CONSISTENCY AND IDENTIFIABILITY IN BAYESIAN ANALYSIS

B. O'NEILL[*], *Australian National University*[**]

WRITTEN 23 SEPTEMBER 2005.

Abstract

The importance of posterior consistency in the robustness of Bayesian analysis is examined and discussed. The notions of sufficient and minimal sufficient parameters are introduced and important consistency results for such parameters are derived. We see that minimal sufficient parameters are fundamental in characterising the relationship between data and parameters. The concept of identifiability is then introduced and several equivalent definitions are given. The relationship between consistency and identifiability is examined and means of establishing identifiability are examined with a view to finding useful practical tests of identifiability. These results are applied to a simple example involving non response.

BAYESIAN STATISTICS; LIKELIHOOD; ROBUSTNESS; CONSISTENCY; MINIMAL SUFFICIENT PARAMETER; IDENTIFIABILITY; NON RESPONSE.

This paper arose out of an examination of survey problems involving non response. In such problems we are faced with the dilemma of attempting to make inferences about some quantity of interest in the presence of nuisance parameters (the nuisance parameters in this case being the probability of response conditional on the quantity of interest). This is nothing new. Problems involving nuisance parameters abound in statistics and, within the Bayesian paradigm, present no more of a theoretical challenge than problems without nuisance parameters.

However, in the case of survey problems involving non response it turns out that under a wide enough specification of possible priors we can be led to any inferences. Moreover, this holds regardless of the amount of data we have observed. Thus, it is clear that such problems are highly dependent upon assumptions regarding the propensity to respond. Intuitively, we may feel that there is an inherent limit on our ability to make inferences in these cases; we may even feel that such problems are so sensitive to assumptions that the data really isn't telling us anything at all. So is there something fundamentally different about these problems from other problems involving nuisance parameters? This paper answers this question using the concept of identifiability. This concept is explored within the Bayesian paradigm and its relationship to consistency and general robustness is explained.

[*] e-mail address: ben.oneill@anu.edu.au
[**] School of Finance and Applied Statistics, Australian National University, ACT, 0200, Australia.

1. The dichotomy between likelihood and prior beliefs

Let $\mathbf{x} \equiv (x_1, x_2, x_3, ...)$ and $\mathbf{x}_k \equiv (x_1, x_2, ..., x_k)$ and suppose that we are interested in making inferences about a quantity θ. From the Radon-Nikodym Theorem, for any quantity (θ, π) for which $\mathcal{P}(\theta, \pi)$ dominates $p(\mathbf{x}_k)$, we have:

$$p(\mathbf{x}_k) = \int p(\mathbf{x}_k | \theta, \pi) d\mathcal{P}(\theta, \pi)$$

From the Representation Theorem of de Finetti (1980) and the work of Fortini, Ladelli and Regazzini (2000) we know that if $\mathbf{x}$ is an exchangeable Polish space then there exists some quantity π so that the elements of $\mathbf{x}|\theta, \pi$ are independent and identically distributed, leading us to the general predictive model:

$$p(\mathbf{x}_k) = \int p(\mathbf{x}_k | \theta, \pi) dP(\theta, \pi) = \int \left(\prod_{i=1}^{k} p(x_i | \theta, \pi) \right) dP(\theta, \pi)$$

with posterior:

$$p(\theta, \pi | \mathbf{x}_k) \propto \left(\prod_{i=1}^{k} L_{x_i}(\theta, \pi) \right) p(\theta, \pi) = \exp\left(\sum_{i=1}^{k} l_{x_i}(\theta, \pi) \right) p(\theta, \pi)$$

where $L_x(\theta, \pi) \equiv p(x | \theta, \pi)$ is the likelihood and $l_x(\theta, \pi) \equiv \ln(L_x(\theta, \pi))$ is the log-likelihood. This result justifies working within the common Bayesian paradigm of a model dichotomized between likelihood and prior. Needless to say, this mathematical form arises under a wide class of Bayesian models[1]. Bernardo and Smith (1994) rightly stress that both the likelihood and prior are essential to the predictive model and are both the consequences of the exchangeability assumption. However, this does not suggest that there are no differences between formulating beliefs about observables and formulating beliefs about unobservable parameters. In particular, due to experience, we should expect that we are more adept at making judgements about quantities that are —at least in principle— observable than those that are not. After all, we receive feedback, by observation, about the correctness of our judgments about the former, but not the latter. Thus, it is not surprising that, within the Bayesian paradigm, greater debate surrounds the choice of an appropriate prior than the choice of an appropriate likelihood.

[1] Further invariance conditions (beyond exchangeability) may ensure that the likelihood takes on a particular parametric form.

2. Robustness testing using sensitivity analysis

Since we are not confident of our ability to make judgements about quantities that are not observable we are lead naturally to ask how much our conclusions depend upon these judgments; this leads us naturally to the concept of robustness testing; that is, testing the sensitivity of our inferences to our prior beliefs. The issue of robustness is critical to Bayesian analysis. Indeed, the most common criticism of the Bayesian paradigm from frequentist practitioners is the subjectivity inherent in making precise —and often fairly arbitrary— prior judgements. These criticisms can be met by using sensitivity analysis to test the robustness of inferences and predictions and by using wide classes of priors to represent ignorance. Unfortunately, in a certain sense, our inferences are *always* highly sensitive to our prior beliefs.

THEOREM 1 (Prior Sensitivity Theorem): For $k \in \mathbb{N}$ and for any posterior belief $\mathcal{P}_k$ on the support for θ of $L_{\mathbf{x}_k}$ there exists a prior belief $\mathcal{P}$ for (θ, π) satisfying:

$$\mathcal{P}_k(\mathcal{A}) \propto \int_{(\theta,\pi):\theta\in\mathcal{A}} L_{\mathbf{x}_k}(\theta,\pi)\, d\mathcal{P}(\theta,\pi) \text{ for all } \mathcal{A}.$$

PROOF: See Appendix. ■

It follows that, with a finite amount of data, by appropriate choice of prior belief we can obtain *any* posterior belief that is not completely contradicted by the data. This means that we are not able to obtain useful results unless we restrict our attention to some class of priors that is smaller than the class of all possible prior beliefs.

Thus, instead of considering all possible prior beliefs we specify some narrower class of prior beliefs (which may still be extremely wide if we consider ourselves completely ignorant) and analyse how our inferences change for these different priors. If our inferences are fairly similar regardless of which prior we use, then we may be confident in our analysis regardless of our confidence in our prior beliefs. If our inferences are very different under different prior beliefs then we may not be so confident. In general we determine our class of reasonable priors

by reference to available prior evidence. In the absence of such evidence, Walley (1991) suggests that we model prior ignorance by using a class of priors that is *vacuous* —in the sense that *a priori* our inferential and predictive probabilities of interest vary over the entire unit interval. However, there are unlimited classes that meet this criterion, each having different consequences *a posteriori* and therefore having different consequences in terms of robustness.

While helpful, this method and others have one important shortcoming in that they may invite further questions of sensitivity. After all, did our robustness analysis depend heavily on the chosen class of priors, leading us to wonder about the robustness of our robustness analysis, and so on, *ad infinitum*?

This shortcoming of sensitivity analysis invites us to pursue more objective methods of testing the sensitivity of our inferences to our prior beliefs; that is, methods that do not require any further assumptions —assumptions that may themselves induce questions of sensitivity. Since the existence of a prior belief inducing a particular posterior belief is ensured only for finite observations it should already be evident that perfect information may hold the key. To explore this line of thought further we introduce the notion of consistency.

3. Posterior consistency

DEFINITION 1 (Consistency): For any hypothesis $\phi \in \mathcal{A}$ and given posterior belief $\mathcal{P}_k(\mathcal{A}) \equiv P(\phi \in \mathcal{A} | \mathbf{x}_k)$ we let $\mathcal{P}_\infty(\mathcal{A}) \equiv \lim_{k\to\infty} \mathcal{P}_k(\mathcal{A})$ be our posterior belief about the hypothesis under perfect information. We say that our beliefs about the hypothesis are *consistent* if and only if $\mathcal{P}_\infty(\mathcal{A}) = I(\phi \in \mathcal{A})$. Further, we say that our beliefs about ϕ are *consistent* if and only if $\mathcal{P}_\infty(\mathcal{A}) = I(\phi \in \mathcal{A})$ for all $\mathcal{A}$.

Consistency asserts that our belief about the parameter converges —under perfect information— to certain belief in the true values of the parameter. At this point it is prudent to raise and rebut a common and flawed objection to the above analysis and to remind the reader of the correct approach. The objection in

question is the assertion that, within the Bayesian paradigm, there is no such thing as a true parameter; that the parameter is not an existent but an abstraction which has no true value. However, the operational approach to statistics —as expounded by de Finetti— renders this argument invalid. Under the operational approach all parameters are functions of the empirical distribution of the appropriate superpopulations and are thereby reducible to some aspect of reality if only in limiting form; parameters are not merely some abstract index to a probabilistic belief. To talk of beliefs about a parameter that cannot be reduced to some aspect of reality is *arbitrary* in the sense expounded by Peikoff (1993):

> An arbitrary statement has no relation to man's means of knowledge. Since the statement is detached from the realm of evidence, no process of logic can assess it.

Such statements are inadmissible; no meaningful discussion can be made concerning a parameter which is defined merely by a floating abstraction. Rather, we must define our terms by reference to reality (operationally) so that parameters are indeed aspects of reality with a true value.

4. Minimal sufficient parameters

In order to determine a useful test of consistency we analyze the way in which data provides us with information regarding the parameters. It is well known that statistical models provide us with information from the data only through the value of the minimal sufficient statistic (with respect to our likelihood and the parameters of interest). However, it is not always appreciated that this property also applies to parameters; that is, statistical models provide us with information about the parameters only through the value of the *minimal sufficient parameter*. To see this we introduce the notions of the sufficient and minimal sufficient parameter following Barankin (1961).

DEFINITION 2 (Sufficient and minimal sufficient parameters): Given a likelihood function L_x with parameter (θ, π), the parameter $\phi \equiv \Lambda(\theta, \pi)$ is said to be a *sufficient parameter* if it is sufficient for x in the usual sense. Moreover, ϕ is said to be a *minimal sufficient parameter* if it is a sufficient parameter and can be expressed as a function of any sufficient parameter.

Sufficient and minimal sufficient parameters are almost entirely analogous to sufficient and minimal sufficient statistics. Indeed, the only difference is that the likelihood function —considered as a function of the parameters— induces only a sigma finite measure rather than a probability measure on the parameter space; this property is not needed for most theorems involving sufficiency and minimal sufficiency so that the same properties generally apply. In particular, we note that there is always a minimal sufficient parameter and therefore a sufficient parameter since the function L_x is itself a minimal sufficient parameter. In fact, proceeding analogously to Lehmann and Scheffé (1951) it can easily be shown that any minimal sufficient parameter ϕ is an injective (that is, one to one) function of L_x. Just as minimal sufficient statistics are able to capture and summarise all relevant information supplied by data, so too, minimal sufficient parameters are able capture and summarise all relevant information about the parameters.

4.1. Minimal sufficient parameters and inference

Sufficient parameters have certain useful informatory properties which ensure that, in a useful sense, they completely characterize the likelihood function. If ϕ is a sufficient parameter then we have:

$$L_x(\theta,\pi) = h(x,\phi)\, g(\theta,\pi)$$

for some functions h and g by the Neyman Factorization Theorem. It follows that:

$$\begin{aligned} p(x,\theta,\pi|\phi) &= p(x|\theta,\pi,\phi)\, p(\theta,\pi|\phi) \\ &= L_x(\theta,\pi)\, p(\theta,\pi|\phi) \\ &= h(x,\phi)\big(g(\theta,\pi)\, p(\theta,\pi|\phi)\big) \end{aligned}$$

so that $x \perp (\theta,\pi)\big|\phi$. That is, regardless of the likelihood function, the data are independent of the parameters given knowledge of any sufficient parameter. Indeed, Dawid (1979) defines sufficiency directly in terms of conditional independence.

This conditional independence relation has three important consequences. Firstly, if ϕ is a sufficient parameter, then letting $L_x(\phi) \equiv p(x|\phi)$ we have:

$$L_x(\theta,\pi) = p(x|\theta,\pi) = p(x|\theta,\pi,\phi) = p(x|\phi) = L_x(\phi)$$

so that the likelihood function depends upon the parameters (θ,π) only through sufficient parameters; this means that the likelihood function can be written in terms of any sufficient parameter.

Secondly, if ϕ is a sufficient parameter, we have:

$$p(\mathbf{x}_k) = \int \left(\prod_{i=1}^{k} L_{x_i}(\phi) \right) dP(\phi) = \mathrm{E}\left(\prod_{i=1}^{k} L_{x_i}(\phi) \right)$$

so that the predictive distribution depends upon the parameters (θ,π) only through sufficient parameters.

Finally, if ϕ is a sufficient parameter, we have:

$$p(\theta,\pi|\mathbf{x}_k) = \int p(\theta,\pi|\phi)\, dP(\phi|\mathbf{x}_k) = \mathrm{E}\left(p(\theta,\pi|\phi) \middle| \mathbf{x}_k \right)$$

so that the inferential distribution of the parameters depends upon the parameters (θ,π) only through sufficient parameters. This holds also for any function of the parameters. In particular, if $\tau = f(\theta,\pi)$ we have $x \perp \tau|\phi$ so that:

$$p(\tau|\mathbf{x}_k) = \int p(\tau|\phi)\, dP(\phi|\mathbf{x}_k) = \mathrm{E}\left(p(\tau|\phi) \middle| \mathbf{x}_k \right)$$

It is pertinent to note that $p(\theta,\pi|\phi)$ and $p(\tau|\phi)$ are determined by our prior beliefs so that these results are of fundamental importance in considering robustness.

Since these properties hold for all sufficient parameters, it follows that the likelihood function and inferential distribution depend upon the parameters only through minimal sufficient parameters. Thus, regardless of the form of the likelihood function —and thus regardless of our model assumptions— we know that the data provides us with information about the parameter of interest only through the minimal sufficient parameter.

4.2. Minimal sufficient parameters and inference under perfect information

In addition to characterizing the likelihood function, minimal sufficient parameters have an important asymptotic property that is of fundamental importance in considering robustness to prior beliefs. This property is an extension of widely known asymptotic results for Bayesian statistics. To facilitate these results we let $\mathcal{P}_k(\mathcal{A}) \equiv P(\phi \in \mathcal{A} | \mathbf{x}_k)$ be our posterior belief for $\phi \in \mathcal{A}$ given knowledge of $\mathbf{x}_k$ and we let $\mathcal{P}_\infty(\mathcal{A}) \equiv \lim_{k\to\infty} \mathcal{P}_k(\mathcal{A})$ be our posterior belief about ϕ under perfect information.

DEFINITION 3 (The Wald function): Given likelihood L_x that has sufficient parameter ϕ with range Φ, let λ be the *Wald function of* ϕ defined by $\lambda(r) \equiv \mathrm{E}(l_x(r)|\phi)$ and let $\lambda^*(\mathcal{A}) \equiv \sup_{r\in\mathcal{A}} \lambda(r)$ and $\Phi^* \equiv \arg\sup_{r\in\Phi} \lambda(r)$.

THEOREM 2 (Convergence of sufficient parameters): If the elements of $\mathbf{x}|\phi$ are independent with likelihood L_x that has sufficient parameter ϕ then —under the regularity conditions of Berk (1970)— if $\lambda^*(\mathcal{A}) < \lambda^*(\Phi)$ then $P(\mathcal{P}_\infty(\mathcal{A}) = 0) = 1$.

THEOREM 3 (Convergence of sufficient parameters): If the elements of $\mathbf{x}|\phi$ are independent with likelihood L_x that has sufficient parameter ϕ then —under the regularity conditions of Berk (1970)— if Φ^* is not empty then the hypothesis $\phi \in \Phi^*$ is almost surely consistent; that is, we have $P(\mathcal{P}_\infty(\Phi^*) = 1) = 1$.

PROOFS: See Appendix. ■

Theorem 3 shows that, under wide regularity conditions, our posterior belief about any sufficient parameter converges —under perfect information— to certain belief in the set of values that maximize the Wald function. It turns out that for minimal sufficient parameters we have a stronger convergence result.

THEOREM 4 (Consistency of minimal sufficient parameters): If the elements of $\mathbf{x}|\phi$ are independent with likelihood L_x that has minimal sufficient parameter ϕ then —under the regularity conditions of Berk (1970)— the parameter ϕ is almost surely consistent; that is, we have $P\left(\mathcal{P}_\infty\left(\{\phi\}\right)=1\right)=1$.

PROOF: See Appendix. ■

Theorem 4 shows that, under wide regularity conditions, our posterior belief about any minimal sufficient parameter converges, under perfect information, to certain belief in the true value of that parameter. We note again that the operational approach to statistics ensures that the minimal sufficient parameter exists in the metaphysical sense and therefore has a true value.

We have now seen the relationship between consistency, sufficiency and minimal sufficiency for parameters. We now proceed to consider the notion of *identifiability* that seeks to determine whether the data gives information regarding the parameter of interest.

5. Identifiability

We have seen that, regardless of our model assumptions, the data provides us with information about the parameter of interest only through the minimal sufficient parameter. We have also seen that our beliefs about the minimal sufficient parameter will converge —under perfect information— to certain belief in the true value of that parameter. Whether this information about the minimal sufficient parameter in turn determines the parameter of interest is determined by the concept of *identifiability*.

DEFINITION 4 (Identifiability): Given a likelihood function L_x with sufficient parameter (θ,π) the parameter θ is said to be *identifiable* if and only if there exists a function g such that $\theta=g(\phi)$ where ϕ is a minimal sufficient parameter.

If θ is identifiable then we may determine the parameter of interest from our minimal sufficient parameter and so the data provides information directly about θ. Conversely, if θ is unidentifiable then we cannot determine the parameter of interest from any minimal sufficient parameter and so the only information about θ provided by the data is through the relationship between the minimal sufficient parameter and the parameter of interest (which is determined by our prior beliefs). We note finally that the choice of minimal sufficient parameter ϕ in the definition of identifiability is immaterial since all minimal sufficient parameters are injective transformations of one another. The notion of identifiability is expressed somewhat differently in the literature. Rothenberg (1971) and Bowden (1973) define *global identifiability* in terms of *observational equivalence* of parameters.

DEFINITION 5 (Observational equivalence): Given a likelihood function L_x the parameter values ϕ' and ϕ'' are said to be *observationally equivalent* if they are pairwise sufficient for x in the usual sense; that is, if $L_x(\phi') = L_x(\phi'')$ almost surely with respect to the sampling measure based on either parameter value.

DEFINITION 6 (Identifiability): Given a likelihood function L_x with sufficient parameter (θ, π) the parameter θ is said to be *globally identifiable* if and only if all observationally equivalent parameter values have the same value of θ.

Halmos and Savage (1949) and Bahadur (1954) show that sufficiency is equivalent to pairwise sufficiency for dominated sampling measures; even for measures that are not dominated, sufficiency implies pairwise sufficiency. It follows that all parameter values that correspond to the same value of a sufficient parameter are observationally equivalent. Thus we see that, for dominate sampling measures, the definition of identifiability in Definition 4 is equivalent to the definition of global identifiability in Definition 6, used by Rothenberg (1971), Bowden (1973) and others. In practice we generally deal only with sampling measures using a density or mass function so that we are always dealing with dominated sampling measures. We will therefore take Definitions 4 and 6 as equivalent definitions of *identifiability*.

5.1. Identifiability and inference under perfect information

Using the above posterior convergence results, the concept of identifiability leads us easily to another useful asymptotic result. To facilitate this result we will let $\mathcal{G}_k(\mathcal{A}) \equiv P(\theta \in \mathcal{A} | \mathbf{x}_k)$ be our posterior belief for $\theta \in \mathcal{A}$ given knowledge of $\mathbf{x}_k$ and we let $\mathcal{G}_\infty(\mathcal{A}) \equiv \lim_{k\to\infty} \mathcal{G}_k(\mathcal{A})$ be our posterior belief about θ under perfect information.

THEOREM 5 (Identifiability and consistency): If the elements of $\mathbf{x}|\phi$ are independent with likelihood L_x such that θ is identifiable then —under the regularity conditions of Berk (1970)— the parameter θ is almost surely consistent; that is, we have $P(\mathcal{G}_\infty(\{\theta\}) = 1) = 1$.

PROOF: Follows trivially from Theorem 4. ■

Theorem 5 shows that, if the parameter of interest is identifiable then, under wide regularity conditions, our posterior belief about the parameter of interest converges to certain belief in the true value of that parameter. We can see that identifiability is an important property of a statistical model. If the parameter of interest is identifiable then we know that the data is providing us with information about this parameter via the minimal sufficient parameter. However, if the parameter of interest is unidentifiable then the data is providing us with information that determines only the value of the minimal sufficient parameter, which may correspond with several possible values of the parameter of interest.

5.2. Tests for identifiability

An obvious test of identifiability is suggested by Definition 4; namely, determine a minimal sufficient parameter and then determine whether the parameter of interest is a function of that parameter. However, it may be difficult to find a minimal sufficient parameter (for a good algorithm see Johnson (1974)) and it may also be difficult to decide whether the minimal sufficient parameter can

be inverted to obtain the parameter of interest. Instead of relying on minimal sufficient parameters we can test for identifiability using sufficient parameters, which are easy to determine.

THEOREM 6 (Identifiability using sufficient parameters): Given a likelihood function L_x with sufficient parameter $\phi = \Lambda(\theta, \pi)$ the parameter θ is identifiable if and only if $\Lambda(\theta', \pi') = \Lambda(\theta'', \pi'')$ implies that $\theta' = \theta''$.

PROOF: Follows easily from Definition 4 and the fact that any minimal sufficient parameter is a function of ϕ. ■

Alternatively, we may wish to test identifiability according to Definition 4.6. An obvious test of identifiability is again suggested; namely, determine the conditions under which parameter values are observationally equivalent and then determine whether these conditions imply the equivalence of the parameter of interest. Following the work of Rothenberg (1971) and Bowden (1973) we can further simplify this test by using a distance function that measures the distance between two probability densities. Bowden (1973) tests for identifiability using the Kullback-Liebler distance function H defined by:

$$\begin{aligned} H(\theta', \pi'; \theta'', \pi'') &\equiv \mathrm{E}\left(\ln\left(R_x(\theta', \pi'; \theta'', \pi'') \right) \middle| \theta', \pi' \right) \\ &= \mathrm{E}\left(\ln\left(\frac{L_x(\theta', \pi')}{L_x(\theta'', \pi'')} \right) \middle| \theta', \pi' \right) \\ &= \mathrm{E}\left(l_x(\theta', \pi') - l_x(\theta'', \pi'') \middle| \theta', \pi' \right) \\ &= \int \left(l_x(\theta', \pi') - l_x(\theta'', \pi'') \right) dP\left(x \middle| \theta', \pi' \right) \end{aligned}$$

It can be shown that H is a distance function in the usual sense. In particular it can be shown that $H(\theta', \pi'; \theta'', \pi'') \geq 0$ with $H(\theta', \pi'; \theta'', \pi'') = 0$ if and only if (θ', π') and (θ'', π'') are observationally equivalent. This function —or indeed any other distance function— can be used to simplify the test of identifiability based on Definition 4.6 as follows.

THEOREM 7 (Identifiability using distance functions): Given a likelihood function L_x with sufficient parameter (θ,π) and a distance function H, the parameter θ is identifiable if and only if $H(\theta',\pi';\theta'',\pi'')=0$ implies that $\theta'=\theta''$.[2]

PROOF: Follows trivially from Definition 6 and the properties of H. ■

Thus, we have at our disposal, several equivalent definitions leading to methods of testing for identifiability. A more detailed analysis of the conditions required for identifiability is given in Roehrig (1988). We will not have need for a detailed discussion of these conditions since we are concerned here only with the identifiability of models for surveys subject to non response and self selection.

6. Identifiability and non response

The Representation Theorem of de Finetti shows that if **x** is exchangeable with elements having finite range $1,2,...,m$, then the elements follow a multinomial model with the long run proportions of outcomes as parameter (see also Fortini, Ladelli and Regazinni (2000)). We can see that this parameter is identifiable in the model of interest.

EXAMPLE 1: Suppose that **x** is exchangeable with elements having finite range $1,2,...,m$ so that we have likelihood:

$$L_x(\theta)=\prod_{i=1}^{m}\theta_i^{I(x=i)}$$

where $\theta\equiv(\theta_1,\theta_2,...,\theta_m)$ with:

$$\theta_i\equiv\lim_{n\to\infty}\frac{\sum_{j=1}^{n}I(x_j=i)}{n}.$$

Since θ is a sufficient parameter it follows that θ is identifiable. ■

[2] In the case when H is divergent we will —by convention— say that $H(\theta',\pi';\theta'',\pi'')=\infty>0$.

Example 1 shows us that, in the standard multinomial model, the parameter of the long run proportions of outcomes is identifiable. It follows from Theorem 5 that, under perfect information, our posterior beliefs will converge to certain belief in the true long run proportions. However, if instead of observing the elements of $\mathbf{x}$ directly we observe them subject to some possibility of non-response then we have the following model.

EXAMPLE 2: Continuing Example 1, suppose that $\mathbf{m} \equiv (m_1, m_2, m_3, ...)$ is an indicator of whether the associated values of $\mathbf{x}$ are missing and that we observe:

$$y_i \equiv \begin{cases} y_i & \text{if } m_i = 0 \\ \bullet & \text{if } m_i = 1 \end{cases}$$

If $\mathbf{m}$ is exchangeable then we have likelihood:

$$L_y(\theta, \pi) = \left(\prod_{i=1}^{m} (\theta_i \pi_i)^{I(y=i)} \right) \left(1 - \sum_{i=1}^{m} \theta_i \pi_i \right)^{I(y=\bullet)}$$

where $\pi \equiv (\pi_1, \pi_2, ..., \pi_m)$ with:

$$\pi_i \equiv \lim_{n \to \infty} \frac{\sum_{j=1}^{n} I(x_j = i, m_j = 0)}{\sum_{j=1}^{n} I(x_j = i)}$$

To see that θ is not identifiable we note that for any $\theta' \neq \theta''$ and $\pi_i'' = (\theta_i' \pi_i') / \theta_i''$ then we have $L_y(\theta', \pi') = L_y(\theta'', \pi'')$ for all y. ■

Example 2 shows us that, in the model in Example 2, the parameter of the long run proportions of outcomes is *not* identifiable. In fact it can be shown that the above model has minimal sufficient parameter $\phi \equiv (\phi_1, \phi_2, ..., \phi_m)$ with $\phi_i \equiv \theta_i \pi_i$ so that, under perfect information, our posterior beliefs about ϕ converge to certain belief in the true value of ϕ. Since $\theta \geq \phi$ perfect information restricts the possible range of the parameter of interest but it does not allow us to determine the parameter with certainty. This property is in fact what makes non response problems fundamentally different from most other problems involving nuisance parameters. In such problems the data gives us information, not about the parameter of interest, but about a minimal sufficient parameter from which we are unable to obtain the parameter of interest. We have seen that, in such cases, no

amount of data can overcome this problem. However, if we were to place restrictions on the possible values of π then we may obtain an identifiable model.

EXAMPLE 3: Continuing Example 2, suppose that $\pi_i = \pi_j$ for all i, j so that we have likelihood:

$$L_y(\theta, \pi) = \left(\prod_{i=1}^{m} \theta_i^{I(y=i)} \right) \pi^{I(y \neq \bullet)} (1-\pi)^{I(y=\bullet)}$$

To see that θ is identifiable we note that for any $\theta' \neq \theta''$ we have $L_x(\theta', \pi) = L_x(\theta'', \pi)$ for $y = \bullet$. ■

Example 3 shows us that, if we assume that the long run proportion of respondents of each type is the same then θ becomes identifiable in the model. This means that *without* this additional assumption, even under perfect information our inferences are sensitive to our prior assumptions. However, *with* this additional assumption, under perfect information our inferences are not sensitive to our prior assumptions. It should therefore be obvious that the model is sensitive to this assumption in the sense that the absence of the assumption leaves the model unidentifiable and therefore highly sensitive to prior beliefs. From the perspective of robustness testing the question then becomes: are we confident enough in this additional assumption to warrant its inclusion in the likelihood model?

7. An objective robustness test

The above analysis suggests that a test of identifiability itself provides a useful test of robustness to prior beliefs. Aside from the regularity conditions involved (which hold widely and can be easily tested on a case by case basis) tests of identifiability are equivalent to tests of the asymptotic behaviour of the parameter of interest under perfect information; this is of direct interest in its own right. Moreover, tests of identifiability are objective, in that they do not require any assumptions beyond those invariance assumptions that determine the likelihood function. This is also important since it avoids raising further questions of sensitivity to assumptions (the avoidance of which is the point of robustness testing in the first place).

What then are we to make of the results of such a test? If the parameter of interest is identifiable then we know that the data is providing information directly about the parameter of interest, and to such an extent that perfect information would lead to certain belief in the true parameter of interest. This is all good news and we may legitimately conclude that with enough data, our model will be robust to prior assumptions. However, if the parameter of interest is unidentifiable then we know that the data is not providing us with information about the parameter of interest except through the relationship between the parameter of interest and the minimal sufficient parameter (which is entirely determined by our prior beliefs).

The only question that may remain is the sensitivity to prior beliefs for a certain finite amount of data (most obviously the amount actually observed). Unfortunately we have seen that in answering this question any posterior belief is possible and supplying a 'more practical' answer necessarily involves the (possibly arbitrary) limitation of the class of prior beliefs under consideration, which may lead to further questions of sensitivity.

7.1. Making assumptions in order to obtain informativity

We have seen that identifiability is a useful and important concept in determining robustness and thus, in determining the degree to which we can trust inferences from our models. We may even go so far as to say that we cannot trust inferences from unidentifiable models at all. We should therefore make every attempt to ensure that our sampling mechanism and accompanying assumptions lead to a model that is identifiable for the parameters of interest.

However, practitioners should be wary of stipulating assumptions (particularly about unobservable limiting quantities) merely in order to obtain identifiability; after all, a model that is robust to prior assumption but is predicated on flawed likelihood assumptions is no better than a model that is sensitive to prior assumptions. In cases of self selection where high non response rates can be expected it may be more prudent to reject data altogether and admit that no reliable inferences are possible rather than to churn data through a flawed model.

Appendix: Proof of Theorems

PROOF OF THEOREM 1: Let $\mathcal{C}(\mathcal{A}) \equiv \{(\theta,\pi) : \theta \in \mathcal{A}\}$. Since the posterior $\mathcal{P}_k$ is concentrated on the support for θ of $L_{\mathbf{x}_k}$ we may restrict our attention to $\mathcal{A} \subseteq \text{supp}(\theta)$; the equality holds trivially for $\mathcal{A}$ outside this support. Let $\mathcal{G}_k$ be our posterior belief for (θ,π) so that $\mathcal{P}_k(\mathcal{A}) = \int_{\mathcal{C}(\mathcal{A})} d\mathcal{G}_k(\theta,\pi)$. We define $\mathcal{P}(\mathcal{C}(\mathcal{A})) \equiv \int_{\mathcal{C}(\mathcal{A})} L_{\mathbf{x}_k}^{-1}(\theta,\pi)\, d\mathcal{G}_k(\theta,\pi)$ so that $\mathcal{G}_k \gg \mathcal{P}$ and so that $L_{\mathbf{x}_k}^{-1}$ is the Radon-Nikodym derivative $d\mathcal{P}/d\mathcal{G}_k$. Since $L_{\mathbf{x}_k}^{-1}$ is strictly positive over $\mathcal{C}(\text{supp}(\theta))$ we also have $\mathcal{P} \gg \mathcal{G}_k$ so that $d\mathcal{G}_k/d\mathcal{P} = L_{\mathbf{x}_k}$. It follows that:

$$\mathcal{P}_k(\mathcal{A}) = \int_{\mathcal{C}(\mathcal{A})} d\mathcal{G}_k(\theta,\pi) = \int_{\mathcal{C}(\mathcal{A})} \frac{d\mathcal{G}_k}{d\mathcal{P}}\, d\mathcal{P}(\theta,\pi) = \int_{\mathcal{C}(\mathcal{A})} L_{\mathbf{x}_k}\, d\mathcal{P}(\theta,\pi)$$

which was to be shown. ■

PROOF OF THEOREM 2: Theorem 3.3 of Berk (1970) shows that, under the regularity conditions specified there, if $\lambda^*(\mathcal{A}) < \lambda^*(\Phi)$ then $P(\mathcal{P}_\infty(\mathcal{A}) = 0 \,|\, \phi) = 1$. It follows from the Radon-Nikodym Theorem that if $\lambda^*(\mathcal{A}) < \lambda^*(\Phi)$ then:

$$P(\mathcal{P}_\infty(\mathcal{A}) = 0) = \int P(\mathcal{P}_\infty(\mathcal{A}) = 0 \,|\, \phi)\, dP(\phi) = \int dP(\phi) = 1$$

which was to be shown. ■

The regularity conditions for Theorem 2 are required to ensure the existence of the posterior distribution and to ensure that where a sequence of functions approach a limiting function, the arguments that maximize the latter are the limit of the arguments that maximize the former. Berk (1970) gives wide conditions under which this occurs; Wald (1949) gives simpler but narrower conditions (see also Berk (1966) and Huber (1967)). The actual result given in Berk (1970) holds under wider conditions than are given here; we will not have need of the wider theorem since the convergence of the minimal sufficient parameter does not always hold under the wider conditions.

PROOF OF THEOREM 3: For all $i \in \mathbb{N}$ let $\mathcal{A}_i \equiv \left\{ r : \lambda(r) < \lambda^*(\Phi) - 1/i \right\}$ so that $\lambda^*(\mathcal{A}_i) < \lambda^*(\Phi)$. From Theorem 1 we then have:

$$P\left(\mathcal{P}_\infty(\mathcal{A}_i) > 0\right) = 0 \text{ for all } i \in \mathbb{N}.$$

Since Φ^* is not empty and since $\bar{\Phi}^* = \bigcup_{i=1}^{\infty} \mathcal{A}_i$ it follows that:

$$P\left(\mathcal{P}_\infty\left(\Phi^*\right) = 1\right) = 1 - P\left(\left(\forall i \in \mathbb{N}\right) : \mathcal{P}_\infty(\mathcal{A}_i) > 0\right) \geq 1 - \sum_{i=1}^{\infty} P\left(\mathcal{P}_\infty(\mathcal{A}_i) > 0\right) = 1$$

which was to be shown. ■

Theorem 3 shows the convergence of our posterior beliefs about sufficient parameters to the set of values that maximise the Wald function. Following a proof by Wald (1949) we now show that for minimal sufficient parameters the true value of the parameter uniquely maximises the Wald function.

LEMMA 1: If the likelihood L_x has minimal sufficient parameter ϕ then $\Phi^* = \{\phi\}$.

PROOF: For all $r \in \Phi$ we have:

$$\lambda(r) - \lambda(\phi) = \mathrm{E}\left(l_x(r) - l_x(\phi) \middle| \phi\right) = \mathrm{E}\left(\ln\left(\frac{L_x(r)}{L_x(\phi)} \right) \middle| \phi \right)$$

Since ϕ is minimal sufficient, it follows from Lehmann and Scheffé (1951) that $L_x(r) \neq L_x(\phi)$ for all $r \neq \phi$. It then follows from Jensen's Inequality that:

$$\lambda(r) - \lambda(\phi) < \ln\left(\mathrm{E}\left(\frac{L_x(r)}{L_x(\phi)} \middle| \phi \right) \right) = \ln(1) = 0 \text{ for all } r \neq \phi$$

so that $\lambda(\phi) > \lambda(r)$ for all $r \neq \phi$. Thus $\Phi^* = \{\phi\}$ which was to be shown. ■

PROOF OF THEOREM 4: From Lemma 1 we have $\Phi^* = \{\phi\}$. From Theorem 2 we then have $P\left(\mathcal{P}_\infty\left(\{\phi\}\right) = 1\right) = 1$ which was to be shown. ■

References

BAHADUR, R.R. (1954) Sufficiency and statistical decision functions. *Annals of Mathematical Statistics* **22**, pp. 423-462.

BARANKIN, E.W. (1961) Sufficient parameters: solution of the minimal dimensionality problem. *Annals of Mathematical Statistics* **12**, pp. 91-118.

BERK, R.H. (1966) Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics* **37(1)**, pp. 51-58.

BERK, R.H. (1970) Consistency a posteriori. *Annals of Mathematical Statistics* **41(3)**, pp. 894-906.

BERNARDO, J.M. AND SMITH, A.F.M. (1994) *Bayesian Theory*. John Wiley, Chichester.

BOWDEN, R. (1973) The theory of parametric identification. *Econometrica* **41(6)**, pp. 1069-1074.

DAWID. A.P. (1979) Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B (Methodological)* **41(1)**, pp. 1-31.

DE FINETTI, B. (1980) Foresight. Its Logical Laws, Its Subjective Sources (translated). *Studies in Subjective Probability* (Kyberg, H.E. and Smokler, H.E. eds.) Krieger Publishing Company.

DUDLEY, R.M. (2002) *Real analysis and probability*. Cambridge University Press: Cambridge.

FORTINI, S., LADELLI, L. AND REGAZZINI, E. (2000) Exchangeability, predictive distributions and parametric models. *Sankhyā: the Indian Journal of Statistics* **62A**, pp. 86-109.

HALMOS, P.R. AND SAVAGE, L.J. (1949) Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Annals of Mathematical Statistics* **20**, pp. 225-241.

HUBER, P. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability* **1**, pp. 221-233.

JOHNSON, B.R. (1974). An effective method for teaching the Lehmann-Scheffé technique for obtaining a minimal sufficient statistic. *American Statistician* **28(2)**, pp. 59-60.

KULLBACK, S. AND LEIBLER, A. (1951) On information and sufficiency. *Annals of Mathematical Statistics* **22**, pp 79-86.

LEHMANN, E.L. AND SCHEFFÉ, H. (1951) Completeness, similar regions and unbiased estimation. *Sankhyā: the Indian Journal of Statistics* **10**, pp. 305-340.

PEIKOFF, L. (1993) *Objectivism: The Philosophy of Ayn Rand*. Meridian: New York.

ROEHRIG, C.S. (1988) Conditions for Identification in Nonparametric and Parametric Models. *Econometrica* **56(2)**, pp. 433-447.

ROTHENBERG. (1971) Identification in parametric models. *Econometrica* **39(3)**, pp. 577-591.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **20(4)**, pp. 595-601.

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall: London.