Bayesian Inference with Specified Prior Marginals

Michael Lavine 1               Larry Wasserman 2
Duke University               Carnegie Mellon University

Robert L. Wolpert 1
Duke University

July 11, 1990

1Research supported by National Science Foundation grant DMS 8903842
2Research supported by a post-doctoral fellowship from the Natural Sciences
and Engineering Research Council of Canada


Key words and phrases. Frechet classes, Monge-Kantorovich problem, posterior
bounds, robust Bayesian inference.

The order of authorship is not meant to reflect degree of contribution. All three
authors have made major contributions to this work.
Abstract

We show how to find bounds on posterior expectations of arbitrary functions of the parameters when the prior marginals are specified but when the complete joint prior is unspecified. We also give a theorem that is useful for finding posterior bounds in a wide range of Bayesian robustness problems. We apply these techniques to two examples.

The first example involves a recent clinical trial for ECMO (Ware 1989). Our analysis may be regarded as a followup to a detailed Bayesian analysis given by Kass and Greenhouse (1989) who conclude that the posterior probability that the treatment is superior to the control is about 0.95. However, their analysis assumes a priori independence of the parameters. We consider other prior distributions with the same marginals as Kass and Greenhouse, but in which the parameters are not independent, and conclude that as long as a priori independence is at least approximately tenable then ECMO seems superior to the control.

The second example is the product of means problem which has been studied in the Bayesian context by Berger and Bernardo (1989). Here the goal is to find the posterior expectation of $\alpha \beta$ where $\alpha$ and $\beta$ are the means of conditionally independent random variables $X$ and $Y$. Berger and Bernardo recommend a joint prior $\pi_0$ proportional to $(\alpha^2 + \beta^2)^{1/2}$. We find that among all priors with the same marginals as $\pi_0$, the posterior expectation of $\alpha \beta$ can be made arbitrarily large or arbitrarily close to 0. Furthermore, the parameterization is important: with a different parameterization the upper bound is strictly finite.
1 Introduction

When a parameter space is multi-dimensional, Bayesian statistical analysis is plagued by two problems: high dimensional integration and high dimensional prior specification. While the former has received a great deal of attention, the latter has not. Although it may be possible to specify the marginals of a multi-dimensional parameter, specifying the joint distribution may be extremely difficult. In this paper, we show how to compute bounds on posterior quantities when the marginals of the prior are given but when the full prior distribution of the parameters is either completely unknown or known only approximately.

There already exists an extensive literature on finding bounds on the prior expectation of a function of two random variables $E[\phi(\theta_1, \theta_2)]$, say, when the marginal distributions of $\theta_1$ and $\theta_2$ are given. This problem, and some of its generalizations, are known as the Monge-Kantorovich problem. Rachev (1985b) provides an excellent review and bibliography of work up to that date. Explicit solutions are given by Cambanis et al. (1976) and Tchen (1980) for the case where $\phi$ is quasi-monotone (has nonnegative mixed partial derivatives) or quasi-antitone (has nonpositive mixed partial derivatives). Some related work is Strassen (1965) Rachev (1985a) and Kellerer (1984).

Unfortunately, we often work with functions that are not quasi-monotone or quasi-antitone, and we want bounds on the posterior expectations of those functions. The reason the existing work does not help us bound posterior expectations is that the posterior mass distribution can be quite different than the prior mass distribution, and the posterior marginals are not determined
by the prior marginals.

But bounding posterior expectations is important from a Bayesian sensitivity or robust Bayesian viewpoint. As Berger (1987) says,

the bulk of . . . recent work [in robust Bayesian analysis] has been concerned with:

1. Modelling uncertainty in the prior by specifying a class \( \Gamma \), of possible prior distributions; and

2. Determining the range of the posterior quantity of interest as the prior ranges over \( \Gamma \).

Posterior expectations are written \( \int_\Theta \phi(\theta) \pi(d\theta|\text{Data}) \) or \( E_\pi[\phi(\theta)|\text{Data}] \) where \( \theta \in \Theta \), the parameter space, \( \phi : \Theta \to \mathbb{R} \), \( \pi(\cdot|\text{Data}) \) is the conditional measure on \( \Theta \) given the data and \( E_\pi[\cdot] \) means expectation when \( \pi \) is the prior. Some useful choices for \( \phi \) are

- \( \phi(\theta) = \int_B \pi(dx|\theta) \),
- \( \phi(\theta) = \int x \pi(dx|\theta) \),
- \( \phi(\theta) = \theta \),
- \( \phi(\theta) = 1_S(\theta) \),
- \( \phi(\theta) = \ell(a, \theta) \) where \( \ell \) is a loss function and \( a \) is an action and
- \( \phi(\theta) = \pi(dx|\theta)/dx \).

These make the posterior expectation respectively equal to
the predictive probability that the next observation lies in the set $B$,

- the predictive mean,

- the posterior mean,

- the posterior probability that $\theta \in S$,

- the posterior expected loss of $a$ and

- the predictive density at $x$.

In many multiparameter Bayesian problems it is possible to specify marginal prior distributions $\mu_1, \ldots, \mu_n$, for the parameters $\theta_1, \ldots, \theta_n$ say, but difficult to specify the joint distribution $\pi$. This leads us to consider the following class of priors.

Let $\pi_0$ be any fixed joint prior with the given marginals, perhaps the product measure, and define $\Gamma_\epsilon$ to be the $\epsilon$-contamination class $\Gamma_\epsilon = \{(1 - \epsilon)\pi_0 + \epsilon \pi : \pi \in \Gamma\}$ where $\epsilon \in [0,1]$ and $\Gamma = \Gamma_1$ is the set of all joint priors with the given marginals. Uncertainty about $\pi_0$ is reflected in the choice of $\epsilon$. We show how to compute bounds on posterior expectations of $\phi$ over all priors in $\Gamma_\epsilon$. Discussion will be in terms of the upper bound; finding the lower bound is similar. That is, we look for

$$\bar{\rho}_\epsilon \equiv \sup_{\pi \in \Gamma_\epsilon} E_\pi[\phi(\theta)|\text{Data}],$$

where $\theta = (\theta_1, \ldots, \theta_n)^T$, and $\theta_i \in \Theta_i$. For future reference define the prior extreme for any real function $\psi$ to be $\bar{E}_\epsilon[\psi] \equiv \sup_{\pi \in \Gamma_\epsilon} E_\pi[\psi]$ and, setting aside the question of whether the sup is obtained, let $\bar{\pi}_\epsilon \equiv \arg\max_{\pi \in \Gamma_\epsilon} E_\pi[\psi|\text{Data}]$.

Similar notation, with underbars, is used for infima.
In \( n = 2 \) dimensions let \( F_1 \) and \( F_2 \) be the given marginal cdf's. Let 
\[
H(\theta_1, \theta_2) = \max(0, F_1(\theta_1) + F_2(\theta_2) - 1) \text{ and } H(\theta_1, \theta_2) = \min(F_1(\theta_1), F_2(\theta_2)).
\]
Both \( H \) and \( H \) are supported on one-dimensional curves. It is well known
(Tchen 1980) that \( \Gamma \) is the set of all joint priors \( \pi \) whose cdf's \( H \) satisfy \( H \leq H \leq H \) and that the extreme bounds on prior expectations of quasi-monotone and quasi-antitone functions are achieved at \( H \) and \( H \). It is also known
(Lindenstrauuss 1965) that even when \( \mu_1 \) and \( \mu_2 \) are absolutely continuous
with respect to Lebesgue measure, the extreme points of \( \Gamma \) are singular.
Typically, the mass of an extreme point is distributed along one or more
one-dimensional curves.

It is not uncommon to construct a multiparameter prior \( \pi_0 \) by specifying
the marginal distributions and assuming prior independence. Our methods
offer a way of examining sensitivity to departures from the independence as-
sumption. An example, involving a clinical trial, will be given in Section 3.1.

2 Methodology

2.1 Theory

Finding \( \bar{\rho}_e \) directly from Equation 1 is difficult because posterior expectations
are nonlinear functions of the prior. We overcome this obstacle by turning
the problem of finding \( \bar{\rho}_e \) into a series of linear problems as in Lavine (1988).
To this end let \( L \) be the likelihood function, define 
\[
c(q, \theta) = L(\theta)(\phi(\theta) - q)
\]
and note that, for any \( q \in \mathbb{R}, \bar{\rho}_e < q \) if and only if \( \mathbb{E}_e[c(q, \theta)] < 0 \) and hence,
by the remark following Theorem 2,

$$\bar{p}_\epsilon = q \iff \bar{E}_\epsilon[c(q, \theta)] \equiv (1 - \epsilon)E_{\pi_0}[c(q, \theta)] + \epsilon E_\pi[c(q, \theta)] = 0. \quad (2)$$

Thus, to find $\bar{p}_\epsilon$ it suffices to evaluate the function $g(q) = \bar{E}_\pi[c(q, \theta)]$.

Note that this technique applies to any situation where we seek extreme posterior expectations over a set of priors $\pi^*$, not just the fixed marginal problem. In any such situation $\bar{p}^* \equiv \sup_{\pi \in \pi^*} E_\pi[\phi|\text{Data}] < q$ if and only if $g^*(q) \equiv \sup_{\pi \in \pi^*} E_\pi[(\phi - q)\text{L}] < 0$, and the following two theorems apply.

**Theorem 1** The function $g^*$ defined above is non-increasing and convex.

**Proof.** $g^*$ is the supremum of non-increasing linear functions. \hfill \blacksquare

**Theorem 2** $\bar{p^*} = \inf\{q : g^*(q) \leq 0\}$.

**Proof.** First we show $\bar{p}^* \leq \inf\{q : g^*(q) \leq 0\}$, or equivalently, $q < \bar{p}^*$ implies $g^*(q) > 0$. If $q < \bar{p}^*$ then there exists a $\pi \in \pi^*$ such that $q < E_\pi[\phi|\text{Data}]$, hence $E_\pi[(\phi - q)\text{L}] > 0$ and $g^*(q) > 0$. Now we show $\bar{p}^* \geq \inf\{q : g^*(q) \leq 0\}$, or equivalently, $q < \inf\{q : g^*(q) \leq 0\}$ implies $\bar{p}^* \geq q$. If $q < \inf\{q : g^*(q) \leq 0\}$ then $g^*(q) > 0$ and there exists a $\pi \in \pi^*$ such that $E_\pi[(\phi - q)\text{L}] > 0$. Hence, $E_\pi[\phi\text{L}] > qE_\pi[\text{L}]$ and, assuming $E_\pi[\text{L}] > 0$, $E_\pi[\phi|\text{Data}] > q$. Therefore, $\bar{p}^* > q$. \hfill \blacksquare

In $\epsilon$-contamination problems there exists an efficient way to find $\bar{p}_\epsilon$ for various values of $\epsilon$: compute $g(q)$ over a grid of values for $q$ and then solve Equation 2 for different values of $\epsilon$. The point is that the intensive calculation
necessary to evaluate $g(q)$ need only be done once, not separately for each $\epsilon$. One possible shortcut in these calculations is to evaluate $g(q)$ for only a few values of $q$ and then fit a smooth function to $g(q)$. Theorem 1 should be of help in suggesting such a smooth function.

Returning to the fixed-marginal problem, there are a few special cases in which either $\bar{\pi}_\epsilon$ or $g(q)$ may be found explicitly. In $n = 2$ dimensions, if $c(q, \theta)$ is a quasi-monotone or quasi-antitone function of $\theta$, or if $c$ factors as $\prod_{i=1}^{2} c_i(q, \theta_i)$ then a result of Cambanis, Simons and Stout (1976) or Tchen (1980) can be used to calculate $g(q)$. Another case where explicit solutions are available is when $L = L_1(\theta_1) + L_2(\theta_2)$ so that $E[L]$ depends only on the marginals and is constant over $\pi \in \Gamma$ and therefore $\bar{\pi}_\epsilon$ depends only on $E_c[\phi \cdot L]$. If, in addition, $\phi \cdot L$ is either quasi-monotone, quasi-antitone or depends only on the marginals, then $E_c[\phi \cdot L]$, and hence $\bar{\pi}_\epsilon$, can be calculated explicitly.

Most of the time, however, we are less fortunate and $g(q)$ must be evaluated or approximated by some other method. One simple, conservative bound for $g(q)$ is easily obtained. For fixed $q$, define $n$ functions $c_i(\theta_i) = \sup_{\theta, j \neq i} c(q, \theta_1, \theta_2, \ldots, \theta_n)$. Clearly, $g(q) \leq E_{\mu_i}[c_i]$ for $i = 1, 2, \ldots, n$. Thus, $g(q) \leq \hat{E}[q]$ where $\hat{E}[q] = \min(E_{\mu_1}[c_1], E_{\mu_2}[c_2], \ldots, E_{\mu_n}[c_n])$. And $\hat{E}[q]$ is simple to compute because it involves only one dimensional expectations.

But $\hat{E}[q]$ may be very far from $g(q)$, and we will usually want to evaluate $g(q)$ more accurately. To do this, we approximate the continuous problem with a discrete one. For each parameter $\theta_i$, partition the corresponding parameter space $\Theta_i$ into $n_i$ partition elements, $\{\Theta_{ij} : j = 1, \ldots, n_i\}$. Let $j = (j_1, \ldots, j_n)^T$, $1 \leq j_i \leq n_i$, and define $\Lambda_j \equiv \Theta_{j_1} \times \Theta_{j_2} \times \cdots \times \Theta_{j_n}$ so
that \( \Lambda_j \) is a cell in \( \Theta \), the cartesian product of partition elements. We choose a representative \( \lambda_j \in \Lambda_j \) and consider discrete priors that put mass \( p_j \) on the point \( \lambda_j \), subject to the restriction that the \( p_j \)'s sum correctly to give the marginal probabilities \( \mu_i(\Theta_{ij}) \).

The error due to discretization can be estimated by finding the ranges of \( \phi(\theta) \) and \( L(\theta) \) over each partition piece. Convergence as the partitions are refined is discussed in the appendix.

Specifying a joint prior for our discrete approximation entails specifying the \( N \equiv \prod n_i \) probabilities \( p_j \). The integral to be maximized in calculating \( g(q) \) is approximated by \( \sum (\phi(\lambda_j) - q)L(\lambda_j)p_j \), a linear function of the \( p_j \)'s. The marginal constraints can also be expressed as linear restrictions: \( \sum \{p_j : j_i = k\} = \pi_i(\Theta_{ik}) \), for example. The problem is now one of maximizing a linear function of \( N \) nonnegative variables subject to \( M \equiv \sum n_i - n + 1 \) linear constraints and thus can be viewed as a linear programming problem and solved by standard techniques.

### 2.2 The Transportation Algorithm

In \( n = 2 \) dimensions the problem of calculating \( g(q) \) can be viewed as that of solving a transportation (or distribution) problem, a well-studied class of linear programming problems for which there exists an efficient numerical algorithm Hadley (1962) Set \( c_{ij} = (q - \phi(\lambda_{ij}))L(\lambda_{ij}) \) and note that \( g(q) = \inf_{\pi \in \Gamma_*} \sum_i \sum_j c_{ij}p_{ij} \) where the class \( \Gamma_* \) consists of those \( \pi = \{\pi_{ij}\} \) satisfying the \( N = n_1n_2 \) positivity constraints \( \pi_{ij} \geq 0 \) and the \( M = n_1+n_2-1 \) linearly independent marginal constraints \( \sum_j p_{ij} = \mu_i(\Theta_{ii}) \) for \( 1 \leq i \leq n_1 \).
and $\sum_i p_{ij} = \mu_2(\Theta_{2j})$ for $1 \leq j \leq n_2$. In the classic transportation problem $c_{ij}$ would be the unit cost of transporting some substance from the $i^{\text{th}}$ of $n_1$ "sources" to the $j^{\text{th}}$ of $n_2$ "destinations," and the objective would be to minimize the total cost $\sum_{ij} c_{ij} x_{ij}$ subject to the marginal constraints expressing limitations on each source's production capacity and minimum requirements for each destination's supply. So long as the constraints are consistent (i.e. production exceeds demand, $\sum_i p_{1i} \geq \sum_j p_{2j}$; both are 1 in our application) there always exists an optimal solution with at most $M$ non-zero shipments $p_{ij}$.

Our implementation of the algorithm first determines a feasible initial assignment of probability to $M = n_1 + n_2 - 1$ of the $n_1 n_2$ points $\lambda_{ij}$ consistent with the positivity and marginal constraints and then proceeds iteratively, determining Lagrange multiplier vectors $u$ and $v$ such that $c_{ij} = u_i + v_j$ for each basis element (i.e. each $ij$ for which $x_{ij} > 0$); points outside the basis with negative cost differentials $c_{ij} - u_i - v_j$ are added successively to the basis (displacing other elements, and maintaining the same marginals) until $c_{ij} \geq u_i + v_j$ for each nonbasic $ij$. It can be shown that this signifies an optimal probability distribution. It seems to be very efficient to begin iterations for successive values of $q$ with the mass distribution attaining the minimum $g(q)$ for the previous $q$. Details of the transportation algorithm are given in many places including Hadley (1962). ANSI C source-code for our implementation is available from one of the authors (RLW).
3 Examples

3.1 A Clinical Trial: ECMO

A recent clinical trial on a treatment for respiratory disease in infants (Ware, 1989) was the subject of a thorough Bayesian analysis by Kass and Greenhouse (1989). The trial involved 9 patients who were given ECMO (extracorporeal membrane oxygenation) of whom all 9 survived, and 10 patients given standard therapy of whom 6 survived. The question of whether randomizing patients in this trial was ethical has been the focus of intense debate. A crucial point was that there was prior information available before the trial. To analyze these data in light of the available prior information, Kass and Greenhouse considered 84 different prior distributions, all involving an assumption of prior independence. They say “We find the independence assumption ... somewhat subtle.”

Let \( p_1 \) be the probability of success under the standard therapy and let \( p_2 \) be the probability of success under ECMO. Let \( \eta_i = \log(p_i/(1 - p_i)) \) be the logit of \( p_i \) and define parameters \( \delta \) and \( \gamma \) by \( \delta = \eta_2 - \eta_1 \) and \( \gamma = (\eta_1 + \eta_2)/2 \). The prior density favored by Kass and Greenhouse was the independence prior \( \pi_0(\delta, \gamma) \) where \( \delta \) is Cauchy\((0, 1.099^2)\) and \( \gamma \) is Cauchy\((0, .419^2)\).

The function of interest \( \phi \) is the indicator function for the event that ECMO is superior to standard treatment, \( \delta > 0 \) (or equivalently, \( p_2 > p_1 \)). Under \( \pi_0 \), \( \Pr(\delta > 0 | \text{Data}) \approx .95 \) indicating substantial evidence in favour of ECMO. The upper and lower bounds \( P_\epsilon \) and \( L_\epsilon \) of \( \Pr(\delta > 0 | \text{Data}) \) for various values of \( \epsilon \) are plotted as solid lines in Figure 1. Note that the bounds are
close to 0 and 1 when $\epsilon = 1$. Thus, the data say nothing if just the marginal prior information is included without the hypothesis of prior independence. It is possible to show analytically that $P_1$ is not quite zero, and that $\bar{P}_1$ is not quite one.

Figures 2, 3 and 4 show the reason for this phenomenon. Figure 2 is a plot of the likelihood contours for $\delta, \gamma \in [-3, 3]$. Figure 3.a shows the prior distribution $\bar{\pi}_1$ that maximizes $Pr(\delta > 0|\text{Data})$ for $\epsilon = 1$. Each circle represents an atom of mass proportional to the area of the circle. Figure 3.b shows the corresponding posterior mass. The point masses occur because we have discretized the problem. Recall that the extremes of $\Gamma$ are singular with mass lying on curves. Therefore a distribution on curves joining the point masses would be a good approximation to the maximizing prior for the continuous version of the problem. Note that in Figure 3.a most of the mass to the left of $\delta = 0$ occurs in places where the likelihood is relatively small. The tail of the marginal prior for the nuisance parameter $\gamma$ is important here. If the tail were smaller then the prior mass to the left of $\delta = 0$ would have a bigger likelihood and the upper bound would decrease. Figure 4 is similar to Figure 3 but shows the prior and posterior that minimize $Pr(\delta > 0|\text{Data})$.

An important point is the way that the bounds depend on $\epsilon$. The bounds do not become unreasonably large until we have extreme values of $\epsilon$. The reason for this is that the average likelihood of the independence prior is much larger than the average likelihood of the maximizing prior $\bar{\pi}_\epsilon$. The interpretation is that the posterior probabilities are not sensitive to small departures from the assumption of prior independence. In particular, note that as long as $\epsilon$ is less than about 0.8, then the posterior probability for
$\delta > 0$ is at least $1/2$. We can go a long way from prior independence before being unsure whether to recommend ECMO.

We consider several other computations for this example. First, we consider the class $\Gamma_\varepsilon = \{(1 - \varepsilon)\pi_0 + \varepsilon \pi; \pi \in \mathcal{P}\}$ where $\mathcal{P}$ is the set of all priors. This $\varepsilon$-contaminated class of priors (Berger 1987) differs from $\Gamma_\varepsilon$ in placing no restrictions on the marginals. Applying Theorem 2, it is easy to show that the upper bound on the posterior expectation of a function $\phi$ is obtained as the solution of $g(q) = 0$ where $g(q) \equiv (1 - \varepsilon)E_{\pi_0}[L(\delta, \gamma)(\phi(\delta, \gamma) - q)] + \varepsilon \sup_{(\delta, \gamma) \in \mathbb{R}^2} L(\delta, \gamma)(\phi(\delta, \gamma) - q)$. The resulting bounds are plotted as dotted lines in Figure 1. We see that the bounds are substantially wider than $\bar{\nu}_\varepsilon$ and $\bar{\nu}_\varepsilon$, indicating that the marginal constraints imposed in $\Gamma_\varepsilon$ are definitely informative. In particular, we need not go so far from independence before becoming unsure whether to recommend ECMO.

We repeated the analysis using another prior suggested by Kass and Greenhouse, a $N(0, 1.63^2)$ for $\delta$ and a $N(0, 0.621^2)$ for $\gamma$. The results were virtually identical to those obtained using Cauchy priors.

Next, we computed posterior bounds on $\Pr(\delta > 0|\text{Data})$ over the class $\Gamma_N$ of bivariate normals with the given (normal) marginals. We simply evaluated the posterior probability at several values of the correlation coefficient $r$. The posterior probability increases from $\bar{\nu}_1 = .409$ at $r = -1$ to $\bar{\nu}_1 = .998$ at $r = 1$. Thus, some reduction in the posterior range is obtained by imposing the strong assumption of joint normality.

We then computed the conservative bounds for $\nu\bar{\nu}_\varepsilon$ and $\bar{\nu}_\varepsilon$ found from $\hat{E}\hat{\|q\|}$. The bounds are plotted as dashed lines in Figure 1. The bounds are surprisingly sharp in this application and substantially easier to compute.
than $\rho_\varepsilon$ and $\bar{\rho}_\varepsilon$. The bounds are always conservative, so it may be useful to calculate them as a first step of a robustness analysis.

So far, we have assumed that the marginals are exactly correct. The final part of this robustness analysis is to assess the sensitivity to the assumed marginals and compare the results to the sensitivity to the prior independence assumption. To this end, let $\mathcal{P}$ be the set of all probability measures on the real line and let $\Gamma_\varepsilon^\delta = \{((1 - \varepsilon)\mu_\delta + \varepsilon P) \times \mu_\gamma; P \in \mathcal{P}\}$ and $\Gamma_\varepsilon^\gamma = \{\mu_\delta \times ((1 - \varepsilon)\mu_\gamma + \varepsilon P); P \in \mathcal{P}\}$ where $\mu_\delta$ and $\mu_\gamma$ are the original Cauchy marginal priors for $\delta$ and $\gamma$ respectively. The variables $\delta$ and $\gamma$ are independent under each prior in both classes; the first embeds $\delta$ in an $\varepsilon$-contaminated class while second embeds $\gamma$ in an $\varepsilon$-contaminated class. Finding the posterior bounds over these classes allows us to assess the sensitivity to the assumed marginal priors. This type of analysis is a special case of Lavine (1989) who analyzes sensitivity to marginal and conditional prior distributions of parameters.

The upper posterior expectations for the two classes are obtained as the solutions of $g_1(q) = 0$ and $g_2(q) = 0$, respectively, where

$$g_1(q) = (1 - \varepsilon)E_{\pi_\theta}[L(\theta)(\phi(\theta) - q)] + \varepsilon \sup_{\delta}E_{\mu_\gamma}L(\delta, \gamma)(\phi(\delta, \gamma) - q)$$

and

$$g_2(q) = (1 - \varepsilon)E_{\pi_\theta}[L(\theta)(\phi(\theta) - q)] + \varepsilon \sup_{\gamma}E_{\mu_\delta}L(\delta, \gamma)(\phi(\delta, \gamma) - q).$$

A plot of the bounds as a function of the $\varepsilon$-contamination of the $\delta$ marginal is in Figure 5a. A similar plot for the $\gamma$ marginal is in Figure 5b. The analysis is remarkably insensitive to the $\gamma$ marginal.

In conclusion, it appears that ECMO may be regarded as being superior to the standard therapy under the prior distributions recommended by Kass
and Greenhouse as long as prior independence and the marginal prior for \( \delta \) are at least approximately tenable.

### 3.2 The Product of Two Means

#### 3.2.1 Background

For a second example we turn to a problem studied in Stein (1982), Efron (1986) and Berger and Bernardo (1989) among others. Let \( X \) and \( Y \) be two independent normally-distributed random variables with uncertain means \( \alpha \) and \( \beta \) and common conditional variance 1 (given \( \alpha \) and \( \beta \)). Both \( \alpha \) and \( \beta \) are known to be positive. The goal is to find the posterior expectation of \( \alpha \beta \).

Stein recommended the use of a prior \( \pi_0 \) proportional to \( (\alpha^2 + \beta^2)^{1/2} \) as a noninformative prior for the problem of estimating the product of independent normal means, as recounted in Efron. Berger and Bernardo follow a quite different line of reasoning (extending Bernardo's notion of reference prior), also leading to a recommendation of \( \pi_0 \).

We will study how sensitive the expected product \( E_\pi[\alpha \beta] \) is to the particular choice of prior \( \pi = \pi_0 \), by considering other priors \( \pi \) with the same marginal distributions for \( \alpha \) and \( \beta \). We will also consider other prior measures \( \pi \) with the same marginals in the transformed polar coordinates, \( r \) and \( \theta \) (which are independent under \( \pi_0 \)).

Note that neither Stein nor Berger and Bernardo based their recommendations of \( \pi_0 \) on any properties of the marginal distributions; our analysis is not directly relevant to the problems they considered and should not be construed as a critique of their methods.
The example is interesting for two reasons. First, it is analytically tractable — exact bounds $\rho_1$ and $\overline{\rho}_1$ can be found. And second, it illustrates that a great deal can depend upon the parameterization. In this example the bounds on the posterior expectation of $\alpha\beta$ for priors whose $\alpha$ and $\beta$ marginals agree with those of $\pi_0$ are $\rho_1 = 0$ and $\overline{\rho}_1 = \infty$. But the upper bound on the posterior expectation of $\alpha\beta = r^2 \sin \theta \cos \theta$ is finite for priors whose polar-coordinate marginals for $r$ and $\theta$ agree with those of $\pi_0$.

Berger and Bernardo considered four possible data sets, each consisting of a single observation of the pair $(X,Y)$: $(x,y) = (0,0), (1,1), (3,3)$ or $(3,0)$. We present only $X = 0, Y = 0$; the results for the other three data sets are qualitatively similar.

When we try to find the $\alpha$ marginal we discover that $\int_0^\infty \pi_0(\alpha, \beta) \, d\beta = \infty$ and we are faced with the question of what it means for a prior $\pi$ to have the "same marginals as $\pi_0$." We are also unable to make sense of the notion of an $\epsilon$-mixture of improper priors so we will restrict our attention to the case of $\epsilon = 1$ and calculate the maximum $\overline{\rho}_1$ and minimum $\rho_1$ value of $E_{\pi}[\alpha\beta]$ as $\pi$ ranges over the class $\Gamma$ of priors with the same marginals as $\pi_0$, in the sense described below.

### 3.2.2 Rectangular Coordinates

The measure $\pi_0$ is $\sigma$-finite, and in particular can be normalized over any compact set. We define $\Gamma$ to be the set of priors $\pi$ such that there exists a compact set $B_\pi$ satisfying

1. $\pi$ and $\pi_0$ have the same marginals on $B_\pi$ and
2. $\pi$ and $\pi_0$ are identical on $B^c_\pi$.

The next theorem shows that with an observation at $(0,0)$ the posterior expectation of $\alpha \beta$ can be made arbitrarily large or arbitrarily close to 0 with a prior from the class $\Gamma$.

**Theorem 3** $\bar{\rho}_1 = \infty$ and $\rho_1 = 0$.

The proof is deferred to the appendix.

### 3.2.3 Polar Coordinates

In polar coordinates $\pi_0$ is proportional to $r^2 \, dr \, d\theta$. We therefore consider the class of priors with uniform $\theta$ marginals on $[0, \pi/2]$ and $r$ marginals proportional to $r^2$ on $[0, \infty)$. The posterior expectation of $\alpha \beta = r^2 \sin \theta \cos \theta$ is the quotient of two integrals:

$$E_\pi[r^2 \sin \theta \cos \theta | (0,0)] = \frac{\int \int r^2 \sin \theta \cos \theta e^{-r^2/2} \pi(r, \theta) \, dr \, d\theta}{\int \int e^{-r^2/2} \pi(r, \theta) \, dr \, d\theta}$$

The denominator depends on $\pi$ only through the $r$ marginal and is the same ($(\pi/2)^{3/2}$, with the indicated normalization) for all priors in our class. Bounding the posterior expectation is therefore equivalent to bounding the numerator.

Because $\sin \theta \cos \theta \leq 1/2$ the numerator is no greater than $K \equiv (\pi/4) \int r^4 e^{-r^2/2} \, dr = (3/2)(\pi/2)^{3/2}$.

**Theorem 4** $\bar{\rho}_1 = 3/2$ and $\rho_1 = 0$.

The proof is deferred to the appendix.
This example shows two things. First, the bounds may depend strongly on the parameterization, and second, analytic solutions are sometimes available. However, the analytic techniques are specific to the problem. We know of no general way to get analytic results.

4 Discussion

The construction of priors remains one of the most challenging aspects of Bayesian inference. Methods of including partial prior information are perhaps the most promising avenue for moving away from noninformative priors and moving toward honest, informative priors. In this paper we have explored one possibility in this direction. Constructing one dimensional marginals is an easier task than constructing the entire joint distribution. We have seen that fixing the prior marginals alone does not result in much posterior information, but allowing for approximate knowledge of the joint prior does produce non-trivial posterior inferences. An alternative way to view the results in this paper is as a methodology for studying deviations from the assumption of prior independence. Along this line, it would be interesting to find a way to calculate the derivative of \( \tilde{\pi} \) with respect to \( \epsilon \), evaluated at \( \epsilon = 0 \) as a diagnostic for deviations from the prior independence assumption. Of course, we can use the techniques of this paper to evaluate this diagnostic numerically, but it would be useful to develop simpler methods, perhaps based on the approximation in section 2.1. Also, it would be of interest to make some formal comparison between the sensitivity to prior independence versus sensitivity to the marginals as we did only informally in the ECMO
example.

For the fixed marginal, \( \epsilon \)-contamination problem, the class \( \Gamma_\epsilon \) contains priors that would typically be regarded as unreasonable, either because they have singular components or because they have components that are very far from \( \pi_0 \). Therefore, \( \bar{p}_\epsilon - \underline{p}_\epsilon \) may overstate the true uncertainty about posterior expectations. Nonetheless, we think that our technique is useful. We suggest using it, and, if \( \bar{p}_\epsilon - \underline{p}_\epsilon \) is small enough, proceeding with some confidence. However, if \( \bar{p}_\epsilon - \underline{p}_\epsilon \) is too large then further analysis is called for.

One approach is to use the simplex method for computing the bounds so that additional constraints can be added to the prior. The method of density bounded classes (Lavine, 1988) is one possibility. Here, the densities are bounded above and below so that singular priors are ruled out.

Another possibility is to elicit some information about the joint measure. This could take the form of a few conditional quantiles, a prior for one or more functions of the parameters, or an expectation for one or more functions. All of these can be modelled as linear constraints on the \( p_i \)'s, and hence handled through linear programming techniques. This could potentially restrict the posterior bounds by a large amount. In higher dimensions, however, standard techniques such as the simplex method may become unwieldy. There are \( N = n_1 \times \cdots \times n_k \) variables to keep track of, a number that grows exponentially in the dimension \( k \).

We have addressed the issue of what to do when the joint prior is not known exactly. A related issue is the question of how to think about the joint prior and determine whether independence is appropriate. One bit of advice is to ask whether knowledge of one parameter would change one's
opinion about the other parameter. Sometimes it is possible to transform to a parameterization where independence is reasonable. See Kass and Greenhouse (1989) for their discussion of independence in the ECMO example and Wolpert (1987) for further comments about choice of parameterization. Our advice is to choose the parameterization in which one can make the best estimate of $\pi_0$, then use our technique to evaluate sensitivity to small variations from $\pi_0$.

Acknowledgements. The authors would like to thank the participants of the Bayesian workshop at Carnegie Mellon University, the referees and an associate editor for many useful comments.

5 References


5 REFERENCES


Hadley, G. (1962), Linear Programming, Addison-Wesley, Reading, Massachusetts.


6 Appendix

Convergence of the algorithm when $h$ is uniformly continuous:
Let $h$ be any uniformly continuous function. For each $k = 1, 2, \ldots$ let $M_{jk}$ be a partition of $\Theta_i$ into $n_{ik}$ intervals and $M_k$ be the partition of $\Theta$ formed from the cartesian product of the elements of the $M_{ik}$. Let $j_k = (j_{1k}, \ldots, j_{nk})^T$, $1 \leq j_{ik} \leq n_{ik}$, and define $\Lambda_{jk} = \Theta_{1j_1k} \times \Theta_{2j_2k} \times \cdots \times \Theta_{nj_nk}$ so that $\Lambda_{jk}$ is an element of $M_k$ and assume further that the maximum size of $\Lambda_{jk}$ over all $j$
decreases to 0 as $k$ increases. Let $h_k$ and $\overline{h}_k$ be the functions that replace $h$
respectively with its minimum and maximum on each $\Lambda_{jk}$.

Because $h_k \leq h \leq \overline{h}_k$ it follows that $\sup_{\pi} E_{\pi} h_k \leq \sup_{\pi} E_{\pi} h \leq \sup_{\pi} E_{\pi} \overline{h}_k$.

Given $\epsilon < 0$, by uniform continuity we can choose $k$ large enough so that
$\overline{h}_k - h_k < \epsilon$ and hence, $\sup_{\pi} E_{\pi} \overline{h}_k - \sup_{\pi} E_{\pi} h_k < \epsilon$. Our algorithm yields an
answer between $h_k$ and $\overline{h}_k$ and therefore converges to $\sup_{\pi} E_{\pi} h$.

Unfortunately, we do not have a proof of convergence for more general
functions, such as indicators.

**Proof of Theorem 3:** To show that the expectation can be made arbitrarily large we show that for any $k > 0$
there exists a prior $\pi_k \in \Gamma$ such that
$E_{\pi_k} |(\alpha \beta)| (0, 0) | \geq k$.

Let $B_1$ be the box with lower left corner at the point $(a, a)$, with sides
parallel to the axes, and with sides of length $b$. Let $B_2$ be the box with lower
left corner at the origin, with sides on the axes, and with sides of length $c$.
We always take $c = a + b$ so that $B_1 \subset B_2$. Let $B_3 = B_2 + (c, 0)$ be $B_2$ shifted
right by $c$ units, $B_4 = B_2 + (c, c)$ and $B_5 = B_2 + (0, c)$. Think of $B_1$ as fixed
and $c$ as variable. Construct the prior $\pi_c \in \Gamma$ by starting with $\pi_0$, sliding all
the mass in $B_2 \cap B_1^c$ units to the right, and sliding the compensating mass
in $B_4$ to the left. Inside $B_3 \cup B_4 \cup B_5$ and $\pi_c$ and $\pi_0$ have the same marginals.
Outside the box they are identical.

Letting $C = B_2 \cup B_3 \cup B_4 \cup B_5 \cap B_1^c$, $D = (B_2 \cup B_3 \cup B_4 \cup B_5)^c$ and $L$ be
the likelihood function we get

$$E_{\pi_c} |(\alpha \beta)| (0, 0) | = \frac{\int \alpha \beta L(\alpha, \beta) \pi_c(\alpha, \beta) \, d\alpha \, d\beta}{\int L(\alpha, \beta) \pi_c(\alpha, \beta) \, d\alpha \, d\beta}$$

$$= \frac{\int_{B_1^c} \cdots + \int_C \cdots + \int_D \cdots}{\int_{B_1^c} \cdots + \int_C \cdots + \int_D \cdots}$$

$$\geq \frac{\int_{B_1^c} \cdots + \int_C e^{-\epsilon/2} e^{-\alpha/2} e^{-\beta/2} \, d\alpha \, d\beta}{\int_{B_1^c} \cdots + \int_C \cdots}$$

The numerator and the first term in the denominator do not depend on $c$.
The other two denominator terms go to 0 as $c$ goes to infinity. Therefore the
posterior expectation of $\alpha \beta$ under $\pi_c$ goes to the posterior expectation under
$\pi_0$ restricted to the set $B_1$. As long as $a > \sqrt{k}$ the limit is also bigger than
$k$, so $\rho_1 = \infty$.

To show that $\rho_1 = 0$ we show that for any $\epsilon > 0$ there exists a prior
$\pi_\epsilon \in \Gamma$ such that $E_{\pi_\epsilon} |(\alpha \beta)| (0, 0) | \leq \epsilon$. Here we argue as for the upper bound
but take $a = 0$ and $b < \sqrt{\epsilon}$.

\[ \]
Proof of Theorem 4: Instead of a formal proof we simply give the idea behind the proof. For small values of \( r, r < R \) say, we put all the prior mass in a wedge near the line \( \theta = \pi/4 \). We use intermediate values of \( r, R < r < S \), to adjust the \( \theta \) marginal, putting mass near the axes. For \( r > S \), we use the prior \( \pi_0 \). The likelihood downweights the mass far from the origin, so the posterior is dominated by small values of \( r \) where the prior is concentrated near \( \theta = \pi/4 \). Therefore, the numerator in the posterior expectation can be made arbitrarily close to \( K \equiv (\pi/4) \int r^4 e^{-r^2/2} dr = (3/2)(\pi/2)^{5/2} \).

A similar argument works for the lower bound. For small values of \( r \) we put all the prior mass in wedges near the axes, making \( r^2 \sin \theta \cos \theta \) as small as we like. The posterior will be dominated by the mass closest to the origin, so the posterior expectation can be made arbitrarily small. Thus, the lower bound on the posterior expectation is 0.
Figure 1. Bounds on the posterior probability that $\delta > 0$ for the ECMO example as a function of $\epsilon$. The solid line is exact, the dashed line is crude. The dotted lines show the bounds when the marginals are not fixed.

Figure 2. Contours of the likelihood function for the ECMO example over the region of the parameter space where the priors are non-negligible.

Figure 3. Mass plots of the distribution that maximizes $\Pr[\delta > 0 | \text{Data}]$: (a) prior; (b) posterior.

Figure 4. Mass plots of the distribution that minimizes $\Pr[\delta > 0 | \text{Data}]$: (a) prior; (b) posterior.

Figure 5. Bounds on the posterior probability that $\delta > 0$ for the ECMO example as a function of $\epsilon$. The first plot is for contaminating the $\delta$ marginal; the second plot is for contaminating the $\gamma$ marginal.
Figure 3a. Mass plot of the prior distribution that maximizes 
$\Pr(\delta > 0|\text{Data})$.

Figure 3b. Mass plot of the posterior distribution that maximizes 
$\Pr(\delta > 0|\text{Data})$. 
Figure 4a. Mass plot of the prior distribution that minimizes \( \Pr[\delta > 0|\text{Data}] \).

Figure 4b. Mass plot of the posterior distribution that minimizes \( \Pr[\delta > 0|\text{Data}] \).