THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #265

Bayesian Inference in Partially Identified Models:

Is the Shape of the Posterior Density Useful?

Paul Gustafson

September 2011

Bayesian Inference in Partially Identified Models: Is the Shape of the Posterior Density Useful?

Paul Gustafson
Department of Statistics
University of British Columbia
gustaf@stat.ubc.ca

September 21, 2011

Abstract

Partially identified models are characterized by the distribution of observables being compatible with a set of values for the target parameter, rather than a single value. From a non-Bayesian point of view, this set, known as the identification region, is the target of inference. In the limit of increasing sample size, this set is revealed to the investigator. On the other hand, a Bayesian obtains a distribution over the identification region. This purports to convey varying plausibility of values across the region. Taking a decision-theoretic view, we investigate the extent to which having a distribution across the identification region is indeed helpful.

Keywords: Bayesian inference; partial identification; posterior distribution.

1 Introduction

Limitations in the form of what variables can be observed may result in a statistical model which is not fully identified. If multiple values of the entire parameter vector give rise to the same distribution of observables but different values of the parameter of inferential interest, then consistent estimation of the target is not possible. On the other hand, data may still be somewhat informative. Let the *identification region* be the set of values for the target that are compatible with a particular distribution of observables. Then the identification region may vary with this distribution, and for a given distribution be a proper subset of the *a priori* possible values for the target. Hence there can still be utility in drawing samples to learn about the distribution of observables.

There is a considerable literature on non-Bayesian approaches to partially identified models. See, for instance, Manski (2003); Imbens and Manski (2004); Romano and Shaikh (2008); Vansteelandt et al. (2006); Zhang (2009); Tamer

(2010). Typically the endeavour is split into two tasks. For a given problem, first one determines the form of the identification region, and this set itself is viewed as the parameter of interest. Then inference is considered as as separate exercise, comprised of estimating the boundaries of the identification region (often the endpoints of an identification interval), and/or reporting a confidence set for the identification region. As a side note, there is an interesting distinction between confidence sets designed to have nominal or better coverage for the true value of the target versus those designed to have nominal or better coverage of the whole identification region. More importantly for present purposes, note that these approaches do not naturally lend themselves to a sense of some target values being more plausible than others in light of the data. Conceptually, if the investigator were handed an infinite number of datapoints, and hence perfect knowledge of the distribution of observables, then the identification region would be reported as 'the answer.'

On the other hand, identification and inference are more integrated under a Bayesian analysis. Based on a sample of size n, the investigator carries out prior-to-posterior updating, yielding a marginal posterior distribution on the target parameter. As n increases, this distribution converges to a non-degenerate distribution with support equal to the identification region. Given an infinite number of datapoints then, inference is summarized by a relative weighting of points in the identification region. Both Liao and Jiang (2010) and Gustafson (2010) suggest this may be a strength of the Bayesian approach, though Moon and Schorfheide (2011) are more circumspect.

Thus there is a fundamental discrepancy between non-Bayesian and Bayesian inference in partially identified models, which differs from the situation in identified models (where the identification region is a single point and therefore does not admit a weighting of its elements). Given this, we seek to understand the utility of the posterior weighting of points in the identification region. If this weighting were completely driven or pre-ordained by the prior distribution, then the difference between the Bayesian and non-Bayesian answers might be argued to be somewhat superficial. We can quickly see, however, that at least in some problems the weighting is not pre-ordained. Some partially identified models, including the two given as examples in this paper, have the property that distinct points in the parameter space can lead to the *same* identification region for the target, but *different* limiting posterior distributions across this region. Thus the situation is nuanced, and warrants investigation.

We investigate the inferential utility in the 'shape' of the posterior distribution by taking a decision-theoretic view. In the large-sample limit, we can compare the posterior distribution as a summary of knowledge about the target versus an ad-hoc choice of distribution over the identification region, in terms of Bayes risk. The difference in Bayes risk can be decomposed into a term representing updating based on knowledge of the identification region only and a term representing the 'extra' information arising because different limiting posterior distributions can correspond to the same identification region. This decomposition is worked out for two examples. One involves a model for imperfect compliance in a randomized trial, while the other deals with inference about a gene-environment interaction when full data cannot be observed but some assumptions can be made.

2 Methodology

Let $\pi(\theta, d)$ denote the joint density of a parameter vector θ and observable data d, as arises from the product of a proper and smooth prior density $\pi(\theta)$ and a statistical model density $\pi(d|\theta)$. Assume that θ comprises a 'scientifically intuitive' parameterization of the model, such that investigators would feel comfortable specifying a prior distribution for θ, as opposed to specifying a prior in some other parameterization. Also assume that the primary inferential interest lies in some scalar aspect of θ, denoted as the estimand $\psi = g(\theta)$. When useful, we write d_n to emphasize observable data comprised of n observations which are independent and identically distributed given θ. Also, we use upper-case D_n and Θ when it is helpful to stress a random variable interpretation of data and parameters, e.g., inside expectations.

Our interest focusses on problems lacking identification, with the distribution of the data depending on θ only through $\phi = s(\theta)$, such that $(D_n|\Phi = \phi)$ constitutes a 'regular' parametric model admitting standard $\sqrt{n}$-consistent estimation of ϕ. Then, if the true parameter values are $\theta = \theta_0$, the large n limit of $(\Theta|D_n)$ is characterized by Φ having a point-mass distribution at $\phi_0 = s(\theta_0)$, combined with the conditional prior distribution for $(\Theta|\Phi = \phi_0)$. We restrict attention to problems where the target ψ is not completely determined by ϕ, so that the large-sample limit of the marginal posterior distribution on ψ will not be a point-mass.

For a finite sample size n, say a family of density functions $h(\cdot;\cdot)$ is used such that $h(\cdot; d_n)$ is a probabilistic estimate or 'forecast' of ψ when data $D_n = d_n$ are observed. The estimation loss incurred can be taken as the entropy loss $-\log h(\psi; d_n)$, i.e., the utility of estimation is the log height of the forecast density at the target value. By averaging the loss across repeated joint realizations of $(\Theta, D_n) \sim \pi$, we obtain the Bayes risk as a summary of estimator h's performance:

$$BR^{(n)}_{\pi,h} \quad = \quad -E_{\pi}\{\log h(\Psi; D_n)\}. \tag{1}$$

It is well known that the choice of h minimizing the Bayes risk is the posterior density of the target with respect to prior, i.e., $h(\psi; d_n) = \pi(\psi|d_n)$.

Because we are studying problems in which the posterior distribution of the target converges to a non-degenerate distribution as the sample size grows, the limiting version of (1) is immediate. Observation of an infinite data set corresponds to knowledge of ϕ, so we are now concerned with a family of density functions of the form $h(\cdot; \phi)$, and the corresponding Bayes risk

$$BR^{\infty}_{\pi,h} \quad = \quad -E_{\pi}\{\log h(\Psi; \Phi)\}. \tag{2}$$

Bear in mind here that ψ and ϕ are both functions of θ, and the expectation is with respect to $\Theta \sim \pi$. The same standard argument applies to verify that

(1) is minimized by $h(\psi;\phi) = \pi(\psi|\phi)$, which is the large-sample limit of the posterior density on the target. Henceforth we refer to $\pi(\psi|\phi)$ as the density characterizing the *limiting posterior distribution* (LPD). For this choice of h we will denote the Bayes risk as $BR^{\infty}_{\pi,B}$ (here 'B' is for Bayes). Note that $\exp(-BR^{\infty}_{\pi,B})$ can be interpreted as the typical height of the limiting posterior density at the target, across a sequence of realizations with different values of θ generated according to π. This provides a summary of how much learning about the parameter is taking place despite the lack of identification.

Much of the non-Bayesian literature on partial identification treats the identification region itself as the target of inference, with the consequent notion that knowledge of the region is all that could be gleaned upon observation of an infinite-sized dataset. Consider a situation where every θ gives rise to an interval of positive but finite length as the identification region. The endpoints of the interval can necessarily only depend on θ through ϕ, so we write the interval as $\phi^* = \{\phi^*_L(\phi), \phi^*_R(\phi)\}$. Thus it might be viewed that knowledge of ϕ^* is just as good as knowledge of ϕ, even if the map from ϕ to ϕ^* is not invertible.

We quantify the utility of the limiting Bayesian distribution over the identification interval, $\pi(\psi|\phi)$, by comparing its Bayes risk to that of a probabilistic estimate depending only on ϕ^*, i.e. a family of density functions $h(\cdot;\phi^*)$ indexed by ϕ^*. One possible h, jibing with the notion of learning only the identification region, is the uniform density over the identification interval. We denote its Bayes risk as $B^{\infty}_{\pi,U}$ (here 'U' is for uniform).

A uniform distribution across the identification interval doesn't take advantage of knowledge of π. That is, π is used in the evaluation of estimator performance, but not in the construction of the estimator. A naive way to access this information would be to combine the marginal prior on the target with knowledge of the identification region via truncation:

$$h(\psi;\phi^*) \quad = \quad \frac{\pi(\psi)I\{\phi^*_L < \psi < \phi^*_R\}}{\int_{\phi^*_L}^{\phi^*_R} \pi(s)ds}.$$

We denote the Bayes risk of this limiting estimator as $B^{\infty}_{\pi,T}$ (here T is for 'truncated').

While truncation of the marginal prior is intuitive, it does not correspond to optimal use of the information encoded by ϕ^*. Considering only h of the form $h(\cdot;\phi^*)$, it is immediate that (2) is minimized by $h(\psi;\phi^*) = \pi(\psi|\phi^*)$. We denote the corresponding Bayes risk as $BR^{\infty}_{\pi,C}$, where 'C' indicates possibly 'coarsened' dependence on ϕ^* rather than ϕ. Clearly this quantity describes the large n limit of performance if one extracts from the data only the information about the identification region.

Now we are in a position to try to understand the worth of the shape of the LPD across the identification interval. Say a uniform distribution across the interval is our point of reference. Then we can write the gain of the LPD relative to this reference point as

$$BR^{\infty}_{\pi,U} - BR^{\infty}_{\pi,B} \quad = \quad (BR^{\infty}_{\pi,U} - BR^{\infty}_{\pi,C}) + (BR^{\infty}_{\pi,C} - BR^{\infty}_{\pi,B}), \qquad (3)$$

where both terms on the right-hand side are nonnegative. The first of these terms reflects the value of Bayesian updating based only on information about the identification region, relative to an ad-hoc use of this information. The second term represents the value of using all the information in the data, not just the information about the identification region. Put another way, the second term reflects information 'left on the table' by supposing that the data can only speak to the location of the identification interval. This term is of particular interest, since non-Bayesian approaches to partially identified models are predicated on the idea that knowledge of the identification region is indeed all that can be obtained in the limit of infinite sample size. Yet another interpretation is that the second term reflects the utility of the fact that multiple θ values can lead to the *same* identification region but *different* limiting posterior distributions over this region. In the special case that the map from ϕ to ϕ^* is invertible, there is only one limiting distribution corresponding to a given identification interval, and the second term in (3) is zero.

Of course the analogous decomposition could also be applied starting with our other ad-hoc estimator, namely the prior truncated to the identification region. In this case,

$$BR^{\infty}_{\pi,T} - BR^{\infty}_{\pi,B} \quad = \quad (BR^{\infty}_{\pi,T} - BR^{\infty}_{\pi,C}) + (BR^{\infty}_{\pi,C} - BR^{\infty}_{\pi,B}). \tag{4}$$

Note that both terms in (4) are left invariant if an invertible function of ψ is taken as the target parameter instead of ψ itself. Conversely, this is not true of the first term in (3). In the examples of the next two sections we refer to the uniform distribution over the identification interval and the prior truncated to the region as 'pre-ordained' limiting estimators, in the sense that the shape of the density is pre-ordained and the data speak only to the location of the identification region.

3 Example: Imperfect Compliance in a Randomized Trial

Here we consider a version of the imperfect compliance model with binary variables considered by various authors, including Chickering and Pearl (1996), Imbens and Rubin (1997), Pearl (2000, Ch. 8), and Richardson et al. (2011). Trial subjects are randomly sampled from a population comprised of never-takers, always-takers, and compliers, in unknown proportions ω_{NT}, ω_{AT}, and $\omega_{CO} = 1 - \omega_{NT} - \omega_{AT}$ respectively. Each subject is randomly assigned to either control or treatment. As the labels suggest, never-takers will not take treatment regardless of their assignment, always-takers will take treatment regardless of their assignment, and compliers will follow their assignment. We exclude the possibility of defiers in the population, though the general version of the problem allows for them.

Assume that a patient's binary response is Y_0 if treatment is not taken, and Y_1 if treatment is taken, regardless of treatment assignment. Then a subject's outcome is $Y = (1-X)Y_0 + XY_1$, where X indicates reception of treatment, whereas Z indicates assignment to treatment. For compliance type $C \in \{NT, AT, CO\}$, let $\gamma_{C,i}$ by the mean of Y_i amongst the sub-population of that type. We consider inference about the population average causal effect (ACE), given as

$$\psi \quad = \quad \omega_{NT}(\gamma_{NT,1} - \gamma_{NT,0}) + \omega_{AT}(\gamma_{AT,1} - \gamma_{AT,0}) + \omega_{CO}(\gamma_{CO,1} - \gamma_{CO,0}),$$

based on a prior distribution under which $\omega \sim \text{Dirichlet}(1,1,1)$ and independently the components of γ follow $\text{Unif}(0,1)$ distributions.

Observable data reveal the $(Y, X|Z)$ distribution, which depends on the unknown parameters $\theta = (\omega, \gamma)$ only through $\phi = (\omega, \gamma_{NT,0}, \gamma_{AT,1}, \gamma_{CO,0}, \gamma_{CO,1})$. The invertible map from ϕ to $(Y, X|Z)$ cell probabilities is given via

$$\begin{aligned}
pr(X=1|Z=0) &= \omega_{AT} \\
pr(X=1, Y=1|Z=0) &= \omega_{AT}\gamma_{AT,1} \\
pr(X=0, Y=1|Z=0) &= \omega_{CO}\gamma_{CO,0} + \omega_{NT}\gamma_{NT,0} \\
pr(X=0|Z=1) &= \omega_{NT} \\
pr(X=0, Y=1|Z=1) &= \omega_{NT}\gamma_{NT,0} \\
pr(X=1, Y=1|Z=1) &= \omega_{CO}\gamma_{CO,1} + \omega_{AT}\gamma_{AT,1}.
\end{aligned}$$

It is unsurprising that the parameters absent from ϕ, namely $\gamma_{NT,1}$ and $\gamma_{AT,0}$, are the intuitively unestimable quantities: the mean outcomes for never-takers who take treatment and for always-takers who don't take treatment.

To describe the LPD, let

$$\begin{aligned}
a(\phi) &= \omega_{CO}(\gamma_{CO,1} - \gamma_{CO,0}) + \omega_{NT}(1/2 - \gamma_{NT,0}) + \omega_{AT}(\gamma_{AT,1} - 1/2), \\
b(\phi) &= (\omega_{NT} + \omega_{AT})/2, \\
c(\phi) &= |\omega_{NT} - \omega_{AT}|/2.
\end{aligned}$$

Gustafson (2011) shows that the identification interval for the target parameter is $(\phi^*_L, \phi^*_R) = \{a(\phi) - b(\phi), a(\phi) + b(\phi)\}$, with the LPD having a trapezoidal-shaped density over this interval. The 'top' of the density spans $a(\phi) \pm c(\phi)$, and commensurately the height of the density is $\{b(\phi) + c(\phi)\}^{-1}$.

In this problem the map from ϕ to ϕ^* is not invertible, and multiple values of ϕ can lead to the same identification region but different limiting distributions across the region. This is illustrated in Figure 1. While the form of the LPD $\pi(\psi|\phi)$ is mathematically very simple, determination of the coarsened LPD $\pi(\psi|\phi^*)$ is somewhat more involved - see Appendix A for details.

The various Bayes risks, which are computed simply by averaging across a large number of Monte Carlo draws of $\Theta \sim \pi$, appear in Table 1. Note that we have 'simulation-significant' evidence that $BR^{\infty}_{\pi,\pi} < BR^{\infty}_{\pi,C}$, in line with the theory. The difference between these two Bayes risks corresponds to the LPD being typically 2.4% higher than the coarsened LPD, in terms of density at the true value of the target. Thus the data do contain a little information beyond

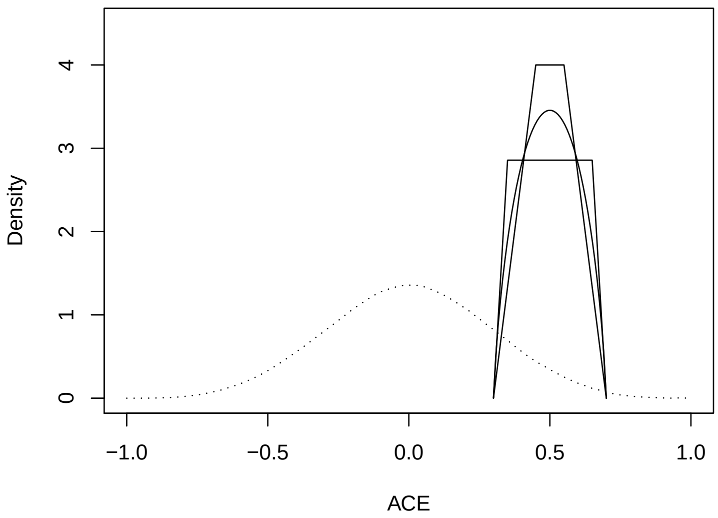

Figure 1: Prior (dotted curve) and limiting posterior densities for the ACE. In all cases $\omega_{CO} = 0.6$ and $a(\phi) = 0.5$, hence the identification region is 0.5 ± 0.2. The less (more) concentrated flat-topped density corresponds to the LPD when $\omega_{AT} = 0.05$ ($\omega_{AT} = 0.15$), while the remaining curve is the coarsened LPD for this identification region.

Table 1: Bayes risks in the treatment compliance example, computed as empirical averages across 10,000 Monte Carlo realizations of $\Theta \sim \pi$. Simulation standard errors are given in parentheses.

$-BR^{\infty}_{\pi,\pi}$	0.6127 (0.0063)
$BR^{\infty}_{\pi,C} - BR^{\infty}_{\pi,\pi}$	0.0235 (0.0023)
$BR^{\infty}_{\pi,T} - BR^{\infty}_{\pi,C}$	0.1409 (0.0052)
$BR^{\infty}_{\pi,U} - BR^{\infty}_{\pi,C}$	0.0880 (0.0036)

that which reveals the identification region. The gaps in Bayes risk between the pre-ordained estimators and the coarsened LPD is quite substantial, which speaks to the utility of the shape of the limiting posterior distribution. For instance, the ratio of the coarsened LPD at the target to the truncated prior density at the target has a typical value of $\exp(0.14) \approx 1.15$.

4 Example: Inferring Gene-Environment Interaction

Consider binary disease status Y, binary environmental exposure X, and binary genotype G. As a variant of a problem studied by Gustafson (2010) and Gustafson and Burstyn (2011), the task is to infer the $(Y|X,G)$ relationship when only (Y,G) data are available, but certain assumptions can be made. The first of these is the Mendelian randomization assumption of independence between X and G in the source population. Second, the disease risk amongst the unexposed is assumed to not vary by genotype, i.e., any impact of genotype is only via modification of the exposure effect, a so-called gene-environment interaction. Third, while (Y,X,G) data are not available, information about the X prevalence in the population is presumed to be available. So the problem can be viewed as inferring a property of the joint (Y,G,X) distribution from information about the (Y,G) and X marginals. Specifically, let the inferential target be $Pr(Y=1|X=1,G=1) - Pr(Y=1|X=0,G=1)$, the risk difference associated with exposure amongst those with genotype $G=1$.

To parameterize this problem, let $\mu_0 = Pr(Y=1|X=0) = Pr(Y=1|X=0,G=g)$, for $g=0,1$, and let $\mu_{1g} = Pr(Y=1|X=1,G=g)$, for $g=0,1$. We exemplify with a prior distribution under which $(\mu_0,\mu_{10},\mu_{11})$ are independent and identically distributed as Beta(k,k), using $b_k()$ to denote the corresponding density function. Also, let $r = Pr(X=1)$ be known.

To make clear the partial identification at play, consider a reparameterization from $\theta = (\mu_0,\mu_{10},\mu_{11})$ to (ϕ_0,ϕ_1,ψ), where

$$\begin{aligned}\phi_g &= Pr(Y=1|G=g) \\ &= (1-r)\mu_0 + r\mu_{1g},\end{aligned}$$

for $g = 0, 1$, while

$$\psi \;=\; \mu_{11} - \mu_0$$

is the target of inferential interest. Upon computing the Jacobian of this linear reparameterization, the prior density in the new parameterization is

$$\begin{aligned} \pi(\phi_0, \phi_1, \psi) \;=\; & r^{-1} b_k(\phi_1 - r\psi) b_k(\phi_1 + (1-r)\psi) \;\times \\ & b_k(r^{-1}(\phi_0 - (1-r)\phi_1 + r(1-r)\psi)). \end{aligned} \tag{5}$$

Clearly $(Y|G)$ data are completely informative about $\phi = (\phi_0, \phi_1)$, while ψ is absent from the likelihood. Thus the LPD on the target is the conditional for $(\Psi|\Phi_0, \Phi_1)$ implied by (5), with conditioning on the true value of ϕ. At least up to a normalizing constant, we can 'read off' the conditional prior density simply by viewing (5) as a function of ψ, with ϕ fixed. The support of this conditional density, or equivalently the identification interval for the target, has endpoints:

$$\phi_L^* \;=\; -\min\left\{(1-r)^{-1}\phi_1, r^{-1}(1-\phi_1), r^{-1}((1-r)^{-1}\phi_0 - \phi_1)\right\}, \tag{6}$$

$$\phi_R^* \;=\; \min\left\{r^{-1}\phi_1, (1-r)^{-1}(1-\phi_1), (1-r)^{-1} + r^{-1}(\phi_1 - (1-r)^{-1}\phi_0)\right\} \tag{7}$$

In the special case that the hyperparameter $k = 1$ is used, it is immediate that for every ϕ, $\pi(\psi|\phi)$ is a uniform density on (ϕ_L^*, ϕ_R^*), so different parameter values that give rise to the same identification region also give rise to the same LPD. Hence, for every ϕ^*, $\pi(\psi|\phi^*)$ is also a uniform density on the identification region, and $BR^{\infty}_{\pi,C} = BR^{\infty}_{\pi,B}$.

For $k \neq 1$, the situation is more nuanced. In Appendix B we prove that for every value of ϕ^* there is either (i) two distinct point solutions to $h(\phi) = \phi^*$, which we denote as ϕ^A, ϕ^B, or, (ii), a 'line segment' of solutions of the form $\{\phi : \phi_{0L}^A < \phi_0 < \phi_{0R}^A, \phi_1 = \phi_1^A\}$ plus one further point solution ϕ^B. Consequently, in case (i),

$$\pi(\psi|\phi^*) \;=\; (1-w)\pi(\psi|\phi = \phi^A) + \; w\pi(\psi|\phi = \phi^B),$$

where $w = \pi(\phi^B)/\{\pi(\phi^A) + \pi(\phi^B)\}$. In case (ii),

$$\pi(\psi|\phi^*) \;\propto\; \int_{\phi_{0L}^A}^{\phi_{0R}^A} \pi(\psi|\phi = (s, \phi_1^A))\pi(s, \phi_1^A)ds. \tag{8}$$

Note that as one of infinitely many solutions, the further point solution ϕ^B does not contribute to (8).

In the $k = 2$ case, Figure 2 compares the four limiting estimators to both the true value of the target and the marginal prior density of the target, for some selected values of θ. Note that the LPD and the coarsened LPD are virtually indistinguishable in each case, and they are quite different from the truncated limiting estimate and the uniform limiting estimate.

As in the previous example, the various Bayes risks are computed by averaging across a large number of draws of $\Theta \sim \pi$, with results reported in Table

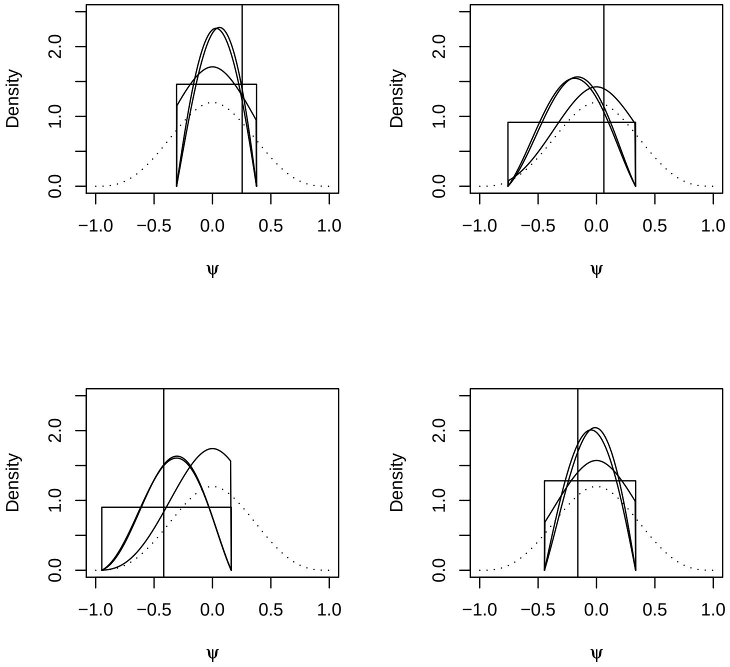

Figure 2: Limiting estimates compared to the target value for four draws of $\Theta \sim \pi$. In all cases the LPD and the coarsened LPD are nearly indistinguishable by eye. The estimator based on truncation is distinguished by its mode at zero. The uniform density over the identification interval is also shown. For reference, the marginal prior density is shown (dotted curve).

Table 2: Bayes risks in the gene-environment example, computed as empirical averages across 10,000 Monte Carlo realizations of $\Theta \sim \pi$. Simulation standard errors are given in parentheses.

-$BR^{\infty}_{\pi,\pi}$	0.2680 (0.0055)
$BR^{\infty}_{\pi,C} - BR^{\infty}_{\pi,\pi}$	0.00293 (0.00068)
$BR^{\infty}_{\pi,T} - BR^{\infty}_{\pi,C}$	0.1540 (0.0046)
$BR^{\infty}_{\pi,U} - BR^{\infty}_{\pi,C}$	0.1892 (0.0046)

2. We indeed have 'simulation significance' to attest to $BR^{\infty}_{\pi,\pi} < BR^{\infty}_{\pi,C}$, in line with the theory. However, the difference between the two Bayes risks is so small as to be negligible in any practical sense. That is, an infinite-sized dataset is seen to contain only a tiny amount of information superseding that used to determine the identification region.

On the other hand, the Bayes risk of the coarsened LPD is substantially lower than that of the two ad-hoc estimators, pointing to a substantial utility in the shape of the Bayesian posterior over the identification interval. In blunt terms, if we choose to average with respect to π when evaluating the performance of the estimator, then we gain a lot by using π as a prior distribution in the construction of a Bayesian estimator.

5 Robustness

Of course the argument for the optimality of any Bayesian procedure relies on the use of the same distribution over the parameter space as both *nature's* prior distribution used to average the expected risk across the parameter space and the *investigator's* prior distribution used to determine the posterior distribution upon receipt of data. Thus it is of interest to see to what extent the performance of the Bayesian procedure degrades as nature's prior and the investigator's prior deviate from one another.

We retain $\pi(\theta)$ as notation for the investigator's prior, but consider what happens when nature's prior is $\pi^*(\theta|\lambda)$ for some choice of λ. We assume the class of possible choices for nature's prior is centered around the investigator's prior, i.e., $\pi^*(\theta|0) = \pi(\theta)$. Specifically we look at the comparison between the LPD and the uniform distribution over the identification region, as the investigator's prior stays fixed but nature's prior moves away from it. Let

$$t(\lambda) \quad = \quad E^*_{\lambda} \left\{ \log \pi(\Psi|\Phi) + \log \left(\Phi^*_R - \Phi^*_L \right) \right\}, \tag{9}$$

where the expectation is with respect to $\pi^*(\theta|\lambda)$. Clearly then $t(0) = BR^{\infty}_{\pi,U} - BR^{\infty}_{\pi,\pi} \geq 0$, and the magnitude of λ required to make $t(\lambda) < 0$ reflects the 'stability' of the usefulness of the shape of the LPD, compared to a uniform distribution over the identification interval.

When λ has more than one component, it may become complicated to evaluate (9) in many different directions away from $\lambda = 0$. Thus we propose computing the gradient

$$t'(0) \quad = \quad E_{\pi}\left[s(\Theta)\left\{\log \pi(\Psi|\Phi) + \log\left(\Phi_R^* - \Phi_L^*\right)\right\}\right], \tag{10}$$

where $s(\theta) = \partial \log \pi^*(\theta|\lambda)/\partial\lambda|_{\lambda=0}$. Then evaluating (9) for values of λ proportional to this gradient corresponds to looking along the direction in which (9) changes most rapidly with λ, locally at zero.

Returning to the compliance example of Section 3, we consider nature's prior distribution to be $\omega \sim \text{Dirichlet}(1+\lambda_1, 1+\lambda_2, 1+\lambda_3)$, whereas the investigator's prior is simply Dirichlet$(1,1,1)$, as before. Both priors use a uniform distribution on γ. Numerical evaluation of (10) indicates that $t'(0) \propto (0,1,1)'$. Thus we focus attention on the case that nature's prior is Dirichlet$(1, 1+\lambda, 1+\lambda)$, for a scalar value of λ. For selected values of λ, $t(\lambda)$ is given in Figure 3. We see that the advantage of the Bayesian procedure is maintained even when the discrepancy between nature's prior and the investigator's prior is given by $\lambda = -0.9$. This suggests considerable robustness, since in practical terms the Dirichlet$(1, 0.1, 0.1)$ distribution is fairly extreme and far from Dirichlet$(1,1,1)$. In particular, this distribution puts considerable weight on extremely small values of ω_{NT} and ω_{AT}.

6 Discussion

Theoretically, the answer to the question posed in the title of this paper is clear: the shape of the posterior density in partially identified models *is* useful. More specifically, if we choose to measure performance in terms of average log density at the target value, where this average is with respect to a distribution π over parameter values, then the posterior distribution of the target arising from π as a prior distribution is optimal. Since the posterior on the target has a non-degenerate large-sample limit, this optimality carries over directly to the limiting case. In both examples considered, the limiting posterior distribution exhibited a substantial advantage over ad-hoc distributions, such as a uniform distribution over the identification interval or the prior for the target truncated to this interval. In particular, a tendency for the posterior density to be 10% to 15% higher at the target than the ad-hoc densities was seen. This speaks to jointly inferring the identification interval and the relative plausibility of values within the interval as being worthwhile.

Additionally, the gain of the posterior density compared to an ad-hoc density was seen to partition intuitively, into a component based on optimal use of information about the identification region and a component based on extra information beyond knowledge of the identification region. In both examples, and particularly in the second example, the latter component is small compared to the former. In a practical sense then, the 'extra' information is not crucial. In a theoretical and conceptual sense, however, it gives pause for thought. The common intuition, particularly in a non-Bayesian sense, is that the identification

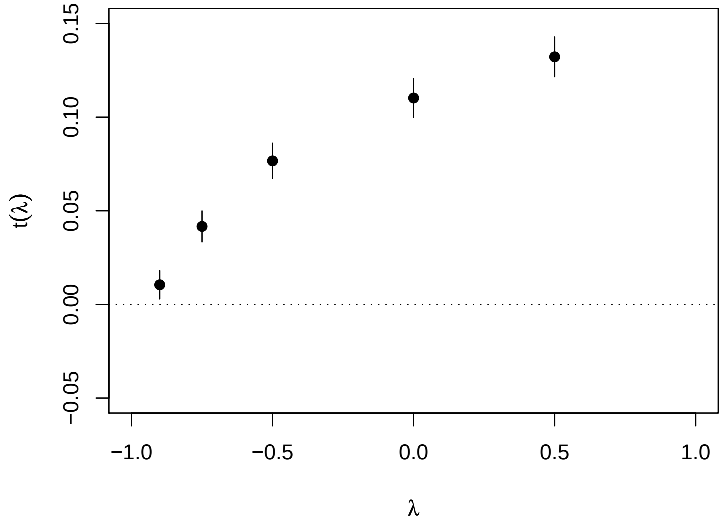

Figure 3: Robustness of the LPD when nature's prior is $\omega \sim$ Dirichlet$(1, 1 + \lambda, 1 + \lambda)$ and the investigator's prior is $\omega \sim$ Dirichlet$(1, 1, 1)$. The difference in Bayes risk, $t(\lambda) = BR^{\infty}_{\pi^*,U} - BR^{\infty}_{\pi^*,\pi}$, is given as a function of λ. The vertical bars are 95% simulation confidence intervals based on the 5,000 Monte Carlo realizations from nature's prior.

region for the target *is* the object of inference, and that knowledge of the region is all that can be obtained in the limit of increasing sample size. The results in this paper give a perspective from which this is not correct.

Appendix A

To determine the coarsened LPD in the compliance model, note that we can write $\psi = \mu + \epsilon$, where $\mu = a(\phi)$, while

$$\epsilon \quad = \quad w_{NT}(\gamma_{NT,1} - 1/2) + w_{AT}(1/2 - \gamma_{AT,0}).$$

Thus the coarsened LPD will be distributed as the conditional prior density $\pi(\psi|\mu, \omega_{CO})$, and it suffices to determine the conditional density of $\pi(\epsilon|\mu, w_{CO})$, which in turn can be determined from $\pi(\mu, \epsilon|w_{CO})$. Defining $\lambda = w_{NT}/(1 - w_{CO})$, it is easy to verify that

$$\pi(\mu, \epsilon|w_{CO}) \quad = \quad \int \pi(\mu|\lambda, w_{CO})\pi(\epsilon|\lambda, w_{CO}))\pi(\lambda)d\lambda.$$

This holds since, with λ and w_{CO} fixed, μ and ϵ depend on disjoint subvectors of γ, whose elements are a priori independent of one another.

Thus the task is reduced to evaluating the conditional prior densities $\pi(\mu|\lambda, w_{CO}) = \pi(\mu|\omega)$ and $\pi(\epsilon|\lambda, w_{CO}) = \pi(\epsilon|\omega)$. Toward this, let $g_s()$ denote the trapezoidal density function of $s(U_1 - 1/2) + (1 - s)(U_2 - 1/2)$, when U_1, U_2 are independent and identically distributed as Unif$(0, 1)$. Then the $(\mu|\omega)$ conditional has a stochastic representation as

$$\mu \quad = \quad 2\omega_{CO}Z_1 + (1 - \omega_{CO})Z_2,$$

where Z_1 and Z_2 are independent with $Z_1 \sim g_{0.5}$, and $Z_2 \sim g_s$ with $s = w_{NT}/(1 - w_{CO})$. Thus the $(\mu|\omega)$ conditional density can be computed exactly via convolution of $g_{0.5}$ and g_s, where the integration is straightforward since these are piecewise linear densities. The evaluation of $\pi(\epsilon|\omega)$ is simpler since convolution is not involved. Particularly, $\pi(\epsilon|\omega) = (1 - \omega_{CO})^{-1}g_s(\epsilon/(1 - \omega_{CO}))$, where again $s = w_{NT}/(1 - w_{CO})$.

Appendix B

Let h be the map from ϕ to ϕ^*. For a given c^* in the image of h, we need to characterize solutions to $h(\phi) = c^*$. Note that the domain of h is the subset of the unit square U given by $S = \{\phi \in U : |\phi_0 - \phi_1| < r\}$. The form of (6) and (7) is such that S can be partitioned as $S = A \cup B \cup C$ as depicted in the left panel of Figure 4, with ϕ^*_L being continuous and piecewise-linear on these subsets. Similarly, $S = D \cup E \cup F$ as in the right panel, with ϕ^*_U being linear on these partition sets. The two dotted reference lines on both panels are the $\phi^*_L = 0$ and $\phi^*_U = 0$ level sets, with $\phi^*_L > 0$ above the upper reference line and

$\phi_U^* < 0$ below the lower reference line. Let $S_1 \subset S$ be the region between the reference lines, for which the identification interval crosses zero. Note that the gradient of ϕ_L^* points straight up on B and straight down on A. Thus a level set for a negative value of ϕ_L^* has an 'open parallelogram' shape, as exemplified in the left panel of Figure 4. We can then speak unambiguously of the 'bottom,' 'spine,' and 'top' of such a level set. In contrast, a level set for a positive value of ϕ_L^* corresponds to a line parallel to and above the upper reference line. A 'mirror-image' situation applies to ϕ_U^*, as depicted in the right panel of the figure.

Let $\tilde{\phi}$ be one solution to $h(\phi) = c^*$. (If it is helpful, one can think of $\tilde{\phi}$ as the 'true' value of ϕ.) Then we have three possible cases.

Case 1. Say that $\tilde{\phi} \in S - S_1$, i.e., the identification region is to one side of zero. Without loss of generality, say $\tilde{\phi}$ lies above the upper reference line. Then ϕ_L^* remains constant along the line through $\tilde{\phi}$ which is parallel to the upper reference line. Along this line, ϕ_U^* takes the value one at the boundary between D and E, decreasing linearly from here in both directions. Moreover, it is simply verified that ϕ_U^* has a common value at both intersections of this line with the boundary of S. Therefore, there must be exactly two point solutions to $h(\phi) = c^*$ in total.

Case 2. Say that $\tilde{\phi} \in S_1 \cap B^C \cap E^C$. By inspection, it must be that either $\tilde{\phi} \in A \cap F \cap S_1$ or $\tilde{\phi} \in D \cap C \cap S_1$. Without loss of generality, assume the former. Then the base of the level set for ϕ_L^* intersects the spine of the level set for ϕ_U^* at $\tilde{\phi}$. Given this, exactly one further solution is generated, as either the spine extends up far enough to hit the top of the level set for ϕ_L^*, or, failing this, the top of the level set for ϕ_U^* hits the spine for ϕ_L^*.

Case 3. Say that $\tilde{\phi} \in S_1 \cap (B \cup E)$. Without loss of generality, say that $\tilde{\phi}$ is in B rather than E. Then, intersecting the tops of the level sets for both ϕ_L^* and ϕ_U^* gives a horizontal line segment of solutions of the form $\phi_0 \in (1 - r, \tilde{\phi}_1)$, $\phi_1 = \tilde{\phi}_1$. We can also see from the shape of the level sets that there will be an additional point solution somewhere to the 'southwest' of B, where the the higher of the two bases of the two level sets crosses the spine of the other.

As claimed then, for a given c^* in the image of h, either there are two point solutions to $h(\phi) = c^*$, or one horizontal line segment of solutions plus an additional point solution.

References

Chickering, D. and Pearl, J. (1996). A clinician's tool for analyzing noncompliance. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96),* Portland, OR, volume 2, pages 1269–1276.

Gustafson, P. (2010). Bayesian inference for partially identified models. *International Journal of Biostatistics* **6,** issue 2 article 17.

Gustafson, P. (2011). Comment on 'Transparent parameterizations of models for potential outcomes,' by Richardson, Evans, and Robins. In Bernardo,

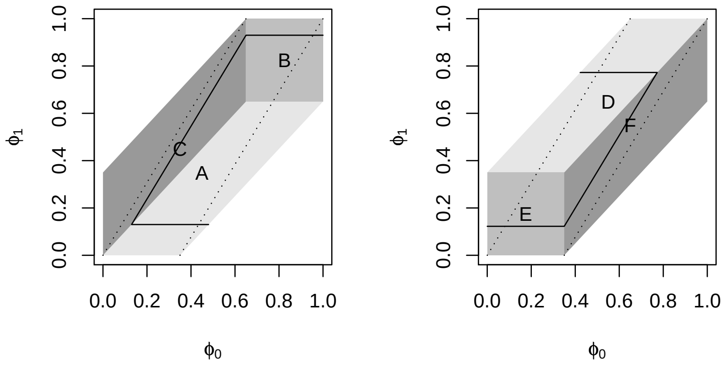

Figure 4: The partition of S into sets on which ϕ_L^* is piecewise linear (left panel) and ϕ_U^* is piecewise linear (right panel), when $r = 0.35$. The upper and lower dotted reference lines appearing on both panels correspond to $\phi_L^* = 0$ and $\phi_R^* = 0$. The level set for $\phi_L^* = -0.2$ is indicated on the left panel and the level set for $\phi_R^* = 0.35$ on the right panel.

J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics 9*, page in press. Oxford University Press.

Gustafson, P. and Burstyn, I. (2011). "Bayesian inference of gene-environment interaction from incomplete data: what happens when information on environment is disjoint from data on gene and disease? *Statistics in Medicine* **30,** 877–889.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72,** 1845–1857.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* **25,** 305–327.

Liao, Y. and Jiang, W. (2010). Bayesian analysis in moment inequality models. *Annals of Statistics* **38,** 275–316.

Manski, C. F. (2003). *Partial Identification of Probability Distributions.* Springer.

Moon, H. R. and Schorfheide, F. (2011). Bayesian and frequentist inference in partially identified models. *Econometrica, to appear* .

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

Richardson, T. S., Evans, R. J., and Robins, J. M. (2011). Transparent parameterizations of models for potential outcomes. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics 9*, page in press. Oxford University Press.

Romano, J. P. and Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference* **138,** 2786–2807.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics* **2,** 167–195.

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica* **16,** 953–979.

Zhang, Z. (2009). Likelihood-based confidence sets for partially identified parameters. *Journal of Statistical Planning and Inference* **139,** 696–710.