Exploring Imprecise Probability Assessments Based on Linear Constraints

RADU LAZAR
University of Minnesota, USA

GLEN MEEDEN
University of Minnesota, USA

Abstract

For many problems there is only suf£cient prior information for a Bayesian decision maker to identify a class of possible prior distributions. In such cases it is of interest to £nd the range of possible values for the prior expectation for some real valued function of the parameter of interest. Here we show how this can be done when the imprecise prior assessment is based on linear constraints. In particular we £nd the joint range of possible values for a pair of such functions. We also study the joint range of the posterior expectation for a pair of functions.

Keywords

linear constraints, probability assessment, Bayesian inference, Metropolis-Hastings algorithm

1 Introduction

Consider the usual statistical inference problem where a subjective Bayesian must select a prior probability distribution which re¤ects their prior knowledge and beliefs about the unknown state of nature. Often one is unable to actually choose a single prior even though some prior information is present. When this occurs the Bayesian often selects a family of possible prior distributions. In such cases one could be interested in the range of possible values for some function de£ned on the the family of possible priors. More generally one could be interested in the set of possible values for a pair of functions. By judiciously selecting different pairs of functions such a graphical representation could help the Bayesian assess how sensible their initial choice for the family of priors really is. These graphical representations could help a pair of experts resolve possible con¤icting prior beliefs. It could be used to check for areas of disagreement and to see how adjustments of some of their beliefs could lead to a merging of opinions.

In section two we assume that the parameter space contains only a £nite number of points, say k. If $p = (p_1, \ldots, p_k)$ denotes a typical prior distribution we assume that the prior information can be expressed through linear equalities and inequalities involving the p_i's. This restricts the class of possible priors to a convex subset, $\mathcal{C}$ say, of the $k-1$ dimensional simplex of possible probability vectors of length k. Note $\mathcal{C}$ must be a convex polytope generated by a £nite number of extreme points or vertices. Let ϕ denote a real valued function de£ned on the parameter space. Then its prior expectation is a linear function on $\mathcal{C}$. Dickey (see Dickey (2003)) has developed an interactive computing environment which computes the minimum and maximum of its expectation over $\mathcal{C}$. This programs allows the statistician to incorporated their prior information in stages and see how the range of ϕ changes. Here we are interested in £nding the range of a pair of such functions. Because the prior expectations of the two functions are linear functions of p the range of possible values must be a convex set. Moreover, its extreme points

must be contained in set of points which are images of the extreme points of $\mathcal{C}$. Hence, this problem is easily solved if one knows the extreme points of $\mathcal{C}$. But these can be found using a program that developed by Fuduka. See Fuduka (2003).

When considering the posterior expectation of such functions our problem becomes much more dif£cult analytically since the posterior expectation is no longer a linear function over $\mathcal{C}$. However, knowing the extreme points of $\mathcal{C}$ lets one £nd an approximate solution quite easily. Using these points one can generate random values in $\mathcal{C}$ by assigning random weights to them. Then one £nds the values of the two functions at each of the realizations and plots these pairs of values.

In section three we consider the situation where the parameter is a r dimensional vector. We assume that it belongs to a convex polytope in r-dimensional Euclidean space de£ned by some known linear equalities and inequalities which re¤ect some of prior information of the statistician. A Bayesian needs to select a prior or possibly a family of prior distributions over this set to further re¤ect their uncertainty. For a given prior one is interested in computing prior and posterior expectations of some function of the parameter. In practice, it is usually not possible to make independent draws from a probability density de£ned on such a set. In such cases statisticians often employ Markov chain Monte Carlo methods to generate dependent samples from the posterior from which expectations can be computed approximately. Here, we use the Metropolis-Hastings algorithm to construct dependent samples drawn from a prior of interest. If the statistician selects as their family of priors all possible convex combinations of some £nite collection of priors de£ned on the parameter space then for any pair of functions de£ned on the parameter space one can £nd the range of all possible values of their prior or posterior expectations.

As far as we know statisticians have not really addressed problems where imprecise knowledge is expressed through linear constraints. In section four we will note examples of somewhat similar problems that have been studied in the operations research litera-

ture. Finally, we will point out some of the dif£culties when either the dimension of $\mathcal{C}$ or the parameter space gets too large.

2 The parameter space is £nite.

We begin by assuming that the parameter space contains only £nitely many points, say k. In k-dimensional Euclidean space let Λ denote the $k-1$ dimensional simplex of p vectors with $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. We assume that the known relations among the p_i's can be expressed by

$$Ap = a \tag{1}$$

where A is a known $r \times k$ matrix and a is a known vector of length r and

$$Bp \leq b \tag{2}$$

$$p \geq 0 \tag{3}$$

where B is a known $s \times k$ matrix and b is a known vector of length s. The set of $p \in \Lambda$ which satisfy the above equations form a closed convex subset of Λ which we will denote by $\mathcal{C}$.

Note that the interior of $\mathcal{C}$ is empty but we assume that properly considered $\mathcal{C}$ will have a nonempty interior in some smaller dimensional Euclidean space with a dimension of at least two. If ϕ_1 and ϕ_2 are two functions de£ned on the parameter space we let $\mathcal{C}(\phi_1, \phi_2)$ denote their range of possible values over $\mathcal{C}$. Our problem is to £nd this set. To see what could happen in practice we considered the following simple example.

Example 1 *We let $k = 10$ and imposed two equality constraints and two inequality constraints. The equality constraints were $p_5 = p_6$ and $\sum_{i=1}^{10} i p_i = 5.5$ while the inequality constraints were $p_1 \leq p_2$ and $\sum_{i=1}^{4} p_i \leq 0.5$. When doing the posterior calculations we assumed that the probabilities of seeing the observed data under the 10 possible parameter values were 0.1, 0.15, 0.09, 0.2, 0.3, 0.2, 0.1, 0.05, 0.07 and 0.02.*

As we noted in the introduction $\mathcal{C}(\phi_1,\phi_2)$ is a convex set whose extreme points are contained in the image of all the extreme points of $\mathcal{C}$. Hence, this becomes an easy problem once we know the extreme points of $\mathcal{C}$.

Fortunately for many problems the extreme points of $\mathcal{C}$ can be found easily using a program that is available over the Internet. See Fuduka (2003). It turns out for our example that $\mathcal{C}$ has 28 extreme points.

For de£niteness we let $\phi_1(i) = (i-5.5)^2$ for $i = 1,\ldots,10$ and ϕ_2 be the indicator function of the set $\{2,3,4,5\}$.

Plotting ϕ_1 and ϕ_2 at the extreme points of $\mathcal{C}$ we found the seven extreme points of $\mathcal{C}(\phi_1,\phi_2)$.

Since we know the extreme points of $\mathcal{C}$ we can use the Dirichlet family of distributions to generate a fairly ¤exible class of distributions which take values in $\mathcal{C}$ and from which one can simulate directly. If $q^1,\ldots,q^m$ are the extreme points of $\mathcal{C}$ and W is Dirichlet(α) where $\alpha = (\alpha_1,\ldots,\alpha_m)$ then $\sum_{i=1}^{M} W_i q^i$ de£nes a random distribution on $\mathcal{C}$. In some cases one may be able to make a judicious choice of α if some partial information about the ϕ's are available. For our example we generated 1,000 random values using the Dirichlet distribution with each of the 10 parameter values set equal to 0.1.

In the upper plot of Figure 1 we plotted the posterior expectations of the ϕ's for these 1,000 pairs of values along with the seven extreme points for the prior expectations. The prior expectations are darker and four of these points are clearly visible. They form the lower boundary of $\mathcal{C}(\phi_1,\phi_2)$. Another, (14.03,0.56), is clearly visible as well but maybe hard to identify because the plot is quite small. The other two, (3.82,0.64) and (4.58,0.67), are totally obscured by the posterior expectations. As to be expected the posterior expectations form a smaller set than the prior expectations.

Being able to £nd the extreme points of $\mathcal{C}$ is a powerful tool. In our somewhat limited experience the program we used seems quite good. In one constrained problem we considered in a different context it found over 28,000 extreme points. Using the Dirichlet distribution on the set of weights associated with the extreme points

is a convenient distribution to sample from. These distributions are known as multivariate B-splines and are well studied. See for example Dahmen and Micchelli (1983). If one had a closed form expression for their densities then one could use importance sampling to approximate expectations under other densities. Unfortunately this can only be done in practice for very small problems. See for example Choudhuri (2003).

In practice one would use a much larger sample that 1,000 when studying the posterior expectations. But if the dimension of C gets to large one may not be able to take a large enough sample to get a reasonable approximation. In such cases one can £nd the minimum or maximum of the posterior expectation of a particular ϕ using a random search. The basic idea underlying random searches is well know and is quite simple. See for example Swann (1974).

3 The parameter space is a convex polytope

Here we will consider cases where the parameter space is no longer £nite. We will assume that the parameter is a m dimensional vector, θ and that any possible choice for θ must satisfy

$$A\theta = a \tag{4}$$

where A is a known $r \times m$ matrix and a is a known vector of length r and

$$B\theta \leq b \tag{5}$$

where B is a known $s \times m$ matrix and b is a known vector of length s.

These constraints represent some of the statistician's prior information about θ. The parameter space, Θ, is the set of all θ which satisfy the above two equations. It is a convex polytope in m dimensional Euclidean space with an empty interior.

We begin by assuming that the statistician can select a prior density f over the parameter space to re¤ect the rest of their prior in-

formation. After discussing this case we will consider the situation where the statistician can only determine a family of possible prior densities.

Let ϕ denote some function de£ned on Θ. Then we are interested in £nding

$$\mu = \int_{\Theta} \phi(\theta) h(\theta)\, d\theta \tag{6}$$

approximately. Interesting choices of h include f and the posterior density of θ under f given the data.

For most cases of interest this means employing Markov chain Monte Carlo methods. Here we will use the Metropolis-Hastings algorithm. With the Metropolis-Hastings algorithm one generates dependent observations, $Y_1, Y_2, \ldots$ from a suitable chosen Markov chain with values in Θ and then calculates

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i)$$

In nice situations $\hat{\mu}_n \to \mu$ almost surely.

If the current value of the chain is θ^* then one selects a proposal value for the next possible value in the sequence, θ', according to some probability distribution. Whether or not this new value is used depends on this probability distribution and the value of h at the two points. If $q(u,v)$ denotes the probability of selecting v as the proposal point when u is the current value we let

$$R = \frac{h(v)q(v,u)}{h(u)q(u,v)} \tag{7}$$

and accept the proposal with probability $\min(1,R)$. From this equation we see that the density h needs to be know only up to a constant since the value of R does not depend on this value.

We now explain how this can be done for our problem. Suppose θ^* is our current state which is assumed to lie in the relative interior of Θ. In particular, this means that it yields a strict inequality in all the equations in 5. There is an intuitive two step process by which we can choose the proposal. First, we select a random direction,

d, in Θ. Remember that even though d is a vector of length m it must remain in Θ which is essentially a lower dimensional set. The distribution for choosing d will not depend on the value θ^* Next, we £nd the set of points in Θ which lie either in the direction d or $-d$ from θ^* and choose the proposal at random from this set.

It is important to note that for this scheme if θ' is the proposal then

$$q(\theta^*, \theta') = q(\theta', \theta^*)$$

This follows because if d is the direction from θ^* to θ' then the only way to move from θ^* to θ' is if the directions d or $-d$ were selected in the £rst step. Of course the same is true if we are moving from θ' to θ^*. Clearly, the second step has the same distribution no matter which of the two points was chosen as the initial point.

To implement this scheme we proceed as follows. To get our direction d we choose at random a vector from $\mathcal{S}$, the unit sphere in m dimensional Euclidean space and then project it onto the null space of A. Next, we normalize this vector so that its length is one and denote it by d. Remember that $d \in \mathcal{S}$. Let $\mathcal{S}(A)$ denote the the subset $\mathcal{S}$ consisting of all vectors which can be generated in this way. Since we use the uniform distribution on the surface of $\mathcal{S}$ to choose the £rst vector, the distribution of d must be uniform on $\mathcal{S}(A)$. This follows by symmetry. The probability of d falling in any region of some £xed shape is independent of the location of the region in $\mathcal{S}(A)$.

Let θ^* denote a point which lies in the relative interior of Θ and consider vectors of the form

$$\theta^* + \alpha d \tag{8}$$

where α is a real number. Note if d does not belong to the null space of A this point cannot belong to Θ whenever $\alpha \neq 0$. On the other hand if $Ad = 0$ and α is suf£ciently close to zero then this point will belong to Θ. Hence, α should be selected from the set of possible values that satisfy all the constraints of equation 5. This is a total of s constraints. Each constraint will result in either an upper or lower bound for α. Consider the interval formed by the maximum of these lower bounds to the minimum of these upper bounds. This is the

range of possible values for α for which the vector in equation 8 will belong to Θ. Given a θ^* and a d in this equation then one just selects a value for α from the uniform distribution on its interval of possible values resulting in the proposal

$$\theta' = \theta^* + \alpha d$$

To recap, this is essentially a very simple procedure. Given a current value in the interior of Θ we first pick a random direction in Θ and then find how far we can move either in this or the opposite direction and still remain in Θ. Note that this method of selecting a proposal point does not depend in any way on the function h although from equation 7 the probability of it being accepted does depend on h. When h is the uniform distribution then the proposal is always accepted.

Suppose now the statistician does not have enough prior information to select a single prior density over Θ but can only specify a family of possible priors on Θ. A convenient and often sensible choice for such a family is all possible convex combinations of some finite set of densities. Let $f_1, \ldots, f_n$ denote such a finite set of densities and Π the be the family of all possible convex combinations of the f_i's. For a function ϕ define on Θ let $\Pi(\phi)$ be the range of possible values of the prior expectation of ϕ as the prior ranges over all members of Π. Since for each f_i we can find its ϕ expectation approximately we can find $\Pi(\phi)$ approximately. It is just the interval from the minimum to the maximum of these n prior expectations. Suppose instead we wished to find the joint range of the prior expectations of two functions as the prior ranged over Π, say $\Pi(\phi_1, \phi_2)$. This is easily done since it is just the convex hull of all points of the form $(\int \phi_1 f_i, \int \phi_2 f_i)$ for $i = 1, \ldots, n$. Posterior calculations are handled in exactly the same way since for any prior in Π the posterior is just a convex combination of the n posteriors. Finally, we emphasize the importance of only needing to know any probability density up to a constant since for most of these kind of problems the normalizing constant will be unknown.

Example 2 *We let* $m = 5$ *and supposed* θ *is the parameter for the*

multinomial distribution. We imposed the constraints $\theta_1 \leq \theta_2$, $(\theta_1 + \theta_2)/2 \leq (\theta_3 + \theta_4 + \theta_5)/3$ *and* $\theta_4 \leq \theta_5$ *to get the parameter space* Θ. *We assumed the class of possible priors is all possible convex combinations of three Dirichlet distributions restricted to* Θ. *These were taken to be Dirichlet(2,2,2,2,2), Dirichlet(0.5,0.5,0.5,0.5,0.5) and Dirichlet(0.5,1.0,1.5,2,2.5). The two functions of interest were* $\phi(\theta) = \theta_4 - \theta_3$ *and* $\phi_2(\theta) = \theta_5 - \theta_3$. *To compare the prior and posterior expectations we assumed that a random sample of size 20 had been observed with the observed counts of states 1, 2, 3, 4 and 5 being 2, 3, 4, 3 and 8 respectively.*

Using our methods we found approximately the three extreme points of $\Pi(\phi_1, \phi_2)$ and the corresponding extreme points for the posterior problem. As we noted before it is crucial that the densities on Θ need only be known up to a constant to £nd these integrals approximately. The results are given in the lower plot of Figure 1. The prior expectations are marked by 0 and the posterior expectations are marked by *x*. As to be expected the posterior range is much smaller than the prior range. Using our approach and considering various sets of constraints, families of priors, choices of ϕ_1 and ϕ_2 and hypothetical samples can be useful in helping one select a sensible representation of their prior beliefs for a particular problem.

We end this section by noting that our method of picking a proposal point can be adapted to the random search method we mentioned in section 2. Formally, the two spaces $\mathcal{C}$ and Θ are essentially the same. In the random search algorithm given a point in the interior of $\mathcal{C}$ one needs to be able to choose a point at random from a small neighborhood that contains it. Hence after a direction has been selected at random rather than allowing a move that is as far as possible in either direction one restricts the new point to be no more than ε away in either direction from the current point where $\varepsilon > 0$ is £xed.

4 Discussion

Although constraints seem a natural way to incorporate prior information into an inference problem they have not been widely considered in the statistical literature. The main reason seems to be that they are dif£cult to deal with both theoretically and practically. Betrò and Guglielmi (2000) considered robust Bayesian analysis under moment constraints in a fairly abstract setting and concluded that none of the current algorithms were good enough to be adopted for routine use. Generating random samples from distributions de£ned over bounded subsets of m dimensional Euclidean has been considered in a variety of contexts. Smith (1984) considers the problem of generating independent uniform observations from a bounded region while Belisle, Romeijin and Smith (1993) considers algorithms for generating observations from a general multivariate distribution. They assume that the region of interest is open with a nonempty interior which will not work here. Boender et al. (1991) and Chen and Schmeiser (1993) consider somewhat related problems.

At the present time Markov chain Monte Carlo methods seem to be the best way to handle the types of problems considered in section 3. They come with no guarantees however. If run long enough they will converge to the correct answer but in a given example it can be very dif£cult to know when to stop. When m is large it is impossible to visualize Θ. From our experience, it seems one should select a starting value for θ that is somewhere in the "center" of Θ. In some cases it seems that it is possible for the chain to spend long periods trapped in a corner near the boundary of Θ. If you start in the center then any region of Θ you eventually reach you will also eventually leave. When trying to compute equation 6 approximately it is not necessary to visit every niche and corner of Θ especially those where h puts little weight. But in examples we have studied we have seen that it can take a very long time to reach certain regions very near the boundary. As be noted by many authors when m increases we must deal with the "curse of dimensionality". For a helpful discussion on the convergence of Markov chain Monte Carlo simulations see

Geyer (1992) and Gelman and Rubin (1992).

The methods discussed here not only can help a single statistician evaluate the consequences of their prior assessments but could help a pair of experts resolve possibly con¤icting prior beliefs. It can be used to check how areas of disagreement will effect their inferences. It could help them study how adjustments of some of their beliefs could lead to a merging of opinions. Clearly this is not a theory of how to readjust one beliefs when face with new information but a way to explore the consequences of readjustments of linear constraints. Here we have emphasized exploring jointly the prior or posterior expectations for a pair of functions. The methods work for jointly exploring three and even more functions however convenient graphical representations are no longer possible.

Our simulations were done using *R*. We are preparing a small package that would make it easy for others to implement these methods. Once it is £nished we will submit it to the *R* archives for public distribution. We hope to complete this sometime this year.

Finally, the authors wish to thank Charles Geyer for many helpful discussions.

References

[1] C. J. P. Belisle, H. E. Romeijn and R. L. Smith. Hit-and-Run Algorithms for Generating Multivariate Distributions. *Mathematics of Operations Research*, 18:255-266, 1993.

[2] B. Betrò and A. Gugliemi. Methods for Global Prior Robustness under Generalized Moment Conditions. In *Robust Bayesian Analysis* (D. R. Insua and F. Ruggeri eds.) 273-293, Springer 2000.

[3] C. G. E. Boender, R. J. Caron, J. F. McDonald, A. H. G Rinnooy Kan, H. E. Romejin and R. L. Smith. Shake-and-Bake Algorithms for Generating Uniform Points on the Boundary of Bounded Polyhedra. *Operations Research* 39: 945-954, 1991.

[4] M. Chen and B. Schmeiser. Performance of the Gibbs, Hit-and-Run and Metropolis Samplers. *Journal of Computational and Graphical Statistics*, 2, 251-272, 1993.

[5] N. Choudhuri. Computing Multivariate B-splines: A Simulation Based Approach. Technical Report, Case Western Reserve University, 2003.

[6] W. Dahmen and C. A. Micchelli. Recent Progress in Multivariate Splines. In *Approximation Theory IV* (C. K. Chui, L. L. Schumaker and J. D. Ward, eds.) 17-121. Academic Press, New York 1983.

[7] J. M. Dickey. Convenient Interactive Computing for Coherent Imprecise Prevision Assessment. submitted to ISIPTA '03, 2003.

[8] K. Fuduka. cdd and cddplus Homepage for locating vertices. http://www.cs.mcgill.ca/∼fukuda/soft/cdd_home/cdd.html, 2003.

[9] A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7, 457:472, 1992.

[10] C. J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7, 473-483, 1992.

[11] W. H. Swann. Constrained Optimization by Direct Search. In *Numerical Methods for Constrained Optimization* (P. E. Gill and W. Murray, eds.) 191-217. Academic Press, New York 1974.

[12] R. L. Smith. Ef£cient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions. *Operations Research* 32:1296-1308, 1984.

Radu Lazar is a graduate student in the School of Statistics at the University of Minnesota, Minneapolis, MN 55455, USA. E-mail: lazar@stat.umn.edu

Glen Meeden is with the School of Statistics at the University of Minnesota, Minneapolis, MN 55455, USA. E-mail: glen@stat.umn.edu

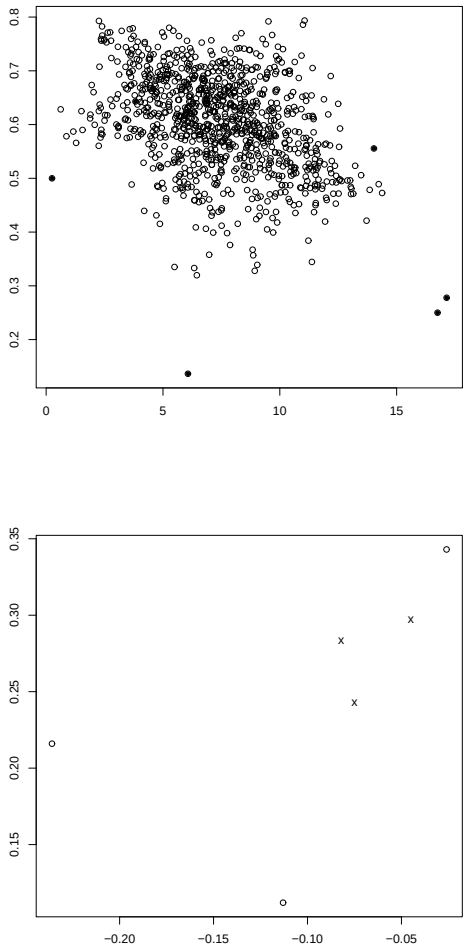

Figure 1: The upper plot contains the seven extreme points for the prior expectations and the plot of 1,000 posterior expectations for the functions ϕ_1 and ϕ_2 in Example 1. The lower plot gives the extreme points for prior (marked by 0) and posterior (marked by x) expectations of (ϕ_1, ϕ_2) for Example 2. In both cases the horizonal and vertical axes are the ϕ_1 and ϕ_2 expectations respectively.