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

Paul Gustafson

Department of Statistics, University of British Columbia
Vancouver, B.C. Canada V6T 1Z2
*gustaf@stat.ubc.ca*

May 27, 2002

**Abstract**

When a candidate model for data is nonidentifiable, conventional wisdom dictates that the model must be simplified somehow, in order 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 an identifiable model. 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, we compare the properties of estimators obtained from identifiable models to those of estimators obtained from nonidentifiable models in tandem with crude prior information. Both asymptotic theory and simulations with MCMC-based estimators are used to draw comparisons. A technical point which arises is that the asymptotic behaviour 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.

**Keywords:** 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 modelling problem at hand, but this model is not identifiable. That is, multiple values of the parameter vector lead to the same distribution for the observable data. Conventional wisdom dictates that the investigator must select a simpler sub-model with fewer than $p$ parameters that is identifiable, though of course this process of *model contraction* may lead to a model involving

dubious assumptions, or a model which is less realistic in some other way. A less intuitive approach to gaining identifiability is to make the initial model larger. As we will see, however, there are natural situations where the initial model has an identifiable super-model with more than $p$ unknown parameters.

To be more specific, this article considers two inferential scenarios involving mismeasured variables. The first scenario involves two imperfect schemes for assessing whether or not a study subject is 'exposed', where the error probabilities describing these schemes and the prevalence of exposure in the study population are unknown. We consider three plausible models in this context, with nonidentifiable Model A nested within identifiable Model B nested within non-identifiable 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 we consider a nested sequence of plausible models, with identifiable Model D 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 focus on 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 model contraction or model expansion.

In both scenarios we compare properties of estimators based on the different models, under a variety of actual data-generating mechanisms. The motivation for doing so 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.

Our comparisons 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 can shed no light on the properties of estimators based on nonidentifiable models, so that simulation is the only possibility for studying the properties of estimators based on Models A, C, and E. On the contrary, we show that standard asymptotic theory can describe the behaviour of posterior means arising from nonidentifiable models. In particular, the requisite trick is to use iterated expectation to re-express the posterior mean in terms of the identifiable part of the model alone.

# 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. For instance, Hui and Walter (1980) give an example involving two tests (the Mantoux test and the Tine test) for the detection of tuberculosis. Drews, Flanders and Kosinski (1993) consider 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) consider a study in which both a serology test and stool examination are used to test for a particular parasitic infection. In addition to the assessment schemes being imperfect, often the classification probabilities which 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 (or 'tests') for $E$. We consider the realistic scenario in which the *sensitivity* $p_i = Pr(T_i = 1|E = 1)$ and *specificity* $q_i = Pr(T_i = 0|E = 0)$ of each test are not known. If we observe $(T_1, T_2)$ for subjects sampled from the population of interest having unknown exposure prevalence $r$, then Model A postulates that

$$\begin{aligned} f_A(t_1, t_2|\theta) &= Pr_A(T_1 = t_1, T_2 = t_2|\theta) \\ &= rp_1^{t_1}(1-p_1)^{1-t_1}p_2^{t_2}(1-p_2)^{1-t_2} + \\ &\quad (1-r)q_1^{1-t_1}(1-q_1)^{t_1}q_2^{1-t_2}(1-q_2)^{t_2}, \end{aligned} \tag{1}$$

where $\theta = (p_1, p_2, q_1, q_2, r)$ is the unknown parameter vector. In particular, Model A invokes the common assumption that the two tests outcomes are independent given the true exposure status. Clearly Model A is nonidentifiable, as the data comprise a $2 \times 2$ table from which at most three parameters can be estimated consistently, whereas in fact five parameters are unknown.

Starting with Model A, one way to develop an identifiable model is to pre or post-stratify the population of interest according to some binary trait $X$ which is thought to be associated with exposure status $E$. For instance, say random samples of size $n_1$ and $n_2$ are taken from the $X = 0$ and $X = 1$ sub-populations respectively. Model B postulates that (1) holds with prevalence $r = r_1$ in the first sub-population, and with prevalence $r = r_2$ in the second sub-population. As well, Model B implicitly assumes the 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 sub-populations. 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 $2 \times 2$ tables for each sub-population, so there is hope of consistently estimating six parameters. Indeed, Hui and Walter (1980) illustrate that subject to some minor caveats Model B is a 'regular' model leading to estimators with standard asymptotic properties.

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. Moreover, in the absence of observations on $E$, the assumption cannot be checked empirically. All the 'degrees-of-freedom' are used up in the estimation of the six-dimensional parameter vector $\theta$. The plausibility of the conditional independence assumption and the effects of incorrectly invoking it are discussed by Fryback (1978), Vacek (1985), Brenner (1996), and Torrance-Rynard and Walter (1997).

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

$$Pr(T_1 = a, T_2 = b|E = 0) = (1-q_1)^a q_1^{1-a}(1-q_2)^b q_2^{1-b} + (-1)^{|a-b|}\delta_0,$$

and similarly,

$$Pr(T_1 = a, T_2 = b|E = 1) \;\; = \;\; p_1^a(1-p_1)^{1-a}p_2^b(1-p_2)^{1-b} + (-1)^{|a-b|}\delta_1.$$

Under this model $p_i$ and $q_i$ retain their interpretations as the sensitivity and specificity of the $i$-th test, but now $\delta_0 = Cov(T_1, T_2|E = 0)$ and $\delta_1 = Cov(T_1, T_2|E = 1)$ are additional unknown parameters. In the special case that $\delta_0 = 0$ and $\delta_1 = 0$ we recover Model B. As scenarios under which $T_1$ and $T_2$ are negatively associated given $E$ are hard to imagine, we restrict to $\delta_0 \in [0, \delta_{MAX}(q_1, q_2)]$ and $\delta_1 \in [0, \delta_{MAX}(p_1, p_2)]$, where $\delta_{MAX}(a, b) = \min\{a, b\} - ab$ is the maximal covariance between two binary random variables with 'success' probabilities $a$ and $b$.

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

$$\begin{aligned} \rho_0 &= Cor(T_1, T_2|E = 0) \\ &= \frac{\delta_0}{\{q_1(1-q_1)q_2(1-q_2)\}^{1/2}}, \end{aligned}$$

and

$$\begin{aligned} \rho_1 &= Cor(T_1, T_2|E = 1) \\ &= \frac{\delta_1}{\{p_1(1-p_1)p_2(1-p_2)\}^{1/2}}, \end{aligned}$$

with the range of dependence now expressed as $\rho_0 \in [0, \rho_{MAX}(q_1, q_2)]$ and $\rho_1 \in [0, \rho_{MAX}(p_1, p_2)]$.

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 $2 \times 2$ 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 behaviour 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 $p_1 + q_1 > 1$ and $p_2 + q_2 > 1$, to avoid the trivial nonidentifiability arising because $f_A(t_1, t_2|\theta)$ is unchanged upon replacing $p_i$ with $1 - q_i$, $q_i$ with $1 - p_i$, and $r_i$ with $1 - r_i$. In practice the restriction is very mild, as an assessment scheme that is worse than chance, i.e. $p_i + q_i < 1$, 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, for instance, Hui and Walter 1980). In situations where Model B is correctly specified, a maximum likelihood or Bayes estimate $\hat{\psi}$ of $\psi = \psi(\theta)$ is consistent, with a readily computed asymptotic variance.

Say that data are actually generated under Model B, with the parameter values $p_1 = 0.8$, $p_2 = 0.8$, $q_1 = 0.75$, $q_2 = 0.9$, $r_1 = 0.3 - \Delta/2$ and $r_2 = 0.3 + \Delta/2$. For later reference we refer to these specific values as DGM (i), where DGM stands for *data generating mechanism*. For simplicity, and without any real loss of generality, assume that $Pr(X = 1) = 0.5$ is known, so that the population exposure prevalence is $r = (r_0 + r_1)/2 = 0.3$, which is estimated by

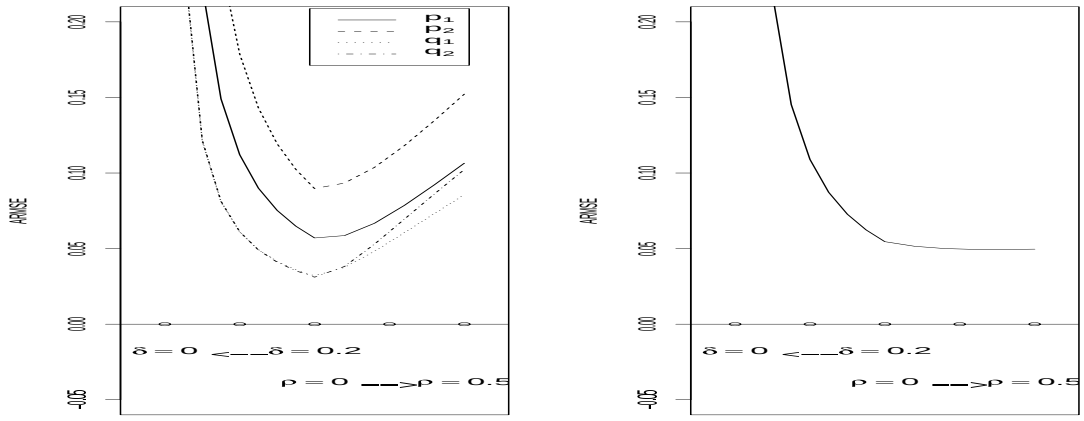


Figure 1: ARMSE for $\hat{p}_1$, $\hat{p}_2$, $\hat{q}_1$, $\hat{q}_2$ (left panel) and $\hat{r}$ (right panel) under Model B and DGM (i), with a sample size of $n = 2000$. The left side of each panel corresponds to the true parameter values in DGM (i), with varying values of $\Delta = r_2 - r_1$. The right side of each panel corresponds to to data generated under Model C, with $\Delta = 0.2$ and varying values of $\rho = Cor(T_1, T_2|E = 0) = Corr(T_1, T_2|E = 1)$.

$\hat{r} = (\hat{r}_1 + \hat{r}_2)/2$. The left halves of the panels in Figure 1 give the 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 sample size of $n = 2000$ (arising say from samples of size $n_1 = n_2 = 1000$ from each sub-population). Specifically, the ARMSE is displayed as a function of $\Delta$, the difference between the prevalences in the two sub-populations.

As $\Delta$ approaches zero each ARMSE diverges to infinity. This is not surprising, as the 'no free lunch' principle dictates that the nonidentifiability in Model A cannot be overcome by simply dichotomizing the population at random and then using Model B. Mathematically it is clear that when $r_1 = r_2$ the corresponding rows (columns) of the 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 $\Delta$ 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 $\Delta = 0.2$. In general, it appears that designing studies to be analyzed via Model B is fraught with peril. Unless there is good prior knowledge to suggest that the sub-populations will have markedly different prevalences, there is a risk of obtaining very poor estimates even with considerable sample sizes.

The right halves of the panels in Figure 1 gives the ARMSE of estimators based on Model B when the data are 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 $\nu$ instead of $\theta$, where $\nu$ is comprised of 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 $h()$ is easily evaluated.

It is also possible to evaluate $h^{-1}()$, although the expressions are are extremely cumbersome (Hui and Walter, 1980). For true parameter values $(\theta, \delta_0, \delta_1)$ under Model C, we compute $\nu^*$, the probabilities characterizing $T_1, T_2|X$ 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

$$\begin{aligned} A_{ij}(\theta) &= E_C\left\{\partial^2 \log f_B(T_1, T_2; \theta)/\partial\theta_i\partial\theta_j\right\}, \\ B_{ij}(\theta) &= E_C\left\{\partial \log f_B(T_1, T_2; \theta)/\partial\theta_i \cdot \partial \log f_B(T_1, T_2; \theta)/\partial\theta_j\right\}, \end{aligned}$$

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 $\hat{\theta}$, we can compute the ARMSE for $\hat{\psi} = g(\hat{\theta})$ at a particular sample size.

As an aside, this route to determining the asymptotic behaviour of Model B estimators when Model C is correct is not fully general. For some Model C parameter values, especially with larger values of $\delta_0$ and $\delta_1$, the $T_1, T_2|X$ probabilities $\nu^*$ can fall outside the Model B parameter space. That is, $\nu^*$ can lie outside the image under $h()$ of the Model B parameter space for $\theta$. In such a situation $\theta^*$, the large-sample limit of the Model B based estimator, cannot be determined as $h^{-1}(\nu^*)$. We have not pursued this here, but in such instances numerical methods are required to determine $\theta^*$ as the value of $\theta$ which minimizes the Kullback-Leibler divergence between the actual distribution of $(T_1, T_2)$ and the distribution postulated under Model B. A finite-sample version of this problem can arise from model misspecification and/or sampling variability. For this reason Drews, Flanders and Kosinski (1993) propose EM algorithm fitting of Model B rather than the 'closed-form' approach of Hui and Walter (1980).

Returning to Figure 1, the right halves of the panels are again based on the DGM (i) values for $(p_1, p_2, q_1, q_2, r)$ with $\Delta = 0.2$. A common value $\rho$ for both $\rho_0$ and $\rho_1$ is varied, from $\rho = 0$ to $\rho = 0.5$. We note in passing that $\rho = 0.5$ is quite close to the upper bound of $\rho_0 \leq \rho_{MAX}(q_1, q_2) = 0.577$ for the specified values of $q_1$ and $q_2$. Conversely, values up to one are possible for $\rho_1$, since $p_1 = p_2$.

The format of Figure 1 is chosen to contrast the two potential pitfalls of using Model B for inference. The centre of each panel corresponds to a good situation, in that the sup-population prevalences are quite disparate ($\Delta = 0.2$), and the assumption of conditional independence is met ($\rho = 0$). Moving to the left, the DGM approaches the nonidentifiable Sub-Model A as $\Delta$ decreases to zero. Moving to the right, Model B becomes increasingly misspecified as the conditional dependence between the two tests increases. Perhaps 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 only moderately identified 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 behaviour is similar for other values, Figure 2 gives ARMSE values for DGM (ii), defined by $p_1 = 0.95$, $p_2 = 0.9$, $q_1 = 0.65$, $q_2 = 0.85$, $r = 0.15$. The overall impression is again that Model B being correct but only moderately identified is more damaging than Model B being incorrect because of dependence between tests.

The concern about the performance of Model B under moderately small values of $\Delta =$

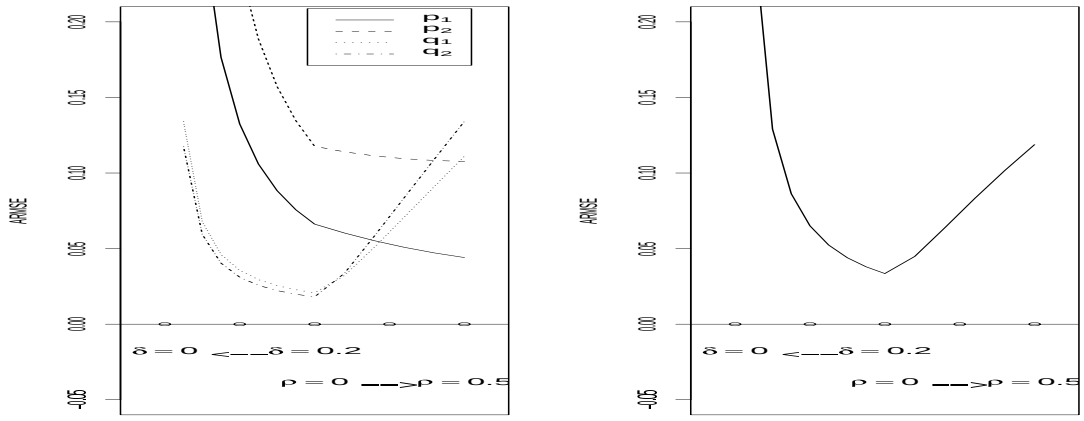


Figure 2: ARMSE for $\hat{p}_1$, $\hat{p}_2$, $\hat{q}_1$, $\hat{q}_2$ (left panel) and $\hat{r}$ (right panel) under Model B and DGM (ii). The format is the same as Figure 1.

$r_2 - r_1$ suggests that dichotomizing the population to gain identifiability is not a panacea. Thus we consider using Model A for inference, its nonidentifiability notwithstanding. Of course we cannot hope to obtain reasonable inferences from a nonidentifiable model with a diffuse prior distribution. We speculate, however, that a crude subjective prior might go some distance towards producing reasonable inferences. Before looking specifically at Model A, however, we develop an asymptotic approach to studying the performance of posterior means arising from nonidentifiable models.

## 2.2 Asymptotic Behaviour of Posterior Means in Nonidentifiable Models

The asymptotic behaviour of Bayes estimates arising from nonidentifiable models has received very little attention in the literature. Neath and Samaniego (1997), Gustafson, Le and Saskin (2001) study 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 $\theta$ 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 $\phi_I$, and not on the nonidentifiable part $\phi_N$. We call such a parameterization *transparent*, as it is intended to make apparent the impact of nonidentifiability. Of course we can commensurately transform the prior distribution $f(\theta)$ as specified in the original parameterization to $f(\phi)$ in the transparent parameterization. Indeed, following Dawid (1979), it is useful to think of the prior in terms of the marginal density $f(\phi_I)$ and the conditional density $f(\phi_N|\phi_I)$, so that immediately we have

$$f(\phi_I|\text{data}) \propto f(\text{data}|\phi_I)f(\phi_I), \tag{2}$$

$$f(\phi_N|\phi_I, \text{data}) = f(\phi_N|\phi_I). \tag{3}$$

Thus (2), the posterior marginal distribution for $\phi_I$, is typically governed by the usual asymptotic theory which applies in the identifiable case. On the other hand, (3), 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 $\theta$ will often lead to prior dependence between $\phi_I$ and $\phi_N$, and consequently

$$\begin{aligned} f(\phi_N|\text{data}) &= \int f(\phi_N|\phi_I) f(\phi_I|\text{data}) d\phi_I \\ &\neq f(\phi_N). \end{aligned}$$

That is, marginally there can be some learning about $\phi_N$. We refer to this as *indirect learning*, as it is learning about $\phi_N$ that results only because of learning about $\phi_I$. We also note that asymptotically (2) will concentrate to a point mass at the true value of $\phi_I$, so that the posterior marginal distribution of $\phi_N$ will tend to the (non-degenerate) prior conditional distribution (3) evaluated at this value of $\phi_I$.

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

$$\begin{aligned} E(\psi|\text{data}) &= \int\int g(\phi_I, \phi_N) f(\phi_I, \phi_N|\text{data})\ d\phi_N\ d\phi_I \\ &= \int\int g(\phi, \lambda) f(\phi_N|\phi_I)\ d\phi_N\ \ f(\phi_I|\text{data})\ d\phi_I \\ &= E(\tilde{g}(\phi_I)|\text{data}), \end{aligned}$$

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 $\phi_I$ alone. Thus under weak regularity conditions its asymptotic behaviour will be described by the the usual asymptotic theory applied to this model. That is, if the model is correct and if an *iid* sample of size $n$ yields $\hat{\psi}_n = E(\psi|\text{data})$, then

$$n^{1/2}\{\hat{\psi}_n - \tilde{g}(\phi_I)\} \Rightarrow N\left[0, \{\tilde{g}'(\phi_I)\}^T I(\phi_I)^{-1}\{\tilde{g}'(\phi_I)\}\right].$$

Moreover, the RMSE incurred by estimating $\psi$ with $E(\psi|\text{data})$ at a given sample size $n$ can be approximated as

$$ARMSE = \left[\{\tilde{g}(\phi_I) - g(\phi_I, \phi_N)\}^2 + n^{-1}\{\tilde{g}'(\phi_I)\}^T I(\phi_I)^{-1}\{\tilde{g}'(\phi_I)\}\right]^{1/2}, \tag{4}$$

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.

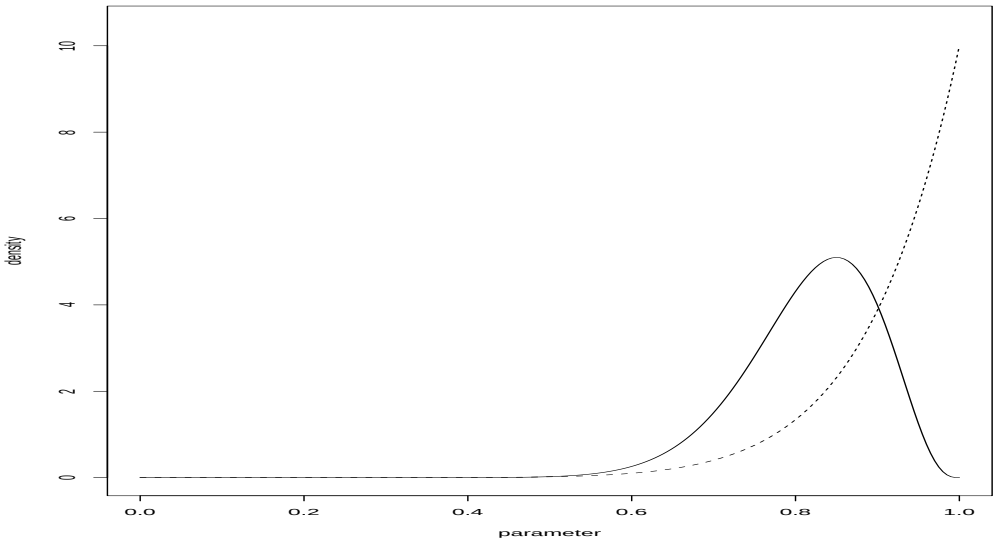


Figure 3: Prior distributions for the probability of correct classification. The solid curve is the Beta(18,4) density function, and the dotted curve is the Beta(10,1) density function

## 2.3 Performance of Model A Estimators

Investigators often have a rough idea about the extent of the mismeasurement in a mismeasured variable scenario. Say, for instance, that in Scenario I the investigators are comfortable with an assessment that the two tests are 'pretty good, but not perfect'. 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 $r \sim Unif(0, 1)$ for the population prevalence. For later reference we refer to this prior 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 a single test. In the absence of any very substantive prior knowledge, the four probabilities of correct classification which govern the two tests are treated interchangeably.

For Model A we obtain a transparent parameterization $\phi = (\phi_I, \phi_N)$ 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 we can easily evaluate the prior density $f(\phi)$ at any given point, but it is not simple to give an

expression for this density.

As in Section 3.1 we focus on $r$ as the parameter of interest. The posterior mean of $r = \phi_{N,1}$ is identically the posterior mean of

$$\begin{aligned} \tilde{g}(\phi_I) &= \int\int \phi_{N,1} f(\phi_N|\phi_I)\ d\phi_{N,1}\ d\phi_{N,2} \\ &= \frac{\int\int \phi_{N,1} f(\phi_I, \phi_N) d\phi_{N,1} d\phi_{N,2}}{\int\int f(\phi_I, \phi_N) d\phi_{N,1} d\phi_{N,2}} \end{aligned} \tag{5}$$

with respect to the identifiable sub-model parameterized by $\phi_I$ alone. Unfortunately, $\tilde{g}$ cannot be evaluated in closed-form. We can, however, using 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 $E(r|\text{data})$ under Model A as an estimator of $r$.

Of course we also need to evaluate the first derivatives of $\tilde{g}(\phi)$ in order to determine the asymptotic variance of $E(r|\text{data})$, as in (4). Analytic differentiation of (5) 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 $\tilde{g}(\phi)$ and $\tilde{g}(\phi + \epsilon(1\ 0\ 0))$ using the same quadrature points, in order to approximate $\partial\tilde{g}(\phi)/\partial\phi_1$. In doing so we take care to check that $\epsilon$ is small enough and the number of quadrature points is large enough to obtain stable approximations to the derivatives. We also note that we are interested in large sample sizes at which the asymptotic variance is small relative to the asymptotic bias, hence exacting precision in computing the derivatives is not required. Thus we can evaluate the ARMSE (4) in the present context, albeit with some numerical effort.

We re-consider the DGM (i) parameter values given in Section 3.1. With prior (i) and a sample size of $n = 2000$ we compute $ARMSE = 0.015$ for estimating the prevalence $r$. This compares *very* favourably to the corresponding ARMSE values for Model B appearing in Figure 1, being considerably smaller when $\Delta = r_2 - r_1$ is large, and very much smaller when $\Delta$ is small. Even at this large sample size, infusing crude prior information into Model A may be preferable to dichotomizing 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 we consider a second prior under which all four classification probabilities are assigned Beta(10,1) prior distributions. This prior, henceforth referred to as prior (ii), has a much different shape as indicated in Figure 3, though it still reflects a crude notion of the two tests being good but perhaps not perfect. It does turn out that the posterior mean of $r$ performs less well under prior (ii), with $ARMSE = 0.037$. However, this is still quite favourable relative to the Model B performance displayed in Figure 1, especially when $\Delta = r_2 - r_1$ is not very large.

We also consider DGM (ii) given in Section 3.1. With this DGM we obtain $ARMSE = 0.079$ with prior (i), and $ARMSE = 0.050$ with prior (ii). In the former case this is better than the Model B performance given in Figure 2 if $\Delta$ is less than about 0.1, and in the latter case it is better if $\Delta$ is less than about 0.15. Again, crude prior information infused into Model A can guard against the small $\Delta$ pitfall associated with the Model B estimator.

| 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 | unif. | 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 | unif. | 0.122 | 0.1332 | (0.0038) | 23% | 0.22 |
| | B | (i) | 0.069 | 0.0717 | (0.0015) | 6% | 0.11 |

Table 1: Performance of four posterior means for $r$ in a simulation study. 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 $\Delta = r_2 - r_1 = 0.7$, and $n = 2000$.

## 2.4 Simulation Comparisons in Scenario I

### 2.4.1 Performance of Model A and B estimators

We carry out a small simulation study to augment the asymptotic comparisons made thus far. We consider both DGM (i) and DGM (ii) introduced in Section 3.1, with the difference between sub-population prevalences taken to be $\Delta = 0.07$. This corresponds to a setting where there is a practical difference between the sub-population prevalences but the asymptotic analysis suggests that the difference may not be large enough to yield good estimates. We simulate 200 datasets under each DGM, and for each dataset 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. assigns a uniform distribution to $(r_1, r_2)$]. Each estimate is obtained from 25000 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, our informal monitoring of the simulation runs indicates that the mixing performance of the Gibbs sampler in Models A and B is tolerable but not ideal. Gelfand and Sahu (1999) note 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 25000 seems to yield tolerable mixing for the purposes of this simulation study, there is some possibility of slightly improving the reported performance in Table 1 by either further increasing the MCMC sample size or by 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 problem similar to Model A, see Gustafson, Le and Saskin (2001).

In examining Table 1, note first that the empirical RMSE observed for $E_A(r|\text{data})$ agrees quite closely with the asymptotic RMSE, 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 even at $n = 2000$ the Model B asymptotics have not fully 'kicked in' yet. For DGM (i) the empirical RMSE for $E_B(r|\text{data})$ 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), though asymptotically the prior does not matter. Thus another aspect of the 'moderate $\Delta$' problem has emerged. Even though Model B is governed by regular asymptotics, if $r_1$ and $r_2$ are moderately close together the asymptotics may not yield a good approximation to the actual estimator performance unless the sample size is ridiculously large.

In comparing the RMSE values in Table 1 we see that Model A with either crude prior yields a lower RMSE than Model B with a flat prior. On the other hand, 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.

With regard to the credible intervals described in Table 1, we simply note that extreme under-coverage and over-coverage arises. 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 under-coverage. Again this speaks to the asymptotics not yet being accurate for Model B, even with $n = 2000$.

### 2.4.2 Bayes Performance

Our asymptotic and simulation comparisons thus far have considered average performance of estimators in repeated sampling with fixed underling 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 take the decision-theoretic point of view that nature generates parameter values (and consequently datasets) 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 datasets, in each instance first drawing a parameter vector and then simulating a dataset 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 dataset we compute the posterior mean and the 80% equal-tailed credible interval for the prevalence $r$ using 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 $\Delta$, 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 results in a similar RMSE for estimating $r$. And when nature uses prior (ii), Model B with a flat prior

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

Table 2: Bayes performance of $r$ estimators under various settings. Nature employs either prior (i) or prior (ii) to generate 200 datasets 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 and average length of the 80% equal-tailed credible interval are reported.

has a lower RMSE than Model A with either crude prior. Thus in aggregate the advantage of identifiability which arises when $\Delta$ 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 datasets generated under small values of $\Delta$. Figure 4 plots $|\hat{r} - r|$ versus $|\Delta|$ in the various scenarios considered in Table 2. Clearly the absolute error varies much less with $\Delta$ under Model A and a crude prior than under Model B with a flat prior, especially when nature employs prior (i).

Of course if nature and the investigator use the same prior then credible intervals will have exactly their nominal coverage, irrespective of whether the model is identifiable or not. The results in Table 2 are in accord with this fact. When the wrong crude prior is used, the coverage varies, though the deviations from nominal coverage are much less extreme than was exhibited for frequentist coverage in Table 1.

### 2.4.3 Performance of Model C Estimators

We briefly 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, $\delta_0$ is assigned an exponential distribution with rate $k(q_1, q_2)$, truncated to the interval $[0, \delta_{MAX}(q_1, q_2)]$. Similarly, $\delta_0$ is assigned an exponential distribution with rate $k(p_1, p_2)$, truncated to the interval $[0, \delta_{MAX}(p_1, p_2)]$. 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, for instance, a value of $\delta_0$ corresponding to conditional correlation $\rho_0$ has a prior density which is $\exp(-c\rho_0)$ times the prior density of $\delta_0 = 0$ which corresponds to $\rho_0 = 0$. For the sake of illustration we take $c = 4 \log 4$, so that $\rho_i = 0.25$ is four times less likely *a priori* than $\rho_i = 0$ in this sense.

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 we simulate data under DGM (ii) to see if fitting Model C can ameliorate this problem. Specifically,

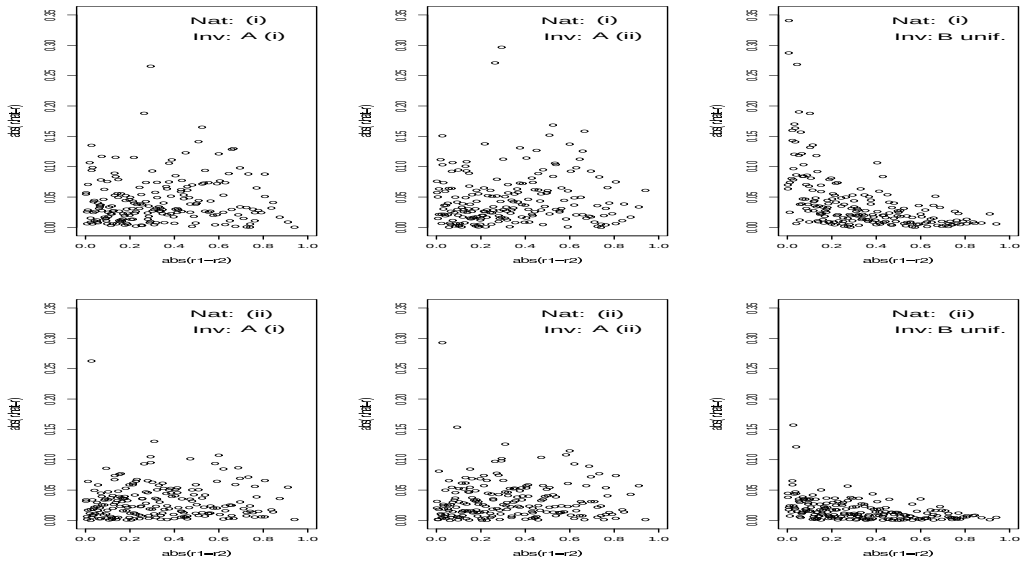


Figure 4: Absolute error $|\hat{r} - r|$ versus absolute difference in sub-population prevalences $|\Delta|$ for the simulated datasets in Section 2.4.2.

we assign a common value $\rho$ to both $\rho_0$ and $\rho_1$, and simulate data under $\rho = 0$, $\rho = 0.125$, and $\rho = 0.25$. 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 two in terms of RMSE, and by about a factor of three in terms of the length of credible interval. As the correlation increases the RMSE discrepancy decreases, but does not disappear, while the difference in credible interval length persists. The advantage of using Model C when correlation is in fact present is a better coverage rate, as one might expect from a procedure which admits the possibility of positive correlation.

While brevity precludes a full description of the MCMC algorithm used to fit Model C, we note that the algorithm was designed to limit the impact of nonidentifiability on MCMC performance. Specifically, $(q_1, q_2, \delta_0)$ is updated simultaneously given the other parameters and

| | Model B | | | Model C | | |
|---|---|---|---|---|---|---|
| $\rho$ | 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 |

Table 3: Comparison of prevalence estimators based on Models B and C, with the prior distributions described in Section 2.4.3. The three rows correspond to DGM (i) with three different underlying values of $\rho$, 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 25000 MCMC iterations after 1000 burn-in iterations.

the true exposure status, as is $(p_1, p_2, \delta_1)$. Thus our approach to fitting Model C differs from that of Dendukuri and Joseph (2001), both in the prior downweighting of higher correlations and in the approach to MCMC fitting. We still find that for some datasets, particularly those generated under high correlations, that our MCMC algorithm can mix somewhat poorly. Again, more research on good MCMC algorithms for nonidentified problems is required.

# 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

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

We refer to this model as Model E. Note that this model invokes the common assumption of nondifferential measurement error, as $X^*$ and $Y$ are conditionally independent given $X$. Note as well that the given parameterization makes $r = Var(X^*|X)/Var(X)$ 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 $\theta = (\beta_0, \beta_1, \sigma^2, \mu, \lambda^2, r)$ are unknown then the model is nonidentifiable. Intuitively, one can consistently estimate only five parameters: an intercept, slope and residual variance describing the distribution of $Y|X^*$, along with a mean and variance describing 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 alternate route to gaining identifiability is through model expansion. In a recent paper, Huang and Huwang (2001) demonstrate that the model

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

is identifiable, even if all seven parameters $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \mu, \lambda^2, r)$ are unknown. Henceforth we refer to this model as Model F. Initially it seems remarkable that replacing the linear regression function in Model E with the quadratic regression function in Model F leads to identifiability. The key is that under Model F $Var(Y|X^*)$ is no longer constant. It is now a quadratic function of $X^*$. Roughly speaking then, we can consistently estimate two parameters describing $Var(Y|X^*)$, three parameters describing $E(Y|X^*)$, and two parameters describing the marginal distribution of $X^*$. Unfortunately, the distribution of $(Y, X^*)$ implied by Model F does not have a closed-form, and therefore one cannot readily evaluate the Fisher information matrix in order

to study the asymptotic behaviour of estimates generated by Model F. However, Model F can be fit to data with iterated algorithms, including MCMC algorithms.

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

## 3.1 Performance of Model E Estimators

For Model E a transparent parameterization $\phi = (\phi_I, \phi_N)$ is obtained if we take the components of $\phi_I$ to be

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

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

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

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 $\beta_0 \sim \mathcal{N}(0,1)$, $\beta_1 \sim \mathcal{N}(0,1)$, $\sigma^2 \sim \mathcal{IG}(0.5, 0.5)$, $\mu \sim \mathcal{N}(0,1)$, $\lambda^2 \sim \mathcal{IG}(0.5, 0.5)$, and $r \sim Beta(\alpha_1, \alpha_2)$. The use of standard normal priors might be appropriate if the data are standardized prior to analysis, while the choice of hyperparameters for the inverse gamma distributions can be interpreted as giving 'unit-information' priors, roughly in the spirit of Kass and Wasserman (1995). We also consider three choices of hyperparameters for the prior on $r$. Prior (i) uses $(\alpha, \beta) = (1, 1)$, i.e. in the absence of any subjective knowledge $r$ is assigned a uniform prior. Prior (ii) uses $(\alpha, \beta) = (7.6, 14.1)$, giving $E(r) = 0.35$, $SD(r) = 0.10$, while the more concentrated prior (iii) uses $(\alpha, \beta) = (24.9, 58.1)$, giving $E(r) = 0.30$, $SD(r) = 0.05$.

We consider what happens when the data-generating mechanism involves the parameter values $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2 = 0.25$, $\mu = 0$, $\lambda^2 = 1$, $r = 0.25$. Note that the true value of $r$ is one standard deviation below the mean with respect to both priors (ii) and (iii). Thus in a rough sense we can view these priors as being 'typically' representative of the truth, though of course prior (iii) represents stronger knowledge about $r$ than does prior (ii).

Following Section 2.2, the key to describing the asymptotic behaviour 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

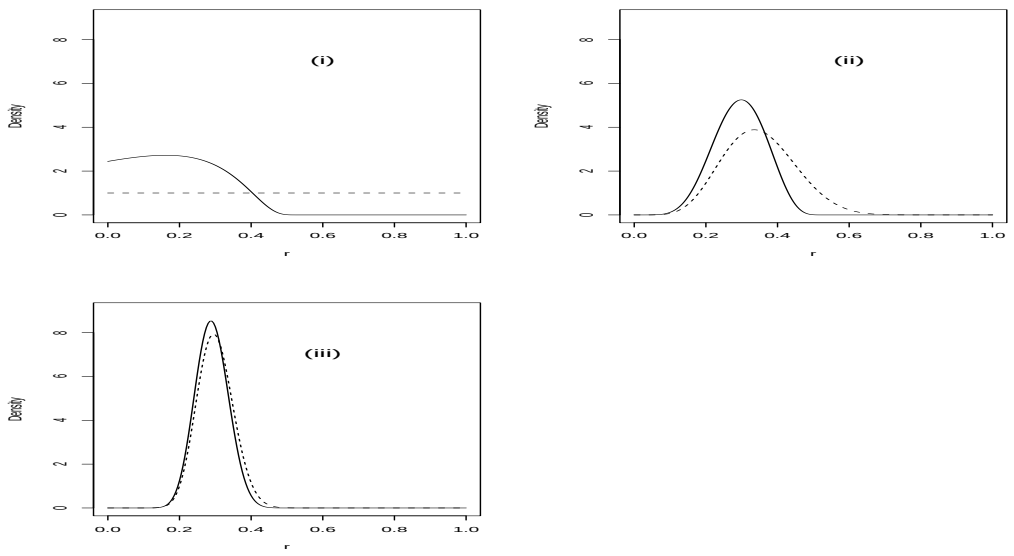


Figure 5: Prior (dashed curve) and limiting posterior (solid curve) densities for $r$ under Model E. The true value of $r$ is 0.25, while the true values of the other parameters are given in the text.

of the $\phi \to \theta$ mapping is 1, and consequently

$$\begin{aligned} f(r|\beta_0^*, \beta_1^*, \sigma_*^2, \mu^*, \lambda_*^2) \quad &\propto \quad f_N(\beta_1^*(1+r)) f_{IG}(\sigma_*^2 - r(\beta_1^*)^2\lambda_*^2) f_{IG}(\lambda_*^2/(1+r)) \ \times \\ &\qquad r^{\alpha_1 - 1}(1-r)^{\alpha_2 - 1} I_{(0, m(\beta_1^*, \sigma_*^2, \lambda_*^2))}(r), \end{aligned} \tag{6}$$

where $f_N()$ is the $\mathcal{N}(0,1)$ density function, $f_{IG}()$ is the $\mathcal{IG}(0.5, 0.5)$ 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

$$\tilde{g}(\beta_0^*, \beta_1^*, \sigma_*^2, \mu^*, \lambda_*^2) \quad = \quad \int g(\beta_0^*, \beta_1^*, \sigma_*^2, \mu^*, \lambda_*^2, r) f(r|\beta_0^*, \beta_1^*, \sigma_*^2, \mu^*, \lambda_*^2)\, dr$$

via one-dimensional numerical integration. Thus the bias and asymptotic variance of $\hat{\psi} = E(\psi|\text{data})$ as in (4) are readily evaluated.

Of course we can interpret (6) evaluated at the true value of $\phi_I$ as the large-sample limiting posterior density of $r$. For each prior (i) through (iii) the prior density of $r$ and the limiting posterior density of $r$ under the illustrative DGM appear in Figure 5. 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 $\beta_1$ by their posterior means. For priors (i) through (iii) the asymptotic bias, variance and RMSE corresponding to a sample size of $n = 250$ are given in Table 4. As we might expect, both the bias and variance of $\hat{r}$ decrease as the prior distribution for $r$ improves. The surprising feature, which relates back the indirect updating witnessed in Figure 5, is that the performance of $\hat{r}$ based on the flat 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

| | $\hat{r}$ | | | | $\hat{\beta}_1$ | | | |
|---|---|---|---|---|---|---|---|---|
| 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) |

Table 4: Asymptotic bias and variance of posterior means under Model E. Results are given for both $\hat{r} = E(r|\text{data})$ and $\hat{\beta}_1 = E(\beta_1|\text{data})$, under priors (i), (ii) and (iii) described in the text. The approximate standard deviation (SD) and RMSE are based on a sample size of $n = 250$. The RMSE incurred under Model D, if $r$ is assumed known and equal to the prior mean, also appears in parentheses. The underlying parameter values are $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2 = 0.25$, $\mu = 0$, $\lambda = 1$, $r = 0.25$.

of $r$. For instance, with prior (i) we simply have $\hat{r} = 0.5$, regardless of the data. The RMSE under Model E with a full 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, in order to reap the benefit of indirect learning about $r$.

Table 4 also shows that the bias in estimating $\beta_1$ 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 of course the asymptotic variance of $\hat{\beta}_1^*$ is unaffected by the prior. However there is a negative asymptotic covariance between the two, which decreases in magnitude as the prior improves fast enough to cause the slight increase overall. Surprisingly, then, we can estimate $\beta_1$ about as well using the flat prior (i) as we can with 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 we use simulation to evaluate the performance of Model F estimators. We extend the DGM used above by taking $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2 = 0.25$, $\mu = 0$, $\lambda^2 = 1$, $r = 0.25$ as before, and considering (i) $\beta_2 = 0$, (ii) $\beta_2 = 0.125$, and (iii) $\beta_2 = 0.25$. The latter value is deliberately chosen to maximize the curvature of the true regression function subject to the function being monotone on the interval from $\mu - 2\lambda = -2$ to $\mu + \lambda = 2$ containing 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 $|\beta_2| \leq (2c)^{-1}|\beta_1|$. Since we expect many relationships of practical interest to be monotone, in a practical sense DGM (iii) represents an extreme degree of curvature.

Under each DGM (i) through (iii) we simulate 200 datasets. For each dataset posterior means and credible intervals for $r$ and $\beta_1$ are computed, under each prior (i) through (iii). The results are summarized in Table 5. We omit details of the relatively straightforward MCMC algorithm used to fit Model F, but note that this algorithm seems to mix quite well in most situations, but less well in the case of DGM (i) and prior (i), i.e. no curvature and no prior information. As mentioned earlier, this is not surprising in light of other MCMC experience

| | | estimating $r$ | | | estimating $\beta_1$ | | |
|---|---|---|---|---|---|---|---|
| | | DGM | | | 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 |

Table 5: Performance of Model F posterior distribution in estimating $r$ and $\beta_1$. DGMs (i) through (iii) correspond to increasing curvature in the regression function, and priors (i) through (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 datasets of size $n = 250$. Each posterior mean and credible interval is computed using 20000 MCMC iterations after 1000 burn-in iterations.

with nonidentified or weakly identified scenarios.

In examining Table 5 we see that for both estimands increased curvature in the regression function leads to shorter credible intervals with closer to nominal coverage. Thus there is a benefit associated with the degree of identifiability. On the other hand, the RMSE performance of the posterior means in fact tends to increase somewhat as the curvature increases. Given this, and given that DGM (iii) represents an extreme degree of curvature in the sense described above, we conclude that this benefit is quite modest.

Table 5 also indicates that when there is no curvature (i.e. Model E is in fact correct), the Model F performance is very similar to the Model E performance quantified in Table 4. On the other hand, the RMSE performance of the estimators is more favourably affected by an improved prior distribution when curvature is present than when it is absent.

# 4 Discussion

The conventional view of identifiability might be crudely summarized as 'identifiability good, nonidentifiability bad'. The findings in this paper, however, suggest that a more nuanced view be adopted. We have seen that it may not take very much prior information to yield reasonable inferences in a nonidentified model, and we have seen that it may take ridiculously large sample sizes and/or underlying parameter values which are very far from a nonidentifiable sub-model to yield reasonable inferences in an identifiable model. These issues are particularly germane in

problems involving measurement error and misclassification, where lack of knowledge about the extent to which variables are mismeasured often raises questions about identifiability.

We have seen that there can be potential for considerable 'indirect' Bayes learning about a nonidentifiable parameter. This speaks in favour of assigning a crude prior to such a parameter if possible, rather than fixing its value at a 'best guess' value. In Scenario II, for instance, Model E with a crude prior substantially outperforms Model D with a corresponding best guess. The use of such a prior might be regarded as 'soft' model contraction, whereas presuming a subset of the parameters to be known is a 'hard' contraction. 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.

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 sub-model comprised of the original nonidentifiable model. Thus in either case we have troubling design issues in scenarios similar to those studied here. Further research is needed to make general recommendations about study design and analysis in specific scenarios.

# References


Brenner, H. (1996). How independent are multiple 'independent' diagnostic classifications? *Statistics in Medicine* **22**, 2987-3003.

Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society 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. *Biomedical Research* **11**, 423-434.

Gelfand, A.E. and Sahu, S.K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association* **94**, 247-253.

Gustafson, P., Le, N.D., and Saskin, R. (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* **57**, 598-609.

Hui, S.L. and Walter, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics* **36**, 167-171.

Huang, Y.H.S. and Huwang, L. (2001). On the polynomial structural relationship. *Canadian Journal of Statistics* **29**, 495-512.

Johnson, W.O., Gastwirth, J.L., and Pearson, L.M. (2001). Screening without a "gold-standard": the Hui-Walter paradigm revisited. *American Journal of 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 Journal of 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. *Journal of the American Statistical Association* **90**, 928- 934.

Neath, A.E., and Samaniego, F.J. (1997). On the efficacy of Bayesian inference for nonidentifiable models. *American Statistician* **51**, 225-232.

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. *Econometrika* **50**, 1-25.