Bayesian Inference for Partially Identified Models

Paul Gustafson, University of British Columbia
Bayesian Inference for Partially Identified Models

Paul Gustafson

Abstract

Identification can be a major issue in causal modeling contexts, and in contexts where observational studies have various limitations. Partially identified models can arise, whereby the identification region for a target parameter -- the set of values consistent with the law of the observable data -- is strictly contained in the set of a priori plausible values, but strictly contains the single true value. The first part of this paper reviews the use of Bayesian inference in partially identified models, and describes the large-sample limit of the posterior distribution over the target parameter. This limiting distribution will have the identification region as its support. The second part of the paper focuses on the informativeness of the shape of the limiting distribution. This provides a point of departure with non-Bayesian approaches, since they focus on inferring the identification region without attempting to speak to relative plausibilities of values inside the identification region. The utility of the shape is investigated in several partially identified models.

KEYWORDS: asymptotics, Bayesian inference, nonidentified models

Author Notes: Research supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Institutes for Health Research (Funding Reference Number 62863).
1 Introduction

A statistical model is identified if multiple sets of parameter values cannot correspond to the same distribution of observables. The vast majority of the statistical modeling literature addresses situations where the model in question is identified, usually with the consequence that the envisioned data structure leads to consistent estimation of target parameters, i.e., the targets are uniquely determined by the distributional law of the observed data. In causal inference contexts, however, one is often operating ‘around’ the boundary between identified and nonidentified models, and discussions abound on what modeling assumptions are plausible versus what assumptions yield identification and consistent estimation. In observational epidemiology contexts, an argument can be made that realistic modeling of study limitations, such as unobserved confounding, selection bias and exposure measurement error, will tend to result in a nonidentified model. This is particularly true when, as is typical, knowledge of the magnitude of these limitations is imperfect. It is also particularly true when the modeling is realistic in terms of embedding a real causal effect as the target of inference.

While conventional statistical thinking might deem it controversial, one can produce inferences based on a nonidentified model, though necessarily such inferences cannot attain all the nice properties (such as consistency) enjoyed when identified and correctly specified models can be used. It is natural to consider Bayesian inference in a nonidentified model setting, since (i) there may well be both need and opportunity to supply prior information to partially mitigate the lack of identification, and (ii) the algebra of Bayes theorem for updating a prior distribution to a posterior distribution via a likelihood is ‘blind’ to whether or not the model is identified. Some examples of applying Bayesian inference in settings where a study or data limitation yields nonidentification include Dendukuri and Joseph (2001), Gustafson, Le and Saskin (2001), Georgiadis et al. (2003), Hanson et al. (2003), Scharfstein, Daniels and Robins (2003), and Gustafson and Greenland (2006). Some in-depth examples of applying Bayesian (or near-Bayesian) inference while acknowledging the lack of identification arising via multiple study limitations are provided by Greenland (2003, 2005). Some general discussion on the utility of Bayesian inference in nonidentified models can be found in Neath and Samaniego (1997), Poirer (1998), Gustafson (2005a), and Xie and Carlin (2006).

In fact, many nonidentified models arising in causal and/or epidemiological contexts might rather be termed as partially identified. For present purposes, partially identified is taken to mean that the target parameter cannot be consistently estimated, but the possible set of values for the target which are...
consistent with the observed data law is smaller than the *a priori* set of possible values (at least for some such laws, if not all). That is, the observed data law gives an *identification region* for the target, rather than the single correct value. In Bayesian terms, partial identification means that as the sample size goes to infinity, the support of the marginal posterior distribution on the target converges to a set which is smaller than the corresponding prior support, but larger than a single value.

A question of obvious import for a given partially identified model is how large is the identification region for the target parameter? A follow-up issue is how finite-sample inference ‘approaches’ the identification region as the sample size grows. In Bayesian terms, the primary issue is how concentrated the large-sample limit of the posterior distribution on the target is around the true value. Section 2 describes the properties of the large-sample limit of the posterior distribution in partially identified models, and reviews some recent literature on this topic. Section 3 then takes up a more specific question. The identification region is common to Bayesian and non-Bayesian inferences in partially identified models. However, non-Bayesian approaches tend to ‘stop here,’ in the sense that inferring the identification region is regarded as all that can be done. The Bayesian approach is fundamentally different in character, since the large-sample answer (the limit of the posterior marginal distribution on the target) is a distribution over the identification region. Thus the question arises of how helpful the *shape* of this distribution is in inferring the target, over and above the identification region itself. This is investigated in detail in Section 3, with two examples provided. The paper concludes with a short discussion in Section 4.

2 Posterior Distributions in Partially Identified Models

2.1 Transparent Parameterization

Many modeling situations which give rise to nonidentified models lend themselves to quite a simple and illuminating display of how the lack of identification impacts inference. Say that modelling assumptions, which perhaps involve unobservable variables, induce a statistical model for an observable data vector \( D_i \) on the \( i \)-th unit, with the distribution of \( D_i \) depending on a vector of unknown parameters \( \theta \). These parameters are assigned a prior distribution \( \theta \sim \pi \). A dataset of size \( n \) is then presumed to consist of independent and identically distributed realizations on \( n \) units, denoted as \( D = (D_1, \ldots, D_n) \).
That is, the data arise from the distribution \( (D|\theta = \theta^*) \), where hereafter the star notation is used to denote true values of parameters, when such emphasis is helpful. A lack of identification is manifested by multiple values of \( \theta \) corresponding to the same distribution of \( D \).

Now say it is possible to reparameterize from \( \theta \) to \( \phi = (\phi_I, \phi_N) \) such that the distribution of \( (D|\phi) \) depends only on \( \phi_I \), with the model \( (D|\phi_I) \) satisfying standard regularity conditions for parametric models. Consequently, \( n^{1/2} \) consistent estimation of \( \phi_I \) is possible. In such situations we refer to \( \phi \) as a transparent reparameterization (following Gustafson 2005a), and it is easy to ascertain the consequences of nonidentification for Bayesian inference. Particularly, the chosen prior distribution in the original parameterization \( \theta \sim \pi \) induces a prior distribution for \( \phi \) and hence a joint distribution for \( (\phi_I, \phi_N, D) \), with the property that \( \phi_N \) and \( D \) are conditionally independent given \( \phi_I \). Under weak assumptions, such as the induced prior for \( \phi_I \) being well behaved, the posterior marginal on \( \phi_I \) will converge to a point mass at the true value \( \phi_I^* \), as \( n \to \infty \). On the other hand, the posterior conditional for \( (\phi_N|\phi_I, D) \) will be the same as the prior conditional for \( (\phi_N|\phi_I^*) \), and hence completely unaffected by data values and sample size. In the \( n \to \infty \) limit then, the posterior for \( \phi \) is characterized as a point mass distribution for \( \phi_I \) at \( \phi_I^* \), along with the prior conditional \( (\phi_N|\phi_I = \phi_I^*) \). Thus the structure of the limiting posterior distribution (LPD) is clear.

For a transparent parameterization, the impact of nonidentification on point estimation is readily seen (Gustafson 2005a). A scalar parameter of interest can be written as \( \psi = g(\phi) \). Using the posterior mean as a point estimator of the target, we have

\[
\hat{\psi} = E\{g(\phi)|D\} = E[E\{g(\phi)|\phi_I\}|D] = E\{\tilde{g}_\pi(\phi_I)|D\},
\]

where \( \tilde{g}_\pi(\phi_I) = E_\pi\{g(\phi)|\phi_I\} \) is the expectation of the target under the prior conditional distribution. Thus our estimator of the target \( g(\phi) \) is consistent for a different (and prior-dependent) quantity \( \tilde{g}_\pi(\phi_I) \), and hence a large-sample bias of \( \tilde{g}_\pi(\phi_I^*) - g(\phi^*) \) is incurred. In general then the bias depends on both the underlying parameter values and the choice of prior distribution. Examples of determining this bias in a variety of problems can be found in Gustafson (2005ab, 2007) and Gustafson and Greenland (2006).

A common thread in the examples is that the bias can be less than intuitively anticipated. This can arise because a sensible prior specified in the original parameterization \( \theta \), which perhaps does not involve strong dependencies, may induce a prior for \( \phi \) which does involve a strong dependence between...
In some problems the support of the conditional prior for \((\phi_N|\theta_I)\) is smaller than the support of the marginal prior for \(\phi_N\). Consequently, the support of the limiting posterior for the target \(\psi\) can be intermediate - strictly contained in the support of the prior for \(\psi\), but strictly containing the true value \(\psi^*\). This provides a point of contact with the econometrics literature on bounds and partial identification (see, for instance, Manski 2003).

Faced with a lack of identification, the fact that the support of the limiting posterior on the target is wider than a single value is a ‘feature,’ not a ‘bug.’ Since the lack of identification rules out the possibility of consistent estimation of the target, a method which gives interval estimates not shrinking to a single point as \(n \to \infty\) is behaving appropriately. One nice feature of Bayesian interval estimates is their calibration property with respect to the prior distribution. That is, say \(I_\alpha(D)\) is a credible interval for \(\psi\) with posterior probability content \(\alpha\). This could be an equal-tailed interval, a highest-posterior-density interval, or some other interval with posterior probability content \(\alpha\). Then joint sampling of parameters and data from the prior on \(\phi\) plus the model for \((D|\phi)\) yields probability \(\alpha\) for the event \(\{\psi \in I_\alpha(D)\}\). There are no regularity conditions here - this sense of correct Bayesian coverage holds exactly for any sample size, whether or not the model is identified. In thinking about the \(n \to \infty\) case, a credible interval based on the LPD for \(\psi\) can be expressed as \(I_\alpha(\theta_I)\), since observation of an infinite data sample equates with knowledge of \(\theta_I\). Then the calibration property can be thought of as the set of \(\phi\) values for which \(\{\psi(\phi) \in I_\alpha(\theta_I)\}\) having prior probability \(\alpha\). Whereas there is no route to obtaining intervals with correct frequentist coverage in nonidentified model settings, correct Bayesian coverage is automatic. This point is discussed at length by Gustafson and Greenland (2009). These authors also contrast Bayesian interval estimators with estimators designed to have conservative frequentist coverage when prior information in the form of bounds on parameters is available, as discussed by Vansteelandt et al. (2006).

### 2.2 The Rationale for LPD Determination

Having sketched out the limiting behaviour in parametric models which admit a transparent parameterization, we take a step back, and discuss why it can be important to determine the LPD arising from a given model and prior distribution. Some pertinent comments are as follows:

- As an intermediate question which impinges on some of the more general questions below, it can be important to assess whether the utility of the LPD as knowledge about the target varies considerably across the
parameter space. It has been noted in several settings (Gustafson 2005b, 2007) that the extent to which the LPD concentrates near the true value of the target $\psi^*$ can vary dramatically with the values of the underlying true parameters $\phi^*$. In such instances, it is more challenging to characterize the overall utility of posterior inference via the nonidentified model under study.

- A more fundamental issue is generally whether one tends (across different values of the underlying parameters) to obtain a LPD which is usefully narrow or a LPD which is uselessly wide. In the latter case, finite-sample data collection would not be worthwhile, since even an infinite sample would likely not be very informative about the target of interest. In the former case, the secondary question arises of how big a sample size is needed to obtain a posterior on the target which is nearly as narrow as the LPD.

- Another important issue is what drives the LPD. If concentration of the LPD arises primarily via its support being much smaller than the prior support, then this concentration will not be strongly governed by the particular choice of prior on the original parameters $\theta$. On the other hand, if the nature of the LPD is strongly governed by the choice of prior (via the shape of the LPD rather than the support), then this dependence should be identified. Questions of prior sensitivity are clearly more important in nonidentified models, since the typical arguments about enough data swamping the prior no longer apply.

- Again in the bigger picture, a key question is whether Bayesian inference arising from the nonidentified model at hand seem like a good strategy compared to other strategies. In keeping with statistical dogma, one obvious strategy is to force identification by strengthening model assumptions. Then the comparison at hand devolves to comparing inference from a nonidentified but correct model to inference from an identified but potentially misspecified model. In the $n \to \infty$ limit this corresponds to comparing the LPD (which is typically not a point mass) to a point-mass limiting posterior which may be at the wrong value due to model misspecification. This sort of comparison is particularly emphasized in Gustafson (2005b, 2007) and Gustafson and Greenland (2006), where the identified but misspecified model arises from making a strong modelling assumption corresponding to a point-mass prior on a certain parameter, while the nonidentified but correctly specified counterpart model relaxes the point-mass prior to a prior quantifying the notion that a big depar-
ture from the assumption in question is not likely. Often a ‘double-win’ is observed in such contexts. That is, the mean of the LPD arising from the correct but nonidentified model tends to be closer to the target than the point-mass limit arising from an identified but misspecified model. Moreover, the LPD may have reasonable coverage properties (as discussed above). In contrast, if misspecification induces a point-mass LPD on the target at the wrong value, interval estimator coverage will necessarily tend to zero as the sample size grows.

2.3 Further Considerations: Study Design and Sample Size Criteria

Inferences which grow more precise at a $n^{1/2}$ rate are of course central to much of statistical thinking. Thus posterior distributions which may concentrate to some extent as data accumulate but then settle at a non-degenerate limit do not fit with this dogma. Gustafson (2006) attempts to clarify thinking in this regard by comparing optimal sample sizes when the same sample size criterion is applied to both a nonidentified model for the real target of interest $g(\phi)$ and the embedded identified model for the wrong target $\tilde{g}_r(\phi_I)$. The crux of the issue is that under the nonidentified model the posterior variance of the target will behave as $a + bn^{-1}$ (with $a > 0$, $b > 0$). Thus the gain from additional data will become negligible more quickly in the nonidentified case than in the identified case (with posterior variance behaving as $cn^{-1}$). This is quantified via a fairly broad family of criterion functions which balance off information gain from the accumulation of data in the present study with the gain of allocating resources to other studies of different phenomena. With respect to any criterion function in this family, the optimal sample size under the nonidentified model is shown to be smaller than the optimal sample size for the corresponding identified model. Thus, in taking the lack of identification which can arise from study limitations seriously, one must think differently about the accumulation of data.

2.4 Further Considerations: Non-transparent Problems

Some nonidentified models cannot easily be reparameterized along the lines described in Section 2.1. Determining the LPD without the aid of a transparent reparameterization seems difficult, and a general approach is lacking. Gustafson (2009) gives an algorithm which works on problems where an amenable ‘overparameterization’ can be found. That is one can work with
\( \phi_I \) and \( \phi_N \) such that, as in transparent problems, the distribution of observables depends only on \( \phi_I \), and the original parameters \( \theta \) can be obtained (analytically, or at least computationally) from \( (\phi_I, \phi_N) \). However, it may be the case that \( \dim(\phi_I) + \dim(\phi_N) > \dim(\theta) \), and that the parameter space expressed in terms of \( (\phi_I, \phi_N) \) involves constraints. The algorithm proposed in Gustafson (2009) yields a weighted Monte Carlo sample which represents the LPD. This is applied to a nonidentified model involving two correlated binary diagnostic tests applied to three populations, and to a nonidentified model involving a binary instrumental variable. It should be mentioned that there are classical results on local identifiability concerning the Jacobian of the mapping from \( \theta \) to \( \phi_I \) (Rothenberg 1971; Goodman 1974), and these can be useful in approaching difficult nonidentified models which do not readily yield a transparent reparameterization. A very interesting application of these tools to models arising in diagnostic testing settings is given by Jones et. al. (2009), and there are links between this work and Gustafson (2009).

3 Shape of the Limiting Posterior Distribution

3.1 Summarizing the Performance of the Limiting Posterior Distribution

Having outlined some general features of Bayesian inference in nonidentified models, we now consider the specific question of how much the shape of the posterior is useful over and above the support of the posterior. Generically, let \( \delta(\cdot; \phi_I) \) be the large sample limit of a distribution describing a posteriori knowledge of target parameter \( \psi \), where conditioning on data in the \( n \to \infty \) limit equates with conditioning on \( \phi_I \). Two choices of \( \delta() \) are considered. The first is the LPD on \( \psi \), which, as discussed, is identically the prior conditional distribution of \( \psi|\phi_I \) evaluated at the true value of \( \phi_I \). This will have the identification region for \( \psi \) as its support. The second limiting estimator is simply the prior marginal distribution of \( \psi \) truncated to the identification region. We refer to this as the truncation-only distribution (TOD). The idea is that the TOD ignores any information in the data about the relative plausibility of different points inside the identification region. Thus any extent to which the LPD tends to be more concentrated around the true value than the TOD reflects a useful contribution of the shape of the LPD, above and beyond the support of the LPD. Put another way, a difference between the two concentrations represents information that is ‘left on the table’ by an analysis which
regards the identification region itself as the sole and final target of inference.

For a choice of $\delta(\cdot; \phi_I)$, such as the LPD or the TOD, the uncertainty about the target $\psi$ which arises when the true parameter vector is $\theta$ is summarized via

$$
    u_\pi(\theta) = E_\delta \left\{ (\Psi - \psi)^2 | \phi_I \right\} \\
    = \left\{ E_\delta(\Psi|\phi_I) - \psi \right\}^2 + Var_\delta(\Psi|\phi_I).
$$

(1)

Note that here $u_\pi(\theta)$ actually depends on $\theta$ only through $\phi_I$ and $\psi$. Noting also that both the LPD and the TOD depend on the specified prior $\theta \sim \pi$, an across-parameter-space summary of how closely the post-data knowledge matches the target arises by averaging (1) with respect to the prior. That is,

$$
    \bar{u}_\pi = E_\pi \{ u_\pi(\theta) \} \\
    = E_\pi \left[ \left\{ E_\delta(\Psi|\phi_I) - \psi \right\}^2 \right] + E_\pi Var_\delta(\Psi|\phi_I)
$$

(2)

summarizes the (across-parameter) average of the second moment of the limiting distribution about the true value.

Several remarks about (2) are in order. First, the first term on the right-hand side can be regarded as the Bayes risk for point estimation of $\psi$ under squared-error loss, which is indeed minimized by the posterior mean (or, more specifically, the mean of the LPD in the limiting case). There may be tertiary interest in evaluating how much excess Bayes risk the TOD carries compared to the LPD, but by itself this does not address the main issue of whether the LPD tends to be usefully more concentrated than the LPD.

A second comment is that (2) must be considered carefully, in that by itself it does not penalize a falsely-precise $\delta$, and in fact (2) as a whole would be minimized by taking $\delta$ to be a point-mass located at the mean of the LPD. Thus it only makes sense to consider (2) for a $\delta$ which does not give a falsely precise representation of knowledge concerning the target $\psi$. The LPD is indeed not falsely precise, since it has the usual calibration property of a Bayesian procedure with respect to joint sampling of parameters and data. In the limiting case this calibration can be expressed as the prior distribution inducing a uniform distribution for $h(\phi_I, \psi) = Pr_\delta(\Psi < \psi|\phi_I)$. By construction the TOD cannot be falsely precise, since it inherits its shape directly from the prior distribution.
3.2 Example: Prevalence of a Binary Trait Subject to Misclassification

Consider estimating the prevalence of a binary trait $X$ from data on an imperfect surrogate $X^*$. The quality of the surrogate is described by its sensitivity, $SN = Pr(X^* = 1|X = 1)$, and its specificity, $SP = Pr(X^* = 0|X = 0)$. Thus the parameter of interest, $r = Pr(X = 1)$, is linked to the prevalence of the surrogate, $\tilde{r} = Pr(X^* = 1)$, according to:

$$\tilde{r} = rSN + (1 - r)(1 - SP).$$

(3)

Now, say that bounds are available on the extent of misclassification. That is, it is known a priori that $SN > a$ and $SP > b$. We presume $a + b > 1$ (i.e., the bounds exclude a purely random ‘coin-flip’ classification scheme). From the form of (3) plus knowledge of $\tilde{r}$ we have the identification region for $r$ as

$$1 - \max \left\{ \frac{1 - \tilde{r}}{b}, 0 \right\} < r < \min \left\{ \frac{\tilde{r}}{a}, 1 \right\}.$$ 

(4)

That is, the a priori bounds plus an infinite sample of $X^*$ values (which reveals $\tilde{r}$) yields (4) as an inferential statement about the target of inference $r$.

In a similar vein, consider a Bayesian analysis arising from uniform prior distributions consistent with the a priori bounds. That is, the prior

$$\pi(r, SN, SP) = (1 - a)^{-1}(1 - b)^{-1}I_{(0,1)}(r)I_{(a,1)}(SN)I_{(b,1)}(SP)$$

(5)

is employed. The resulting large-sample limit of the marginal posterior distribution on $r$ is then simply the conditional distribution of $(r|\tilde{r})$ induced by (5), evaluated at the true value of $\tilde{r}$. Direct manipulation starting with (5) gives this conditional density as

$$\pi(r|\tilde{r}) \propto r^{-1} \left( \min \left\{ \frac{\tilde{r} - ar}{1 - r}, 1 - b \right\} - \max \left\{ \frac{\tilde{r} - r}{1 - r}, 0 \right\} \right)$$

$$\propto \min \left\{ \frac{1 - a}{1 - r}, \frac{1 - b}{r}, \frac{\tilde{r} - ar}{r(1 - r)}, (1 - \tilde{r}) - b(1 - r) \right\}.$$ 

(6)

over the support defined by (4). Some examples of the shape of the density (6) are given in Figure 1, for various values of $\tilde{r}$, when the prior bounds on the classification parameters are given by $a = 0.7$ and $b = 0.9$. A uniform prior on the interest parameter $r$ is seen to yield a somewhat or very ‘pointy’ limiting posterior distribution over the identification region.

The contribution of the shape of the LPD to inference can be quantified along the lines described in Section 3.1. Again using the prior based on $(a, b) =$
Figure 1: Limiting posterior marginal distribution of trait prevalence $r$ arising for various values of $\tilde{r}$. The prior bounds the classification sensitivity to be above $a = 0.7$ and the specificity to be above $b = 0.9$. (0.7, 0.9) for illustration, 40 000 values of $\theta = (r, SN, SP)$ are drawn from the prior, with the LPD (6) being evaluated in each instance. The average uncertainty is seen to be $\tilde{u}_\pi = 0.083^2$ (reported in this format for ease of interpretation as an expected squared deviation). In contrast, the TOD, which in this example is a uniform distribution over the identification region, achieves $u_\pi$ which is 18% higher. Thus a non-negligible amount of information is ‘left on the table’ if the focus is simply on aiming at the identification region itself as the target of inference.

The excess uncertainty of the TOD relative to the LPD is seen to vary with the choice of prior (i.e., the choice of a priori bounding values $a$ and $b$). For DOI: 10.2202/1557-4679.1206
instance, $\bar{u}_x$ is 27% higher for the TOD than the LPD when $(a, b) = (0.9, 0.9)$, 28% higher when $(a, b) = (0.8, 0.8)$, 27% higher when $(a, b) = (0.9, 0.9)$, but only 9% higher when $(a, b) = (0.6, 0.95)$.

3.3 Example: Gene-Environment Interaction with Data on Marginals only

Gustafson and Burstyn (2009) (hereafter GB09) characterize the LPD in a specific model for gene-environment interaction with binary variables. Their characterization is used here to evaluate the utility of the shape of the LPD in inferring the target parameter, along the lines of Section 3.1. The framework of GB09 involves a binary disease outcome $Y$ linked to a binary ‘environment’ variable $X$ (regarded as absence/presence of an environmental exposure), and a binary ‘gene’ variable $G$, according to

$$\text{logit} Pr(Y = 1|X, G) = \beta_0 + \beta_x X + \beta_{xg} XG.$$  \hspace{1cm} (7)

Note that this describes a situation where genotype is thought to perhaps modify the risk associated with the environmental exposure, but genotype does not affect the disease risk in the absence of environmental exposure, i.e., the model lacks a main effect for $G$. Secondly, the prevalence of environmental exposure is described as

$$\text{logit} Pr(X = 1|G) = \gamma.$$  \hspace{1cm} (8)

That is, the gene and environment variables are known to be independent in the study population. There is a considerable literature on situations where gene-environment independence is a plausible assumption and can be exploited to statistical advantage. Such an assumption can lead to more efficient inference in case-control studies (see, for instance, Umbach and Weinberg 1997, Chatterjee and Carroll 2005). There is also a considerable literature on exploiting independence of genotype and confounders, typically under the heading of Mendelian randomization (see for instance Davey Smith and Ebrahim 2003, 2004). GB09 discuss at length the two central model assumptions at play here, namely the absence of a $G$ term in (7), and the same omission in (8).

Together (7) and (8) define a model for $(Y, X|G)$ containing four unknown parameters. Clearly a cohort study in which $(Y, X, G)$ are all observed yields consistent estimation of all parameters. However, GB09 study the problem of inference from a much more limited data source. In particular, only the $(Y|G)$
conditional distribution and the \( X \) marginal distribution (or more precisely the \( (X|G) \) conditional which is known to not depend on \( G \)) are learned from data. Clearly this is a nonidentified model setting. The \( (Y|G) \) distribution allows estimation of two parameters and the \( X \) distribution allows estimation of one parameter. Thus identification is ruled out, and it is not possible to estimate all four unknown parameters consistently.

Let \( \theta = (\beta_0, \beta_x, \beta_{xg}, \gamma) \) denote the unknown parameters, and focus on \( \beta_{xg} \) as the target of inference. GB09 determine that this is a partially identified problem, in that the identification region for \( \beta_{xg} \) is smaller than the whole parameter space (presumed to be the real line). In particular, presuming that the true value of \( \beta_{xg} \) is non-zero, the identification region is bounded away from (and on the correct side of) zero.

To elucidate the bound, GB09 reparameterize to \((p_0, p_1, q, r)\), where \( p_i = \text{Pr}(Y = 1|G = i) \), \( q = \text{Pr}(Y = 1|G = 0, X = 0) \), and \( r = \text{Pr}(X = 1) \). As sample sizes go to infinity, conditioning on \((Y|G)\) and \((X)\) samples corresponds to conditioning on correct values of \((p_0, p_1, r)\). GB09 determine the induced prior distribution for \((q|p_0, p_1, r)\) arising from a ‘full support’ prior on the original parameters, and show that the support of this conditional distribution will be smaller than the unit interval for some values of \((p_0, p_1, r)\). Moreover, since the target parameter can be written as

\[
\beta_{xg} = h(q; p_0, p_1, r) = \logit \left( \frac{p_1}{r} \right) - \logit \left( \frac{p_0 - (1-r)q}{r} \right), \quad (9)
\]

the LPD for the target can be determined via transformation of the prior conditional for \((q|p_0, p_1, r)\). Clearly if \( p_0 = p_1 \) then the target LPD will be a point-mass at \( \beta_{xg} = 0 \). Thus we focus on the case \( p_0 \neq p_1 \), considering \( p_0 < p_1 \) without loss of generality.

Since \( h(\cdot; p_0, p_1, r) \) can be verified to be symmetric about, and minimized at \( q^*(p_0, p_1, r) = (p_0 + p_1 + r)/(2(1-r)) \), it immediately follows that the left endpoint of the support of the target LPD is given by \( h() \) evaluated either at the left endpoint of the \((q|p_0, p_1, r)\) support (if \( q^* \) is further left than this), or at \( q^* \) (if \( q^* \) is inside the support), or at the right endpoint on the support (if \( q^* \) is further right than this). Working through the algebraic details, the left endpoint of the support of the LPD for \( \beta_{xg} \) can be expressed as:

\[
b(p_0, p_1, r) = \begin{cases} 
\logit(p_1/r) - \logit(p_0/r) & \text{if } p_0 + p_1 \leq r, \\
2\logit((r + p_1 - p_0)/(2r)) & \text{if } r < p_0 + p_1 < 2 - r, \\
\logit((1 - p_0)/r) & \text{if } p_0 + p_1 > 2 - r.
\end{cases}
\]
This constitutes an identified bound away from zero on the target \( \beta_{xg} \), i.e.,
\[ b(p_0, p_1, r) < \beta_{xg} < \infty \]
is the identification region. It is also apparent that in the case \( r < p_0 + p_1 < 2 - r \), the LPD for \( \beta_{xg} \) will have infinite density at the left endpoint \( b(p_0, p_1, r) \). This is an immediate consequence of (9) being minimized at a point \( q^* \) in the interior of the support for \( (q|p_0, p_1, r) \).

As an example, consider the prior specification \( \logit \beta_0 \sim U(0, 1) \), \( \beta_x \sim N(0, 1) \), \( \beta_{xg} \sim N(0, 1) \). Such a prior could be quite reasonable in many epidemiological settings where odds-ratios beyond \( \exp(\pm 2) \) would indeed be implausible. It is easily seen that the LPD and TOD will not depend on the choice of prior for \( \gamma \) (since this parameter is consistently estimated). The resulting LPD and TOD for various values of the underlying parameters appear in Figure 2. For many values it transpires that \( b(p_0, p_1, r) \) is quite a tight lower bound for \( \beta_{xg} \). The combination of a tight lower bound plus the tendency for the LPD to put much mass near the bound results in an LPD which may do somewhat well for inferring the target. The improvement of the LPD over the TOD for estimating \( \beta_{xg} \) is fairly evident from Figure 2, at least for the parameter values considered.

To add generality to the findings, we simulate 50 values of \( \theta \) from the prior distribution described above for \( \beta \), augmented with \( \text{expit} \gamma \sim U(0, 1) \). The aggregate performances of the TOD and the LPD are reported in Table 1. For this problem the LPD is much better than the TOD, with more than a fourfold difference in \( \bar{u}_\pi \). The table also reports the two components of \( \bar{u}_\pi \), i.e., the Bayes risk of the mean of the limiting distribution as a point estimator, and the average variance of the limiting distribution. The roughly fourfold improvement of the LPD relative to the TOD is seen to hold for both components. That is, the shape of the LPD is useful both in terms of closeness of the mean to the true value and smallness of the variance. This is also reflected in Figure 3, which plots, for the fifty simulated values of \( \theta \), the means of the LPD and the TOD against the target, and against each other.

<table>
<thead>
<tr>
<th></th>
<th>BR</th>
<th>AV</th>
<th>( \bar{u}_\pi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOD</td>
<td>0.52</td>
<td>0.50</td>
<td>0.72²</td>
</tr>
<tr>
<td>LPD</td>
<td>0.21</td>
<td>0.27</td>
<td>0.34²</td>
</tr>
</tbody>
</table>

Table 1: Performance of the TOD and the LPD in estimating \( \beta_{xg} \). The last column gives \( \bar{u}_\pi \), the across parameter average of the limiting distribution’s second moment around the true value. The first and second columns decompose \( \bar{u}_\pi \) into the two constituent terms in (2), namely the Bayes risk (BR) associated with the limiting mean and the average variance (AV) of the limiting distribution. For ease of interpretation, all entries are expressed as squares.
Figure 2: Densities of the LPD and TOD for $\beta_{xg}$, arising from various underlying parameter values. In all cases the true value is $\beta_{xg} = 0.5$. The rows correspond to $\gamma = \text{logit} 0.1$, $\gamma = \text{logit} 0.25$, and $\gamma = \text{logit} 0.4$ respectively. The columns correspond to $(\beta_0, \beta_x) = (-2, 0.5)$, $(\beta_0, \beta_x) = (-1, 0.5)$, and $(\beta_0, \beta_x) = (-2, 1)$ respectively. In each panel the shading indicates values outside the identification region, while the dotted vertical line indicates the true value ($\beta_{xg} = 0.5$). In all cases the TOD density has a thicker tail than the LPD density.

4 Discussion

In brief, inference in partially identified models is partially helpful. An infinite data sample does not reveal the true value of the target parameter, but it does rule out some values. Moreover, the limiting posterior distribution over the identification region on aggregate has a shape which is better concentrated...
Figure 3: Comparison of limiting estimates and true value for $\beta_{xg}$. The first panel plots the mean of the TOD against the true value. The second panel plots the mean of the LPD against the true value. The third panel plots the mean of the TOD against the mean of the LPD. The plotted points arise from sampling parameter values from the prior distribution described in the text.

around the true value than is the prior distribution truncated to the identification region. Generally it is important to understand both the infinite and finite sample size performance of Bayesian inference in nonidentified models, since such models are inevitably encountered when trying to realistically encapsulate the study and data limitations arising in causal modeling of observational data.

Interestingly, Moon and Schorfheide (2009), in contrasting Bayesian and frequentist inference for partially identified models, write that:
The main challenge to Bayesian inference is to control the shape of the prior distribution on the identified set conditional on the reduced form parameter to avoid highly informative priors on the identified set induced by nonlinearities of parameter transformations and to document the sensitivity of posterior inference to the choice of prior even in large samples.

The suggestion to check sensitivity to the prior is well taken. However, the findings for the two examples in Section 3 run counter to Moon and Schorfheide’s suggestion about trying to control the shape of the prior induced by reparameterization. Starting with a defensible and weakly informative prior in the original parameterization can indeed lead to a ‘strongly shaped’ prior for \((\phi_N|\phi_I)\), but this shape is helpful in aggregate, not detrimental. In particular, the limit of the posterior distribution on the target parameter tends to be better concentrated around the true value than does the prior distribution truncated to the identification region. Thus Bayes theorem applied to partially identified models works on two fronts. As data accumulate, values of the target outside the identification region get discredited, and values inside get weighted appropriately with respect to one another.

Much work remains to be done in the area of Bayesian inference for partially identified models. One avenue for further research involves the application of such inference in a wider array of applied problems, and particularly in problems which are central to causal inference. Experience with the Bayesian approach currently lags behind that with non-Bayesian methods. For instance, Manski (2003, Ch. 7-10) investigates identification regions arising when counterfactual outcome models are applied to study treatment-response relationships, but, as far as this author is aware, the corresponding Bayesian solutions remain to be developed. Another very open question would seem to be that of model comparison or model averaging in the face of multiple plausible models, when one or more of these models is not fully identified. Right away, the non-standard asymptotics of nonidentified models precludes the use of the asymptotically motivated model-choice criteria that are often used with Bayesian models (e.g., BIC or DIC). Fully Bayesian model comparison (for instance, using Bayes factors) would seem to be a possibility, but the large-sample properties of such a procedure are unclear in the nonidentified case, and computational challenges are likely to arise. Work on this topic is clearly needed.
References


DOI: 10.2202/1557-4679.1206