Electronic Journal of Statistics
Vol. 8 (2014) 476–496
ISSN: 1935-7524
DOI: 10.1214/14-EJS891

Bayesian inference in partially identified models: Is the shape of the posterior distribution useful?

Paul Gustafson*

Department of Statistics, University of British Columbia
e-mail: gustaf@stat.ubc.ca

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. This set is often referred to as an *identification region*. From a non-Bayesian point of view, the identification region is the object revealed to the investigator in the limit of increasing sample size. Conversely, a Bayesian analysis provides the identification region plus the limiting posterior distribution over this 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.

MSC 2010 subject classifications: Primary 60K35, 62F15; secondary 62F12.
Keywords and phrases: Bayesian inference, partial identification, posterior distribution.

Received September 2013.

1. Introduction

1.1. Partial identification

Limitations in terms of what variables can be observed, and how well they can be measured, can result in a statistical model which is nonidentified. That is, multiple values of the parameters can give rise to the same distribution of observables. Say the statistical model at hand is parameterized by $\theta \in \Theta$, where θ has p components. If the likelihood function depends on θ only through $\phi = s(\theta)$ having $q < p$ components, for some non-injective function $s()$, then the model is nonidentified. In this paper we consider situations where the model for the data given ϕ obeys standard asymptotic regularity conditions, so that $\sqrt{n}$-consistent estimation of ϕ is possible. We also presume that interest focusses on a scalar inferential target, denoted $\psi = g(\theta)$.

In what follows, we generically use square brackets to denote taking the image of a set under a function, as opposed to simply evaluating the function at a point in its domain. The *identification region* for the target parameter is defined as

*Research supported by the Natural Sciences and Engineering Research Council of Canada.

$I(\phi) = g[\{\theta \in \Theta : s(\theta) = \phi\}]$. Intuitively, say the true parameter values giving rise to the data are $\theta = \theta_0$, with $\phi_0 = s(\theta_0)$. In the large-sample limit the data reveal the value of ϕ_0. Thus the corresponding identification region $I(\phi_0)$ is all values of the target that remain compatible with the data in this limit.

For simplicity of exposition, and without very much loss of generality, we restrict our interest to models under which the identification region is in fact guaranteed to be an interval of finite length, i.e., $I(\phi)$ is an interval for all $\phi \in s[\Theta]$. More fundamentally, we consider only the sub-class of nonidentified models and choices of target parameter for which the target is *partially identified* in the following sense. By construction, for every $\theta \in \Theta$, $g[\{\theta\}] \subseteq I(s(\theta)) \subseteq g[\Theta]$. We say the target is partially identified if $g[\{\theta\}] \subsetneq I(s(\theta)) \subsetneq g[\Theta]$, for at least one $\theta \in \Theta$. Note that this corresponds very literally to a sense of partial information. For a sequence of data arising under such a θ, at least one *a priori* plausible value of the target is ruled out as the data accumulate, while at least one incorrect value of the target remains plausible.

As a very simple example of a partially identified model, say that interest lies in the population prevalence μ of a binary trait Y. However, only Y^*, a binary surrogate for Y, is observable. Furthermore, say the surrogate is known to have perfect specificity, i.e., $Pr(Y^* = 0|Y = 0) = 1$. However, the sensitivity $\eta = Pr(Y^* = 1|Y = 1)$ is only known to exceed a bound b. So the problem is parameterized by $\theta = (\mu, \eta) \in (0, 1)\times(b, 1)$, with target parameter $\psi = g(\mu, \eta) = \mu$. The likelihood function clearly depends only on $\phi = s(\mu, \eta) = Pr(Y^* = 1) = \mu\eta$. Thus the identification interval is $I(\phi) = (\phi, \min\{\phi/b, 1\})$. Note that even this simple problem has a typical feature of partial identification: depending on where we are in the parameter space, we learn less or more about the target. For instance, say that $b = 0.8$. If it happens to be that $(\mu, \eta) = (0.75, 0.81)$, then, as data accumulate, we learn that $\mu \in I(\phi) = (0.608, 0.760)$. But if it happens to be that $(\mu, \eta) = (0.95, 0.99)$, then we draw the much sharper inference that $\mu \in I(\phi) = (0.9405, 1)$.

1.2. Example: Imperfect compliance in a randomized trial

As a more involved example of a partially identified model, we consider a version of the imperfect compliance model with binary variables considered by various authors, including Chickering and Pearl [3], Imbens and Rubin [16], Pearl [24, Ch. 8], and Richardson et al. [26]. Clinical trial participants 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 participant's binary response is $Y^{(0)}$ if treatment is not taken, and $Y^{(1)}$ if treatment is taken, regardless of treatment assignment. Then a

participant's observable response is $Y = (1 - X)Y^{(0)} + XY^{(1)}$, where X indicates receipt of treatment. Against this, let Z indicate randomization to treatment, with the possibility that $X \neq Z$. For compliance type indicated by $C \in \{NT, AT, CO\}$, let $\gamma_{C,i}$ be the mean of $Y^{(i)}$ amongst the sub-population of that type. We consider inference about the population average causal effect (ACE), given as

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

It is easy to verify that the present set-up gives a nonidentified model, with $p = 8$, $q = 6$, $\theta = (\omega_{NT}, \omega_{AT}, \gamma)$, and $\phi = (\omega_{NT}, \omega_{AT}, \gamma_{NT,0}, \gamma_{AT,1}, \gamma_{CO,0}, \gamma_{CO,1})$. Particularly, the form of the invertible map from ϕ to the $(Y, X|Z)$ cell probabilities is readily established (see Appendix A for details). Unsurprisingly, 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.

It is also straightforward to verify that this model is partially identified when the ACE is the inferential target. Defining

$$a(\phi) = \omega_{CO}(\gamma_{CO,1} - \gamma_{CO,0}) + \omega_{NT}(1/2 - \gamma_{NT,0}) + \omega_{AT}(\gamma_{AT,1} - 1/2), \quad (1)$$
$$b(\phi) = (\omega_{NT} + \omega_{AT})/2, \quad (2)$$

the identification interval for the ACE is $I(\phi) = (a(\phi) - b(\phi), a(\phi) + b(\phi))$. Thus, unless the population happens to contain only compliers, uncertainty about the ACE will remain no matter how much data accumulates. We will investigate the limiting behavior of Bayesian inference for the ACE in Section 3.1.

1.3. Example: Inferring gene-environment interaction

As another example of a partially identified model, consider binary disease status Y, binary environmental exposure X, and binary genotype G. As a variant of a problem studied by Gustafson [10] and Gustafson and Burstyn [13], interest lies in the $(Y|X, G)$ relationship when only (Y, G) data are available, but certain assumptions can be invoked. The first of these is the gene-environment independence assumption, that X and G are independent 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 one of "ecological inference," as we wish to infer a property of the joint (Y, G, X) distribution from information about the (Y, G) and X marginals. As one example of an inferential target, say the task is to estimate $\psi = Pr(Y = 1|X = 1, G = 1) - Pr(Y = 1|X = 0, G = 1)$, the risk difference associated with exposure amongst those with genotype $G = 1$.

To gain a foothold in this problem, let $\theta = (\mu_0, \mu_{10}, \mu_{11})$ parameterize the distribution of $(Y|X, G)$, according to $\mu_0 = Pr(Y = 1|X = 0) = Pr(Y = 1|X = 0,$

$G = g)$, for $g = 0, 1$, and $\mu_{1g} = Pr(Y = 1|X = 1, G = g)$, for $g = 0, 1$. We take $r = Pr(X = 1)$ as a fixed constant, and define $\phi_g = Pr(Y = 1|G = g) = (1 - r)\mu_0 + r\mu_{1g}$, for $g = 0, 1$. Thus the likelihood arising from the $(Y|G)$ data depends on θ only through $\phi = (\phi_0, \phi_1)$. Hence we have a nonidentified model with $p = 3$ and $q = 2$. (Note that here we have left the marginal distribution of G unmodeled, but it makes no material difference if we include $Pr(G = 1)$ as a further parameter and then have a nonidentified model with $p = 4$ and $q = 3$.) In Section 3.2 we will determine the identification interval $I(\phi)$ for the target parameter $\psi = \mu_{11} - \mu_0$, and we will consider the limiting behavior of Bayesian inference in this setting.

1.4. Inferential approaches to partially identified models

There is a considerable literature on non-Bayesian approaches to partially identified models. See, for instance, Manski [21], Imbens and Manski [15], Romano and Shaikh [27], Vansteelandt et al. [29], Zhang [30], Tamer [28]. Typically the endeavor is split into two tasks. For a given problem, first one determines the form of the identification interval. Then the interval endpoints are viewed as the parameters of interest. Inference is considered as a separate exercise, comprised of estimating the endpoints and/or reporting a confidence set for the identification interval. 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 interval. More importantly for present purposes, 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 and the value of ϕ, then the identification interval $I(\phi)$ would simply be reported as "the answer."

Bayesian inferences in partially identified settings, and nonidentified models in general, have received considerable attention recently. In part this is due to needs arising in observational epidemiology. Study and data limitations which preclude identification are commonplace in this field. Works promulgating Bayesian or Bayes-like inference in such settings includes Joseph et al. [17], Dendukuri and Joseph [5], Greenland [7], Hanson et al. [14], Greenland [8], MacLehose et al. [20]. One theme in the broader literature is that identification and inference are very integrated under a Bayesian analysis (see, for instance, Barankin [1], Kadane [18], Dawid [4], Neath and Samaniego [23], Poirier [25], Gustafson [9]). 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 interval. Given an infinite number of datapoints then, "the answer" is this limiting posterior distribution, which constitutes a relative weighting of points in the identification interval.

Thus there is a fundamental discrepancy between non-Bayesian and Bayesian inference in partially identified models. This discrepancy is more extreme than

for identified models, where the identification interval is typically a single point, and therefore does not admit a weighting of its elements. As an example of this, Gustafson [12] considers the large-sample limit of frequentist coverage for Bayesian $(1-\alpha)$ credible intervals, in the partially identified context. He shows this limit is one over a large subset of the parameter space, and zero over its complement, where large means having prior probability $1-\alpha$. More generally, both Liao and Jiang [19] and Gustafson [10] suggest that obtaining a posterior distribution across the identification interval is a strength of the Bayesian approach. In contrast Moon and Schorfheide [22], who draw some large-sample comparisons between Bayes and non-Bayes procedures, are much more guarded about the prospect of reporting a posterior distribution across an identification interval as opposed to simply estimating the interval. None of these authors, however, attempt any sort of quantification of the potential utility of the shape of the posterior distribution over the identification interval.

It might be tempting to intuit that the force of the data is completely used up in determining the identification interval, so that the shape of the limiting posterior distribution across the interval is driven exclusively by the choice of prior distribution. Indeed, this is the case in some problems. In the simple example of Section 1.1, for instance, the relationship between ϕ and $I(\phi)$ is clearly bijective. Consequently, with a fixed prior distribution over Θ, knowledge of the identification interval for the target completely determines the limiting posterior distribution over the interval.

We can quickly establish, however, that other problems, such as the examples in Sections 1.2 and 1.3, exhibit more complex behavior. There can be distinct points ϕ_1 and ϕ_2 in $s[\Theta]$ such that $I(\phi_1) = I(\phi_2)$, but, starting with the same prior distribution over Θ, the limiting posterior distribution arising for true values of θ such that $s(\theta) = \phi_1$ differs from that arising if $s(\theta) = \phi_2$. This directly corresponds to the data having a say in the *shape* of the limiting posterior distribution of the target, as well as having a say in the support of this distribution. In turn this gives a sense in which there can be more to take away from an (infinite-sized) dataset than just the identification interval, whereas non-Bayesian approaches seem to suppose the opposite. Thus the situation is nuanced, and warrants investigation.

In the remainder of the paper we investigate the inferential utility of the shape of the posterior distribution, by taking a decision-theoretic view. In the large-sample limit, we focus on the typical height of the marginal posterior density for the target parameter, at the true value of the target. This can be compared to the typical height of other densities over the identification interval. More technically, we consider the expected score for a probabilistic forecast of the target parameter, under a logarithmic scoring rule. The difference in expected score between the Bayesian forecast and an ad-hoc choice of distribution over the identification interval can be decomposed into two terms. The first term speaks to the value of Bayesian processing of the information in the data about the identification interval. The second term reflects the additional information that can be recovered from using all the data. This decomposition is worked out for both the trial compliance example and the gene-environment interaction

example. It is hoped that this will be found relevant by researchers from both Bayesian and non-Bayesian backgrounds. For the Bayesian, it seems important to recognize that the usual decision-theoretic optimality of Bayesian procedures has a different "look-and-feel" in the partially identified case, particularly as the relevant posterior distributions are not converging to point masses. For the non-Bayesian, it seems important to recognize that there may be a sense in which the data speak beyond just estimating the identification interval.

2. Methodology

Starting with the parameter vector θ, say the investigator is willing to specify a prior distribution having a smooth density function $\pi(\theta)$ over Θ. With respect to an appropriate measure we write the density of data given parameters as $\pi(d|\theta)$, also assumed to be a smooth function of θ. Thus we can unambiguously refer to the joint density $\pi(d,\theta) = \pi(d|\theta)\pi(\theta)$ induced by the prior and model. In what follows, when useful we will write d_n to emphasize observable data comprised of n observations which are independent and identically distributed given θ. Also, we occasionally use upper-case notation for functions of data and parameters when it is helpful to stress random variable interpretations, e.g., inside expectations.

We frame our discussion in terms of how well we can generate a probabilistic forecast for the target parameter $\psi = g(\theta)$. For a finite sample of size n, say a family of density functions $h(\cdot;\cdot)$ is used, such that $h(\cdot;d_n)$ is the probabilistic forecast of the value of ψ, based on observing data $D_n = d_n$. We summarize the utility of the forecasting procedure by the expected score (ES) under a logarithmic scoring rule,

$$ES^{(n)}_{\pi,h} \quad = \quad E_\pi\{\log h(\Psi; D_n)\}. \tag{3}$$

Note here that by taking the expectation with respect to π, we are evaluating what would happen on average across repeated instantiations of *both* parameter and data values, with this ensemble of parameter values distributed according to the prior distribution. Note also that (3) is following the well-studied path of preferring forecasts with the highest expected score, which is essentially the same idea as preferring inferential schemes with the highest expected utility. But we can also think more prosaically simply in terms of $exp(ES^{(n)}_{\pi,h})$ being the typical height of the forecast density at the true value, with typical being in the sense of a geometric mean. For a general discussion of scoring rules for density forecasts, see Gneiting and Raftery [6]. Also, Bernardo [2] gives a sense in which all members of a class of scoring rules with desirable properties are equivalent to a logarithmic scoring rule.

The usual decision-theoretic optimality of Bayesian procedures applies here, despite the fact that this optimality is more commonly seen expressed for estimators than for probabilistic forecasts. The choice of family of densities h which maximizes (3) is the marginal posterior density of the target, $h(\psi; d_n) =$

$\pi(\psi|d_n)$. This can be seen as an immediate consequence of the non-negativity of Kullback-Leibler divergence.

Because we are studying partially identified problems in which the posterior distribution of the target converges to a non-degenerate distribution as the sample size grows, the limiting version of (3) is immediate. Observation of an infinite amount of data corresponds to knowledge of ϕ, so in the limit we are concerned with a family of density functions of the form $h(\cdot;\phi)$, and the corresponding expected score:

$$ES^{\infty}_{\pi,h} \quad = \quad E_{\pi}\{\log h(\Psi;\Phi)\}. \tag{4}$$

Bear in mind here that ψ and ϕ are both functions of θ, and the expectation is with respect to the prior density $\pi(\theta)$. Again the non-negativity of Kullback-Leibler divergence immediately implies that (4) is maximized by $h(\psi;\phi) = \pi(\psi|\phi)$. That is, the optimal probabilistic forecast is the conditional prior distribution of ψ given the value of ϕ that is gleaned from the data. Equivalently, this is the limit of the marginal posterior distribution of ψ, as the sample size goes to infinity. Thus the optimality of the marginal posterior distribution on the target extends smoothly in the limit. Note also that the lack of full identification typically implies that $\pi(\psi|\phi)$ is not degenerate, hence the maximized value of (4) will be finite. Continuing to work in the large-sample limit, we can use (4) as a starting point for understanding the "information flow" in the partially identified model. Henceforth it is useful to write the identification interval explicitly as $I(\phi) = (\phi^*_L(\phi), \phi^*_R(\phi))$, so that the two-component parameter ϕ^* is itself the identification interval.

Much of the non-Bayesian literature on partial identification treats the identification interval as the bivariate target of inference, with the consequent notion that knowledge of this interval is either all that should be gleaned, or all that can be gleaned, upon observation of an infinite-sized dataset. 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. Bearing this in mind, we term a family of densities indexed by ϕ^* to be an *ad-hoc* probabilistic forecast for the target, in the limiting case.

To fix ideas, one example of an ad-hoc scheme would be

$$h(\psi;\phi^*) = \frac{I\{\phi^*_L \le \psi \le \phi^*_R\}}{\phi^*_R - \phi^*_L},$$

corresponding to a uniform distribution over the identification interval. This would arise as the large-sample limit of forecasting a uniform distribution between estimates of the identification interval endpoints ϕ^*. Or, given that performance is measured on average with respect to prior π, an ad-hoc attempt to "do better where it counts" would involve truncating the prior distribution to the identification interval, i.e.,

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

Again, this could be thought of as the large-sample limit arising from truncating the prior distribution according to estimated values for ϕ^*.

For a given ad-hoc procedure $h(\psi;\phi^*)$, let $ES^{\infty}_{\pi,AH} = E_\pi\{\log h(\Psi;\Phi^*)\}$ be the expected score with respect to prior π. In contrast, let $ES^{\infty}_{\pi,B}$ be the optimal expected score arising from $h(\psi;\phi) = \pi(\psi|\phi)$, with the subscript B reminding us that this is the limit of the Bayesian procedure. We want to decompose $ES^{\infty}_{\pi,B} - ES^{\infty}_{\pi,AH} \geq 0$ in a way that sheds light on the utility of the shape of the limiting posterior distribution over the identification interval.

In investigating ad-hoc schemes, we are considering taking only information about ϕ^* from the data, which may "leave behind" some information about ϕ. We can easily elucidate that the best possible ad-hoc scheme is $h(\psi;\phi^*) = \pi(\psi|\phi^*)$, i.e., the same argument used above immediately reveals that the prior conditional distribution of the target given ϕ^* maximizes $E_\pi\{\log h(\Psi;\Phi^*)\}$. We refer to this as the coarsened Bayes (CB) procedure, and denote its expected score as $ES^{\infty}_{\pi,CB}$.

Note that in some problems it may literally be possible to regard the CB procedure as arising in the limit when Bayesian inference is applied only to a coarsened version of the data. That is, there might exist a function t such that coarsened data $D^*_n = t(D_n)$ have the following properties: (i), the distribution of D^*_n depends on ϕ only through ϕ^*, and (ii), the distribution of D^*_n given ϕ^* supports $\sqrt{n}$-consistent estimation of ϕ^*. By construction then, using only data D^*_n suffices to estimate the identification interval for the target, and the posterior distribution of $(\Psi|D^*_n)$ must converge to the conditional prior distribution given by $\pi(\psi|\phi^*)$. However, regardless of whether we can actually exhibit such a function $t()$, we can interpret $\pi(\psi|\phi^*)$ as the limiting Bayesian knowledge about the target were we to extract just enough of the data to estimate the identification interval, and not an iota more.

Now we are in a position to try to understand the worth of the shape of the limiting posterior distribution of the target across the identification interval. For a given ad-hoc procedure, we immediately have

$$ES^{\infty}_{\pi,B} - ES^{\infty}_{\pi,AH} \quad = \quad (ES^{\infty}_{\pi,CB} - ES^{\infty}_{\pi,AH}) + (ES^{\infty}_{\pi,B} - ES^{\infty}_{\pi,CB}), \qquad (5)$$

where both terms on the right-hand side of (5) are guaranteed to be nonnegative. This follows trivially from the optimality of the CB procedure amongst AH procedures, and the global optimality of the Bayesian procedure. So the first term on the right in (5) reflects the value of Bayesian processing of information about the identification interval, relative to ad-hoc processing of this information. The second term represents the value of using all the information in the data, not just the information about the identification interval. 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. The second term is of particular interest, since non-Bayesian approaches to partially identified models are predicated on the idea that knowledge of the identification interval is indeed all that can be obtained in the limit of infinite sample size.

Yet another interpretation is that the second term in (5) reflects the utility of the fact that multiple θ values in Θ can lead to the *same* identification interval

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 posterior distribution corresponding to a given identification interval, and the second term in (5) is zero. However, when there is indeed coarsening (i.e, the map from ϕ to ϕ^* is not invertible), there is no general reason to expect the second term to be zero.

In the next section, we simply compute the decomposition (5) in the examples from Sections 1.2 and 1.3. In doing so, we will refer to the first term as describing the *first-order Bayes advantage*. In particular, $exp(ES^{\infty}_{\pi,CB} - ES^{\infty}_{\pi,AH})$ is interpreted as the typical density ratio comparing the coarsened posterior density to the ad-hoc density, at the true value of the target parameter. Thus we can think of the first-order advantage as follows. Presuming that we choose to measure performance by averaging across the parameter space with respect to $\pi(\theta)$, and presuming that we are only allowed to take information about $I(\phi)$ from the data, then we are quantifying the gain from doing a Bayesian analysis with $\pi(\theta)$ as the prior. Similarly, we can view $\exp(ES^{\infty}_{\pi,B} - ES^{\infty}_{\pi,CB})$ as quantifying *the second-order Bayes advantage*, again on the density ratio scale. This gives us the further gain achieved if we allow ourselves to hear all that the data have to say, rather than just taking the information about the identification interval.

3. Examples

3.1. Imperfect compliance in a randomized trial, continued

Continuing the example of Section 1.2, recall that the model is parameterized by $\theta = (\omega_{NT}, \omega_{AT}, \gamma)$ having $p = 8$ components, while the distribution of the data depends on θ only through $\phi = (\omega_{NT}, \omega_{AT}, \gamma_{NT,0}, \gamma_{AT,1}, \gamma_{CO,0}, \gamma_{CO,1})$ having $q = 6$ components. Also, as mentioned earlier, the identification interval for the ACE takes the form $\phi^* = (a(\phi) - b(\phi), a(\phi) + b(\phi))$, with $a()$ and $b()$ given in (1) and (2) respectively.

We consider Bayesian inference under a uniform prior distribution; particularly, a prior under which $\omega \sim \text{Dirichlet}(1, 1, 1)$ and independently each of the six components of γ follow a $\text{Unif}(0, 1)$ distribution. Under this prior, Gustafson [11] shows that the limiting posterior distribution of the target, $\pi(\psi|\phi)$, has a trapezoidal-shaped density. In particular, whereas the bottom edge of the trapezoid is the identification interval $a(\phi) \pm b(\phi)$, the top edge of the trapezoid is $a(\phi) \pm c(\phi)$, where $c(\phi) = |\omega_{NT} - \omega_{AT}|/2 \leq b(\phi)$. Commensurately, the height of the density over $\psi \in a(\phi) \pm c(\phi)$ is $\{b(\phi) + c(\phi)\}^{-1}$. In this problem it is readily apparent that multiple values of $\phi \in s[\Theta]$ can produce the same values of $a()$ and $b()$ but different values of $c()$. Or, put another way, the mapping from ϕ to ϕ^* is not invertible. An explicit illustration of this appears in Figure 1. While for this problem the form of $\pi(\psi|\phi)$ is very simple to characterize, determination of the coarsened limiting posterior distribution, $\pi(\psi|\phi^*)$, is rather more involved, as described in Appendix A. The coarsened distribution is also depicted in Figure 1.

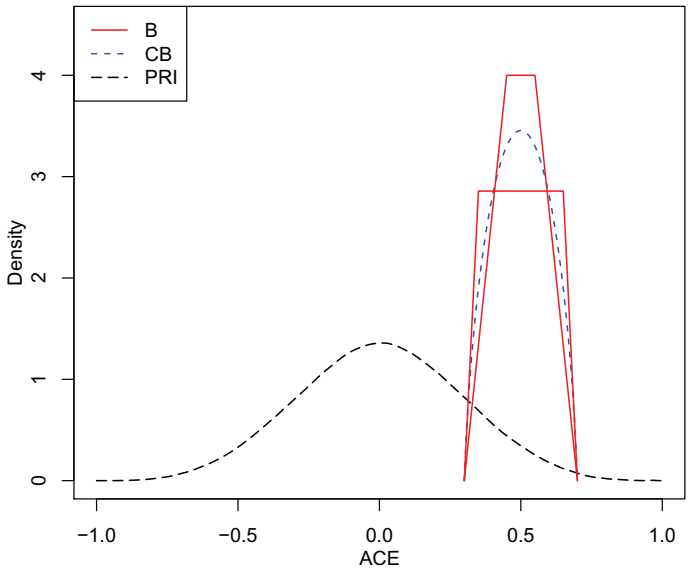

FIG 1. *Prior, limiting coarsened posterior, and limiting full posterior distributions for the ACE. In all cases* $\omega_{CO} = 0.6$ *and* $a(\phi) = 0.5$*, hence the identification interval is* 0.5 ± 0.2*. Of the two full posterior distributions, the less (more) concentrated distribution arises from* $\omega_{AT} = 0.05$ *(*$\omega_{AT} = 0.15$*).*

The various limiting expected scores for this problem are reported in Table 1. These are computed in a direct Monte Carlo fashion. That is, we generate m independent and identically distributed realizations $\theta^{(1)}, \ldots, \theta^{(m)}$ according to the prior $\pi(\theta)$. Then, for each realization we determine the information available from an infinite-sized dataset, $\phi^{(i)} = s(\theta^{(i)})$, and the target value $\psi^{(i)} = g(\theta^{(i)})$. For any probabilistic forecast then, the expected score is numerically approximated by the Monte Carlo average $m^{-1} \sum_{i=1}^{m} \log h(\psi^{(i)}; \phi^{(i)})$. Moreover, the numerical error involved is easily quantified by the standard error associated with this average, and similarly the error involved in computing the difference in two expected scores is described by the standard error arising from averaging m differences.

TABLE 1

Expected scores in the imperfect compliance example. These are computed as Monte Carlo averages across 10,000 realized values of θ *drawn from the prior distribution. Simulation standard errors are given in parentheses. The labels AH(TP) and AH(U) refer to the ad-hoc truncated prior and ad-hoc uniform procedures respectively*

$ES^{\infty}_{\pi,B}$	0.6127 (0.0063)
$ES^{\infty}_{\pi,B} - ES^{\infty}_{\pi,CB}$	0.0235 (0.0023)
$ES^{\infty}_{\pi,CB} - ES^{\infty}_{\pi,AH(TP)}$	0.1409 (0.0052)
$ES^{\infty}_{\pi,CB} - ES^{\infty}_{\pi,AH(U)}$	0.0880 (0.0036)

From Table 1 we see that the first-order Bayes advantage is appreciable in this problem. Using only information about the identification interval, the coarsened posterior density over this interval is typically considerably higher at the true value than an ad-hoc density (with a typical density ratio of $\exp(0.14) \approx 1.15$ compared to the truncated prior density and 1.09 compared to a uniform density). Finally, we see there is a modest second-order Bayes advantage. By using all the information rather than just the information about $I(\phi)$, the fully Bayesian posterior garners a further 2.4% improvement over the coarsened posterior, on the density ratio scale. Moreover, this improvement is calculated with simulation-significance, i.e., we have computed with sufficient accuracy to be convinced that $ES^{\infty}_{\pi,B} > ES^{\infty}_{\pi,CB}$. Generally then we see the shape of the posterior distribution is helpful in this problem. The magnitude of the first-order effect corresponds to a practical advantage in an applied statistics sense. The second-order effect is much more modest in magnitude, but the important point here is that the data can make a helpful contribution beyond the direct information they convey about the identification interval.

3.2. Inferring gene-environment interaction, continued

Recall that the initial parameterization for this problem introduced in Section 1.3 is in terms of $\theta = (\mu_0, \mu_{10}, \mu_{11}) \in (0,1)^3$, where $\mu_0 = Pr(Y = 1|X = 0)$ and $\mu_{1g} = Pr(Y = 1|X = 1, G = g)$, for $g = 0, 1$. Also recall that the likelihood depends on θ only through $\phi = (\phi_0, \phi_1)$, where $\phi_g = Pr(Y = 1|G = g) = (1-r)\mu_0 + r\mu_{1g}$, for $g = 0, 1$, and the inferential target is $\psi = g(\theta) = \mu_{11} - \mu_0$.

In this problem it is easy to see that $(\phi, \psi) = (\phi_0, \phi_1, \psi)$ constitutes a linear reparameterization of $\theta = (\mu_0, \mu_{10}, \mu_{11})$, with the inverse of the mapping given as

$$\begin{pmatrix} \mu_0 \\ \mu_{10} \\ \mu_{11} \end{pmatrix} = r^{-1} \begin{pmatrix} 0 & r & -r^2 \\ 1 & -(1-r) & r(1-r) \\ 0 & r & r(1-r) \end{pmatrix} \begin{pmatrix} \phi_0 \\ \phi_1 \\ \psi \end{pmatrix}. \tag{6}$$

Thus for given ϕ the identification interval is all values of ψ under which (6) yields a value in $[0,1]^3$, i.e., the identification interval endpoints are:

$$\phi^*_L = -\min\left\{\frac{\phi_1}{r}, \frac{1-\phi_1}{1-r}, \frac{\phi_0 - (1-r)\phi_1}{r(1-r)}, 1\right\}, \tag{7}$$

and

$$\phi^*_R = \min\left\{\frac{\phi_1}{r}, \frac{1-\phi_1}{1-r}, \frac{r + (1-r)\phi_1 - \phi_0}{r(1-r)}, 1\right\}. \tag{8}$$

As we will demonstrate explicitly below, the mapping from ϕ to ϕ^* is not invertible, which leaves open the possibility that $ES^{\infty}_{\pi,B} > ES^{\infty}_{\pi,CB}$.

We consider Bayesian inference using the prior distribution having $(\mu_0, \mu_{10}, \mu_{11})$ independent and identically distributed as Beta(k_1, k_2). In what follows,

$b_{k_1,k_2}()$ is used to denote the density of this Beta distribution. The linear map between θ and (ϕ, ψ) immediately gives the joint prior density of (ϕ, ψ) as

$$\begin{aligned} \pi(\phi_0, \phi_1, \psi) &= r^{-1} b_{k_1,k_2}(\phi_1 - r\psi)\, b_{k_1,k_2}(\phi_1 + (1-r)\psi) \times \\ &\quad b_{k_1,k_2}(r^{-1}(\phi_0 - (1-r)\phi_1 + r(1-r)\psi)), \end{aligned} \tag{9}$$

with support restricted to (ϕ, ψ) such that $\psi \in I(\phi)$. The limiting posterior distribution for the target is given by the conditional prior $\pi(\psi|\phi)$, with conditioning on the true value of ϕ. At least up to a normalizing constant, this conditional density can be read off from the joint density (9), by regarding this expression as a function of ψ for fixed ϕ.

An obvious choice of prior specification for this problem is $(k_1, k_2) = (1, 1)$, corresponding to a uniform distribution on each of the three outcome probabilities. In this special case, it is immediate from (9) that for every $\phi \in s[\Theta]$, $\pi(\psi|\phi)$ is the uniform distribution on the identification interval $I(\phi)$. The limiting posterior is therefore always the same as the coarsened limiting posterior, and there can be no second-order Bayes advantage.

For other choices of prior distribution, the situation is more nuanced. As an example, in Appendix B we examine in detail the specification $k_1 = k_2 = 2$, which gives slightly more prior weight to mid-range values of the response probabilities. Using $f()$ to denote the map from ϕ to ϕ^*, we prove that for every value of $\phi^* \in s[\Theta]$ there is either (i), two distinct point solutions to $f(\phi) = \phi^*$, which we denote as ϕ^A, ϕ^B, or (ii), a line-segment of solutions of the form $\{\phi : \phi^A_{0L} < \phi_0 < \phi^A_{0R}, \phi_1 = \phi^A_1\}$ 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),

$$\pi(\psi|\phi^*) \propto \int_{\phi^A_{0L}}^{\phi^A_{0R}} \pi(\psi|\phi = (s, \phi^A_1))\pi(s, \phi^A_1)ds. \tag{10}$$

Note that as one of infinitely many solutions, the further point solution ϕ^B does not contribute to (10). We are then able to compute $\pi(\psi|\phi^*)$ for a given ϕ^*.

For the $(k_1, k_2) = (2, 2)$ case, Figure 2 compares the Bayes, coarsened Bayes, and ad-hoc probabilistic forecasts 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 full and coarsened limiting posterior densities are virtually indistinguishable in each case, while being quite different from both the ad-hoc forecasts (uniform distribution over the identification interval, prior marginal distribution truncated to the identification interval).

As in the previous example, the various expected scores are computed as Monte Carlo averages across a large number of draws of θ from $\pi(\theta)$, with results reported in Table 2. Again we have "simulation significance" to attest to

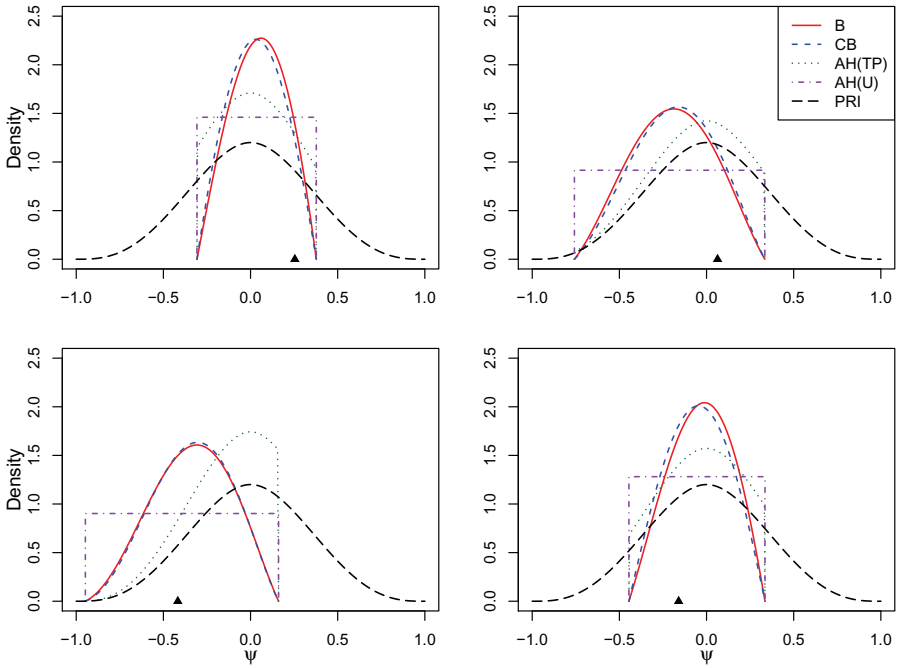

FIG 2. *Limiting probabilistic forecasts for the risk difference under hyperparameters* $(k_1, k_2) = (2, 2)$. *Each panel corresponds to a different underlying value of* θ *drawn from its prior distribution. The true value of the target is indicated in each panel (triangle on the horizontal axis).*

$ES^{\infty}_{\pi,B} > ES^{\infty}_{\pi,CB}$. However, the difference between these two expected scores is so small as to be negligible in any practical sense. This jibes with the close agreement seen between $\pi(\psi|\phi)$ and $\pi(\psi|\phi^*)$ in Figure 2. So there is a tiny, but non-zero, second-order Bayes advantage. On the other hand, the first-order Bayes advantage is very substantial in this example. The optimal-shaped density over the identification interval tends to be 17% higher at the true value than truncated marginal prior, and 21% higher compared to the uniform distribution over the identification interval.

TABLE 2

Expected scores in the gene-environment example with hyperparameters $k_1 = k_2 = 2$. *These are computed as Monte Carlo averages across 10,000 realized values of* θ *drawn from the prior distribution. Simulation standard errors are given in parentheses. The labels AH(TP) and AH(U) refer to the ad-hoc truncated prior and ad-hoc uniform procedures respectively*

$ES^{\infty}_{\pi,B}$	0.2680 (0.0055)
$ES^{\infty}_{\pi,B} - ES^{\infty}_{\pi,CB}$	0.00293 (0.00068)
$ES^{\infty}_{\pi,CB} - ES^{\infty}_{\pi,AH(TP)}$	0.1540 (0.0046)
$ES^{\infty}_{\pi,CB} - ES^{\infty}_{\pi,AH(U)}$	0.1892 (0.0046)

4. Robustness

In general the decision-theoretic optimality of any Bayesian procedure stems from using the same distribution over the parameter space ("the prior") to both (i), average the procedure's performance across possible true parameter values, and (ii), use as an input to form the posterior distribution for a given dataset. An obvious question to ask then is how quickly does the optimality fade if two different distributions are used? That is, what happens if "Nature's prior distribution" used to average performance across the parameter space differs from the "investigator's prior distribution" used to determine the posterior distribution upon receipt of data.

We retain $\pi(\theta)$ as notation for the investigator's prior, but consider what happens when Nature's prior is $\pi^*_\lambda(\theta)$ 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., $\pi^*_0(\theta) = \pi(\theta)$. Specifically we look at the comparison between the limiting posterior marginal distribution over the identification interval and a uniform distribution over the identification interval, as the investigator's prior stays fixed but Nature's prior moves away from it. Let

$$t(\lambda) \quad = \quad E^*_\lambda \left\{ \log \pi(\Psi|\Phi) + \log \left(\Phi^*_R - \Phi^*_L\right) \right\}, \tag{11}$$

where the expectation is with respect to $\pi^*_\lambda(\theta)$. Clearly then $t(0) = ES^\infty_{\pi,B} - ES^\infty_{\pi,AH(U)} > 0$, and the magnitude of λ required to make $t(\lambda) \leq 0$ reflects the stability of the Bayes advantage.

When λ has more than one component, it may become complicated to evaluate (11) in many different directions away from $\lambda = 0$. Thus we propose computing the gradient

$$t'(0) \quad = \quad E_\pi \left[s(\Theta) \left\{ \log \pi(\Psi|\Phi) + \log \left(\Phi^*_R - \Phi^*_L\right) \right\} \right], \tag{12}$$

where $s(\theta) = \partial \log \pi^*_\lambda(\theta) / \partial \lambda|_{\lambda=0}$. Then evaluating (11) for values of λ proportional to this gradient corresponds to looking along the direction in which (11) changes most rapidly with λ, locally at zero.

Returning to the compliance example of Section 3.1, we assume that Nature and the investigator agree on a uniform prior for the components of γ. However, whereas the investigator uses $\omega = (\omega_{CO}, \omega_{NT}, \omega_{AT}) \sim \text{Dirichlet}(1,1,1)$, Nature uses $\omega \sim \text{Dirichlet}(1+\lambda_1, 1+\lambda_2, 1+\lambda_3)$. Numerical evaluation of (12) indicates that $t'(0) \propto (0,1,1)'$. Thus we focus attention on the case that Nature's prior is $\text{Dirichlet}(1, 1+\lambda, 1+\lambda)$, 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 $\text{Dirichlet}(1, 0.1, 0.1)$ distribution is fairly extreme and far from $\text{Dirichlet}(1,1,1)$. In particular, this distribution puts considerable weight on extremely small values of ω_{NT} and ω_{AT}.

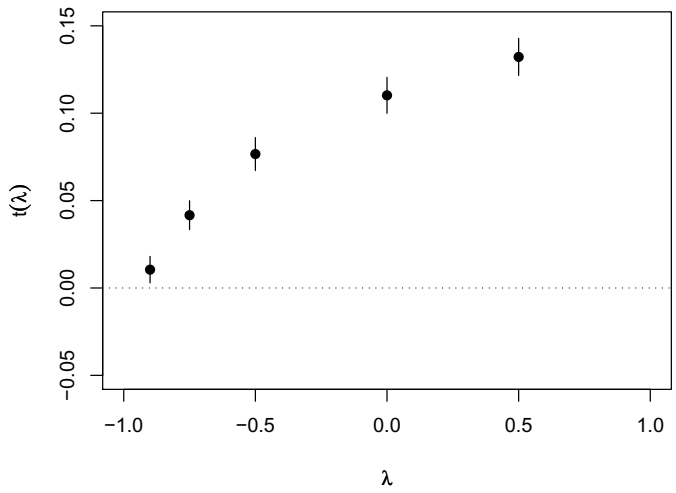

FIG 3. *Robustness of the limiting posterior distribution when Nature's prior is* $\omega \sim Dirichlet(1, 1+\lambda, 1+\lambda)$ *and the investigator's prior is* $\omega \sim Dirichlet(1, 1, 1)$. *The difference in expected score for the Bayes procedure compared to the ad-hoc uniform procedure,* $t(\lambda)$, *is computed for selected values of* λ. *The vertical bars are 95% simulation confidence intervals based on the 5,000 Monte Carlo realizations from Nature's prior used to compute* $t(\lambda)$.

5. Discussion

Returning to the question posed in the title of this paper, we have seen that the shape of the posterior distribution in partially identified models is indeed useful. Most of the benefit lies in what we have termed the first-order Bayes advantage. If we take only the data that inform the identification interval, then processing these data in a Bayesian way, as we have termed the coarsened Bayes procedure, tends to yield a higher posterior density for the target, evaluated at the true value. In the two examples, gains of 10–20% in density height were seen, relative to ad-hoc choices of distributions across the identification interval.

We have also seen, both theoretically and empirically, that there can be a further second-order advantage that arises from using all the data. This arises as in some problems different datasets (in the limiting sense) can produce the same identification interval but different posterior distributions over this interval. In turn, this gives a tangible sense in which the data themselves speak to the relative plausibility of different values inside the identification interval, and a tangible sense in which the shape of the posterior distribution is not just a consequence of the choice of prior distribution. Of course in both examples, and particularly in the second example, the second-order advantage is small, both in absolute terms and in comparison to the first-order advantage. Thus the impact lies in the conceptual and theoretical understanding of how inference works in the partially identified setting, rather then in finding hitherto inaccessible gains in estimator performance.

One issue arising is that the first-order Bayes advantage might in part be a self-fulfilling prophecy, since the same prior distribution is used to both derive the posterior distribution and weight the averaging of performance across the parameter space. And of course this is the general underpinning of the decision-theoretic optimality of all Bayesian procedures. While a full look at this issue is beyond the scope of this article, a modest evaluation of "robustness" was conducted in Section 4, examining what happens when two different distributions over the parameter space are used to fulfill the two roles mentioned above.

Of course any assessment of a statistical procedure involves choices concerning how performance is quantified. Arguably our choice of logarithmic scoring of probabilistic forecasts for the target parameter has intuitive appeal. If two forecast distributions always have the same support, then it seems appealing to deem the one with highest average density at the true value as having the more useful shape. Obviously other assessments could be made though, say based on point-estimator performance. For instance, say $m_h(\phi) = \int \psi h(\psi;\phi)$ is the mean of the probabilistic forecast, reducing to the limiting posterior mean when $h(\psi;\phi) = \pi(\psi|\phi)$. Then performance of various schemes (Bayesian or not) could be based on average mean-squared error $E_\pi[\{m_h(\Phi) - \Psi\}^2]$. While investigation of this is beyond the scope of the present paper, it is easy to note that this would change things substantially in our imperfect compliance example. Here the Bayes, coarsened Bayes, and truncated uniform procedures all yield distributions which are symmetric about $a(\phi)$ as given in (1). Hence all three give rise to $m_h(\phi) = a(\phi)$, and consequently identical performance.

A referee has also raised the question of evaluating procedures in terms of predictive distributions. For instance, in the Bayesian case the forecast distribution of the next data point D^* given the first n datapoints d_n tends to $\pi(d^*|\phi) = \int \pi(d^*|\theta)\pi(\theta|\phi)d\theta$, as n goes to infinity. In general this is not so tightly connected to the present work, since our focus is on $\pi(\lambda|\phi)$, which is only a marginal distribution associated with $\pi(\theta|\phi)$ required to form the predictive distribution. In some cases the two do coincide though, particularly should (ϕ, ψ) constitute a reparameterization of θ (this happens in the gene-environment interaction example, but not in the imperfect compliance example). When they do coincide, a possibility would be to shift the evaluation of performance from (4) to $E_\pi\{\log \int \pi(D^*|\Phi,\lambda)h(\lambda;\Phi)d\lambda\}$.

A limitation of this work is that only the asymptotic limit is considered. We point out, however, that this limit is often approached rather quickly, in the following sense. For given data d_n, the posterior variance of the target parameter is

$$\mathrm{Var}_\pi(\Psi|D_n = d_n) \quad = \quad E_\pi\left\{\mathrm{Var}_\pi(\Psi|\Phi)|D_n = d_n\right\} + \mathrm{Var}_\pi\left\{E_\pi(\Psi|\Phi)|D_n = d_n\right\}.$$

The first term on the right-side will tend to $\mathrm{Var}_\pi(\Psi|\Phi = \phi) > 0$, while the second term falls off as n^{-1}. As soon as the second term is small relative to the first we are "near" the limit, with minimal scope for further reduction in the width of the target posterior distribution as further data are collected. Thus we can reasonably expect that what we learn in the limit applies at realistic sample sizes.

A final comment is that the general topic of inference in the absence of identification may be perceived by some as rather esoteric. Indeed, there is often a feeling that in order to tackle an applied problem, one needs to make enough modeling assumptions so as to attain an identified model. Unfortunately, in many scientific domains this can promote the use of quite dubious modeling assumptions. Arguably then, we need to make peace with models that are only partially identified for the target parameter, and we need to understand the workings of inference in such settings.

Appendix A: Further details of the imperfect compliance example

The map from ϕ to $(Y, X|Z)$ cell probabilities is given via

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

Expressed in this form, the mapping is readily seen to be invertible.

To determine the coarsened limiting posterior distribution in this example, note that we can write $\psi = \mu + \epsilon$, where $\mu = a(\phi)$, while

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

Thus the coarsened limiting posterior distribution 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

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

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

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

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

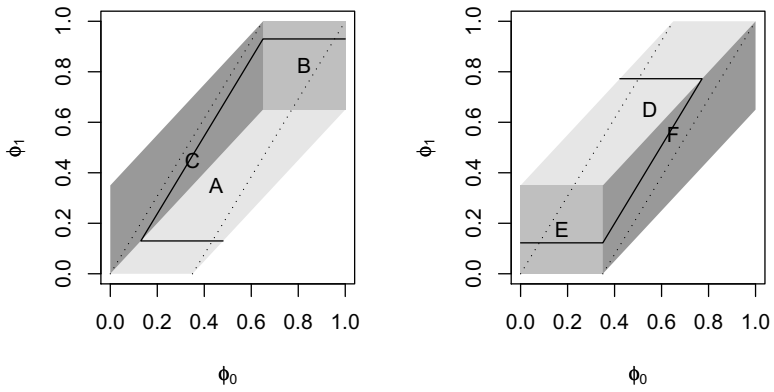

FIG 4. *The partition of S into sets on which ϕ_L^* is piecewise linear (left panel) and ϕ_U^* is piecewise linear (right panel), when $r = 0.35$. The upper and lower dotted reference lines appearing on both panels correspond to $\phi_L^* = 0$ and $\phi_R^* = 0$. The level set for $\phi_L^* = -0.2$ is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

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: Details of the gene-environment model with hyperparameters $k_1 = k_2 = 2$

Recall that f is the map from ϕ to ϕ^*. For a given c^* in the image of f, we need to characterize solutions to $f(\phi) = c^*$. Note that the domain of f is the subset of the unit square U given by $S = \{\phi \in U : |\phi_0 - \phi_1| < r\}$. The form of (7) and (8) 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 $f(\phi) = c^*$. (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 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 f, either there are two point solutions to $f(\phi) = c^*$, or one horizontal line segment of solutions plus an additional point solution.

References

[1] Barankin, E. W. (1960). Sufficient parameters: Solution of the minimal dimensionality problem. *Annals of the Institute of Mathematical Statistics*, 12:91–118. MR0126308

[2] Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, 7:686–690. MR0527503

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

[4] Dawid, A. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, 41:1–31. MR0535541

[5] Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*, 57(1):158–167. MR1833302

[6] Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. MR2345548

[7] Greenland, S. (2003). The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes

and magnetic fields to childhood leukemia. *Journal of the American Statistical Association*, 98:47–55. MR1977199

[8] Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society, Series A*, 168:267–306. MR2119402

[9] Gustafson, P. (2005). On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with discussion). *Statistical Science*, 20:111–140. MR2183445

[10] Gustafson, P. (2010). Bayesian inference for partially identified models. *International Journal of Biostatistics*, 6(2), article 17. MR2602560

[11] Gustafson, P. (2011). Comment on 'Transparent parameterizations of models for potential outcomes,' by Richardson, Evans, and Robins. 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: Proceedings of the Ninth Valencia International Meeting*. Oxford University Press.

[12] Gustafson, P. (2012). On the behaviour of bayesian credible intervals in partially identified models. *Electronic Journal of Statistics*, 6:2107–2124. MR3020258

[13] 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. MR2767805

[14] Hanson, T., Johnson, W., and Gardner, I. (2003). Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics*, 8(2):223–239.

[15] Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72:1845–1857. MR2095534

[16] Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, 25:305–327. MR1429927

[17] Joseph, L., Gyorkos, T., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–272.

[18] Kadane, J. B. (1974). The role of identification in Bayesian theory. In Fienberg, S. E. and Zellner, A., editors, *Studies in Bayesian Econometrics and Statistics, In Honor of Leonard J. Savage*, page 175. MR0483124

[19] Liao, Y. and Jiang, W. (2010). Bayesian analysis in moment inequality models. *Annals of Statistics*, 38:275–316. MR2589323

[20] MacLehose, R., Olshan, A., Herring, A., Honein, M., Shaw, G., Romitti, P., et al. (2009). Bayesian methods for correcting misclassification: An example from birth defects epidemiology. *Epidemiology*, 20(1):27-35.

[21] Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer. MR2151380

[22] Moon, H. R. and Schorfheide, F. (2012). Bayesian and frequentist inference in partially identified models. *Econometrica*, 80:755–782. MR2951948

[23] Neath, A. and Samaniego, F. (1997). On the efficacy of bayesian inference for nonidentifiable models. *American Statistician*, 51:225–232. MR1467551

[24] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press. MR1744773

[25] Poirier, D. (1998). Revising beliefs in nonidentified models. *Econometric Theory*, 14(4):483–509. MR1650041

[26] 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: Proceedings of the Ninth Valencia International Meeting*, pages 569–610. Oxford University Press.

[27] 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. MR2422399

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

[29] 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. MR2281311

[30] Zhang, Z. (2009). Likelihood-based confidence sets for partially identified parameters. *Journal of Statistical Planning and Inference*, 139:696–710. MR2479821