mid l_0, \lambda_1, \ldots, \lambda_{k-1}).$$

Unfortunately, this is rarely the case, but the group members could proceed to inform each other of their new, revised opinions and repeat the Bayesian updating process. In fact, this process of sharing and revising could continue indefinitely. However, convergence to a unanimous opinion does not necessarily take place. French (1981) and others (DeGroot, 1974; Chatterjee and Senata, 1977) have considered the question of convergence with indefinite revision, especially in terms of stochastic models. French (1981, p.336) concluded that he ". . . can find no reasonable sets of condition to ensure that we should (reach a consensus)."

As a second approach, French (1981) introduces the concept of a group decision maker (GDM), which French views as the group considered in the sense of a single rational being. For the Bayesian updating scheme, the GDM begins with no knowledge of the probability of A, so that its prior is vague (e.g., $l_0(A) = 0$). Then, the GDM evaluates the forecasting ability of the group members via $\underline{\lambda} \mid A \sim N_k(\underline{m}, V)$ and $\underline{\lambda} \mid \overline{A} \sim N_k(\overline{\underline{m}}, V)$, where $\underline{m} = [m_1, \ldots, m_k]'$ and $\overline{\underline{m}} = [\overline{m}_1, \ldots, \overline{m}_k]'$. Note that $(\underline{\lambda} \mid A, l_0) = (\underline{\lambda} \mid A)$ and $(\underline{\lambda} \mid \overline{A}, l_0) = (\underline{\lambda} \mid \overline{A})$ because of the vague prior log-odds of the GDM. Finally, the posterior log-odds for the GDM, through application of Result 5.3, is

$$l_0(A \mid \underline{\lambda}) = (\underline{m} - \overline{\underline{m}})V^{-1}\left[\underline{\lambda} - \frac{1}{2}(\underline{m} + \overline{\underline{m}})\right] + l_0(A).$$

This posterior log-odds for the GDM is then the group consensus log-odds which was the original goal.

It is this GDM approach to the group consensus problem which is used to evaluate the French-Lindley method across several examples in Section 5.3. In deterministic simulation modeling, often a person (or group) wishes to make inference about events involving particular model parameters. This person, armed with knowledge of the model but usually no expert knowledge of the parameters in question, solicits expert opinion about all model parameters. In the deterministic simulation model framework, French's first group pooling approach is not feasible, since the experts whose opinions are surveyed are scattered throughout different fields and geographic locations; they may not even be aware of the model which maps their opinion into the parameter space of interest. Thus, we turn to French's GDM approach. The group decision maker, like the person who collects expert opinions, contributes no knowledge to the events and parameters of interest. Through French's Bayesian updating scheme, the GDM can update its noninformative prior by treating expert opinions as data.

Two significant problems still must be overcome before applying French's GDM approach to the deterministic simulation models of Section 5.3. First, the group decision maker must accurately evaluate the forecasting ability of all group members by estimating the parameters of the normal distributions outlined in Results 5.1 and 5.2. Not only must this ubiquitous, omniscient decision maker accurately quantify each expert's bias, inaccurate precision evaluations, and correlations with other experts, but the decision maker must make these evaluations using the log-odds scale. In the case of the GDM, the group members are required to form a consensus opinion about their own abilities as forecasters. For instance, French recommends using the group's history, if one exists, in evaluating similar events. In general, obtaining beliefs for the GDM which are agreeable to all group

members is a monumental task. LTB raise this concern as well, realizing that the assessment of $p_0(\pi \mid A)$ could prove to be a large hurdle from a practical standpoint. We found it extremely difficult to set reasonable, intuitive parameters in applications of the French-Lindley approach in Section 5.3.

Second, the French-Lindley approach, which was originally developed for binary events, must be adapted to handle the continuous parameters which are commonly found in deterministic simulation models. If we begin with a grid of points in the parameter space of interest, we can consider the series of binary events $A_i = \{\text{event that } \phi < \phi_i\}$ which leads to the cumulative distribution function. Through multiple applications of the French-Lindley method, a pooled distribution can be approximated. Further details on this adaptation to continuous variables are contained in Sections 5.3.1 and 5.3.2.

5.3 Examples

In this section, the group decision maker version of the French-Lindley supra-Bayesian method is applied to the examples from Chapter 4. One significant hurdle to the practical implementation of the French-Lindley method is the definition of reasonable hyperparameter values. The interpretations of the hyperparameters, as discussed in Section 5.2, are not often quantities easily appraised or intuitively known by the supra-Bayesian, which in this case is the group itself. As a result, the effect of the choice of hyperparameter values on the resulting consensus prior is illustrated in the succeeding sections. Although focus is constrained to $\boldsymbol{\phi}$-space, similar hyperparameter effects would be seen in $\boldsymbol{\theta}$-space.

5.3.1 $\Re^1 \rightarrow \Re^1$ Examples

In the examples of Chapter 4 the inputs and outputs are all continuous parameters; the French-Lindley method, however, was developed for binary events. By considering a finely nested set of half-intervals in ϕ-space, we can reconstruct

cumulative distribution functions using the methodology of French and Lindley to find approximate consensus prior distributions for ϕ. Specifically, define an equally spaced (width v) grid of ϕ's, $\phi_1, \phi_2 = \phi_1 + 1v, \ldots, \phi_n = \phi_1 + (n-1)v$, such that $p_1^*(\phi)$ and $p_2(\phi)$ are negligible for $\phi < \phi_1$ and $\phi > \phi_n$. Consider a series of n binary events $A_1, \ldots, A_n$ such that $A_i = \{\text{event that } \phi < \phi_i\}$. Thus, $P(A_i) = F(\phi_i)\ \forall i = 1, \ldots, n$ where $F(\phi)$ is the appropriate cumulative distribution function. Next, compute $F_{p_1^*}(\phi_i)$ and $F_{p_2}(\phi_i)$ for $i = 1, \ldots, n$. Note that $F_{p_1^*}$ and F_{p_2} are known for the introductory example, and F_{p_2} is known for the cusp example. Although no closed-form expression for $F_{p_1^*}$ exists in the cusp example, p_1^* is known and can be used to obtain $F_{p_1^*}$ numerically. From this, the log-odds for both experts can be found for each of our n events:

$$\lambda_1(A_i) = \log\left(\frac{F_{p_1^*}(\phi_i)}{1 - F_{p_1^*}(\phi_i)}\right) \text{ and } \lambda_2(\mathrm{A_i}) = \log\left(\frac{\mathrm{F_{p_2}}(\phi_i)}{1 - \mathrm{F_{p_2}}(\phi_i)}\right)$$

for $i = 1, \ldots, n$. Finally, as described in the group decision maker (GDM) approach of Section 5.2, the posterior consensus log-odds can be found using:

$$l_0(A \mid \underline{\lambda}) = (\underline{m} - \overline{\underline{m}})V^{-1}\left[\underline{\lambda} - \frac{1}{2}(\underline{m} + \overline{\underline{m}})\right] + l_0(A) \tag{5.1}$$

for $i = 1, \ldots, n$, where the group (size 2 here) must decide on values for 7 hyperparameters. m_1 is the log-odds which Expert 1 would ascribe to event A when A subsequently occurs, and $\overline{m}_1$ is the log-odds which Expert 1 would ascribe to A when A subsequently does not occur. m_2 and $\overline{m}_2$ are similarly defined for Expert 2. s_1 is the standard deviation for Expert 1's log-odds predictions about A, s_2 is the standard deviation for Expert 2's log-odds predictions about A, and ρ is the correlation between the two experts. Finally, let $l_0(A) = 1$, indicating a noninformative prior for the GDM.

At this point, a distribution for French-Lindley's consensus log-odds has been derived, but often a probability density will be more useful when forming inference

about ϕ. To this end, a consensus cumulative probability distribution function can be formed via

$$F_{p_0}(\phi_i \mid \underline{\lambda}) = \frac{\exp(l_0(A_i \mid \underline{\lambda}))}{1 + \exp(l_0(A_i \mid \underline{\lambda}))}.$$

The consensus pdf $p_0(\phi \mid \underline{\lambda})$ (which will be referred to as $p(\phi)$ for short) can then be approximated using differences $F_{p_0}(\phi_{i+1} \mid \underline{\lambda}) - F_{p_0}(\phi_i \mid \underline{\lambda})$ and scaling $p(\phi)$ so that it integrates to 1. Where $F_{p_0}(\phi_{i+1} \mid \underline{\lambda}) - F_{p_0}(\phi_i \mid \underline{\lambda}) < 0$, a failure of relative propensity consistency (see Section 2.2.1 and discussed further in Section 5.4) has been encountered. In these cases, $p(\phi_{i+1})$ is set to 0. Although this approach is somewhat ad-hoc, we felt that setting $p(\phi_{i+1}) = 0$ provided a simple, sensible fix to a serious flaw in the French-Lindley methodology. Obviously, as a probability distribution, $p(\phi_{i+1})$ can not have negative values. Another potential fix would be to add a constant value equal to $c_{\min} = \min(F_{p_0}(\phi_{i+1} \mid \underline{\lambda}) - F_{p_0}(\phi_i \mid \underline{\lambda}))$ to each $p(\phi_{i+1})$ and scale $p(\phi)$ so it integrates to 1. Unfortunately, this fix carries the undesirable side effect of changing probability relationships among points. For instance, if $p(\widehat{\phi_{j_1}}) = 0.3$ and $p(\widehat{\phi_{j_2}}) = 0.1$ but $c_{\min} = 0.1$ is used to adjust negative probabilities, then ϕ_{j_1} is only 2 times more likely than ϕ_{j_2} under the adjusted probability distribution instead of 3 times more likely under the unadjusted distribution. Setting negative $p(\phi_{i+1})$ to 0 preserves probability relationships while applying a band-aid to the RPC failure of the French-Lindley method. As a result, violations of RPC often get hidden even though they are a real and common problem as we investigate in Section 5.4.

Some adjustments to this procedure are required in higher dimensions; these are discussed in Section 5.3.2.

Introductory Example

The sensitivity of the pooling to various values of the hyperparameters for the introductory $\Re^1 \rightarrow \Re^1$ example of Section 4.2 is illustrated in Figures 5.1-

5.7. In developing these figures, we chose what we felt to be a reasonable set of hyperparameters from which variations could be explored. Our original baseline set of values was $m_1 = m_2 = 2$, $\overline{m}_1 = \overline{m}_2 = -2$, $s_1 = s_2 = 1$, and $\rho = 0$. With these settings, the mean probability that an expert assigns to an event A which subsequently occurs is .88 with approximate 95% confidence interval (.50, .98), assuming normality in the log-odds. The mean probability that an expert assigns to an event A which subsequently does not occur is .12, with approximate 95% confidence interval (.02, .50). In addition, the two experts are assumed to be independent with respect to their information sources which contribute to their prior beliefs. Unfortunately, these values, especially $s_1 = s_2 = 1$ and $\rho = 0$, produce extremely narrow, unrealistic pooled distributions. Thus, for illustrative purposes, variations are explored from the set of hyperparameters $m_1 = m_2 = 2$, $\overline{m}_1 = \overline{m}_2 = -2$, $s_1 = s_2 = 2$, and $\rho = 0.7$.

Figure 5.1 shows the effect of changes in variability, assuming that $s_1 = s_2 = s$ and letting this common standard deviation s vary from 0.5 to 4.0. As expected, the variability in the consensus prior increases as s increases. It is disturbing, however, that the pooling can be considerably more precise than either of the original priors, as is found in plots (a) and (b) with $s \leq 1.0$. Plots for s =0.5, 1.0, and 2.0 show consensus priors with modes between the modes of $p_2(\phi)$ and $p_1^*(\phi)$ as one might expect, but plot (d) for $s = 4$ shows an unusual bimodal consensus prior that does not correspond to an intuitive notion of what a pooling should look like.

Figure 5.2 allows the correlation ρ to vary from 0 to 0.9. As ρ increases, the mode of the consensus prior remains fairly constant while the variability increases, likely due to the decrease in information about ϕ when the two experts are highly correlated. The consensus distribution under the independent case ($\rho = 0$), shown in plot (a), again features a distribution which is more precise than either of the

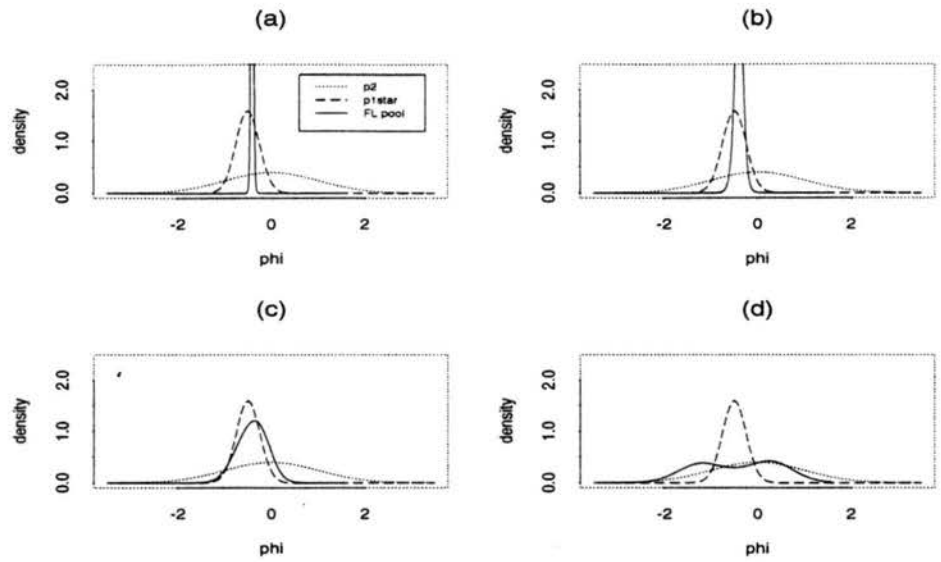

Figure 5.1: Introductory example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $m_1 = m_2 = 2$, $\overline{m}_1 = \overline{m}_2 = -2$, and $\rho = 0.7$. Variable hyperparameters: (a) $s_1 = s_2 = 0.5$, (b) $s_1 = s_2 = 1.0$, (c) $s_1 = s_2 = 2.0$, (d) $s_1 = s_2 = 4.0$.

original expert priors—a curious result given the incoherence between $p_2(\phi)$ and $p_1^*(\phi)$.

Figure 5.3 requires that $m_1 = m_2 = -\overline{m}_1 = -\overline{m}_2 = m$, while allowing the common value m to vary. Thus, higher values of m indicate that both experts offer higher probabilities for A when A subsequently occurs and lower probabilities when A subsequently does not occur. As m increases, the variability of the consensus prior decreases, and its mode shifts slightly toward the mode of p_1^*. In addition, increasing m, especially $m \geq 4$, produces a pooling which is more precise than either expert individually. So, unrealistic poolings result from the GDM attributing too much expertise (as measured by m) to the experts, even though the levels of expertise may appear reasonable on the surface to the GDM.

Figure 5.4 is similar to Figure 5.3, except that m_2 and $-\overline{m}_2$ are fixed at 2 and only m_1 and $-\overline{m}_1$ are allowed to vary, under the constraint $m_1 = -\overline{m}_1 = m$. Thus, the relative expertise of Expert 1 (the narrower prior) compared to Expert 2 is increasing as m increases. For $m = 1$, the consensus prior is basically equivalent

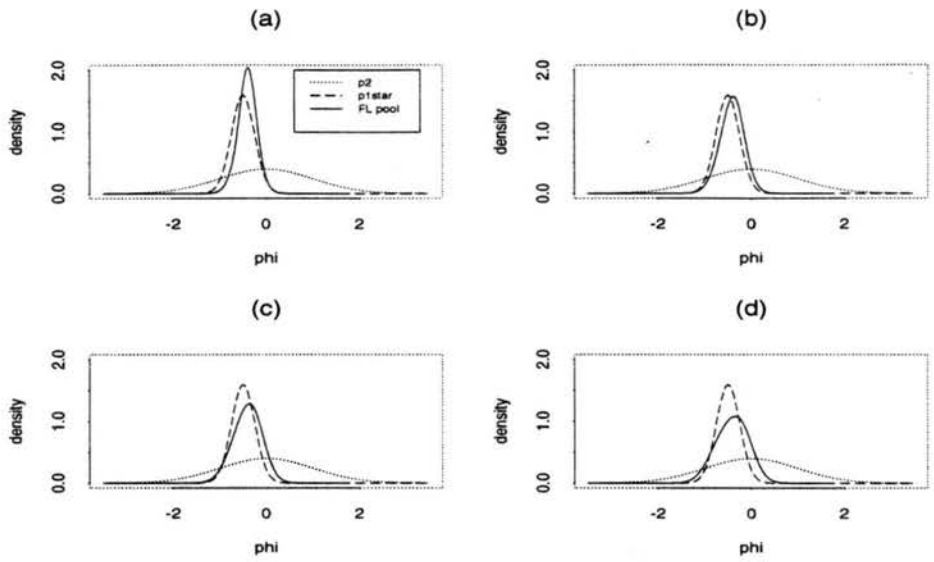

Figure 5.2: Introductory example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $m_1 = m_2 = 2, \overline{m}_1 = \overline{m}_2 = -2$, and $s_1 = s_2 = 2.0$. Variable hyperparameters: (a) $\rho = 0$, (b) $\rho = 0.3$, (c) $\rho = 0.6$, (d) $\rho = 0.9$.

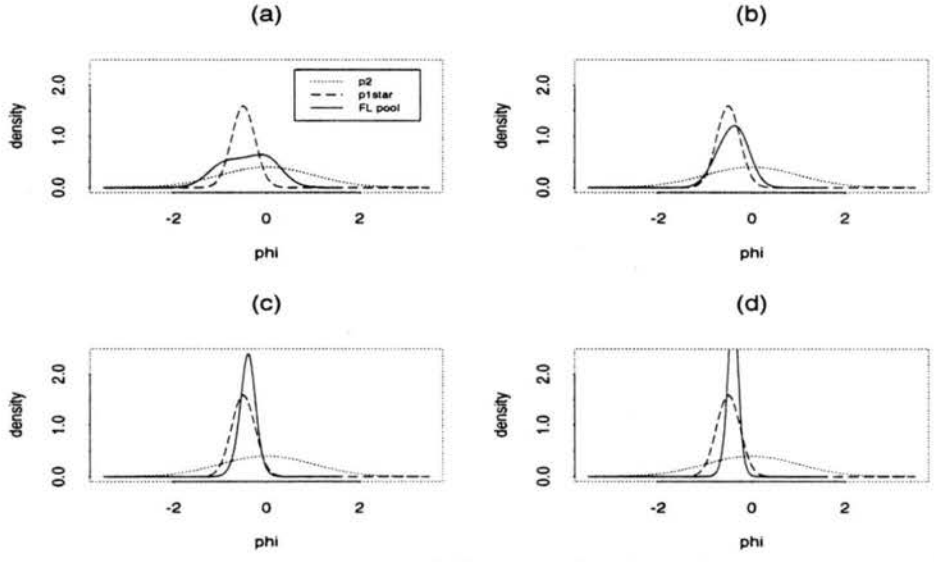

Figure 5.3: Introductory example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $s_1 = s_2 = 2.0$, and $\rho = 0.7$. Variable hyperparameters: (a) $m_1 = m_2 = -\overline{m}_1 = -\overline{m}_2 = 1$, (b) $m_1 = m_2 = -\overline{m}_1 = -\overline{m}_2 = 2$, (c) $m_1 = m_2 = -\overline{m}_1 = -\overline{m}_2 = 4$, (d) $m_1 = m_2 = -\overline{m}_1 = -\overline{m}_2 = 6$.

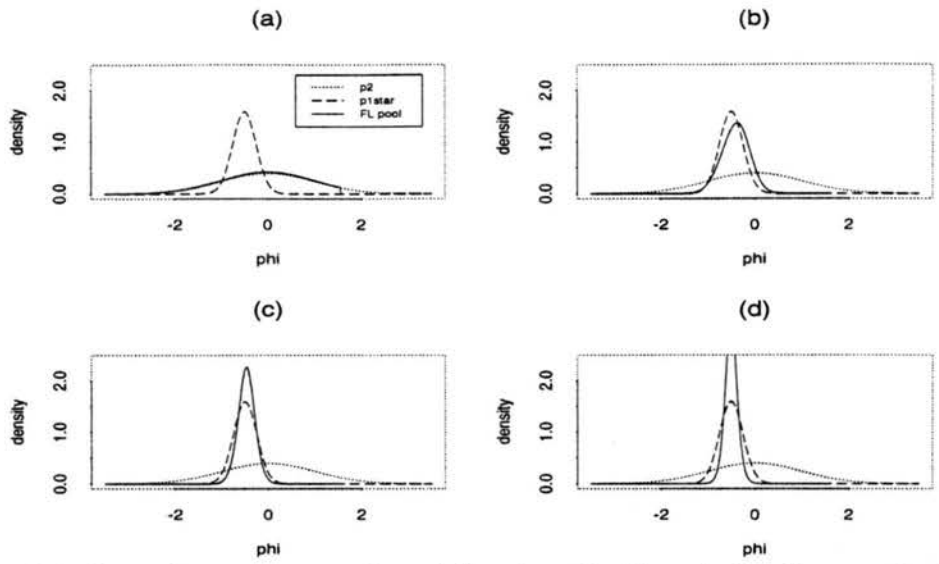

Figure 5.4: Introductory example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $m_2 = -\overline{m}_2 = 2$, $s_1 = s_2 = 2.0$, and $\rho = 0.5$. Variable hyperparameters: (a) $m_1 = -\overline{m}_1 = 1$, (b) $m_1 = -\overline{m}_1 = 2$, (c) $m_1 = -\overline{m}_1 = 3$, (d) $m_1 = -\overline{m}_1 = 4$.

to $p_2(\phi)$, while for higher values of m the consensus prior becomes a narrow distribution centered at the mode of $p_1^*(\phi)$. The large fluctuations in the pooled distributions resulting from slight changes in the relative expertise level of the two experts illustrate an unsettling lack of robustness to hyperparameter settings for the French-Lindley approach. Figure 5.5 is analogous to Figure 5.4, except that the mean log-odds for Expert 1 remains fixed while that of Expert 2 varies. For $m = 1$, the consensus prior is essentially equivalent to $p_1^*(\phi)$, while for higher values of m the consensus prior becomes a narrow distribution centered at the mode of $p_2(\phi)$.

Figure 5.6 allows s_1 to vary while all other hyperparameters remain fixed. Therefore, the mean log-odds for Expert 1's statement about event A when A subsequently occurs is fixed at 2, but the variability in this log-odds is allowed to vary between 0.5 and 4.0. For small s_1, the consensus prior is an extremely narrow distribution centered at the mode of $p_1^*(\phi)$, while for large s_1 the consensus prior approaches $p_2(\phi)$. Again, pooling produced by the French-Lindley method exhibit

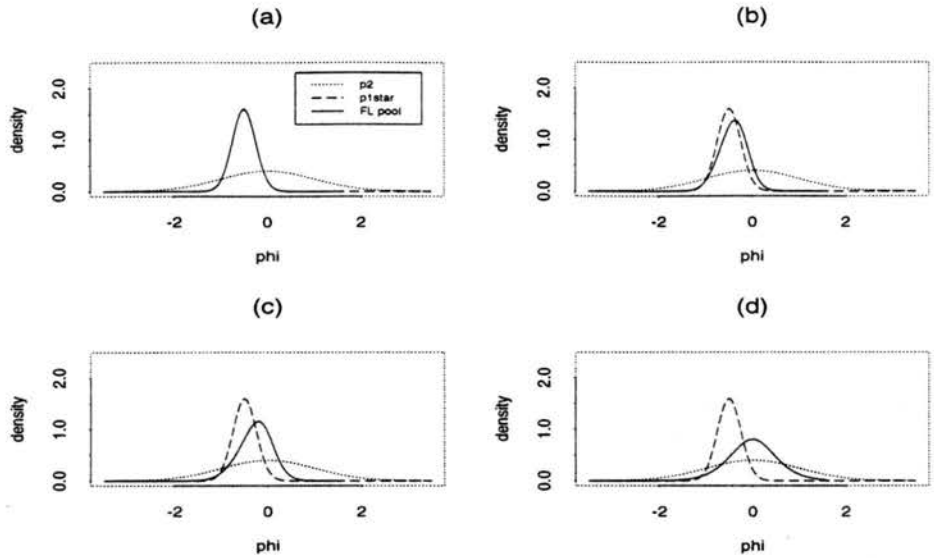

Figure 5.5: Introductory example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $m_1 = -\overline{m}_1 = 2$, $s_1 = s_2 = 2.0$, and $\rho = 0.5$. Variable hyperparameters: (a) $m_2 = -\overline{m}_2 = 1$, (b) $m_2 = -\overline{m}_2 = 2$, (c) $m_2 = -\overline{m}_2 = 3$, (d) $m_2 = -\overline{m}_2 = 4$.

little robustness to variations in precision settings, and unreasonable consensus distributions can result from the unfortunate selection of s_1 and s_2.

Figure 5.7 is analogous to Figure 5.6, except that s_2 is the only hyperparameter allowed to vary. As with Figure 5.6, the consensus prior for large s_2 approaches $p_1^*(\phi)$, but unusual things occur for small s_2. At $s_2 = 1$, the consensus prior is a narrow distribution centered near the mode of $p_2(\phi)$, but at $s_2 = 0.5$, the consensus mode shifts to the right of the mode of $p_2(\phi)$. Again this suggests that any overstatement of precision in the hyperparameters can lead to nonsensical French-Lindley poolings.

Cusp Example

The sensitivity of the pooling to various values of the hyperparameters for the cusp example of Section 4.3 are illustrated in Figures 5.8- 5.10. In Figure 5.8, s_1 and s_2 are both fixed at 2. Then, plots (a) and (b) show the effect of ρ. As ρ increases so that the two experts are more strongly dependent, the consensus prior flattens out, and its mode shifts left. Greater variability in the pooled

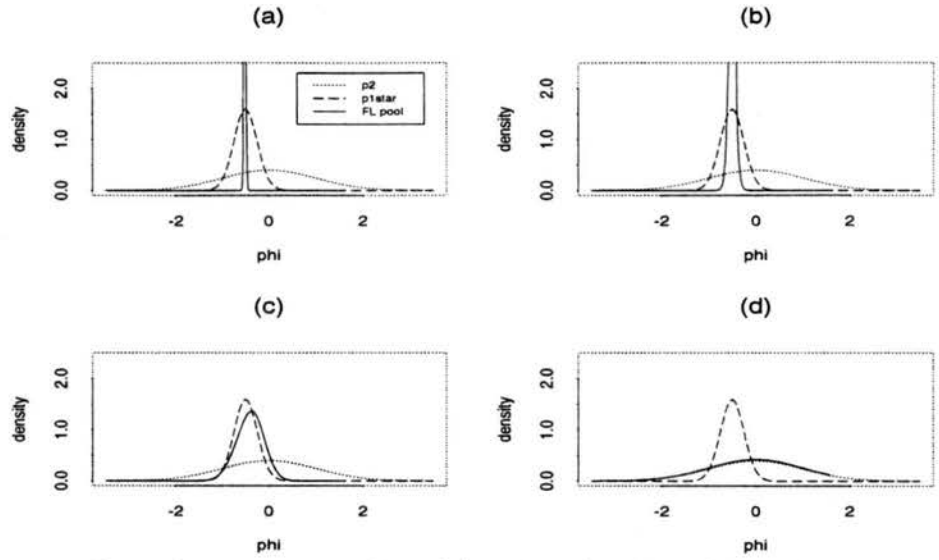

Figure 5.6: Introductory example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $m_1 = m_2 = 2$, $\overline{m}_1 = \overline{m}_2 = -2$, $s_2 = 2$, and $\rho = 0.5$. Variable hyperparameters: (a) $s_1 = 0.5$, (b) $s_1 = 1.0$, (c) $s_1 = 2.0$, (d) $s_1 = 4.0$.

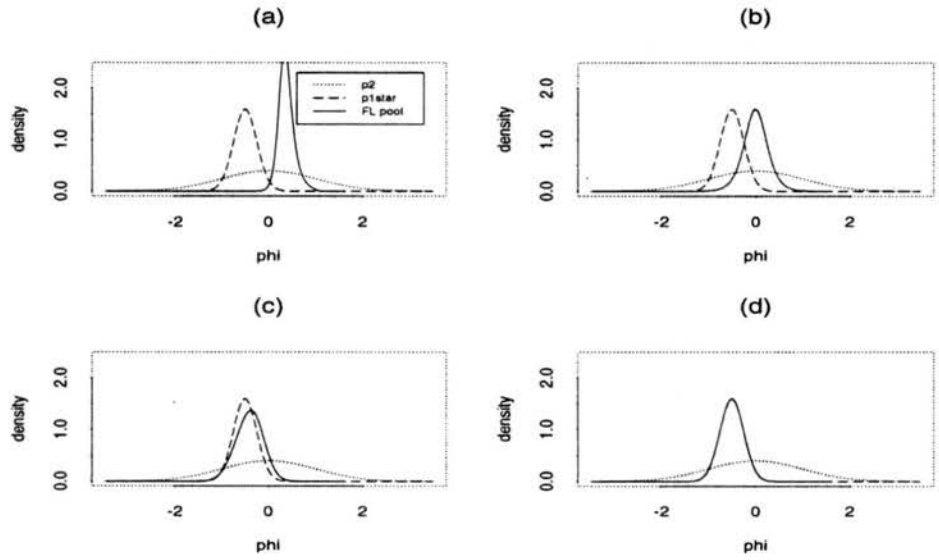

Figure 5.7: Introductory example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $m_1 = m_2 = 2$, $\overline{m}_1 = \overline{m}_2 = -2$, $s_1 = 2$, and $\rho = 0.5$. Variable hyperparameters: (a) $s_2 = 0.5$, (b) $s_2 = 1.0$, (c) $s_2 = 2.0$, (d) $s_2 = 4.0$.

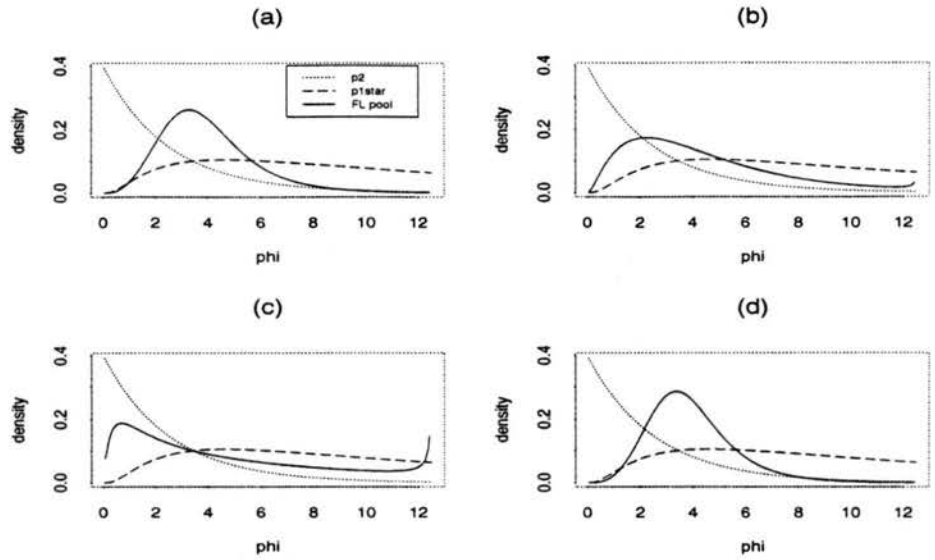

Figure 5.8: Cusp example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $s_1 = s_2 = 2$. Variable hyperparameters: (a) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.1$; (b) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.9$; (c) $m_1 = -\overline{m}_1 = 1$, $m_2 = -\overline{m}_2 = 1$, and $\rho = 0.5$; (d) $m_1 = -\overline{m}_1 = 3$, $m_2 = -\overline{m}_2 = 3$, and $\rho = 0.5$.

distribution with greater correlation is expected since less information about ϕ is available from the two priors. Plots (c) and (d) fix $\rho = 0.5$ but allow m to vary, where $m_1 = m_2 = -\overline{m}_1 = -\overline{m}_2 = m$. As m increases, the consensus prior becomes more concentrated about $\phi \doteq 3.5$, but for $m = 1$ the consensus prior flattens out, with a mode just below $\phi = 1$ and an unusual spike near $\phi = 12.5$. Reflecting a lower estimate of the experts' abilities, $m = 1$ leads to more variability in the pooled distribution as expected. However, the truncation inherent to this example is handled poorly, since $p(\phi)$ should not be large for $\phi > 12$.

In Figure 5.9, $m = 2$ and $\rho = 0.5$ are fixed, and different combinations of s_1 and s_2 are analyzed. When $s_1 = s_2$, as in plots (a) and (b), an effect is seen similar to that in Figure 5.8 (c,d) where m was allowed to vary. Low s_1 and s_2 produce a narrow distribution with mode just below $\phi = 4$, similar to the consensus prior produced with high m. On the other hand, the very disturbing consensus prior with secondary mode near $\phi = 12.5$ is observed when s_1 and s_2 are high (as when m is low). In plot (c), s_1 is low and s_2 is high, and the consensus distribution

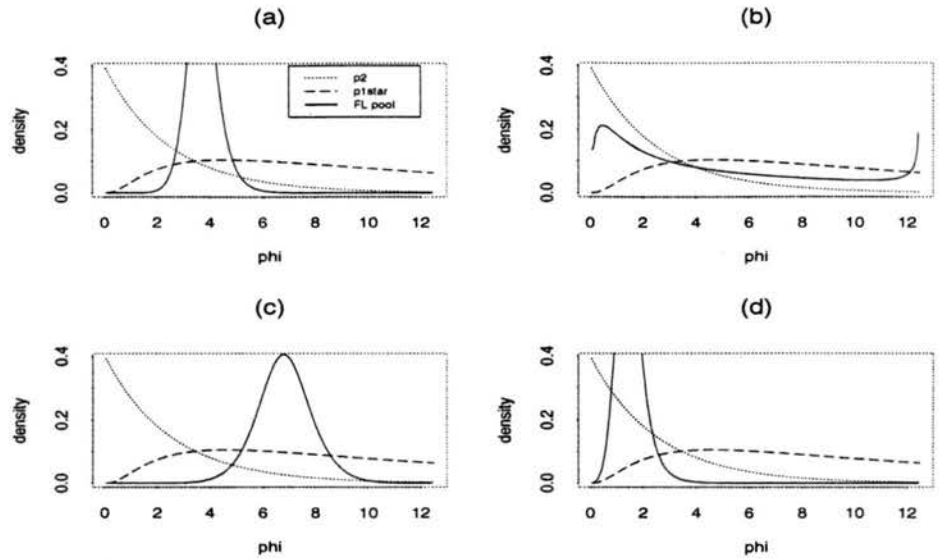

Figure 5.9: Cusp example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $m_1 = -\overline{m}_1 = m_2 = -\overline{m}_2 = 2$, and $\rho = 0.5$. Variable hyperparameters: (a) $s_1 = 1, s_2 = 1$; (b) $s_1 = 3, s_2 = 3$; (c) $s_1 = 1, s_2 = 3$; (d) $s_1 = 3, s_2 = 1$.

shifts right, away from $p_2(\phi)$. In plot (d), s_2 is low and s_1 is high, and the reverse is seen—a shift left, towards $p_2(\phi)$. It seems odd, however, that the consensus distribution would have a mode as high as $\phi = 6.5$, since Expert 1 assigns greater probability to smaller ϕ, and Expert 2 assigns almost no probability near $\phi = 6.5$. It is also troubling that the poolings shown here bear very little relationship to the original priors, especially in terms of shape.

In Figure 5.10, s_1 and s_2 are fixed, and different combinations of m_1, m_2, $-\overline{m}_1$, and $-\overline{m}_2$ are analyzed for $\rho = 0.2$ and $\rho = 0.7$, subject to the constraint that $m_1 = -\overline{m}_1$ and $m_2 = -\overline{m}_2$. Plots (a) and (b) feature superior performance by Expert 1 (i.e. $m_1 > m_2$ and $\overline{m}_1 < \overline{m}_2$). With $\rho = 0.7$ in plot (a), the consensus prior has a mode near $\phi = 12$, while the mode shifts left for $\rho = 0.2$ in plot (b). Again these results are not at all appealing; the mode near $\phi = 12$ is especially illogical. Plots (c) and (d) feature superior performance by Expert 2. As one might expect, these consensus distributions resemble $p_2(\phi)$ much more than $p_1^*(\phi)$. The mode shifts right as ρ decreases, which seems logical as Expert 1 receives more

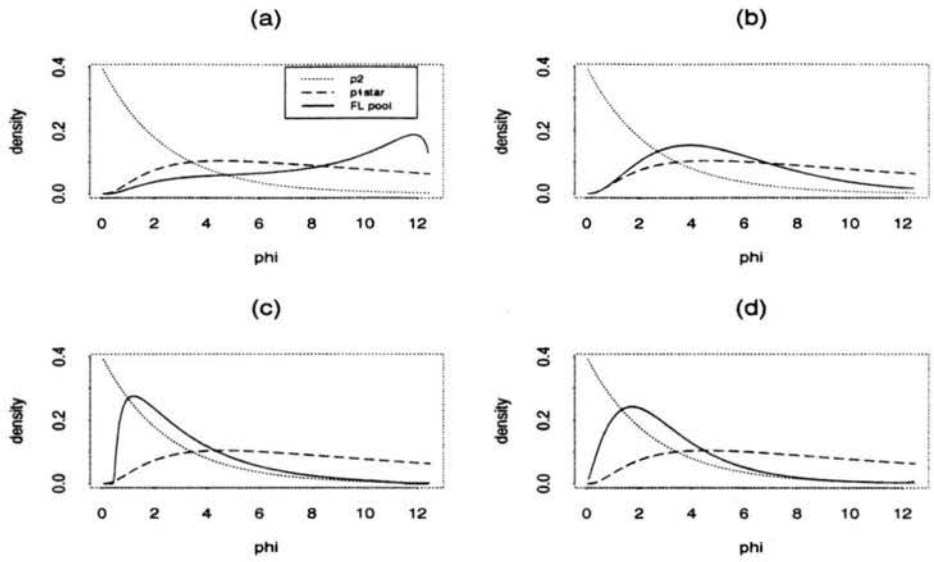

Figure 5.10: Cusp example: $p(\phi)$ using the French-Lindley method. Fixed hyperparameters: $s_1 = s_2 = 2$. Variable hyperparameters: (a) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 1$, and $\rho = 0.7$; (b) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 1$, and $\rho = 0.2$; (c) $m_1 = -\overline{m}_1 = 1$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.7$; (d) $m_1 = -\overline{m}_1 = 1$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.2$.

weight than in the case where he is highly correlated with a supposedly superior Expert 2.

5.3.2 $\Re^2 \rightarrow \Re^2$ Examples

Application of the French-Lindley method to examples with 2 inputs and outputs follows the procedure outlined in Section 5.3.1 with a few modifications. First, an equally spaced 2-dimensional grid of ϕ-points $(\phi_{1,i}, \phi_{2,j})$ is defined for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, where $\phi_{1,1} < \phi_{1,2} < \ldots < \phi_{1,m}$ and $\phi_{2,1} < \phi_{2,2} < \ldots < \phi_{2,n}$. Consider a set of mn binary events, labeled A_{ij}, which correspond to 2-dimensional cumulative probabilities. Then $P(A_{ij}) = P(\phi_1 < \phi_{1,i}, \phi_2 < \phi_{2,j}) = F(\phi_{1,i}, \phi_{2,j})$ where $F(\phi_1, \phi_2)$ is the appropriate cdf. Note that $F_{p_1^*}$ and F_{p_2} are known for the linear example, and F_{p_2} is known for the noninvertible example. Although no closed-form expression for $F_{p_1^*}$ exists in the noninvertible example, p_1^* is known and can be used to obtain $F_{p_1^*}$ numerically. Equation 5.1 can again be used to find the GDM's posterior log-odds (i.e. the

consensus log-odds for the experts) if the GDM has chosen values for the 7 hyperparameters $(m_1, \overline{m}_1, m_2, \overline{m}_2, s_1, s_2, \rho)$. In this case, the hyperparameters relate to the experts' abilities to state probabilities about areas A_{ij} in ϕ-space where $A_{ij} = \{\text{event that } \phi_1 < \phi_{1,\mathrm{i}} \text{ and } \phi_2 < \phi_{2\mathrm{j}}\}$.

As in the $\Re^1 \rightarrow \Re^1$ examples, consensus prior densities $p(\phi)$ can be derived from the consensus prior log-odds $l_0(A_{ij} \mid \underline{\lambda})$ to allow more natural inference regarding ϕ to be formed. Recall that estimation of the French-Lindley pooled probability distribution in $\Re^1 \rightarrow \Re^1$ problems followed these steps:

1. Obtain $F_1(A_2)$, $F_1(A_1)$, $F_2(A_2)$, and $F_2(A_1)$, where $\phi_2 = \phi_1 + \epsilon$ and $F_i(A_j) = P_i(\phi < \phi_j)$ for $i = 1, 2$ and $j = 1, 2$.

2. Calculate log-odds $\lambda_i(A_j) = \log\left(\frac{F_i(A_j)}{1-F_i(A_j)}\right)$.

3. Derive the French-Lindley GDM posterior log-odds $l_0(A_j \mid \underline{\lambda}) = c_1\lambda_1(A_j) + c_2\lambda_2(A_j) + c_3$, where the constants $c_i, i = 1, 2, 3$, are functions of the 7 hyperparameters.

4. Convert the GDM's posterior log-odds to cumulative probabilities according to

$$F_0(A_j \mid \underline{\lambda}) = \frac{\exp(l_0(A_j \mid \underline{\lambda}))}{1 + \exp(l_0(A_j \mid \underline{\lambda}))}.$$

5. Estimate $p_0(\phi_2 \mid \underline{\lambda})$ by $F_0(A_2 \mid \underline{\lambda}) - F_0(A_1 \mid \underline{\lambda})$.

So that the estimate of $p_0(\phi_2 \mid \underline{\lambda})$ at our final step is legitimate (i.e. non-negative), it is vital that the following conditions hold at each step:

1. $F_1(A_2) \geq F_1(A_1)$ and $F_2(A_2) \geq F_2(A_1)$. This is true because F_1 and F_2 are cumulative probability distributions and $\phi_2 > \phi_1$.

2. $\lambda_1(A_2) \geq \lambda_1(A_1)$ and $\lambda_1(A_2) \geq \lambda_1(A_1)$. This holds because:

$$F_1(A_2) \geq F_1(A_1) \Rightarrow$$

$$\frac{F_1(A_2)}{F_1(A_1)} \geq 1 \text{ and } \frac{1-F_1(A_1)}{1-F_1(A_2)} \geq 1 \Rightarrow$$
$$\frac{F_1(A_2)}{1-F_1(A_2)} \cdot \frac{1-F_1(A_1)}{F_1(A_1)} \geq 1 \Rightarrow$$
$$\log\left(\frac{F_1(A_2)}{1-F_1(A_2)}\right) - \log\left(\frac{F_1(A_1)}{1-F_1(A_1)}\right) \geq 0 \Rightarrow$$
$$\lambda_1(A_2) - \lambda_1(A_1) \geq 0.$$

3. $l_0(A_2 \mid \underline{\lambda}) \geq l_0(A_1 \mid \underline{\lambda})$. This holds if relative propensity consistency holds.

4. $F_0(A_2 \mid \underline{\lambda}) \geq F_0(A_1 \mid \underline{\lambda})$. This holds because:

$$l_0(A_2 \mid \underline{\lambda}) \geq l_0(A_1 \mid \underline{\lambda}) \Rightarrow$$
$$\log\left(\frac{F_0(A_2 \mid \underline{\lambda})}{1-F_0(A_2 \mid \underline{\lambda})}\right) - \log\left(\frac{F_0(A_1 \mid \underline{\lambda})}{1-F_0(A_1 \mid \underline{\lambda})}\right) \geq 0 \Rightarrow$$
$$\frac{F_0(A_2 \mid \underline{\lambda})}{1-F_0(A_2 \mid \underline{\lambda})} \cdot \frac{1-F_0(A_1 \mid \underline{\lambda})}{F_0(A_1 \mid \underline{\lambda})} \geq 1 \Rightarrow$$
$$F_0(A_2 \mid \underline{\lambda}) - F_0(A_2 \mid \underline{\lambda})F_0(A_1 \mid \underline{\lambda}) \geq F_0(A_1 \mid \underline{\lambda}) - F_0(A_2 \mid \underline{\lambda})F_0(A_1 \mid \underline{\lambda}) \Rightarrow$$
$$F_0(A_2 \mid \underline{\lambda}) \geq F_0(A_1 \mid \underline{\lambda}).$$

5. $p_0(\phi_2 \mid \underline{\lambda}) = F_0(A_2 \mid \underline{\lambda}) - F_0(A_1 \mid \underline{\lambda}) \geq 0$ following Item 4.

Thus, in $\Re^1 \rightarrow \Re^1$ examples it follows that, using this estimation procedure, $p_0(\phi_2 \mid \underline{\lambda}) \geq 0$ if and only if RPC holds. The remaining conditions above are always true.

To extend this estimation process to 2-dimensional $\boldsymbol{\phi}$, one would naturally consider using the following relationship:

$$\begin{aligned} p(\phi_{1,i}, \phi_{2,j}) \doteq\ & F_{p_0}(\phi_{1,i}, \phi_{2,j}) - F_{p_0}(\phi_{1,i-1}, \phi_{2,j}) \\ & - F_{p_0}(\phi_{1,i}, \phi_{2,j-1}) + F_{p_0}(\phi_{1,i-1}, \phi_{2,j-1}). \end{aligned}$$

Unfortunately, using this relationship can lead to $p(\phi_{1,i}, \phi_{2,j}) < 0$ even in cases when RPC does hold. This is because

$$l_0(A_4 \mid \underline{\lambda}) - l_0(A_3 \mid \underline{\lambda}) - l_0(A_2 \mid \underline{\lambda}) + l_0(A_1 \mid \underline{\lambda}) \geq 0$$

does not imply

$$F_0(A_4 \mid \underline{\lambda}) - F_0(A_3 \mid \underline{\lambda}) - F_0(A_2 \mid \underline{\lambda}) + F_0(A_1 \mid \underline{\lambda}) \geq 0.$$

In other words, there is no analogous condition to condition 4 above if we attempt to use this conventional 4-event method for estimating 2-dimensional probability distributions.

To obtain valid French-Lindley estimates for the consensus prior $p(\boldsymbol{\phi})$ with 2-dimensional $\boldsymbol{\phi}$, we must consider just 2 carefully selected events. This is because relative propensity consistency will hold for any two events A_2 and A_1 whether these events are constructed in one dimension or many, but logical extensions involving more than two events do not necessarily hold. Thus, we should consider the two events $A_{ij} = \{\text{event that } \phi_1 < \phi_{1,\mathrm{i}} \text{ and } \phi_2 < \phi_{2,\mathrm{j}}\}$ and $B_{ij} = \{\text{event that } (\phi_1 < \phi_{1,\mathrm{i}-1} \text{ and } \phi_2 < \phi_{2,\mathrm{j}}) \text{ or } (\phi_1 < \phi_{1,\mathrm{i}} \text{ and } \phi_2 < \phi_{2,\mathrm{j}-1})\}$. Then we can use the process described above, and estimate the French-Lindley pooled prior $p_0(\phi_{1,i}, \phi_{2,j} \mid \underline{\lambda})$ with $F_0(A_{ij} \mid \underline{\lambda}) - F_0(B_{ij} \mid \underline{\lambda})$. If we find that $p_0(\phi_{1,i}, \phi_{2,j} \mid \underline{\lambda}) < 0$, then RPC has been violated, and the pooled prior is set to 0 at those points.

Linear $\Re^2 \rightarrow \Re^2$ Example

The sensitivity of the pooling to various values of the hyperparameters for the linear $\Re^2 \rightarrow \Re^2$ example of Section 4.4 is illustrated in Figures 5.11- 5.13, using the same baseline set of hyperparameters described in the introductory example section. The French-Lindley joint consensus distribution is illustrated by the image plot, with darker shading indicating regions of higher $p(\boldsymbol{\phi})$. Hyperparameter settings in Figures 5.11- 5.13 are identical to those in Figures 5.8- 5.10 for the cusp example. Thus, in Figure 5.11, plots (a) and (b) show the effect of ρ; plot (b), with $\rho = 0.9$, shows greater variability in $p(\boldsymbol{\phi})$ since less information about $\boldsymbol{\phi}$ is provided by highly dependent experts. More variability is also seen in plot (c)

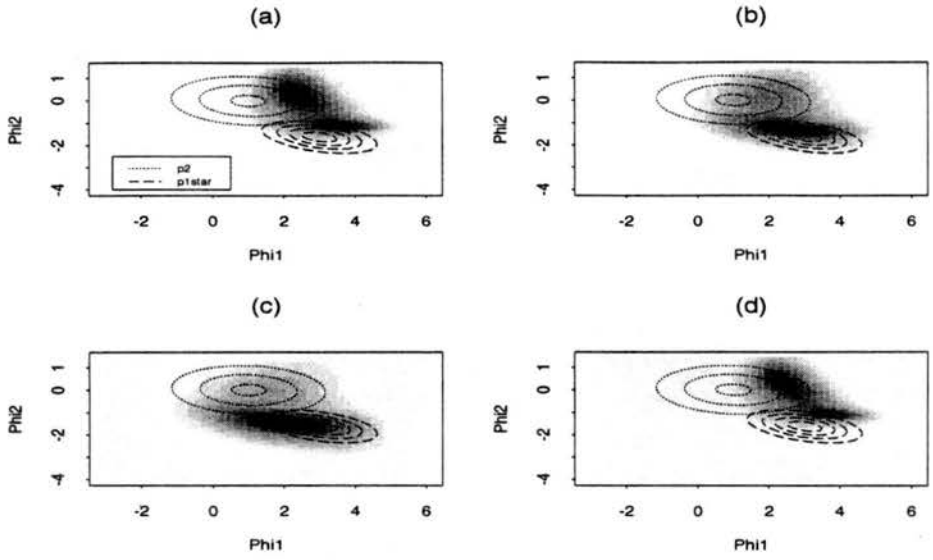

Figure 5.11: Linear $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) using the French-Lindley method. Fixed hyperparameters: $s_1 = s_2 = 2$. Variable hyperparameters: (a) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.1$; (b) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.9$; (c) $m_1 = -\overline{m}_1 = 1$, $m_2 = -\overline{m}_2 = 1$, and $\rho = 0.5$; (d) $m_1 = -\overline{m}_1 = 3$, $m_2 = -\overline{m}_2 = 3$, and $\rho = 0.5$.

compared to plot (d) due to a more pessimistic evaluation of the mean predictive abilities of both experts.

Note that the mode of $p(\phi)$ does not lie between the modes of $p_1^*(\phi)$ and $p_2(\phi)$ in these plots as one might expect. This troublesome observation persists throughout the remaining linear $\Re^2 \rightarrow \Re^2$ example plots. In general, the application of the French-Lindley approach to higher dimensional models seems incapable of producing sensible poolings for most seemingly reasonable hyperparameter values.

In Figure 5.12, $m = 2$ and $\rho = 0.5$ are fixed, and different combinations of s_1 and s_2 are analyzed. A tight distribution is observed in plot (a) with $s_1 = s_2 = 1$, whereas much more spread is apparent in plot (b) with $s_1 = s_2 = 3$. The mode of the consensus distribution shifts down and to the left as the variances increase, a trend which is not easily explained. In plot (c), with Expert 1's variance on log-odds smaller than that of Expert 2, the consensus distribution approaches $p_1^*(\phi)$. Conversely, in plot (d) with $s_1 > s_2$, the consensus distribution approaches

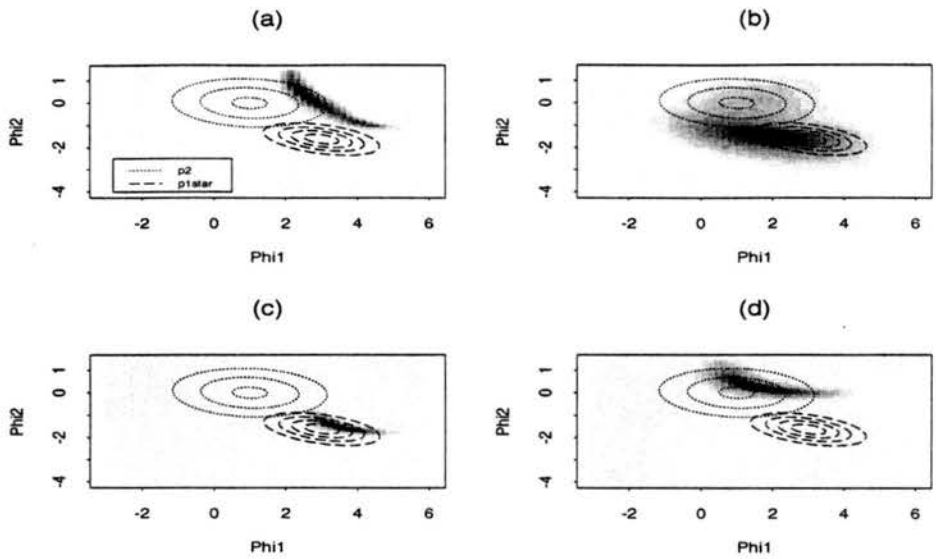

Figure 5.12: Linear $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) using the French-Lindley method. Fixed hyperparameters: $m_1 = -\overline{m}_1 = m_2 = -\overline{m}_2 = 2$, and $\rho = 0.5$. Variable hyperparameters: (a) $s_1 = 1, s_2 = 1$; (b) $s_1 = 3, s_2 = 3$; (c) $s_1 = 1, s_2 = 3$; (d) $s_1 = 3, s_2 = 1$.

$p_2(\phi)$, although in both cases the French-Lindley pool is located above and to the right of the original prior's mode. These plots illustrate that any misstatement of precision can largely impact the resulting pooled distribution.

In Figure 5.13, plots (a) and (b) feature superior performance by Expert 1 (i.e. $m_1 > m_2$ and $\overline{m}_1 < \overline{m}_2$), evaluated at $\rho = 0.7$ and $\rho = 0.2$. The mass of both consensus joint priors is located near the mass of $p_1^*(\phi)$, with a shift to the southwest as ρ increases. In comparison, plots (c) and (d) feature superior performance by Expert 2, also evaluated at $\rho = 0.7$ and $\rho = 0.2$. The mass of $p(\phi)$ for $\rho = 0.2$ approaches $p_2(\phi)$, although we see again the unsatisfactory trend for $p(\phi)$ to lie above and to the right of what one might expect. In plot (c), with $\rho = 0.7$, the French-Lindley method completely falls apart, due in part to a failure of relative propensity consistency and in part to an inability to process seemingly reasonable sets of hyperparameter values.

Noninvertible $\Re^2 \rightarrow \Re^2$ Example

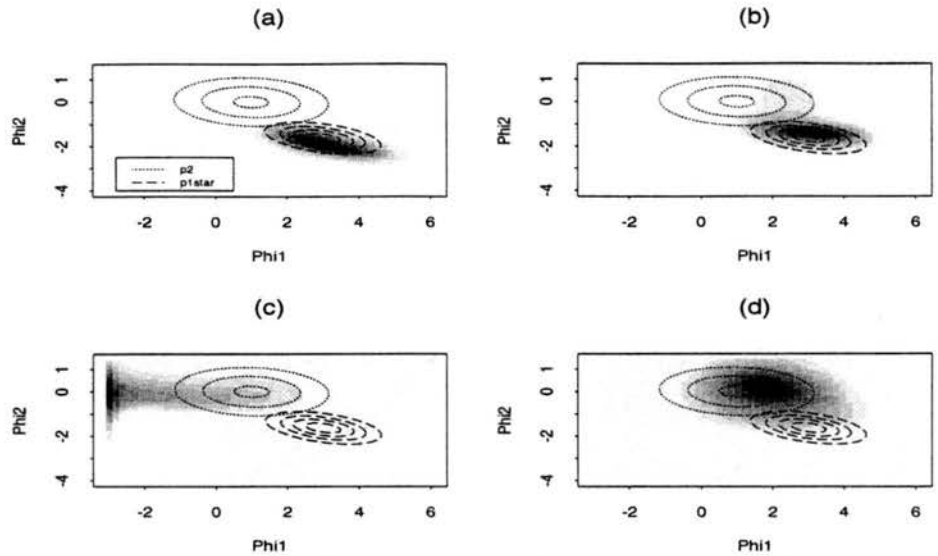

Figure 5.13: Linear $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) using the French-Lindley method. Fixed hyperparameters: $s_1 = s_2 = 2$. Variable hyperparameters: (a) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 1$, and $\rho = 0.7$; (b) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 1$, and $\rho = 0.2$; (c) $m_1 = -\overline{m}_1 = 1$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.7$; (d) $m_1 = -\overline{m}_1 = 1$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.2$.

The sensitivity of the pooling to various values of the hyperparameters for the noninvertible $\Re^2 \rightarrow \Re^2$ example of Section 4.5 is illustrated in Figures 5.14-5.16. Hyperparameter settings in these plots are identical to those in the past two examples. Plots (a) and (b) of Figure 5.14 illustrate increased variability in $p(\phi)$ with increased ρ, just as noted in previous examples. Because of the dispersion in $p_1^*(\phi)$, the poolings produced in this example are largely influenced by Expert 2. Plots (c) and (d) also show a tendency observed in previous plots—that of increased variability in $p(\phi)$ with lower performance of both experts, as measured by the log-odds conditional on A. Note that this noninvertible model produces a line of truncation in ϕ-space, where $p_1^*(\phi) = 0$ above and to the left of the line. Many hyperparameter combinations produce areas of high density along the line of truncation which are difficult to observe in the plots.

In Figure 5.15, effects of different combinations of s_1 and s_2 are analyzed. As might be expected, small $s_1 = s_2$ produces a concentrated $p(\phi)$ as in plot (a), while larger $s_1 = s_2$ produce more dispersed $p(\phi)$ as in plot (b). In addition, low

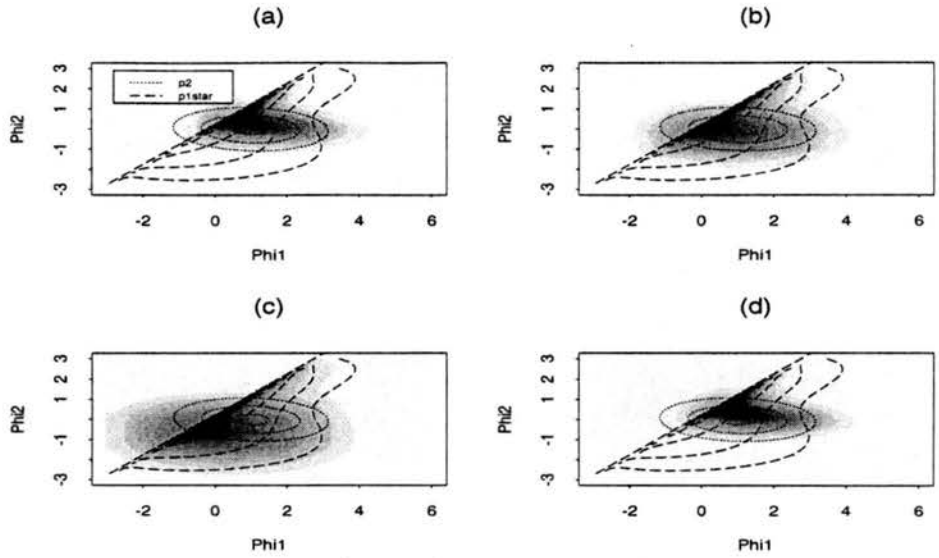

Figure 5.14: Noninvertible $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) using the French-Lindley method. Fixed hyperparameters: $s_1 = s_2 = 2$. Variable hyperparameters: (a) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.1$; (b) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.9$; (c) $m_1 = -\overline{m}_1 = 1$, $m_2 = -\overline{m}_2 = 1$, and $\rho = 0.5$; (d) $m_1 = -\overline{m}_1 = 3$, $m_2 = -\overline{m}_2 = 3$, and $\rho = 0.5$.

s_1 compared to s_2, such as in plot (c), leads to $p(\phi)$ concentrated near areas of high mass for $p_1^*(\phi)$, while low s_2 compared to s_1 as in plot (d) leads to $p(\phi)$ concentrated near areas of high mass for $p_2(\phi)$. Changes in s_1 and s_2 have the anticipated effects on the pooled distributions, but $p(\phi)$ in most cases seems to skirt the mode of Expert 2's beliefs high and right or low and left, generating a consensus which is somewhat less than ideal.

Plots (a) and (b) of Figure 5.16 feature superior performance by Expert 1 (i.e. $m_1 > m_2$ and $\overline{m}_1 < \overline{m}_2$). With $\rho = 0.7$ as in plot (a), mass of $p(\phi)$ is concentrated near the truncation line as expected, although some mass is inexplicably located where both ϕ_1 and ϕ_2 are unusually high or unusually low. With $\rho = 0.2$ as in plot (b), more of a balance between $p_1^*(\phi)$ and $p_2(\phi)$ is noted. Plots (c) and (d) feature superior performance by Expert 2, and not surprisingly the mass of $p(\phi)$ is concentrated near $p_2(\phi)$, especially when correlation is high. $p(\phi)$ in these plots seems to be a reasonable synthesis of the beliefs of Experts 1 and 2 given the hyperparameter settings. Note that the French-Lindley method as

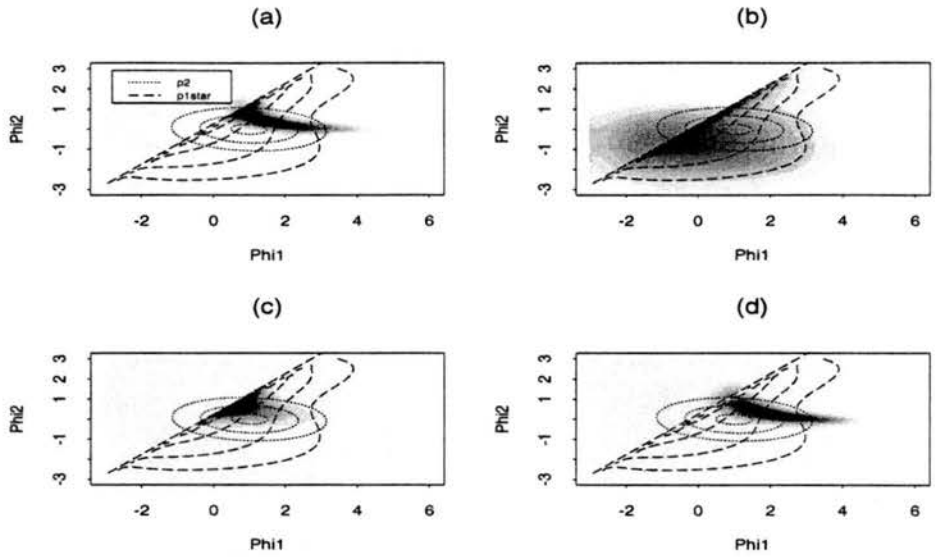

Figure 5.15: Noninvertible $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) using the French-Lindley method. Fixed hyperparameters: $m_1 = -\overline{m}_1 = m_2 = -\overline{m}_2 = 2$, and $\rho = 0.5$. Variable hyperparameters: (a) $s_1 = 1, s_2 = 1$; (b) $s_1 = 3, s_2 = 3$; (c) $s_1 = 1, s_2 = 3$; (d) $s_1 = 3, s_2 = 1$.

applied to continuous parameters does not possess the strong version of the zero preservation property, so that $p(\phi)$ sometimes has significant mass in regions above the truncation line, where Expert 1 assigns no probability.

5.4 Relative Propensity Consistency

One problem with the French-Lindley method is that it violates relative propensity consistency in certain cases. Recall that relative propensity consistency (RPC) (Section 2.2.1), states that if all experts favor an event A_2 over another event A_1, then their pooled opinion must also favor A_2 over A_1. Although RPC is often stated in terms of probabilities, it can be shown to also hold when probabilities are replaced by log-odds. Although masked by our convention of setting $p(\phi)$ to 0 whenever the French-Lindley estimated pooled density was negative, RPC failure did occur in several of our examples. In this section, we investigate when and how these violations occur. This demonstration will focus on the basic model, in which a single decision maker N wishes to update her opinion in light of a second opinion solicited from an expert E. Similar findings

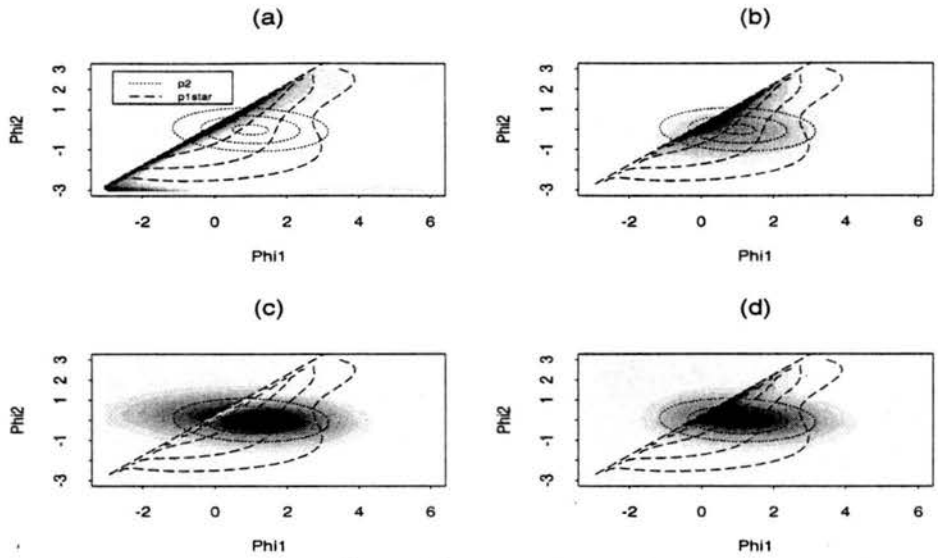

Figure 5.16: Noninvertible $\Re^2 \rightarrow \Re^2$ example: $p(\boldsymbol{\phi})$ (shaded density) using the French-Lindley method. Fixed hyperparameters: $s_1 = s_2 = 2$. Variable hyperparameters: (a) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 1$, and $\rho = 0.7$; (b) $m_1 = -\overline{m}_1 = 2$, $m_2 = -\overline{m}_2 = 1$, and $\rho = 0.2$; (c) $m_1 = -\overline{m}_1 = 1$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.7$; (d) $m_1 = -\overline{m}_1 = 1$, $m_2 = -\overline{m}_2 = 2$, and $\rho = 0.2$.

hold for the GDM model for combining beliefs from a group of individuals (see Section 5.4.4).

Consider that RPC holds under the assumptions for the French-Lindley model with just two experts. If, for instance, the decision maker N assigns higher log-odds to event A_2 than event A_1, and if the expert E also assigns higher log-odds to event A_2, then by relative propensity consistency the consensus posterior distribution (N's beliefs updated in light of E's opinion) should also favor A_2 over A_1. In terms of the notation of this chapter, if the French-Lindley model satisfies RPC, then when $l_0(A_1) < l_0(A_2)$ and $\lambda_1(A_1) < \lambda_1(A_2)$, it must be true that $l_0(A_1 \mid \lambda_1) < l_0(A_2 \mid \lambda_1)$ for all $\underline{m} > \underline{0}$, $\overline{m} < \underline{0}$, and V. The restrictions on $\underline{m}$ and $\overline{m}$ are necessary to ensure that both experts are in positive accord with the truth; if an expert's log-odds for event A become smaller as A becomes more likely, then RPC need not hold. We believe, however, that most practical situations will feature $\underline{m} > \underline{0}$ and $\overline{m} < \underline{0}$. No restrictions are required for V, the

covariance matrix describing the variability and correlation which the decision maker expects in her own probability statements and those of the expert.

We will show that RPC does not always hold under the French-Lindley approach, depending on the beliefs about events A_1 and A_2 of the decision maker and the expert, and on the hyperparameters set by the decision maker. Assume $l_0(A_1) < l_0(A_2)$, $\lambda_1(A_1) < \lambda_1(A_2)$, $\underline{m} > \underline{0}$, and $\overline{\underline{m}} < \underline{0}$. Also, assume:

$$\underline{l} \mid A \sim N(\underline{m}, V) = N_2\left(\begin{bmatrix} m_0 \\ m_1 \end{bmatrix}, \begin{bmatrix} V_{00} & V_{01} \\ V_{10} & V_{11} \end{bmatrix}\right),$$

$$\underline{l} \mid \overline{A} \sim N(\overline{\underline{m}}, V) = N_2\left(\begin{bmatrix} \overline{m}_0 \\ \overline{m}_1 \end{bmatrix}, \begin{bmatrix} V_{00} & V_{01} \\ V_{10} & V_{11} \end{bmatrix}\right),$$

and

$$\rho = V_{10}/\sqrt{V_{00}V_{11}}.$$

From Result 5.3 in the case of two experts, we have

$$\begin{aligned} l_0(A_1 \mid \lambda_1) &= [\mu(l_0(A_1)) - \overline{\mu}(l_0(A_1))]\, W^{-1} \left[\lambda_1(A_1) - \frac{1}{2}(\mu(l_0(A_1)) + \overline{\mu}(l_0(A_1)))\right] \\ &\quad + l_0(A_1) \end{aligned}$$

where:

$$\begin{aligned} \lambda_1 \mid A_1, l_0 &\sim N(\mu(l_0(A_1)), W), \\ \lambda_1 \mid \overline{A_1}, l_0 &\sim N(\overline{\mu}(l_0(A_1)), W), \\ \mu(l_0(A_1)) &= m_1 + \tfrac{V_{10}}{V_{00}}(l_0(A_1) - m_0), \\ \overline{\mu}(l_0(A_1)) &= \overline{m}_1 + \tfrac{V_{10}}{V_{00}}(l_0(A_1) - \overline{m}_0), \text{ and} \\ W &= V_{11} - \tfrac{V_{10}{}^2}{V_{00}} = V_{11}(1 - \rho^2). \end{aligned}$$

Similarly, we have

$$\begin{aligned} l_0(A_2 \mid \lambda_1) &= [\mu(l_0(A_2)) - \overline{\mu}(l_0(A_2))]\, W^{-1} \left[\lambda_1(A_2) - \frac{1}{2}(\mu(l_0(A_2)) + \overline{\mu}(l_0(A_2)))\right] \\ &\quad + l_0(A_2). \end{aligned}$$

Based on the assumptions and derivations above, the difference in the consensus posterior probability distributions can be simplified as follows:

$l_0(A_2 \mid \lambda_1) - l_0(A_1 \mid \lambda_1)$

$$
\begin{aligned}
&= \left[m_1 + \frac{V_{10}}{V_{00}}(l_0(A_2) - m_0) - \overline{m}_1 - \frac{V_{10}}{V_{00}}(l_0(A_2) - \overline{m}_0)\right] \frac{1}{V_{11}(1-\rho^2)} \\
&\quad \left[\lambda_1(A_2) - \frac{1}{2}(m_1 + \frac{V_{10}}{V_{00}}(l_0(A_2) - m_0) + \overline{m}_1 + \frac{V_{10}}{V_{00}}(l_0(A_2) - \overline{m}_0))\right] + l_0(A_2) - \\
&\quad \left[m_1 + \frac{V_{10}}{V_{00}}(l_0(A_1) - m_0) - \overline{m}_1 - \frac{V_{10}}{V_{00}}(l_0(A_1) - \overline{m}_0)\right] \frac{1}{V_{11}(1-\rho^2)} \\
&\quad \left[\lambda_1(A_1) - \frac{1}{2}(m_1 + \frac{V_{10}}{V_{00}}(l_0(A_1) - m_0) + \overline{m}_1 + \frac{V_{10}}{V_{00}}(l_0(A_1) - \overline{m}_0))\right] - l_0(A_1) \\
&= \left[(m_1 - \overline{m}_1) - \rho\sqrt{\frac{V_{11}}{V_{00}}}(m_0 - \overline{m}_0)\right] \frac{1}{V_{11}(1-\rho^2)} \\
&\quad \left[\lambda_1(A_2) - \rho\sqrt{\frac{V_{11}}{V_{00}}}l_0(A_2) - \frac{1}{2}\left[(m_1 + \overline{m}_1) - \rho\sqrt{\frac{V_{11}}{V_{00}}}(m_0 + \overline{m}_0)\right]\right] + l_0(A_2) - \\
&\quad \left[(m_1 - \overline{m}_1) - \rho\sqrt{\frac{V_{11}}{V_{00}}}(m_0 - \overline{m}_0)\right] \frac{1}{V_{11}(1-\rho^2)} \\
&\quad \left[\lambda_1(A_1) - \rho\sqrt{\frac{V_{11}}{V_{00}}}l_0(A_1) - \frac{1}{2}\left[(m_1 + \overline{m}_1) - \rho\sqrt{\frac{V_{11}}{V_{00}}}(m_0 + \overline{m}_0)\right]\right] - l_0(A_1) \\
&= c_1\left[(\lambda_1(A_2) - \lambda_1(A_1)) - \rho\sqrt{\frac{V_{11}}{V_{00}}}\,(l_0(A_2) - l_0(A_1))\right] + (l_0(A_2) - l_0(A_1)) \\
&\quad \text{if } c_1 = \frac{(m_1 - \overline{m}_1) - \rho\sqrt{\frac{V_{11}}{V_{00}}}(m_0 - \overline{m}_0)}{V_{11}(1-\rho^2)} \\
&= c_1\left[\lambda_1(A_2) - \lambda_1(A_1)\right] + \left(1 - c_1\rho\sqrt{\frac{V_{11}}{V_{00}}}\right)\left[l_0(A_2) - l_0(A_1)\right].
\end{aligned}
$$

Under our assumptions, $\lambda_1(A_2) - \lambda_1(A_1) > 0$, $l_0(A_2) - l_0(A_1) > 0$, and, if RPC holds, $l_0(A_2 \mid \lambda_1) - l_0(A_1 \mid \lambda_1) > 0$. This implies

$$
c_1\left[\lambda_1(A_2) - \lambda_1(A_1)\right] > \left(c_1\rho\sqrt{\frac{V_{11}}{V_{00}}} - 1\right)\left[l_0(A_2) - l_0(A_1)\right],
$$

which leads to the result

$$
\begin{cases}
\frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)} > \rho\sqrt{\frac{V_{11}}{V_{00}}} - \frac{1}{c_1} & \text{if } c_1 > 0 \\
\frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)} < \rho\sqrt{\frac{V_{11}}{V_{00}}} - \frac{1}{c_1} & \text{if } c_1 < 0.
\end{cases}
$$

Hereafter, let $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)}$ and $K = \rho\sqrt{\frac{V_{11}}{V_{00}}} - \frac{1}{c_1}$. Relative propensity consistency, therefore, does not hold in the basic French-Lindley model with one decision maker and one expert if $R < K$ when $c_1 > 0$ or if $R > K$ when $c_1 < 0$. This condition depends on the beliefs l_0 of the decision maker, the beliefs λ_1 of the

expert, and the hyperparameters describing the abilities of the decision maker and the expert.

5.4.1 Constraint Set 1

By setting constraints on the hyperparameters, we can illustrate for which values of $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)}$ RPC failure will occur. We chose two simple but sensible sets of assumptions under which to investigate RPC failure. Under constraint set 1, four constraints are imposed on the hyperparameters used to assess the abilities of the two experts:

1. $m_i = -\overline{m}_i$ for $i = 0, 1$. For each expert, the mean probability for success predicted for an event A which subsequently occurs equals the mean probability for failure predicted for an event A which subsequently does not occur. In other words, N is equally adept at forecasting successes and failures, as is E.

2. $V_{00} = V_{11}$. The variability in the experts' log-odds are identical.

3. $m_0 = am_1$. The parameter a describes the relative expertise between N and E. For instance, $a > 1$ suggests that N offers higher log-odds on average than E for events A which subsequently occur.

4. $V_{11} = bm_1$. The parameter b allows the variability of E's opinions to be expressed as a function of his mean log-odds.

With these constraints, the expression for RPC failure becomes

$$\begin{cases} \frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)} < \rho - \frac{b(1-\rho^2)}{2(1-a\rho)} & \text{if } \rho < \frac{1}{a} \\ \frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)} > \rho - \frac{b(1-\rho^2)}{2(1-a\rho)} & \text{if } \rho > \frac{1}{a} \end{cases}$$

after solving for K and simplifying the expressions for $c_1 < 0$ and $c_1 > 0$.

Using this expression, we illustrate scenarios for RPC failure in Figures 5.17, 5.18, and 5.19. Each plot represents a different fixed value for the parameter b which

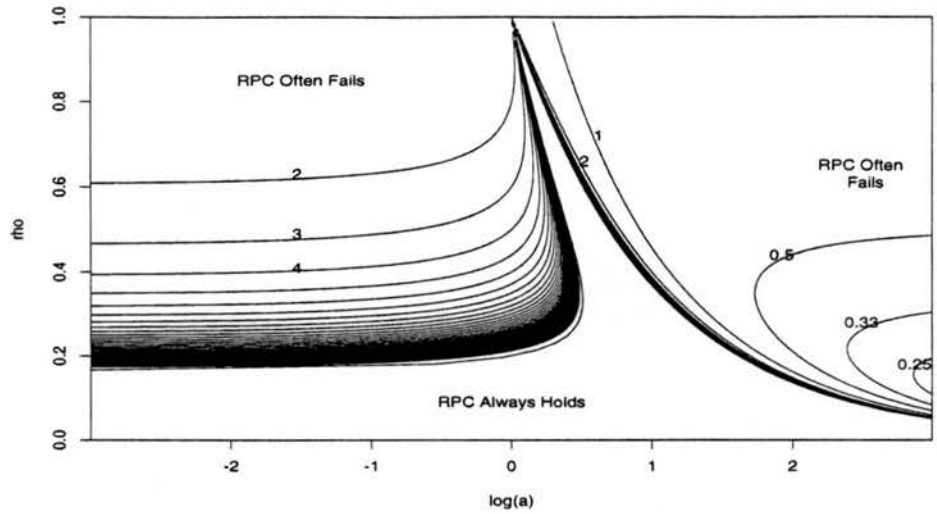

Figure 5.17: RPC assessment: constraint set 1, $b = \frac{1}{3}$. Contour levels correspond to $\frac{1}{K}$ where $\rho < \frac{1}{a}$ and K where $\rho > \frac{1}{a}$ so that smaller contour levels are linked with higher chances for RPC failure. RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)}$.

describes the relationship between N's assessment of the mean beliefs for E (expressed as log-odds) and the variance of those beliefs. In Figure 5.17 $b = \frac{1}{3}$, in Figure 5.18 $b = 1$, and in Figure 5.19 $b = 3$. In these figures, a is allowed to vary between .05 and 20 and is presented on the x-axis as $\log(a)$. ρ, the correlation between the decision maker and the expert, is allowed to vary between 0 and 1 and is presented on the y-axis. The contours in Figures 5.17- 5.19, then, illustrate combinations of a and ρ (given a value for b) for which RPC likely holds or for which RPC likely fails. RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)}$. Contours have been sketched at levels corresponding to $\frac{1}{K}$ in case 1 (where $\rho < \frac{1}{a}$) and K in case 2 (where $\rho > \frac{1}{a}$) so that smaller contour levels are linked with higher chances for RPC failure. Note that case 1 corresponds to the left region of each figure and case 2 corresponds to the right region. Although we denote in these figures where RPC failure in most likely, it is possible for RPC to fail under any combinations of a and ρ which do not fall in the region marked "RPC Always Holds".

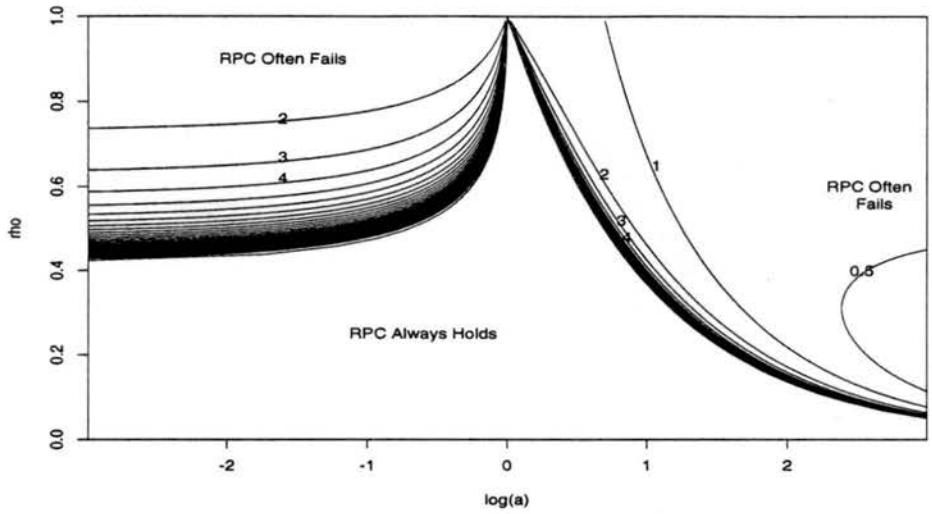

Figure 5.18: RPC assessment: constraint set 1, $b = 1$. Contour levels correspond to $\frac{1}{K}$ where $\rho < \frac{1}{a}$ and K where $\rho > \frac{1}{a}$ so that smaller contour levels are linked with higher chances for RPC failure. RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2) - \lambda_1(A_1)}{l_0(A_2) - l_0(A_1)}$.

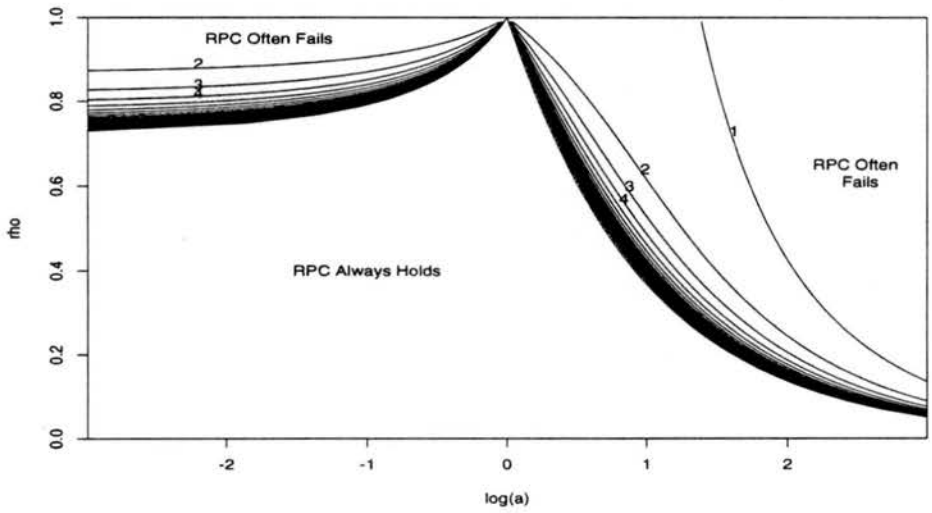

Figure 5.19: RPC assessment: constraint set 1, $b = 3$. Contour levels correspond to $\frac{1}{K}$ where $\rho < \frac{1}{a}$ and K where $\rho > \frac{1}{a}$ so that smaller contour levels are linked with higher chances for RPC failure. RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2) - \lambda_1(A_1)}{l_0(A_2) - l_0(A_1)}$.

In Figure 5.17 $b = \frac{1}{3}$, so that V_{11}, the decision maker's expected variance in the expert's statements, is just one-third the value of N's expected log-odds for E; thus, N expects E to be very consistent in stating $\lambda_1(A)$ when A is subsequently true. In this figure, RPC always holds in the lower region defined by small correlations between N and E (e.g. $\rho < 0.2$). For combinations of a and ρ in this lower region, relative propensity consistency is never violated, regardless of the value of R. However, Figure 5.17 is dominated by regions in which RPC will often fail depending, of course, of the particular value of R. Failure of relative propensity consistency is likely to occur when some dependence between N and E exists (e.g. $\rho > 0.5$) regardless of a. RPC failure is also likely whenever a is high, regardless of ρ; high values of a correspond to situations in which N tends to predict much higher log-odds than E for events A which are subsequently found to be true.

In Figure 5.18 $b = 1$, so that $V_{11} = m_1$. Under this setting, there is again one large region in which RPC always holds, regardless of the value of R. As in Figure 5.17, this RPC region is associated with small values of ρ, although in this case it extends to $\rho < 0.4$ for $a < 1$. There are also large regions in which RPC will often fail. Failure of relative propensity consistency is most likely to occur when a is small (E tends to offer higher log-odds than N) and ρ is large, or when a is large (N tends to offer higher log-odds than E) and the correlation is non-zero. As before, failure of RPC is dependent on R.

In Figure 5.19 $b = 3$, so that $V_{11} = 3m_1$—i.e. N expects that E's log-odds for a true event A will be wildly variant. In this case, the region in which RPC always holds is larger than in the two previous plots; therefore, failure of RPC is less likely with higher b. Still, failure of RPC can occur when N and E are highly correlated, or when N perceives that she has greater expertise than E (so that a is high).

5.4.2 Constraint Set 2

A second set of constraints is used to illustrate other conditions under which RPC failure can occur. As with constraint set 1, four constraints are imposed on the hyperparameters for expert assessment in constraint set 2:

1. $m_i = -\overline{m}_i$ for $i = 0, 1$ For each expert, the mean probability for success predicted for an event A which subsequently occurs equals the mean probability for failure predicted for an event A which subsequently does not occur. In other words, N is equally adept at forecasting successes and failures, as is E.

2. $m_0 = m_1$. N and E offer the same log-odds on average for events A which subsequently occur.

3. $\sqrt{V_{11}} = a\sqrt{V_{00}}$. The parameter a describes the relative precision between N and E. For instance, $a > 1$ suggests that N's opinions have less variability than E's.

4. $V_{11} = bm_1$. The parameter b allows the variability of E's opinions to be expressed as a function of his mean log-odds.

With these constraints, the expression for RPC failure becomes

$$\begin{cases} \frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)} < a\rho - \frac{b(1-\rho^2)}{2(1-a\rho)} & \text{if } \rho < \frac{1}{a} \\ \frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)} > a\rho - \frac{b(1-\rho^2)}{2(1-a\rho)} & \text{if } \rho > \frac{1}{a} \end{cases}$$

after solving for K and simplifying the expressions for $c_1 < 0$ and $c_1 > 0$.

RPC failure under these assumptions is illustrated in Figures 5.20, 5.21, and 5.22. As before, each plot represents a different fixed value for the parameter b, and contours have been sketched at levels corresponding to $\frac{1}{K}$ in case 1 (where $\rho < \frac{1}{a}$) and K in case 2 (where $\rho > \frac{1}{a}$) so that smaller contour levels are linked

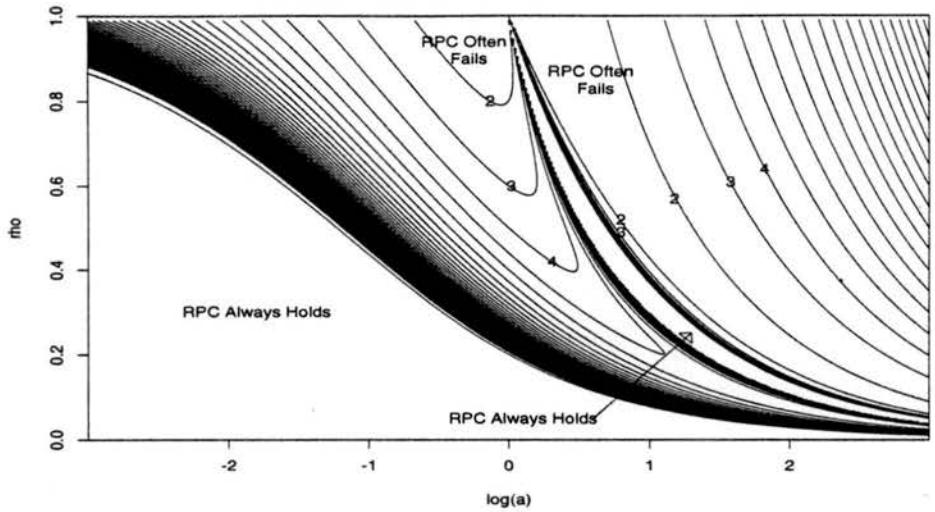

Figure 5.20: RPC assessment: constraint set 2, $b = \frac{1}{3}$. Contour levels correspond to $\frac{1}{K}$ where $\rho < \frac{1}{a}$ and K where $\rho > \frac{1}{a}$ so that smaller contour levels are linked with higher chances for RPC failure. RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2) - \lambda_1(A_1)}{l_0(A_2) - l_0(A_1)}$.

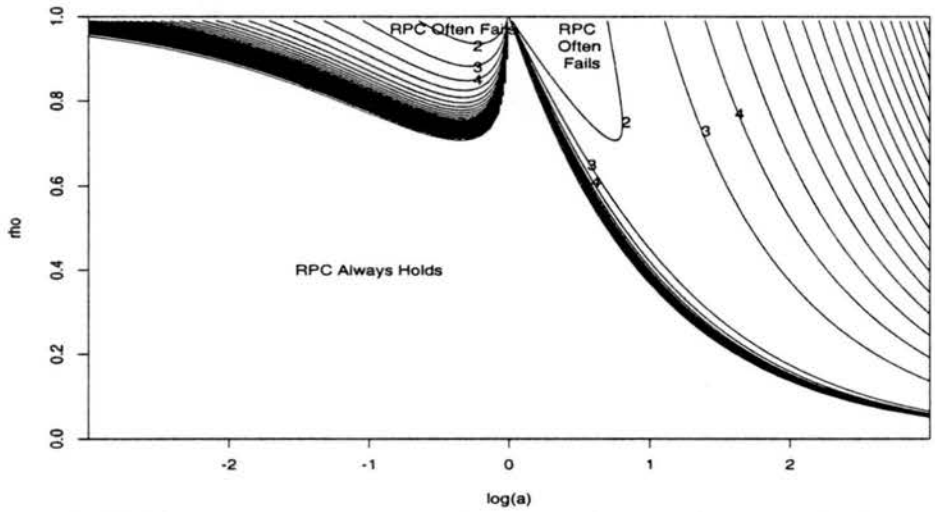

Figure 5.21: RPC assessment: constraint set 2, $b = 1$. Contour levels correspond to $\frac{1}{K}$ where $\rho < \frac{1}{a}$ and K where $\rho > \frac{1}{a}$ so that smaller contour levels are linked with higher chances for RPC failure. RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2) - \lambda_1(A_1)}{l_0(A_2) - l_0(A_1)}$.

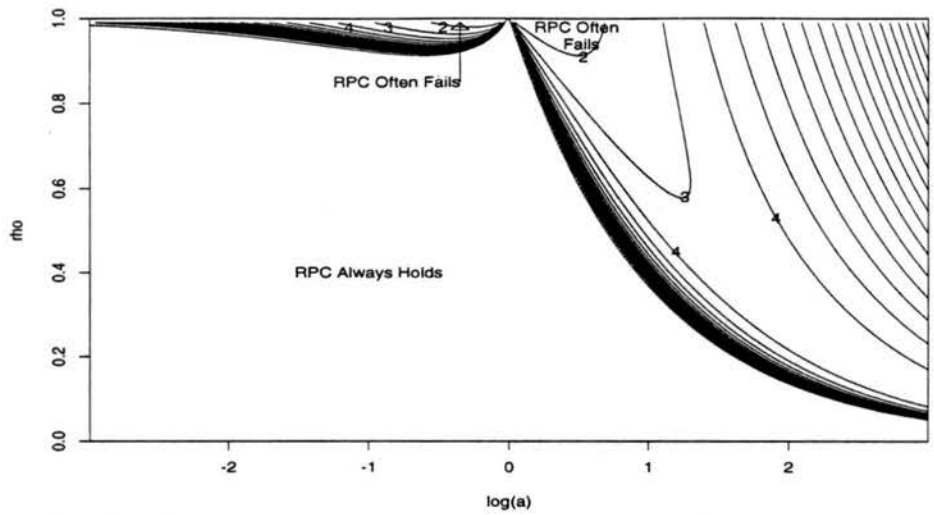

Figure 5.22: RPC assessment: constraint set 2, $b = 3$. Contour levels correspond to $\frac{1}{K}$ where $\rho < \frac{1}{a}$ and K where $\rho > \frac{1}{a}$ so that smaller contour levels are linked with higher chances for RPC failure. RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)}$.

with higher chances for RPC failure. Note that case 1 corresponds to the left region of each figure and case 2 corresponds to the right region.

In Figure 5.20 $b = \frac{1}{3}$, so that N expects E to be very consistent in stating $\lambda_1(A)$ when A is subsequently true. In this figure, there are two regions in which RPC always holds, associated primarily with smaller values of both a and ρ. For combinations of a and ρ falling in these two regions, relative propensity consistency is never violated, regardless of the value of R. However, there are also two large regions in which RPC will often fail, depending on the particular value of R. Failure of RPC is most likely to occur when a is near 1 ($\log(a)$ is near 0), so that N and E have similar levels of expertise, and ρ is fairly high. In addition, RPC failure is common even for low ρ when a is above 1—i.e. when N tends to predict much higher log-odds than E for events A which are subsequently found to be true.

In Figure 5.21 $b = 1$, so that $V_{11} = m_1$. Under this setting, there is one large region in which RPC always holds, regardless of the value of R. As in Figure 5.20,

this RPC region is associated with smaller values of ρ and a, although in this case it extends to $\rho < 0.8$ for $\log(a) < 1$. There are also two regions in which RPC will often fail. Failure of relative propensity consistency is most likely to occur when a is near 1, so that m_0 and m_1 are in reasonable agreement, and ρ is fairly high.

In Figure 5.22 $b = 3$, so that $V_{11} = 3m_1$—i.e. N expects that E's log-odds for a true event A will be wildly variant. In that case, the region in which RPC always holds is larger than in the two previous plots; therefore, failure of RPC is less likely with higher b. As with $b = 1$, RPC failure is most likely when a is near 1 and ρ is high.

5.4.3 Special Case: Zero Correlation

In the special case in which $\rho = 0$, it can be shown that RPC failure is impossible when $m_1 > 0$ and $\overline{m}_1 < 0$. This follows because

$$c_1 = \frac{m_1 - \overline{m}_1}{V_{11}} > 0 \text{ and}$$

$$K = \rho\sqrt{\frac{V_{11}}{V_{00}}} - \frac{1}{c_1} = -\frac{1}{c_1} < 0.$$

Thus, we must have $\frac{\lambda_1(A_2)-\lambda_1(A_1)}{l_0(A_2)-l_0(A_1)} > \rho\sqrt{\frac{V_{11}}{V_{00}}} - \frac{1}{c_1}$, which, when $c_1 > 0$, ensures that RPC holds regardless of the other hyperparameter values. As a result, any violations of RPC which were observed in this chapter occurred where $\rho > 0$.

5.4.4 RPC under the Group Decision Maker Approach

Conditions for the failure of relative propensity consistency in the group decision maker version of the French-Lindley method can be investigated using similar techniques to those we employed with the basic French-Lindley method. Assume that a group wishes to arrive at a consensus opinion about two events A_1 and A_2. A group decision maker who begins with no opinion about A_1 and A_2 solicits opinions (denoted $\lambda_1(A_i)$ and $\lambda_2(A_i)$ for $i = 1, 2$) from two experts.

The GDM can obtain group consensus opinions for these events by deriving its posterior log-odds for A_1 and A_2 in light of the expert opinions according to $l_0(A \mid \underline{\lambda}) = (\underline{m} - \overline{\underline{m}})V^{-1}\left[\underline{\lambda} - \frac{1}{2}(\underline{m} + \overline{\underline{m}})\right]$ as discussed in Section 5.2.2. In this case, the GDM version of the French-Lindley method is said to satisfy relative propensity consistency, if $l_0(A_1 \mid \underline{\lambda}) < l_0(A_2 \mid \underline{\lambda})$ whenever $\lambda_1(A_1) < \lambda_1(A_2)$ and $\lambda_2(A_1) < \lambda_2(A_2)$ for all $\underline{m} > \underline{0}$, $\overline{\underline{m}} < \underline{0}$, and V.

To derive theoretical properties under which the relative propensity consistency holds, first assume $\lambda_1(A_1) < \lambda_1(A_2)$, $\lambda_2(A_1) < \lambda_2(A_2)$, $m_1 > 0$, $m_2 > 0$, $\overline{m}_1 < 0$, and $\overline{m}_2 < 0$. In addition, assume

$$\underline{\lambda} \mid A \sim N(\underline{m}, V) = N_2\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{bmatrix}\right) \text{ and}$$

$$\underline{\lambda} \mid \overline{A} \sim N(\overline{\underline{m}}, V) = N_2\left(\begin{bmatrix} \overline{m}_1 \\ \overline{m}_2 \end{bmatrix}, \begin{bmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{bmatrix}\right).$$

Recall that in the GDM case, $\underline{\lambda} \mid A, l_0 = \underline{\lambda} \mid A$ and $\underline{\lambda} \mid \overline{A}, l_0 = \underline{\lambda} \mid \overline{A}$, so that Result 5.3 simplifies to:

$$\begin{aligned} l_0(A \mid \underline{\lambda}) &= \begin{bmatrix} m_1 - \overline{m}_1 \\ m_2 - \overline{m}_2 \end{bmatrix} V^{-1} \begin{bmatrix} \lambda_1(A) - \frac{1}{2}(m_1 + \overline{m}_1) \\ \lambda_2(A) - \frac{1}{2}(m_2 + \overline{m}_2) \end{bmatrix} \\ &= \frac{1}{s_1^2 s_2^2(1-\rho^2)}\Big[s_2^2(m_1 - \overline{m}_1)\lambda_1(A) - \frac{1}{2}s_2^2(m_1^2 - \overline{m}_1^2) \\ &\quad -\rho s_1 s_2(m_2 - \overline{m}_2)\lambda_1(A) + \frac{1}{2}\rho s_1 s_2(m_1 + \overline{m}_1)(m_2 - \overline{m}_2) \\ &\quad +s_1^2(m_2 - \overline{m}_2)\lambda_2(A) - \frac{1}{2}s_1^2(m_2^2 - \overline{m}_2^2) \\ &\quad -\rho s_1 s_2(m_1 - \overline{m}_1)\lambda_2(A) + \frac{1}{2}\rho s_1 s_2(m_1 - \overline{m}_1)(m_2 + \overline{m}_2)\Big]. \end{aligned}$$

The difference in the GDM's posterior probability distributions can be simplified as follows:

$$\begin{aligned} l_0(A_2 \mid \underline{\lambda}) - l_0(A_1 \mid \underline{\lambda}) &= \frac{1}{s_1^2 s_2^2(1-\rho^2)}\Big[(\lambda_1(A_2) - \lambda_1(A_1))\left(s_2^2(m_1 - \overline{m}_1)\right. \\ &\quad \left. -\rho s_1 s_2(m_2 - \overline{m}_2)\right) + (\lambda_2(A_2) - \lambda_2(A_1)) \\ &\quad \cdot \left(s_1^2(m_2 - \overline{m}_2) - \rho s_1 s_2(m_1 - \overline{m}_1)\right)\Big]. \end{aligned}$$

If RPC holds, $l_0(A_2 \mid \underline{\lambda}) - l_0(A_1 \mid \underline{\lambda}) > 0$, which implies:

$$\begin{cases} \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)} > \frac{\rho s_1 s_2(m_1-\overline{m}_1)-s_1^2(m_2-\overline{m}_2)}{s_2^2(m_1-\overline{m}_1)-\rho s_1 s_2(m_2-\overline{m}_2)} & \text{if } s_2^2(m_1-\overline{m}_1) > \rho s_1 s_2(m_2-\overline{m}_2) \\ \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)} < \frac{\rho s_1 s_2(m_1-\overline{m}_1)-s_1^2(m_2-\overline{m}_2)}{s_2^2(m_1-\overline{m}_1)-\rho s_1 s_2(m_2-\overline{m}_2)} & \text{if } s_2^2(m_1-\overline{m}_1) < \rho s_1 s_2(m_2-\overline{m}_2). \end{cases}$$

Let $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)}$ as before and $K = \frac{\rho s_1 s_2(m_1-\overline{m}_1)-s_1^2(m_2-\overline{m}_2)}{s_2^2(m_1-\overline{m}_1)-\rho s_1 s_2(m_2-\overline{m}_2)}$. Then relative propensity consistency does not hold if $R < K$ when $s_2^2(m_1-\overline{m}_1) > \rho s_1 s_2(m_2-\overline{m}_2)$ or if $R > K$ when $s_2^2(m_1-\overline{m}_1) < \rho s_1 s_2(m_2-\overline{m}_2)$.

Using constraints analogous to those used in Sections 5.4.1 and 5.4.2, we can illustrate the values of $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)}$ for which RPC fails to hold. Constraint set 1G imposes three constraints on the French-Lindley hyperparameters:

1. $m_i = -\overline{m}_i$ for $i = 1, 2$.

2. $s_1^2 = s_2^2$.

3. $m_2 = am_1$.

The fourth assumption which introduced the parameter b in Section 5.4.1 is unnecessary here. We need only consider parameter a, which describes the relative expertise between the two experts. With constraint set 1G, the expression for RPC failure becomes:

$$\begin{cases} \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)} < \frac{\rho-a}{1-a\rho} & \text{if } \rho < \frac{1}{a} \\ \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)} > \frac{\rho-a}{1-a\rho} & \text{if } \rho > \frac{1}{a}. \end{cases}$$

Using this expression, RPC failure is illustrated in Figure 5.23. As in Figures 5.17- 5.22, the contours illustrate combinations of a and ρ for which RPC likely holds or for which RPC likely fails. Remember that RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)}$. Contours have been sketched at levels corresponding to $\frac{1}{K}$ in case 1 (where $\rho < \frac{1}{a}$) and K in case 2 (where $\rho > \frac{1}{a}$) so that smaller contour levels are linked with higher chances for RPC failure. Note that case 1 corresponds to the left region of each figure and case 2 corresponds to the right region.

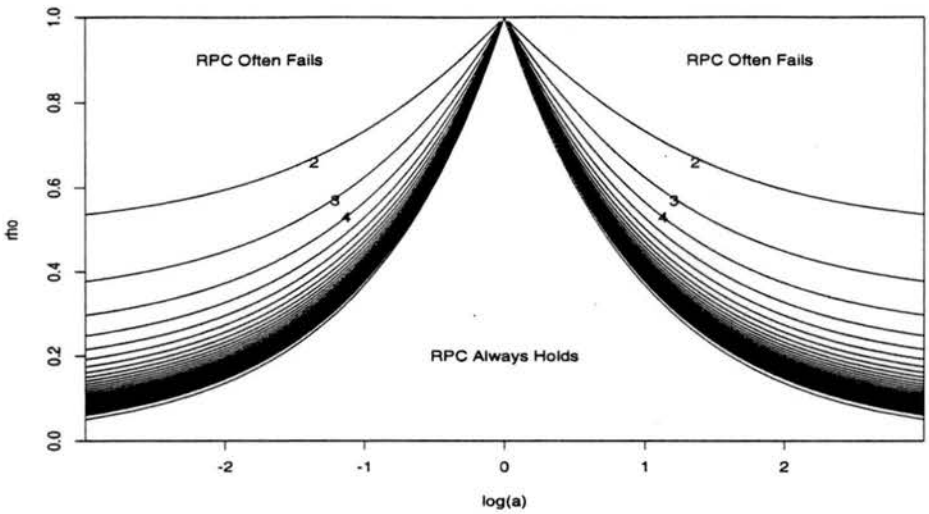

Figure 5.23: RPC assessment under group decision maker approach: constraint set 1G. Contour levels correspond to $\frac{1}{K}$ where $\rho < \frac{1}{a}$ and K where $\rho > \frac{1}{a}$ so that smaller contour levels are linked with higher chances for RPC failure. RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)}$.

In Figure 5.23, relative propensity consistency always holds for the lower region defined by small correlations and/or similar levels of expertise between the two experts. On the other hand, failure of relative propensity consistency is likely to occur when some dependence between the two experts exists (e.g. $\rho \geq 0.6$) and one expert is considered to have greater expertise by the group decision maker. Expertise in this case is measured by comparing the log-odds which the GDM expects each expert to state for an event A which subsequently occurs. Although the scenario above describes when RPC is most likely to fail, it is possible for RPC to fail for any combination of ρ and a which does not fall in the triangular region in the center of Figure 5.23.

A second set of constraints analogous to those in Section 5.4.2 further illustrate conditions under which failure of relative propensity consistency can occur. Constraint set 2G imposes these three constraints on the hyperparameters for expert assessment:

1. $m_i = -\overline{m}_i$ for $i = 1, 2$.

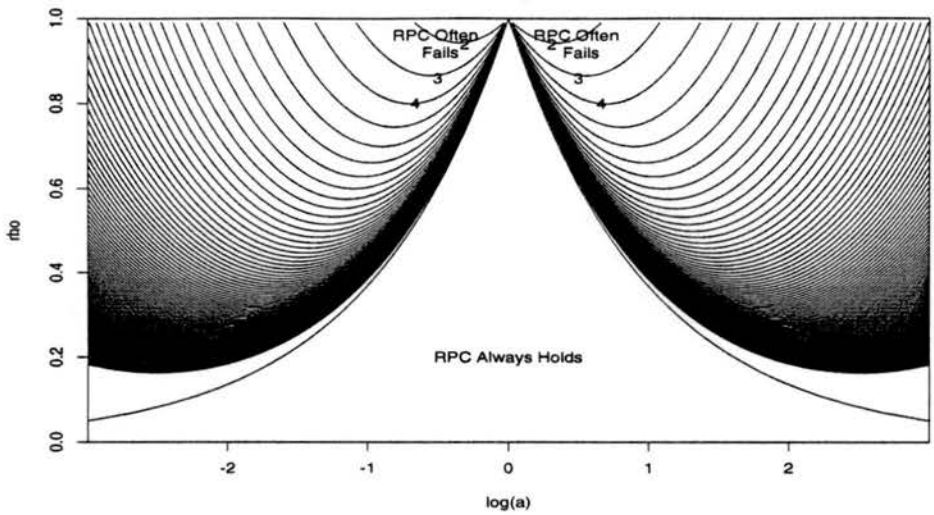

Figure 5.24: RPC assessment under group decision maker approach: constraint set 2G. Contour levels correspond to $\frac{1}{K}$ where $\rho < a$ and K where $\rho > a$ so that smaller contour levels are linked with higher chances for RPC failure. RPC failure ultimately depends on the value of $R = \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)}$.

2. $m_1 = m_2$.

3. $s_1 = as_2$.

Again parameter b is unnecessary, and here a describes the relative precision between the two experts. With constraint set 2G, the expression for RPC failure becomes:

$$\begin{cases} \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)} < \frac{a\rho-1}{a(a-\rho)} & \text{if } \rho < a \\ \frac{\lambda_1(A_2)-\lambda_1(A_1)}{\lambda_2(A_2)-\lambda_2(A_1)} > \frac{a\rho-1}{a(a-\rho)} & \text{if } \rho > a. \end{cases}$$

RPC failure under these assumptions is illustrated in Figure 5.24. Again contours have been sketched at levels corresponding to $\frac{1}{K}$ in case 1 (where $\rho < a$) and K in case 2 (where $\rho > a$) so that smaller contour levels are linked with higher chances for RPC failure. Note that, unlike previous plots, case 1 corresponds to the right region of each figure and case 2 corresponds to the left region.

In Figure 5.24, relative propensity consistency always holds for the lower region defined by small correlations and/or similar levels of precision between the two experts. On the other hand, failure of relative propensity consistency is likely

to occur when some dependence between the two experts exists (e.g. $\rho \geq 0.8$) and one expert is considered to have better precision by the group decision maker. Precision in this case is measured by comparing the variability in the log-odds which the GDM expects from each expert. Although the scenario above describes when RPC is most likely to fail, it is possible for RPC to fail for any combination of ρ and a which does not fall in the triangular region in the center of Figure 5.24.

5.4.5 Conclusions

The constraint sets and plots of this section were developed for illustrative purposes, to show that under the French-Lindley model failure of relative propensity consistency is not only possible, but it can frequently occur under seemingly reasonable scenarios. In the examples discussed in Section 5.3, failure of RPC led to illogical, unsatisfactory results in several instances. The frequency of RPC failure under the group decision maker setting is especially vexing, since this setting mostly closely approximates the situation in which priors to be pooled are linked by a deterministic simulation model. In those cases, often a person or group will collect opinions from experts on each of the model parameters and then form a consensus from those opinions about an event or parameter of interest. But, like French's GDM, the opinion collector in a deterministic simulation model setting usually has no expert opinion to offer about any of the model parameters. Therefore, for inference problems involving deterministic simulation models, as well as other problems in which consensus distributions must be formed, it is difficult to place confidence in a method such as French-Lindley which leads to violations of a basic axiom such as RPC in relatively common situations.

5.5 Summary

Our examples and consideration of RPC suggest that the French-Lindley supra-Bayesian method is unsuitable for pooling expert opinions to make inference

in deterministic simulation models. The examples in Section 5.3 were designed to allow examination of potential pooling methods in small-scale versions of deterministic simulation models. Typically, a single decision maker (perhaps the statistician overseeing the process) will obtain priors from one or more experts in input-space and one or more experts in output-space. In order to make inference about outputs, for instance, the decision maker should pool the prior distribution from the output experts with the implicit prior distribution (through the model) about the outputs from the input experts.

Although developed with single, discrete events in mind, the French-Lindley approach is one viable option for performing this pooling, especially for those who adopt the supra-Bayesian perspective and prefer to consider priors solicited from experts as data to be used by the ultimate decision maker in forming inference. By considering a series of cumulative distribution functions, we adapted French and Lindley's ideas to the case with continuous parameters, which is a more common scenario in deterministic simulation models. However, in applying these methods to the examples of Section 5.3, several serious problems emerged.

First, the supra-Bayesian decision maker or GDM must estimate the values of a series of hyperparameters to quantify each expert's bias, inaccurate precision evaluations, and correlations with other experts. Moreover, these hyperparameters are based on the log-odds scale. Unfortunately, the examples in this chapter show that the pooled distributions formed by French-Lindley methods lack robustness to changes in hyperparameter values. Thus, small errors made by the supra-Bayesian, either through imperfect knowledge of the experts or imperfect understanding of the interpretation of the hyperparameters, will greatly impact the consensus distribution. In fact, even when the supra-Bayesian has not erred, seemingly reasonable hyperparameter values can produce nonsensical poolings; we discovered this while attempting to define a baseline set of values for plots in

Section 5.3. For instance, small s (the assessed variability in an expert's belief statements), small ρ (the assessed correlation between experts), and large m (the assessed average log-odds an expert would state about an event A which subsequently occurs) often led to $p(\phi)$ which were much more precise the corresponding individual expert priors. Other factors frequently led to $p(\phi)$ with mass concentrated in regions of ϕ-space which would not intuitively be considered as regions of agreement between the experts. These factors included $s_i < s_j$, $m_i < m_j$ with high correlation, truncation in ϕ-space, and higher dimensions in general.

Second, the French-Lindley approach does not guarantee adherence to relative propensity consistency. That is, even if all experts believe that event A_1 is more likely than event A_2, the French-Lindley consensus may state that A_2 is more likely than A_1. RPC violations are especially noticeable when applying the French-Lindley approach to continuous variables, since they can result in $p(\phi) < 0$ for certain values of ϕ. Greater detail can be found in Section 5.4.

Third, the dependence on log-odds makes it difficult to adapt the French-Lindley approach to pooling priors linked by deterministic simulation models, which primarily contain continuous parameters. In addition to the large barriers it creates for a decision maker attempting to set hyperparameter values, the use of log-odds makes extension into higher dimensions more difficult. As discussed in Section 5.3.2, conventional methods for calculating cumulative distributions in two dimensions often produce estimated French-Lindley pooled prior distributions which violate laws of probability.

Fourth, the assumption that N is a coherent individual, while the expert E is potentially incoherent, is a tenuous one at best. LTB argue that coherence at some point in the process is necessary to ensure the coherence of reconciled values, and that one does not reach too far to assume the existence of a decision maker N who can detect and reconcile her own inconsistencies with respect to probability

statements. Even though coherent experts are envisioned for many applications of deterministic simulation models, the elicitation of prior opinions is an inexact science at best, and a certain degree of incoherence is perhaps unavoidable.

The overwhelming conclusion from our analysis is that the French-Lindley approach is unsuitable for pooling prior distributions in deterministic simulation models. A person preferring to maintain the supra-Bayesian perspective would be best suited by exploring other alternatives.

Chapter 6

The Lindley-Winkler Supra-Bayesian Method

6.1 The Lindley Method

6.1.1 Univariate Case

Lindley (1983) motivates his method with the following example. Assume that a submarine commander is interested in the range θ of a possible target. The commander obtains estimates of θ from several sources (e.g. sonar, radar), each of which also provides a measure of precision. How should the commander combine this information to form her own opinion about θ? In general, Lindley asks how should one form a single consensus probability distribution using several expert opinions about θ, where each expert offers a mean and standard deviation, although not necessarily a particular distributional form. The absence of a requirement that experts specify a distributional form for their prior beliefs about θ is a strength of the Lindley approach.

In notational form, Lindley answers the question: How should N update her opinion of θ in light of the information provided by $E_i, i = 1, \ldots, k$? In this case, N is the decision-maker interested in a quantity θ, and $E_i, i = 1, \ldots, k$, are the k experts consulted by N. Lindley applies supra-Bayesian reasoning, forming a likelihood function by treating each E_i's opinion about θ as data, and then combining this likelihood with N's prior on θ to form a posterior probability distribution for

θ. This can be written in the form "posterior $\propto$ (likelihood)(prior)" as

$$p(\theta \mid \underline{t}_1, \ldots, \underline{t}_k, H) \propto p(\underline{t}_1, \ldots, \underline{t}_k \mid \theta, H) p(\theta \mid H),$$

where H is the background information (other than E_i's information) possessed by N, $p(\theta \mid H)$ is N's prior opinion regarding θ, and $\underline{t}_i = (t_{i,1}, \ldots, t_{i,n})$ is a vector of numbers provided by E_i to describe his probability distribution. The likelihood $p(\underline{t}_1, \ldots, \underline{t}_k \mid \theta, H)$ includes not only each expert's prior probability for θ, but it also incorporates N's opinion of the accuracy and precision of each expert's statement. This will be accomplished through the introduction of several hyperparameters, which will be described shortly.

Start with the case of k experts and a single quantity of interest θ. Assume $t_{i,1} = m_i$ and $t_{i,2} = s_i$, where these quantities will typically represent the mean and standard deviation, respectively, for E_i's prior for θ. Then,

$$\begin{aligned} p(\theta \mid m_1, \ldots, m_k, s_1, \ldots, s_k) &\propto p(m_1, \ldots, m_k, s_1, \ldots, s_k \mid \theta) p(\theta) \\ &= p(m_1, \ldots, m_k \mid s_1, \ldots, s_k, \theta) p(s_1, \ldots, s_k \mid \theta) p(\theta) \end{aligned}$$

where N's background information H has been suppressed for now. If we let $\underline{m} = (m_1, \ldots, m_k)'$ and $\underline{s} = (s_1, \ldots, s_k)'$, then we have $p(\theta \mid \underline{m}, \underline{s}) \propto p(\underline{m} \mid \underline{s}, \theta) p(\underline{s} \mid \theta) p(\theta)$. Assume:

Assumption 1. $p(\underline{s} \mid \theta)$ does not depend on θ so that $p(\underline{s} \mid \theta) = p(\underline{s})$. Thus, $s_1, \ldots, s_k$ by themselves provide no information about θ.

Assumption 2.

$$\begin{aligned} p(\underline{m} \mid \underline{s}, \theta) &\sim N_k \left(\begin{bmatrix} \alpha_1 + \beta_1 \theta \\ \vdots \\ \alpha_k + \beta_k \theta \end{bmatrix}, \begin{bmatrix} \gamma_1^2 s_1^2 & \rho_{12}\gamma_1 s_1 \gamma_2 s_2 & \cdots & \rho_{1k}\gamma_1 s_1 \gamma_k s_k \\ \rho_{12}\gamma_1 s_1 \gamma_2 s_2 & \gamma_2^2 s_2^2 & \cdots & \rho_{2k}\gamma_2 s_2 \gamma_k s_k \\ \vdots & \vdots & & \vdots \\ \rho_{1k}\gamma_1 s_1 \gamma_k s_k & \rho_{2k}\gamma_2 s_2 \gamma_k s_k & \cdots & \gamma_k^2 s_k^2 \end{bmatrix} \right) \\ &= N_k(\underline{\mu}, \Sigma). \end{aligned}$$

In other words, $p(\underline{m} \mid \underline{s}, \theta)$ is multivariate normal with means $\alpha_i + \beta_i \theta$, standard deviations $\gamma_i s_i$, and correlations ρ_{ij}. Let $\underline{\alpha}$ be a $k \times 1$ vector of additive bias terms,

$\underline{\beta}$ be a $k \times 1$ vector of multiplicative bias terms, $\underline{\gamma}$ be a $k \times 1$ vector of precision adjustment terms, and R be a $k \times k$ matrix of correlations between experts. These hyperparameters are set by N and will be explained in greater detail later.

Assumption 3. $p(\theta)$ is an improper, flat (noninformative) prior. This is similar to Lindley's third assumption (1983), which also implies that N's knowledge of θ is weak compared with the information provided by the experts.

Under these three assumptions the following result holds.

Theorem 6.1 *If E_i states that θ has mean m_i and standard deviation s_i, $(i = 1, \ldots, k)$, and Assumptions 1-3 apply, then the Lindley-Winkler supra-Bayesian pooled prior distribution for θ is normal with mean $\sum_{i,j} \beta_i \sigma^{ij} (m_j - \alpha_j) / \sum_{i,j} \beta_i \sigma^{ij} \beta_j$ and variance $(\sum_{i,j} \beta_i \sigma^{ij} \beta_j)^{-1}$, where σ_{ij} is the $(i,j)^{\text{th}}$ element of Σ, $(i = 1, \ldots, k; j = 1, \ldots, k)$, and σ^{ij} is the $(i,j)^{\text{th}}$ element of Σ^{-1}.*

Proof.

$$
\begin{aligned}
p(\theta \mid \underline{m}, \underline{s}) &\propto p(\underline{m}, \underline{s} \mid \theta) p(\theta) \\
&= p(\underline{m} \mid \underline{s}, \theta) p(\underline{s} \mid \theta) p(\theta) \\
&\propto p(\underline{m} \mid \underline{s}, \theta) p(\theta) \text{ by Assumption 1} \\
&\propto p(\underline{m} \mid \underline{s}, \theta) \text{ by Assumption 3} \\
&= N_k(\underline{\mu}, \Sigma) \text{ by Assumption 2} \\
&= \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} \left(\underline{m} - \left(\underline{\alpha} + \underline{\beta}\theta \right) \right)' \Sigma^{-1} \left(\underline{m} - \left(\underline{\alpha} + \underline{\beta}\theta \right) \right) \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \left((\underline{m} - \underline{\alpha}) - \underline{\beta}\theta \right)' \Sigma^{-1} \left((\underline{m} - \underline{\alpha}) - \underline{\beta}\theta \right) \right\} \\
&= \exp\left\{ -\frac{1}{2} \left[\theta^2 \underline{\beta}' \Sigma^{-1} \underline{\beta} - 2\theta \underline{\beta}' \Sigma^{-1} (\underline{m} - \underline{\alpha}) + (\underline{m} - \underline{\alpha})' \Sigma^{-1} (\underline{m} - \underline{\alpha}) \right] \right\} \\
&= \exp\left\{ -\frac{1}{2} \left(\underline{\beta}' \Sigma^{-1} \underline{\beta} \right) \left[\theta^2 - 2\theta \frac{\underline{\beta}' \Sigma^{-1} (\underline{m} - \underline{\alpha})}{\underline{\beta}' \Sigma^{-1} \underline{\beta}} + \right. \right. \\
&\quad \left. \left. \frac{(\underline{m} - \underline{\alpha})' \Sigma^{-1} (\underline{m} - \underline{\alpha})}{\underline{\beta}' \Sigma^{-1} \underline{\beta}} \right] \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \exp\left\{-\frac{1}{2}\left(\underline{\beta}'\Sigma^{-1}\underline{\beta}\right)\left[\theta^2 - 2\theta\frac{\underline{\beta}'\Sigma^{-1}(\underline{m}-\underline{\alpha})}{\underline{\beta}'\Sigma^{-1}\underline{\beta}} + \right.\right. \\
&\quad \left.\left.\frac{\underline{\beta}'\Sigma^{-1}\underline{\beta}\,(\underline{m}-\underline{\alpha})'\,\Sigma^{-1}(\underline{m}-\underline{\alpha})}{\left(\underline{\beta}'\Sigma^{-1}\underline{\beta}\right)^2}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\underline{\beta}'\Sigma^{-1}\underline{\beta}\right)\left[\theta^2 - 2\theta\frac{\underline{\beta}'\Sigma^{-1}(\underline{m}-\underline{\alpha})}{\underline{\beta}'\Sigma^{-1}\underline{\beta}} + \frac{\left(\underline{\beta}'\Sigma^{-1}(\underline{m}-\underline{\alpha})\right)^2}{\left(\underline{\beta}'\Sigma^{-1}\underline{\beta}\right)^2}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\frac{\left(\theta - \frac{\underline{\beta}'\Sigma^{-1}(\underline{m}-\underline{\alpha})}{\underline{\beta}'\Sigma^{-1}\underline{\beta}}\right)^2}{\left(\underline{\beta}'\Sigma^{-1}\underline{\beta}\right)^{-1}}\right\} \\
&\propto N\left(\frac{\underline{\beta}'\Sigma^{-1}(\underline{m}-\underline{\alpha})}{\underline{\beta}'\Sigma^{-1}\underline{\beta}}, \left(\underline{\beta}'\Sigma^{-1}\underline{\beta}\right)^{-1}\right),
\end{aligned}
$$

which is a univariate normal distribution with mean $\sum_{i,j}\beta_i\sigma^{ij}(m_j-\alpha_j)/\sum_{i,j}\beta_i\sigma^{ij}\beta_j$ and variance $(\sum_{i,j}\beta_i\sigma^{ij}\beta_j)^{-1}$ □

Aside from the priors themselves, Assumption 2 is the dominant influence in the pooled prior because it forces N's posterior distribution for θ to be normally distributed. In addition, the decision maker N is required by Assumption 2 to set values for several hyperparameters which together express N's opinion of the expertise of $E_1, \ldots, E_k$. α_i is the amount by which N expects E_i to overestimate θ (thus a negative α_i reflects a tendency for underestimation). β_i is a multiplicative bias term for E_i, allowing the bias to vary linearly with θ. The hyperparameter γ_i allows N to express how accurately she feels that E_i has assessed his own precision. For instance, $\gamma_i > 1$ indicates that N believes that E_i overconfidently states his precision in estimating θ. Finally, the correlation terms ρ_{ij} in the matrix R allow N to evaluate the similarities between each pair of experts in estimating θ—high ρ_{ij} corresponds to pairs of experts who share a common knowledge base or who frequently offer similar opinions about parameters like θ. In all, N must choose values for $3k + \frac{k(k-1)}{2}$ hyperparameters, which requires N to possess vast knowledge of the assessment abilities of the k experts.

The requirement that a supra-Bayesian decision maker sets reasonable values for all hyperparameters is both the primary strength and weakness of the supra-Bayesian approach. It is a strength in that, theoretically at least, the supra-Bayesian can adjust for potential disruptive features such as lack of calibration, dishonesty, and significant correlation. These features are important concerns in much of the supra-Bayesian literature since the solicitation and production of expert opinions is often an inexact science. Certain experts may be consistently optimistic, or overconfident in their estimation abilities, or prone to rely on the same knowledge base as another expert. Under the supra-Bayesian approach, an informed, intelligent supra-Bayesian decision maker has the ability to adjust for these potential disturbances.

However, it is probably very unlikely that any decision maker has the knowledge or the access to relevant data which would allow him to accurately assess values for all hyperparameters. Without knowing the truth about θ, the decision maker would need to accurately assess additive bias, multiplicative bias, precision, and correlation for all experts—quite a difficult task. Consider, for instance, the pooling problem resulting from attempts to form Bayesian inference about parameters in the bowhead whale population model introduced in Section 4.6. The decision maker solicits opinions from one or more laboratory biologists about the input parameters and from one or more oceanographers about the output parameters. Then, the model is used to express opinions in the same parameter space, say output space. In that case, the laboratory biologists expressed opinions in terms of carrying capacity and productivity, but their opinions are being expressed in terms of population size and replacement yield. The biologists and oceanographers, besides working in different fields and different locations with diverse research goals, may be unaware that their areas of knowledge can be linked by a deterministic population model. The decision maker is unlikely to be an

expert in all of these areas, and hence it is difficult to imagine one person being able to adequately evaluate all experts.

6.1.2 Multivariate Case

Lindley did not consider the case in which $\underline{\theta}$ is a vector of parameters rather than just a single quantity of interest. However, multivariate $\underline{\theta}$ is a common and important possibility. Deterministic simulation models used to model phenomena such as global climate, AIDS transmission, and wildlife population can easily involve tens or even hundreds of output and input variables. To form Bayesian inference about parameters in these models. a decision maker needs to pool two multidimensional distributions. As a result, we extended Lindley's ideas to the case of multivariate $\underline{\theta}$. Before stating Theorem 6.2 for multivariate $\underline{\theta}$, new notation must be defined and several assumptions made.

Let α_{ij} be an additive bias term, β_{ij} be a multiplicative bias term, and γ_{ij} be a precision adjustment term for E_i's estimate of $\theta_j, i = 1, \ldots, k$ and $j = 1, \ldots, m$. In addition, let ρ_{iajb} be N's assessment of the correlation between E_i's estimate of θ_a and E_j's estimate of θ_b, $i \neq j$, whereas ρ_{iab} is E_i's own assessment of correlation between his estimates of θ_a and θ_b. Then, consider the case in which the supra-Bayesian decision maker N solicits opinions from k experts regarding the $m \times 1$ vector of parameters $\underline{\theta}$. Each expert $E_i, i = 1, \ldots, k$, provides a mean vector $\underline{m_i} = (m_{i1}, \ldots, m_{im})'$ and a covariance matrix

$$S_i = \begin{bmatrix} s_{i1}^2 & \rho_{i12} s_{i1} s_{i2} & \cdots & \rho_{i1m} s_{i1} s_{im} \\ \rho_{i12} s_{i1} s_{i2} & s_{i2}^2 & \cdots & \rho_{i2m} s_{i2} s_{im} \\ \vdots & \vdots & & \vdots \\ \rho_{i1m} s_{i1} s_{im} & \rho_{i2m} s_{i2} s_{im} & \cdots & s_{im}^2 \end{bmatrix}$$

for his estimate of the joint prior of $\underline{\theta}$. The necessary assumptions can be generalized from those above as follows:

Assumption 1'. $p(S_1, \ldots, S_k \mid \underline{\theta}) = p(S_1, \ldots, S_k)$ so that there is no scale information.

Assumption 2′. $p(\underline{m}_1, \ldots, \underline{m}_k \mid, S_1, \ldots, S_k, \underline{\theta})$ is distributed according to the following multivariate normal distribution:

$$N_{mk}\left(\begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \\ \vdots \\ \underline{\mu}_k \end{bmatrix}, \begin{bmatrix} A_1 & B_{12} & \ldots B_{1k} \\ B_{21} & A_2 & \ldots B_{2k} \\ \vdots & \vdots & \vdots \\ B_{k1} & B_{k2} & \ldots A_k \end{bmatrix}\right) = N_{mk}(\underline{\mu}, \Sigma).$$

where

$$\underline{\mu}_i = \begin{bmatrix} \alpha_{i1} + \beta_{i1}\theta_1 \\ \alpha_{i2} + \beta_{i2}\theta_2 \\ \vdots \\ \alpha_{im} + \beta_{im}\theta_m \end{bmatrix},$$

for $i = 1, \ldots, k$,

$$A_i = \begin{bmatrix} \gamma_{i1}^2 s_{i1}^2 & \rho_{i12}\gamma_{i1}s_{i1}\gamma_{i2}s_{i2} & \cdots & \rho_{i1m}\gamma_{i1}s_{i1}\gamma_{im}s_{im} \\ \rho_{i12}\gamma_{i1}s_{i1}\gamma_{i2}s_{i2} & \gamma_{i2}^2 s_{i2}^2 & \cdots & \rho_{i2m}\gamma_{i2}s_{i2}\gamma_{im}s_{im} \\ \vdots & \vdots & & \vdots \\ \rho_{i1m}\gamma_{i1}s_{i1}\gamma_{im}s_{im} & \rho_{i2m}\gamma_{i2}s_{i2}\gamma_{im}s_{im} & \cdots & \gamma_{im}^2 s_{im}^2 \end{bmatrix}$$

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

$$B_{ij} = \begin{bmatrix} \rho_{i1j1}\gamma_{i1}s_{i1}\gamma_{j1}s_{j1} & \rho_{i1j2}\gamma_{i1}s_{i1}\gamma_{j2}s_{j2} & \cdots & \rho_{i1jm}\gamma_{i1}s_{i1}\gamma_{jm}s_{jm} \\ \rho_{i2j1}\gamma_{i2}s_{i2}\gamma_{j1}s_{j1} & \rho_{i2j2}\gamma_{i2}s_{i2}\gamma_{j2}s_{j2} & \cdots & \rho_{i2jm}\gamma_{i2}s_{i2}\gamma_{jm}s_{jm} \\ \vdots & \vdots & & \vdots \\ \rho_{imj1}\gamma_{im}s_{im}\gamma_{j1}s_{j1} & \rho_{imj2}\gamma_{im}s_{im}\gamma_{j2}s_{j2} & \cdots & \rho_{imjm}\gamma_{im}s_{im}\gamma_{jm}s_{jm} \end{bmatrix}$$

for $i = 1, \ldots, k$, $j = 1, \ldots, k$, and $i \neq j$. Note that $B_{ij} = B'_{ji}$, so that $\rho_{iajb} = \rho_{jbia}$.

Assumption 3′. $p(\underline{\theta})$ is an improper, flat (noninformative) prior; in other words, N's knowledge of $\underline{\theta}$ is weak compared with the information provided by the experts.

We can then generalize Theorem 6.1 to multivariate $\underline{\theta}$.

Theorem 6.2 *If E_i states that $\underline{\theta}$ has mean $\underline{m}_i$ and covariance matrix $S_i, (i = 1, \ldots, k)$, and Assumptions $1' - 3'$ apply, then the Lindley-Winkler supra-Bayesian pooled prior distribution for $\underline{\theta}$ is multivariate normal with mean vector*

$$\underline{c} = VZ'\mathrm{Diag}(\underline{\beta})\Sigma^{-1}(\underline{\mathrm{m}} - \underline{\alpha})$$

and covariance matrix

$$V = \left[Z' \mathrm{Diag}(\underline{\beta}) \Sigma^{-1} \mathrm{Diag}(\underline{\beta}) \mathrm{Z} \right]^{-1},$$

where

$$\underline{\alpha} = \begin{bmatrix} \alpha_{11} \\ \vdots \\ \alpha_{im} \\ \alpha_{21} \\ \vdots \\ \alpha_{2m} \\ \vdots \\ \alpha_{k1} \\ \vdots \\ \alpha_{km} \end{bmatrix}, \underline{\beta} = \begin{bmatrix} \beta_{11} \\ \vdots \\ \beta_{im} \\ \beta_{21} \\ \vdots \\ \beta_{2m} \\ \vdots \\ \beta_{k1} \\ \vdots \\ \beta_{km} \end{bmatrix}, \text{and } \mathrm{Z} = \begin{bmatrix} I_m \\ I_m \\ \vdots \\ I_m \end{bmatrix}.$$

Also, $\mathrm{Diag}(\underline{\beta})$ *is a* $km \times km$ *matrix with* β_{ij} *terms along the diagonal and zeros elsewhere, and* I_m *is an* $m \times m$ *identity matrix.*

Proof.

$$\begin{aligned}
& p(\underline{\theta} \mid \underline{m}_1, \ldots, \underline{m}_k, S_1, \ldots, S_k) \\
& \quad \propto \; p(\underline{m}_1, \ldots, \underline{m}_k, S_1, \ldots, S_k \mid \underline{\theta}) p(\underline{\theta}) \\
& \quad = \; p(\underline{m}_1, \ldots, \underline{m}_k \mid S_1, \ldots, S_k, \underline{\theta}) p(S_1, \ldots, S_k \mid \underline{\theta}) p(\underline{\theta}) \\
& \quad \propto \; p(\underline{m}_1, \ldots, \underline{m}_k \mid S_1, \ldots, S_k, \underline{\theta}) p(\underline{\theta}) \text{ by Assumption } 1' \\
& \quad \propto \; p(\underline{m}_1, \ldots, \underline{m}_k \mid S_1, \ldots, S_k, \underline{\theta}) \text{ by Assumption } 3' \\
& \quad = \; N_{mk}(\underline{\mu}, \Sigma) \text{ by Assumption } 2' \text{ for } \underline{\mu}, \Sigma \text{ defined previously} \\
& \quad = \; \frac{1}{(2\pi)^{mk/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} \left(\underline{m} - \left(\underline{\alpha} + \mathrm{Diag}(\underline{\beta}) \mathrm{Z} \underline{\theta} \right) \right)' \Sigma^{-1} \left(\underline{m} - \left(\underline{\alpha} + \mathrm{Diag}(\underline{\beta}) \mathrm{Z} \underline{\theta} \right) \right) \right\} \\
& \quad \propto \; \exp\left\{ -\frac{1}{2} \left[(\underline{m} - \underline{\alpha})' \Sigma^{-1} (\underline{m} - \underline{\alpha}) - 2 (\underline{m} - \underline{\alpha})' \Sigma^{-1} \mathrm{Diag}(\underline{\beta}) \mathrm{Z} \underline{\theta} + \right. \right. \\
& \qquad \left. \left. \underline{\theta}' Z' \mathrm{Diag}(\underline{\beta}) \Sigma^{-1} \mathrm{Diag}(\underline{\beta}) \mathrm{Z} \underline{\theta} \right] \right\} \qquad (6.1)
\end{aligned}$$

At this point, let $V = \left(Z'\text{Diag}(\underline{\beta})\Sigma^{-1}\text{Diag}(\underline{\beta})\text{Z}\right)^{-1}$ and $\underline{c} = VZ'\text{Diag}(\underline{\beta})\Sigma^{-1}(\underline{\text{m}} - \underline{\alpha})$. Note that

$$\begin{aligned}
\underline{\theta}'V^{-1}\underline{\theta} &= \underline{\theta}'Z'\text{Diag}(\underline{\beta})\Sigma^{-1}\text{Diag}(\underline{\beta})\text{Z}\underline{\theta} \\
\underline{c}'V^{-1}\underline{\theta} &= (\underline{m} - \underline{\alpha})'\Sigma^{-1}\text{Diag}(\underline{\beta})\text{Z}\underline{\theta} \\
\underline{c}'V^{-1}\underline{c} &= (\underline{m} - \underline{\alpha})'W^{-1}(\underline{m} - \underline{\alpha}),
\end{aligned}$$

where $W^{-1} = \Sigma^{-1}A[A'\Sigma^{-1}A]^{-1}A'\Sigma^{-1}$ and $A = \text{Diag}(\underline{\beta})\text{Z}$, and assuming that Σ is nonsingular and $\beta_{ij} \neq 0$ for all $i = 1, \ldots, k$ and $j = 1, \ldots, m$.

Therefore, Equation 6.1 can be rewritten as

$$\exp\left\{-\frac{1}{2}\left[\underline{c}'V^{-1}\underline{c} - 2\underline{c}'V^{-1}\underline{\theta} + \underline{\theta}'V^{-1}\underline{\theta}\right]\right.$$

$$\left. - \frac{1}{2}\left[(\underline{m} - \underline{\alpha})'\Sigma^{-1}(\underline{m} - \underline{\alpha}) - (\underline{m} - \underline{\alpha})'W^{-1}(\underline{m} - \underline{\alpha})\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}(\underline{\theta} - \underline{c})'V^{-1}(\underline{\theta} - \underline{c})\right\} \propto N_m(\underline{c}, V)$$

which is a multivariate normal distribution with mean vector $\underline{c} = VZ'\text{Diag}(\underline{\beta})\Sigma^{-1}(\underline{\text{m}} - \underline{\alpha})$ and covariance matrix $V = \left(Z'\text{Diag}(\underline{\beta})\Sigma^{-1}\text{Diag}(\underline{\beta})\text{Z}\right)^{-1}$ □

For multivariate $\underline{\theta}$, the number of hyperparameters required from the supra-Bayesian decision maker N is considerable. Similar to the case with univariate θ, the decision maker must assess the additive bias, multiplicative bias, and error in the precision evaluation for each expert and each outcome variable. In addition, the decision maker must also assess the correlation between each pair of outcomes within each pair of experts. With k experts and m outcome variables, this translates to $3mk + \frac{1}{2}m^2k(k-1)$ hyperparameter assessments for the decision maker—a daunting task indeed. In fact, this model could be extended even farther to allow N to adjust the within-expert correlation terms (ρ_{iab} for $i = 1, \ldots, k$, $a = 1, \ldots, m$, and $b = 1, \ldots, m$). The multivariate version of the method presented here requires N to accept the ρ_{iab}-values assessed by E_i, but N could be given more flexibility with the introduction of even more hyperparameters.

Assumptions $1' - 3'$ can be relaxed to make the method more realistic in certain situations. For instance, Assumption $1'$ would be unrealistic if N feels

that certain experts will have more difficulty estimating θ_J for large values of θ_J. Lindley offers two suggestions in this case—apply a logarithmic transformation to θ_J and assume no scale information, or make a distributional assumption for s_J^2, such as $\nu\lambda^2\theta_J^2/s^2 \sim \chi_\nu^2$ for suitable λ and ν. Assumption 3′ can also be relaxed, especially to cases in which $p(\theta)$ is normal, conjugate to the likelihood.

Assumption 2′ is potentially the most unrealistic, yet at the same time the most influential assumption. While admittedly convenient, the assumption of multivariate normality forces the resulting posterior distribution for $\underline{\theta}$ to also be multivariate normal. This may be unrealistic, as we illustrate with several examples in Section 6.4. Moreover, the dependence on a supra-Bayesian decision maker to provide accurate values for the vast array of hyperparameters is extremely optimistic. A few suggestions can be made in this regard. First, model restrictions can be imposed to limit the number of hyperparameters the supra-Bayesian must define. For instance, no bias would imply $\underline{\alpha} = \underline{0}$ and $\underline{\beta} = \underline{1}$, or no correlation between any pair of experts would imply $\rho_{iajb} = 0$ for $i \neq j$, $a = 1, \ldots, m$, and $b = 1, \ldots, m$. Second, any uncertainty in the determination of certain hyperparameter values can be incorporated through a hierarchical Bayes model. In fact, Lindley considers γ to be the toughest parameter to assess, so in the case with $k = 1$ and $m = 1$, he proposes assuming that $\nu c^2/\gamma^2 \sim \chi_\nu^2$ for suitable ν and c. The density for γ is therefore proportional to $\exp[-\nu c^2/2\gamma^2]\,\gamma^{-(\nu+1)}$, so that $\log(\gamma)$ is approximately normal with mean log(c) and variance $(2\nu)^{-1}$. With this hierarchical Bayes model, the posterior distribution of N's judgment about θ has a t-distribution instead of a normal distribution, and Lindley illustrates several advantages of the t-distribution. When $m = 1$ and $k > 1$, Winkler (1981) suggests using an inverse Wishart distribution to model the covariance matrix of $\underline{m}$. Although this assumption leads to a t-distribution for the posterior of θ, it is severely limited by only having one parameter available to express uncertainty.

6.2 The Winkler Method

At this point, a short summary of Winkler (1981) is appropriate, since Lindley and Winkler developed overlapping methods after attacking the problem from different perspectives. Lindley required each expert to provide the first two moments of the prior distributions for the quantity of interest θ, and then he focused on the parameterization of a supra-Bayesian decision maker's opinion of the experts. The decision maker's updated opinion of θ (in light of the expert opinions) is formed using Bayes' rule, treating the expert opinions as data. Winkler, on the other hand, focuses on stochastic dependence among the experts' errors of estimation. Winkler assumes no particular general form for the consensus distribution. If the decision maker has a prior density $h_o(\theta)$ for a single parameter of interest θ, then the posterior consensus distribution for θ is expressed as:

$$h(\theta \mid g_1, \ldots, g_k, f, h_o) \propto h_o(\theta) f(\mu_1 - \theta, \ldots, \mu_k - \theta),$$

where f = the decision maker's density of $(\mu_1 - \theta, \ldots, \mu_k - \theta)$, g_i = Expert i's prior distribution for θ, μ_i = the mean of Expert i's prior, and $\mu_i - \theta$ = the error of estimation for Expert i. Winkler assumes that (i) $\underline{u} = (u_1, \ldots, u_k)' = (\mu_1 - \theta, \ldots, \mu_k - \theta)'$ is location-invariant, and (ii) g_i has been calibrated to reflect biases or incorrect precision assessment which the decision maker feels exists in Expert i's prior.

In particular, Winkler applies his pooling method to the case in which f is restricted to the family of k-variate normal densities such that

$$f(\underline{\mu} - \theta) \sim N_k \left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & \sigma_k^2 \end{bmatrix} \right) = N_k(\underline{0}, \Sigma)$$

Thus, Expert i's (calibrated) distribution of θ is approximately $N(\mu_i, \sigma_i^2)$, while the dependence between Experts i and j is captured by the correlation term $\rho_{ij} = \sigma_{ij}/\sigma_i\sigma_j$ for $i \neq j$.

With Σ known and a noninformative prior for θ, the posterior density for θ is proportional to $N(\mu^*, \sigma^{*2})$ where $\mu^* = \underline{e}'\Sigma^{-1}\underline{\mu}/\underline{e}'\Sigma^{-1}\underline{e}$, $\sigma^{*2} = 1/\underline{e}'\Sigma^{-1}\underline{e}$, and $\underline{e}$ = the $k \times 1$ vector $[1, \ldots, 1]'$. This is the same result produced by the Lindley method for the case with $m = 1$, $k > 1$, $\underline{\alpha} = \underline{0}$, $\underline{\beta} = \underline{e}$, $\underline{\gamma} = \underline{e}$, and R assumed known—i.e. the case in which the decision maker makes no adjustments to the experts' priors, or the experts' priors are perfectly calibrated.

When Σ is unknown, Winkler recommends the inverted Wishart distribution, a natural-conjugate prior distribution for the covariance matrix of a normal process, as a satisfactory approximation for the distribution of Σ. In this case, with $f(\underline{\mu} - \theta) \propto N_k(\underline{0}, \Sigma)$ and a noninformative prior, the consensus distribution is a t-density.

6.3 Extensions to the Lindley-Winkler Method

Several enhancements or additions have been proposed to this Lindley-Winkler method in recent years. Agnew (1985) extended the method under Winkler's framework to multivariate $\underline{\theta}$. In the general case, Agnew assumes that the $mk \times 1$ error vector $\underline{u} = \underline{\mu} - \underline{\theta} \sim N_{mk}(X\underline{\theta}, \Sigma)$, where $X = I_m \otimes \underline{e}_k$. ($I_m$ is an $m \times m$ identity matrix, $\underline{e}_k$ is a $k \times 1$ vector of ones, and $\otimes$ is a Kronecker product.) With Σ supplied by the decision maker (i.e., Σ is known) and a noninformative prior, the decision maker's posterior distribution for $\underline{\theta}$ is proportional to $N_m((X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\underline{\mu}, (X'\Sigma^{-1}X)^{-1})$. This parallels Theorem 6.2 for the case in which the decision maker decides that no adjustment of the experts' priors is necessary. If Σ is unknown, an inverted Wishart distribution for Σ can be used, leading to a multivariate-t posterior distribution for $\underline{\theta}$.

Agnew discusses in detail some special cases of his general multivariate normal model for dependence. In particular, he considers dependent experts with independent random variables, and independent experts with dependent random variables. Under the former scenario, Agnew obtains the posterior density for θ proportional to $N(\mu^*, \sigma^{*2})$ from Winkler's model extended to multivariate $\underline{\theta}$. This is the same result we derived based on the Lindley framework, given that $\rho_{iajb} = 0$ for $i \neq j$, $a = 1, \ldots, m$, and $b = 1, \ldots, m$.

West (1988) extends the Lindley framework to situations in which the experts express their opinions about θ in partial terms, using quantiles of the distribution for θ. A key aspect of West's development is the assumption that the distribution of these quantiles is implicitly determined by a Dirichlet process with precision parameter δ. Then, if $p(\theta)$ represents the decision maker's prior for θ, the decision maker's posterior distribution is

$$p(\theta \mid H) \propto p(\theta) \exp(-\delta D(\theta)),$$

where

$$D(\theta) = \int_{-\infty}^{\infty} \alpha[F(X)] f(X) \log \left[\frac{f(X)}{m_Y(X)} \right] dX.$$

$D(\theta)$ is a measure of divergence between the target density $M_\theta(x)$ and the stated expert density $f(x)$ for each θ. $M_\theta(x)$ is the anticipated form of the distribution of the median conditional on the true value θ, adjusting for the bias and lack of calibration. $\alpha[U]$ is a density function whose domain is the unit interval. If $M_\theta(x)$ is $(x \mid \theta) \sim N(c + \theta, W)$ and if a single expert offers a mean and standard deviation for his prior distribution of θ, then West's method produces the same likelihood and posterior distributions as Lindley's method with a single expert and a single quantity of interest. West's method is not restricted to distributions in the exponential family (which are strongly dependent on location parameters), which

he illustrates with an example involving the Cauchy distribution, but extension of West's method beyond a single expert and single outcome would not be trivial.

Gelfand (1995) builds on both Lindley and West. He allows experts to express their beliefs about a quantity of interest θ in one of two partial ways—by providing probabilities for disjoint, exhaustive intervals in the domain of Θ, or by providing a set of quantiles for the distribution of Θ. Gelfand advocates handling opinions of both types using the family of mixtures of beta distributions. He illustrates extensions of his model to multiple experts, but not to multivariate $\underline{\theta}$. This approach is more computationally demanding than other supra-Bayesian approaches outlined here, with its reliance on the Gibbs sampler to handle missing-data likelihoods.

6.4 Examples

In this section, we apply the Lindley-Winkler supra-Bayesian method to the examples from Chapter 4. One significant hurdle to the practical implementation of the Lindley-Winkler method is the definition of reasonable hyperparameter values. The interpretations of the hyperparameters, as discussed in Section 6.1.1, are not often quantities easily appraised or intuitively known by the supra-Bayesian. As a result, we investigate the effect of the choice of hyperparameter values on the resulting consensus prior, using variations from a set of baseline parameter values. For examples with one input and one output, we set $\alpha_1 = \alpha_2 = 0$, $\beta_1 = \beta_2 = 1$, $\gamma_1 = \gamma_2 = 1$, and $\rho = 0$. With these settings, the supra-Bayesian adds no information to the two expert priors, as she considers each expert to be unbiased, uncorrelated, and a perfect assessor of his own precision. Thus, as a baseline reference point we parallel the logarithmic pooling method of Chapter 4, in which no attempt is made to adjust for systematic imperfections in the experts. These particular baseline settings are also most appropriate for many of the deterministic simulation model inference problems which might be considered. Often the

supra-Bayesian who solicits and collects expert priors on model parameters is not an expert herself on the model parameters. Without thorough knowledge of the field of each expert, she can not accurately quantify characteristics such as bias and precision. In addition, one might argue that less expert calibration is required in the deterministic simulation model setting, because experts surveyed have often spent years or decades studying the model parameters in question, and they consider it a natural exercise to express their beliefs about the parameters in terms of probability distributions.

Settings for the Lindley-Winkler model applied to $\Re^2 \rightarrow \Re^2$ examples are merely multivariate extensions of these $\Re^1 \rightarrow \Re^1$ settings. Also note that, although focus is constrained to ϕ-space, similar hyperparameter effects would be seen in θ-space.

6.4.1 $\Re^1 \rightarrow \Re^1$ Examples

The Lindley-Winkler method was developed for continuous variables, so application of this method to the two $\Re^1 \rightarrow \Re^1$ examples requires no major adjustments. To begin, the first two moments of $p_1^*(\phi)$ and $p_2(\phi)$ must be derived. In both examples, the mean and standard deviation of p_2 are known. In the introductory example the mean and standard deviation of p_1^* are also known, while for the cusp example these values can be easily estimated numerically. A consensus prior is then given by Theorem 6.1, which states that

$$p(\theta \mid \underline{m}, \underline{s}) \propto N\left(\frac{\underline{\beta}'\Sigma^{-1}(\underline{m} - \underline{\alpha})}{\underline{\beta}'\Sigma^{-1}\underline{\beta}}, \left(\underline{\beta}'\Sigma^{-1}\underline{\beta}\right)^{-1}\right),$$

where values for $\underline{\alpha}$, $\underline{\beta}$, and certain components of Σ must be determined by Lindley's Bayesian decision maker N. With $\Re^1 \rightarrow \Re^1$ models, N must assign values to 7 hyperparameters: $\alpha_i, i = 1, 2$, is the additive bias for Expert i; $\beta_i, i = 1, 2$, is the multiplicative bias for Expert i; $\gamma_i, i = 1, 2$, is a multiplicative precision adjustment for Expert i; and, ρ is the correlation between experts.

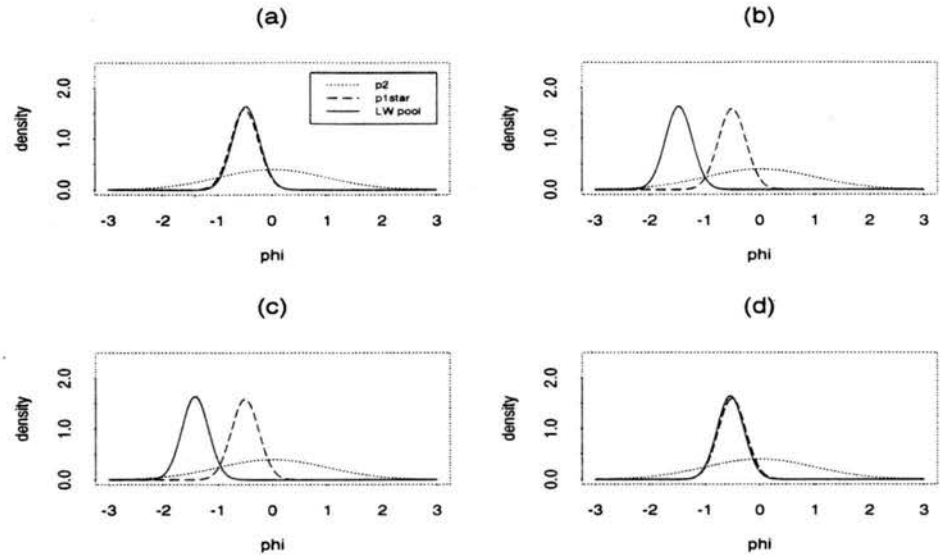

Figure 6.1: Introductory example: $p(\phi)$ using the Lindley-Winkler method. Fixed hyperparameters: $\beta_1 = \beta_2 = 1$, $\gamma_1 = \gamma_2 = 1$, and $\rho = 0$. Variable hyperparameters: (a) $\alpha_1 = 0, \alpha_2 = 0$, (b) $\alpha_1 = 1, \alpha_2 = 1$, (c) $\alpha_1 = 1, \alpha_2 = 0$, (d) $\alpha_1 = 0, \alpha_2 = 1$.

Introductory Example

The effects on the pooled Lindley-Winkler prior of varying hyperparameter values for the introductory $\Re^1 \rightarrow \Re^1$ example of Section 4.2 are illustrated in Figures 6.1- 6.3. Figure 6.1 shows the effects of the additive bias term α. Plot (a) represents baseline settings of all hyperparameters, reflecting a judgment by the decision maker N that the experts are unbiased, uncorrelated, and accurate estimators of their own precision. These settings produce $p(\phi)$ which closely resembles $p_1^*(\phi)$. In plot (b), both experts were judged to have overestimated ϕ ($\alpha_1 = \alpha_2 = 1$), so that $p(\phi)$ is shifted left when compared to $p(\phi)$ in plot (a). Plots (c) and (d) allow one expert at a time to have nonnegative additive bias. $\alpha_1 = 1$ shifts $p(\phi)$ left by approximately one unit, but $\alpha_2 = 1$ has little effect on $p(\phi)$. These panels show that the Lindley-Winkler pooled prior is influenced mainly by the more precise of the two original priors, which in this example is that of Expert 1. In addition, caution must be exercised when setting hyperparameter values to reflect lack of calibration by an expert, since the pooled prior quickly moves into regions not supported by the original two priors. The resulting pooled

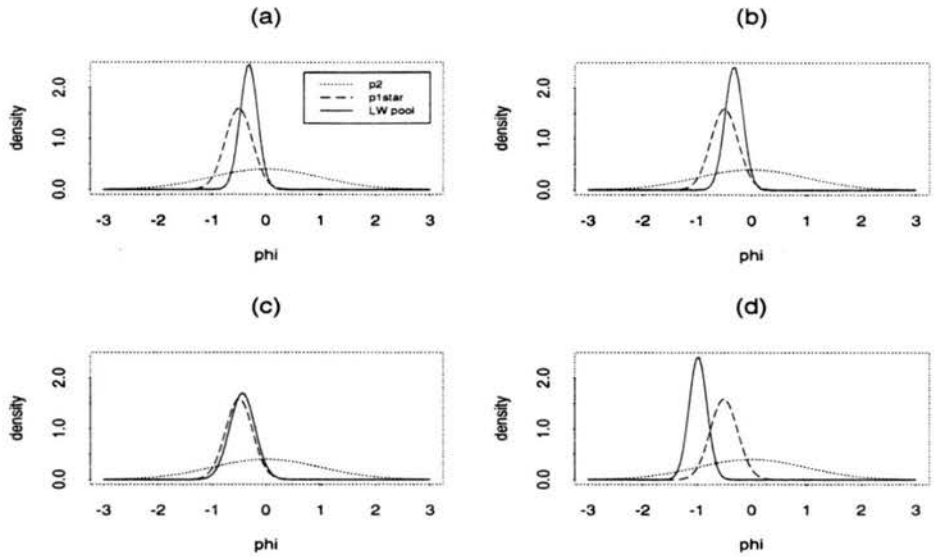

Figure 6.2: Introductory example: $p(\phi)$ using the Lindley-Winkler method. Fixed hyperparameters: $\alpha_1 = \alpha_2 = 0$ (except (d)), $\gamma_1 = \gamma_2 = 1$, and $\rho = 0$. Variable hyperparameters: (a) $\beta_1 = 1.5, \beta_2 = 1.5$, (b) $\beta_1 = 1.5, \beta_2 = 1$, (c) $\beta_1 = 1, \beta_2 = 1.5$, (d) $\alpha_1 = 1, \beta_1 = 1.5, \beta_2 = 1$.

distribution can change greatly based on a single judgment of N—a judgment for which data rarely exists.

Figure 6.2 illustrates the effects of the multiplicative bias term β. Multiplicative bias of 1.5, for instance, indicates that the expert tends to overestimate ϕ by 50%. In plot (a), $\beta_1 = \beta_2 = 1.5$, which pulls $p(\phi)$ towards 0 and reduces its variability. Plots (b) and (c) allow one expert at a time to have multiplicative bias of 1.5. $\beta_1 = 1.5$ has basically the same impact as $\beta_1 = \beta_2 = 1.5$, while $\beta_2 = 1.5$ produces only slight location and scale changes in $p(\phi)$. Again the pooled prior is influenced mainly by Expert 1, the more precise of the two experts. Finally, in plot (d), Expert 1 has both additive and multiplicative bias ($\alpha_1 = 1, \beta_1 = 1.5$). The multiplicative bias, in this case, tempers the shift left caused by $\alpha_1 = 1$ and reduces the variability in $p(\phi)$.

Figure 6.3 illustrates the effects of the precision adjustment terms (γ_1 and γ_2) and the correlation term (ρ). A precision adjustment of 2, for instance, indicates that the expert was overconfident in stating his precision, and that standard

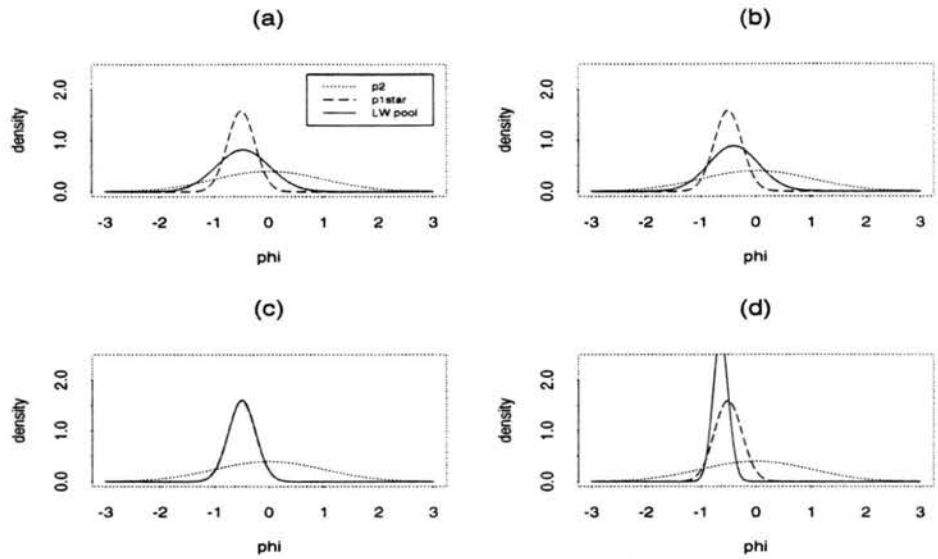

Figure 6.3: Introductory example: $p(\phi)$ using the Lindley-Winkler method. Fixed hyperparameters: $\alpha_1 = \alpha_2 = 0$, and $\beta_1 = \beta_2 = 1$. Variable hyperparameters: (a) $\gamma_1 = 2, \gamma_2 = 2, \rho = 0$, (b) $\gamma_1 = 2, \gamma_2 = 1, \rho = 0$, (c) $\gamma_1 = 1, \gamma_2 = 2, \rho = 0$, (d) $\gamma_1 = 1, \gamma_2 = 1, \rho = 0.9$.

deviation terms must be adjusted up by a factor of 2. In plot (a), $\gamma_1 = \gamma_2 = 2$, which increases the variability in $p(\phi)$. Plots (b) and (c) require one expert at a time to adjust his precision by a factor of 2. $\gamma_1 = 2$ has basically the same impact on the variability of $p(\phi)$ as $\gamma_1 = \gamma_2 = 2$, while also producing a slight shift right. $\gamma_2 = 2$ has very little impact on $p(\phi)$, causing a very small variance increase and shift left when compared to $\gamma_2 = 1$. Finally, plot (d) introduces strong dependence between the two experts ($\rho = 0.9$). The mode of $p(\phi)$ is actually below the modes for both $p_1^*(\phi)$ and $p_2(\phi)$. In addition, $p(\phi)$ exhibits a reduced variance, which seems counterintuitive since high correlation implies that less information is provided than by the two independent experts.

Cusp Example

Hyperparameter sensitivity for the cusp example of Section 4.3 is illustrated in Figure 6.4. Plot (a) represents basic settings of all hyperparameters—essentially perfectly calibrated and independent experts. $p(\phi)$ is a normal distribution centered near $\phi = 4.5$. One thing to immediately note about the application of the

Lindley-Winkler approach to this example is that the zero preservation property, even the sensible weak form, is not followed. Because of the Lindley-Winkler method's requirement (through Assumption 2) that the pooled prior be normally distributed, $p(\phi)$ assigns non-negligible probability to $\phi < 0$, where both p_1^* and p_2 are 0, and to $\phi > 12.5$, where p_1^* is 0. This feature, especially the fact that the pooled probability can be positive where all experts assigned zero probability, is disturbing. Furthermore, visually and intuitively, this consensus prior does not seem to capture the regions of greatest agreement between the two experts, probably because it is difficult for a normal distribution to adequately describe the pooled beliefs about ϕ of the two experts in this example.

In plot (b) $\beta_2 = 0.5$ and in plot (c) $\gamma_1 = 2$; both settings display the effects one would expect based on Figures 6.2- 6.3 for the introductory example. The pooled density shifts away from 0 and spreads out in plot (b) as Expert 2's prior was adjusted for alleged underestimation of ϕ. In plot (c), the pooled density flattens slightly as Expert 1's prior is adjusted for alleged overly optimistic precision estimates. Plot (d) features highly dependent experts ($\rho = 0.9$), and $p(\phi)$ shifts toward the more precise original prior (p_2) while also displaying greater variability because of the information loss from dependent sources. Note that in plots (c) and (d), significant amounts of pooled probability are nonsensically located where $\phi < 0$, a region in which both experts assigned zero probability.

Since the restriction that $p(\theta \mid \underline{m}, \underline{s})$ be normally distributed produced inadequate consensus distributions in this example, we considered a modification of the Lindley-Winkler framework. We primarily hoped to obtain priors which better represented a consensus about ϕ between the two experts while avoiding values of ϕ which both experts agreed had no probability of occurring. As a result, we modified the Lindley-Winkler approach so that $p(\theta \mid \underline{m}, \underline{s})$ is distributed as a log-normal distribution. First, assume that each expert offers an opinion

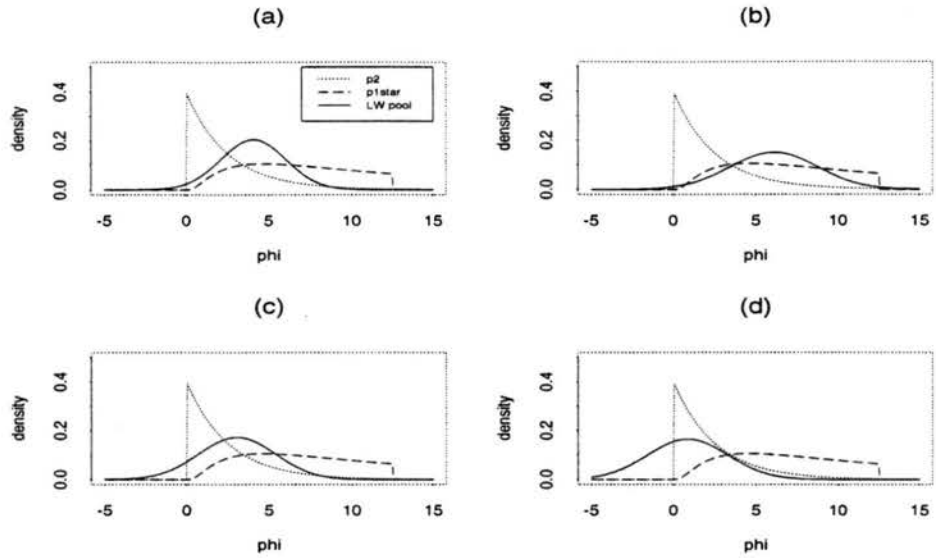

Figure 6.4: Cusp example: $p(\phi)$ using the Lindley-Winkler method with normality. Fixed hyperparameters: $\alpha_1 = \alpha_2 = 0$, $\beta_1 = 1$, and $\gamma_2 = 1$. Variable hyperparameters: (a) $\beta_2 = 1, \gamma_1 = 1, \rho = 0$, (b) $\beta_2 = 0.5, \gamma_1 = 1, \rho = 0$, (c) $\beta_2 = 1, \gamma_1 = 2, \rho = 0$, (d) $\beta_2 = 1, \gamma_1 = 1, \rho = 0.9$.

about $\xi = \log(\phi)$ in the form of a mean (m) and standard deviation (s). Then, Assumptions 1-3 must be made, where in each case ξ is substituted for θ, so that:

1. $p(\underline{s} \mid \xi) = p(\underline{s})$

2. $p(\underline{m} \mid \underline{s}, \xi)$ is multivariate normal with means $\alpha_i + \beta_i \xi$, standard deviations $\gamma_i s_i$, and correlations ρ_{ij}; and,

3. $p(\xi)$ is an improper, flat (noninformative) prior

Through an application of Theorem 6.1,

$$p(\xi \mid \underline{m}, \underline{s}) \propto N\left(\frac{\underline{\beta}'\Sigma^{-1}(\underline{m} - \underline{\alpha})}{\underline{\beta}'\Sigma^{-1}\underline{\beta}}, \left(\underline{\beta}'\Sigma^{-1}\underline{\beta}\right)^{-1}\right).$$

Thus, ϕ is distributed log-normal, or

$$p(\phi \mid \underline{m}, \underline{s}) \propto \frac{1}{\phi} \exp\left\{-\frac{1}{2}\frac{\left(\log(\phi) - \frac{\underline{\beta}'\Sigma^{-1}(\underline{m}-\underline{\alpha})}{\underline{\beta}'\Sigma^{-1}\underline{\beta}}\right)^2}{\left(\underline{\beta}'\Sigma^{-1}\underline{\beta}\right)^{-1}}\right\}.$$

Hyperparameters in this case describe the decision maker's evaluation of each expert's stated opinion about $\xi = \log(\phi)$. One immediate concern, therefore, is

that the decision maker must not only evaluate the abilities of each expert, but she now must base hyperparameter values on effects of the experts' shortcomings on their evaluation of $\log(\phi)$. As an example, non-zero additive bias on the log scale turns into nonlinear bias with respect to ϕ. Nevertheless, sensitivity of the Lindley-Winkler log-normal pool to choices of the hyperparameters is illustrated in Figures 6.5- 6.7. Figure 6.5 shows the effects of the additive bias term α. Plot (a) contains basic settings for all hyperparameters; for these basic settings, the log-normal pool appears to do a better job than the normal pool at defining a region of consensus and avoiding $\phi < 0$ where $p_1^*(\phi) = p_2(\phi) = 0$. Distribution functions other than log-normal are also feasible.

In plot (b) both experts were judged to have overestimated $\log(\phi)$ and predictably the mass of $p(\phi)$ shifts left, although $p(\phi)$ is now more precise than either of the two original priors. In plot (c), Expert 1 overestimated $\log(\phi)$ but interestingly (and counterintuitively) the mass of $p(\phi)$ shifts right when compared to an unbiased Expert 1 from plot (a). In fact, despite evidence through the hyperparameters than Expert 1 overestimated $\log(\phi)$ (and in turn overestimated ϕ), the pooled density in plot (c) has considerable mass where $\phi > 12.5$, a region of zero probability according to Expert 1. In contrast, plot (d) illustrates the impact of overestimation by Expert 2, and the mass of $p(\phi)$ shifts left of the comparable $p(\phi)$ featuring an unbiased Expert 2.

Figure 6.6 illustrates the effects of the multiplicative bias term β. Interpretation of β-parameters under log-normality is especially difficult. As β increases, $p(\phi)$ should concentrate near $\phi = 1$. As β decreases, however, the effect on $p(\phi)$ must be described in two cases. If $\phi < 1, p(\phi)$ shifts left, while if $\phi > 1$ $p(\phi)$ shifts right. In all cases, the effect of changes in β on $p(\phi)$ are nonlinear in nature. In plot (a), with 50% overestimation by both experts, the mass of $p(\phi)$ is indeed pulled toward 1. With overestimation by only a single expert (plots (b) and (c)),

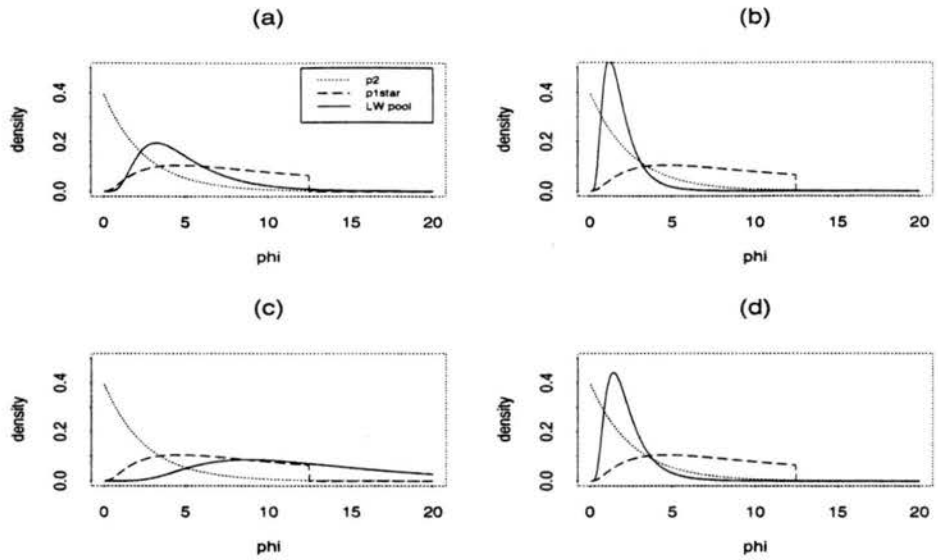

Figure 6.5: Cusp example: $p(\phi)$ using the Lindley-Winkler method with log-normality. Fixed hyperparameters: $\beta_1 = \beta_2 = 1$, $\gamma_1 = \gamma_2 = 1$, and $\rho = 0$. Variable hyperparameters: (a) $\alpha_1 = 0, \alpha_2 = 0$, (b) $\alpha_1 = 1, \alpha_2 = 1$, (c) $\alpha_1 = 1, \alpha_2 = 0$, (d) $\alpha_1 = 0, \alpha_2 = 1$.

$p(\phi)$ shifts left as in plot (a), but to lesser degrees. Overestimation by Expert 1 has a larger effect on the consensus distribution than similar overestimation by Expert 2. Finally, with Expert 1 exhibiting both additive and multiplicative bias on $\log(\phi)$, $p(\phi)$ is pulled strongly toward $\phi = 1$ (plot (d)), and the pooled density becomes highly concentrated.

Figure 6.7 illustrates the effects of the precision adjustment terms (γ_1 and γ_2) and the correlation term (ρ). Again, the hyperparameters relate to experts' assessments of $\log(\phi)$, so interpretation in terms of ϕ is not very intuitive. Plots (a)-(c) exhibit an adjustment of one or both experts' standard deviation claims up by a factor of 2. As expected, these precision adjustments increased the variance of $p(\phi)$, in a few cases indicating significant pooled probability for $\phi > 12.5$, where Expert 1 assigned zero probability originally. As we have noted previously, when the variability in Expert 1's prior increases, Expert 2's prior has a stronger influence on the consensus distribution, and vice versa. This effect is illustrated in plots (b) and (c). Finally, plot (d) exhibits a disturbingly counterintuitive

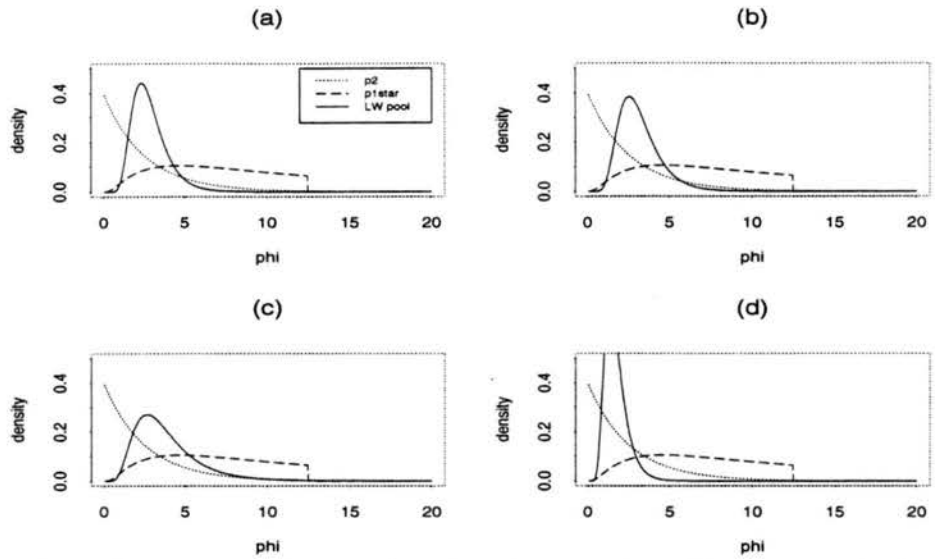

Figure 6.6: Cusp example: $p(\phi)$ using the Lindley-Winkler method with log-normality. Fixed hyperparameters: $\alpha_1 = \alpha_2 = 0$ (except (d)), $\gamma_1 = \gamma_2 = 1$, and $\rho = 0$. Variable hyperparameters: (a) $\beta_1 = 1.5, \beta_2 = 1.5$, (b) $\beta_1 = 1.5, \beta_2 = 1$, (c) $\beta_1 = 1, \beta_2 = 1.5$, (d) $\alpha_1 = 1, \beta_1 = 1.5, \beta_2 = 1$.

$p(\phi)$ formed under strong dependence between the two experts—the probability associated with ϕ peaks near $\phi = 10$ and basically ignores smaller values of ϕ where both experts assign fair amounts of probability.

6.4.2 $\Re^2 \to \Re^2$ Examples

Extension of the Lindley-Winkler method to multidimensional parameter space is described in Section 6.1.2. Here we examine some applications for deterministic simulation models with 2 inputs and 2 outputs. Each expert must report a mean vector with corresponding covariance matrix to describe his or her beliefs about $\boldsymbol{\phi}$. In both examples, the mean vectors and covariance matrices of p_2 are known. In the linear example the mean vector and covariance matrix of p_1^* are also known, while for the noninvertible example these values can be easily estimated numerically. A consensus prior is then given by Theorem 6.2, which states that:

$$p(\underline{\phi} \mid \underline{m}_1, \underline{m}_2, S_1, S_2) \propto \exp\left\{-\frac{1}{2}(\underline{\theta} - \underline{c})' V^{-1} (\underline{\theta} - \underline{c})\right\} \propto N_m(\underline{c}, V)$$

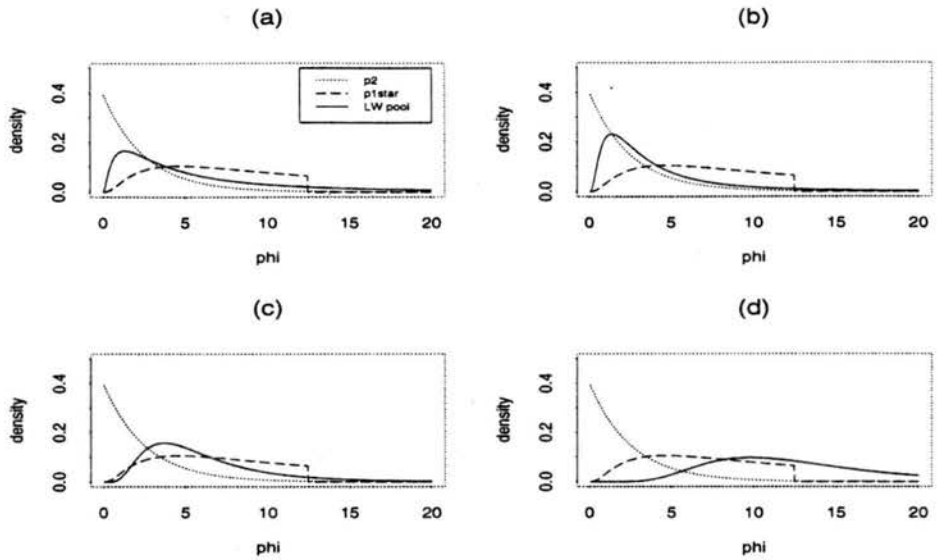

Figure 6.7: Cusp example: $p(\phi)$ using the Lindley-Winkler method with log-normality. Fixed hyperparameters: $\alpha_1 = \alpha_2 = 0$, and $\beta_1 = \beta_2 = 1$. Variable hyperparameters: (a) $\gamma_1 = 2, \gamma_2 = 2, \rho = 0$, (b) $\gamma_1 = 2, \gamma_2 = 1, \rho = 0$, (c) $\gamma_1 = 1, \gamma_2 = 2, \rho = 0$, (d) $\gamma_1 = 1, \gamma_2 = 1, \rho = 0.9$.

which is a multivariate normal distribution with mean vector $\underline{c}_{2\times 1} = VZ'\text{Diag}(\underline{\beta})\Sigma^{-1}(\underline{\text{m}} - \underline{\alpha})$ and covariance matrix $V_{2\times 2} = \left(Z'\text{Diag}(\underline{\beta})\Sigma^{-1}\text{Diag}(\underline{\beta})\text{Z}\right)^{.-1}$ Hyperparameter values again must be determined by the decision maker N. With $\Re^2 \rightarrow \Re^2$ models, N must assign values to 16 hyperparameters: $\underline{\alpha} = [\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}]'$, where α_{ij} is the additive bias for expert i and outcome variable j; $\underline{\beta} = [\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}]'$, where β_{ij} is the multiplicative bias for expert i and outcome variable j; $\underline{\gamma} = [\gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}]'$, where γ_{ij} is the multiplicative precision adjustment for expert i and outcome variable j; and, $\underline{\rho} = [\rho_{11}, \rho_{12}, \rho_{21}, \rho_{22}]'$, where ρ_{ij} is the correlation between Expert 1's assessment of outcome variable i and Expert 2's assessment of outcome variable j.

Linear $\Re^2 \rightarrow \Re^2$ Example

The sensitivity of the Lindley-Winkler pool for the linear $\Re^2 \rightarrow \Re^2$ example of Section 4.4 to hyperparameter values is illustrated in Figures 6.8- 6.11. The Lindley-Winkler joint consensus distribution is illustrated by the grey-scale image plot, with darker shading indicating regions of higher $p(\phi)$. The contours repre-

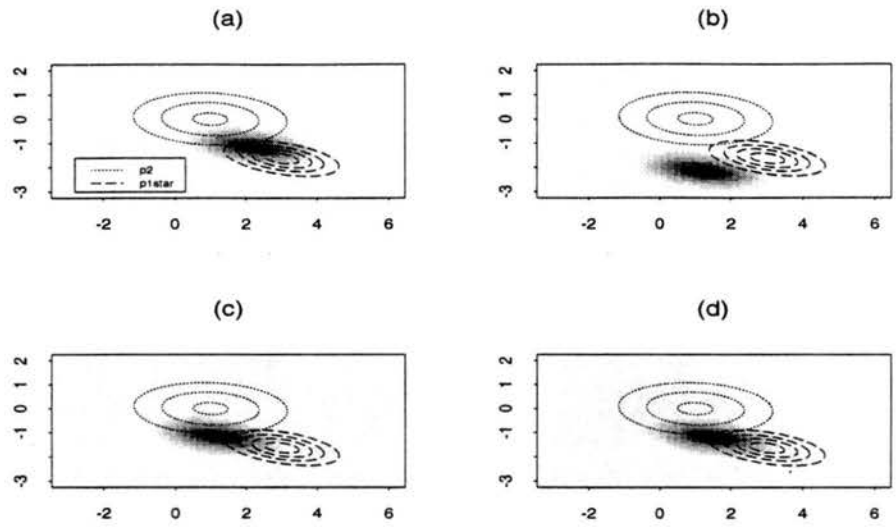

Figure 6.8: Linear $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) using the Lindley-Winkler method. Fixed hyperparameters: $\underline{\beta} = \underline{1}$, $\underline{\gamma} = \underline{1}$, and $\underline{\rho} = \underline{0}$. Variable hyperparameters: (a) $\underline{\alpha} = [0, 0, 0, 0]'$, (b) $\underline{\alpha} = [1, 1, 1, 1]'$, (c) $\underline{\alpha} = [1, 0, 1, 0]'$, (d) $\underline{\alpha} = [1, 0, 0, 0]'$.

sent the original priors p_1^* and p_2. Figure 6.8 shows the effects of the additive bias term $\underline{\alpha}$. Plot (a) represents basic settings of all hyperparameters, reflecting a judgment by the decision maker N that the experts are unbiased, uncorrelated, and accurate estimators of their own precision with respect to both outcome variables. The basic settings produce a $p(\phi)$ which appears to be a reasonable compromise between $p_1^*(\phi)$ and $p_2(\phi)$; its closer proximity to $p_1^*(\phi)$ stems from the smaller level of uncertainty associated with Expert 1's opinion. In plots (b)-(d), additive bias is assigned to various combinations of experts and outcome variables. As expected, $p(\phi)$ undergoes a location shift but not a scale shift. For instance, in plot (b) $p(\phi)$ is shifted both in the ϕ_1 and ϕ_2 directions, while in plot (c) $p(\phi)$ is only shifted in the ϕ_1-direction since additive bias only exists for the ϕ_1-terms. Although changes in $\underline{\alpha}$ produce the changes in $p(\phi)$ which would be intuitively expected, the plots in Figure 6.8 underscore the extreme caution the decision maker must use when choosing values for the hyperparameters. The introduction of bias

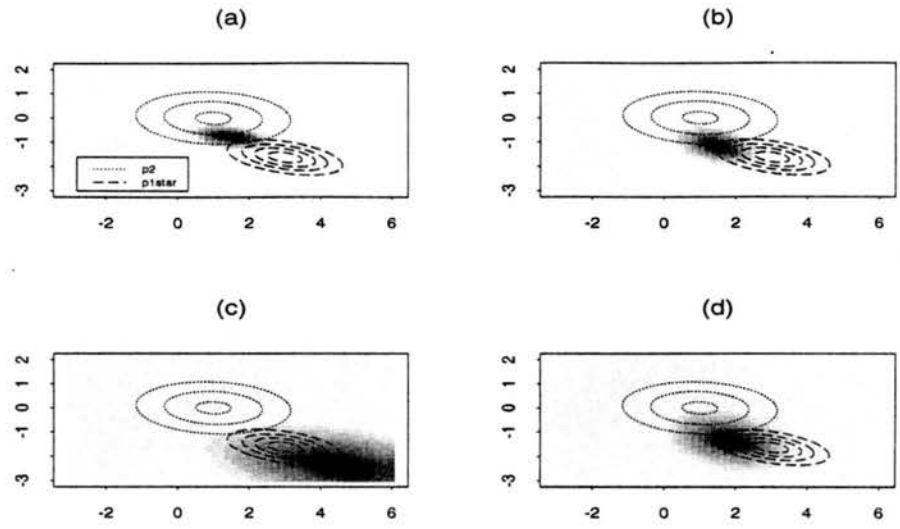

Figure 6.9: Linear $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) using the Lindley-Winkler method. Fixed hyperparameters: $\underline{\alpha} = \underline{0}$, $\underline{\gamma} = \underline{1}$, and $\underline{\rho} = \underline{0}$. Variable hyperparameters: (a) $\underline{\beta} = [1.5, 1.5, 1.5, 1.5]'$, (b) $\underline{\beta} = [1.5, 1, 1.5, 1]'$, (c) $\underline{\beta} = [0.5, 0.5, 0.5, 0.5]'$, (d) $\underline{\beta} = [1, 0.5, 1, 1]'$.

correction, often based on a single, uninformed judgment by the supra-Bayesian, changes $p(\phi)$ markedly and must not be taken lightly.

Figure 6.9 shows the effects of the multiplicative bias term $\underline{\beta}$. Changes in $\underline{\beta}$ affect both the location and scale of $p(\phi)$. For example, with all β_{ij} terms set to 1.5 in plot (a), indicating 50% overestimation by each expert on each variable, the adjusted $p(\phi)$ is less variable and has modes closer to 0 on both axes. In contrast, with all β_{ij} terms set to 0.5 in plot (c), indicating 50% underestimation by each expert on each variable, the adjusted $p(\phi)$ is more variable and has modes farther from 0 on both axes. In plots (b) and (d), it can be seen that the effect of a change in β_{ij} with respect to just one variable is restricted to that one variable. For instance, in plot (b) $\beta_{11} = \beta_{21} = 1.5$, and $p(\phi)$ undergoes location (left) and scale (reduced) shifts along the ϕ_1-axis but not the ϕ_2-axis. These plots in Figure 6.9 illustrate the large effect on $p(\phi)$ that the introduction of multiplicative bias can cause, and again suggest that the supra-Bayesian must be extremely confident and careful when setting hyperparameter values.

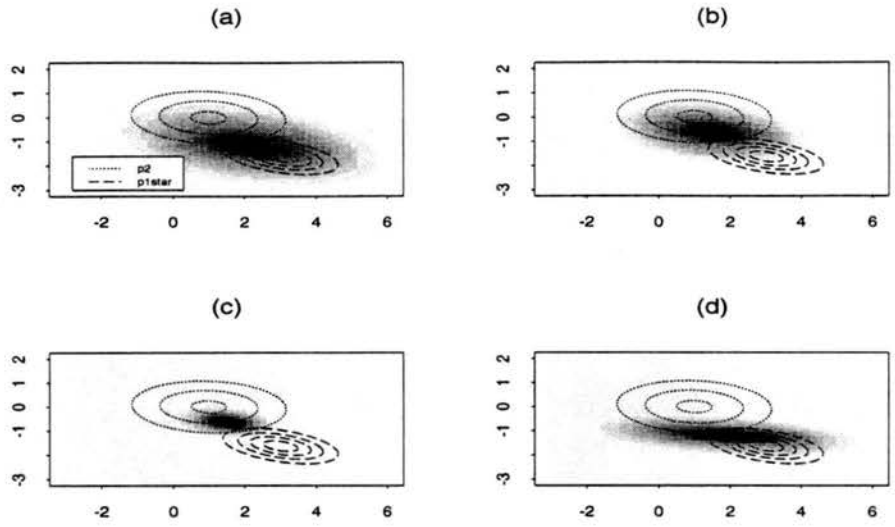

Figure 6.10: Linear $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) using the Lindley-Winkler method. Fixed hyperparameters: $\underline{\alpha} = \underline{0}$, $\underline{\beta} = \underline{1}$, and $\underline{\rho} = \underline{0}$. Variable hyperparameters: (a) $\underline{\gamma} = [2, 2, 2, 2]'$, (b) $\underline{\gamma} = [2, 2, 1, 1]'$, (c) $\underline{\gamma} = [1, 1, 0.5, 0.5]'$, (d) $\underline{\gamma} = [2, 1, 2, 1]'$.

Figure 6.10 shows the effects of the precision adjustment term $\underline{\gamma}$. In plot (a) all γ_{ij} terms are set to 2, reflecting N's evaluation that each expert overstated their confidence with respect to each variable, so that all stated standard deviations should be doubled. As expected, the adjusted $p(\phi)$ (compared to fixing all γ_{ij}'s at 1) shows greater variance in all directions and no location shift. In plots (b) and (c), the precision statements of only one of the two experts are adjusted, and in these cases changes in both location and scale are noted. In plot (b), $\gamma_{11} = \gamma_{12} = 2$, and $p(\phi)$ is more variable and shifted closer to $p_2(\phi)$because of the greater uncertainty in Expert 1's beliefs. In plot (c), $\gamma_{21} = \gamma_{22} = 0.5$, and $p(\phi)$ is less variable. In addition, $p(\phi)$ again shifts toward $p_2(\phi)$, reflecting the increased relative confidence associated with Expert 2's opinion. Finally, plot (d) allows $\gamma_{11} = \gamma_{21} = 2$. Thus, both experts overstated their confidence regarding their opinion about ϕ_1. In this case, $p(\phi)$ expands along the ϕ_1-axis to reflect greater uncertainty, but its location remains the same. Changes in $\underline{\gamma}$ produced the corresponding changes in $p(\phi)$ which one might intuitively expect, but care

must be exercised when introducing a precision adjustment, since it will alter $p(\boldsymbol{\phi})$ in terms of overall spread and which expert's opinion exerts more influence.

Figure 6.11 shows the effects of the correlation term $\underline{\rho}$. In general, changes in $\underline{\rho}$ alter location slightly, but their biggest impact are in the size and shape of the density cloud for $p(\boldsymbol{\phi})$. In plot (a) all ρ_{ij} terms are set to 0.9, indicating a high level of dependence between experts for each pair of variables. The resulting $p(\boldsymbol{\phi})$ exhibits more spread, likely from a decrease in available information compared to independent experts, and a positive correlation, due to high levels of ρ_{12} and ρ_{21}. When all ρ_{ij} terms are set to 0.2 in plot (b), the resulting $p(\boldsymbol{\phi})$ looks similar to $p(\boldsymbol{\phi})$ under no correlation, except with less of a negative correlation. In plot (c), correlation is assumed to exist only between experts' opinions on the same variable. The resulting $p(\boldsymbol{\phi})$ has a similar negative correlation but greater overall spread than $p(\boldsymbol{\phi})$ under zero correlation. Finally, plot (d) evaluates $p(\boldsymbol{\phi})$ for a $\underline{\rho}$ which could be realistic in certain situations—strong correlation between experts with respect to ϕ_1, fair correlation between experts with respect to ϕ_2, and low (but non-zero) correlation between experts with respect to opposite ϕ's.

Noninvertible $\Re^2 \rightarrow \Re^2$ Example

The sensitivity of the Lindley-Winkler pool for the noninvertible $\Re^2 \rightarrow \Re^2$ example of Section 4.5 to hyperparameter values is illustrated in Figures 6.12- 6.15. Hyperparameter settings in these plots are identical to those in the linear $\Re^2 \rightarrow \Re^2$ example. One thing to note about the application of the Lindley-Winkler approach to this example is that the strong form of the zero preservation property is not followed. Even though $p_1^*(\boldsymbol{\phi})= 0$ in the region above $\phi_1 = \phi_2 - \frac{1}{4}$, the Lindley-Winkler $p(\boldsymbol{\phi})$ need not be 0 for these $\boldsymbol{\phi}$-points; in fact, some hyperparameter settings produce $p(\boldsymbol{\phi})$ with significant mass in that region. Of course, the strong form of ZPP is probably too restrictive to be desirable, so the lack of strong ZPP is

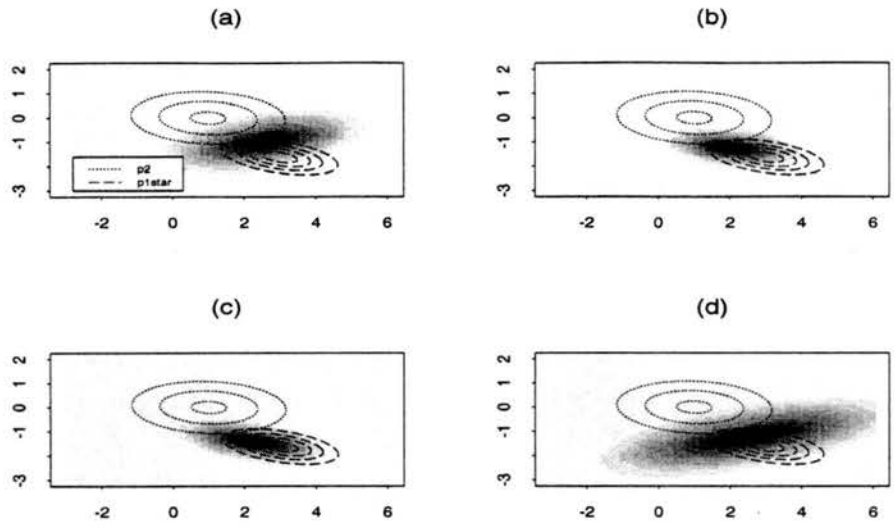

Figure 6.11: Linear $\Re^2 \rightarrow \Re^2$ example: $p(\boldsymbol{\phi})$ (shaded density) using the Lindley-Winkler method. Fixed hyperparameters: $\underline{\alpha} = \underline{0}$, $\underline{\beta} = \underline{1}$, and $\underline{\gamma} = \underline{1}$. Variable hyperparameters: (a) $\underline{\rho} = [0.9, 0.9, 0.9, 0.9]'$, (b) $\underline{\rho} = [0.2, 0.2, 0.2, 0.2]'$, (c) $\underline{\rho} = [0.5, 0, 0, 0.5]'$, (d) $\underline{\rho} = [0.9, 0.2, 0.2, 0.5]'$.

not necessarily a negative feature, but rather a result of the normality assumption employed.

Figure 6.12 shows the effects of $\underline{\alpha}$. Plot (a) represents basic settings of all hyperparameters, and $p(\boldsymbol{\phi})$ in this case follows very closely to $p_2(\boldsymbol{\phi})$; because $p_1^*(\boldsymbol{\phi})$ has its probability mass spread over a much broader region than $p_2(\boldsymbol{\phi})$, $p_1^*(\boldsymbol{\phi})$ has very little influence over the pooled distribution. In plots (b)-(d), additive bias is assigned to various combinations of experts and outcome variables. As expected, $p(\boldsymbol{\phi})$ undergoes a location shift but not a scale shift. For instance, in plot (b) $p(\boldsymbol{\phi})$ is shifted both in the ϕ_1 and ϕ_2 directions, while in plot (c) $p(\boldsymbol{\phi})$ is only shifted in the ϕ_1-direction since additive bias only exists for the ϕ_1-terms. In plot (d), a correction for Expert 1's overestimate of ϕ_1 produces only a small shift left in $p(\boldsymbol{\phi})$. As in the other examples in this chapter, Figure 6.12 illustrates that simple changes in the hyperparameter settings can cause significant changes in $p(\boldsymbol{\phi})$. Sometimes a certain shift in $p(\boldsymbol{\phi})$ will be expected and desired by the supra-Bayesian, but troubling changes such as in plot (c) can also occur, where

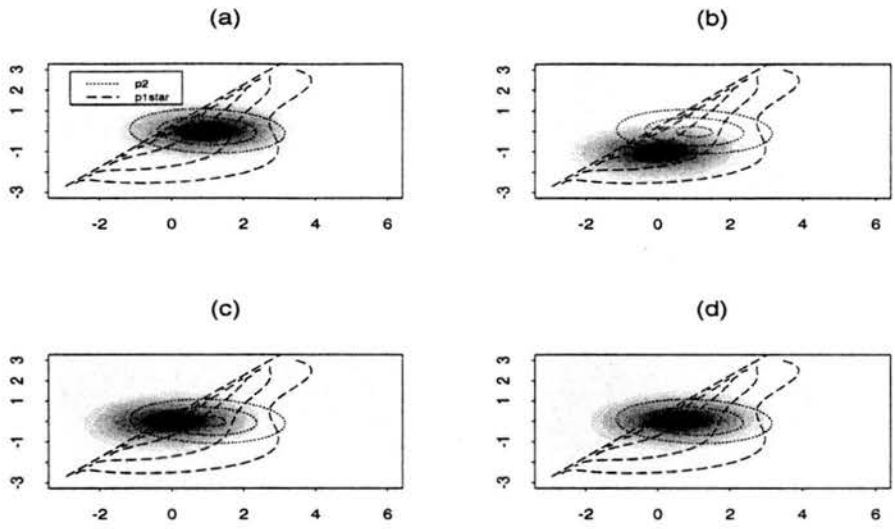

Figure 6.12: Noninvertible $\Re^2 \rightarrow \Re^2$ example: $p(\underline{\phi})$ (shaded density) using the Lindley-Winkler method. Fixed hyperparameters: $\underline{\beta} = \underline{1}$, $\underline{\gamma} = \underline{1}$, and $\underline{\rho} = \underline{0}$. Variable hyperparameters: (a) $\underline{\alpha} = [0,0,0,0]'$, (b) $\underline{\alpha} = [1,1,1,1]'$, (c) $\underline{\alpha} = [1,0,1,0]'$, (d) $\underline{\alpha} = [1,0,0,0]'$.

nearly half of the mass of $p(\underline{\phi})$ is in a region of $\underline{\phi}$-space originally considered impossible by Expert 1.

Figure 6.13 shows the effects of $\underline{\beta}$. Changes in $\underline{\beta}$ affect both the location and scale of $p(\underline{\phi})$. For example, with all β_{ij} terms set to 1.5 in plot (a), indicating 50% overestimation by each expert on each variable, the adjusted $p(\underline{\phi})$ is less variable and has mode closer to 0 on the ϕ_1-axis. (The mode of $p(\underline{\phi})$ on the ϕ_2-axis is essentially 0 under the basic settings.) In contrast, with all β_{ij} terms set to 0.5 in plot (c), indicating 50% underestimation by each expert on each variable, the adjusted $p(\underline{\phi})$ is more variable and has ϕ_1-mode farther from 0. In plots (b) and (d), it can be seen that the effect of a change in β_{ij} with respect to just one variable is restricted to that one variable. For instance, in plot (b) $\beta_{11} = \beta_{21} = 1.5$, and $p(\underline{\phi})$ undergoes location (left) and scale (reduced) shifts along the ϕ_1-axis but not the ϕ_2-axis. The shifts in $p(\underline{\phi})$ observed in Figure 6.13 match closely with intuition, but one also notes the dominance of Expert 2's opinion and the potential for significant changes in $p(\underline{\phi})$.

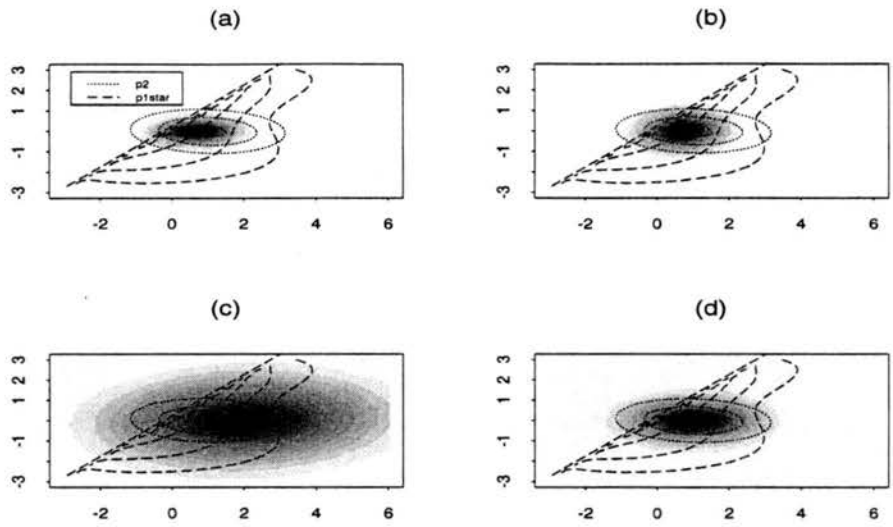

Figure 6.13: Noninvertible $\Re^2 \rightarrow \Re^2$ example: $p(\underline{\phi})$ (shaded density) using the Lindley-Winkler method. Fixed hyperparameters: $\underline{\alpha} = \underline{0}$, $\underline{\gamma} = \underline{1}$, and $\underline{\rho} = \underline{0}$. Variable hyperparameters: (a) $\underline{\beta} = [1.5, 1.5, 1.5, 1.5]'$, (b) $\underline{\beta} = [1.5, 1, 1.5, 1]'$, (c) $\underline{\beta} = [0.5, 0.5, 0.5, 0.5]'$, (d) $\underline{\beta} = [1, 0.5, 1, 1]'$.

Figure 6.14 shows the effects of the precision adjustment term $\underline{\gamma}$. In plot (a) all γ_{ij} terms are set to 2, reflecting N's evaluation that each expert overstated their confidence with respect to each variable, so that all stated standard deviations should be doubled. As expected, the adjusted $p(\underline{\phi})$ (compared to fixing all γ_{ij}'s at 1) shows greater variance in all directions and no location shift. In plots (b) and (c), the precision statements of only one of the two experts are adjusted, and in these cases changes in both location and scale are noted. In plot (b), $\gamma_{11} = \gamma_{12} = 2$, and $p(\underline{\phi})$ is more variable, reflecting the greater uncertainty in Expert 1's beliefs. In plot (c), $\gamma_{21} = \gamma_{22} = 0.5$, and $p(\underline{\phi})$ is less variable. No location shift as occurred with these settings in the linear $\Re^2 \rightarrow \Re^2$ example is noticeable, possibly due to the lack of influence of $p_1^*(\underline{\phi})$. Finally, plot (d) allows $\gamma_{11} = \gamma_{21} = 2$. Thus, both experts overstated their confidence regarding their opinion about ϕ_1. In this case, $p(\underline{\phi})$ expands along the ϕ_1-axis to reflect greater uncertainty, but its location remains the same. Figure 6.14 again illustrates the

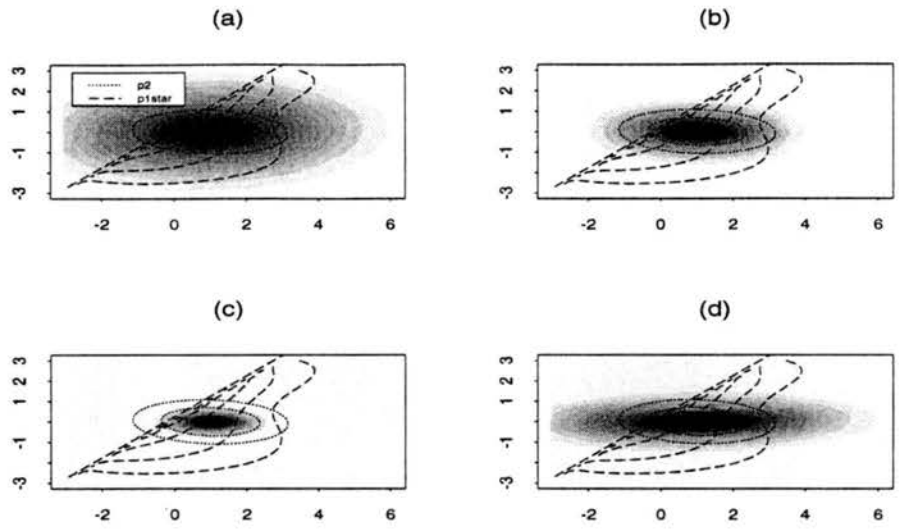

Figure 6.14: Noninvertible $\Re^2 \rightarrow \Re^2$ example: $p(\underline{\phi})$ (shaded density) using the Lindley-Winkler method. Fixed hyperparameters: $\underline{\alpha} = \underline{0}$, $\underline{\beta} = \underline{1}$, and $\underline{\rho} = \underline{0}$. Variable hyperparameters: (a) $\underline{\gamma} = [2, 2, 2, 2]'$, (b) $\underline{\gamma} = [2, 2, 1, 1]'$, (c) $\underline{\gamma} = [1, 1, 0.5, 0.5]'$, (d) $\underline{\gamma} = [2, 1, 2, 1]'$.

strong influence of Expert 2 and the potential changes in $p(\underline{\phi})$ which adjustments in precision can bring through one judgment of the decision maker.

Figure 6.15 shows the effects of the correlation term $\underline{\rho}$. In general, changes in $\underline{\rho}$ alter location slightly, but their biggest impact are in the size and shape of the density cloud for $p(\underline{\phi})$. In plot (a) all ρ_{ij} terms are set to 0.9, indicating a high level of dependence between experts for each pair of variables. The resulting $p(\underline{\phi})$ exhibits more spread, likely from a decrease in available information compared to independent experts, and a positive correlation, due to high levels of ρ_{12} and ρ_{21}. When all ρ_{ij} terms are set to 0.2 in plot (b), the resulting $p(\underline{\phi})$ looks similar to $p(\underline{\phi})$ under no correlation, except with a small positive correlation. In plot (c), correlation is assumed to exist only between experts' opinions on the same variable. The resulting $p(\underline{\phi})$ has greater overall spread than $p(\underline{\phi})$ under zero correlation. Finally, plot (d) evaluates $p(\underline{\phi})$ for a $\underline{\rho}$ which could be realistic in certain situations—strong correlation between experts with respect to ϕ_1, fair correlation between experts with respect to ϕ_2, and low (but non-zero) correlation

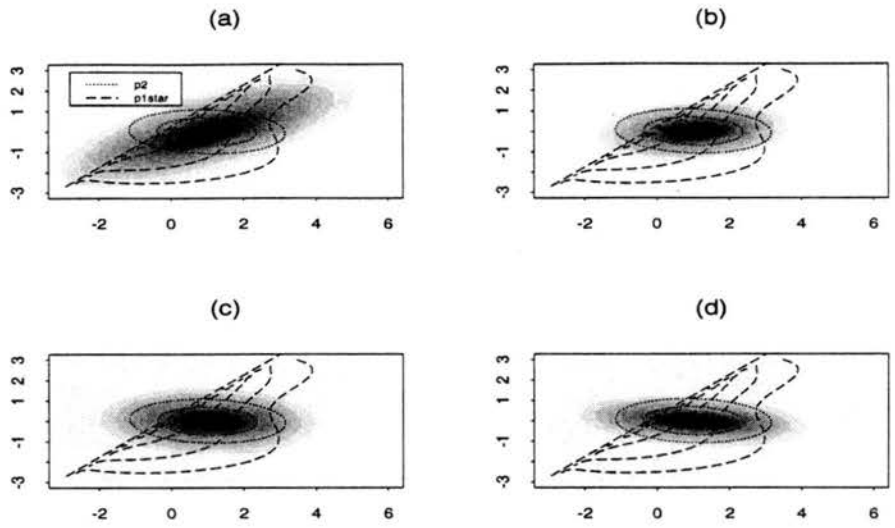

Figure 6.15: Noninvertible $\Re^2 \rightarrow \Re^2$ example: $p(\underline{\phi})$ (shaded density) using the Lindley-Winkler method. Fixed hyperparameters: $\underline{\alpha} = \underline{0}$, $\underline{\beta} = \underline{1}$, and $\underline{\gamma} = \underline{1}$. Variable hyperparameters: (a) $\underline{\rho} = [0.9, 0.9, 0.9, 0.9]'$, (b) $\underline{\rho} = [0.2, 0.2, 0.2, 0.2]'$, (c) $\underline{\rho} = [0.5, 0, 0, 0.5]'$, (d) $\underline{\rho} = [0.9, 0.2, 0.2, 0.5]'$.

between experts with respect to opposite ϕ's. Because of the strong influence of $p_2(\underline{\phi})$, changes in $\underline{\rho}$ had little effect until ρ_{ij} approached 1.

6.5 Summary

In general, the Lindley-Winkler supra-Bayesian method for pooling expert opinion performed reasonably well in certain cases, but limitations of the method prevented it from performing well in other cases. In the following paragraphs, the strengths and limitations of the Lindley-Winkler method will be outlined in the context of pooling priors linked by a deterministic simulation model.

The requirement that the pooled density be normally distributed can be unduly restrictive. The normality assumption simplifies the mathematics, allows for a closed-form expression of the decision maker's updated prior, and provides for intuitively appealing hyperparameters. These features are appealing compared to the French-Lindley framework. However, the bottom line is still to obtain a pooled distribution which represents a reasonable consensus between the beliefs

of the two experts, and the assumption of normality can prevent achievement of the bottom line. In some situations, such as the noninvertible $\Re^2 \rightarrow \Re^2$ and cusp examples, a normal pooled density does not adequately represent certain aspects of the problem, such as truncated distributions or decidedly non-normal expert priors. Under normality, the Lindley-Winkler method does not necessarily adhere to the weak version of the zero preservation property, which states that the pooled probability should be zero wherever all individual expert priors are zero. In the cusp example, we felt compelled to relax the normality assumption in favor of an assumption of log-normality. Our resulting poolings seemed to better represent a consensus between experts, but removing the normality assumptions brought other problems, such as in the interpretation of hyperparameters. By extending it to different or more general classes of distributions, the Lindley-Winkler framework becomes more applicable to a variety of practical situations, as long as the hyperparameters remain intuitive to the decision maker.

As with any supra-Bayesian method, careful and accurate assessments of expert abilities and the expression of these assessments with appropriate values of hyperparameters is of extreme importance. A strength of the Lindley-Winkler method—one supported by the examples in this chapter—is the intuitive nature of the hyperparameters. Changes in hyperparameter values lead to the expected corresponding changes in the pooled densities, in most cases. However, the resulting Lindley-Winkler poolings are tied too closely to the final hyperparameter values to make this method useful in most inference problems involving deterministic simulation models. In all the examples we investigated, small changes in a single hyperparameter could produce significant changes in the pooled prior. Thus, one judgment by the supra-Bayesian decision maker, often made in the absence of adequate data, can greatly impact the final result. In deterministic simulation modeling, the decision maker is often someone (such as a statistician)

who possesses some general knowledge of the field but none of the specific knowledge regarding individual model parameters that the surveyed experts do. In addition, the "input experts" and the "output experts" may work in mostly unrelated fields. As a result, the chances are small that a single, omniscient decision maker holds enough knowledge and insight to accurately evaluate each expert with respect to bias, precision, and correlation. And without such a decision maker, the Lindley-Winkler poolings can not easily be trusted. A misinformed or biased decision maker can too easily generate pooled densities which either do not sufficiently represent a consensus opinion or too closely reflect the beliefs of the decision maker.

Deterministic simulation models create another complication for the Lindley-Winkler or any other supra-Bayesian method. If we are interested in obtaining a consensus opinion about output parameters ϕ, then the supra-Bayesian decision maker must determine hyperparameter values for all experts in terms of ϕ. However, in a deterministic simulation model, one of the opinions to be pooled comes from an expert who expressed an opinion about input parameters θ, and then his opinion is transformed through the model into ϕ-space. Evaluating the ability of this "input expert" to state opinions about ϕ is a nearly impossible task, since the "input expert" may not even have knowledge of the model which maps inputs into outputs. Hyperparameters set in θ-space would require an appropriate transformation into ϕ-space and would probably lose interpretability in the process.

Another characteristic of the Lindley approach which is both an advantage and a shortcoming is the requirement that each expert offer a mean and standard deviation to reflect his or her beliefs about the parameter of interest. On one hand, it greatly simplifies the prior solicitation process by not requiring experts to express their beliefs in terms of entire distributions. It is much more realistic

to expect that experts can convert their opinions into two descriptive statistics rather than an entire probability distribution. On the other hand, the mean and standard deviation do not adequately describe many common distributions. In both the noninvertible $\Re^2 \rightarrow \Re^2$ and the cusp examples, information about the shapes of the priors was lost by soliciting only a mean and a standard deviation. As discussed in Section 6.3, the Lindley-Winkler framework is currently being extended to allow experts to express their beliefs in terms of sets of quantiles or sets of probabilities for disjoint intervals. These efforts should make the Lindley-Winkler method applicable to a wider range of situations.

In conclusion, the Lindley-Winkler method performs nicely under controlled circumstances when the normality assumption does not become too restrictive, but it is limited by its reliance on sound, sensible hyperparameters determined by the supra-Bayesian decision maker.

Chapter 7

A Comparison of Four Pooling Approaches

7.1 Description

In Chapter 2, several methods for combining expert opinion were summarized, their strengths and weaknesses sketched out. Then, in Chapters 3- 6, three specific approaches were discussed in greater detail, and their performances in several examples were evaluated.

In this chapter, we compare and contrast pooled prior distributions stemming from various approaches with respect to five different examples. The pooling approaches being compared are: logarithmic pooling, the French-Lindley supra-Bayesian approach, the Lindley-Winkler supra-Bayesian approach, and linear pooling. The pooled priors under these approaches will be denoted by $p_{\log}(\phi)$, $p_{\mathrm{FL}}(\phi)$, $p_{\mathrm{LW}}(\phi)$, and $p_{\mathrm{lin}}(\phi)$, respectively. Although linear pooling was not covered in detail in Chapters 3- 6, it is a common, easily-implemented approach which is natural to include in a large comparison of possible approaches to the problem of combining expert opinion. The five examples—introductory, cusp, linear $\Re^2 \to \Re^2$, noninvertible $\Re^2 \to \Re^2$, and whale population model—are first discussed in Chapter 4.

The goal of this comparison is to examine which method(s) are most suitable for pooling priors linked by deterministic simulation models. We consider suitability to depend primarily on three factors. First, the theoretical properties of a method must be considered. Properties such as external Bayesianity, rela-

tive propensity consistency, and the zero preservation property are introduced in Chapter 2, and their merits and desirability are discussed in Section 2.3. Methods which satisfy many of these desirable properties reduce the chances of encountering unanticipated difficulties with an application. Second, theoretically justifiable methods must exhibit robustness in practical situations—performing well over a wide variety of examples and applications. In this chapter, we evaluate each method over 5 examples ranging from single dimensions, normally distributed priors, and linear mappings to higher dimensions, non-normally distributed priors with truncation, and noninvertible models. Ideally, a good method will provide a pooled prior in each example which represents a reasonable consensus between experts. Third, the selection and sensitivity of hyperparameters or weightings incorporated in a method must be evaluated. The hyperparameters or weightings should be intuitively meaningful and easy to set, their influence on the pooled prior should match expectations, and the pooled prior should not be highly sensitive to small changes in these parameters.

One tricky yet unavoidable issue is the level at which hyperparameter values should be set for the supra-Bayesian approaches in these comparisons. A priori, even before generating the examples contained in Chapters 5 and 6, we selected values for all hyperparameters according to our best intuition in light of the definition and purpose of each hyperparameter. For the Lindley-Winkler approach with one input and one output, we set $\alpha_1 = \alpha_2 = 0$, $\beta_1 = \beta_2 = 1$, $\gamma_1 = \gamma_2 = 1$, and $\rho = 0$. With these settings, the supra-Bayesian essentially adds no information to the two expert priors, as she considers each expert to be unbiased, uncorrelated, and a perfect assessor of his own precision. Settings for the Lindley-Winkler model applied to $\Re^2 \rightarrow \Re^2$ examples are merely multivariate extensions of these $\Re^1 \rightarrow \Re^1$ settings. For the group decision maker approach of French-Lindley, we set: $m_1 = m_2 = 2$, $\overline{m}_1 = \overline{m}_2 = -2$, $s_1 = s_2 = 1$, and $\rho = 0$. Unfortunately, the

French-Lindley approach does not have a convenient, meaningful set of default hyperparameters as in the Lindley-Winkler approach. With the settings chosen, the mean probability that an expert assigns to an event A which subsequently occurs is .88 with approximate 95% confidence interval (.50, .98), assuming normality in the log-odds. The mean probability that an expert assigns to an event A which subsequently does not occur is .12, with approximate 95% confidence interval (.02, .50). In addition, the two experts are assumed to be independent with respect to their information sources which contribute to their prior beliefs. For linear pooling and logarithmic poolings, both experts are assigned equal weights of $\frac{1}{2}$.

7.2 Introductory Example

Pooled priors from each of the four approaches applied to the introductory example of Section 4.2 are presented in Figure 7.1. The logarithmic pool is a reasonable compromise between the two experts—$p_{\log}(\phi)$ modifies $p_1^*(\phi)$ to account for the greater uncertainty and higher mode of $p_2(\phi)$. The linear pool $p_{\text{lin}}(\phi)$ is also a fairly reasonable compromise, with a larger right tail to reflect the beliefs of Expert 2. The consensus prior under the Lindley-Winkler approach seems to overweight Expert 1's opinion, in that $p_{\text{LW}}(\phi)$ closely follows $p_1^*(\phi)$ with an even smaller variance. One would expect a larger variance in $p(\phi)$ than $p_1^*(\phi)$, however, since knowing the opinion of Expert 2 adds uncertainty to the beliefs of Expert 1 instead of providing solid confirmation. Finally, the consensus prior under the French-Lindley approach is the most extreme, exhibiting a reasonable mode but an unreasonably small variance.

7.3 Cusp Example

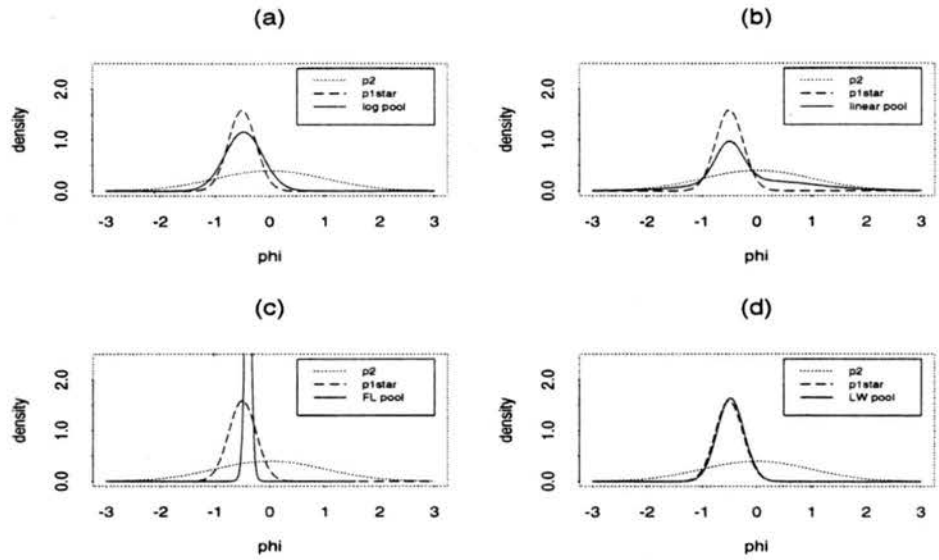

Figure 7.1: Introductory example: $p(\phi)$ under the (a) logarithmic pool, (b) linear pool, (c) French-Lindley pool, (d) Lindley-Winkler pool.

Pooled priors from each of the four approaches applied to the cusp example of Section 4.3 are presented in Figure 7.2. This example is difficult to assess because of the strong incoherence between the two experts, so what represents a reasonable consensus is quite subjective. $p_{\log}(\phi)$ is perhaps the most sensible of the four poolings. It provides balance between the two experts' assessments with its large concentration of mass below $\phi = 4$, where Expert 2 assigns nearly all of his probability, and its long right tail to reflect Expert 1's uncertainty. $p_{\text{lin}}(\phi)$ is maximized near $\phi = 0$ and minimized at $\phi = 12.5$, but the probability at the maximum point is only 2 or 3 times that at the minimum. Although a reasonable representation of a consensus, $p_{\text{lin}}(\phi)$ fails to recognize regions of high agreement between the experts. $p_{\text{LW}}(\phi)$ is centered near $\phi = 4$, a level of ϕ to which both experts assign a reasonable level of probability, but its normal shape is inadequate for capturing the agreement between these two experts. One may consider relaxing the Lindley-Winkler assumptions to allow non-normal distributions for the pooled prior to be explored in this example. Finally, $p_{\text{FL}}(\phi)$ is concentrated even more tightly around $\phi = 4$, ignoring the uncertainty and incoherence expressed by the experts.

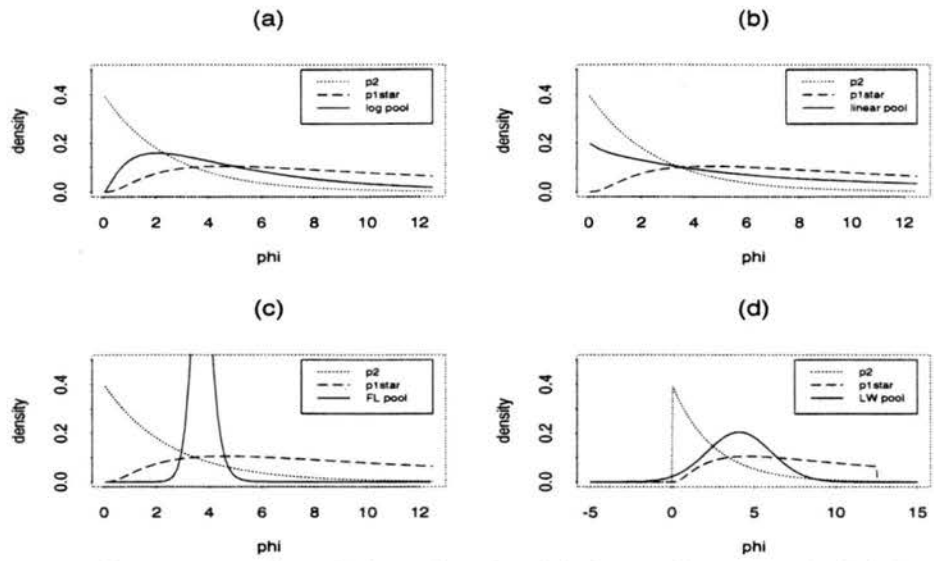

Figure 7.2: Cusp example: $p(\phi)$ under the (a) logarithmic pool, (b) linear pool, (c) French-Lindley pool, (d) Lindley-Winkler pool.

7.4 Linear $\Re^2 \to \Re^2$ Example

Pooled priors from each of the four approaches applied to the linear $\Re^2 \to \Re^2$ example of Section 4.4 are presented in Figure 7.3. The logarithmic pool and the Lindley-Winkler method give very similar pooled priors which are both reasonable consensuses of the two expert opinions. The modes of both $p_{\log}(\boldsymbol{\phi})$ and $p_{\mathrm{LW}}(\boldsymbol{\phi})$ lie between the modes of $p_2(\boldsymbol{\phi})$ and $p_1^*(\boldsymbol{\phi})$, shading toward $p_1^*(\boldsymbol{\phi})$ since Expert 1 has less uncertainty than Expert 2 with respect to beliefs expressed in $\boldsymbol{\phi}$-space. $p_{\mathrm{lin}}(\boldsymbol{\phi})$ is too similar to $p_1^*(\boldsymbol{\phi})$, overlapping Expert 1's prior for the most part and conceding the differing opinion of Expert 2 only through an extension toward $p_2(\boldsymbol{\phi})$ in the tail. Finally, $p_{\mathrm{FL}}(\boldsymbol{\phi})$ does a poor job of coherizing the two experts' beliefs, concentrating mass in regions of $\boldsymbol{\phi}$-space in which ϕ_2 is too large to be likely under $p_1^*(\boldsymbol{\phi})$ and ϕ_1 too large to be likely under $p_2(\boldsymbol{\phi})$.

7.5 Noninvertible $\Re^2 \to \Re^2$ Example

Pooled priors from each of the four approaches applied to the noninvertible $\Re^2 \to \Re^2$ example of Section 4.5 are presented in Figure 7.4. $p_{\log}(\boldsymbol{\phi})$ and $p_{\mathrm{lin}}(\boldsymbol{\phi})$

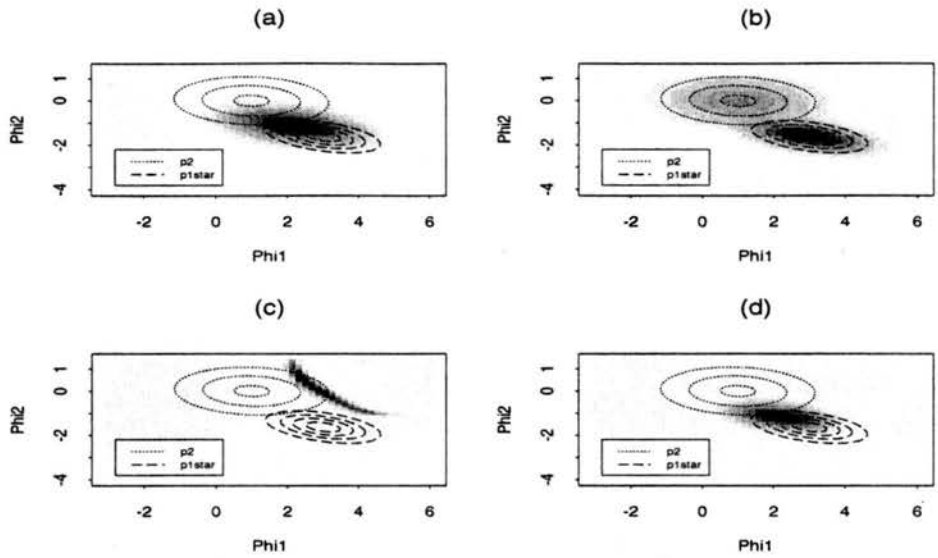

Figure 7.3: Linear $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) under the (a) logarithmic pool, (b) linear pool, (c) French-Lindley pool, (d) Lindley-Winkler pool.

have similar attributes; the largest densities are found near the intersection of the truncation line of $p_1^*(\phi)$ and $p_2(\phi)$, and further density contours fan out toward the maximum density of $p_2(\phi)$. As a result, these two approaches probably offer the most sensible poolings in this example. $p_{\mathrm{FL}}(\phi)$ is also somewhat similar to $p_{\log}(\phi)$ and $p_{\mathrm{lin}}(\phi)$, except that it lies above and to the right of the other two pooled densities, which makes little sense since in doing so it avoids those ϕ to which Expert 2 ascribes the highest probability. Finally, $p_{\mathrm{LW}}(\phi)$ basically mirrors $p_2(\phi)$, with the only effect of the opinion of Expert 1 being a small positive correlation. By constraining the pooled density to be multivariate normal, the Lindley-Winkler method also allows significant mass to be assigned to the region in ϕ-space which Expert 1 considers to have zero probability.

7.6 Whale Population Model

Pooled priors from each of the four approaches applied to the bowhead whale population model of Section 4.6 are presented in Figures 7.5. As with the cusp example, this example is difficult to assess because of the strong incoherence between the two experts when their beliefs are both expressed in ϕ-space. In Figure 7.5,

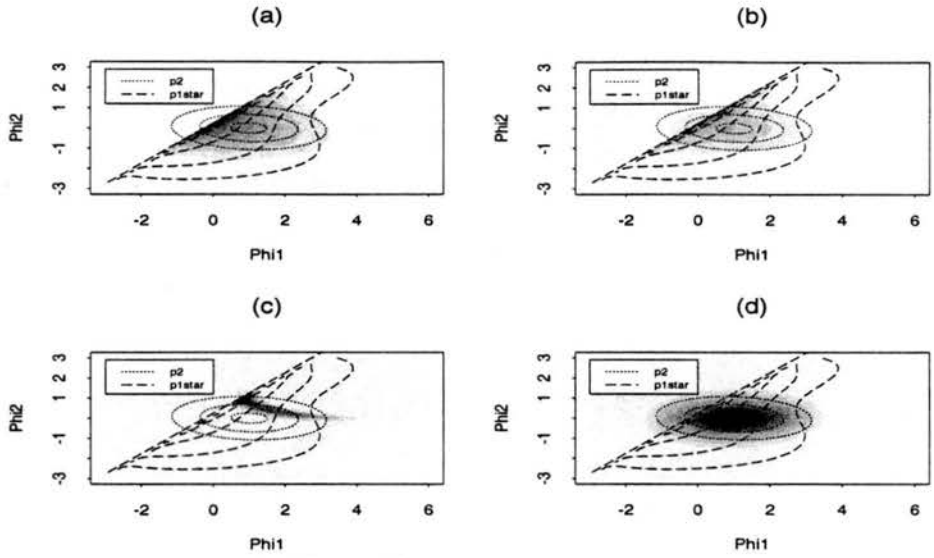

Figure 7.4: Noninvertible $\Re^2 \rightarrow \Re^2$ example: $p(\phi)$ (shaded density) under the (a) logarithmic pool, (b) linear pool, (c) French-Lindley pool, (d) Lindley-Winkler pool.

the contour plots representing $p_1^*(\phi)$ are not smooth; most of the mass of $p_1^*(\phi)$ is concentrated along a horizontal line at $\phi_2 \doteq 35$ running from $\phi_1 = 12,000$ to $\phi_1 = 25,000$. Regions of lower probability extend from this line, so that those points which intersect $p_2(\phi)$ have fairly low probability under $p_1^*(\phi)$. $p_{\log}(\phi)$ appears to best capture the regions of agreement between Experts 1 and 2, although it tends to overweight $p_2(\phi)$ slightly. $p_{\mathrm{LW}}(\phi)$ also does a fairly reasonable job of quantifying a set of pooled beliefs, although its normality restriction prevents it from fully capturing the uncertainty and shape of the consensus region. $p_{\mathrm{lin}}(\phi)$ and $p_{\mathrm{FL}}(\phi)$ suffer from the same flaws which plagued them in the linear $\Re^2 \rightarrow \Re^2$ example—$p_{\mathrm{lin}}(\phi)$ follows too closely to $p_1^*(\phi)$, recognizing $p_2(\phi)$ only with small tail probabilities, and $p_{\mathrm{FL}}(\phi)$ is located in regions where both experts assign negligible probability.

7.7 Summary: Recommendations on Pooling Methods

An examination of these five examples, which constitute just a small set of pooling problems which emerge in inferential situations involving deterministic

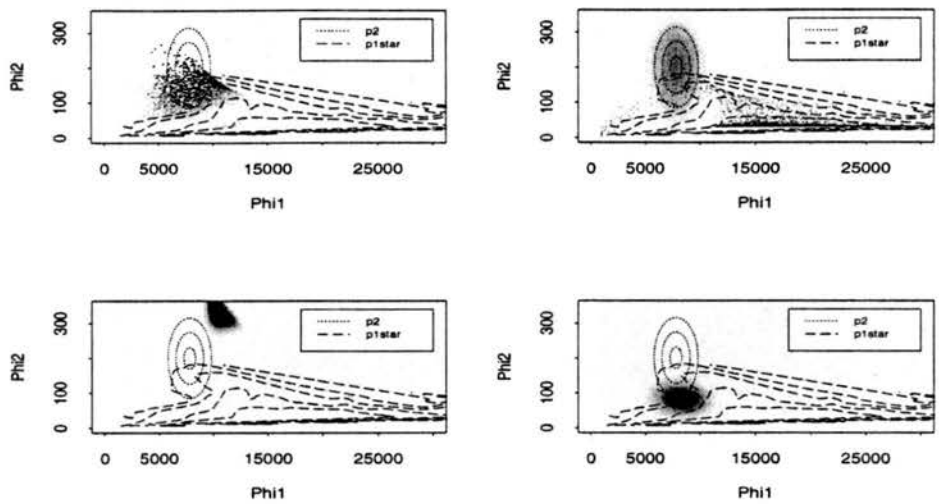

Figure 7.5: Whale population model: $p(\phi)$ (shaded density) under the (a) logarithmic pool, (b) linear pool, (c) French-Lindley pool, (d) Lindley-Winkler pool.

simulation models, the logarithmic pool seems to be the most consistently reasonable coherization of the two experts' beliefs. However, as described in Section 7.1, performance over a range of appropriate examples is just one factor in determining which method(s) are most suitable for pooling priors linked by deterministic simulation models. In the following paragraphs, we evaluate each of the four pooling methods in light of all important suitability factors. A summary of this discussion can be found in Table 7.1.

7.7.1 Logarithmic Pooling

Theoretical Properties. The logarithmic pool stands on strong theoretical ground. It is the unique pooling method which is externally Bayesian, so that the order of pooling and updating does not matter when forming Bayesian probabilities. It is the unique relative propensity consistent pooling method, ensuring that the pooled distribution must favor event A_2 over A_1 whenever all experts favor A_2 over A_1. It also possesses the strong version of the zero preservation property, so that $p(\phi)$ is 0 whenever any expert assign 0 probability to ϕ; this version of ZPP can be a bit restrictive.

	Log pool	Linear pool	FL pool	LW pool
Theoretical Properties:				
externally Bayesian	yes	no	yes[a]	no[b]
RPC	yes	yes	no	N/A[c]
ZPP (weak)	yes	yes	yes	no[d]
ZPP (strong)	yes	no	no[e]	no[d]
Performance in Examples:				
linear $\Re^1$	good	good	terrible	poor
cusp $\Re^1$	good	fair	terrible	fair[f]
linear $\Re^2$	good	fair	terrible	good
noninvertible $\Re^2$	good	fair	poor	poor
whale $\Re^2$	fair	fair	terrible	poor
Hyperparameters:				
number required	1	1	7	$m(m+6)$[g]
intuitive / easy to set	yes	yes	no	somewhat
robustness	yes[h]	yes[h]	no	no

[a] Obeys a variation of external Bayesianity defined by French (1985).

[b] However, Lindley (1985) argues that external Bayesianity is not a desirable property under the supra-Bayesian framework.

[c] Not applicable because Lindley-Winkler framework only requires first two moments from each expert.

[d] Conceptually, non-normal models using the Lindley-Winkler method can sometimes be found to obtain ZPP; however, deriving such a parametric form can be difficult or impossible in practice.

[e] No strong ZPP for the manner in which we apply the French-Lindley method to continuous parameters; however, strong ZPP does hold for individual events.

[f] Performance could be improved by using log-normal rather than normal model (see Section 6.4.1).

[g] Assumes m outcome variables.

[h] Examples not described in this dissertation suggest that this method is not overly sensitive to reasonable changes in the pooling hyperparameter.

Table 7.1: Comparison of 4 pooling methods.

Robustness. Over the range of our examples, the logarithmic pool consistently produced reasonable consensus distributions.

Hyperparameters. Only one parameter must be set when pooling priors according to the logarithmic pool—α such that $p(\phi) \propto p_2(\phi)^{\alpha} p_1^*(\phi)^{1-\alpha}$. α can be thought of as the relative levels of faith the decision maker has in the two experts, although how to choose a specific α based on data or informed opinions is not entirely clear. We have consistently set $\alpha = \frac{1}{2}$ for two reasons: (i) often there is no justifiable a priori reason to assume that one expert is superior to another, and (ii) results are invariant to relabeling for any M with $\alpha = \frac{1}{2}$. Nevertheless, the logarithmic pool is not overly sensitive to reasonable changes in α.

7.7.2 Linear Pooling

Theoretical Properties. The linear pool possesses some desirable theoretical properties but lacks others. It is not externally Bayesian, so that a Bayesian posterior distribution formed under linear pooling might be different depending on whether one pools expert opinions together and then updates the pooled distribution with a data-driven likelihood function, or whether one updates each expert's prior distribution individually and then pools the expert posteriors. However, the linear pool is relative propensity consistent, and it possesses the weak version of the zero preservation property, so that $p(\phi)$ is 0 whenever all experts assign 0 probability to ϕ.

Robustness. Over the range of our examples, the linear pool produced consensus distributions which were usually appealing. The biggest negative is that one expert can often exert undue influence on the linear pool, marked by long tails and locations driven almost exclusively by one expert. In addition, the linear pool can sometimes be bimodal, which reflects its tendency to average probabilities rather than zeroing in on regions of agreement.

Hyperparameters. Only one parameter must be set when pooling priors according to the logarithmic pool—α such that $p(\phi) \propto \alpha \cdot p_2(\phi) + (1-\alpha) \cdot p_1^*(\phi)$. α can be thought of as the relative levels of faith the decision maker has in the two experts, although how to choose a specific α based on data or informed opinions is not entirely clear. As in logarithmic pooling, we consistently set $\alpha = \frac{1}{2}$ to reflect an a priori belief of equal experts even though the linear pool is not overly sensitive to reasonable changes in α.

7.7.3 French-Lindley supra-Bayesian Pooling

Theoretical Properties. The French-Lindley method follows the weak form of the zero preservation property. On the other hand, a pooled prior formed under the French-Lindley approach does not possess external Bayesianity. However, Lindley (1985) argues that, under the supra-Bayesian philosophy of treating expert priors as data for the decision maker, external Bayesianity is not necessarily desirable. In addition, French (1985) shows that the French-Lindley pool adheres to a variation of external Bayesianity assuming that the decision maker knows the new data which the experts have observed. The French-Lindley pool's biggest theoretical shortcoming is the failure of RPC, which can arise depending on the hyperparameter settings. These failures have been described in Section 5.4.

Robustness. Consensus distributions from the French-Lindley pool were, in general, very unsatisfying. In $\Re^1 \to \Re^1$ examples, for instance, the pooled priors had too little variability, while in the $\Re^1 \to \Re^1$ examples the pooled priors were often concentrated in regions favored by neither expert.

Hyperparameters. The hyperparameters required to evaluate the experts' bias, correlation, and precision assessments are difficult to use. Their interpretations are not intuitive, seemingly reasonable settings are capable of producing unreasonable poolings, and the pooled distribution is extremely sensitive to changes

in the hyperparameters. For instance, some could argue that we happened to select a poor set of default hyperparameter values for the examples in this chapter, and that a wiser choice would have presented the approach in a better light. Although further experimentation shows that more reasonable pooled densities can emerge from the French-Lindley approach in some cases (like the introductory example), in other cases (like the linear $\Re^2 \rightarrow \Re^2$ example) we were unable to form any reasonable pooling. Furthermore, it is inappropriate to experiment post-hoc until a pooled density seems satisfactory with real examples. Thus, the French-Lindley method seems ill-suited for the multi-dimensional, multi-event pooling required for use with deterministic simulation models.

7.7.4 Lindley-Winkler supra-Bayesian Pooling

Theoretical Properties. As with the French-Lindley supra-Bayesian approach, the Lindley-Winkler pool does not possess external Bayesianity, although Lindley (1985) argues that external Bayesianity is not a desirable property for supra-Bayesian methods. For the Lindley-Winkler approach, RPC is not necessarily applicable, since experts are not required to offer event probabilities or entire distributions expressing their beliefs—they must only summarize their beliefs via a mean and a standard deviation. The biggest theoretical shortcoming of the Lindley-Winkler approach is its failure to adhere to the zero preservation property in its weak form, although this is not a fatal flaw. Because of the normality assumption, the Lindley-Winkler pool can ascribe significant probability in regions where no expert has assigned any positive probability.

Robustness. Where the normality assumption is appropriate, the Lindley-Winkler pool performs well, but it other cases it performs rather poorly. In earlier experiments with the cusp example of Section 4.3 we relaxed the normality assumption to allow the pooled prior to be distributed log-normal; this seemed to

produce a better consensus distribution, but at the expense of parameter interpretability and generalizability.

Hyperparameters. The Lindley-Winkler hyperparameters are much more intuitively meaningful than their French-Lindley counterparts, and they appear to have the expected effect on the pooled prior. However, choosing a sound set of hyperparameter values is still a monumental task for the decision maker because of the sheer number of them and the vast knowledge base required to evaluate all experts. In addition, the pooled prior is very sensitive to changes in the hyperparameters, placing further stress on the already tenuous task of assigning appropriate values to the hyperparameters.

7.7.5 Conclusion.

After examining all factors, we recommend logarithmic pooling as the most suitable method for pooling prior distributions linked by a deterministic model. In obtaining log pooled prior distributions, we recommend the adaptive version of the algorithm described in Chapter 3, especially in many inference problems involving deterministic simulation models when one is limited to few iterations and/or small numbers of model simulations.

Chapter 8

An Application of Log Pooled Inference Methods to an AIDS Transmission Model

8.1 Hethcote's Model: Background

AIDS (Acquired Immunodeficiency Syndrome) is caused by a retrovirus, HIV-1, which infects T-helper cells. This virus eventually causes T-helper cell counts to fall, which allows opportunistic infections that are signs of the immune deficiency to appear. The virus can also invade the brain, giving sign of dementia. HIV is transmitted by sexual intercourse, by transfusion of infected blood, and by sharing needles during drug use. The first known American case of AIDS was a male homosexual living in Los Angeles whose symptoms were diagnosed in March, 1978, but some evidence points to unrecognized cases as far back as the 1960s (Pickering *et al.*, 1986; Jacquez *et al.*, 1988). The early cases in the United States were primarily homosexual men and IV drug users, but it spread quickly beyond those two groups. As of late 1996, the United States has had the highest number of reported AIDS cases in the world, with over 581,429 cases and 326,000 deaths (Centers for Disease Control, 1997).

Early in the epidemic, statisticians and mathematical modelers began collaborating with epidemiologists, medical researchers, and public health officials to gain understanding and predict the course of the epidemic. The modeling efforts have been reviewed by Isham (1988) and Foulkes (1998). Models have been useful in understanding the progression of the HIV infection, in recognizing social

and biological variables which can affect transmission rates, in forecasting future AIDS cases, and in evaluating clinical assumptions or providing guidance on policy choices.

In particular, Hethcote *et al.* (1991b; 1991a) developed an epidemiological simulation model for HIV transmission. The Hethcote model requires parameterization of the population structure, interaction patterns, transmission dynamics, and disease progression in order to describe the system by which AIDS spreads. The model of Hethcote *et al.* describes the transmission of AIDS in the population of homosexual and bisexual males in San Francisco. Because blood samples were saved from participants in a hepatitis B vaccine trial starting in 1978, San Francisco is the only place where pre-1984 data on HIV incidence are available (Hethcote *et al.*, 1991a), so data from San Francisco has been used in several AIDS modeling efforts. In addition, San Francisco has proven to be a good site for behavioral surveys because of the large open community of homosexual men there. Models developed using the San Francisco data can often be generalized to other populations.

The progression from HIV infection to AIDS and death can be described in several stages. Longini *et al.* (1989; 1990) modeled the natural history of HIV as a 5-stage process: 1) infected but antibody negative, 2) antibody positive but asymptomatic, 3) symptoms begin developing, especially abnormal hematological indicators, 4) clinical AIDS, and 5) death. The course of infection and progression to AIDS is highly variable. Longini has estimated median waiting times of 1.5 months before antibodies are developed, 36.5 months before symptoms appear, 43.6 months before the T-helper cell count drops enough that the opportunistic infections that signal clinical AIDS begin, and 16.3 months until eventual death due to AIDS. Identification of these stages is important in model development because the infectiousness and the sexual activity patterns of infected persons

can vary considerably in the different stages. In addition to the stages of HIV progression, two assumptions are made which further differentiate AIDS models from typical models of transmission for infectious diseases. First, the latent period (the time from when a person is infected with HIV to when he can pass on the disease to others) is assumed to be short enough to be ignored in a model. Second, HIV infectivity is assumed to continue for life, since no evidence exists that the virus ever leaves the host, even if it never develops into clinical AIDS.

Hethcote's model is compartmental, where the compartments are defined by not only the stages of progression from HIV infection to AIDS, but also by the sexual activity of members of the study population. The population is divided into men who have many different sexual partners (very active), those who have only a few different partners (active), and a third category for those who emigrate. The flow of subjects among the compartments is governed by many parameters. These input parameters can be divided into three groups—those related to population size, those related to stages of infection, and those related to sexual activity levels. The population category contains parameters such as

- total population size,
- fraction of population who are sexually very active,
- ratio of partnership rates for very active and active men,
- natural mortality rate, and
- transfer rate between very active and active states.

The stages of infection category contains parameters such as

- number of infectious stages,
- rate of progression from stage k to stage $k+1$,

- relative infectivity of stage k men compared with asymptomatic men, and
- relative sexual activity of stage k men compared with asymptomatic men.

The sexual activity category contains parameters such as

- starting date of the epidemic,
- probability of transmission to partners by infected asymptomatic men,
- average number of partners per month at the start,
- starting date for reduction in average number of partners per month (as awareness about AIDS increased),
- stopping date for reductions,
- yearly reduction factor, and
- fraction of new partnerships distributed by proportionate mixing.

8.1.1 Our Version of Hethcote's Model

For our purposes, we consider a version of Hethcote's model with 6 input variables and 6 output variables, fixing all other variables at levels suggested by Hethcote. Specifically, we consider the following 6 input variables:

1. θ_1 (PPM) = average number of partners per month before an overall decrease in sexual activity was observed (i.e. prior to August 1981).
2. θ_2 (RDN) = yearly reduction factor in number of partners per month between August, 1981, and December, 1986. Sexual activity as measured by PPM is assumed to have been constant until August, 1981, when awareness of the seriousness of the AIDS epidemic and its transmission methods became more widespread. Sexual activity decreased geometrically from then at rate RDN until December, 1986, at which point it leveled off.

3. θ_3 (EMF) = external mixing fraction, or the proportion of sexual partnerships which were governed by proportionate mixing between active and very active groups. 1-EMF, then, represents the percentage of contacts which were strictly internal to one's sexual activity group.

4. θ_4 (MMP) = monthly migration percentage.

5. θ_5 (PT) = probability of transmission when an infected individual has sexual intercourse with a non-infected (susceptible) individual.

6. θ_6 (PHM) = total size of population of homosexual males in San Francisco. The population is assumed to remain constant at this level throughout the epidemic.

The 6 output variables which we consider are:

1. ϕ_1 (HIVINC) = 1995 HIV incidence, where incidence refers to the number of new HIV infections occurring among the homosexual male population in San Francisco in 1995.

2. ϕ_2 (AIDINC) = 1995 AIDS incidence—the number of HIV-infected individuals who developed clinical AIDS in 1995.

3. ϕ_3 (AIDDTH) = 1995 deaths due to AIDS.

4. ϕ_4 (STG2VA) = prevalence of very active HIV-infected men who were in Stage 2 at the end of 1995. Prevalence refers to the total number of very active men in Stage 2, whether or not they entered Stage 2 in 1995 or earlier. Recall that Stage 2 is the point at which an individual is antibody positive but not showing symptoms.

5. ϕ_5 (STG2A) = prevalence of active HIV-infected men in Stage 2 at the end of 1995.

6. ϕ_6 (STG2EM) = prevalence of emigrated HIV-infected men in Stage 2 at the end of 1995.

The input variables were chosen because a significant amount of uncertainty was associated with their values or because they exhibited high sensitivity to HIV and AIDS incidence estimates. In addition, obtaining good estimates of these 6 inputs would be very important to a physician, medical researcher, epidemiologist, or public health official. Knowing the distribution of partners per month (PPM) would help health officials decide if a behavioral change in sexual activity, as measured by partners per month, had potential to reduce the spread of AIDS. The reduction rate (RDN) provides a measure of how successful AIDS awareness campaigns were between 1981 and 1986. The external mixing fraction (EMF) would interest an epidemiologist attempting to describe the sexual activity patterns of the homosexual male population through mixing models. The monthly migration percentage (MMP) helps an epidemiologist estimate how quickly an epidemic like AIDS can spread from one core population. The probability of transmission (PT) is vital in forecasting the rate of spread of the epidemic and in understanding the risks associated with unprotected sexual intercourse with a potentially infected individual. Finally, the population size (PHM) is important when one wishes to generalize to populations other than homosexual males in San Francisco.

The 6 output variables were also chosen because of their importance to personnel from the health field as well as to the general public. At the time Hethcote's model was developed (1990), predictions to 1995—5 years out—would have been very interesting to persons concerned about the continuing impact of the AIDS epidemic. In particular, health officials were interested in estimates of new cases of HIV infection (HIVINC), cases of clinical AIDS (AIDINC), and deaths due to AIDS (AIDDTH) in 1995 to forecast resource needs. In addition, recall that Stage 2 in the progression from HIV infection to AIDS includes patients who are

antibody positive but asymptomatic. Since they are infected but not displaying symptoms, these men in Stage 2 are most likely to spread AIDS to their partners, but because they are antibody positive, their infectivity could be diagnosed at this stage and their sexual activity discouraged. Thus, estimates of the number of men in Stage 2 represent those at risk of spreading the disease who could potentially be diagnosed through an aggressive awareness and testing program. We stratify the estimate of Stage 2 men into three compartments—those sexually very active (STG2VA), those sexually active (STG2A), and those who have emigrated (STG2EM).

8.2 Prior Distributions for the Bayesian Model

In each case above, for both inputs and outputs, it is helpful to express knowledge about a particular parameter in terms of an entire probability distribution rather than just a point estimate. Consider an example. A health official may allocate resources for AIDS medical care and counseling programs based on a point estimate for the HIV incidence level in 1995. However, depending on how accurate this estimate is, the AIDS programs may be vastly under-resourced, or vital resources may be wasted. With a single, imperfect prediction guiding policy, in the best case officials will find themselves scrambling to add or redeploy resources as the time approaches. If an estimate of the HIV incidence level was instead given as a complete probability distribution, then officials could formulate a plan which covers the range of likely scenarios for HIV incidence. For instance, they could base plans on a "worst case scenario"—say the 95th percentile of the probability distribution for HIV incidence—and scale back if it looks like the worst case will not be realized.

Hethcote used numerous sources to estimate values of input and output parameters. He realized that the information he gathered was subject to uncertainty,

so he performed sensitivity analyses to investigate the impact of this uncertainty. In these sensitivity analyses, Hethcote held all variables except one constant, allowed the test variable to vary within its range of possible values, and checked the impact those changes had on output values. Our Bayesian inferential approach outlined in this dissertation provides a more systematic and formal approach to achieving the goals of Hethcote's sensitivity analysis. Our method requires experts to provide prior probability distributions which express their beliefs about realistic values for model parameters which lie in their areas of expertise. Then, these expert priors are all transformed into the same parameter space, where they can be pooled into a single prior for that parameter. The benefits of this Bayesian pooling approach for a simulation model such as Hethcote's AIDS transmission model are twofold. First, better estimates of important outputs (like 1995 AIDS incidence) can be obtained—estimates which combine the expertise of all experts from whom priors were solicited and which provide an entire probability distribution for likely values of the output. Second, we can check if values for input parameters seem to be reasonable in light of the opinions of other experts and the model. If, for instance, the prior from the expert on probability transmission (PT) does not agree with the implicit prior about PT from the output experts, then the model or the individual priors may need to be revisited.

Hethcote's goals in creating a simulation model for AIDS transmission were: to organize and coalesce many sources of data, to obtain good parameter estimates to better explain how the AIDS epidemic is spread, to investigate the sensitivity of the model by evaluating the effect of uncertainties in the input parameters, and to compare outputs of the modeling process to independent assessments of output parameters. All these goals can be addressed using our Bayesian analysis framework for making inferences with deterministic simulation models. Thus, we set out to apply our adaptive importance sampling algorithm for obtaining

log pooled priors to Hethcote's AIDS transmission model. The first step is to define priors for the 6 input variables and the 6 output variables in our version of Hethcote's model. For some parameters sources were plentiful, and for others there was less information available. Greater details on the selection of prior distributions is contained in the next two sections.

8.2.1 Input Variables

Our prior distribution $p_1(\boldsymbol{\theta})$ for the input parameters will be a multivariate normal distribution $N_6(\underline{\mu}, \Sigma)$, where the mean vector $\underline{\mu}$, the standard deviation vector $\underline{\sigma}$, and the correlation matrix R are all based on expert opinions from the AIDS literature, and $\Sigma = \text{Diag}(\underline{\sigma})\text{RDiag}(\underline{\sigma})$, where $\text{Diag}(\underline{\sigma})$ is a square matrix with $\underline{\sigma}$-terms along the diagonal and 0's elsewhere.

Information on average number of partners per month (PPM) comes from several sources. Research and Decisions Corporation, in a report prepared for the San Francisco AIDS Foundation (1984), found that homosexual males in San Francisco had sex with an average of 2.6 different men over the past 30 days (N=492). In the San Francisco City Clinic Cohort (SFCCC), a group of homosexual men participating in a hepatitis B study between 1978 and 1980, interviews showed that the mean number of non-steady partners during a 4-month period in 1978 was 29.3 per person (median 16) (Centers for Disease Control, 1987), although there is evidence that these men were more sexually active than most homosexual men in San Francisco. Participants in an SFCCC vaccine trial reported PPM=1.4 before 1981, and those in the San Francisco Men's Health Study (SFMHS) reported PPM=1.7 before 1984 (Hethcote *et al.*, 1991b). Finally, Pickering (1986) cites a poll by McKusick which places the average number of sex partners per month at 5.9 in November, 1982. Taking all 5 of these studies into consideration,

we decided that our input prior distribution $p_1(\boldsymbol{\theta})$ should reflect a mean of 2.5 and a standard deviation of 1 for θ_1, with a lower bound of 0.

Information on yearly reduction rate (RDN) is also plentiful. In the sample from the SFCCC hepatitis B study, median partners per 4-month period decreased from 16 to 1 over a 7-year period from 1978 to 1985, producing a yearly reduction factor of $(1/16)^{1/7} = 0.67$ (Centers for Disease Control, 1987). In these studies, reduction was also expressed in terms of a decreased risk index for anal intercourse with ejaculation with nonsteady partners and insertive anal intercourse with nonsteady partners; these two measures produced estimates for RDN of 0.59 and 0.67, respectively. An additional study about HIV was performed on SFCCC participants who tested negative for hepatitis B. The yearly reduction factor for these men between 1981 and 1987 was 0.43 (Hethcote *et al.*, 1991b). Finally, the poll by McKusick (1986) shows that the average number of sex partners decreased from 5.9 per month in 1982 to 2.5 in 1984, giving an estimated RDN of $(2.5/5.9)^{1/2} = 0.65$. Incorporating all of these studies, we set the prior mean for θ_2 at 0.61, with a standard deviation of 0.11 and a lower bound of 0 (note that an unlikely $\theta_2 > 1$ indicates that an increase in PPM actually occurred between 1978 and 1985).

Jacquez (1988) illustrates how HIV transmission can be effectively modeled using a population with two subgroups (sexually very active and sexually active) and an assumption of preferred mixing. Under preferred mixing, a fraction (1-EMF) of each group's contacts occur strictly internally, while the rest are subject to proportional mixing assumptions. EMF is estimated at .82 according to a study by Hethcote and Yorke (1984) on gonorrhea transmission in a similar population. The prior standard deviation for θ_3 is set at .15 to allow a 95% lower bound near .5. Absolute lower and upper bounds are set at 0 and 1.

The aforementioned study by Research and Decisions Corporation (1984) found that 26% of homosexual males surveyed had lived in San Francisco less than 5 years (N=500). Based on this, a constant immigration rate of 5% was chosen. Furthermore, since the immigration and emigration rates must be equivalent if the assumption of constant total population is to hold, (5/12)% can be used to estimate the monthly emigration rate (MMP). A standard deviation of $\frac{1}{12}\sqrt{\frac{(.50)(.50)}{500}} = .002$ can be used in the prior distribution for θ_4, along with lower and upper bounds of 0 and 1.

Grant *et al.* (1987) found that the mean probability of infection through unprotected receptive anal intercourse was .102 per sex partner for 672 men studied in the SFMHS. He also reports a 95% confidence interval for his estimate of (.043,.160). Since the average relative infectivity (compared to asymptomatic men) of the infected persons contacted by the men in the SFMHS is assumed to be 2, Grant's estimate for PT, the probability of transmission of HIV to sex partners of active asymptomatic infected persons, would be .05 with 95% confidence interval (.021,.080). As a result, the prior distribution should reflect a mean value for θ_5 of .05 with standard deviation of .015 and lower and upper bounds of 0 and 1.

Two reliable estimates for the population of homosexual males in San Francisco (PHM) were cited by Lemp (1990). First, Research and Decisions Corporation provided an estimate of 69,122 openly gay or bisexual males from their telephone surveys (N=500) (1984). This total was based on an estimate that 24% of the 290,377 men over 15 are gay. With an approximate 95% confidence interval for the percentage of gay men of (.20,.28), this translates into a 95% confidence interval of (58,000,81,300) for PHM. Second, an estimate for PHM of 42,509 was obtained by extrapolating the ratio of AIDS cases in the SFMHS to the entire city. Lemp averaged these two estimates together to get 56,000, a figure that the

San Francisco Department of Public Health relies on when projecting AIDS morbidity and mortality. Thus, $p_1(\boldsymbol{\theta})$ should reflect a mean of 56,000 and a standard deviation of 12,000 for θ_6, with a lower bound of 0.

Finally, correlations among the 6 input variables must be estimated. Relatively high correlations (.5–.7) were assumed among $\theta_1 - \theta_3$. A greater starting average number of partners per month (θ_1), for example, provides greater opportunity for high reduction rates (θ_2), and a higher external mixing fraction (θ_3) means that men in the population have a larger network of sexual contacts, leading to more partners and more potential for reduction. The only other strongly positive correlation is between partners per month (θ_1) and total population (θ_6). Certain correlations were zero or near-zero, especially those involving probability of transmission (θ_5), since PT is biologically determined and not affected by population dynamics.

In conclusion, our prior distribution $p_1(\boldsymbol{\theta})$ for the input parameters, based on expert opinions from the AIDS literature, is a multivariate normal distribution with mean vector

$$\underline{\mu} = [2.5, .61, .82, .05/12, .05, 56000]',$$

corresponding standard deviation vector

$$\underline{\sigma} = [1.0, .11, .15, .002, .015, 12000]',$$

and correlation matrix

$$R = \begin{bmatrix} 1 & .5 & .7 & .3 & 0 & .6 \\ .5 & 1 & .5 & .1 & .2 & .1 \\ .7 & .5 & 1 & .3 & 0 & -.1 \\ .3 & .1 & .3 & 1 & 0 & .3 \\ 0 & .2 & 0 & 0 & 1 & 0 \\ .6 & .1 & -.1 & .3 & 0 & 1 \end{bmatrix}.$$

8.2.2 Output Variables

Our prior distribution $p_2(\phi)$ for the output parameters will be a multivariate normal distribution $N_6(\underline{\mu}, \Sigma)$, where the mean vector $\underline{\mu}$, the standard deviation vector $\underline{\sigma}$, and the correlation matrix R are all based on expert opinions from the AIDS literature, and $\Sigma = \text{Diag}(\underline{\sigma})\text{RDiag}(\underline{\sigma})$, where $\text{Diag}(\underline{\sigma})$ is a square matrix with $\underline{\sigma}$-terms along the diagonal and 0's elsewhere.

Whereas information for the input priors was collected mainly from epidemiological and biological sources, information for the output priors comes mainly from survey sampling studies. The Centers for Disease Control (1995) provide yearly data by metropolitan area on the number of AIDS cases diagnosed and the number of deaths due to AIDS. This data can be broken down even further by exposure risk group (e.g. homosexual and bisexual males). In 1995, the estimated AIDS incidence (AIDINC) among homosexual males in San Francisco was 1387, and the corresponding estimated number of deaths due to AIDS (AIDDTH) was 1323. Thus, the prior distribution, $p_2(\phi)$, for the outputs should reflect means of 1387 and 1323 for ϕ_2 and ϕ_3 respectively. A standard deviation of 300 for each variable was selected to represent the potential inaccuracies and biases in the CDC counts, stemming from sources such as underreporting, incorrect diagnoses, and incorrect classification into exposure risk groups.

To obtain priors for ϕ_1, ϕ_4, ϕ_5, and ϕ_6 (HIVINC, STG2VA, STG2A, and STG2EM), we rely on estimated HIV incidences and prevalences from two reports prepared by the San Francisco Department of Public Health. The first (San Francisco Department of Public Health AIDS Office, 1993) provides estimates of the prevalence of HIV infection in San Francisco in 1993. Estimates are based on "... numerous studies conducted in San Francisco in recent years, including probability samples of San Francisco neighborhoods, cohort studies of gay and bisexual men, random digit-dial telephone surveys, clinic-based unlinked serosurveys, studies of young men who have sex with men, surveys of childbearing women, studies of het-

erosexual partners, street-based studies of injection and non-injection drug users, and screening of blood donors." (1993, p. 3). The second (San Francisco Department of Public Health AIDS Office, 1997) describes estimates of the prevalence and incidence of HIV infection in San Francisco in 1997 based on expert opinions of researchers, epidemiologists, and service providers attending the 1997 HIV Consensus Meeting. At the meeting, recent empirical data on HIV prevalence and incidence in San Francisco was presented and discussed, and the opinions of the participants were sythesized using a modified Delphi process, which is a feedback and reassessment technique under which experts reconsider their assessments after being presented with summary statistics regarding the assessments of all experts.

In 1993, HIV prevalence was estimated at 24,978, or 43.1% of the estimated MSM (men who have sex with men) population of 58,000. In 1997, HIV prevalence was estimated at 13,135, or 30.5% of the estimated MSM population of 43,100. No HIV incidence was estimated in the 1993 report, but the panel of experts in 1997 estimated HIV incidence for that year at 336, or 1.1% of the 29,965 susceptible MSMs. The lower level of HIV prevalence in 1997 compared to 1993 was attributed to more accurate information about at-risk populations and decreased rates of new infections. Using linear interpolation, we can estimate the 1995 HIV prevalence at 19,056; this will be used later to construct estimates of the Stage 2 counts. To estimate the 1995 HIV incidence (HIVINC), an estimate for the 1993 HIV incidence must first be obtained. A total of 33,022 susceptible MSMs was estimated in 1993. Furthermore, an HIV incidence rate of 1.5% in 1993 can be assumed—the rate of new infections is known to have decreased during the 1990s, and this represents the high end of the 1997 range. As a result, we have an estimate for 1993 HIV incidence of 495, and linear interpolation gives an estimate of 415 for the 1995 HIV incidence (ϕ_1). A range of 0.5% to 1.5% in

the 1997 estimates for HIV incidence rate suggests that the standard deviation for ϕ_1 might be set at 200.

Prior estimates for stratified Stage 2 prevalances (STG2VA, STG2A, STG2EM) can be derived from the estimated 1995 HIV prevalence of 19,056. Longini's (1990) transition rate estimates describing his 5 stages in the progression of AIDS suggest that approximately 40% of patients infected with HIV in 1995 would be in Stage 2 (asymptomatic). Note that 40% of 19,056 is 7622 men in Stage 2, some of which are classified as very sexually active and others as active. Although 10% of the total population is very active, we assume that 20% of the Stage 2 men in 1995 are very active because of the higher prevalance of HIV among the very active subgroup. This suggests that ϕ_4 = STG2VA can be estimated by 1524 and ϕ_5 = STG2A by 6098. ϕ_6 = STG2EM can be estimated at 381—the product of the assumed emigration rate (5%) and the estimated total Stage 2 prevalence (7622). A range of 24% to 38% in the 1997 estimates for HIV prevalence rate suggests that the standard deviations for ϕ_4, ϕ_5, and ϕ_6 might be set at 400, 1000, and 300, respectively.

Finally, correlations among the 6 output variables must be estimated. In general, enough of a time lag exists between HIV infection (ϕ_1) and both development of clinical AIDS (ϕ_2) and death due to AIDS (ϕ_3) that ϕ_1 should not be highly correlated with either variable. However, the time lag between ϕ_2 and ϕ_3 is short enough that these variables should exhibit a strong positive correlation. Other strong positive correlations ($\rho > .5$) are likely to occur among the variables ϕ_4, ϕ_5, and ϕ_6, since as overall HIV prevlance increases, each of these counts of infected men in Stage 2 will also increase. Smaller positive correlations are likely between an incidence ($\phi_1 - \phi_3$) and a prevalence ($\phi_4 - \phi_6$).

In conclusion, our prior distribution $p_2(\boldsymbol{\phi})$ for the output parameters, based on actual counts and survey sampling results, is a multivariate normal distribution

with mean vector

$$\underline{\mu} = [415, 1387, 1323, 1524, 6098, 381]',$$

corresponding standard deviation vector

$$\underline{\sigma} = [200, 300, 300, 400, 1000, 300]',$$

and correlation matrix

$$R = \begin{bmatrix} 1 & .2 & .2 & .3 & .4 & .3 \\ .2 & 1 & .6 & .2 & .2 & .1 \\ .2 & .6 & 1 & .1 & .1 & 0 \\ .3 & .2 & .1 & 1 & .7 & .5 \\ .4 & .2 & .1 & .7 & 1 & .5 \\ .3 & .1 & 0 & .5 & .5 & 1 \end{bmatrix}.$$

8.3 Generalization of Our Log Pooling Algorithm for Higher Dimensional Models

For the most part, the generalization of the adaptive importance sampling algorithm for log pooling given in Chapter 3 follows naturally. Bivariate densities become multivariate densities in 6 dimensions, two concurrent marginal tests for convergence become six concurrent marginal tests, and a 2-by-2 covariance matrix governing the selection of new points for the envelope becomes a 6-by-6 covariance matrix. Certain aspects of the algorithm do not change at all, such as importance weights and the method for estimating densities with noninvertible $\boldsymbol{M}$. However, a few aspects of the algorithm are different; these differences are described in the remainder of this section.

First, this example illustrated the importance of performing calculations within the algorithm using standardized values (centered and scaled) of variables in any multivariate situation. Problems with unstandardized variables are exaggerated in input space, where the population PHM has scale 10,000 times larger than any other input variable. Specific problems when large differences in scale

exist include misleading calculations of nearest neighbors and biased estimates of densities.

Second, mixture weights $b_{l,r}$ (see Section 3.2.3) for the envelope sampling distribution are now based on the differences between the six marginal distributions $\widehat{F_{p_2,i}}$ and $F_{p_2,i}$, $i = 1, \ldots, 6$. Although mixture weights could still theoretically be based on $\left|\widehat{F_{p_2}} - F_{p_2}\right|$, where both $\widehat{F_{p_2}}$ and F_{p_2} are 6-dimensional cdfs, the computing time required to evaluate the 6-dimensional cdfs is prohibitive. Therefore, for this 6-dimensional example, we set

$$b_{l,r} = \sqrt{\left(\hat{F}_{p_2,1}(\phi_{l,r}) - F_{p_2,1}(\phi_{l,r})\right)^2 + \ldots + \left(\hat{F}_{p_2,6}(\phi_{l,r}) - F_{p_2,6}(\phi_{l,r})\right)^2}$$

where $b_{l,r}$ is the mixture weight for the r^{th} point generated after the l^{th} iteration. Using the adaptive method with these mixture weights should lead to a more precise and efficient estimate of the log pooled prior than the nonadaptive method.

Third, Jacobian estimation proceeds analogously to the case with 2 inputs and 2 outputs (see Section 3.3), but because of the extra complexities introduced by higher dimensions, it is worth specifying a few details. As before, let $|\boldsymbol{J}(\boldsymbol{\theta}_i)|$ denote the Jacobian of $\boldsymbol{\phi}$ with respect to $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}_i$. $|\boldsymbol{J}(\boldsymbol{\theta}_i)|$ is merely the determinant of the matrix J of estimates of partial derivatives, where

$$J = \begin{bmatrix} j_{11} & j_{12} & j_{13} & j_{14} & j_{15} & j_{16} \\ j_{21} & j_{22} & j_{23} & j_{24} & j_{25} & j_{26} \\ j_{31} & j_{32} & j_{33} & j_{34} & j_{35} & j_{36} \\ j_{41} & j_{42} & j_{43} & j_{44} & j_{45} & j_{46} \\ j_{51} & j_{52} & j_{53} & j_{54} & j_{55} & j_{56} \\ j_{61} & j_{62} & j_{63} & j_{64} & j_{65} & j_{66} \end{bmatrix},$$

$j_{kl} = (\phi_{\Delta l,k} - \phi_k)/\epsilon_l$ and $\phi_{\Delta l,k} = \boldsymbol{M}_k(\theta_1, \ldots, \theta_l + \epsilon_l, \ldots, \theta_6)$.

$|\boldsymbol{J}(\boldsymbol{\phi}_i)|$ is the Jacobian of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\phi}$, evaluated at $\boldsymbol{\phi}_i$. As with 2 inputs and 2 outputs, directional derivatives can be used to estimate $|\boldsymbol{J}(\boldsymbol{\phi}_i)|$ in the case with 6 inputs and 6 outputs. Estimating the 36 partial derivatives in the Jacobian matrix requires solving 6 sets of 6 equations and 6 unknowns involving

appropriate directional derivatives. The Jacobian can be constructed one row at a time; each row features the solutions $\underline{x}$ to the linear system $A\underline{x} = \underline{b}$, where

$$A = \begin{bmatrix} \phi_{\Delta 1,1} - \phi_1 & \phi_{\Delta 1,2} - \phi_2 & \phi_{\Delta 1,3} - \phi_3 & \phi_{\Delta 1,4} - \phi_4 & \phi_{\Delta 1,5} - \phi_5 & \phi_{\Delta 1,6} - \phi_6 \\ \phi_{\Delta 2,1} - \phi_1 & \phi_{\Delta 2,2} - \phi_2 & \phi_{\Delta 2,3} - \phi_3 & \phi_{\Delta 2,4} - \phi_4 & \phi_{\Delta 2,5} - \phi_5 & \phi_{\Delta 2,6} - \phi_6 \\ \phi_{\Delta 3,1} - \phi_1 & \phi_{\Delta 3,2} - \phi_2 & \phi_{\Delta 3,3} - \phi_3 & \phi_{\Delta 3,4} - \phi_4 & \phi_{\Delta 3,5} - \phi_5 & \phi_{\Delta 3,6} - \phi_6 \\ \phi_{\Delta 4,1} - \phi_1 & \phi_{\Delta 4,2} - \phi_2 & \phi_{\Delta 4,3} - \phi_3 & \phi_{\Delta 4,4} - \phi_4 & \phi_{\Delta 4,5} - \phi_5 & \phi_{\Delta 4,6} - \phi_6 \\ \phi_{\Delta 5,1} - \phi_1 & \phi_{\Delta 5,2} - \phi_2 & \phi_{\Delta 5,3} - \phi_3 & \phi_{\Delta 5,4} - \phi_4 & \phi_{\Delta 5,5} - \phi_5 & \phi_{\Delta 5,6} - \phi_6 \\ \phi_{\Delta 6,1} - \phi_1 & \phi_{\Delta 6,2} - \phi_2 & \phi_{\Delta 6,3} - \phi_3 & \phi_{\Delta 6,4} - \phi_4 & \phi_{\Delta 6,5} - \phi_5 & \phi_{\Delta 6,6} - \phi_6 \end{bmatrix}.$$

To obtain the first row, let $b' = [\epsilon_1, 0, 0, 0, 0, 0]$, to obtain the second row, let $b' = [0, \epsilon_2, 0, 0, 0, 0]$, and so on. A relative convergence criterion is used to choose the magnitude of discrete perturbations like ϵ_1.

8.4 Results and Interpretations

In order to estimate log pooled prior distributions, the adaptive importance sampling algorithm was applied to our version of Hethcote's AIDS transmission model with 6 inputs and 6 outputs. In this example, two different sources of AIDS information are being combined—epidemiological and biological opinions about input priors for Hethcote's transmission model, and actual counts and survey sampling results regarding outputs from Hethcote's model. By combining this information using our Bayesian inferential methods for priors linked by deterministic simulation models, we can achieve three important goals. First, the incorporation of information from an epidemiological model for AIDS improves our estimates of model outputs, which were previously based on sampling data only. Second, information about the model outputs from the sampling data can help indicate if our expert prior distributions for model inputs are reasonable. Third, we obtain formal assessments of uncertainty for both inputs and outputs.

Priors on the input and output parameters were defined in Sections 8.2.1 and 8.2.2. The adaptive algorithm was stopped after 4 iterations, each with 500 simulation runs. One measure of the quality of the sampling envelope is the

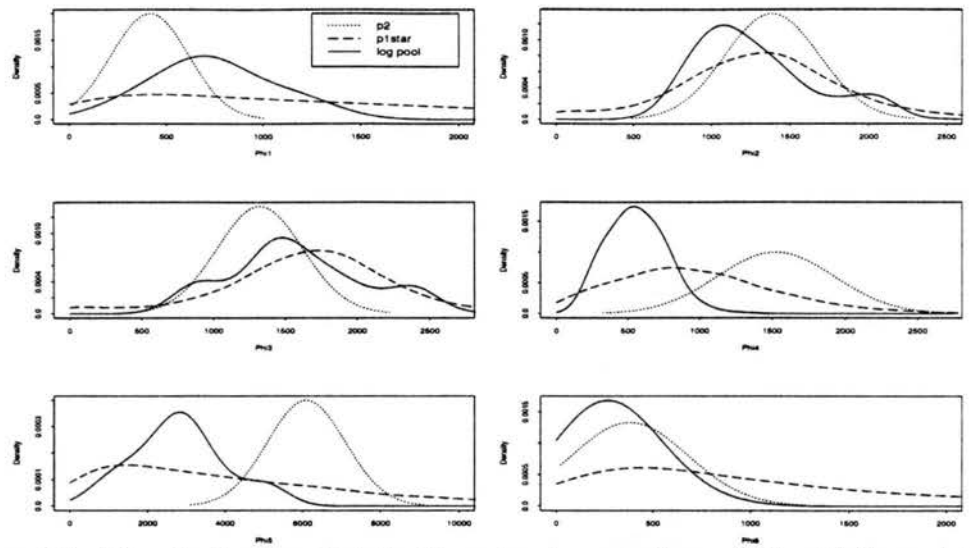

Figure 8.1: Marginal prior distributions in ϕ-space for $p_2(\phi)$, $p_1^*(\phi)$, and the log pooled $p(\phi)$ obtained under adaptive importance sampling.

6-dimensional standardized Euclidean distance from each simulated point to the mean of $p_2(\phi)$. The 420 standardized simulated points following the first iteration had a median distance of 49.1 from the mean of $p_2(\phi)$, while the corresponding 1570 points following the fourth iteration had a median distance of 29.9. This suggests that our adaptive sampling envelope is more representative of the prior opinions of both experts—and hence a more suitable importance sampling envelope for the log pooled prior—than a nonadaptive envelope.

The results of our log pooling estimation in ϕ-space are shown in Figure 8.1, and similar results in θ-space are shown in Figure 8.2. Of particular interest is the inferences which an AIDS researcher might draw from the results contained in Figures 8.1 and 8.2. In the marginal plot for ϕ_1 (Figure 8.1), the input experts' implicit opinion $p_1^*(\phi_1)$ (through the AIDS transmission model) is much broader and extends to much larger HIV incidence levels than the opinion of the output experts. As a result, the marginal log pooled prior $p(\phi_1)$ exhibits a higher mean and greater variability than $p_2(\phi_1)$. The most likely 1995 HIV incidence level, using all available information, is around 700, rather than the estimate of 415 obtained strictly from sampling-based data.

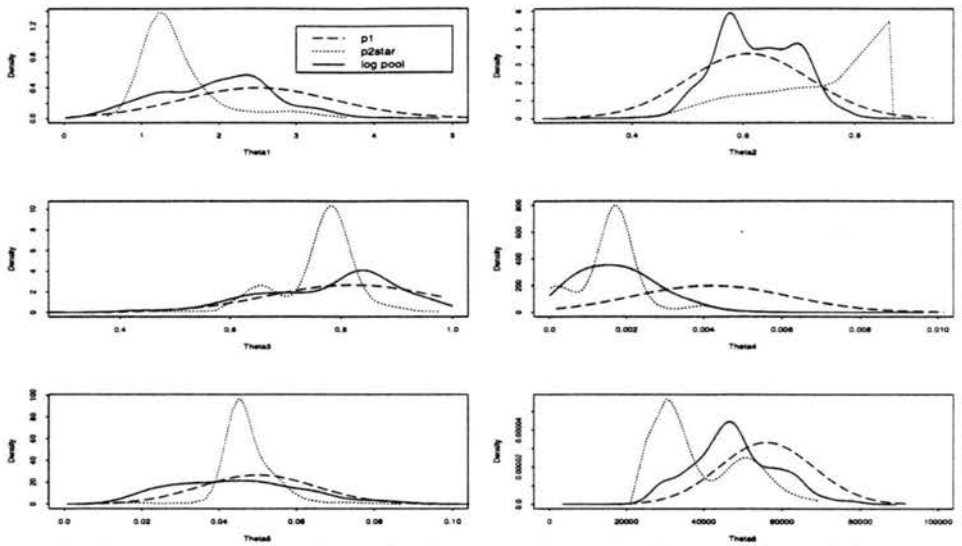

Figure 8.2: Marginal prior distributions in $\boldsymbol{\theta}$-space for $p_1(\boldsymbol{\theta})$, $p_2^*(\boldsymbol{\theta})$, and the log pooled $p(\boldsymbol{\theta})$ obtained under adaptive importance sampling.

The introduction of information from Hethcote's model and the corresponding priors on model inputs produces a few other noticeable marginal changes in $\boldsymbol{\phi}$-space. The mode of the pooled distribution of ϕ_4 (the prevalence of very sexually active Stage 2 men) is about 550, compared to 1524 under $p_2(\boldsymbol{\phi})$. Similarly, the mode of the pooled distribution of ϕ_5 (the prevalence of sexually active Stage 2 men) is about 3000, compared to 6098 under $p_2(\boldsymbol{\phi})$. In both cases, the input experts' implicit opinion ($p_1^*(\boldsymbol{\phi})$) favored smaller prevalences than the corresponding opinion of the output experts. With respect to other marginal distributions, such as those for ϕ_2 and ϕ_3, the pooling of priors led to small location shifts and greater variability in $p(\phi)$ when compared to $p_2(\phi)$.

Note that occasionally marginal results can be counterintuitive. Consider, for instance, results for ϕ_4 in Figure 8.1. In this case, the mode of the pooled prior distribution is smaller than the modes of both $p_2(\phi_4)$ and $p_1^*(\phi_4)$. This illogical result could emerge from several factors. First, the multivariate log pooled prior may be a sensible consensus between the experts even if it does not seem sensible over a particular margin. Second, the adaptive algorithm used to obtain an estimate of the log pooled prior distribution was designed to produce good

estimates of the multivariate $p(\phi)$ without specifically considering the marginal distributions. Third, the precision of log pooled prior estimates could be improved by increasing the number of simulation runs made from the sampling envelope, especially with the high number of dimensions in this AIDS transmission model.

In Figure 8.2, the reasonableness of the input experts priors can be assessed in light of the output experts' implicit beliefs (through the model) about typical values of the input parameters. For example, the output experts' prior beliefs about ϕ tend to correspond to an average number of partners per month (θ_1) near 1.2, which is below the mean of 2.5 ascribed to θ_1 by the input experts. The log pooled marginal prior for θ_1 reflects this disagreement with its skewness to the left. Pooled marginal distributions for θ_4 (monthly migration percentage) and θ_6 (total population size) are also examples of cases in which the input values which lead to the most likely values of $p_2(\phi)$ are not necessarily the input values considered most likely under $p_1(\theta)$. The output experts implicitly favor lower levels of θ_4 than the input experts (modes of .002 vs. .004), and the output experts implicitly favor lower levels of θ_6 than the input expert (modes of 30,000 vs. 56,000). In both cases, $p(\phi)$ strikes a balance between the two differing opinions. For other input variables, such as θ_3 (external mixing fraction) and θ_5 (probability of transmission), the implicit opinion from the output experts ($p_2^*(\theta)$) provides validation of the input experts' prior beliefs. In these cases, the marginal $p(\theta)$ closely resembles marginal $p_1(\theta)$.

Chapter 9

Summary and Conclusions

9.1 Summary

This dissertation focused on making inference about parameters in deterministic simulation models. These models are used by a wide variety of scientists and engineers, in diverse fields such as animal population modeling, epidemiological modeling, global climate modeling, and tree growth and yield modeling. A Bayesian approach provides a good framework for the statistical analysis of deterministic simulation models, permitting:

1. formal assessments of uncertainty about model inputs and outputs,
2. evaluations of the sensitivity of the model to differing input values,
3. comparisons of model output estimates to independent assessments,
4. evaluations the reasonableness of model inputs in light of information about the outputs,
5. identification of applications where the model itself might need rethinking,
6. and systematic organization of all available information about a model.

In Chapters 1 and 2, we illustrate that the problem of making Bayesian inference about parameters in deterministic simulation models is fundamentally related to the issue of aggregating expert opinion. In Chapter 2 we review the literature

about pooling expert opinion, and in Chapters 3-7 we detail and compare 4 particular pooling approaches—linear pooling, logarithmic pooling, French-Lindley supra-Bayesian pooling, and Lindley-Winkler supra-Bayesian pooling. These 4 approaches are compared with respect to three suitability factors—theoretical properties, performance in examples, and the selection and sensitivity of hyperparameters or weightings incorporated in each method. A summary of this comparison is presented in Chapter 7. Based on the three suitability factors described above, we concluded that the logarithmic pool is generally the most appropriate pooling approach when combining expert opinions in the context of deterministic simulation models.

Given that logarithmic pooling is recommended for making inference about model parameters, the next issue becomes how to obtain estimates of log pooled prior distributions for inputs and outputs. To address this issue, we offer a general algorithm in Chapter 3. In addition, we detail an adaptive sampling algorithm which leads to better and more efficient estimates of the log pooled prior than a simpler estimation algorithm, as illustrated in Chapter 4. Our adaptive estimation approach relies on importance sampling methods, density estimation techniques for which we numerically approximate the Jacobian, and nearest neighbor approximations in cases in which the model is noninvertible. In Chapter 8 we apply our adaptive approach for estimating log pooled priors to a large, realistic model for AIDS transmission. Prior distributions for input and output parameters are justified through the AIDS literature, and interpretations of results are offered.

9.2 Contributions

In this dissertation, several unique contributions were made to the statistical discipline. These contributions are described below.

1. Although it has been proposed at meetings of the International Whaling Commission Scientific Committee (Raftery *et al.*, 1996), the idea of applying logarithmic pooling to inference in deterministic simulation models has not appeared in the published literature. The idea of using a Bayesian framework to make inference about parameters in deterministic simulation models was first proposed by Raftery, Givens, and Zeh (1995). However, their inferential method was subject to the Borel paradox, under which different results can occur if parameters are expressed on a different scale (Wolpert, 1995). The Borel paradox results from having two different prior distributions on the same quantity; logarithmic pooling (as well as the other pooling approaches considered in this dissertation) resolves the Borel paradox.

2. The algorithm developed here for estimating log pooled priors (Section 3.1) using an adaptive strategy as outlined in Section 3.2.3 is new. The improvements which the new, adaptive approach offer over a nonadaptive approach are quantified and illustrated in Chapter 4.

3. The Jacobian-based approach to density estimation in this context (Section 3.3), especially in higher dimensions, is a novel approach to the best of our knowledge. The use of directional derivatives to estimate sets of partial derivatives allows higher dimensional Jacobians to be estimated, and this provides an attractive alternative to multivariate histogram density estimation or multivariate kernel density estimation.

4. The extension of the French-Lindley supra-Bayesian methodology to continuous parameters is new to this dissertation, as is the extension of the Lindley-Winkler supra-Bayesian methodology to multivariate parameters.

5. The proofs and illustrations of the failure of Relative Propensity Consistency under the French-Lindley supra-Bayesian approach are new.

9.3 Conclusions

The marriage of statistical ideas and deterministic simulation models through the Bayesian approach outlined in this dissertation is exciting for statisticians and scientists alike. Scientists can combine diverse sources of model information and assess uncertainty and sensitivity in a more formal and thorough manner than current ad-hoc procedures allow; statisticians can witness applications of these ideas to a fascinating variety of scientific models. Hopefully these mutual benefits to statisticians and scientists will encourage the use and improvement of the ideas in this dissertations in the years ahead.

REFERENCES

Agnew, C. E. (1985). Multiple probability assessments by dependent experts. *Journal of the American Statistical Association*, 80:343–347.

Bacharach, M. (1979). Normal Bayesian dialogues. *Journal of the American Statistical Association*, 74:837–846.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 2nd edition.

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden-Day, San Francisco.

Bogstad, B., Tjelmeland, S., Tjelta, T., and Ulltang, O. (1992). Description of a multispecies model for the Barents Sea (MULTSPEC) and a study of its sensitivity to assumptions on food preferences and stock sizes of minke whales and harp seals. Paper SC/44/O9 presented to the IWC Scientific Committee, June, 1992.

Breiwick, J. M., Eberhardt, L. L., and Braham, H. W. (1984). Population dynamics of western arctic bowhead whales (*balaena mysticetus*). *Canadian Journal of Fisheries and Aquatic Sciences*, 41:484–96.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis.* Chapman and Hall, London.

Centers for Disease Control (1987). Self-reported changes in sexual behaviors among homosexual and bisexual men from the San Francisco city clinic cohort. *Morbidity and Mortality Weekly Report, 4/87.*

Centers for Disease Control (1995). CDC National AIDS Clearinghouse. Data available publicly at: hivinsite.ucsf.edu.

Centers for Disease Control (1997). Self-reported changes in sexual behaviors among homosexual and bisexual men from the San Francisco city clinic cohort. *Morbidity and Mortality Weekly Report, 2/97.*

Chatterjee, S. and Senata, E. (1977). Towards consensus: some theorems on repeated averaging. *Journal of Applied Probability*, 14:89–97.

Cooke, R. M. (1990). Statistics in expert resolution: a theory of weights for combining expert opinion. In Cooke, R. and Constantini, D., editors, *Statistics in Science. The Foundations of Statistical Methods in Biology, Physics, and Economics*, pages 41–72. Kluwer Academic Publishers, Netherlands.

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science.* Oxford University Press, New York.

Cox, Jr., L. A. (1991). Knowledge-based resolution of conflicting expert opinions. *Journal of Applied Statistics*, 18:23–34.

Cubasch, U. and Cess, R. D. (1990). Processes and modelling. In Ephraums, J. J., Houghton, J. T., and Jenkins, G. J., editors, *Climate Change, The IPCC Scientific Assessment.* Cambridge University Press, Cambridge.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, 69:118–121.

Foulkes, M. A. (1998). Advances in HIV/AIDS statistical methodology over the past decade. *Statistics in Medicine*, 17:1–25.

French, S. (1980). Updating of belief in the light of someone else's opinion. *Journal of the Royal Statistical Society, Series A*, 143:43–48.

French, S. (1981). Consensus of opinion. *European Journal of Operational Research*, 7:332–340.

French, S. (1985). Group consensus probability distributions: A critical survey. In Bernardo, J. M., DeGroot, M. H., Lindley, D., and Smith, A. F. M., editors, *Bayesian Statistics 2*, pages 183–202. Elsevier Science Publishers B.V., Amsterdam.

Friedman, J. H., Stuetzle, W., and Schroeder, A. (1984). Projection pursuit density estimation. *Journal of the American Statistical Association*, 79:599–608.

Gelfand, A. E., Mallick, B. K., and Dey, D. K. (1995). Modeling expert opinion arising as a partial probabilistic specification. *Journal of the American Statistical Association*, 90:598–604.

Gelman, A., Bois, F., and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, 91:1400–1412.

Genest, C. (1984). A characterization theorem for externally Bayesian groups. *Annals of Statistics*, 12:1100–1105.

Genest, C., Weerahandi, S., and Zidek, J. V. (1984). Aggregating opinions through logarithmic pooling. *Theory and Decision*, 17:61–70.

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography (with discussion). *Statistical Science*, 1:114–148.

George, L. C. and Grant, W. E. (1983). A stochastic simulation model of brown shrimp *Penaeus aztecus* Ives growth, movement, and survival in Galveston Bay, Texas. *Ecological Modelling*, 19:41–70.

Givens, G. H., Zeh, J. E., and Raftery, A. E. (1995). Assessment of the Bering-Chukchi-Beaufort Seas stock of bowhead whales using the BALEEN II population model in a Bayesian synthesis framework. *Report of the International Whaling Commission*, 45:345–364.

Grant, R. M., Wiley, J. A., and Winkelstein, W. (1987). Infectivity of the human immunodeficiency virus: estimates from a prospective study of homosexual men. *The Journal of Infectious Diseases*, 156:189–193.

Green, E. J. and Strawderman, W. E. (1996). A Bayesian growth and yield model for slash pine plantations. *Journal of Applied Statistics*, 23:285–299.

Hammersley, J. M. and Hanscomb, D. C. (1964). *Monte Carlo Methods.* Methuen, London.

Henderson-Sellers, A. and Robinson, P. J. (1986). *Contemporary Climatology.* John Wiley and Sons, New York.

Hethcote, H. W., van Ark, J. W., and Karon, J. M. (1991a). A simulation model of AIDS in San Francisco: II. Simulations, therapy, and sensitivity analysis. *Mathematical Biosciences*, 106:223–247.

Hethcote, H. W., van Ark, J. W., and Longini, Jr., I. M. (1991b). A simulation model of AIDS in San Francisco: I. Model formulation and parameter estimation. *Mathematical Biosciences*, 106:203–222.

Hethcote, H. W. and Yorke, J. A. (1984). *Gonorrhea: Transmission Dynamics and Control (Lecture Notes in Biomathematics 56).* Springer-Verlag, Berlin.

Isham, V. (1988). Mathematical modelling of the transmission dynamics of HIV infection and AIDS: a review. *Journal of the Royal Statistical Society, Series A*, 151:5–30.

Izenman, A. J. (1991). Recenty developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86:205–224.

Jacquez, J. A., Simon, C. P., Koopman, J., *et al.* (1988). Modeling and analyzing HIV transmission: The effect of contact patterns. *Mathematical Biosciences*, 92:119–199.

Jeffreys, H. (1961). *Theory of Probability.* Oxford University Press, London, third edition.

King, J. N. and Johnson, G. R. (1993). Monte carlo simulation models of breeding-population advancement. *Silvae Genetica*, 42:68–78.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Istituto Italiano degli Attuari*, 4:83–91.

Law, A. M. and Kelton, W. D. (1991). *Simulation Modeling and Analysis.* McGraw-Hill, New York, 2nd edition.

Law, A. M. and McComas, M. G. (1988). How simulation pays off. *Manufacturing and Engineering*, 100:37–39.

Lemp, G. F., Payne, S. F., Rutherford, G. W., *et al.* (1990). Projections of AIDS morbidity and mortality in San Francisco. *Journal of the American Medical Association*, 263:1497–1501.

Lindley, D. V. (1983). Reconciliation of probability distributions. *Operations Research*, 31:866–880.

Lindley, D. V. (1985). Reconciliation of discrete probability distributions. In Bernardo, J. M., DeGroot, M. H., Lindley, D., and Smith, A. F. M., editors, *Bayesian Statistics 2*, pages 375–390. Elsevier Science Publishers B.V., Amsterdam.

Lindley, D. V., Tversky, A., and Brown, R. V. (1979). On the reconciliation of probability assessments. *Journal of the Royal Statistical Society, Series A*, 142:146–180.

Longini, I. M., Clark, W. S., Byers, R. H., *et al.* (1989). Statistical analysis of the stages of HIV infections using a Markov model. *Statistics in Medicine*, 8:831–843.

Longini, I. M., Clark, W. S., Haber, M., *et al.* (1990). The stages of HIV infection: waiting times and infection transmission probabilities. In Castillo-Chavez, C., editor, *Mathematical and Statistical Approaches to AIDS Epidemiology (Lecture Notes in Biomathematics 83)*, pages 111–137. Springer-Verlag, New York.

Madansky, A. (1978). Externally Bayesian groups. Unpublished manuscript, University of Chicago.

McConway, K. J. (1978). *The Combination of Experts' Opinions in Probability Assessment: Some Theoretical Considerations.* PhD thesis, University College London.

Morris, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *Annals of Statistics*, 11:515–529.

Morris, P. A. (1977). Combining expert judgements: A Bayesian approach. *Management Science*, 29:679–693.

Oglesby, R. J., Maasch, K. A., and Saltzman, B. (1989). Glacial meltwater cooling of the Gulf of Mexico: GCM implications for Holocene and present-day climates. *Climate Dynamics*, 3:115–133.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Pickering, J., Wiley, J. A., Padian, N. S., *et al.* (1986). Modeling the incidence of AIDS in San Francisco, Los Angeles, and New York. *Mathematical Modelling*, 7:661–688.

Poole, D. and Raftery, A. E. (1998). Assessment of the Bering-Chukchi-Beaufort Seas stock of bowhead whales using Bayesian and full pooling Bayesian synthesis methods. Paper SC/50/AS6 presented to the IWC Scientific Committee, May 1998.

Pritsker, A. A. B. (1986). *Introduction to Simulation and SLAM II.* Systems Publishing Corp., West Lafayette, IN, 3rd edition.

Raftery, A. E., Givens, G. H., and Zeh, J. E. (1995). Inference from a deterministic population dynamics model for bowhead whales (with discussion). *Journal of the American Statistical Association*, 90:402–430.

Raftery, A. E., Poole, D., and Givens, G. H. (1996). The Bayesian synthesis assessment method: Resolving the Borel paradox and comparing the backwards and forwards variants. Paper SC/48/AS16 presented to the IWC Scientific Committee, June, 1996.

Rawlings, J. O. (1988). *Applied Regression Analysis: A Research Tool.* Wadsworth and Brooks, Pacific Grove, CA.

Research and Decisions Corporation (1984). Designing an effective AIDS prevention campaign strategy for San Francisco: results from the first probability sample of an urban gay male community. A report prepared for the San Francisco AIDS Foundation.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837.

Rubin, D. B. (1987). Comment on 'The calculation of posterior distributions by data augmentation'. *Journal of the American Statistical Association*, 82:543–546.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In Bernardo, J. M. *et al.*, editors, *Bayesian Statistics 3*, pages 395–402. Clarendon Press, Oxford.

San Francisco Department of Public Health AIDS Office (1993). Estimates of the number of persons infected with HIV in San Francisco by race/ethnicity. A report prepared on June 28, 1993.

San Francisco Department of Public Health AIDS Office (1997). The HIV consensus report on HIV prevalence and incidence in San Francisco. A report from the 1997 HIV Consensus Meeting.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801.

Schweder, T. and Ianelli, J. N. (1998). Bowhead assessment by likelihood synthesis, methods and difficulties. Paper SC/50/AS2 presented to the IWC Scientific Committee, May 1998.

Simmons, A. J. and Bengtsson, L. (1984). Atmospheric general circulation models, their design and use for climate studies. In Houghton, J., editor, *The Global Climate.* Cambridge University Press, Cambridge.

Smirnov, N. V. (1939). Estimate of deviation between empirical distribution functions of two independent samples (in Russian). *Bulletin of Moscow University*, 2(2):3–16.

Steinhorst, R. K., Morgan, P., and Neuenschwander, L. F. (1985). A stochastic-deterministic simulaton model of shrub succession. *Ecological Modelling*, 29:35–55.

Stone, M. (1961). The opinion pool. *Annals of Mathematical Statistics*, 32:1339–42.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185:1124–1131.

Weerahandi, S. and Zidek, J. V. (1978). Multi-Bayesian statistical decision theory. Report No. 78-34, Institute of Applied Mathematics and Statistics, University of British Columbia.

West, M. (1988). Modelling expert opinion. In Bernardo, J. M. *et al.*, editors, *Bayesian Statistics 3*, pages 493–508. Clarendon Press, Oxford.

Westergaard, H. (1968). *Contributions to the History of Statistics.* Agathon, New York. Page 22.

Winkler, R. L. (1968). The consensus of subjective probability distributions. *Management Science*, 15:B61–B75.

Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science*, 27:479–488.

Wolpert, R. L. (1995). Comment on 'Inference from a deterministic population dynamics model for bowhead whales'. *Journal of the American Statistical Association*, 90:426–7.