Statistical Science
2005, Vol. 20, No. 2, 111–140
DOI 10.1214/088342305000000098

On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables

Paul Gustafson

Abstract. When a candidate model for data is nonidentifiable, conventional wisdom dictates that the model must be simplified somehow so as to gain identifiability. We explore two scenarios involving mismeasured variables where, in fact, model expansion, as opposed to model contraction, might be used to obtain identifiability. We compare the merits of model contraction and model expansion. We also investigate whether it is necessarily a good idea to alter the model for the sake of identifiability. In particular, estimators obtained from identifiable models are compared to those obtained from nonidentifiable models in tandem with crude prior distributions. Both asymptotic theory and simulations with Markov chain Monte Carlo-based estimators are used to draw comparisons. A technical point which arises is that the asymptotic behavior of a posterior mean from a nonidentifiable model can be investigated using standard asymptotic theory, once the posterior mean is described in terms of the identifiable part of the model only.

Key words and phrases: Bayes analysis, identifiability, measurement error, misclassification, nested models, prior information.

1. INTRODUCTION

Say that a particular statistical model with p unknown parameters seems appropriate for a modeling problem at hand, but this model is not identifiable. That is, multiple values of the parameter vector correspond to the same distribution of observable data. Conventional wisdom dictates that a simpler submodel with fewer than p parameters must be selected so as to gain identifiability. Of course this process of *model contraction* may lead to a model that involves dubious assumptions or a model that is less realistic in some other way. A less intuitive approach to gaining identifiability is *model expansion* whereby the initial model is enlarged. As we will see, there are natural situations where the initial nonidentifiable model has an identifiable supermodel with more than p unknown parameters.

To be more specific, two inferential scenarios involving mismeasured variables are considered. The first scenario involves two imperfect schemes for assessing whether a study subject is "exposed," where both the misclassification probabilities describing these schemes and the prevalence of exposure in the study population are unknown. Three plausible models are considered in this context, with nonidentifiable Model A nested within identifiable Model B nested within nonidentifiable Model C. Thus the identifiable Model B might be arrived at by contraction of the nonidentifiable Model C or by expansion of the nonidentifiable Model A.

The second scenario involves regression of a continuous response variable on a continuous explanatory variable, where the explanatory variable is subject to measurement error. Again a nested sequence of plausible models is considered, with identifiable Model D

Paul Gustafson is Associate Professor, Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z2 (e-mail: gustaf@stat.ubc.ca).

nested within nonidentifiable Model E nested within identifiable Model F. Thus if Model E is considered initially, then either model contraction or model expansion can be used to gain identifiability.

In addition to considering different models, the role of prior information is also investigated in both scenarios. Particularly, we investigate how helpful a crude subjective prior distribution can be when faced with a nonidentifiable model. That is, the infusion of prior information into a nonidentifiable model is considered as an alternative to gaining identifiability through either model contraction or model expansion.

In both scenarios estimators based on the different models are compared, under a variety of actual data generating mechanisms. The motivation for these comparisons is curiosity about two questions. First, is the uncommon strategy of gaining identifiability via model expansion appealing, particularly in comparison to the common strategy of gaining identifiability via model contraction. Second, is the conventional wisdom that something must be done to gain identifiability always sound. The comparisons of estimator performance are based on both asymptotic theory and simulation studies. In the former case, both "right-model" and "wrong-model" asymptotic theory is employed. In the latter case, Bayes estimators computed via Markov chain Monte Carlo (MCMC) methods are compared on simulated data.

On the face of it one might think that standard asymptotic theory cannot shed light on the performance of estimators based on nonidentifiable models, so that simulation is the only possibility for studying the estimators based on Models A, C and E. On the contrary, we show that standard asymptotic theory can describe the behavior of posterior means that arise from nonidentifiable models. To do so one must simply use iterated expectation to reexpress the original posterior mean as a posterior mean with respect to the identifiable part of the model alone.

Before proceeding to the first scenario, some care with terminology and definitions surrounding identifiability is required. Say that Model M postulates a distribution $F(\cdot|\theta)$ for the observable data, with the unknown parameter θ being an element of the parameter space Θ. Let A be the subset of Θ defined as

$$A = \{\theta \in \Theta : F(\cdot|\theta) = F(\cdot|\theta^*) \text{ for some } \theta^* \in \Theta \setminus \{\theta\}\}. \tag{1}$$

That is, A consists of all parameter values which do not give rise to a distinctive distribution of observable data. We will say that Model M is

- *fully identified* if A is empty;
- *essentially identified* (with respect to prior distribution Π on Θ) if A is nonempty but $\Pi(A) = 0$;
- *nonidentified* (with respect to prior Π) if $\Pi(A) > 0$.

With regard to these definitions, note that expanding a nonidentified Model M to a supermodel M′ can at best yield essential rather than full identifiability. The parameter space for M′ will have a lower dimensional subspace corresponding to Model M, and some elements of this subspace will necessarily lie in A. Thus the first question about model expansion strategies is whether essential identifiability is strong enough to equate with "practical" identifiability. Certainly if we give credence to the prior distribution, then being able to definitely learn the parameter values from a large enough sample with prior probability 1 seems sufficient. Particularly if we take the view that in practice all models are somewhat misspecified, so that the true parameter values are merely the values minimizing Kullback–Leibler divergence between the true and modeled data distributions, then it seems reasonable to regard the parameter values falling on the lower dimensional subspace as being a probability 0 event.

2. SCENARIO I

In many epidemiological studies the classification of subjects as "unexposed" or "exposed" cannot be done perfectly. To mitigate this problem, it is common to employ several different imperfect classification schemes to test for exposure. For instance, Hui and Walter (1980) gave an example involving two tests (the Mantoux test and the Tine test) for the detection of tuberculosis; Drews, Flanders and Kosinski (1993) considered both patient interviews and medical records to measure various putative binary risk factors in a case-control study of sudden-infant-death syndrome; and Joseph, Gyorkos and Coupal (1995) considered a study in which both a serology test and stool examination were used to test for a particular parasitic infection. In addition to the assessment schemes being imperfect, often the classification probabilities that characterize the degree of imperfection are not known precisely, although there may be some reasonable prior knowledge in this regard.

Let E denote the exposure variable ($E = 0$ for unexposed; $E = 1$ for exposed), and let T_1 and T_2 be two imperfect surrogates (tests) for E. We consider the re-

alistic scenario in which the *sensitivity* $p_j = \Pr(T_j = 1|E = 1)$ and *specificity* $q_j = \Pr(T_j = 0|E = 0)$ of each test are unknown. Say that (T_1, T_2) are observed for subjects sampled from the population of interest, which has unknown exposure prevalence $r = \Pr(E = 1)$. Model A postulates that

$$\begin{aligned}(2)\qquad &\Pr(T_1 = a, T_2 = b|\theta)\\ &= rp_1^a(1-p_1)^{1-a}p_2^b(1-p_2)^{1-b}\\ &\quad + (1-r)q_1^{1-a}(1-q_1)^a q_2^{1-b}(1-q_2)^b,\end{aligned}$$

where $\theta = (p_1, p_2, q_1, q_2, r)$ is the unknown parameter vector. Note that Model A invokes the common assumption that the two test outcomes (T_1, T_2) are conditionally independent given the true exposure status E. Clearly Model A is nonidentifiable, because the (T_1, T_2) data comprise a 2×2 table from which at most three parameters can be estimated consistently, whereas in fact five parameters are unknown. Despite the nonidentifiability, Bayesian inference under Model A is relatively straightforward to implement, as exemplified by Joseph, Gyorkos and Coupal (1995).

Starting with Model A, one way to develop an identifiable model is to pre- or poststratify the sample/population according to some binary covariate X, which is thought to be associated with the exposure E. For instance, say random samples of size n_1 and n_2 are taken from the $X = 0$ and $X = 1$ subpopulations, respectively. Model B postulates that (2) holds with prevalence $r = r_1$ in the first subpopulation and with prevalence $r = r_2$ in the second subpopulation. As well, Model B implicitly assumes the exposure misclassification is *nondifferential*, in that (T_1, T_2) and X are conditionally independent given E. Less formally, the mechanisms which yield misclassification are assumed to operate identically in the two subpopulations. Model B, with six unknown parameters, $\theta = (p_1, p_2, q_1, q_2, r_1, r_2)$, is clearly an expansion of Model A, but now the data can be summarized into separate 2×2 tables for each subpopulation, so there is hope of consistently estimating six parameters. Indeed, Hui and Walter (1980) illustrated that, subject to some minor caveats, Model B is a regular model that leads to likelihood-based estimators with standard asymptotic properties. More precisely, in our terminology Model B is essentially identifiable with respect to any continuous prior distribution. Johnson, Gastwirth and Pearson (2001) explicitly argued in favor of stratification and the use of Model B for the sake of identifiability. We will somewhat loosely phrase our comparisons as being between models, but of course the choice to prestratify or not is more accurately described as being a design issue.

Of course, moving from Model A to Model B might not be viewed as purely a model expansion, since inferences from Model B are based on more data than inferences from Model A. Thus it is perhaps not so surprising that this kind of model and data expansion can yield parameter identifiability. On the other hand, we certainly have a nested model situation where the submodel is nonidentifiable whereas the supermodel is essentially identifiable. In particular, we can view both models as describing the distribution of (T_1, T_2, X), but Model A happens to posit that X contributes no information about the unknown parameters.

Another comment is that starting with Model A, other model and data expansions can lead to identifiability. For instance, the addition of a third conditionally independent surrogate T_3 for E leads to identifiability without stratification. Indeed there is literature that addresses scenarios with a large number of surrogates; see, for instance, Qu, Tan and Kutner (1996) for examples with up to seven surrogates!

The assumption that T_1 and T_2 are conditionally independent given E may not be reasonable in a given application. Indeed, it is easy to imagine that T_1 and T_2 will be positively correlated given E in many practical settings, because subjects who are particularly susceptible to misclassification by one scheme may also be particularly susceptible under the other scheme. Moreover, without observations on E, the assumption is not amenable to formal empirical checking, because all the degrees of freedom are used in the estimation of θ. The plausibility of the conditional independence assumption and the effects of incorrectly invoking it were discussed by Fryback (1978), Vacek (1985), Brenner (1996) and Torrance-Rynard and Walter (1997).

As illustrated by Dendukuri and Joseph (2001) and Georgiadis, Johnson, Gardner and Singh (2003), Bayesian modeling can be used to relax the assumption that two tests are conditionally independent given the true exposure. In the present context we construct Model C, an expansion of Model B, by modeling the distribution of $T_1, T_2|E$ as

$$\begin{aligned}&\Pr(T_1 = a, T_2 = b|E)\\ &\quad = \begin{cases} (1-q_1)^a q_1^{1-a}(1-q_2)^b q_2^{1-b} + (-1)^{|a-b|}\delta_0, \\ \quad \text{if } E = 0, \\ p_1^a(1-p_1)^{1-a}p_2^b(1-p_2)^{1-b} + (-1)^{|a-b|}\delta_1, \\ \quad \text{if } E = 1. \end{cases}\end{aligned}$$

Under this model p_j and q_j retain their interpretations as the sensitivity and specificity of the jth test, but now (δ_0, δ_1) are additional unknown parameters, with $\delta_j = \text{Cov}(T_1, T_2|E = j)$. Model B is recovered if $\delta_0 = \delta_1 = 0$. Whereas it is hard to imagine scenarios under which T_1 and T_2 are negatively associated given E, we restrict to $\delta_0 \in [0, \delta_{\text{MAX}}(q_1, q_2)]$ and $\delta_1 \in [0, \delta_{\text{MAX}}(p_1, p_2)]$, where $\delta_{\text{MAX}}(s, t) = \min\{s, t\} - st$ is the maximal covariance between two binary random variables with "success" probabilities s and t. For future reference, note that the dependence in Model C can also be expressed in terms of correlation, which is more interpretable but complicates the requisite mathematical expressions. Specifically, let $\rho_j = \text{Corr}(T_1, T_2|E = j)$ for $j = 1, 2$, with the range of dependence now expressed as $\rho_0 \in [0, \rho_{\text{MAX}}(q_1, q_2)]$ and $\rho_1 \in [0, \rho_{\text{MAX}}(p_1, p_2)]$.

Model C, with eight unknown parameters $\theta_C = (p_1, p_2, q_1, q_2, \delta_0, \delta_1, r_1, r_2)$, is clearly not identifiable from the data which are still summarized by two 2×2 tables. Thus while Model C may be appealing on the grounds of realism, it is tempting to contract to Model B for the sake of identifiability.

2.1 Performance of Model B Estimators

The behavior of estimates generated by fitting Model B to data can be studied via regular asymptotic theory. It is convenient to restrict the parameter space $\theta \in \Theta$ according to $p_j + q_j > 1$ for $j = 1, 2$ to avoid the trivial nonidentifiability arising because $f(t_1, t_2|\theta)$ is unchanged upon replacing p_j with $1 - q_j$, q_j with $1 - p_j$ and r_j with $1 - r_j$. In practice the restriction is very mild, because an assessment scheme that is worse than chance (i.e., $p_j + q_j < 1$) can usually be ruled out a priori. While it is cumbersome to write down explicit expressions, there is no difficulty in evaluating the Fisher information matrix $I(\theta)$ exactly (see, e.g., Hui and Walter, 1980). In situations where Model B is correctly specified, a maximum likelihood or Bayes estimator of $\psi = \psi(\theta)$ is consistent and has a readily computed asymptotic variance.

To give a specific example, say that data are generated under Model B with $p_1 = 0.8$, $p_2 = 0.8$, $q_1 = 0.75$, $q_2 = 0.9$, $r_1 = 0.3 - \Delta/2$ and $r_2 = 0.3 + \Delta/2$. For later reference this scenario is referred to as DGM (i), where DGM stands for *data generating mechanism*. For simplicity we assume that $\Pr(X = 1)$ is known, thereby giving slightly favorable assessments of estimator variance relative to the more realistic setting where $\Pr(X = 1)$ must also be estimated. In fact we consider the balanced case of $\Pr(X = 1) = 0.5$, so that DGM (i) implies $r = (r_1 + r_2)/2 = 0.3$, which can be estimated by $\hat{r} = (\hat{r}_1 + \hat{r}_2)/2$. As well, in the case of prestratification we assume balanced stratum-specific sample sizes $n_1 = n_2 = n/2$. We note in passing that the impact of $\Pr(X = 1)$ being closer to 0 or 1, and/or (n_1, n_2) being unbalanced, is less than obvious, because the complex nature of the model implies that the estimator of r_j is not based only on the data from the jth stratum.

The left sides of the panels in Figure 1 give the approximate (asymptotic) root-mean-squared error (ARMSE) for maximum likelihood or Bayes estimators of the classification probabilities (p_1, p_2, q_1, q_2) and the prevalence r, assuming a large overall sample size of $n = 2000$. Specifically, the ARMSE is displayed as a function of $\Delta = r_2 - r_1$, the difference between the prevalences in the two subpopulations.

As suggested by Figure 1, each ARMSE diverges to infinity as Δ approaches zero. This is not surprising if we think about the $r_1 = r_2$ scenario as corresponding to a completely artificial stratification based on randomization. We would be getting something for nothing—a "free lunch"—if allocating subjects to strata on the basis of coin flips would yield identifiability and hence estimator consistency. Put another way, if X and E are independent, then the Model B parameters lie on the lower dimensional subspace corresponding to Model A. Mathematically it is clear that when $r_1 = r_2$, the corresponding rows (columns) of the Model B Fisher information matrix $I(\theta)$ are identical and, hence, the information matrix is singular in the $r_1 = r_2$ limit. What is surprising, however, is that Δ need not be very close to zero before each ARMSE is quite large. When $\Delta = 0.1$, for instance, ARMSE$[\hat{r}]$ is about 0.12, perhaps large enough to render a study of the population prevalence futile despite the large sample size. Moreover, this is about double the ARMSE attained when $\Delta = 0.2$. This sounds a cautionary note about study design in the Model B framework. Unless there is good prior knowledge to indicate that the subpopulations will have markedly different prevalences, there is a risk of obtaining very poor estimates even with considerable sample sizes.

It should also be mentioned that our experience with simulated data sets suggests that the sample size at which the asymptotic estimator performance becomes a good approximation to the actual performance itself increases as the difference in prevalences Δ gets smaller. This cautions against taking the small Δ regions of Figure 1 too literally, although it does not di-

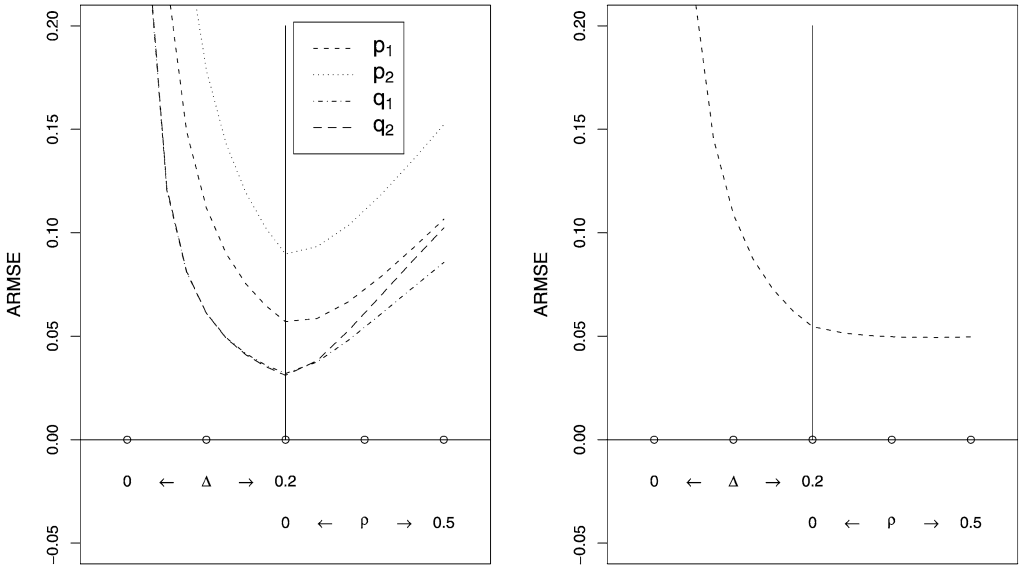

FIG. 1. *ARMSE for* $\hat{p}_1$, $\hat{p}_2$, $\hat{q}_1$, $\hat{q}_2$ (left panel) *and* $\hat{r}$ (right panel) *under Model* B *and DGM* (i), *with a sample size of* $n = 2000$. *The left side of each panel corresponds to the true parameter values in DGM* (i), *with varying values of* $\Delta = r_2 - r_1$. *The right side of each panel corresponds to data generated under Model* C, *with* $\Delta = 0.2$ *and varying values of* $\rho = \text{Corr}(T_1, T_2|E)$.

minish the point that estimator performance may be very poor if Δ is small, even if n is large. Another manifestation of this point will be emphasized in Section 2.4.1, namely that when Δ is small, the prior distribution may have considerable influence on posterior quantities even at large sample sizes.

The right sides of the Figure 1 panels give the ARMSE of estimators based on Model B when the data are actually generated under Model C. Thus they describe the impact of incorrectly assuming conditional independence of T_1 and T_2 given E. Standard wrong-model asymptotic theory (e.g., White, 1982) is used to compute the ARMSE in this scenario. In particular, Model B can be parameterized in terms of ν instead of θ, where ν comprises three probabilities which characterize the distribution of $T_1, T_2|X = 0$, along with three probabilities which characterize the distribution of $T_1, T_2|X = 1$. We write $\nu = h(\theta)$, where the function $h(\cdot)$ is easily evaluated. It is also possible to evaluate $h^{-1}(\cdot)$, although the expressions are extremely cumbersome (Hui and Walter, 1980). For true parameter values $(\theta, \delta_0, \delta_1)$ under Model C, we compute ν^*, the probabilities that characterize $T_1, T_2|X$ under Model C. Then $\theta^* = h^{-1}(\nu^*)$ will be the large-sample limit of $\hat{\theta}$ obtained when fitting the incorrect Model B to the data. Thus in estimating $\psi = g(\theta)$, the asymptotic bias incurred because of model misspecification is $g(\theta^*) - g(\theta)$. Moreover, following White (1982), the asymptotic variance of $\hat{\theta}$ is given as $A(\theta^*)^{-1}B(\theta^*)A(\theta^*)^{-1}$, where

$$A_{ij}(\theta) = E_C\{\partial^2 \log f_B(T_1, T_2; \theta)/\partial\theta_i\, \partial\theta_j\},$$

$$B_{ij}(\theta) = E_C\{\partial \log f_B(T_1, T_2; \theta)/\partial\theta_i \cdot \partial \log f_B(T_1, T_2; \theta)/\partial\theta_j\},$$

with the notation chosen to emphasize that the f inside the expectations is from the incorrect Model B, while the expectations themselves are with respect to the actual distribution of (T_1, T_2) given by Model C. Armed with the asymptotic bias and asymptotic variance of $\hat{\theta}$, the ARMSE for $\hat{\psi} = g(\hat{\theta})$ at a particular sample size is readily computed.

As an aside, this route to determining the asymptotic behavior of Model B estimators when Model C is correct is not fully general. For some Model C parameter values, especially with larger values of δ_0 and δ_1, the $T_1, T_2|X$ probabilities ν^* can fall outside the Model B parameter space. That is, ν^* can lie outside the image under $h(\cdot)$ of the Model B parameter space for θ. In such a situation θ^*, the large-sample limit of the Model B-based estimator, cannot be determined as $h^{-1}(\nu^*)$. We have not pursued this here, but such instances require numerical methods to find θ^* as the value of θ which minimizes the Kullback–Leibler divergence between the actual distribution of (T_1, T_2) and the distribution postulated under

Model B. In fact, a similar problem can arise in fitting Model B even when it is correct. That is, sampling variability alone can lead to observed cell proportions for the (T_1, T_2) table that lie outside the image of $h(\cdot)$, so that $h^{-1}(\cdot)$ cannot be applied to these proportions as a route to generating parameter estimates. For this reason Drews, Flanders and Kosinski (1993) proposed expectation–maximization (EM) algorithm fitting of Model B rather than the closed-form approach of Hui and Walter (1980).

Returning to Figure 1, the right sides of the panels are again based on the DGM (i) values for (p_1, p_2, q_1, q_2, r), with Δ fixed at $\Delta = 0.2$. A common value for both ρ_0 and ρ_1 is varied from $\rho = 0$ to $\rho = 0.5$. We note in passing that $\rho_0 = 0.5$ is quite close to the upper bound of $\rho_0 \leq \rho_{\mathrm{MAX}}(q_1, q_2) = 0.577$ for the specified values of q_1 and q_2. Conversely, values up to 1 are possible for ρ_1, since $p_1 = p_2$. The right sides of the panels in Figure 1 plot ARMSE as a function of ρ.

The format of Figure 1 is chosen to contrast the two potential pitfalls of using Model B for inference. The center of each panel corresponds to a good situation, in that the subpopulation prevalences are quite disparate ($\Delta = 0.2$) and the conditional independence assumption is satisfied ($\rho = 0$). Moving to the left, the DGM approaches the nonidentifiable Model A as Δ decreases to zero. Moving to the right, Model B becomes increasingly misspecified as the conditional correlation between the two tests increases. Surprisingly, the increase in ARMSE to the left tends to be more dramatic than the increase to the right. That is, Model B being correct but with parameter values in the vicinity of the nonidentifiable submodel is more damaging than Model B being incorrect due to conditional dependence between the two tests.

Of course Figure 1 pertains to specific underlying values of (p_1, q_1, p_2, q_2, r). To suggest that the qualitative behavior is similar for other values, Figure 2 gives ARMSE values for DGM (ii), defined by $p_1 = 0.95$, $p_2 = 0.9$, $q_1 = 0.65$, $q_2 = 0.85$ and $r = 0.15$. The overall impression is again that Model B being correct with parameter values near Model A is worse than Model B being incorrect because of dependence between tests. Experimentation with further sets of underlying parameter values (results not shown) also supports this view.

The concern about the performance of Model B under moderately small values of $\Delta = r_2 - r_1$ suggests that stratifying the population to gain identifiability is not a panacea. Thus we consider using Model A for inference, its nonidentifiability notwithstanding. Of course, it seems unreasonable to expect reasonable inferences from a nonidentifiable model with a diffuse prior distribution. We speculate, however, that a crude subjective prior might go some distance toward producing good inferences. Before looking specifically at Model A, we develop an asymptotic approach to studying the performance of posterior means that arise from nonidentifiable models in general.

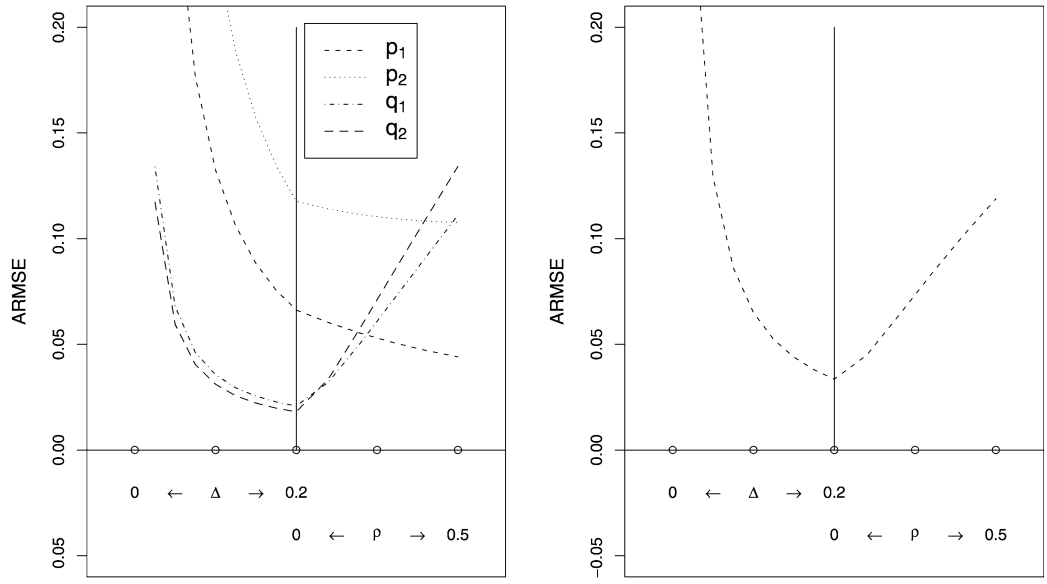

FIG. 2. *ARMSE for* $\hat{p}_1$, $\hat{p}_2$, $\hat{q}_1$, $\hat{q}_2$ (left panel) *and* $\hat{r}$ (right panel) *under Model* B *and DGM* (ii). *The format is the same as Figure* 1.

2.2 Asymptotic Behavior of Posterior Means in Nonidentifiable Models

The asymptotic behavior of Bayes estimates arising from nonidentifiable models has received very little attention in the literature. Whereas Neath and Samaniego (1997) and Gustafson, Le and Saskin (2001) studied the issue in the context of specific models, here we describe a more general approach.

To gain insight into a nonidentified model we seek to reparameterize from the original parameter vector θ to $\phi = (\phi_I, \phi_N)$ in such a way that $f(\text{data}|\phi) = f(\text{data}|\phi_I)$. That is, the distribution of the data depends only on the identifiable part of the parameter vector ϕ_I, and not on the nonidentifiable part ϕ_N. We call such a parameterization *transparent*, because it is intended to make apparent the impact of nonidentifiability. We also assume that a proper prior distribution with density $f(\theta)$ has been specified in the original parameterization. Of course, this induces a prior density $f(\phi)$ in the transparent parameterization. Indeed, following Dawid (1979) and Gelfand and Sahu (1999), it is useful to think of the prior for ϕ in terms of the marginal density $f(\phi_I)$ and the conditional density $f(\phi_N|\phi_I)$. Then immediately we have

$$f(\phi_I|\text{data}) \propto f(\text{data}|\phi_I) f(\phi_I), \tag{3}$$

$$f(\phi_N|\phi_I, \text{data}) = f(\phi_N|\phi_I). \tag{4}$$

Thus (3), the posterior marginal distribution for ϕ_I, is typically governed by the usual asymptotic theory which applies in the identifiable case. On the other hand, (4), the posterior conditional distribution for $\phi_N|\phi_I$, is identical to the prior conditional distribution. That is, there is no Bayesian learning whatsoever about the conditional distribution of $\phi_N|\phi_I$. We emphasize, however, that a natural or obvious prior for θ will often lead to prior dependence between ϕ_I and ϕ_N, and consequently

$$f(\phi_N|\text{data}) = \int f(\phi_N|\phi_I) f(\phi_I|\text{data})\, d\phi_I \neq f(\phi_N).$$

That is, marginally there can be some learning about ϕ_N. We refer to this as *indirect learning*, because it is learning about ϕ_N that results only because of learning about ϕ_I. Also, note that under typical regularity conditions, (3) will concentrate to a point mass at the true value of ϕ_I as the sample size grows. Thus the posterior marginal distribution of ϕ_N will tend to the (nondegenerate) prior conditional distribution (4) evaluated at the true value of ϕ_I. For general discussion of regularity conditions governing the asymptotic normality of posterior distributions and Bayes estimators, see, for instance, Bernardo and Smith (1994, Section 5.3) and Lehmann and Casella (1998, Section 6.8). Outwardly it appears that the indirect learning is quite subjective in nature, because it is driven by the conditional prior distribution of ϕ_N given ϕ_I. In both scenarios studied here, however, the nature of the transparent parameterization is such that the support of this distribution depends on ϕ_I, so that the indirect learning is quite intrinsic to the problem at hand.

Now say that with respect to the transparent parameterization the parameter of interest can be expressed as $\psi = g(\phi) = g(\phi_I, \phi_N)$. Then

$$\begin{aligned} &E(\psi|\text{data}) \\ &\quad = \iint g(\phi_I, \phi_N) f(\phi_I, \phi_N|\text{data})\, d\phi_N\, d\phi_I \\ &\quad = \iint g(\phi_I, \phi_N) f(\phi_N|\phi_I)\, d\phi_N f(\phi_I|\text{data})\, d\phi_I \\ &\quad = E\big(\tilde{g}(\phi_I)|\text{data}\big), \end{aligned}$$

where

$$\tilde{g}(\phi_I) = \int g(\phi_I, \phi_N) f(\phi_N|\phi_I)\, d\phi_N.$$

In particular, the posterior mean of interest is now expressed as a posterior mean in the identifiable model parameterized by ϕ_I alone. Thus its large-sample behavior will typically be described by the standard asymptotic theory based on Fisher information. That is, if the model is correct and an i.i.d. sample of size n yields $\hat{\psi}^{(n)} = E(\psi|\text{data})$, then

$$\begin{aligned} &n^{1/2}\{\hat{\psi}^{(n)} - \tilde{g}(\phi_I)\} \\ &\quad \Rightarrow \text{N}\big[0, \{\tilde{g}'(\phi_I)\}^T I(\phi_I)^{-1}\{\tilde{g}'(\phi_I)\}\big] \end{aligned}$$

in distribution, as $n \to \infty$. Thus the RMSE incurred when estimating ψ by $\hat{\psi}^{(n)}$ can be approximated as

$$\begin{aligned} &\text{ARMSE} \\ &\quad = \big[\{\tilde{g}(\phi_I) - g(\phi_I, \phi_N)\}^2 \\ &\qquad + n^{-1}\{\tilde{g}'(\phi_I)\}^T I(\phi_I)^{-1}\{\tilde{g}'(\phi_I)\}\big]^{1/2}, \end{aligned} \tag{5}$$

where the first term describes the asymptotic bias and the second term describes the asymptotic variance. Although it is trivial to establish, this approach to quantifying the frequentist performance of a posterior mean in a nonidentified model does not seem to have been used previously in the literature. We emphasize that these asymptotic developments assume a proper joint prior distribution, and extensions to improper priors are

not altogether obvious. Indeed, subtle issues surround nonidentified models and improper priors, as emphasized by Gelfand and Sahu (1999).

2.3 Performance of Model A Estimators

In mismeasured variable scenarios, investigators often have a rough idea about the extent of the mismeasurement. In the present context, for instance, say that the investigators are comfortable with an assessment that the two tests T_1, T_2 are "pretty good, but not perfect" surrogates for the actual exposure E. This might be encapsulated by assigning the same prior distribution to each of (p_1, p_2, q_1, q_2). As an illustration, consider assigning a Beta$(18, 4)$ prior distribution to each of these parameters independently, along with the prior $r \sim \text{Unif}(0, 1)$ for the population prevalence. For later reference we refer to this as prior (i). Also for reference, the Beta$(18, 4)$ density function appears in Figure 3. The crudeness in this prior specification derives in part from the inherent uncertainty in the Beta$(18, 4)$ distribution, but more from the lack of any discrimination between the two tests or any discrimination between the sensitivity and specificity of either test. In the absence of very substantive prior knowledge, the four classification probabilities (p_1, p_2, q_1, q_2) are treated exchangeably. We note that as a special case of exchangeability the assumption of prior independence is made for the sake of convenience. Conceivably, a more realistic attempt to elicit an exchangeable prior distribution on these four parameters might result in positive dependence a priori.

For Model A a transparent parameterization $\phi = (\phi_I, \phi_N)$ obtains by taking

$$\begin{aligned}
\phi_{I,1} &= rp_1p_2 + (1-r)(1-q_1)(1-q_2),\\
\phi_{I,2} &= rp_1(1-p_2) + (1-r)(1-q_1)q_2,\\
\phi_{I,3} &= r(1-p_1)p_2 + (1-r)q_1(1-q_2),
\end{aligned}$$

which directly determine the distribution of (T_1, T_2). It is then convenient to complete the parameterization by taking $\phi_N = (r, p_1)$. The prior density $f(\phi)$ is determined by transformation of prior (i) in the original parameterization. In doing so it is quite messy to determine the requisite Jacobian of the $\phi \to \theta$ mapping directly. Thus we work with the simply determined Jacobian of the $\theta \to \phi$ mapping instead, and use implicit differentiation. The net result is that the prior density $f(\phi)$ is readily evaluated at any given ϕ, but it is not simple to give an expression for this density.

As previously we focus on r as the parameter of interest. Following the development of Section 2.2, the posterior mean of r is identically the posterior mean of

$$\begin{aligned}
\tilde{g}(\phi_I) &= \int\int \phi_{N,1} f(\phi_N|\phi_I)\, d\phi_{N,1}\, d\phi_{N,2}\\
&= \frac{\int\int \phi_{N,1} f(\phi_I, \phi_N)\, d\phi_{N,1}\, d\phi_{N,2}}{\int\int f(\phi_I, \phi_N)\, d\phi_{N,1}\, d\phi_{N,2}}
\end{aligned} \tag{6}$$

with respect to the identifiable submodel parameterized

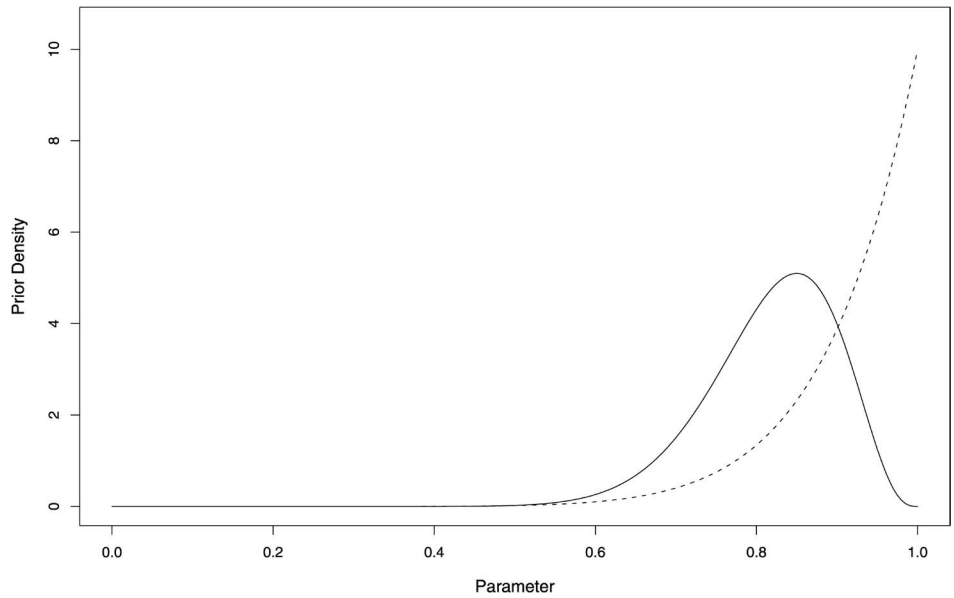

FIG. 3. *Prior distributions for the probability of correct classification. The solid curve is the* Beta$(18, 4)$ *density function and the dotted curve is the* Beta$(10, 1)$ *density function.*

by ϕ_I alone. Unfortunately, $\tilde{g}$ cannot be evaluated in closed form. We can, however, use two-dimensional numerical integration to evaluate both the numerator and denominator integrals. Thus we can compute $\tilde{g}(\phi_I) - g(\phi_I, \phi_N)$, the asymptotic bias incurred by $\hat{r}_A = E(r|\text{data})$.

Of course the first derivatives of $\tilde{g}(\phi)$ are also needed to determine the asymptotic variance of the posterior mean as in (5). Analytic differentiation of (6) is problematic because of the lack of a closed-form expression for $f(\phi)$ alluded to above. Hence we content ourselves with numerical differentiation, for instance, evaluating both $\tilde{g}(\phi_I)$ and $\tilde{g}(\phi_I + \varepsilon(1\ 0\ 0))$ using the same quadrature points, so as to approximate $\partial\tilde{g}(\phi_I)/\partial\phi_{I,1}$. In doing so we take care to check that ε is small enough and the number of quadrature points is large enough to obtain stable approximations to the derivatives. Also, since we are interested in large sample sizes at which the asymptotic variance is typically small relative to the asymptotic bias, exacting precision in computing the derivatives is not required. Thus the ARMSE (5) can be computed in the present context, albeit with some numerical effort.

We reconsider the DGM (i) parameter values given in Section 2.1. With prior (i) and a sample size of $n = 2000$ we compute $\text{ARMSE}[\hat{r}_A] = 0.015$. This compares *very* favorably to the $\text{ARMSE}[\hat{r}_B]$ values given in Figure 1, being considerably smaller when $\Delta = r_2 - r_1$ is large and very much smaller when Δ is small. Even at this large sample size, infusing crude prior information into Model A may be preferable to stratifying the population and using Model B for the sake of identifiability.

Of course this comparison may reflect some luck in choosing a crude prior that happens to yield a small asymptotic bias for the DGM (i) parameter values. Thus prior (ii) is considered, under which all four classification probabilities are assigned Beta$(10, 1)$ prior distributions. This prior has a much different shape as indicated in Figure 3, although it still reflects a crude notion of the two tests being good but perhaps not perfect. It does turn out that $\hat{r}_A$ performs less well under prior (ii), with $\text{ARMSE} = 0.037$. However, this is still quite favorable relative to $\text{ARMSE}[\hat{r}_B]$ as displayed in Figure 1, especially when $\Delta = r_2 - r_1$ is not very large.

We also reconsider DGM (ii) used in Section 2.1. With this DGM we obtain $\text{ARMSE}[\hat{r}_A] = 0.079$ with prior (i) and $\text{ARMSE}[\hat{r}_A] = 0.050$ with prior (ii). In the former case this is better than the Model B performance given in Figure 2 if Δ is less than about 0.1, and in the latter case it is better if Δ is less than about 0.15. Again, crude prior information infused into Model A can guard against the small Δ pitfall associated with the Model B estimator.

We have concentrated on the performance of prevalence estimators, although the same approach can be employed to assess the performance of sensitivity and specificity estimators under Model A. An interesting finding in related nonidentified models is that under low prevalence the data provide much more information about specificity than about sensitivity (Johnson and Gastwirth, 1991; Johnson, Gastwirth and Pearson, 2001; Gustafson, Le and Saskin, 2001).

2.4 Simulation Comparisons

2.4.1 *Performance of Model A and B estimators.* A small simulation study is carried out to augment the asymptotic comparisons made thus far. Data generating mechanisms (i) and (ii) are considered again, with the difference between subpopulation prevalences taken to be $\Delta = 0.07$. This corresponds to a setting where there is a practical difference between the subpopulation prevalences, but the asymptotic analysis suggests that the difference may not be large enough to yield good estimates. We simulate 200 data sets of size $n = 2000$ under each DGM, and for each data set we estimate the prevalence r using Model A with prior (i), Model A with prior (ii), Model B with uniform priors on all six parameters and Model B with prior (i) suitably extended [i.e., with uniform distributions on (r_1, r_2)]. Each estimate is obtained from 25,000 Gibbs sampler iterations after 1000 burn-in iterations. Under both Models A and B the Gibbs sampler is simple to implement once the parameter space is expanded to include the unobserved true exposure status of the subjects, along the lines of Joseph, Gyorkos and Coupal (1995) or Johnson, Gastwirth and Pearson (2001), for instance. The simulation results are summarized in Table 1.

As an aside, informal monitoring indicates that the mixing behavior of the Gibbs sampler in Models A and B is tolerable but not ideal. Gelfand and Sahu (1999) noted that the Gibbs sampler can mix poorly in posterior distributions based on nonidentifiable likelihoods, and this appears to be an issue of some concern in the present situation, both for nonidentifiable Model A and moderately identified Model B. While the MCMC sample size of 25,000 seems reasonable given the mixing behavior, there is some possibility of slightly improving the reported performance in Table 1 by further increasing the MCMC sample size or by

TABLE 1
Performance of four posterior means for r in a simulation study

DGM	Model	Prior	Bias	RMSE	(SIM SE)	COV	ALEN
(i)	A	(i)	−0.012	0.0186	(0.0009)	100%	0.13
	A	(ii)	−0.032	0.0366	(0.0012)	100%	0.19
	B	Uniform	0.047	0.0638	(0.0028)	93%	0.22
	B	(i)	−0.005	0.0208	(0.0010)	98%	0.12
(ii)	A	(i)	0.074	0.0755	(0.0011)	4%	0.11
	A	(ii)	0.050	0.0536	(0.0013)	91%	0.15
	B	Uniform	0.122	0.1332	(0.0038)	23%	0.22
	B	(i)	0.069	0.0717	(0.0015)	6%	0.11

NOTE. Performance is summarized by bias and RMSE, along with the coverage (COV) and average length (ALEN) of the nominal 80% equal-tailed credible interval. These quantities are estimated via 200 simulated data sets. In the case of RMSE, a simulation standard error is also given. The upper half of the table concerns data generated under DGM (i); the lower half concerns DGM (ii). In both cases $\Delta = r_2 - r_1 = 0.07$ and $n = 2000$.

choosing a different MCMC algorithm in light of the identifiability issue. For an example of designing an MCMC algorithm to work well in a nonidentified context similar to Model A, see Gustafson, Le and Saskin (2001).

In examining Table 1, note first that the empirical RMSE observed for the Model A posterior mean agrees quite closely with the ARMSE for both choices of prior and both choices of DGM. Thus the asymptotic analysis of the posterior mean under a nonidentified model is reflecting actual estimator performance. On the other hand, it is clear that for the Model B posterior mean the asymptotics have not fully kicked in yet, even with $n = 2000$. For DGM (i) the empirical RMSE under the flat prior is far smaller than the asymptotic value given in Figure 1. Moreover, the empirical RMSE is clearly very sensitive to the choice of a flat prior versus prior (i), although asymptotically the prior does not matter. Thus another aspect of the "moderate Δ" problem has emerged. Even though Model B is governed by regular asymptotics, if r_1 and r_2 are moderately close together, then the asymptotic approximation to the sampling distribution of estimators may be very inaccurate unless the sample size is extremely large.

In comparing across models in Table 1 we see that Model A with either crude prior yields a lower RMSE than Model B with a flat prior. However, Model B with prior (i) yields very similar performance to Model A with prior (i). Put succinctly, in this scenario the key to successful inference is a reasonable prior. Whether or not identifiability obtains seems to be of little import. In this regard, a referee has raised an interesting question. Is there a sense in which Models A and B can be shown to yield the same estimator performance if the same prior distributions are employed. If in fact $\Delta \neq 0$, then in one sense the answer is clearly no, because we have seen that the two estimators are governed by quite different asymptotic behavior. Model B gives asymptotically unbiased estimators, but they may have high variance if Δ is not far from zero, and the finite sample bias and variance may be influenced by the choice of prior until the sample size is very large. In contrast, Model A gives asymptotically biased estimators, but the bias and variance may both be modest. On the other hand, it may be possible to develop arguments which quantify the behavior of Model B estimators when $\Delta = 0$, and it may be possible to link this to the corresponding performance of Model A estimators, because presumably if $\Delta = 0$, then the impact of the prior distribution on Model B estimators will not diminish with sample size.

With regard to the credible intervals described in Table 1, we simply note that extreme undercoverage and overcoverage arise. Of course, with a nonidentifiable model there is no reason to expect Bayesian credible intervals to have approximately matching frequentist coverage. Theory dictates that the credible intervals from identifiable Model B have asymptotic matching frequentist coverage, yet with DGM (ii) we see extreme undercoverage. Again this speaks to the asymptotics not yet being accurate for Model B, even with $n = 2000$.

Of course both the asymptotic and simulation comparisons between Model A and Model B estimators are based on illustrative values of the true parameter values and, particularly for Model A, illustrative values of

the hyperparameters. Clearly the choice of prior distribution will impact the performance of estimators from nonidentified models even at very large sample sizes. Thus remarks about the surprisingly good performance of Model A estimators and surprisingly poor performance of Model B estimators must be tempered somewhat. We note, however, that the two prior distributions considered are arguably of somewhat modest precision. Moreover, the true parameter values in the DGMs considered are not strikingly consistent with these prior distributions. In essence, we have tried to illustrate with priors of practical accuracy, in terms of both variability and closeness to true parameter values. In fact, a companion manuscript (Gustafson, 2005) greatly extends the comparison of Model A and Model B estimators, considering many more combinations of prior distributions and true parameter values, in what are intended to be realistic configurations. The findings are quite consistent with the limited comparisons drawn here.

2.4.2 *Bayes performance.* Our asymptotic and simulation comparisons thus far have considered average performance of estimators in repeated sampling with fixed underlying parameter values in the true model. For several reasons we now turn attention to average performance across different underlying parameter values. One reason for so doing is to verify that our findings are not overly sensitive to the parameter values which have been arbitrarily chosen for the sake of illustration. Second, we wish to contrast the frequentist coverage of credible intervals with their Bayesian coverage.

We adopt the decision-theoretic point of view that nature generates parameter values (and consequently data sets) from a prior distribution, while the investigator uses a possibly different prior distribution to construct a posterior distribution. For each choice of nature's prior we simulate 200 data sets, in each instance first drawing a parameter vector from the prior and then simulating a data set of size $n = 2000$ under Model B. Bearing in mind that nature's prior assigns a uniform distribution to (r_1, r_2), $\Delta = r_2 - r_1$ has a symmetric triangular-shaped prior density on $(-1, 1)$. For each data set the posterior mean and the 80% equal-tailed credible interval for r are computed under three different model–prior combinations: Model A with prior (i), Model A with prior (ii) and Model B with a flat prior. The results appear in Table 2.

Since we are now considering average performance across small and large underlying values of Δ, we no longer expect to see Model A with a crude prior substantially outperform Model B with a flat prior. Indeed, when nature uses prior (i), all three model–prior combinations result in a similar RMSE for estimating r, and when nature uses prior (ii), Model B with a flat prior has a lower RMSE than Model A with either crude prior. Thus in aggregate the advantage of identifiability which arises when Δ is quite large slightly outweighs the benefit of crude prior information. Of course this does not mitigate the fact that estimates generated from Model B with a flat prior can be quite poor for data sets generated under small values of Δ. Figure 4 plots the absolute error $|\hat{r} - r|$ versus $|\Delta|$ for the simulated data sets under the various scenarios considered in Table 2. Clearly the error varies much less with Δ under Model A and a crude prior than under Model B with a flat prior, especially when nature employs prior (i).

TABLE 2
Bayes performance of r estimators under various settings

Nature's prior	Model	Investigator's prior	RMSE	COV	ALEN
(i)	A	(i)	0.057	77%	0.14
	A	(ii)	0.063	90%	0.21
	B	Uniform	0.065	82%	0.12
(ii)	A	(i)	0.043	62%	0.07
	A	(ii)	0.047	82%	0.11
	B	Uniform	0.025	80%	0.06

NOTE. Nature employs either prior (i) or prior (ii) to generate 200 data sets under Model B. The investigator uses either Model A with prior (i), Model A with prior (ii) or Model B with a uniform prior to obtain a posterior distribution for r. The RMSE of the posterior mean and the coverage (COV) and average length (ALEN) of the 80% equal-tailed credible interval are reported.

Of course if nature and the investigator use the same prior, then credible intervals will have exactly their nominal frequentist coverage, irrespective of whether the model is identifiable or not. The results in Table 2 are in accord with this fact. The coverage does vary when the investigator's prior does not match nature's prior, although the deviations from nominal coverage are much less extreme than those exhibited for frequentist coverage in Table 1.

2.4.3 *Performance of Model* C *estimators.* In principle the approach of Section 2.2 could be used to quantify the performance of posterior means based on nonidentifiable Model C and to determine whether there is appreciable indirect learning. However, the requisite expressions are extremely unwieldy, so we do

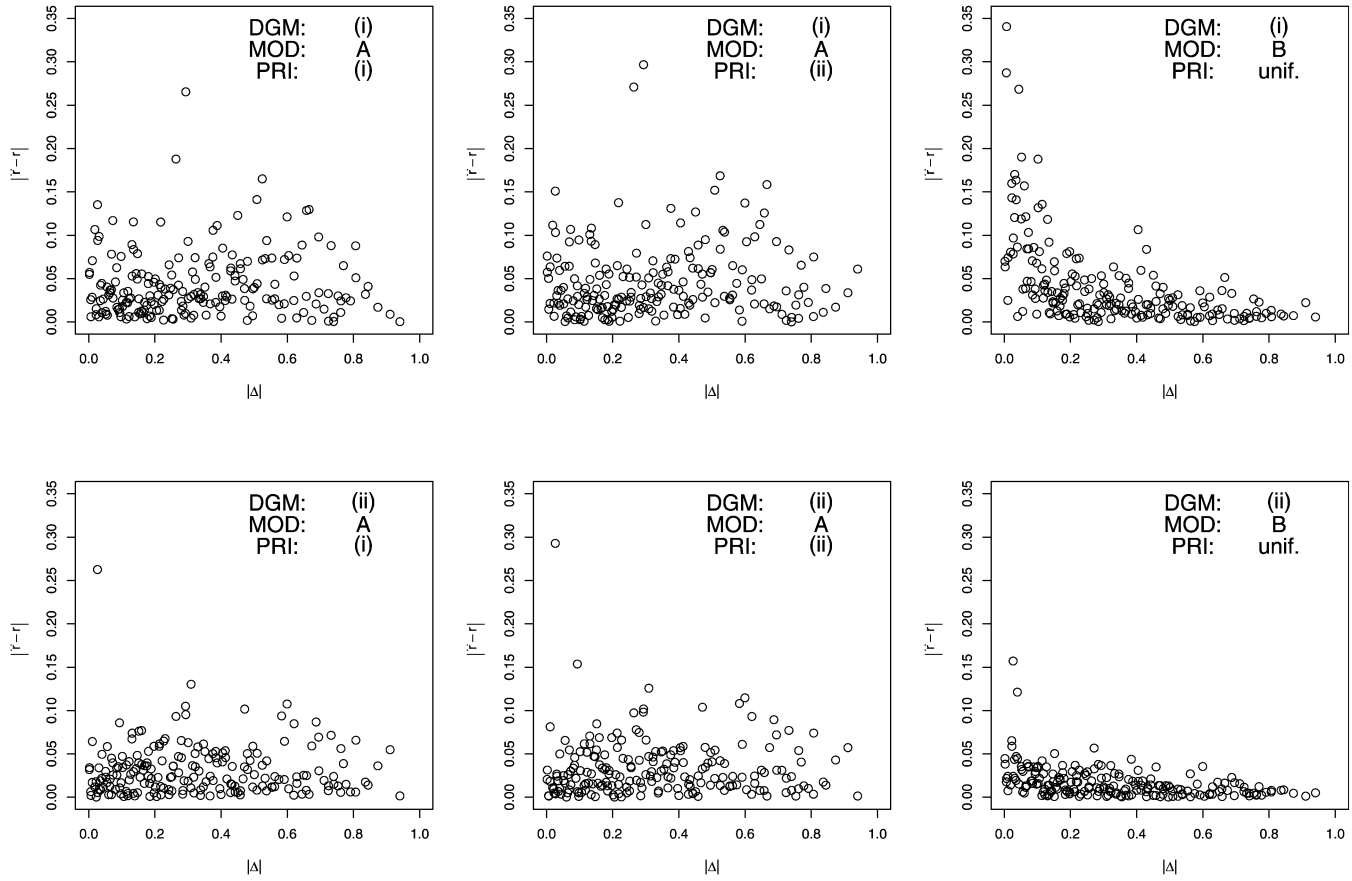

FIG. 4. *Absolute estimation error $|\hat{r} - r|$ versus absolute difference in subpopulation prevalences $|\Delta|$ for the simulated data sets in Section* 2.4.2.

not pursue this here. Rather, we carry out a small simulation study.

We consider fitting Model C to data, using a uniform prior on $(p_1, p_2, q_1, q_2, r_1, r_2)$ along with a crude prior on $(\delta_0, \delta_1 | p_1, p_2, q_1, q_2, r_1, r_2)$ that reflects a belief of "not too much" dependence between the two tests given the true exposure status. Specifically, δ_0 is assigned an exponential distribution with rate parameter $k(q_1, q_2)$, truncated to the interval $[0, \delta_{\mathrm{MAX}}(q_1, q_2)]$. Similarly, δ_1 is assigned an exponential distribution with rate $k(p_1, p_2)$, truncated to the interval $[0, \delta_{\mathrm{MAX}}(p_1, p_2)]$. To give this prior an interpretation in terms of downweighting higher correlation between T_1 and T_2 given E, we take $k(a, b) = c/[\{a(1-a)b(1-b)\}^{1/2}]$. Then a value of δ_j corresponding to conditional correlation ρ_j has a prior density which is $\exp(-c\rho_j)$ times the prior density of $\delta_j = 0$, which corresponds to $\rho_j = 0$. For the sake of illustration we take $c = 4 \log 4$, so that $\rho_j = 0.25$ is four times less likely a priori than $\rho_j = 0$ in this sense. By specifying prior densities for δ_j which are positive at zero and monotonically decreasing, we hope to avoid spurious inferential claims of dependence when, in fact, none is present. A referee has suggested assigning some prior probability to negative values of δ_j as a further safeguard against this problem.

The investigation in Section 2.1 suggests that estimation of prevalence using Model B is adversely affected by between-test correlation under DGM (ii), but not under DGM (i). Thus data are simulated under DGM (ii) with different values of ρ_j. The results appear in Table 3. We see that in the absence of correlation the Model C estimator is worse than the Model B estimator by about a factor of 2 in terms of RMSE, and by about a factor of 3 in terms of credible interval length. As the correlation increases, the difference in RMSE decreases but does not disappear, while the difference in credible interval length persists. The advantage of using Model C when there is, in fact, dependence is a better coverage rate for credible intervals, as one might expect from a model which admits the possibility of dependence.

TABLE 3
Comparison of prevalence estimators based on Models B *and* C*, with the prior distributions described in Section* 2.4.3

	Model B			Model C		
ρ	RMSE	COV	ALEN	RMSE	COV	ALEN
0	0.033	72%	0.070	0.074	93%	0.230
0.125	0.064	4%	0.070	0.093	74%	0.230
0.25	0.093	0%	0.070	0.101	23%	0.200

NOTE. The three rows correspond to DGM (i) with three different underlying values of ρ, the conditional correlation of $T_1, T_2|E$ for both values of E. For the posterior mean of r under both models, the RMSE and the coverage (COV) and average length (ALEN) of the nominal 80% equal-tailed are reported, based on 200 simulated data sets of sample size $n = 2000$. Each posterior distribution is based on 25,000 MCMC iterations after 1000 burn-in iterations.

While brevity precludes a full description of the MCMC algorithm used to fit Model C, we note that the algorithm is designed to limit the impact of nonidentifiability on MCMC performance. Specifically, (q_1, q_2, δ_0) and (p_1, p_2, δ_1) are both updated in blocks given the other parameters and the true exposure status. Thus our approach to admitting dependence between tests differs from that of Dendukuri and Joseph (2001), both in the prior downweighting of higher correlations and in the approach to MCMC fitting. For some data sets, particularly those generated under larger values of ρ_j, we do see somewhat poor mixing of the MCMC algorithm. Again, more research on good MCMC algorithms for nonidentified models is required.

As a final comment on Model C, we note that in a recent paper Black and Craig (2002) considered Bayesian model averaging across Models B and C (and several other models as well). That is, they assigned prior probabilities to the competing models which are then updated to posterior probabilities. These constitute weights for averaging across models to obtain final inferences. This approach seems interesting, although presumably it cannot obviate the nonidentifiability of Model C. In particular, there is no reason to expect the posterior probability on the correct model to increase to 1 as the sample size grows.

3. SCENARIO II

Our second scenario involves a continuous response variable and a continuous predictor variable subject to measurement error. Let Y be the response variable, let X be the unobservable predictor variable of interest and let X^* be the observable surrogate variable for X. A typical normal measurement error model might postulate that the joint distribution of (X^*, Y, X) follows

$$\begin{aligned} X^*|X, Y &\sim \mathrm{N}(X, r\lambda^2), \\ Y|X &\sim \mathrm{N}(\beta_0 + \beta_1 X, \sigma^2), \\ X &\sim \mathrm{N}(\mu, \lambda^2). \end{aligned}$$

We refer to this model as Model E. Note that this model invokes the common assumption of nondifferential measurement error, because X^* and Y are conditionally independent given X. Note as well that the given parameterization makes $r = \mathrm{Var}(X^*|X)/\mathrm{Var}(X)$ interpretable as the measurement error variance expressed as a fraction of the variance in the predictor itself.

Of course Model E implies a joint distribution for the observable quantities (Y, X^*) and thus yields a likelihood function. However, it is well known that if all six parameters $\theta = (\beta_0, \beta_1, \mu, r, \sigma^2, \lambda^2)$ are unknown, then the model is nonidentifiable. Intuitively, one can consistently estimate only five parameters: an intercept, slope and residual variance that describe the distribution of $Y|X^*$, along with a mean and variance that describe the distribution of X^*. Therefore, contracting the model by taking the value of one parameter to be known is a route to identifiability. For instance, if enough is known about the measurement error process, then r might be presumed known. We refer to the model obtained by fixing the value of r as Model D.

An alternative route to gaining identifiability is through model expansion. In a recent paper, Huang and Huwang (2001) demonstrated that the model

$$\begin{aligned} X^*|X, Y &\sim \mathrm{N}(X, r\lambda^2), \\ Y|X &\sim \mathrm{N}(\beta_0 + \beta_1 X + \beta_2 X^2, \sigma^2), \\ X &\sim \mathrm{N}(\mu, \lambda^2) \end{aligned}$$

is identifiable, even if all seven parameters $\theta = (\beta_0, \beta_1, \beta_2, \mu, r, \sigma^2, \lambda^2)$ are unknown. Henceforth we refer to this model as Model F. We also clarify that in our terminology Model F is essentially identified, because its parameter space does contain the lower dimensional subspace that corresponds to the nonidentified Model E. Initially it seems remarkable that replacing the linear regression function in Model E with the quadratic regression function in Model F leads to essential identifiability. The key is that under Model F, $\mathrm{Var}(Y|X^*)$ is no longer constant. It is now a quadratic

function of X^*. Roughly speaking then, we can consistently estimate two parameters that describe $\mathrm{Var}(Y|X^*)$, three parameters that describe $E(Y|X^*)$ and two parameters that describe the marginal distribution of X^*. Unfortunately, the distribution of (Y, X^*) implied by Model F does not have a closed form, so the Fisher information matrix is not readily evaluated to describe the performance of Model F estimators. However, Model F is readily fit to data via MCMC algorithms.

Of course identifiable Model F reduces to nonidentifiable Model E when $\beta_2 = 0$. Thus it may not be wise to use Model F if one suspects that β_2 is close to zero. Alternatively, if some prior information about r is available, then one might use Model E, its lack of identifiability notwithstanding. Alternatively, to avoid the specification of a prior one might simply fix r at a "best guess" value and use the identifiable Model D.

As an illustrative example we consider a subset of data from the HARVEST study (Palatini, Pessina and Dal Palu, 1993) as described and reported by Schork and Remington (2000). Along the lines of an example considered by the latter authors, we examine the relationship between systolic blood pressure (SBP) and heart rate (HR) among the $n = 311$ subjects with clinical measurements of both at the 5-year follow-up examination. In fact we take Y and X^* to be linearly rescaled versions of the recorded HR and $\log(\mathrm{SBP} - 50)$ values, where the latter transformation is commonly applied to blood pressure measurements (Carroll, Ruppert and Stefanski, 1995, Section 4.5). The rescaling is for the sake of convenience, so that both variables have sample mean 0 and sample variance 1. Recognizing that there is substantial short-term fluctuation in blood pressure measurements, we are implicitly viewing the unobserved X as a long-term average of X^*. A scatterplot of the data appears in Figure 5. While the plot does not indicate obvious curvature, one point emphasized in Gustafson (2002) is that measurement error reduces the power of diagnostic plots to detect departures from model assumptions.

In fitting Models D, E and F to these data, we assign independent $\mathrm{N}(0, 1)$ priors to the regression coefficients. In light of the data standardization, these are relatively diffuse priors. The variance components σ^2 and λ^2 are taken to have $\mathrm{IG}(1/2, 1/2)$ priors. In the spirit of Kass and Wasserman (1995) these can be interpreted as unit-information priors, with prior guesses of 1 for both variances.

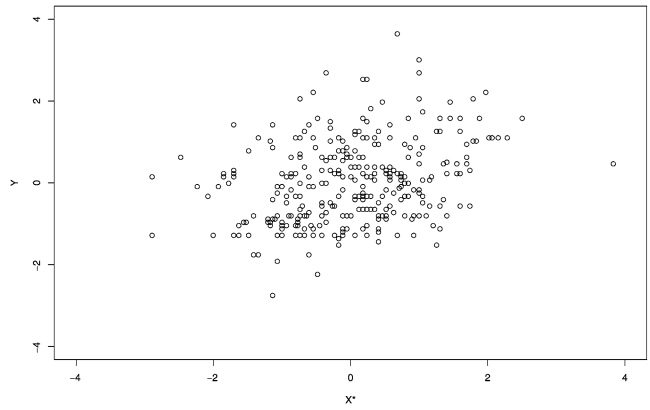

FIG. 5. *Scatterplot of standardized heart rate* (Y) *versus standardized transformed systolic blood pressure* (X^*) *from the HARVEST study.*

As an initial guess for r we note that in a somewhat similar study discussed by Carroll, Ruppert and Stefanski (1995, Section 4.5), repeated SBP measurements produced an estimated value of $r = 0.17$ for the ratio of $\mathrm{Var}(X^*|X)$ to $\mathrm{Var}(X)$. Thus in fitting Model D we take $r = 0.17$ as known. As an alternative which acknowledges the uncertainty in this guess, we fit Model E with the prior $r \sim \mathrm{Beta}(3.21, 11.79)$. This prior has its mode at $r = 0.17$ but encapsulates considerable uncertainty around this value. Finally Model F is fit with both (i) $r \sim \mathrm{Beta}(3.21, 11.79)$ and (ii) $r \sim \mathrm{Unif}(0, 1)$. The latter choice is made to investigate whether the essential identifiability of Model F obviates the need for an informative prior on r.

We note that with these choices of priors Models D, E and F are readily fit to the data via MCMC methods. In particular, most quantities can be updated via the Gibbs sampler, although more specialized Metropolis–Hastings updates are needed to update r in Models E and F, and also X in Model F. Informal monitoring of the MCMC output indicates reasonable Markov chain mixing. The following results are based on 50,000 MCMC iterations.

Kernel-density estimates of posterior densities for β_1 are given in the left panel of Figure 6. Clearly the posterior distribution of β_1 is very similar under Models D, E and F with the informative prior (i). In particular, there is only a slight increase in posterior variance associated with the loss of identifiability in moving from Model D to Model E. Further expanding to Model F while keeping the same prior on r leaves the posterior distribution of β_1 virtually unchanged, despite the formal gain of identifiability. In addition, using Model F

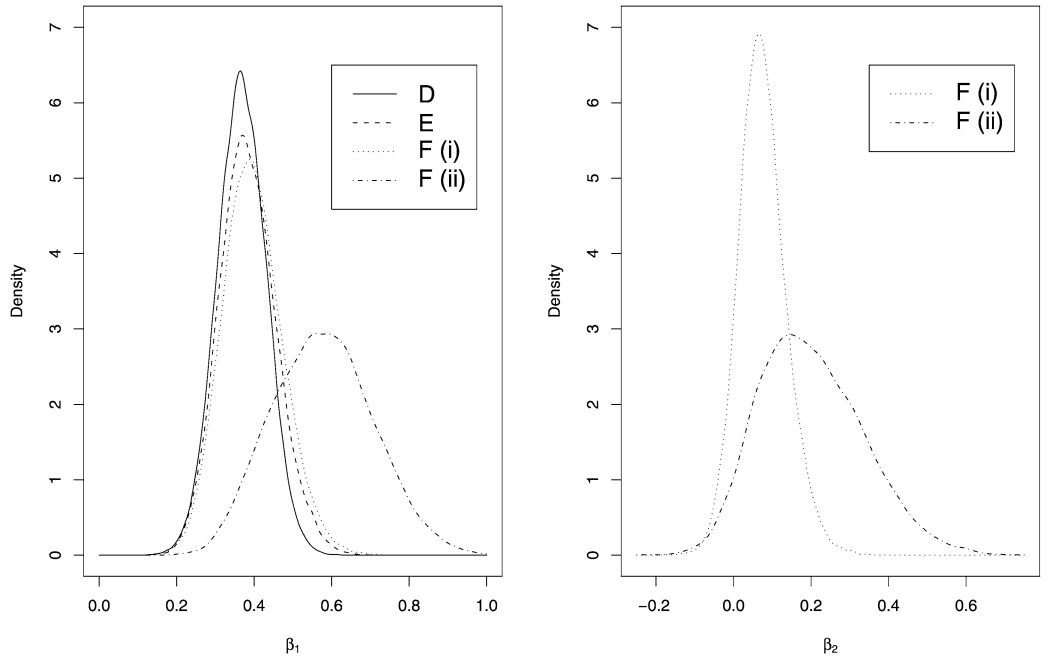

FIG. 6. *Posterior distributions on* β_1 (left panel) *and* β_2 (right panel) *for the HARVEST example.*

with a flat prior on r yields a substantial increase in posterior variance, indicating that the gain in identifiability cannot replace the strength of the informative prior. The lack of advantage with Model F is surprising given that the posterior distribution of β_2 is centered away from zero (right panel of Figure 6), particularly under prior (ii). In all, it seems that prior information is quite important in this example, while identifiability may not be particularly crucial. The remainder of Section 3 aims to shed more light on these notions.

3.1 Performance of Model E Estimators

Clearly Model E postulates joint normality of (X^*, X, Y) and, hence, joint normality of the observable quantities (X^*, Y), with parameters readily determined as functions of the six parameters governing Model E. In fact, standard multivariate normal theory yields the distributions of $Y|X^*$ and X^*, which form the basis for a transparent parameterization $\phi = (\phi_I, \phi_N)$, with the components of ϕ_I taken to be

$$\begin{aligned}
\beta_0^* &= \beta_0 + \mu\beta_1/(1+r),\\
\beta_1^* &= \beta_1/(1+r),\\
\mu_* &= \mu,\\
\sigma_*^2 &= \sigma^2 + \beta_1^2\lambda^2 r/(1+r),\\
\lambda_*^2 &= \lambda^2(1+r),
\end{aligned}$$

while $\phi_N = r$. Then the distribution of the observable data (X^*, Y) depends on ϕ only through ϕ_I according to

$$\begin{aligned}
Y|X^* &\sim \mathrm{N}(\beta_0^* + \beta_1^* X^*, \sigma_*^2),\\
X^* &\sim \mathrm{N}(\mu_*, \lambda_*^2).
\end{aligned}$$

The developments of Section 2.2 can be applied to study the asymptotic performance of posterior means arising from Model E. As an illustrative example, suppose the analyst assigns independent priors to the six original parameters, specifically $\beta_0 \sim \mathrm{N}(0,1)$, $\beta_1 \sim \mathrm{N}(0,1)$, $\mu \sim \mathrm{N}(0,1)$, $\sigma^2 \sim \mathrm{IG}(0.5, 0.5)$, $\lambda^2 \sim \mathrm{IG}(0.5, 0.5)$ and $r \sim \mathrm{Beta}(\alpha_1, \alpha_2)$. We consider three particular choices of hyperparameters (α_1, α_2) for the prior on r. Prior (i) sets $(\alpha_1, \alpha_2) = (1,1)$; that is, r is assigned a uniform prior in the absence of any subjective knowledge. Prior (ii) sets $(\alpha_1, \alpha_2) = (7.6, 14.1)$, giving $E(r) = 0.35$, $\mathrm{SD}(r) = 0.10$, while the more concentrated prior (iii) sets $(\alpha_1, \alpha_2) = (24.9, 58.1)$, giving $E(r) = 0.30$, $\mathrm{SD}(r) = 0.05$.

We consider what happens when the data are generated according to $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2 = 0.25$, $\mu = 0$, $\lambda^2 = 1$ and $r = 0.25$. Note that for this DGM the true value of r is 1 standard deviation below the mean with respect to both priors (ii) and (iii). Thus in a rough sense both priors are typically representative of the truth, although of course prior (iii) represents stronger knowledge.

Following Section 2.2, the key to describing the asymptotic behavior of Model E estimators is the prior distribution of $\phi_N|\phi_I$. In the present context it is easy to check that the Jacobian of the $\phi \to \theta$ mapping is 1 and, consequently,

$$
\begin{aligned}
&f(r|\beta_0^*, \beta_1^*, \mu_*, \sigma_*^2, \lambda_*^2) \\
&\quad \propto f_{\mathrm{N}}(\beta_1^*(1+r))\, f_{\mathrm{IG}}\big(\sigma_*^2 - r(\beta_1^*)^2\lambda_*^2\big) \\
&\quad\quad \cdot f_{\mathrm{IG}}\big(\lambda_*^2/(1+r)\big) \\
&\quad\quad \cdot r^{\alpha_1-1}(1-r)^{\alpha_2-1} I_{(0,m(\beta_1^*,\sigma_*^2,\lambda_*^2))}(r),
\end{aligned} \tag{7}
$$

where $f_{\mathrm{N}}(\cdot)$ is the $\mathrm{N}(0,1)$ density function, $f_{\mathrm{IG}}(\cdot)$ is the $\mathrm{IG}(0.5, 0.5)$ density function and $m(\beta_1^*, \sigma_*^2, \lambda_*^2) = \min[\sigma_*^2/\{(\beta_1^*)^2\lambda_*^2\}, 1]$. Thus for a given quantity of interest $\psi = g(\phi)$, one can easily evaluate both the value and the first derivatives of

$$
\begin{aligned}
&\tilde{g}(\beta_0^*, \beta_1^*, \mu_*, \sigma_*^2, \lambda_*^2) \\
&\quad = \int g(\beta_0^*, \beta_1^*, \mu_*, \sigma_*^2, \lambda_*^2, r) \\
&\quad\quad \cdot f(r|\beta_0^*, \beta_1^*, \mu_*, \sigma_*^2, \lambda_*^2)\, dr
\end{aligned}
$$

via one-dimensional numerical integration. Thus the bias and asymptotic variance of $\hat{\psi} = E(\psi|\text{data})$ as in (5) are readily evaluated.

Of course (7) evaluated at the true value of ϕ_I is the large-sample limiting posterior density of r. For each prior (i)–(iii) the prior density and limiting posterior density of r under the illustrative DGM appear in Figure 7. In the case of the uniform prior (i) there is a surprising amount of indirect learning about r. There is slightly less such learning under prior (ii), and almost none at all under the sharper prior (iii).

To be more specific, consider estimating r and β_1 by their posterior means. While r is not likely to be of direct inferential interest, it is clearly the crucial intermediary quantity in learning about the (Y, X) relationship from observations on (Y, X^*). Following up on Figure 7, it would be useful to have a quantification of how much can be learned about this parameter. Of course, in most problems inference about β_1 is likely to be more scientifically relevant.

For priors (i)–(iii) the asymptotic bias, variance and RMSE for a sample size of $n = 250$ appear in Table 4. As expected, both the bias and variance of $\hat{r} = E(r|\text{data})$ decrease as the prior distribution for r improves. The surprising feature, which relates back to the indirect updating witnessed in Figure 7, is that the performance of $\hat{r}$ under the uniform prior (i) is not terrible, either in absolute terms or relative to priors (ii) and (iii). For the sake of comparison, Table 4 also gives the RMSE if Model D is employed, with the fixed value of r taken to be the prior mean of r. For instance, with prior (i) we simply have $\hat{r} = 0.5$, regardless of the

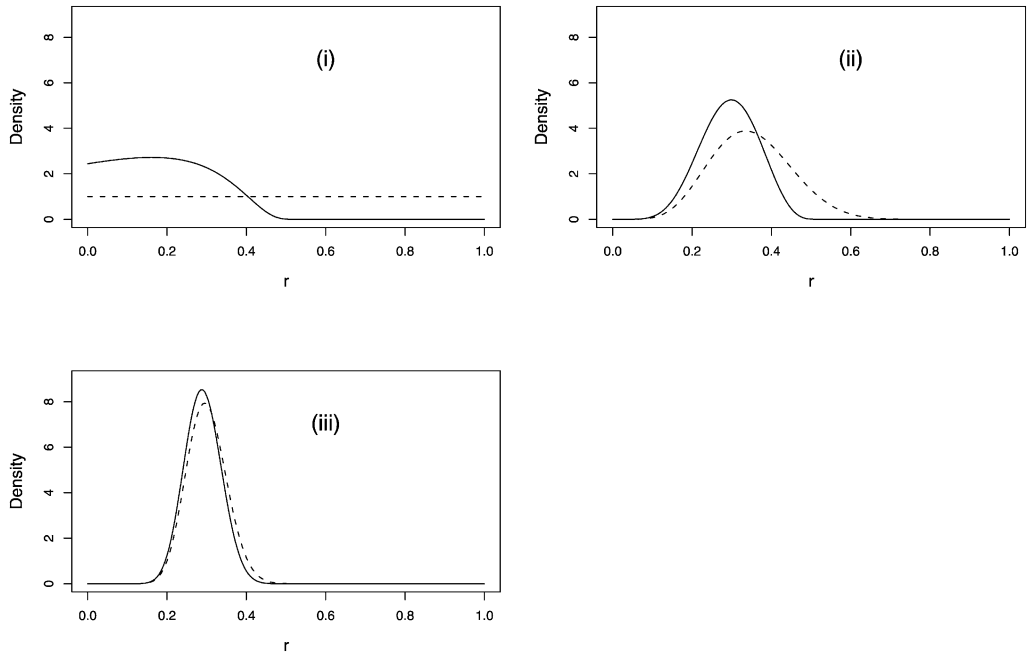

FIG. 7. *Prior* (*dashed curve*) *and limiting posterior* (*solid curve*) *densities for* r *under Model* E, *with the illustrative DGM and priors* (i)–(iii). *The true value of* r *is* 0.25.

TABLE 4
Asymptotic bias and variance of posterior means under Model E *under the illustrative DGM and priors* (i)–(iii)

	Estimating r				Estimating β_1			
Prior	**Bias**	**SD**	**RMSE**	**(RMSE-D)**	**Bias**	**SD**	**RMSE**	**(RMSE-D)**
(i)	−0.056	0.036	0.066	(0.250)	−0.045	0.038	0.059	(0.208)
(ii)	0.044	0.028	0.052	(0.100)	0.036	0.041	0.055	(0.095)
(iii)	0.040	0.008	0.041	(0.050)	0.032	0.046	0.056	(0.064)

NOTE. Results are given for both $\hat{r} = E(r|\text{data})$ and $\hat{\beta}_1 = E(\beta_1|\text{data})$. The approximate standard deviation (SD) and RMSE are based on a sample size of $n = 250$. The RMSE incurred under Model D, with r assumed known and equal to the prior mean, also appears in parentheses.

data. The RMSE under Model E with a prior is consistently lower than under Model D with a corresponding best guess, with the difference being very large in some cases. Thus it is clearly worthwhile to formulate a prior and use Model E rather than simply fix r at a best guess value in Model D. Particularly, the use of a prior allows one to benefit from indirect learning about r.

Table 4 also shows that the bias in estimating β_1 decreases as the prior distribution for r improves. In terms of RMSE, however, this improvement is offset by a corresponding increase in variance. To understand this, note that in terms of the transparent parameterization, $\beta_1 = \beta_1^*(1 + r)$. The asymptotic variance of $\hat{r}$ decreases as the prior improves, and by definition the asymptotic variance of $\hat{\beta}_1^*$ is unaffected by the prior. However, there is a negative asymptotic covariance between $\hat{r}$ and $\hat{\beta}_1^*$. This covariance decreases in magnitude as the prior improves, thereby producing the overall increase in variance. Surprisingly then, β_1 can be estimated about as well under the flat prior (i) as under the sharp prior (iii).

3.2 Performance of Model F Estimators

Model F, while identifiable, does not have a closed-form Fisher information matrix. Thus a simulation study is undertaken to evaluate the performance of Model F estimators obtained via MCMC methods. The illustrative Model E priors are extended by taking $\beta_2 \sim \text{N}(0, 1)$. Three DGMs are constructed by extending the DGM of the previous section. In particular, the same values of $(\beta_0, \beta_1, \mu, r, \sigma^2, \lambda^2)$ are used in tandem with (i) $\beta_2 = 0$, (ii) $\beta_2 = 0.125$ and (iii) $\beta_2 = 0.25$. The last value is deliberately chosen to maximize the curvature of the true regression function subject to the function being monotone on the interval from $\mu - 2\lambda$ to $\mu + 2\lambda$ that contains the bulk of the X distribution. To be more specific, it is easy to verify that the regression function will be monotone on $(-c, c)$ if $|\beta_2| \le (2c)^{-1}|\beta_1|$. Since many relationships of interest are monotone, in a practical sense DGM (iii) represents an extreme degree of curvature.

Under each DGM (i)–(iii), we simulate 200 data sets of size $n = 250$. Posterior means and credible intervals for r and β_1 are computed for each data set under each prior (i)–(iii). The results are summarized in Table 5. We note that the MCMC algorithm seems to mix quite well in most cases, but less well in the case of DGM (i) and prior (i), that is, no curvature and no prior informa-

TABLE 5
Performance of Model F *posterior distributions in estimating* r *and* β_1

		Estimating r DGM			Estimating β_1 DGM		
Prior		**(i)**	**(ii)**	**(iii)**	**(i)**	**(ii)**	**(iii)**
(i)	RMSE	0.064	0.083	0.080	0.055	0.064	0.062
	ALEN	0.337	0.276	0.212	0.274	0.227	0.185
	COV	0.970	0.910	0.795	0.980	0.925	0.855
(ii)	RMSE	0.055	0.062	0.063	0.052	0.056	0.056
	ALEN	0.202	0.187	0.165	0.189	0.178	0.167
	COV	1.000	0.870	0.810	0.960	0.915	0.885
(iii)	RMSE	0.039	0.041	0.041	0.051	0.051	0.051
	ALEN	0.119	0.116	0.109	0.152	0.150	0.151
	COV	1.000	0.900	0.805	0.885	0.860	0.875

NOTE. DGMs (i)–(iii) correspond to increasing curvature in the regression function and priors (i)–(iii) correspond to increasing prior information about r as described in the text. Under each condition the RMSE of the posterior mean, as well as the average length (ALEN) and coverage (COV) of the 80% equal-tailed credible interval are reported, based on 200 simulated data sets of size $n = 250$. Each posterior mean and credible interval is computed using 20,000 MCMC iterations after 1000 burn-in iterations.

tion. As mentioned earlier, this is not surprising in light of other MCMC experience in nonidentified or weakly identified scenarios.

Table 5 reveals that, for both estimands, increased curvature in the regression function leads to shorter credible intervals with closer to nominal coverage. Thus a benefit does accrue when the Model F parameter values are further away from the submodel that corresponds to Model E. On the other hand, the RMSE of the posterior mean actually increases somewhat with β_2. Given this and given that DGM (iii) represents an extreme degree of curvature in the sense described above, we conclude that the benefit associated with curvature in the regression function that leads to identifiability is quite modest. As in Scenario I, model expansion is not a panacea.

Table 5 also indicates that the Model F estimators under the three priors perform quite similarly to their Model E counterparts, particularly under DGM (i). Again this speaks to prior information, and not identifiability, being the key to successful inference.

4. DISCUSSION

The conventional view of identifiability might be crudely summarized as "identifiability good, nonidentifiability bad." The findings in this paper, however, indicate that sometimes a more nuanced view is required. We have exhibited realistic scenarios where a moderate amount of prior information leads to reasonable inferences from a nonidentified model, and scenarios where ridiculously large sample sizes may be required to obtain reasonable inferences from an identified model. These issues are particularly germane to problems involving measurement error and misclassification, where a lack of knowledge about the extent of mismeasurement often raises questions about identifiability. Indeed, what-if analyses, which give inferences under a variety of assumed magnitudes of mismeasurement, are common. If a more definitive analysis is required, our findings support the "honest" approach of formulating the most realistic model and prior possible, without particular regard for whether the model is identifiable or not. If it happens to be nonidentified, then either contracting or expanding the model for the sake of identifiability can in fact lead to poorer estimator performance, particularly if one resorts to flat prior distributions for the sake of objectivity.

It is self-evident that gaining identifiability by model expansion cannot be very beneficial if the true parameter values in the expanded model lie close to the original submodel. That is, there is a "danger zone" of parameter values in the expanded model under which estimators will perform poorly. More precisely, the danger zone corresponds to parameter values close to the lower dimensional parameter subspace given in (1). What has been demonstrated, in Scenario I particularly, is that these danger zones can be surprisingly large. Thus the appropriateness of model expansion can only be judged relative to prior knowledge. If there is good reason to believe a priori that the true parameter values will lie outside the danger zone, then use of the expanded model and a flat prior will likely be effective. Without such prior knowledge, however, this is a risky strategy. Put another way, using model expansion to gain identifiability does not equate with unfettered freedom to use flat priors.

While the development in Section 2.2 is very simple, it may come as a surprise that one can use standard asymptotic theory to describe the performance of estimators obtained from a nonidentifiable model and a particular prior distribution, and to characterize how much indirect learning about nonidentifiable parameters can occur. The notion of indirect learning speaks in favor of assigning a crude subjective prior to nonidentifiable parameters rather than fixing such parameters at best guess values. We might describe this as a "soft" rather than "hard" model contraction. In Scenario II, for instance, Model E with a crude prior for r substantially outperforms Model D with a corresponding best guess for r.

A limiting feature of our investigation is the inability to make conclusions which hold broadly over large ranges of underlying parameter values. In the case of nonidentifiable models, estimator performance must be evaluated on a prior by prior and DGM by DGM basis, and substantial variation can arise. Similarly, if an identifiable model is obtained by expanding a nonidentifiable model, performance can vary considerably with the distance of the underlying parameter values in the identifiable model from the original submodel. For the sake of brevity we have not attempted comprehensive evaluations of estimator performance across many parameter values and prior distributions. However, our findings speak to the need for such evaluations so as to make focused recommendations about study design and analysis in scenarios akin to those examined here. In the case of Models A and B from Scenario I, Gustafson (2005) undertakes a more comprehensive comparison.

Looking beyond mismeasured variable scenarios, the issues raised here and the simple tools developed in Section 2.2, in particular, are probably relevant to Bayesian analysis in many other contexts. The MCMC methods permit Bayesian analysis without much thought about important issues. A prime example is that even the fundamental question of posterior propriety can be swept under the rug, possibly with disastrous consequences (see, e.g., Hobert and Casella, 1996). Similarly, MCMC methods can be applied without thought about whether the model is identifiable, whether a danger-zone problem might exist or which parameters require a crude subjective prior rather than a flat prior to make reasonable inferences. Many complex models with numerous parameters that are fit with MCMC methods may, in fact, be nonidentifiable or close to nonidentifiable, and this issue seems deserving of closer scrutiny.

ACKNOWLEDGMENTS

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Institutes of Health Research. The author is grateful to the editors and referees for comments which led to an improved manuscript.

REFERENCES

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

BLACK, M. A. and CRAIG, B. A. (2002). Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* **21** 2653–2669.

BRENNER, H. (1996). How independent are multiple "independent" diagnostic classifications? *Statistics in Medicine* **15** 1377–1386.

CARROLL, R. J., RUPPERT, D. and STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC, Boca Raton, FL.

DAWID, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 1–31.

DENDUKURI, N. and JOSEPH, L. (2001). Bayesian approaches to modelling the conditional dependence between multiple diagnostic tests. *Biometrics* **57** 158–167.

DREWS, C. D., FLANDERS, W. D. and KOSINSKI, A. S. (1993). Use of two data sources to estimate odds-ratios in case-control studies. *Epidemiology* **4** 327–335.

FRYBACK, D. G. (1978). Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research* **11** 423–434.

GELFAND, A. E. and SAHU, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *J. Amer. Statist. Assoc.* **94** 247–253.

GEORGIADIS, M. P., JOHNSON, W. O., GARDNER, I. A. and SINGH, R. (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Appl. Statist.* **52** 63–76.

GUSTAFSON, P. (2002). On the simultaneous effects of model misspecification and errors-in-variables. *Canad. J. Statist.* **30** 463–474.

GUSTAFSON, P. (2005). The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. *Statistics in Medicine* **24** 1203–1217.

GUSTAFSON, P., LE, N. D. and SASKIN, R. (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* **57** 598–609.

HOBERT, J. P. and CASELLA, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91** 1461–1473.

HUANG, Y. H. S. and HUWANG, L. (2001). On the polynomial structural relationship. *Canad. J. Statist.* **29** 495–512.

HUI, S. L. and WALTER, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics* **36** 167–171.

JOHNSON, W. O. and GASTWIRTH, J. L. (1991). Bayesian inference for medical screening tests: Approximations useful for the analysis of acquired immune deficiency syndrome. *J. Roy. Statist. Soc. Ser. B* **53** 427–439.

JOHNSON, W. O., GASTWIRTH, J. L. and PEARSON, L. M. (2001). Screening without a "gold-standard": The Hui–Walter paradigm revisited. *American J. Epidemiology* **153** 921–924.

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 J. Epidemiology* **141** 263–272.

KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90** 928–934.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York.

NEATH, A. A. and SAMANIEGO, F. J. (1997). On the efficacy of Bayesian inference for nonidentifiable models. *Amer. Statist.* **51** 225–232.

PALATINI, P., PESSINA, A. C. and DAL PALU, C. (1993). The hypertension and ambulatory recording venetia study (HARVEST): A trial on the predictive value of ambulatory blood pressure monitoring for the development of fixed hypertension in patients with borderline hypertension. *High Blood Pressure* **2** 11–18.

QU, Y., TAN, M. and KUTNER, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52** 797–810.

SCHORK, M. A. and REMINGTON, R. D. (2000). *Statistics with Applications to the Biological and Health Sciences*, 3rd ed. Prentice–Hall, Upper Saddle River, NJ.

TORRANCE-RYNARD, V. L. and WALTER, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* **16** 2157–2175.

VACEK, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* **41** 959–968.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25.

Comment

Alan E. Gelfand and Sujit K. Sahu

This impressive paper presents two elaborately discussed examples that illustrate nested model settings which relate model expansion and contraction to identifiability. The examples are novel in that they illuminate identifiability in nonhierarchical settings. In most Bayesian modeling, nonidentifiability is implicit in hierarchical models where data and hyperparameters are conditionally independent given first stage parameters. Also attractive is the sandwiching of a nonidentifiable model between two identifiable models in the first example and vice versa in the second.

The conclusions drawn are very useful and have some connection to results reported in Gelfand and Sahu (1999). For instance, the author uses reparameterization to "the identifiable part" with iterated expectation to argue for application of standard asymptotic theory to posterior means. Such reparameterizations need not be straightforward for models that are nonhierarchical. However, with this parameterization Gelfand and Sahu considered posterior propriety. They argued that an improper posterior for ϕ_I, where $\phi = (\phi_I, \phi_N)$, will have a *unique* proper posterior distribution if the likelihood can be written as a function of ϕ_I only and the parameter space can be written as a product of the spaces for ϕ_I and ϕ_N. The posterior distribution of ϕ_I, called the *embedded* posterior, is not affected by the implied prior distribution on the nonidentifiable parameters ϕ_N.

In this case, a related question they investigated is, "Will the MCMC output for the identifiable parameters ϕ_I converge when a Gibbs sampler is run on ϕ with ϕ having an improper posterior?" For generalized linear models they showed that the embedded proper posterior distributions can be reconstructed from the output of an MCMC sampler implemented to obtain "samples" from the joint improper posterior distribution, although some precaution (e.g., recentering of the parameters) is needed.

It is perhaps worth reiterating that under proper priors there is no identifiability problem within the Bayesian framework, perhaps only that there may be no Bayesian learning for some parameters. This suggests that one should fit the model(s) that one is interested in; the role of expansion or contraction would be viewed primarily with regard to facilitating computation. Indeed this seems to be the author's message in Section 2.1, where the difficulties that can arise by working with Model B are well demonstrated.

Section 2.3 illuminates this even further, noting that even crude prior information in Model A can "guard against the small Δ pitfall associated with Model B." In this regard, it would be interesting to see the implications of the crude prior in terms of the induced prior on the identifiable part of the transformation. These priors are not easily accessible analytically but can be studied through simulation.

With nested nonhierarchical models it is often the case that, when the reduced model is true, the Bayes factor for the reduced model relative to the full model tends to ∞ as sample size grows large. Why do you believe this is not the case for Models B and C?

In Scenario II the version of the measurement error model you discuss is the so-called MEM model. It is interesting to look at the alternative Berkson model here. (See, e.g., Carroll, Ruppert and Stefanski, 1995, for full discussion of the Berkson version.) In the setting of the paper, it takes the form $f(Y|X)f(X|X^*)$ with $X|X^* \sim \mathrm{N}(X^*, r\lambda^2)$. The distribution for X^* need not be modeled; the X^*'s can be taken as fixed. Then regardless of whether $f(Y|X)$ is as in Model E or F, we still have nonidentifiability unless r is fixed; we cannot separate r and λ^2. In some sense this is more natural than the MEM specification.

Last, the use of the Kullback–Leibler (KL) divergence as a distance between models is well-established. In recent work which has some connection to the author's, Sahu and Cheng (2003) used it to determine the number of components in mixture distributions fitted to data. They argue that if the data genuinely arise from a model with a lower dimensional parameter space, then the additional parameters in the expanded model with a higher dimensional parameter space will not represent true structure, and as a result the expanded model can be collapsed. Sahu and Cheng (2003) developed an easy to implement upper bound on the KL divergence measure between two mixture densities and performed a Bayesian test to decide whether to contract the ex-

Alan E. Gelfand is Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, USA (e-mail: alan@stat.duke.edu). Sujit K. Sahu is Senior Lecturer, School of Mathematics, S^{3}RI, University of Southampton, Southampton SO17 1BJ, UK (e-mail: S.K.Sahu@maths.soton.ac.uk).

panded model. This idea makes sense in the current context as well. The KL divergence measure can be used to characterize the danger zone between the expanded model and its submodel.

Comment

Wesley O. Johnson and Timothy E. Hanson

1. INTRODUCTION

We congratulate Professor Gustafson on an exceptionally timely and interesting article. Historically statisticians have tended to avoid models that lack identifiability for obvious reasons. However, it is clear that realistic situations arise where there is simply insufficient data to estimate all quantities of interest, as in the situations discussed here. Bayesian modeling is especially useful under these circumstances, provided subject matter experts are available to provide useful information for incorporation into the data analysis and most particularly to provide information to make up for *missing* data. The author has emphasized frequentist properties of Bayesian point estimators under models that lack identifiability, but which incorporate modest prior information, and has argued that the incorporation of prior information may be preferable to expanding the model and the data to achieve identifiability. He cites Johnson, Gastwirth and Pearson (2001), who argued for model expansion to achieve identifiability and consistency.

We totally agree with Gustafson that a little (or a lot) of prior information can be a very good thing. It is unfortunate that Johnson, Gastwirth and Pearson perhaps underemphasized the role of the prior in the published version due to final editing. We will attempt to rectify this here by highlighting Gustafson's argument. We believe that when *good* prior information is available, it should always be incorporated.

Section 2 comments on Models A, B and C from our point of view. Section 3 discusses two simple illustrations where the asymptotic conditional posteriors of $\phi_N|\phi_I$ are easily obtained. Section 4 discusses computational issues. Section 5 presents a geometric argument for the lack of identifiability of Model C, even with the addition of more samples from distinct populations, and discusses a hierarchical model for which consistent estimates of the sensitivity and specificity parameters exist. Section 6 gives concluding remarks.

Wesley O. Johnson is Professor, Department of Statistics, University of California, Irvine, California 95697, USA (e-mail: wjohnson@uci.edu). Timothy E. Hanson is Assistant Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87123, USA.

2. MODELS A, B AND C REVISITED

Professor Gustafson's main example illustrates the potential difficulty that is associated with the use of identifiable Model B to replace nonidentifiable Model A, where Model A relies on the use of substantive, though not particularly precise, prior information. He argues successfully that, with very large samples, a little good prior information using Model A (which does not wash out as the sample size increases) can result in better (frequentist) point estimates than if one uses the identifiable model B, with the sample size so large (effectively infinity compared to perhaps more realistic sample sizes) that the prior information has washed out. We would like to emphasize his point that part of the reason for the success of Model A over Model B is that in making the comparison, the prior information in Model A remains while the prior information in Model B has been eliminated through the use of standard asymptotics.

It is subsequently pointed out in Section 2.4 that the asymptotics for Model B may require very large sample sizes before the information in the prior is actually eliminated. Of course, if the prior information is of good quality (by good quality, we simply mean that it is consistent with the truth, but not necessarily highly focused on the truth), as is understood in Gustafson's arguments, the fair comparison between the two models would involve leaving the prior information in both models, as is done in the comparisons based on simulated data. It is noted in Section 2.4 that Model B seems to fare much better than when the prior information has been eliminated. We would argue that these comparisons, though not as elegant as the ones presented earlier, are more appropriate. That is not to say that we were not interested in the earlier results, since it is the comparison between the analytic results and those based on simulation that makes it clear that the asymp-

totics may not be very useful and in fact may do harm if real prior information is available.

There do exist formal asymptotic *Bayesian* methods that allow for the prior to remain as the sample size tends to infinity, for models that are identifiable or not (Yee, Johnson and Samaniego, 2002; Su and Johnson, 2005). Yee, Johnson and Samaniego (2002) also considered the two-test one-population diagnostic screening problem and data presented in Joseph, Gyorkos and Coupal (1995). The two-population version could have been similarly handled. We mention here that the exact Bayesian methods developed by Johnson, Gastwirth and Pearson (2001) for the two-test scenario only had convergence problems when we considered large sample sizes. In this instance, the asymptotic methods in Yee, Johnson and Samaniego (2002) gave very accurate results virtually instantaneously since they only involve algebraic manipulations and a simple iterative scheme (somewhat like the expectation–maximization algorithm) with, of course, no Monte Carlo sampling.

Moving on to nonidentifiable Model C, Georgiadis, Johnson, Gardner and Singh (2003) have argued based on a very large collection of simulated data, that moderate precision in the prior for at least two parameters is necessary in order to make reasonable inferences for all eight parameters. Moreover, they also argued that if the sensitivity and specificity of one of the tests are sufficiently large (>0.98) or have relatively small correlations (0.1–0.2), the effect of correlation between the tests may be practically irrelevant. With larger correlations (0.5–0.6) and smaller test accuracies, they found that using Model B could be disastrous. They used informative priors, both correctly and incorrectly specified and both precise and imprecise, on at least two parameters.

Gustafson found an appreciable loss in mean squared error associated with using Model C for estimating the overall prevalence when Model B was appropriate and found that with a correlation of 0.25, the frequentist properties of point estimation using Models B and C are comparable. He has used uniform priors on the prevalences and on the test accuracies, and somewhat informative priors on the test covariances. His simulation invites this kind of prior since he considers performance when averaged over a variety of different circumstances. The real utility of this simulation applies to a scenario where a data analyst uses the above models to analyze a variety of data sets, where the only subject matter input is that there may be a little correlation between the tests. The use of uniform priors on the prevalences implies that it is expected that prevalences will cover the continuum with equal plausibility and that, similarly, test accuracies are equally unpredictable and can as easily be below 0.5 as above.

While we understand Gustafson's purpose in performing the simulation, and we find the results interesting, we would like to assert our belief that in many if not most real problems, the data analyst will be working with someone who is able to obtain independent information about the test accuracies and/or the prevalences so that appropriate "prior" information can be incorporated into the analysis. The author's forceful defense of Model A is implicitly based on this premise and so we would simply like to emphasize this point here as well. Clearly, good prior information is going to enhance the frequentist properties of the Model C estimators as it did for Model A estimators.

Our bottom line conclusion regarding the use of Models A, B and C is that we would use the model that seems most appropriate at the outset, whereas in the past we might have worked harder to expand the model to achieve consistency, which now seems somewhat moot. So, if we only had data from a single population and if the tests could be regarded a priori as conditionally independent, we would use Model A. We thus would *now* probably not go out of our way to find a variable to use to create two populations. If an obvious one were available, we would probably not shy away from using it, even if we thought the prevalences might be within 0.1 of each other. We would either use Model B or C, depending on whether or not the tests could be regarded as conditionally independent based on biological considerations, or we would modify the model to allow for point mass at $\Delta = 0$. We would work hard to obtain independent information that could be formulated as the prior.

3. TWO SIMPLE ILLUSTRATIONS OF THE CONDITIONAL DISTRIBUTION OF $\phi_N|\phi_I$

A very interesting aspect of the article is the discussion of the parameterization (ϕ_I, ϕ_N) and the simple characterization of the posterior for $\phi_N|\phi_I$. We believe that the study of this distribution in specific problems should warrant investigation as suggested by Gustafson. In this section we highlight two very simple illustrations where no computation is necessary to obtain this distribution, and in Section 4 we observe the potential simplicity of exploring this distribution for selected values of ϕ_I by using WinBUGS.

First consider the classic model with $x \sim \text{Bin}(n, \phi_I)$ and where x is the observed data. Then assume that y is not observed but that $y|x \sim \text{Bin}(x, \phi_N/\phi_I)$, $\phi_I \sim \text{Beta}(c, d)$ and $\phi_N/\phi_I \sim \text{Beta}(a, b)$, where $a + b = c$. This corresponds to the missing data problem where

$(y, x-y, n-x) \sim \text{Mult}(n, \{\phi_N, \phi_I - \phi_N, 1-\phi_I\})$ and $\{\phi_N, \phi_I - \phi_N, 1-\phi_I\} \sim \text{Dirichlet}(a, b, d)$. Then it follows that the conditional distribution of $\phi_N|\phi_I$ is simply $\phi_I * \text{Beta}(a, b)$. So if $a = b = 1$, the conditional posterior is $\text{U}(0, \phi_I)$.

Johnson and Gastwirth (1991) discussed Bayesian inference for prevalence and test accuracy based on use of a single screening test for HIV, and Johnson, Gastwirth and Pearson (2001) discussed identifiability in the context of Models A and B. These papers all discussed (asymptotic) approximations to posteriors for prevalence and test accuracy under low prevalence and high accuracy assumptions. It is particularly interesting to note that under these assumptions, it is shown that the posterior for the two sensitivities in the Hui and Walter (1980) model is simply the prior, while there is clearly direct information in the data for the two specificities and the prevalence. Thus (asymptotically) we have $\phi_N = (p_1, p_2)$ and $\phi_I = (r, q_1, q_2)$.

4. COMPUTATION IN MODELS A, B AND C

We have rarely had difficulty with convergence of our algorithms. We believe that the parametrization developed in Georgiadis, Johnson, Gardner and Singh (2003) is particularly useful for this purpose since it leads to well-known full conditional distributions (in conjunction with modeling the latent disease status). Moreover, we also found that implementing the parametrization developed by Dendukuri and Joseph (2001) in WinBUGS (without accounting for the latent data) also converged very nicely if slightly slower than the parameterization in Georgiadis et al. In subsequent work (Branscum, Gardner and Johnson, 2005), we implement the latter approach and have advertised it broadly for use in the veterinary community due to its simplicity and directness. We expect that part of our lack of difficulty with convergence is due to having reasonable prior information and also due to small–moderate sample sizes. On the other hand, when we used so-called flat priors for regression coefficients in a logistic regression model with (diagnostic test) error in the response (McInturff, Johnson, Cowling and Gardner, 2004), we had horrendous difficulties with convergence, while using a mildly informative prior of the form used by Bedrick, Christensen and Johnson (1997) rendered convergence a nonissue.

It is possible to study the conditional distribution of $\phi_N|\phi_I$, as discussed by Gustafson, in WinBUGS. When say $x/n \to \phi_I$, simply model x with a very large n, or use a standard normal approximation to x/n, and input x/n to be a value ϕ_I of interest. The posterior for ϕ_N is then approximately (4). For example, with a standard one-test one-sample screening problem and with $x \sim \text{Bin}(n, \phi_I)$ with $\phi_I = rp + (1-r)(1-q)$, so that $\phi_N = (r, q)$, we have (approximately) $y \equiv x/n \sim N(\phi_I, \phi_I(1-\phi_I)/n)$. Sample WinBUGS code is

$$\begin{aligned}\text{model}\{&\mu \leftarrow n * \phi_I, \tau \leftarrow 1/(\mu * (1-\phi_I))\\ &y \sim \text{dnorm}(\mu, \tau)\\ &\phi_I \leftarrow r * p + (1-r) * (1-q)\\ &p \sim \text{dbeta}(18, 4)I(\phi_I,), \bar{\phi}_I \leftarrow 1 - \phi_I\\ &q \sim \text{dbeta}(18, 4)I(\bar{\phi}_I,), \phi_I \sim \text{dbeta}(a, b)\},\\ \text{list}(&x = 50000, n = 100000),\end{aligned}$$

where the indicator notation truncates the corresponding distribution, for example, $I(\phi_I)$ denotes that the corresponding distribution is truncated below at ϕ_I. With these data we have (for practical purposes) set $\phi_I = x/n = 0.5$. Adding a line of code to specify an independent beta distribution for the prior, it is possible to compare the conditional and marginal distributions for ϕ_N. Any parameter $g(\phi_I, \phi_N)$ can be specified with an additional line of code. Then one can simply take different data sets to assess the effect of the "data" on inferences for ϕ_N. The prior on g is also trivial to induce and compare with the posterior.

5. IDENTIFIABILITY IN MODEL C WITH MULTIPLE POPULATIONS AND ITS HIERARCHICAL EXTENSION

We thought it might be interesting to present a geometric argument that makes clear why Model C lacks identifiability, even if it is extended to allow for say k sampled populations. With two populations, the issue is obvious as pointed out by Gustafson. However, with three populations, there are degrees of freedom and nine parameters, so it might be tempting to think that identifiability could be bought, notwithstanding Gustafson's arguments against expanding, by expanding the model and the data in this way.

We define η to be the vector of four probabilities of two-test outcomes conditional on being "diseased," and define θ to be the corresponding vector of probabilities conditional on being "not diseased." The components of each must sum to 1. Note that there are bijections between η and (p_1, p_2, δ_1), and between θ and (q_1, q_2, δ_0). The prevalence in population i is r_i for $i = 1, \ldots, k$. Let $r = (r_1, \ldots, r_k)$. The data consist of k four-vectors of observations, say $x_i, i = 1, \ldots, k$, corresponding to the two-test outcomes in each of the sampled populations. We assume

that $x_i \sim \text{Mult}(n_i, \phi_{Ii})$ and note that

$$\phi_{Ii} = r_i\eta + (1 - r_i)\theta.$$

We have $\phi_I = \{\phi_{Ii} : i = 1, \ldots, k\}$. The ϕ_{Ii}'s are all convex combinations of η and θ, and as vectors, they all lie on a line segment (on which components sum to 1) in between η and θ (draw a picture of vectors in two dimensions). The question is, "Is there a unique solution in (η, θ, r) to these k equations?" In fact, there are an infinite number of solutions, so the model lacks identifiability.

Let ϕ_{I1} and ϕ_{Ik} be the "outer" two vectors. Then all ϕ_{Ii}, $1 < i < k$, lie on the line that connects ϕ_{I1} and ϕ_{Ik}. Let l_1 and l_2 denote the line segments:

$$l_1 = \{y \in (0, 1)^4 : y = x\phi_{I1} + (1 - x)\phi_{Ik}, x < 0\},$$

$$l_2 = \{y \in (0, 1)^4 : y = x\phi_{I1} + (1 - x)\phi_{Ik}, x > 1\}.$$

Then *any* combination of $\eta \in l_1$ and $\theta \in l_2$ (*or* $\eta \in l_2$ and $\theta \in l_1$) can generate the vectors $\phi_{I1}, \ldots, \phi_{Ik}$. Corresponding to each pair (η, θ) is a unique k vector r such that the above system is satisfied. There is no unique solution here.

If the prevalences are known, this model can be used to estimate η and θ with only *two* populations. Otherwise, if good prior information is available for the prevalences and the priors are fairly "tight" around the r_i's, posterior inferences should be fine; *however*, posterior means will not be consistent.

It also may be the case that two of the cell-probability maximum likelihood estimators MLEs $\hat{\phi}_{Ii} = x_i/n_i$ and $\hat{\phi}_{Ij} = x_j/n_j$ are at opposite ends near the boundary of the parameter space $(0, 1)^4$. This effectively "seals off" the possible maximal likely values for η and θ, and inferences resulting from the Bayesian approach should be dead on.

Finally, consider an alternative model where everything is as above only now we also have $r_i|\mu, \gamma \overset{\text{i.i.d.}}{\sim} \text{Beta}(\mu\gamma, (1 - \mu)\gamma)$. Hanson, Johnson and Gardner (2003) developed Bayesian methodology for this model and also established that there exist (non-Bayesian) consistent estimators of (η, θ, r), where the limit is taken as $n_i \to \infty$, $k \to \infty$. This result is not too surprising, since as k grows, under the assumption on the r_i's, we should be ultimately sampling populations with small and large prevalences. If the Beta distribution is overly concentrated away from 0 or 1, then it may be a very long time before that would happen however, so it is possible that extraordinarily large sample sizes would be necessary.

6. CONCLUDING REMARKS

We thank Professor Gustafson for writing a very stimulating paper. We hope it is clear that it was not our intent to criticize, but to complement his work. We believe that his elaboration on these issues was clear and insightful, and will provide a basis for evaluating models that lie on the boundary between identifiable and not, and for obtaining insight into how much information there is in the data for ϕ_N. We also look forward to his future insights on these and other topics.

Comment

Lawrence Joseph

1. INTRODUCTION

Paul Gustafson has written a very provocative paper with results that may surprise some statisticians. The main conclusion, demonstrated through two illustrative examples, is that under certain conditions, non-identifiable models can sometimes outperform identifiable models. In this commentary I will focus on some practical issues not given much attention by Gustafson.

Lawrence Joseph is Associate Professor, Division of Clinical Epidemiology, Montreal General Hospital, Department of Medicine, Montreal, Quebec, Canada H3G 1A4, and Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada H3A 1A2 (e-mail: Lawrence.Joseph@mcgill.ca).

2. BAYESIAN VERSUS FREQUENTIST ANALYSES OF DIAGNOSTIC TEST DATA: PRIOR DISTRIBUTIONS VERSUS POINT ESTIMATES AS INPUTS

For identifiable models such as Model B in Section 2 of Gustafson, both maximum likelihood and Bayesian estimation with a suitable choice of "flat" prior will produce similar inferences, at least numerically. For

nonidentifiable models such as Models A or C, however, Bayesian and frequentist inferences operate quite differently.

Frequentist methods do not include prior distributions and, therefore, do not have the luxury of choosing whether or not to use a nonidentified model: the model must be made identifiable in order to estimate the unknown parameters. The added flexibility in modeling is a major advantage for Bayesians, as Gustafson has shown that a Bayesian analysis of nonidentifiable models using a "crude" prior can have lower mean squared error compared to related expanded or contracted models that are identifiable. In practice, what are the frequentist options in such situations and how should a Bayesian handle the problem of nonunique "crude" prior distributions?

Consider the case of Model A. Gustafson expands to Model B to produce identifiability, but this is not the only option for a frequentist. With five unknown parameters but only 3 degrees of freedom, frequentists can still use Model A, inserting point estimates for any two of the five unknown parameters (two sensitivities, two specificities and the prevalence) in order to estimate the other three parameters via maximum likelihood (Walter and Irwig, 1988). As the two parameters given fixed values are almost never known exactly, the analyst would typically do a sensitivity analysis to this choice and, thus, the end result is a table that displays different sets of estimates, one for each choice of "fixed" parameter values. Overall conclusions are difficult to derive from this collection of values, however, and total uncertainty is underestimated, since confidence intervals from each line of the table omit the inherent uncertainty in the fixed parameter values (Joseph, Gyorkos and Coupal, 1995).

Bayesians do not need to fix these parameter values, so overall conclusions are available which include all inherent uncertainty. The situation is not as ideal as it seems, however, because there is the ever-present problem of choice of prior distribution. While Gustafson claims the models work well with crude priors, how crude are they really? Section 3.2 suggests a Beta$(18, 4)$ prior is crude, even though it is quite peaked near 0.85. Starting from the entire $[0, 1]$ parameter space, the 95% highest density interval from a Beta$(18, 4)$ distribution is $(0.66, 0.96)$, which has length of 0.3. The Beta$(10, 1)$ density also used by Gustafson has a 95% highest density interval of even smaller length, at $(0.74, 1.0)$. "Crudeness" is surely a relative term, and sensitivity to the prior distribution is important in any analysis using a nonidentifiable model.

Thus, in real practice we see that the Bayesian analysis has similarities to the frequentist analysis, in that sensitivity to prior inputs (whether point estimates or prior distributions) is important. In either case a table must be created that gives inferences across a range of prior inputs, rendering overall conclusions difficult.

In the context of clinical trials, Spiegelhalter, Freedman and Parmar (1994) have suggested creating a family of prior distributions, ranging from optimistic to pessimistic or, in this context, from highest to lowest "reasonable" values for the prevalence, sensitivities and specificities of the tests given the available prior information. While a "strict" interpretation of Bayesian analysis mandates a single choice of prior distribution, if results are to be reported and accepted by a wide audience, a range of prior distributions is usually needed. This, however, raises many questions: If a nonidentifiable model is to be used, what range of prior distributions should be input? How should overall conclusions be derived given this range of reasonable posterior inferences? One ad-hoc suggestion might be to average over all results using some weighting scheme on the various choice of priors, and another might be to take the minimum lower and maximum upper highest density interval limits to derive a conservative or "all inclusive" interval, which should contain all reasonable values. While this problem occurs in all Bayesian analyses, it is especially pertinent in the context of nonidentifiable models, where the importance of prior information does not decrease as the sample size increases and often posterior $\approx$ prior for a subset of parameters.

3. STUDY DESIGN

As Gustafson points out in Section 2.1, the choice between using Models A or B is made easier by prior knowledge about the prevalences of the subpopulations. In general, if we are to analyze studies using nonidentifiable models, we need to plan accordingly at the design stage. Given prior distributions on the subpopulation prevalences, an interesting optimal design problem arises about whether the extra data required for Model B should be collected. Further, if Model A is chosen, how should a sample size be selected? Standard sample size methods do not apply, and in nonidentifiable models some marginal posterior distributions do not converge to a single point even with an infinite sample size, so that a desired accuracy may

never be reached (Rahme, Joseph and Gyorkos, 2000; Dendukuri, Rahme, Bélisle and Joseph, 2004). Further design questions arise if results are to be summarized across a range of prior distributions and one wishes to claim a certain accuracy will be attained at the end of the study. Clearly, further work is required here.

4. CONCLUDING COMMENTS

As is often the case with the best scientific work, this article opens many avenues for further research. There is much to be done if statisticians are to convince mainstream users to apply nonidentified models in their analyses. The choice of prior distribution remains a hurdle, even if almost any reasonable choice provides better inferences compared to a contracted or expanded identifiable model. Other than the two commonly occurring situations discussed in detail by Gustafson, there are probably many examples where nonidentifiable models with some prior information perform better than related identified models. Clearly, further experience with these models is needed before we can be confident that we are producing solid inferences and before routine users will incorporate them into their daily practice.

Comment

Jaeyong Lee

I congratulate Paul Gustafson on carefully laying out issues involved in Bayesian modeling with identifiability. I think this paper is one of the few that has investigated the identifiability issue in Bayesian analysis seriously.

In Bayesian statistical modeling, there seem to exist two slightly different views on identifiability. One is the position, as the author describes, "identifiability good, nonidentifiability bad," in which identifiability is considered as one of the minimum requirements. This view seems to be based on the fact that one cannot accurately pin down the actual parameter value from which the data were generated even with infinitely many observations. The other position is a rather casual attitude toward identifiability because, in posterior analysis of nonidentifiable statistical models, the posterior can be computed without difficulty and the meaning of posterior does not change, even with nonidentifiability.

Paul Gustafson's paper gives careful thought to this issue of identifiability in Bayesian modeling using two examples. In particular, using these examples, the author raises two interesting and potentially controversial points:

POINT 1. As opposed to conventional wisdom (i.e., to gain identifiability, the model needs to be simplified somehow), it is possible to gain identifiability by expanding the model and to lose identifiability by contracting it.

POINT 2. As opposed to conventional wisdom (i.e., "identifiability good, nonidentifiability bad"), an identifiable model can perform dramatically worse than a nonidentifiable model.

Below I discuss the author's points individually.

POINT 1

The author argues that identifiability can be obtained by expanding an unidentifiable model, rendering Scenario I, as an example, where simpler Model A is not identifiable while more complex Model B is identifiable. The author argues that this contradicts conventional wisdom. The conventional wisdom, as the author calls it, comes from the following simple fact. Suppose observable y follows a density f_0 with unknown parameter θ and suppose model M_i postulates $\theta \in \Theta_i$ for $i = 1, 2$ and $\Theta_1 \subset \Theta_2$. When model M_1 is not identifiable (i.e., there exist $\theta_1 \neq \theta_2$ in Θ_1 with $f_{\theta_1} = f_{\theta_2}$), expanding the parameter space to Θ_2 does not help to gain identifiability. On the other hand, if model M_1 is identifiable, expansion of the model to M_2 may result in nonidentifiability. The author's example is clever in that it does not exactly fit this situation, because by adding an additional variable X, the observable (T_1, T_2) changes from one 2×2 table to two 2×2 tables.

Jaeyong Lee is Assistant Professor, Department of Statistics, Seoul University, Sillimdong Kwanakgu, Seoul, 151-742, Republic of Korea (e-mail: leej@stat.psu.edu).

The situation in Scenario II is slightly different because the parameter space of Model E is the prior probability 0 subset of the expanded Model F, and Model F gains "essential identifiability."

POINT 2

In Scenario I the author considers estimation of $r = \Pr(E = 1)$ in nonidentifiable Model A and expanded identifiable Model B with comparable priors for each model which the author calls prior (i). For Model B the author adopts the standard asymptotic theory and shows the range of $\mathrm{ARMSE}(\hat{r_B})$ is from 0.05 to ∞, with ∞ occurring at $\Delta = 0$, which corresponds to nonidentifiable Model A (see the left side of the right panel in Figure 1). For Model A the author sets up a framework to evaluate the performance of the Bayes estimator for nonidentifiable models and shows that $\mathrm{ARMSE}(\hat{r_A}) = 0.015$. These numbers are remarkable, because by taking nonidentifiable Model A over identifiable Model B, one can achieve almost four times better results even than the best situation of Model B. Based on this result, the author claims that "...either contracting or expanding the model for the sake of identifiability can in fact lead to poorer estimator performance...."

A closer look at the analysis casts doubt on this claim. First of all, while Model A is exactly the same as Model B with $\Delta = 0$, how can two ARMSEs be so different? The ARMSE of r is ∞ under Model B with $\Delta = 0$ and is 0.015 under Model B. This is because the calculation of ARMSEs was not fair to these models. In ARMSE calculation for Model A, the effect of the prior stays in asymptotic calculation, while ARMSE calculation of Model B does not involve the prior effect. In fact, if we adopt the standard ARMSE calculation, which is used for Model B, for Model A the ARMSE will be ∞, because the corresponding diagonal element of the Fisher information will be 0. On the other hand, had more accurate higher-order asymptotic calculation been adopted for the ARMSE calculation of Model B, I expect that both ARMSEs would be similar. In fact, the author's simulation results in Table 1 support my point. In Table 1 RMSEs of Model A with prior (i) and Model B with prior (i) are almost same in both DGM (i) and (ii). The small difference may be due to the difference in priors. Prior (i) for Model A uses $r \sim \mathrm{U}(0, 1)$, while prior (i) for Model B uses $r_1 \sim \mathrm{U}(0, 1)$ and $r_2 \sim \mathrm{U}(0, 1)$, resulting in a triangular prior distribution for $r = (r_1 + r_2)/2$ which is more concentrated at $1/2$; note that the true value is 0.3.

I hope the results in this paper, as the author also emphasizes, sound a cautionary note on the issue of identifiability. When an identifiability problem arises, simply making the model identifiable by either contraction or expansion does not make the problem go away. A lesson from this paper is that identifiability is important. Practically nonidentifiable and theoretically identifiable models (which the author calls danger zone) are as dangerous as nonidentifiable models. A similar example arises in a selection model setting (Lee and Berger, 2001), where the model is theoretically identifiable but the posterior does not distinguish two very different scenarios; however, reasonable prior information can help us to draw a rather sound informative posterior. Another lesson from the paper with which I agree with the author is that we should not retreat from nonidentifiable models, because by carefully eliciting our prior, we can get an informative posterior—of course, careful elicitation should be underlined.

Rejoinder

Paul Gustafson

I thank all the discussants for making insightful remarks and for sharing some of their own related findings. I am pleased to detect considerable commonality in how these modeling issues are perceived. We all seem to align on the point that identifiability can be a rather nuanced issue. None of us argues that an identified model necessarily excuses one from formulating and using prior information, nor do any of us subscribe to the view that nonidentified models are necessarily useless and that identifiability *must* be "bought" at any price. On more specific issues raised, I will respond to the discussants in turn.

RESPONSE TO PROFESSORS GELFAND AND SAHU

The ideas in Gelfand and Sahu (1999) are very interesting. I am particularly intrigued by one of the con-

ditions for a unique proper posterior arising from an improper prior, namely that the parameter space for ϕ_I and ϕ_N be a product space. A quirk of many situations involving mismeasured variables is that the support of ϕ_N depends on ϕ_I, and this can be the primary source of "indirect learning" when it occurs.

Gelfand and Sahu asked about the marginal prior distribution induced on ϕ_I, noting that this could be ascertained by simulation. To give an example of this, Figure R1 displays this distribution for the first crude prior used with Model A. There is considerable structure in this prior distribution. Note that sensibly the prior distribution for the probabilities of discordant observations, $\phi_{I,2} = \Pr(T_1 = 1, T_2 = 0)$ and $\phi_{I,3} = \Pr(T_1 = 0, T_2 = 1)$, is centered at quite a low value, whereas the prior distribution for $\phi_{I,1} = \Pr(T_1 = 1, T_2 = 1)$ is rather diffuse. [By symmetry, the same prior will describe $\Pr(T_1 = 0, T_2 = 0)$.]

It was presumptuous of me to say that in the model averaging scenario there is no reason to expect the posterior probability on the correct model to tend to 1 as the sample size grows. More accurately, I have no idea about the extent to which such results might still hold when one of the models lacks identifiability, and I am not aware of literature that addresses this question. This seems like an obvious area for more research.

I agree that the identifiability issues with Models D, E and F would be much different under a Berkson measurement error model. I am puzzled though by the suggestion that the Berkson formulation would be more natural. I would think that generally the choice between the two formulations must be driven by the subject-area context and the data-gathering mechanism. The Berkson model is appropriate when modeling X given X^* makes sense, typically in a controlled experiment. For instance, say the experimenter sets an imperfect thermostat to temperature X^*, but the impact of this is to actually achieve temperature X. Conversely, the non-Berkson model is usually appropriate in an observational context. For instance, say the actual temperature is X, but the observer's imperfect thermometer records X^*. While the distinction seems

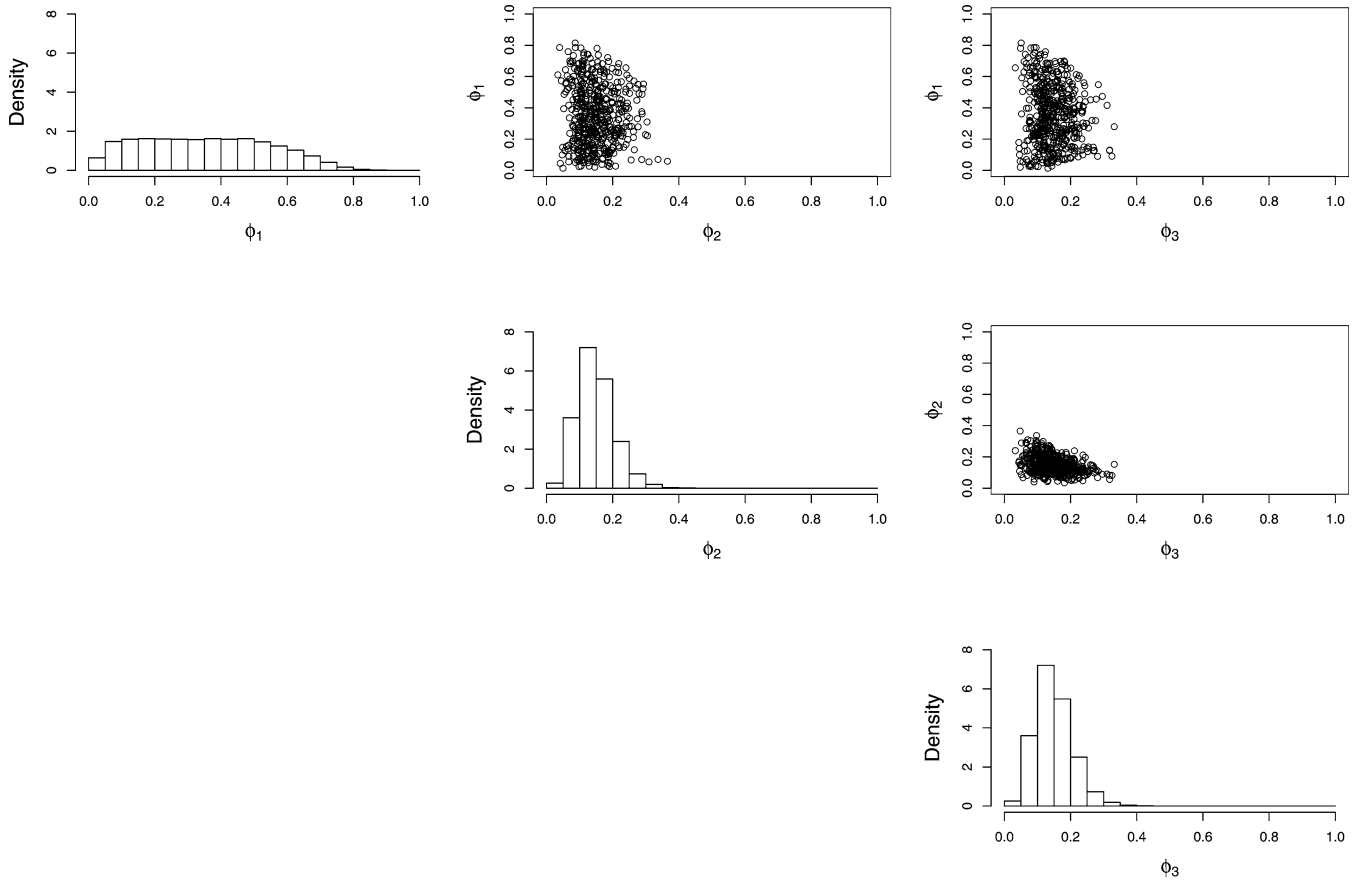

FIG. R1. *Prior distribution on ϕ_I for the first crude prior under Model* A. *Recall that* $\phi_{I,1} = \Pr(T_1 = 1, T_2 = 1)$, $\phi_{I,2} = \Pr(T_1 = 1, T_2 = 0)$, $\phi_{I,3} = \Pr(T_1 = 0, T_2 = 1)$.

slight, its implications can be considerable. In particular, an analysis which ignores the measurement error can be much more misleading in the non-Berkson scenario.

RESPONSE TO PROFESSORS JOHNSON AND HANSON

If the goal is to compare Models A and B with the same amount of prior information infused in both, then I agree that the simulations in Section 2.4 are much fairer comparisons than the asymptotic comparisons in Section 2.3 However, the asymptotics do establish that (i) Model A plus somewhat crude prior information can lead to reasonable inferences, and they suggest that (ii) Model B may not perform well without prior information and (iii) the asymptotics may not kick in until prohibitive sample sizes. These points all push toward a preference for a principled Bayesian analysis without particular emphasis on whether the model is identifiable or not.

The discussants' remarks about computational issues and the WinBUGS example are encouraging. I had come to the view that MCMC for nonidentified models is generally hard and that special-purpose samplers are often needed, but Professor Johnson and Professor Hanson have a lot of experience with MCMC fitting of models like these and they indicate that the situation is not as dire as I imagined.

I found the discussion of nonidentifiability in Model C extended to more than two populations to be particularly insightful. I had wondered about this question, to the point of trying to fit such a model in the three-population case, without much success (Gustafson, 2003). There I had waffled on the question of identifiability, simply noting that the number of parameters was consistent with, but did not prove, identifiability. Johnson and Hanson have elegantly resolved the question, providing a good example of why it does not suffice just to count parameters when assessing identifiability of a model.

RESPONSE TO PROFESSOR JOSEPH

I admit to hiding behind "illustrative" analyses, deferring the real and difficult question of which prior to actually use in a real problem to a subject-area specialist. I certainly agree that it is wise to try a few different prior distributions, or perhaps even undertake a more formal assessment of prior influence. In the latter vein, Gustafson and Clarke (2004) considered a partitioning of posterior variance to assess prior influence, using a context similar to Model A as one of their examples. In a more direct and practical vein, adaptation of the Spiegelhalter, Freedman and Parmar three-prior approach makes eminent sense.

If multiple priors are identified but an overall result is required, I would argue in favor of averaging the priors and letting the posterior fall where it may, rather than combining the multiple posteriors in an ad hoc way. I say this particularly because the unpredictable nature of indirect learning in these sorts of models could imply a nontrivial change in the weighting of the constituent posterior distributions relative to the weighting of the constituent prior distributions.

I certainly agree that design questions are difficult in these sorts of models, and as far as I know Professor Joseph and his colleagues are the only ones who have been brave enough to tackle this head on. Dendukuri, Rahme, Bélisle and Joseph (2004) should be eye-opening for many readers, with its finding that mismeasurement may imply that no finite sample will be adequate under plausible design criteria.

Another aspect of the design question which has interested me recently is as follows. The two-term decomposition of ARMSE as in (5) indicates diminishing returns once the sample size suffices to make the second (variance) term small relative to the first (bias) term. Under some formulations of the design problem this may call for relatively small sample sizes to reach this point, with resources then conserved for use elsewhere. This fits with one of the messages in Greenland (2005) on what he terms "bias-modelling" in epidemiological settings. Larger and/or more observational studies on a particular exposure–disease relationship will be of very limited value if there is already enough data to control random variability relative to biases arising from sources such as mismeasurement, selection bias and unobserved confounding.

RESPONSE TO PROFESSOR LEE

I am not so convinced that the two examples differ fundamentally in how essential identifiability is gained from model expansion. In some sense a data expansion accompanies the model expansion in both cases, or, put in reverse, both model contractions involve data contractions. If we start with Model B and set $r_1 = r_2$, we obtain Model A and render the distinction between the two separate (T_1, T_2) tables irrelevant. If we start with Model F and set $\beta_2 = 0$, we obtain Model E and, roughly speaking, make the $(X^*)^2$ column of the design matrix irrelevant in the model for $Y|X^*$. While this

is imprecise since the model for $Y|X^*$ lacks a closed form, it seems clear that in both cases the model contraction makes part of the data structure redundant.

While the ARMSE comparisons are fair in the technical sense of being $o(n^{-1})$ approximations to the MSE under either model, I agree they are not fair in the practical sense that it seems a larger sample size may be needed for the approximation to kick in with Model B, particularly when the parameter values fall in the danger zone. The fairer comparisons are probably those in the simulation study. As mentioned in the response to Professor Johnson and Professor Hanson though, I think the asymptotics are informative in a number of ways.

As for the small difference in the simulation between the two models when the same prior is used, perhaps this is induced by the slightly different prior on r. However, there are other factors potentially at play as well, such as the weak information conveyed by X versus the usual increase in estimator variance associated with using a bigger model.

ADDITIONAL REFERENCES

BEDRICK, E. J., CHRISTENSEN, R. and JOHNSON, W. O. (1997). Bayesian binomial regression: Predicting survival at a trauma center. *Amer. Statist.* **51** 211–218.

BRANSCUM, A. J., GARDNER, I. A. and JOHNSON, W. O. (2005). Estimation of diagnostic test sensitivity and specificity through Bayesian modeling. *Preventive Veterinary Medicine* **68** 145–163.

DENDUKURI, N., RAHME, E., BÉLISLE, P. and JOSEPH, L. (2004). Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics* **60** 388–397.

GREENLAND, S. (2005). Multiple-bias modelling for analysis of observational data (with discussion). *J. Roy. Statist. Soc. Ser. A* **168** 267–306.

GUSTAFSON, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology*: *Impacts and Bayesian Adjustments*. Chapman and Hall/CRC, Boca Raton, FL.

GUSTAFSON, P. and CLARKE, B. (2004). Decomposing posterior variance. *J. Statist. Plann. Inference* **119** 311–327.

HANSON, T. E., JOHNSON, W. O. and GARDNER, I. A. (2003). Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold-standard. *J. Agric. Biol. Environ. Stat.* **8** 223–239.

LEE, J. and BERGER, J. O. (2001). Semiparametric Bayesian analysis of selection models. *J. Amer. Statist. Assoc.* **96** 1397–1409.

MCINTURFF, P., JOHNSON, W. O., COWLING, D. W. and GARDNER, I. A. (2004). Modeling risk when binary outcomes are subject to error. *Statistics in Medicine* **23** 1095–1109.

RAHME, E., JOSEPH, L. and GYORKOS, T. (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Appl. Statist.* **49** 119–128.

SAHU, S. K. and CHENG, R. C. H. (2003). A fast distance-based approach for determining the number of components in mixtures. *Canad. J. Statist.* **31** 3–22.

SPIEGELHALTER, D., FREEDMAN, L. and PARMAR, M. (1994). Bayesian approaches to randomized trials (with discussion). *J. Roy. Statist. Soc. Ser. A* **157** 357–416.

SU, C. L. and JOHNSON, W. O. (2005). Large sample joint posterior approximations when the full conditionals are approximately normal: Applications to generalized mixed models. Submitted for publication.

WALTER, S. D. and IRWIG, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *J. Clinical Epidemiology* **41** 923–937.

YEE, J., JOHNSON, W. O. and SAMANIEGO, F. J. (2002). Asymptotic approximations to posterior distributions via conditional moment equations. *Biometrika* **89** 755–767.