Econometric Theory, **14**, 1998, 483–509. Printed in the United States of America.

REVISING BELIEFS IN NONIDENTIFIED MODELS

DALE J. POIRIER
University of Toronto

A Bayesian analysis of a nonidentified model is always possible if a proper prior on all the parameters is specified. There is, however, no Bayesian free lunch. The "price" is that there exist quantities about which the data are uninformative, i.e., their marginal prior and posterior distributions are identical. In the case of improper priors the analysis is problematic—resulting posteriors can be improper. This study investigates both proper and improper cases through a series of examples.

1. INTRODUCTION

> In passing it might be noted that unidentifiability causes no real difficulty in the Bayesian approach.
>
> —D.V. Lindley

The concept of identification is the same in Bayesian and non-Bayesian analyses and goes back to Koopmans and Reiersøl (1950). Dréze (1974) and Kadane (1974) provided Bayesian synopses. In particular, Kadane (1974) forcefully argued that identification is a property of the likelihood function and is the same whether considered from classical or Bayesian perspectives. The "solution" to an identification problem is also in common agreement: more information is required, and this usually means non-data-based information. Here, however, paths begin to diverge. Non-Bayesians introduce dogmatic restrictions on the unknown parameters until identification is achieved. Bayesians may do so as well, but they also have the option of introducing nondogmatic information via a prior distribution. Lindley's quote at the beginning of the paper (Lindley, 1971, p. 46, fn. 34) refers to the fact that a Bayesian analysis of a nonidentified model is always possible if a proper prior on all the parameters is specified.

There is, however, no Bayesian free lunch. The "price" is that there exist quantities about which the data are uninformative, i.e., their prior and posterior distributions are identical. Although this issue has been discussed elsewhere in the literature (e.g., Dréze, 1974; Dréze and Richard, 1983; Leamer, 1978), there are some subtle issues that warrant further discussion. This study illustrates these issues by means of a series of examples in Section 3. Section 4 then turns to the more problematic case of improper priors.

Address correspondence to: Dale J. Poirier, Department of Economics, University of Toronto, 150 St. George Street, Toronto, Ontario M5S 3G7, Canada; e-mail: poirier@chass.utoronto.ca.

© 1998 Cambridge University Press 0266-4666/98 $9.50

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

The next section characterizes the general situation and suggests two senses in which the data are "uninformative": either conditional or marginal of other parameters. The quantities about which no marginal learning occurs depend on both prior beliefs and the nature of the nonidentifiability in the data. The dependency of informativeness on prior beliefs distinguishes it from the concept of identification (see Kadane, 1974, pp. 180–184).

Throughout this discussion density functions are denoted generically by $f(\cdot)$, leaving the arguments to indicate the random variables to which the density applies. When such notation might cause confusion, subscripts will be added to clarify matters. The notation $N_T(\mu, \Sigma)$ denotes a T-dimensional multivariate normal distribution with mean μ and covariance matrix Σ. The corresponding multivariate normal density is denoted $\phi_T(\cdot \mid \mu, \Sigma)$. When $T = 1$, the subscript is dropped. Finally, $\Phi(\cdot)$ is used to denote the univariate normal cumulative distribution function (c.d.f.).

2. PROPER PRIORS: GENERAL RESULTS

Given data y, consider a regular likelihood function $\mathcal{L}(\theta; y)$, where $\theta \in \Theta$ is an unknown $K \times 1$ vector, and $\Theta \subseteq \mathcal{R}^K$. Suppose θ is nonidentified. In other words, for all $\theta^{(1)} \in \Theta$ there exists $\theta^{(2)} \in \Theta$, $\theta^{(1)} \neq \theta^{(2)}$, such that $\mathcal{L}(\theta^{(1)}; y) = \mathcal{L}(\theta^{(2)}; y)$ for all y, i.e., $\theta^{(1)}$ and $\theta^{(2)}$ are observationally equivalent. Suppose, however, that $K - 1$ parametric functions are identified. Let $\theta = [\psi', \lambda]'$, where ψ is $(K-1) \times 1$ and λ is 1×1. Without loss of generality, assume ψ is identified and λ is nonidentified, i.e., give $\theta^{(1)} = [\psi^{(1)\prime}, \lambda^{(1)}]'$, $\theta^{(2)} = [\psi^{(2)\prime}, \lambda^{(2)}]' \in \Theta$, $\mathcal{L}(\theta^{(1)}; y) = \mathcal{L}(\theta^{(2)}; y)$ implies $\psi^{(1)} = \psi^{(2)}$. The term $\lambda^{(1)}$ need not equal $\lambda^{(2)}$. Let $\Lambda(\psi) = \{\lambda \mid [\psi', \lambda]' \in \Theta\}$ denote the permissible values of λ given $\psi = \psi^{(1)} = \psi^{(2)}$. Because ψ is identified and λ is nonidentified, the likelihood is constant over $\lambda \in \Lambda(\psi)$, and so it can be written simply as a function of ψ, say $\mathcal{L}_*(\psi; y)$, not depending on the nonidentified λ. This section describes what a Bayesian researcher with a proper prior probability density function (p.d.f.) $f(\theta)$, $\theta \in \Theta$ can and cannot learn from such data. Extensions to the case in which more than one parameter is nonidentified are straightforward and are largely ignored in the subsequent discussion until Section 3.9.

Care must be taken with notation, especially when the parameter space is not a product space across ψ and λ, i.e., when ψ and λ are not variation-free. Let $\Psi = \{\psi \mid [\psi', \lambda]' \in \Theta \text{ for some } \lambda\}$ and $\Lambda = \{\lambda \mid [\psi', \lambda]' \in \Theta \text{ for some } \psi\}$ be the permissible values in Θ of ψ and λ, respectively. The terms ψ and λ are variation-free iff $\Theta = \Psi \times \Lambda$ with no restrictions across Ψ and λ. Analogous to $\Lambda(\psi)$, let $\Psi(\lambda) = \{\psi \mid [\psi', \lambda]' \in \Theta\}$ be the permissible values of ψ given λ. Although the assumed nonidentifiability of λ implies that $\mathcal{L}_*(\psi; y)$ does not vary with λ, it does not rule out the possibility that its domain $\Psi(\lambda)$ depends on λ.

Because the likelihood $\mathcal{L}_*(\psi; y)$ does not vary with λ, it might be thought that the only information available on λ is contained in its prior and so perhaps the data are uninformative in the sense that the data to not revise prior beliefs. To see

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

that this is not necessarily the case, it is useful to distinguish whether the prior marginal density of λ, $f(\lambda)$, or the prior conditional density of λ given ψ, $f(\lambda|\psi)$, are identical to their posterior counterparts $f(\lambda|y)$ and $f(\lambda|\psi,y)$, respectively.

DEFINITION. *The data y are marginally uninformative for λ iff $f(\lambda|y) = f(\lambda)$. The data y are conditionally uninformative for λ given ψ iff $f(\lambda|\psi,y) = f(\lambda|\psi)$.*

Consider the marginal posterior density of λ:

$$f(\lambda|y) = \frac{f(\lambda)f(y|\lambda)}{f(y)}, \qquad \lambda \in \Lambda, \tag{1}$$

where

$$\begin{aligned} f(y|\lambda) &= \int_{\Psi(\lambda)} f(\psi|\lambda)\mathcal{L}_*(\psi;y)\,d\psi, \\ &= E_{\psi|\lambda}[\mathcal{L}_*(\psi;y)], \qquad \lambda \in \Lambda, \end{aligned} \tag{2}$$

$$f(y) = \int_{\Lambda}\int_{\Psi(\lambda)} f(\psi,\lambda)\mathcal{L}_*(\psi;y)\,d\psi\,d\lambda. \tag{3}$$

The data are marginally informative for λ iff $f(\lambda|y) \neq f(\lambda)$, or equivalently, $f(y|\lambda) \neq f(y)$. Because the likelihood in (2) does not depend on λ, any dependence of $f(\lambda|y)$ on λ arises from the conditional prior distribution $\psi|\lambda$: through either the density $f(\psi|\lambda)$ or its support $\Lambda(\psi)$. In the absence of such dependence, the data are marginally uninformative for λ. This leads directly to the following elementary proposition.

PROPOSITION 1. *If ψ and λ are a priori independent (denoted $\psi \perp \lambda$), then the data are marginally uninformative for λ. Equivalently, if the data are marginally informative for λ, then ψ and λ are a priori dependent.*

The converse of Proposition 1 is not necessarily valid, although it will be seen in the examples that follow that when ψ and λ are a priori dependent, then except for irregular cases, the data are marginally informative for λ.[1] In other words, prior dependence is a necessary but not a sufficient condition for informativity. The terms ψ and λ can be dependent whereas $f(y|\lambda)$ in (2) does not depend on λ.

Note, however, regardless of the prior dependence between ψ and λ, the conditional beliefs $\lambda|\psi$ are not updated:

$$\begin{aligned} f(\lambda|\psi,y) &= \frac{f(\psi,\lambda|y)}{f(\psi|y)} = \frac{[\,f(\psi)f(\lambda|\psi)\mathcal{L}_*(\psi;y]/f(y)}{\left[\int_{\Lambda(\psi)} f(\psi)f(\lambda|\psi)\mathcal{L}_*(\psi;y)\,d\lambda\right]\Big/f(y)} \\ &= \frac{f(\psi)f(\lambda|\psi)\mathcal{L}_*(\psi;y)}{f(\psi)\mathcal{L}_*(\psi;y)} \\ &= f(\lambda|\psi), \qquad \lambda \in \Lambda(\psi), \qquad \psi \in \Psi. \end{aligned} \tag{4}$$

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

Given ψ, the data do not modify prior beliefs about λ. Equivalently, λ and y are conditionally independent given ψ. This implies that although the data may or may not be marginally uninformative for nonidentified parameters, *data are always conditionally uninformative for nonidentified parameters given the identified parameters*. This observation is not new. For example, it manifests itself in the context of linear simultaneous equations models where prior beliefs about the structural parameters given the reduced form parameters are not revised by the data (see, e.g., Dréze, 1974; Dréze and Richard, 1983, pp. 521–522, 536–537). For clarity the following immediate implication of (4) is noted.

PROPOSITION 2. *The data are always conditionally* (*on* ψ) *uninformative for* λ.

Using Proposition 2, it is convenient to rewrite posterior (1) as

$$\begin{aligned} f(\lambda|y) &= \int_{\Psi(\lambda)} f(\lambda|\psi,y) f(\psi|y)\, d\psi, \\ &= \int_{\Psi(\lambda)} f(\lambda|\psi) f(\psi|y)\, d\psi, \\ &= E_{\psi|y}[\, f(\lambda|\psi)], \qquad \lambda \in \Lambda. \end{aligned} \tag{5}$$

In other words, posterior beliefs about the nonidentified λ are the average prior conditional beliefs given the identified ψ, averaged with respect to the posterior of ψ. Also note that (5) is the posterior expectation of a prior density not involving the data.

Although the data can be marginally informative for the nonidentified parameter λ, there necessarily exist other quantities for which the data are marginally uninformative. The following proposition (apparently new) describes these quantities.

PROPOSITION 3. *Let* $f(\psi,\lambda)$ *be a proper joint prior p.d.f., and denote the conditional p.d.f. and c.d.f. of* λ *given* ψ *by* $f(\lambda|\psi)$ *and* $F(\lambda|\psi)$, *respectively. Define a new random variable* $\nu = g(\psi,\lambda)$ *that is the probability integral transform based on the conditional distribution of* λ *given* ψ, *i.e.*, $g:\Theta \to [0,1]$ *and* $\nu = g(\psi,\lambda) \equiv F(\lambda|\psi)$.

(a) ψ *and* ν *are independent.*
(b) ν *is uniformly distributed over* $[0,1]$.
(c) *The data are marginally uniformative for* ν.
(d) *The data are marginally uninformative for any known one-to-one function of* ν.

Proof. Because $\partial g(\psi,\lambda)/\partial\lambda = f(\lambda|\psi) > 0$, the inverse transformation $\lambda = g^{-1}(\psi,\nu)$ is well defined. Using the change-of-variable technique,

$$\begin{aligned} f(\psi,\nu) &= \left[\frac{\partial\lambda}{\partial\nu}\right] f(\psi) f(\lambda|\psi) \\ &= [\, f(\lambda|\psi)]^{-1} f(\psi) f(\lambda|\psi) \\ &= f(\psi), \qquad \psi \in \Psi, \qquad \nu \in [0,1]. \end{aligned} \tag{6}$$

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

Therefore, (a) and (b) follow immediately after integrating (6) with respect to ψ. (a) implies (c) as noted previously. Finally, (d) follows immediately from (c). ■

Proposition 3 provides the unavoidable price a Bayesian pays for updating marginal beliefs for the nonidentified λ: nothing is learned about $\nu = F(\lambda/\psi)$. Four points are worth noting.

First, Manski (1994) offered a plethora of examples in econometrics where the sampling scheme is such that some parameters ψ are identified, others λ are nonidentified, and data-free bounds for the latter given the former are warranted (see, e.g., Section 3.5). Such bounds serve as the support $\Lambda(\psi)$ for the prior conditional distribution of λ given ψ. Bounds depending on unknown parameters, however, are of limited usefulness. The point of Proposition 2 is that the posterior conditional inferences for λ given ψ are identical to prior inferences: the data are conditionally uninformative for λ. However, the existence of the previously noted bounds necessitates prior dependence between ψ and λ, and the data are potentially marginally informative for λ. Of course, Proposition 3 implies there is always a reparameterization into ν for which the data are marginally uninformative. The fundamental roles played by prior conditional distributions $\psi|\lambda$ (in Proposition 1) and $\lambda|\psi$ (in Propositions 2 and 3) are apparent.

Second, the fact that "although the likelihood does not vary with λ the data may nonetheless be marginally informative for λ" does not violate the likelihood principle as suggested by Vijverberg (1993, p. 75).[2] When the parameter space Θ is not a product space across ψ and λ, the support $\Psi(\lambda)$ depends on λ. The likelihood $\mathcal{L}_*(\psi; y)$ and its domain $\Psi(\lambda)$ can be expressed equivalently as $\mathcal{L}_{**}(\psi; y) \equiv [\mathbf{1}_{\Psi(\lambda)}(\psi)]\mathcal{L}_*(\psi; y)$, where $\mathbf{1}_{\Psi(\lambda)}(\psi)$ is an indicator function equal to unity if $\psi \in \Psi(\lambda)$ and equal to zero otherwise.[3] The informal interpretation of the likelihood principle (e.g., Poirier, 1995, p. 227) that the likelihood function contains all the evidence in the data concerning the unknown parameters is obvious in terms of $\mathcal{L}_{**}(\psi; y)$.

Third, because the data are marginally uninformative for ν, Bayes factors equal unity for tests involving ν, regardless of the data. For example, consider testing H_0: $\nu = \frac{1}{2}$ versus H_1: $\nu \neq \frac{1}{2}$, with prior specification: $\text{Prob}(H_0) = \underline{\pi} > 0$, $\text{Prob}(H_1) = 1 - \underline{\pi} > 0$, and prior density $f(\nu|H_1)$. Because ψ and ν are a priori independent, $f(\nu|H_0) = f(\nu|H_1) = f(\nu)$, and

$$\begin{aligned} f(y|H_1) &= \int_\Psi \int_0^1 f(\psi) f(\nu) \mathcal{L}_*(\psi; y)\, d\nu\, d\psi \\ &= \int_\Psi f(\psi) \mathcal{L}_*(\psi; y) \left[\int_0^1 f(\nu)\, d\nu \right] d\psi \\ &= \int_\Psi f(\psi) \mathcal{L}_*(\psi; y)\, d\psi = f(y|H_0) = f(y). \end{aligned} \tag{7}$$

Therefore, the Bayes factor for H_0 versus H_1 is $f(y|H_0)/f(y|H_1) = 1$. In other words, the posterior odds equal the prior odds (also see Dréze and Richard, 1983, pp. 538–539).

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

Finally, the interpretation of the data being uninformative for a parameter, marginally or conditionally, depends on the context. In general the parameters of interest are functions of both ψ and λ, and so it is not always clear what to make of these results.

Section 3 provides nine simple examples to help crystallize the discussion. Section 3.1 provides a basic case demonstrating Propositions 1–3 that will prove valuable in later discussion (Section 4) of improper priors. Section 3.2 emphasizes that identifiability is a property of the likelihood and any ambiguity over the likelihood implies ambiguity over identifiability. Section 3.3 takes up the classic case of perfect multicollinearity and suggests that determining the quantity ν in Proposition 3 may aid in prior elicitation. Sections 3.4–3.6 involve examples in which prior dependence between identified and nonidentified parameters arises naturally. In Section 3.4 the dependency results from a positive definiteness constraint on a covariance matrix (reminiscent of Koop and Poirier, 1997). The example in Section 3.5 is representative of those studied in Manski (1994). In both Sections 3.4 and 3.5, flat priors, $f(\psi|\lambda) = c_\psi$ for $\lambda \in \Lambda(\psi)$ where c_ψ is a constant, are investigated to determine the quantity ν in Proposition 3 about which the data are uninformative. Section 3.6 analyzes a Bernoulli model with classification error in which prior dependency also naturally arises. Section 3.7 considers a classic simultaneous equation model and investigates the imposition of an identifying exclusion restriction in a stochastic fashion. Section 3.8 turns to the probit model and offers a possibly shocking conclusion regarding seemingly innocuous normalization restriction. Finally, Section 3.9 considers a multinomial nonparametric model involving numerous nonidentified parameters.

3. PROPER PRIORS: EXAMPLES

3.1. A Basic Case

A simple example of a nonidentified model is $y|\beta,\lambda \sim N(\beta + \lambda, \sigma^2)$, where $\beta,\lambda \in \mathcal{R}$ are unknown and $\sigma^2 \in \mathcal{R}^+$ is known. Clearly, neither β nor λ is identified, but $\psi = \beta + \lambda$ is identified. Suppose β and λ are a priori independent with $\beta \sim N(0,\tau^2)$, $\lambda \sim N(0,\omega^2)$, and $\tau^2,\omega^2 \in \mathcal{R}^+$ known. Then the posterior density $f(\beta,\lambda|y)$ is bivariate normal with mean $(\sigma^2+\tau^2+\omega^2)^{-1}y[\tau^2,\omega^2]'$ and variance

$$(\sigma^2+\tau^2+\omega^2)^{-1}\begin{bmatrix} \tau^2(\sigma^2+\omega^2) & -\tau^2\omega^2 \\ -\tau^2\omega^2 & \omega^2(\sigma^2+\tau^2) \end{bmatrix}.$$

The implied joint prior distribution for ψ and λ is bivariate normal with zero means and variance

$$\Sigma = \begin{bmatrix} \tau^2+\omega^2 & \omega^2 \\ \omega^2 & \omega^2 \end{bmatrix}.$$

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

Hence, the posterior density $f(\psi,\lambda|y)$ is bivariate normal with mean $(\sigma^2+\tau^2+\omega^2)^{-1}y[\tau^2+\omega^2,\omega^2]'$ and variance

$$(\sigma^2+\tau^2+\omega^2)^{-1}\begin{bmatrix}\sigma^2(\tau^2+\omega^2) & -\sigma^2\omega^2\\ -\sigma^2\omega^2 & \omega^2(\sigma^2+\tau^2)\end{bmatrix}.$$

Marginal beliefs about the nonidentified λ are updated from $N(0,\omega^2)$ to $N(\omega^2 y[\sigma^2+\tau^2+\omega^2]^{-1}, \omega^2[\sigma^2+\tau^2][\sigma^2+\tau^2+\omega^2]^{-1})$. Proposition 2, however, implies there is no conditional updating: $\lambda|\psi \sim \lambda|\psi,y \sim N(\mu_{\lambda|\psi},\sigma^2_{\lambda|\psi})$, where $\mu_{\lambda|\psi}=\omega^2\psi[\sigma^2+\tau^2]^{-1}$ and $\sigma^2_{\lambda|\psi}=\omega^2\tau^2[\sigma^2+\tau^2]^{-1}$. Furthermore, Proposition 3 implies the data are marginally uninformative for $\nu=\Phi[(\lambda-\mu_{\lambda|\psi})/\sigma_{\lambda|\psi}]$ and one-to-one transformations.

This example also demonstrates the importance in Proposition 1–3 of working in terms of a parameterization involving an identified parameter ψ and a nonidentified parameter λ. In terms of β and λ, the prior distribution $\beta \sim N(0,\tau^2)$ is marginally updated to $\beta|y \sim N(\tau^2 y,\tau^2[\sigma^2+\omega^2][\sigma^2+\tau^2+\omega^2]^{-1})$. The analog of Proposition 2 does not hold: $\lambda|\beta \sim \lambda \sim N(0,\omega^2)$ is updated to $\lambda|\beta,y \sim N(\omega^2[y-\beta][\sigma^2+\omega^2]^{-1},\sigma^2\omega^2[\sigma^2+\omega^2]^{-1})$. Nor does the analog of Proposition 3 hold: $\nu=F_{\lambda|\beta}(\lambda)=\Phi(\lambda/\omega)$, and as was seen previously, there is marginal updating for the one-to-one transformation $\omega\Phi^{-1}(\nu)=\lambda$.

3.2. Hierarchical Models

Next consider the following simple hierarchical specification for the observable variable y, the unobserved parameters ψ and λ, and the known constants σ_1^2 and σ_2^2: $y|\psi,\lambda \sim N(\psi,\sigma_1^2)$, $\psi|\lambda \sim N(\lambda,\sigma_2^2)$, and $\lambda \sim N(0,1)$. Equivalently, $[y,\psi,\lambda]' \sim N_3(0,\Sigma)$, where $\Psi=\Lambda=\mathcal{R}$, $\Theta=\mathcal{R}^2$, and

$$\Sigma=\begin{bmatrix}\sigma_1^2+\sigma_2^2+1 & \sigma_2^2+1 & 1\\ \sigma_2^2+1 & \sigma_2^2+1 & 1\\ 1 & 1 & 1\end{bmatrix}. \tag{8}$$

Treating the first stage as the likelihood, i.e., $\mathcal{L}_*(\psi;y)=\phi(y|\psi,\sigma_1^2)$, it is clear that ψ is identified and λ is nonidentified. Given ψ, the data y are conditionally uninformative for λ because $f(\lambda|\psi)=f(\lambda|\psi,y)=\phi(\lambda|\psi/[\sigma_1^2+1],\sigma_2^2/[\sigma_2^2+1])$. The data are marginally informative for λ, however, because the prior $\lambda \sim N(0,1)$ is updated to $\lambda|y \sim N(y/[\sigma_1^2+\sigma_2^2+1],\sigma_2^2/[\sigma_2^2+1])$. Proposition 3 implies the data are marginally uninformative for

$$\nu=\Phi\left(\frac{\lambda-\left(\frac{\psi}{\sigma_2^2+1}\right)}{\left(\frac{\sigma_2^2}{\sigma_2^2+1}\right)^{1/2}}\right) \tag{9}$$

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

and any known one-to-one transformations thereof. For example, the data are marginally uninformative for the standardized residual $\Phi^{-1}(\nu)$ from the regression of λ on ψ.

Combining the first two stages and marginalizing with respect to ψ, $y|\lambda \sim N(\lambda, \sigma_1^2 + \sigma_2^2)$, providing the marginal likelihood $\mathcal{L}(\lambda; y) = \phi(y|\lambda, \sigma_1^2 + \sigma_2^2)$ in which λ is identified. This example demonstrates that "identification" is a property of the likelihood, and ambiguity over the appropriate likelihood (see Bayarri, DeGroot, and Kadane, 1988) implies some ambiguity over identification. Note, however, there is no ambiguity to a Bayesian using all three levels of the hierarchy: the same marginal posterior distribution for λ is obtained regardless of whether the first stage is considered the likelihood and the last two stages the prior, or whether the first two stages are combined to yield a marginal likelihood and the third stage is the prior.

3.3. Multicollinearity

Consider the normal linear regression model (with unit variance for simplicity) $y|\beta \sim N_T(X\beta, I_T)$, where y is $T \times 1$, $X = [X_1, X_2]$ is a fixed $T \times K$ matrix, X_1 is $T \times (K-1)$, X_2 is $T \times 1$, $\beta = [\beta_1', \beta_2]' \in \mathcal{R}^K$ is partitioned analogous to X, and I_T is the $T \times T$ identity matrix. Suppose $\text{rank}(X) = K - 1$ and $X_2 = X_1 d$, where d is a given $(K-1) \times 1$ vector. Then β is not identified. Consider

$$X\beta = X_1\beta_1 + X_2\beta_2 = X_1(\beta_1 + \beta_2 d) = X_1\psi, \tag{10}$$

where $\psi \equiv \beta_1 + d\beta_2$ is identified. Among known one-to-one transformations from β to $\theta = [\psi', \lambda]'$, where $\Psi = \mathcal{R}^{K-1}$, $\Lambda = \mathcal{R}$, and $\Theta = \Psi \times \Lambda = \mathcal{R}^K$, there are many choices for the nonidentified component λ. For example, for any $K \times 1$ vector r of constants, $r'\beta$ is nonidentified iff $r'd \neq 0$ (see, e.g., Leamer, 1978, pp. 188–189). For example, $\lambda \equiv \beta_2$ is not identified, nor is any component of β_1 for which the corresponding component of d is nonzero. The likelihood function $\mathcal{L}(\psi, \lambda; y) = \mathcal{L}_*(\psi; y) = (2\pi)^{-T/2} \exp[-\frac{1}{2}(y - X_1\psi)'(y - X_1\psi)]$ depends *only* on ψ.

Suppose the prior distribution for β is a multivariate normal distribution with mean $\underline{b}$ and covariance matrix $\underline{Q}$, denoted $f(\beta) = \phi_K(\beta|\underline{b}, \underline{Q})$. Then the posterior density of β is $\phi_K(\beta|\bar{b}, \bar{Q})$, where $\bar{b} = \bar{Q}(\underline{Q}^{-1}\underline{b} + X'y)$ and $\bar{Q} = (\underline{Q}^{-1} + X'X)^{-1}$. Because the prior and posterior distributions of the nonidentified β are different, "learning" has occurred. As many authors have noted (e.g., Leamer, 1978, pp. 172–173; Malinvaud, 1970, pp. 246–249; Poirier, 1995, p. 558), however, no learning occurs for $\alpha \equiv c'\underline{Q}^{-1}\beta$, where $c = [d', -1]'$. The prior and posterior densities of α are both identically $\phi(\alpha|c'\underline{Q}^{-1}\underline{b}, c'\underline{Q}^{-1}c)$.

To help clarify matters, let

$$\theta = [\psi', \lambda]' = A\beta, \qquad A = \begin{bmatrix} I_{K-1} & d \\ 0_{K-1}' & 1 \end{bmatrix}, \qquad \mu = A\underline{b} = [\underline{b}_1 + d\underline{b}_2, \underline{b}_2]',$$

and

$$\Sigma = A\underline{Q}A' = \begin{bmatrix} \Sigma_{\psi\psi} & \Sigma_{\psi\lambda} \\ \Sigma_{\psi\lambda}' & \Sigma_{\lambda\lambda} \end{bmatrix}.$$

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

The prior density of θ is $f(\theta) = \phi_K(\theta|\mu,\Sigma)$. In general ψ and λ are dependent, and although λ is not identified, the data are marginally informative for it. However, ψ and α are a priori independent, because they are jointly normal and $\mathrm{Cov}(\psi,\alpha) = \mathrm{Cov}([I_{K-1},d]\beta,[d',-1]\underline{Q}^{-1}\beta) = [I_{K-1},d]\underline{Q}\underline{Q}^{-1}(d',-1]' = 0_{K-1}$. Therefore, the data are marginally uninformative for α.[4]

The dependency of α on the prior is potentially useful as an aid in prior specification. For example, suppose $K = 2$, $d = 1$, and

$$\underline{Q} = \begin{bmatrix} 1 & 2\rho \\ 2\rho & 4 \end{bmatrix}, \tag{11}$$

where ρ is the prior correlation between β_1 and β_2. Then $\psi = \beta_1 + \beta_2$ is identified. The data are marginally uninformative about $\alpha = c'\underline{Q}^{-1}\beta = [(4+2\rho)-(1+2\rho)]/[4(1-\rho^2)]$. The standard way of choosing ρ is explicit self-interrogation for the prior correlation between β_1 and β_2, or possibly, using the predictive distribution of y given x. Sometimes, ρ can be chosen implicitly by self-interrogation about the quantity α for which the experiment is uninformative. For example, if the experiment is believed to be uninformative for $\alpha = \beta_1 - \beta_2$, then $\rho = 0$; alternatively, if the experiment is uninformative for $\alpha = \beta_1$, then $\rho = -\frac{1}{2}$. The more questions available for interrogation in prior elicitation, the better served is the researcher.

A moral to draw from this example is that the quantity α, about which the data are uninformative, depends on the prior state of information (through $\underline{Q}^{-1}$) *and* the nature of the collinearity (through c). Whether multicollinearity is a problem in a specific context depends on whether α is of any particular interest (a point emphasized in Poirier, 1995, Sect. 9.14).

3.4. Errors in Variables

A classic topic in econometrics is errors in variables, and it has played a prominent role in discussions of identification. Its history is long, and Bayesian work on it has been especially enlightening (e.g., Erickson, 1989; Florens, Mouchart, and Richard, 1974; Leamer, 1978, Ch. 7; Lindley and Sayyad, 1968; Zellner, 1971, Ch. 5). The topic also provides a good example for discussing the meaning of data being uninformative.

Consider T observations on the linear relationship $y_t = \lambda\tilde{x}_t + \varepsilon_t (t = 1,2,\ldots,T)$. Assume $\tilde{x}_t|\sigma_1^2,\sigma_2^2,\sigma_3^2,\lambda \sim N(0,\sigma_1^2)$, $x_t|\tilde{x}_t,\sigma_1^2,\sigma_2^2,\sigma_3^2,\lambda \sim N(\tilde{x}_t,\sigma_2^2)$, and $y_t|x_t,\tilde{x}_t,\sigma_1^2,\sigma_2^2,\sigma_3^2,\lambda \sim N(\lambda\tilde{x}_t,\sigma_3^2)$ for $t = 1,2,\ldots,T$. Then the likelihood is

$$\mathcal{L}(\sigma_1^2,\sigma_2^2,\sigma_3^2,\lambda;x,y) = \phi_{2T}\left(\begin{bmatrix} x \\ y \end{bmatrix}\middle| 0_{2T}, I_T \otimes \begin{bmatrix} \sigma_1^2+\sigma_2^2 & \lambda\sigma_1^2 \\ \lambda\sigma_1^2 & \lambda^2\sigma_1^2+\sigma_3^2 \end{bmatrix}\right), \tag{12}$$

where $x = [x_1,x_2,\ldots,x_T]'$ and $y = [y_1,y_2,\ldots,y_T]'$ are observed. Reparameterizing in terms of $\psi_1 = \sigma_1^2+\sigma_2^2$, $\psi_2 = \lambda\sigma_1^2$, $\psi_3 = \lambda^2\sigma_1^2+\sigma_3^2$, and λ yields

$$\mathcal{L}_*(\psi_1,\psi_2,\psi_3;x,y) = \phi_{2T}\left(\begin{bmatrix} x \\ y \end{bmatrix}\middle| 0_{2T}, I_T \otimes \begin{bmatrix} \psi_1 & \psi_2 \\ \psi_2 & \psi_3 \end{bmatrix}\right), \tag{13}$$

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

where $\Psi = \{\psi | \psi_1 > 0, \psi_3 > 0, \psi_1\psi_3 - \psi_2^2 > 0\}$ and $\Lambda = \mathcal{R}$. The parameter vector $\psi = [\psi_1, \psi_2, \psi_3]'$ is identified; λ is nonidentified.

Most Bayesian analyses of the errors-in-variables model have employed improper "noninformative" priors. Consider an informative proper prior density $f(\psi)$ (e.g., Wishart), defined over $\psi \in \Psi$, i.e., values for which the covariance matrix in (13) is positive definite. Positive definiteness implies that λ must lie between the direct (y on x) and the reciprocal of the indirect (x on y) population regression slopes, where the sign of ψ_2 determines which slope is the lower bound and which slope is the upper bound (see (14), which follows). An obvious candidate for a proper noninformative conditional prior density for λ given ψ is a uniform density over these respective intervals:[5]

$$f(\lambda|\psi) = \left\{ \begin{array}{lll} \left(\dfrac{\psi_3}{\psi_2} - \dfrac{\psi_2}{\psi_1}\right)^{-1}, & \text{if } \psi_2 > 0 \quad \text{and} \quad \dfrac{\psi_2}{\psi_1} < \lambda < \dfrac{\psi_3}{\psi_2} \\ \left(\dfrac{\psi_2}{\psi_1} - \dfrac{\psi_3}{\psi_2}\right)^{-1}, & \text{if } \psi_2 < 0 \quad \text{and} \quad \dfrac{\psi_3}{\psi_2} < \lambda < \dfrac{\psi_2}{\psi_1} \\ 0, & \text{otherwise} \end{array} \right\}. \tag{14}$$

In other words, λ given ψ is uniform over $\Lambda(\psi) = \{\lambda | \psi_2/\psi_1 < \lambda < \psi_3/\psi_2$ if $\psi_2 > 0$, and $\psi_3/\psi_2 < \lambda < \psi_2/\psi_1$ if $\psi_2 < 0\}$. Coherency demands $\Lambda(\psi)$ reflects these bounds. The uniformity in (14) is not mandatory (e.g., it could be generalized to a beta distribution), but uniformity is an interesting base case. The fact that the data are conditionally uninformative for λ implies that posterior density $f(\lambda|\psi, y)$ is also given by (14). Clearly, λ and ψ are a priori dependent [$\Lambda(\psi)$ depends on ψ], and the data can be shown to be marginally informative for λ. The marginal prior and posterior distributions for λ, however, are fairly complicated.

Proposition 3 implies that the data are marginally uninformative for

$$\nu = \left\{ \begin{array}{ll} \lambda - \dfrac{\psi_2}{\psi_1}, & \text{if } \psi_2 > 0 \\ \lambda - \dfrac{\psi_3}{\psi_2}, & \text{if } \psi_2 < 0 \end{array} \right\}. \tag{15}$$

In other words, given ψ we cannot learn where in the interval the nonidentified parameter λ lies. Also note in contrast to conditional density (14), the support of the marginal distribution of λ, both prior and posterior, may be the whole real line $\mathcal{R}$.

3.5. Censored Sampling

A special case of the censored sampling model of Manski (1994, pp. 22–25) is given by a binary random variable Z_t such that $\text{Prob}(Z_t = 1) = \psi_1$ and $\text{Prob}(Z_t = 0) = 1 - \psi_1$, and the continuous random variable y_t with conditional densities

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

$f(y_t|z_t = 1) = \phi(y_t|\psi_2,1)$ and $f(y_t|z_t = 0) = \phi(y_t|\mu,1)$. The "censoring" is reflected in the assumption that although $Z_t = z_t$ is always observed, y_t is observed iff $z_t = 1$. To put matters into the Manski framework, reparameterize into ψ_1, ψ_2, and

$$\lambda \equiv P(y_t \geq 0) = \psi_1 \Phi(\psi_2) + [1 - \psi_1]\Phi(\mu). \tag{16}$$

Then the likelihood function is

$$\mathcal{L}_*(\psi; y) = \prod_{t=1}^{T} [\psi_1 \phi(y_t|\psi_2,1)]^{z_t}[1 - \psi_1]^{1-z_t}, \tag{17}$$

where $\psi = [\psi_1,\psi_2]'$. Clearly, neither μ nor λ is identified. However, ψ is identified. Although $\Psi = [0,1] \times \mathcal{R}$ and $\Lambda = [0,1]$, λ in (16) has the restricted (given ψ) parameter space $\Lambda(\psi) = \{\lambda|\psi_1\Phi(\psi_2) \leq \lambda \leq \psi_1\Phi(\psi_2) + 1 - \psi_1\}$ because $0 \leq \Phi(\mu) \leq 1$. Clearly, prior beliefs on λ and ψ must be dependent, and so the data are potentially marginally informative for λ.

In effect the point of Manski is that a nonidentified parameter like μ can be reparameterized into λ (depending on the identified ψ and the nonidentified μ) such that the bounded support of λ depends on ψ. Consistent estimators of ψ therefore provide asymptotically valid bounds for the nonidentified λ. This point seems self-evident. Rather than dwell on it, we proceed to exact finite-sample posterior analysis.

Reminiscent of (14), suppose the prior conditional distribution for λ given ψ is uniform over $\Lambda(\psi)$, i.e., $f(\lambda|\psi) = (1 - \psi_1)^{-1}$ if $\lambda \in \Lambda(\psi)$ and zero otherwise. Keeping in mind note 5, the marginal prior density of λ (the probability that y_t is nonnegative, unconditional of observability) is

$$\begin{aligned} f(\lambda) &= \int_{\Psi} (1 - \psi_1)^{-1} \mathbf{1}_{\Lambda(\psi)}(\lambda) f(\psi)\, d\psi \\ &= E_{\psi}[(1 - \psi_1)^{-1} \mathbf{1}_{\Lambda(\psi_1,\psi_2)}(\lambda)], \qquad 0 \leq \lambda \leq 1. \end{aligned} \tag{18}$$

Because $f(\lambda|\psi, y) = f(\lambda|\psi)$, the marginal posterior density for λ can be obtained from (18) by simply replacing $f(\psi)$ with its posterior analog $f(\psi|y)$:

$$\begin{aligned} f(\lambda|y) &= \int_{\Psi} (1 - \psi_1)^{-1} \mathbf{1}_{\Lambda(\psi)}(\lambda) f(\psi|y)\, d\psi \\ &= E_{\psi|y}[(1 - \psi_1)^{-1} \mathbf{1}_{\Lambda(\psi_1,\psi_2)}(\lambda)], \qquad 0 \leq \lambda \leq 1. \end{aligned} \tag{19}$$

The posterior probability that $\lambda \in [a, b] \subseteq \mathcal{R}$, where the endpoints can depend on y, is

$$\begin{aligned} P(a \leq \lambda \leq b|y) &= \int_{\Psi} P(a \leq \lambda \leq b|\psi) f(\psi|y)\, d\psi \\ &= \int_{\Psi} (1 - \psi_1)^{-1} \max[0, \min\{b, \psi_1\Phi(\psi_2) + 1 - \psi_1\} \\ &\quad - \max\{a, \psi_1\Phi(\psi_2)\}] f(\psi|y)\, d\psi. \end{aligned} \tag{20}$$

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

Manski (1994) suggests, as a reasonable estimate for the nonidentified λ, the set $\Lambda(\hat{\psi}) \subseteq [0,1]$, where $\hat{\psi}$ is the maximum likelihood estimate (MLE) of ψ. Whether such bounds should be interpreted as point estimates or interval estimates is an old matter of controversy even in identified cases (see Engle, 1976; Genberg, 1972; Maddala, 1976a, 1976b; Zellner, 1972, 1976a, 1976b). Given a loss function, Bayesian point estimation is straightforward using marginal posterior density (19). The highest posterior density interval of level $1 - \alpha$ can be obtained from (19) in the usual fashion, yielding choices for a and b in (20). In contrast, $\Lambda(\hat{\psi})$ appears motivated from consideration of conditional density $f(\lambda|\psi,y) = f(\lambda|\psi)$, not marginal density (19). The posterior conditional and marginal distributions can be quite different: the former has support $\Lambda(\psi)$, and the latter has support $[0,1]$.

The posterior probability content of $\Lambda(\hat{\psi})$ is unclear in general. To gain some insight, assume that the censoring probability $1 - \psi_1$ is known. This implies the likelihood simplifies to $\mathcal{L}_*(\psi_2;\underline{y}) \propto \prod_{t=1}^{T}[\phi(y_t|\psi_2,1)]^{z_t}$, the length $1 - \psi_1$ of $\Lambda(\psi_1,\hat{\psi}_2)$ is known, and the location is determined by $\hat{\psi}_2$. High censoring yields a broad width for the interval $\Lambda(\psi_1,\hat{\psi}_2)$. Ceteris paribus, $\Lambda(\psi_1,\hat{\psi}_2)$ has greater posterior probability content in the face of high censoring, but its informativeness is lessened because of its width. Ironically, as the censoring disappears (i.e., $\psi_1 \to 1$) and λ becomes identified, $\Lambda(\psi)$ approaches the point $\psi_1\Phi(\psi_2)$, and the posterior probability content of $\Lambda(\psi_1,\hat{\psi}_2)$ is zero. This behavior underlines the ambiguity of whether $\Lambda(\psi_1,\hat{\psi}_2)$ is motivated as a point or as an interval estimate.

With ψ_1 known, densities (18) and (19) simplify somewhat. Furthermore, with a few algebraic manipulations, (20) implies that the posterior probability content of $\Lambda(\psi_1,\hat{\psi}_2)$ is

$$p \equiv P[\lambda \in \Lambda(\psi_1,\hat{\psi}_2)]$$

$$= F_{\psi_2|y}(u) - F_{\psi_2|y}(\ell) - \left(\frac{\psi_1}{1-\psi_1}\right)\int_{\ell}^{u} |\Phi(\hat{\psi}_2) - \Phi(\psi_2)|\, f(\hat{\psi}_2|y)\, d\psi_2, \qquad \textbf{(21)}$$

where $\ell = \Phi^{-1}[\max\{0, \Phi(\hat{\psi}_2) - \psi_1^{-1} + 1\}]$, $u = \Phi^{-1}[\min\{1, \Phi(\hat{\psi}_2) + \psi_1^{-1} - 1\}]$, and $F_{\psi_2|y}(\cdot)$ denotes the marginal posterior c.d.f. of ψ_2. The following development focuses on prior density (18), but similar expressions can be obtained for posterior density (19) by replacing the marginal prior distribution of ψ_2 in the following expressions by the marginal posterior distribution of ψ_2. The integration in (18) can be expressed in terms of the prior c.d.f. $F_{\psi_2}(\cdot)$ by

$$f(\lambda) = (1-\psi_1)^{-1}(F_{\psi_2}[\Phi^{-1}(\min\{1, \lambda/\psi_1\})]$$

$$- F_{\psi_2}[\max\{0, (\lambda - 1 + \psi_1)/\psi_1\}]), \qquad \textbf{(22)}$$

for $0 \le \lambda \le 1$. The nature of prior density (18) for λ depends crucially on the sign of $\psi_1 - \frac{1}{2}$. If $\psi_1 < \frac{1}{2}$ (i.e., a censoring probability $1 - \psi_1 > \frac{1}{2}$), then

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

$$f(\lambda) = (1-\psi_1)^{-1}\begin{Bmatrix} F_{\psi_2}\left[\Phi^{-1}\left(\dfrac{\lambda}{\psi_1}\right)\right], & \text{if } 0 \le \lambda < \psi_1 \\ 1, & \text{if } \psi_1 \le \lambda \le 1-\psi_1 \\ 1 - F_{\psi_2}\left[\Phi^{-1}\left(\dfrac{\lambda+\psi_1-1}{\psi_1}\right)\right], & \text{if } 1-\psi_1 < \lambda \le 1 \end{Bmatrix}. \tag{23}$$

On the other hand, if $\psi_1 > \frac{1}{2}$ (i.e., a censoring probability $1-\psi_1 < \frac{1}{2}$), then

$$f(\lambda) = (1-\psi_1)^{-1}\begin{Bmatrix} F_{\psi_2}\left[\Phi^{-1}\left(\dfrac{\lambda}{\psi_1}\right)\right], & \text{if } 0 \le \lambda < 1-\psi_1 \\ F_{\psi_2}\left[\Phi^{-1}\left(\dfrac{\lambda}{\psi_1}\right)\right] - F_{\psi_2}\left[\Phi^{-1}\left(\dfrac{\lambda+\psi_1-1}{\psi_1}\right)\right], & \text{if } 1-\psi_1 \le \lambda \le \psi_1 \\ 1 - F_{\psi_2}\left[\Phi^{-1}\left(\dfrac{\lambda+\psi_1-1}{\psi_1}\right)\right], & \text{if } \psi_1 < \lambda \le 1 \end{Bmatrix}. \tag{24}$$

When $\psi_1 = \frac{1}{2}$, then the middle ranges for λ in (25) and (26) shrink to the point $\lambda = \frac{1}{2}$, with $f(\frac{1}{2}) = 2$, and the prior density simplifies to

$$f(\lambda) = \begin{Bmatrix} 2F_{\psi_2}[\Phi^{-1}(2\lambda)], & \text{if } 0 \le \lambda < \frac{1}{2} \\ 2(1 - F_{\psi_2}[\Phi^{-1}(2\lambda-1)]), & \text{if } \frac{1}{2} \le \lambda \le 1 \end{Bmatrix}. \tag{25}$$

The flat segment in (23), when $\psi_1 < \frac{1}{2}$ (high censoring) is interesting. It is of length $1-2\psi_1$ and the fraction $(1-2\psi_1)/(1-\psi_1)$ of the probability mass is contained below it. Furthermore, this flat segment remains unchanged in the posterior, the data affecting only the tails of $f(\lambda|y)$. To see how this flat segment arises, consider Figure 1, which plots the lower and upper bounds of the support $\Lambda(\psi)$ as functions of ψ_2 for $\psi_1 = .40$, .50, and .95. The area between these bounds is the support for the joint density $f(\psi_2, \lambda) = (1-\psi_1)^{-1}\mathbf{1}_{\Lambda(\psi_1,\psi_2)}(\lambda)f(\psi_2)$. Iff $\psi_1 < \frac{1}{2}$, there are some values of λ for which $\mathbf{1}_{\Lambda(\psi_1\cdot\psi_2)}(\lambda) = 1$ for all ψ_2. Therefore, integrating prior density $f(\psi_2, \lambda)$ with respect to ψ_2 (to obtain the marginal prior density of λ) involves merely integrating $f(\psi_2)$ over $\mathcal{R}$, obtaining unity, and yielding $f(\lambda) = (1-\psi_1)^{-1}$ for these values of λ.

To be more specific, suppose the prior distribution for ψ_2 is normal with mean zero and variance .2, implying the posterior density $f(\psi_2|y,z) = \phi(\psi_2|T_1\bar{y}(T_1+5)^{-1},(T_1+5)^{-1})$, where $\bar{y}$ is the sample mean based on $T_1 = z_1 + z_2 + \cdots + z_T$ observations. Figure 2 shows prior density $f(\psi_2, \lambda)$ in this case with observability probability $\psi_1 = .40$. Figure 3 shows the corresponding prior density for λ (flat between .4 and .6) together with $f(\lambda|y)$ in the case $T_1 = 25$ and $\bar{y}$ equaling $-.5$, 0, and .5, respectively. Figures 4 and 5 are the analogous graphs for $\psi_1 = .50$ and $\psi_1 = .95$. Note in Figure 3 that the behavior of the tails of the posterior has the effect of lengthening the flat segment as

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

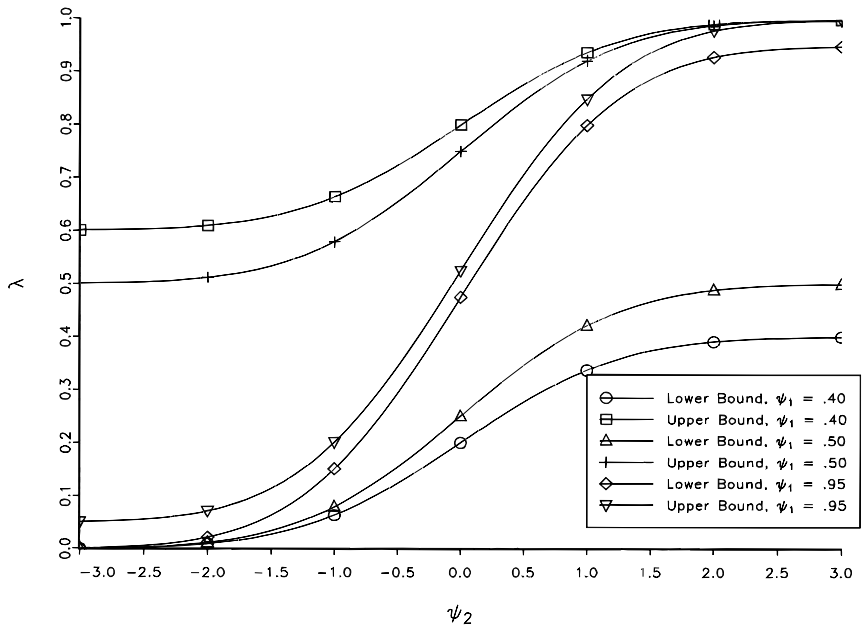


FIGURE 1. Joint support for ψ_2 and λ, $\psi_1 = .40, .50, .95$.

mass shifts in the direction of $\bar{y}$. In Figure 4 a flat segment is in effect created in the posterior. Finally, in Figure 5 with $\psi_1 = .95$, the posterior is distinctly much less dispersed because there is less censoring. On the downside, however, the posterior content of $\Lambda(\psi_1, \hat{\psi}_2)$ is quite variable (see Figure 6) as a result of its minimal length of $1 - \psi_1 = .05$.

Given the uniform conditional prior (21), the data are marginally uninformative for $\nu = [\lambda - \psi_1\psi_2)]/[1 - \psi_1] = \Phi(\mu)$ and known one-to-one transformations of it. Therefore, as in the errors-in-variables problem of Section 3.3, given ψ_1 and ψ_2 we cannot learn where in the interval $\Lambda(\psi)$ the nonidentified parameter λ lies.

3.6. Bernoulli Model with Classification Error[6]

Given $\gamma \in [0,1]$ and $\lambda \in [0, \frac{1}{2})$, consider the latent independent and identically distributed (i.i.d.) Bernoulli observations $\tilde{y}_t (t = 1,2,\ldots,T)$ such that $\text{Prob}(\tilde{y}_t = 1|\gamma,\lambda) = \gamma$ and $\text{Prob}(\tilde{y}_t = 0|\gamma,\lambda) = 1 - \gamma$. The observed random variables $y_t (t = 1,2,\ldots,T)$ are related to these latent variables by $\text{Prob}(y_t = 1|\tilde{y}_t = 1,\gamma,\lambda) = 1 - \lambda$, $\text{Prob}(y_t = 0|\tilde{y}_t = 1,\gamma,\lambda) = \lambda$, $\text{Prob}(y_t = 1|\tilde{y}_t = 0,\gamma,\lambda) = \lambda$, and $\text{Prob}(y_t = 0|\tilde{y}_t = 0,\gamma,\lambda) = 1 - \lambda$. Thus, λ represents the probability of misclassification. The marginal sampling probabilities for the observed data are $\text{Prob}(y_t = 1|\gamma,\lambda) = \psi$ and $\text{Prob}(y_t = 0|\gamma,\lambda) = 1 - \psi$, where $\psi = \gamma(1 - \lambda) + (1 - \gamma)\lambda$. In other words, the observed data are Bernoulli with probabil-

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

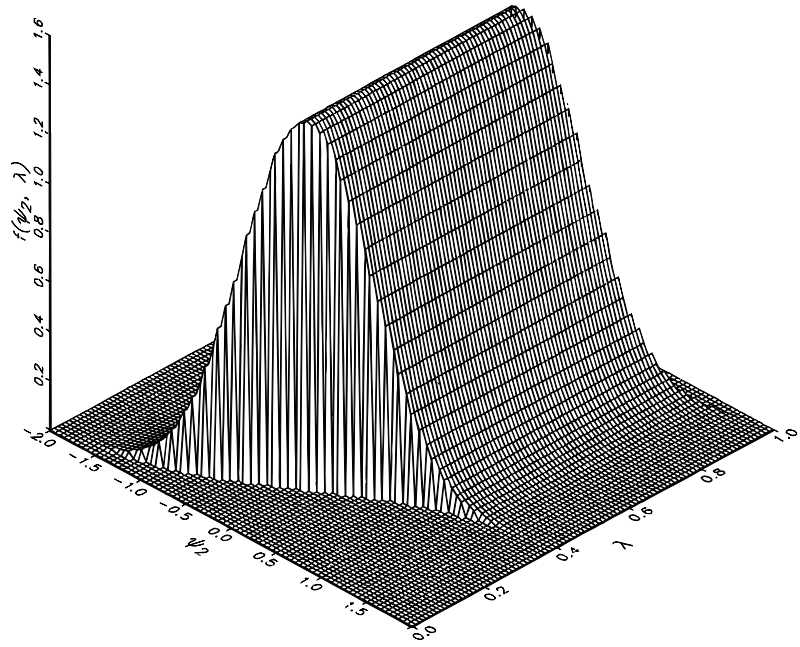


FIGURE 2. Bivariate prior density $f(\psi_2, \lambda) = (1 - .4)^{-1}\mathbf{1}_{\Lambda(.4,\psi_2)}(\lambda)\phi(\psi_2|0, .2)$.

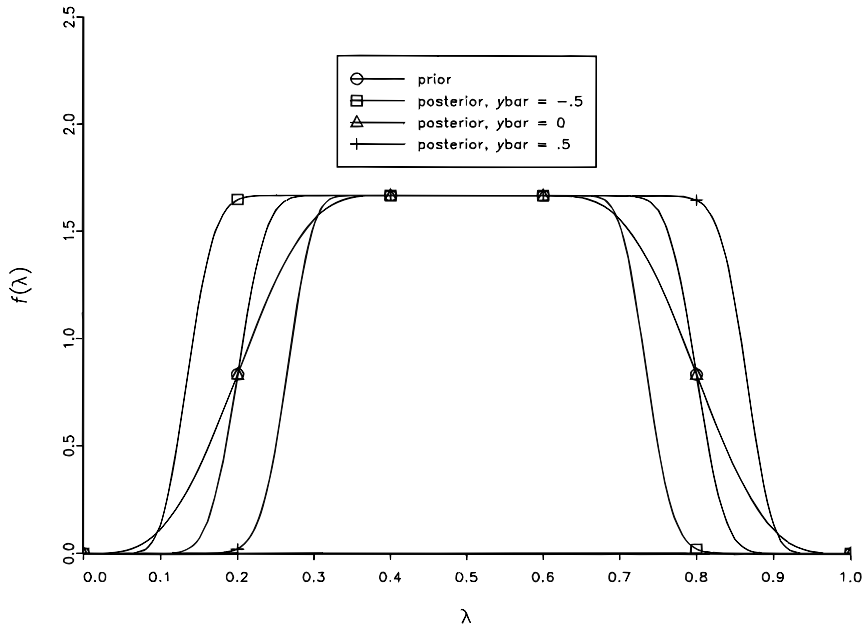


FIGURE 3. Posterior and prior densities for λ; $\psi_1 = .40$, $T_1 = 25$, $\bar{y} = -.5, 0, .5$.

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

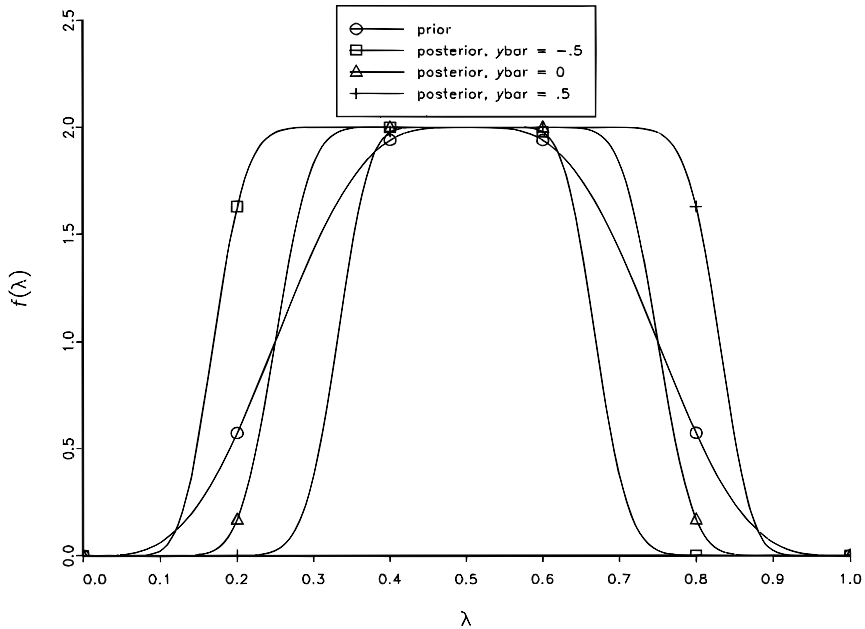


FIGURE 4. Posterior and prior densities for λ; $\psi_1 = .50$, $T_1 = 25$, $\bar{y} = -.5, 0, .5$.

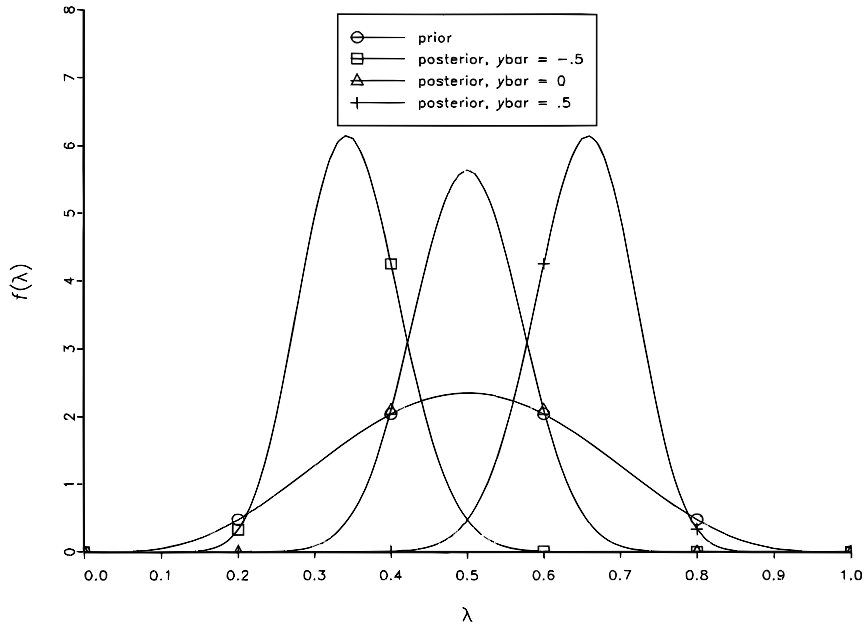


FIGURE 5. Posterior and prior densities for λ; $\psi_1 = .95$, $T_1 = 25$, $\bar{y} = -.5, 0, .5$.

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

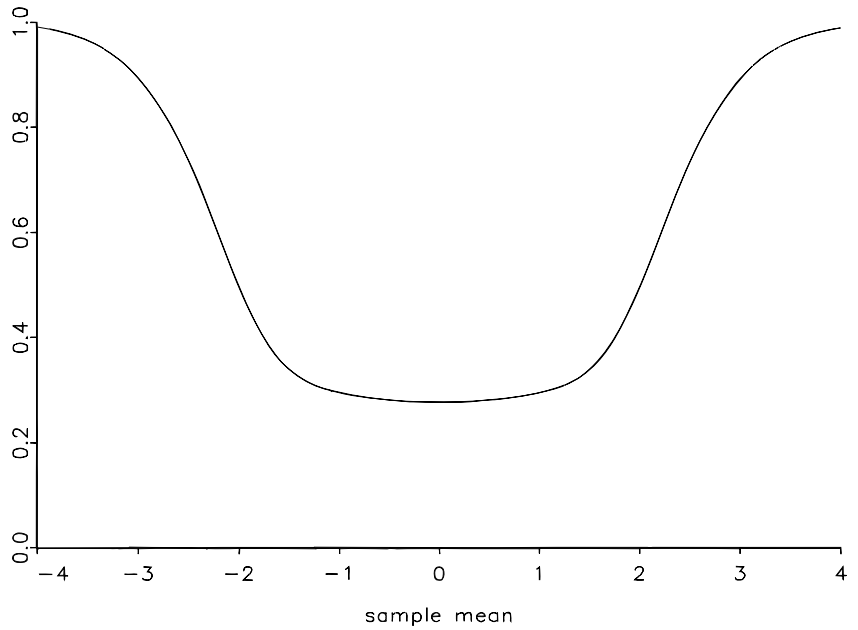


FIGURE 6. Posterior probability content of $\Lambda(.95, \bar{y})$, $T_1 = 25$.

ity of a success ψ, and the latent data are Bernoulli with probability of a success γ. Given a sample of r "1's" and $N - r$ "0's," the likelihood is

$$\mathcal{L}(\gamma, \lambda; y) = \mathcal{L}_*(\psi; y) = \psi^r (1 - \psi)^{N-r}. \tag{26}$$

Even with the restriction of $\lambda \in [0, \frac{1}{2})$, neither γ nor λ is identified, but ψ is identified.

Given a prior density $f_{\gamma,\lambda}(\gamma, \lambda)$ over $[0,1] \times [0, \frac{1}{2})$, the implied joint prior distribution for $\Theta = [\psi, \lambda]'$ is $f(\psi, \lambda) = (1 - 2\lambda)^{-1} f_{\gamma,\lambda}([(\psi - \lambda)/(1 - 2\lambda)], \lambda)$ over the triangular region $\Theta = \{[\psi, \lambda]' | 0 < \psi \leq \frac{1}{2}, 0 \leq \lambda \leq \psi$ or $\frac{1}{2} < \psi < 1, 0 \leq \lambda < 1 - \psi\}$ shown in Figure 7. Other parameter spaces of interest are the conditional supports $\Psi(\lambda) = \{\psi | \lambda < \psi < 1 - \lambda\}$ and $\Lambda(\psi) = \{\lambda | 0 \leq \lambda < \psi$, if $0 < \psi \leq \frac{1}{2}$, and $0 \leq \lambda < 1 - \psi$, if $\frac{1}{2} < \psi < 1\}$, and the unconditional supports $\Psi = \{\psi | 0 < \psi < 1\}$ and $\Lambda = \{\lambda | 0 \leq \lambda < \frac{1}{2}\}$. The marginal density of ψ is

$$f(\psi) = \left\{ \begin{array}{ll} \displaystyle\int_0^{\psi} (1 - 2\lambda)^{-1} f_{\gamma,\lambda}\left(\left[\frac{\psi - \lambda}{1 - 2\lambda}\right], \lambda\right) d\lambda, & \text{if } 0 < \psi \leq \frac{1}{2} \\ \displaystyle\int_0^{1-\psi} (1 - 2\lambda)^{-1} f_{\gamma,\lambda}\left(\left[\frac{\psi - \lambda}{1 - 2\lambda}\right], \lambda\right) d\lambda, & \text{if } \frac{1}{2} < \psi < 1 \end{array} \right\}. \tag{27}$$

The data are marginally informative for γ, λ, and ψ.

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

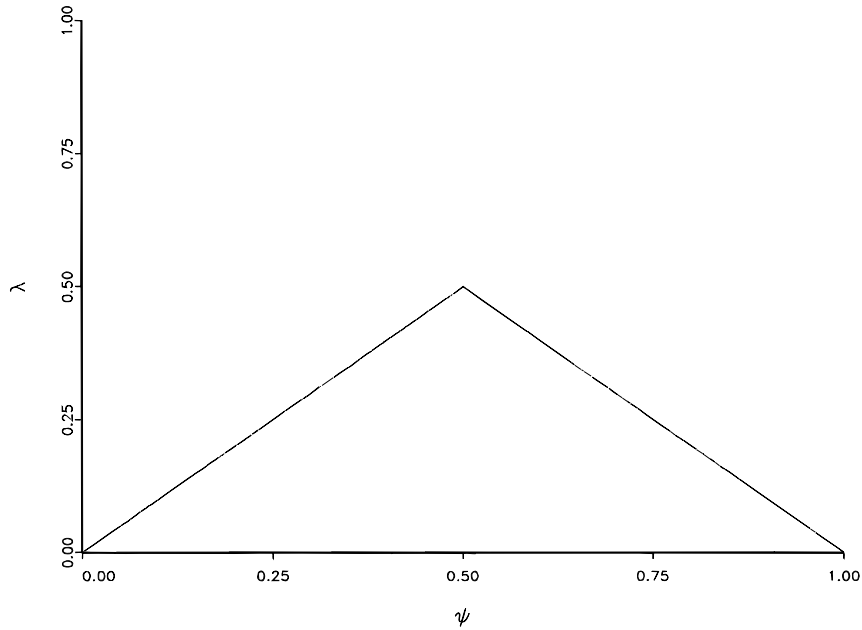


FIGURE 7. Prior support for Θ in Example 3.5.

To demonstrate Proposition 3, suppose γ and λ are a priori independent, γ is uniformly distributed over $[0,1]$, and λ has the following beta density over $[0,\frac{1}{2})$: $f_\lambda(\lambda) = [2\Gamma(a+1)/\Gamma(a)](1-2\lambda)^{a-1}$, given $a \in \mathcal{R}^+$, where $E(\lambda|a) = 1/(1+a)$ (see Figure 8 for $a = 1,2,3$]. Then the joint p.d.f. is $f(\psi,\lambda) = [2\Gamma(a+1)/\Gamma(a)](1-2\lambda)^{a-2}$, and the conditional p.d.f. of λ given ψ is

$$f(\lambda|\psi) = \left\{ \begin{array}{ll} \dfrac{2(1-a)(1-2\lambda)^{a-2}}{(1-2\psi)^{a-1}-1}, & \text{if } 0 < \psi \le \frac{1}{2} \quad \text{and} \quad 0 \le \lambda < \psi \\ \dfrac{2(1-a)(1-2\lambda)^{a-2}}{(2\psi-1)^{a-1}-1}, & \text{if } \frac{1}{2} < \psi < 1 \quad \text{and} \quad 0 \le \lambda < 1-\psi \end{array} \right\} \tag{28}$$

if $a \neq 1$, and

$$f(\lambda|\psi) = \left\{ \begin{array}{ll} \dfrac{-2}{(1-2\lambda)\ln(1-2\psi)}, & \text{if } 0 < \psi \le \frac{1}{2} \quad \text{and} \quad 0 \le \lambda < \psi \\ \dfrac{-2}{(1-2\lambda)\ln(2\psi-1)}, & \text{if } \frac{1}{2} < \psi < 1 \quad \text{and} \quad 0 \le \lambda < 1-\psi \end{array} \right\} \tag{29}$$

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

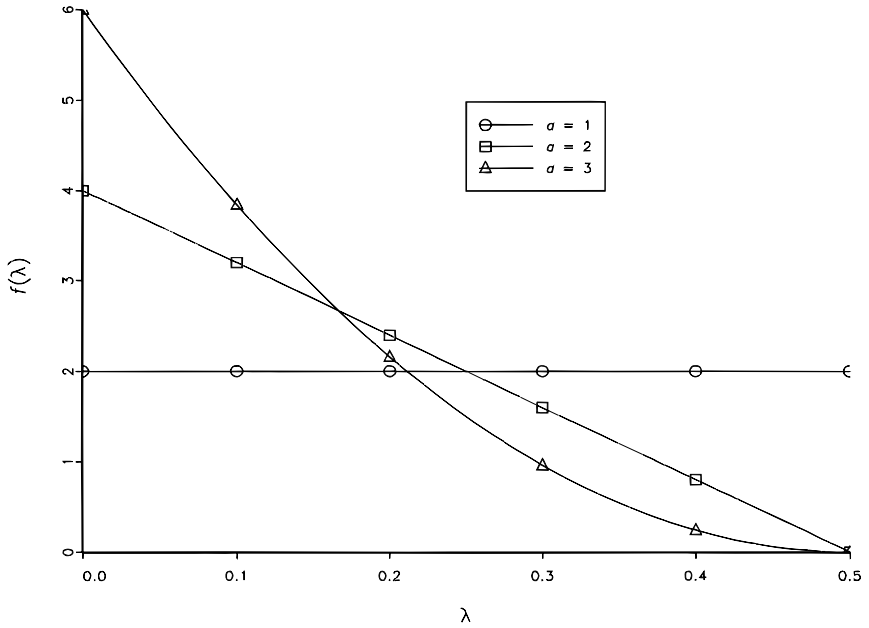


FIGURE 8. Prior marginal density for λ ($a = 1,2,3$) in Example 3.5.

if $a = 1$. If $a \neq 1$, then Proposition 3 implies the data are marginally uninformative for

$$\nu = \left\{ \begin{array}{ll} \dfrac{(1-2\lambda)^{a-1}-1}{(1-2\psi)^{a-1}-1}, & \text{if } 0 < \psi \leq \tfrac{1}{2} \quad \text{and} \quad 0 \leq \lambda < \psi \\ \dfrac{(1-2\lambda)^{a-1}-1}{(2\psi-1)^{a-1}-1}, & \text{if } \tfrac{1}{2} < \psi < 1 \quad \text{and} \quad 0 \leq \lambda < 1-\psi \end{array} \right\}. \tag{30}$$

For example, if $a = 2$, then

$$\nu = \left\{ \begin{array}{ll} \dfrac{\lambda}{\psi}, & \text{if} \quad 0 < \psi \leq \tfrac{1}{2} \\ \dfrac{\lambda}{1-\psi}, & \text{if} \quad \tfrac{1}{2} < \psi < 1 \end{array} \right\}, \tag{31}$$

i.e., the data are marginally uninformative for the probability of misclassification relative to the minimum probability of observability: $\min\{\text{Prob}(y_t = 1|\gamma,\lambda), \text{Prob}(y_t = 0|\gamma,\lambda)\}$.

Note that it is straightforward to extend this example to the multinomial case with the highly structured misclassification probabilities in Harless and Camerer (1994).

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

3.7. Simultaneous Equations

Consider T independent observations from the two-equation reduced form

$$y_{t1} = \pi_1 x_t + u_{t1}, \tag{32}$$

$$y_{t2} = \pi_2 x_t + u_{t2}, \tag{33}$$

where $y_t = [y_{t1}, y_{t2}]'$, x_t is a fixed scalar, $\pi = [\pi_1, \pi_2]'$, $u_t = [u_{t1}, u_{t2}]' | \pi, \Omega \sim N_2(0_2, \Omega)$, and $\Omega = [\omega_{ij}]$. Given $\lambda \in \mathcal{R}$, replace (34) by the structural equation formed by subtracting λ times (35) from (34):

$$y_{t1} = \lambda y_{t2} + (\pi_1 - \lambda\pi_2)x_t + (u_{t1} - \lambda u_{t2}). \tag{34}$$

Here $\psi \equiv [\pi_1, \pi_2, \omega_{11}, \omega_{12}, \omega_{22}]'$ is identified. Unless $\pi_1 - \lambda\pi_2 = 0$, λ is nonidentified. Suppose the prior conditional density of λ given ψ is

$$f(\lambda|\psi) = \phi\left(\lambda \middle| \frac{\pi_1}{\pi_2}, \frac{c\omega_{22}}{\pi_2^2 x'x}\right), \tag{35}$$

where c is a known positive constant. Prior density (37) is centered over the identifying restriction. It also has a variance that is inversely proportional to the noncentrality parameter $\pi_2^2 x'x/\omega_{22}$ of reduced form (35), following the suggestion in Poirier (1996) that the strength of prior beliefs on structural parameters, given reduced form parameters, should increase as the structural parameters become more identified (as measured by the explanatory power of the instrument x in the reduced form equation for y_2). Using Proposition 3, the data are marginally uninformative for

$$\nu = \Phi\left(\frac{\lambda - \dfrac{\pi_1}{\pi_2}}{\sqrt{\dfrac{c\omega_{22}}{\pi_2^2 x'x}}}\right) = \Phi\left(-\frac{\pi_1 - \pi_2\lambda}{\sqrt{\dfrac{c\omega_{22}}{x'x}}}\right) \tag{36}$$

or known one-to-one transformations, such as $[(c\omega_{22})/x'x]^{-1/2}\Phi^{-1}(\nu) = \pi_1 - \pi_2\lambda$, yielding the coefficient of x_t in (36). This corresponds to imposing the identifying restriction stochastically. For further discussion of identification in the context of Bayesian analysis of simultaneous equations, see Dréze and Richard (1983, esp. pp. 522, 537).

3.8. Probit

Suppose $\tilde{y}_t | x_t, \beta, \sigma^{-2} (t = 1, 2, \ldots, T)$ are independent $N(x_t'\beta, \sigma^2)$ latent random variables. If the latent data are available, then β and σ^{-2} are both identified. Suppose, however, only sign information is available on $\tilde{y}_t$, i.e., only $y_t = 1$ if $\tilde{y}_t \geq 0$ and $y_t = 0$ if $\tilde{y}_t < 0$ is observed.[7] Define $\psi = \beta/\sigma$ and $\lambda = \sigma^{-2}$. The likelihood function is

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

$$\mathcal{L}_*(\psi; y) = \prod_{t=1}^{T} [\Phi(x_t'\psi)]^{y_t} [1 - \Phi(x_t'\psi)]^{1-y_t}. \tag{37}$$

Clearly ψ is identified and λ is nonidentified.

Consider the conjugate (for the latent process) normal-gamma prior for β and σ^{-2} (which assumes that β and σ^{-2} are a priori dependent). Then prior beliefs for ψ and λ are also dependent, and the data are marginally informative for λ.[8] This suggests thinking of standard probit analysis as "imposing the normalization $\sigma^{-2} = 1$" is unfortunate: learning about σ^{-2} occurs! In fact, nontrivial posterior odds of $\sigma^{-2} = 1$ versus $\sigma^{-2} \neq 1$ can be computed.

3.9. Nonparametric Multinomial Problem

Let Θ be the unit simplex in $\mathcal{R}^{K+1}$ and suppose a random variable Y follows the sampling mass function

$$P(Y = a_k | \theta_1, \ldots, \theta_{K+1}) = \theta_k \qquad (k = 1,2,\ldots,K+1), \tag{38}$$

where $\theta_k (k = 1,2,\ldots,K+1)$ are unknowns satisfying $0 \leq \theta_k \leq 1$, $\theta_1 + \cdots + \theta_{K+1} = 1$, and $a_k \in \mathcal{R}$ is known $(k = 1,2,\ldots,K+1)$. Given a random sample $y = [y_1, y_2, \ldots, y_T]'$ from (40), let $n_k = \sum_{t=1}^{T} \mathbf{1}(y_t = a_k)$ denote the number of observations equaling a_k $(k = 1,2,\ldots,K+1)$. The likelihood function is

$$\mathcal{L}(\theta; y) \propto \prod_{k=1}^{K+1} \theta_k^{n_k}. \tag{39}$$

A convenient prior for $\theta = [\theta_1, \ldots, \theta_K]'$ is the K-dimensional Dirichlet distribution, denoted $D_K(\underline{b}_1, \ldots, \underline{b}_{K+1})$, with density

$$f(\theta | \underline{b}_1, \ldots, \underline{b}_{K+1}) = c_K(\underline{b}_1, \ldots, \underline{b}_{K+1}) \prod_{k=1}^{K+1} \theta_k^{\underline{b}_k - 1}, \tag{40}$$

where $\underline{b}_k > 0$ $(k = 1,2,\ldots,k+1)$ are known, $\theta_{K+1} = 1 - \iota_K'\theta$, ι_K is a $K \times 1$ vector with each element equal to unity, and the integrating constant is

$$c_K(\underline{b}_1, \ldots, \underline{b}_{K+1}) = \frac{\Gamma(\underline{b}_1 + \cdots + \underline{b}_{K+1})}{\prod_{k=1}^{K+1} \Gamma(\underline{b}_k)}. \tag{41}$$

The attractiveness of prior (40) is that the corresponding posterior is also Dirichlet with hyperparameters $\bar{b}_k = \underline{b}_k + n_k$ $(k = 1,2,\ldots,K+1)$ and integrating constant $c_{K+1}(\bar{b}_1, \ldots, \bar{b}_{K+1})$.

In general $T \ll K$. Given the observed data y, let a_ψ denote the J_ψ distinct mass points appearing in y. Suppose $a_{K+1} \notin a_\psi$ and let a_λ denote the remaining $J_\lambda = K - J_\psi$ mass points not appearing in the observed data y. Align a_k $(k = 1,2,\ldots, K+1)$ into a $(K+1) \times 1$ vector $a \equiv [a_\psi', a_\lambda', a_{K+1}]'$. Let ψ denote the probabilities

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

corresponding to the J_ψ distinct mass points a_ψ; let λ denote the probabilities corresponding to the J_λ mass points a_λ, $\underline{b}_{K+1}$.

Define $\delta \equiv \iota_\psi'\psi$, $\underline{d} \equiv \iota_\psi'\underline{b}_\psi + \underline{b}_{K+1}$ and $\bar{d} \equiv \iota_\psi'\bar{b}_\psi = \underline{d} + \underline{b}_{K+1} + T$, where ι_ψ, and rearrange θ so that $\theta = [\psi', \lambda']'$. Analogously define $\underline{b} = [\underline{b}_\psi', \underline{b}_\lambda', \underline{b}_{K+1}]'$ and $\bar{b} = [\bar{b}_\psi', \bar{b}_\lambda', \bar{b}_{K+1}]'$. Note that $\bar{b}_\lambda = \underline{b}_\lambda$ and $\bar{b}_{K+1} =$ is a $J_\psi \times 1$ vector with each element equal to unity. Then the data are marginally informative for λ: $\lambda \sim D_{J_\lambda}(\underline{b}_\lambda, \underline{d})$ is updated to $\lambda|y \sim D_{J_\lambda}(\underline{b}_\lambda, \bar{d})$. This information is not surprising because $\mathrm{Cov}(\psi, \lambda) = -E(\psi|\underline{b})[E(\lambda|\underline{b})]'[1 + \iota_\psi'\underline{b}_\psi + \iota_\lambda'\underline{b}_\lambda + \underline{b}_{K+1}]^{-1} \neq 0$ implies ψ and λ are dependent, where ι_λ is a $J_\lambda \times 1$ vector with each element equal to unity, $E(\psi|\underline{b}) = [\iota_\psi'\underline{b}_\psi + \iota_\lambda'\underline{b}_\lambda + \underline{b}_{K+1}]^{-1}\underline{b}_\psi$, and $E(\lambda|\underline{b}) = [\iota_\psi'\underline{b}_\psi + \iota_\lambda'\underline{b}_\lambda + \underline{b}_{K+1}]^{-1}\underline{b}_\lambda$. As Proposition 2 suggests, the data are conditionally uninformative for λ given ψ: $(1-\delta)^{-1}\lambda|\psi \sim (1-\delta)^{-1}\lambda|\psi, y \sim D_{J_\lambda}(\underline{b}_\lambda', \underline{b}_{K+1})$. For example, if $\underline{b} = (K+1)^{-1}\iota_{K+1}$, then $\lambda \sim D_{J_\lambda}([K+1]^{-1}\iota_\lambda, [K+1]^{-1}[J_\psi + 1])$ is updated to $\lambda|y \sim D_{j_\lambda}([K+1]^{-1}\iota_\lambda, [K+1]^{-1}[J_\psi + 1 + T])$, and $(1-\delta)^{-1}\Lambda|\psi \sim (1-\delta)^{-1}\Lambda|\psi, y \sim D_{J_\lambda - 1}([K+1]^{-1}\iota_\lambda, [K+1]^{-1})$.

4. IMPROPER PRIORS

Improper distributions have densities whose integrals diverge to infinity.[9] When nonidentified models are combined with improper priors, serious complications can arise (e.g., improper posteriors).[10] In general all bets are off. But depending on the nature of nonidentifiability and the extent to which the prior, albeit improper, provides enough information on the nonidentified parameters, the posterior may be proper. Specifically, $f(\psi, \lambda) \propto f(\lambda|\psi) f(\psi) \mathcal{L}_*(\psi; y)$ is proper if $f(\lambda|\psi)$ and $f(\psi|y) \propto f(\psi)\mathcal{L}_*(\psi; y)$ are proper. Many issues can be appreciated from reconsideration of Example 3.1 as the prior variances τ^2 and ω^2 approach infinity. For further enlightening discussion and examples, see Gelfand and Sahu (1996).

First suppose $\tau \to \infty$, corresponding to impropriety in the marginal prior distribution of β and of the identified parameter ψ. Then the posterior distribution of $[\beta, \lambda]'$ is a proper bivariate normal with mean $[y, 0]'$ and covariance matrix $\begin{bmatrix} \sigma^2 + \omega^2 & -\omega^2 \\ -\omega^2 & \omega^2 \end{bmatrix}$, and the posterior distribution of $[\psi, \lambda]'$ is a proper bivariate normal with mean $[y, 0]'$ and covariance matrix $\begin{bmatrix} \sigma^2 & 0 \\ 0 & \omega^2 \end{bmatrix}$. Analogous to Example 3.1:

(a) we learn about ψ marginally (the prior is improper, but the posterior is a proper $N(y, \sigma^2)$);
(b) we do not learn about λ conditional on ψ (despite the impropriety in the marginal prior for ψ, the prior conditional distribution of λ given ψ is a proper $N(0, \omega^2)$);
(c) we learn about λ given β (the prior is $\lambda|\beta \sim N(0, \omega^2)$, and the posterior is $\lambda|\beta, y \sim N(\omega^2(y-\beta)[\sigma^2 + \omega^2]^{-1}, \sigma^2\omega^2[\sigma^2 + \omega^2]^{-1})$);
(d) we learn about β marginally.

However, unlike in Example 3.1, we do not learn about λ marginally. In effect the impropriety breaks the dependency between ψ and λ that provided the link for updating marginal prior beliefs on the nonidentified parameter λ.

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

Next suppose $\omega \to \infty$, corresponding to impropriety in the marginal prior distribution of the nonidentified parameter λ. Then the posterior distribution of $[\beta, \lambda]'$ is a proper bivariate normal with mean $[0, y]'$ and covariance matrix $\begin{bmatrix} \tau^2 & -\tau^2 \\ -\tau^2 & \sigma^2 + \tau^2 \end{bmatrix}$, and the posterior distribution of $[\psi, \lambda]'$ is a proper bivariate normal with mean $[y, y]'$ and covariance matrix $\begin{bmatrix} \sigma^2 & -\sigma^2 \\ -\sigma^2 & \sigma^2 + \tau^2 \end{bmatrix}$.

Analogous to Example 3.1:

(a) we learn about ψ marginally (the prior is improper, but the posterior is a proper $N(y, \sigma^2)$);
(b) we do not learn about λ given ψ (despite the impropriety in the marginal prior for λ, the prior and posterior conditional distributions of λ given ψ are proper $N(\psi, \tau^2)$);
(c) we learn about λ given β (the prior is improper, but the posterior is a proper $N([y - \beta], \sigma^2)$);
(d) we learn about λ marginally (the prior is improper, but the posterior is a proper $N(y, \sigma^2 + \tau^2)$).

Unlike in Example 3.1, we do not learn about β marginally (the prior and posterior are proper $N(0, \tau^2)$).

Therefore, these two asymptotic cases differ from the earlier results for proper priors only in that the data are marginally uniformative for λ (as $\tau \to \infty$) or β (as $\omega \to \infty$), respectively. Also note the crucial role played by the prior distribution $\lambda | \psi$ in both the proper and improper cases considered in this subsection. We do not revise this distribution after observing y, and we never learn about $\nu = F(\lambda | \psi)$.

Finally, as both $\tau \to \infty$ and $\omega \to \infty$, the marginal posteriors for β and λ are improper. The full conditionals $\beta | \lambda, y \sim N(y - \lambda, \sigma^2)$ and $\lambda | \beta, y \sim N(y - \beta, \sigma^2)$, however, are both proper. Furthermore, the marginal posterior for the identified ψ is proper $N(y, \sigma^2)$.

These examples suggest, not surprisingly, that prior impropriety can affect our ability to learn from the data in nonidentified models. Also there are connections to the well-studied marginalization paradoxes (see, e.g., Dawid, Stone, and Zidek, 1973, 1980; Hartigan, 1983; Jaynes, 1980a, 1980b; Kadane, Schervish, and Seidenfeld, 1996a, 1996b; Kass and Wasserman, 1996, pp. 1358–1359; Robert, 1994, pp. 131–133; Stone, 1976, 1982; Stone and Dawid, 1972; Sudderth, 1980).

Finally, the Gibbs sampler has helped foster a renaissance of interest in Bayesian inference. The Gibbs sampler generates draws from complicated posterior distributions by iteration through the so-called *full conditional posterior distributions* of blocks of parameters given the data and all other parameters. For the Gibbs sampler to provide a converging sequence (in distribution) of draws from the *joint* posterior distribution, it is necessary that not only are all the full conditionals proper, but that the joint posterior itself is proper (see, e.g., Casella and George, 1992; Roberts, Sahu, and Gilks, 1995). Often the propriety of the former, but not the latter, is easy to establish. Cases involving improper priors and nonidentified models, however, are especially tricky (Hobert and Casella, 1996, p. 1462). Gelfand and Sahu (1996) formally discuss conditions under which the

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

output from a Gibbs sampler based on an improper joint posterior for which all the full conditionals are proper can be used to draw inferences on such embedded models.

5. CONCLUSIONS

This study has discussed at some length the inescapable uninformativeness of data that arise from likelihoods with nonidentified parameters. Although the Bayesian approach clearly encompasses non-Bayesian dogmatic "solutions" to nonidentifiability, it shares with its competitors the inevitable fact that there are some parametric quantities about which the data are uninformative, not only conditionally (on identified parameters) but marginally as well. Consideration of improper priors further complicates matters, possibly resulting in improper posteriors, and leads to links with the literature on marginalization paradoxes.

NOTES

1. Carmen Fernández (personal communication) suggested the following (somewhat nonstandard) example demonstrating that the "only if" analog of Proposition 1 is not valid without further restrictions. Suppose ψ and λ are variation free with $\Psi = (0,10)$. Letting $[\cdot]$ denote the greatest integer function, suppose that the conditional distribution of y given ψ is uniform over $([\psi],[\psi]+1)$, and that the prior density for ψ given λ is $f(\psi|\lambda) = g(\psi - [\psi],\lambda)$, where $g(\cdot)$ is any nonnegative function such that $\int_0^1 g(\psi,\lambda)\,d\psi = .10$. Then it is straightforward to show that this prior specification implies $f(y|\lambda) = .10$, $0 < y < 10$. Because $f(y|\lambda)$ does not depend on λ, the data are marginally uninformative for λ. But it is clear from the prior specification that ψ and λ can be dependent.

2. In the reduced form two-regime switching regression model with a probit switching equation of Vijverberg (1993), the across-regime correlation coefficient is nonidentified. Koop and Poirier (1997) noted that the author's flat prior for the three disturbance correlations, defined over the region in which the disturbance covariance matrix is positive definite, induces prior dependency between the across-regime correlation and each of the regime-switching disturbance simple correlations. Posterior updating of the marginal distribution of the across-regime correlation results. Application of Proposition 3, however, implies the data are marginally uninformative for the across-regime *partial* correlation. The prior dependence induced by a positive definiteness restriction on a covariance matrix also plays a role in subsequent examples.

3. James O. Berger and Robert L. Wolpert suggested this interpretation to me.

4. How does ν in Proposition 3 relate to α? Formally, the argument proceeds as follows. First, note

$$A^{-1} = \begin{bmatrix} I_{K-1} & -d \\ 0'_{K-1} & 1 \end{bmatrix}$$

and let

$$\Sigma^{-1} = (A^{-1})'QA^{-1} = \begin{bmatrix} \Sigma^{\psi\psi} & \Sigma^{\psi\lambda} \\ (\Sigma^{\psi\lambda})' & \Sigma^{\lambda\lambda} \end{bmatrix}.$$

Then using $Q^{-1}A^{-1} = A'\Sigma^{-1}$, rewrite α as $\alpha \equiv c'Q^{-1}A^{-1}\theta = (\Sigma^{\psi\lambda})'\psi + \Sigma^{\lambda\lambda}\lambda$. Second, $f(\lambda|\psi) = f(\lambda|\psi,y) = \phi_1(\lambda|\mu_{\lambda|\psi},\sigma^2_{\lambda|\psi})$, where $\mu_{\lambda|\psi} = \mu_\lambda + (\Sigma_{\psi\lambda})'(\Sigma_{\psi\psi})^{-1}(\psi - \mu_\psi)$ and $\sigma^2_{\lambda|\psi} = \Sigma_{\lambda\lambda} - (\Sigma_{\psi\lambda})'(\Sigma_{\psi\psi})^{-1}\Sigma_{\psi\lambda}$. Third, because Σ^{-1} is the inverse of Σ, $\Sigma_{\psi\psi}(\Sigma^{\psi\lambda})' + \Sigma_{\psi\lambda}\Sigma^{\lambda\lambda} = 0$, and so

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

$(\Sigma_{\psi\lambda})'(\Sigma_{\psi\psi})^{-1} = -(\Sigma^{\lambda\lambda})^{-1}\Sigma^{\psi\lambda}$. Fourth, after some algebra, $\lambda - \mu_{\lambda|\psi} = a_0 + b_0\alpha$, where $a_0 = -[\mu_\lambda - (\Sigma_{\psi\lambda})'(\Sigma_{\psi\psi})^{-1}\mu_\psi]$ and $b_0 = (\Sigma^{\lambda\lambda})^{-1}$. Therefore, ν in Proposition 3 can be expressed as $\nu = \Phi[(\lambda - \mu_{\lambda|\psi})/\sigma_{\lambda|\psi}] = \Phi[(a_0 - b_0\alpha)/\sigma_{\lambda|\psi}]$. The data are marginally uninformative about α and any known one-to-one function of it, such as ν.

5. Although (14) constitutes an interesting reference point, uniformity is not invariant to arbitrary monotonic transformations.

6. Winkler and Gaba (1990) analyzed this model with independent beta priors for both γ and λ, both over [0,1]. Gaba and Winkler (1992) considered the case of differing misclassification rates for $\tilde{y}_t = 0$ and for $\tilde{y}_t = 1$.

7. When the latent data are potentially available and not just a convenient construct, then it is reasonable to discuss priors for β and σ^{-2}. An example is where $\tilde{y}_t$ denotes birth weight and y_t denotes low and high birth weights. There is great precedence in the medical and demographic literatures for working with the latter dichotomy even when the former is available.

8. Indeed, $f(\psi|\lambda) = \phi(\psi|\lambda^{1/2}\underline{b}, \underline{Q})$, where $\underline{b}$ and $\underline{Q}$ are as given in Section 3.3. The quantity $\nu = F_{\lambda|\psi}(\lambda|\psi)$, about which no marginal learning occurs, is not easy to interpret in this case.

9. There are two distinct ways to make the concept of improper priors mathematically precise. (i) One may remove the requirement that the probability of a set is at most one. For example, Hartigan (1983) permits sets to have infinite probability, but this implies that many traditional results of probability theory must be reproved or fail to apply to infinite probabilities. (ii) The approach advocated by de Finetti (1974) does not require that probabilities be countably additive. Again many traditional results must be reproved or scrapped entirely, but probabilities are bounded by unity. Improper priors are not necessarily "noninformative" in any particular sense. Indeed they can be dogmatic about quantities of some interest (Bauwens, 1991).

10. For examples of improper priors leading to improper posteriors, see Berger (1985, p. 187), Chamberlain and Imbens (1996), Kass and Wasserman (1996, p. 1359), Mitchell (1967), O'Hagan (1994, p. 79), Ye and Berger (1991). Ibrahim and Laud (1991) provide (i) sufficient and (ii) necessary and sufficient conditions for the propriety of the (i) posterior and (ii) prior distributions for Jeffreys's priors in generalized linear models.

REFERENCES

Bauwens, L. (1991) The "pathology" of the natural conjugate prior density in the regression model. *Annales d'Econometrie et de Statistique* 23.

Bayarri, M.J., M.H. DeGroot, & J.B. Kadane (1988) What is the likelihood function? (with discussion). In S. Gupta & J.O. Berger (eds.), *Statistical Decision Theory and Related Topics IV*, vol. 1, pp. 1–27. New York: Springer-Verlag.

Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag.

Casella, G. & E.I. George (1992) Explaining the Gibbs sampler. *American Statistician* 46, 167–174.

Chamberlain, G. & G.W. Imbens (1996) Nonparametric Applications of Bayesian Inference. Discussion paper 1772, Harvard Institute of Economic Research, Harvard University.

Dawid, A.P., M. Stone, & J.V. Zidek (1973) Marginalization paradoxes in Bayesian and structural inference (with discussion). *Journal of the Royal Statistical Society, Series B* 35, 189–233.

Dawid, A.P., M. Stone, & J.V. Zidek (1980) Comments on Jaynes's paper "Marginalization and prior probabilities." In A. Zellner (ed.), *Bayesian Analysis in Econometrics and Statistics*, pp. 79–82. Amsterdam: North-Holland.

De Finetti, B. (1974) *The Theory of Probability*, vol. 1. New York: Springer-Verlag.

Dréze, J.H. (1974) Bayesian theory of identification in simultaneous equations models. In S.E. Fienberg & A. Zellner (eds.), *Studies in Bayesian Econometrics and Statistics*, pp. 159–174. Amsterdam: North-Holland.

Dréze, J.H. & J.-F. Richard (1983) Bayesian analysis of simultaneous equation systems. In Z. Griliches & M.D. Intriligator (eds.), *Handbook of Econometrics* 2, pp. 517–600. Amsterdam: North-Holland.

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

Engle, R.F. (1976) Constraints often overlooked in analyses of simultaneous equation models: Comment. *Econometrica* 44, 617–618.
Erickson, T. (1989) Proper posteriors from improper priors for an unidentified errors-in-variables model. *Econometrica* 57, 1299–1316.
Florens, J.-P., M. Mouchart, & J.-F. Richard (1974) Bayesian inference in error-in-variables models. *Journal of Multivariate Analysis* 4, 419–452.
Gaba, A. & R.L. Winkler (1992) Implications of errors in survey data: A Bayesian model. *Management Science* 38, 913–925.
Gelfand, A.E. & S.K. Sahu (1996) Identifiability, Propriety and Parameterization with Regard to Simulation-Based Fitting of Generalized Linear Models. Manuscript, University of Connecticut.
Genberg, H. (1972) Constraints on the parameters in two simple simultaneous equation models. *Econometrica* 40, 855–865.
Harless, D.W. & C.F. Camerer (1994) The predictive utility of generalized expected utility theories. *Econometrica* 62, 1251–1289.
Hartigan, J.A. (1983) *Bayes Theory.* New York: Springer-Verlag.
Hobert, J.P. & G. Casella (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* 91, 1461–1473.
Ibrahim, J.G. & P.W. Laud (1991) On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association* 86, 981–986.
Jaynes, E.T. (1980a) Marginalization and prior probabilities (with discussion). In A. Zellner (ed.), *Bayesian Analysis in Econometrics and Statistics*, pp. 43–78. Amsterdam: North-Holland.
Jaynes, E.T. (1980b) Reply to Dawid, Stone and Zidek. In A. Zellner (ed.), *Bayesian Analysis in Econometrics and Statistics*, pp. 83–87. Amsterdam: North-Holland.
Kadane, J.B. (1974) The role of identification in Bayesian theory. In S.E. Fienberg & A. Zellner (eds.), *Studies in Bayesian Econometrics and Statistics*, pp. 175–191. Amsterdam: North-Holland.
Kadane, J.B., M.J. Schervish, & T. Seidenfeld (1996a) Reasoning to a foregone conclusion. *Journal of the American Statistical Association* 91, 1228–1235.
Kadane, J.B., M.J. Schervish, & T. Seidenfeld (1996b) When several Bayesians agree that there will be no reasoning to a foregone conclusion. *Philosophy of Science* 63, 281–289.
Kass, R.E. & L. Wasserman (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 91, 1343–1370.
Koop, G. & D.J. Poirier (1997) Learning about the across-regime correlation in switching regression models. *Journal of Econometrics* 78, 217–227.
Koopmans, T.C. & O. Reiersøl (1950) The identification of structural characteristics. *Annals of Mathematical Statistics* 21, 225–241.
Leamer, E.E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley.
Lindley, D.V. (1971) *Bayesian Statistics: A Review.* Philadelphia: SIAM.
Lindley, D.V. & G.M. El-Sayyad (1968) The Bayesian estimation of a linear functional relationship. *Journal of the Royal Statistical Society Series B* 30, 190–202.
Maddala, G.S. (1976a) Constraints often overlooked in analyses of simultaneous equation models: Comment. *Econometrica* 44, 615–616.
Maddala, G.S. (1976b) Rejoinder. *Econometrica* 44, 625.
Malinvaud, E. (1970) *Statistical Methods in Econometrics*, 2nd ed. Amsterdam: North-Holland.
Manski, C.F. (1994) *Identification Problems in the Social Sciences.* Cambridge: Harvard University Press.
Mitchell, A.F.S. (1967) Comment on "A Bayesian significance test for multinomial distributions." *Journal of the Royal Statistical Society, Series B* 29, 423.
O'Hagan, A.O. (1994) *Kendall's Advanced Theory of Statistics: Bayesian Inference*, vol. 2B. New York: Halsted Press.
Poirier, D.J. (1995) *Intermediate Statistics and Econometrics: A Comparative Approach.* Cambridge: MIT Press.

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press

Poirier, D.J. (1996) Prior beliefs about fit. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (eds.), *Bayesian Statistics 5*, pp. 731–738. Oxford: Oxford University Press.
Robert, C.P. (1994) *The Bayesian Choice: A Decision Theoretic Motivation.* New York: Springer-Verlag.
Roberts, G.O., S.K. Sahu, & W.R. Gilks (1995) Comment. *Statistical Science* 10, 49–51.
Stone, M. (1976) Strong inconsistency from uniform priors (with discussion). *Journal of the American Statistical Association* 71, 114–125.
Stone, M. (1982) Review and analysis of some inconsistencies related to improper priors and finite additivity. In L.J. Cohen, J. Los, H. Pfeiffer, & K.-P. Podewski (eds.), *Logic, Methodology and Philosophy of Science VI*, pp. 413–426. Amsterdam: North-Holland.
Stone, M. & A.P. Dawid (1972) Un-Bayesian implications of improper Bayesian inference in routine statistical problems. *Biometrika* 52, 623–627.
Sudderth, W.D. (1980) Finitely additive priors, coherence and the marginalization paradox. *Journal of the Royal Statistical Society, Series B* 42, 339–341.
Vijverberg, W.P.M. (1993) Measuring the unidentified parameter of the extended Roy model of selectivity. *Journal of Econometrics* 57, 69–89.
Winkler, R.L. & A. Gaba (1990) Inference with imperfect sampling from a Bernoulli process. In S. Geisser, J.S. Hodges, S.J. Press, & A. Zellner (eds.), *Bayesian and Likelihood Methods in Statistics and Econometrics*, pp. 303–317. Amsterdam: North-Holland.
Ye, K. & J.O. Berger (1991) Noninformative priors for inferences in exponential regression models. *Biometrika* 78, 645–656.
Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.
Zellner, A. (1972) Constraints often overlooked in analyses of simultaneous equation models. *Econometrica* 40, 849–853.
Zellner, A. (1976a) Reply. *Econometrica* 44, 619–624.
Zellner, A. (1976b) Further reply. *Econometrica* 44, 627–628.

https://doi.org/10.1017/S0266466698144043 Published online by Cambridge University Press