Including Prior Information in Probit Model Estimation

William E. Griffiths
University of Melbourne

R. Carter Hill
Louisiana State University

Christopher J. O'Donnell
University of New England

September 28, 2001

Abstract

The effects of including different kinds of prior information in estimation of the probit model is examined within the framework of Bayesian inference. Of interest is the effect on posterior information for coefficients, probabilities and elasticities. In a model designed to explain choice between fixed and variable interest-rate mortgages, we show that using Bayesian inference to include inequality information on the signs of coefficients yields inferences about probabilities and elasticities that are substantially different from those obtained using maximum likelihood estimation. In a second model, concerned with state voting behavior, we find that putting prior information on probabilities, rather than coefficients, has a dramatic effect on the posterior density functions for the model coefficients, probabilities and elasticities.

KEY WORDS: Inequality restrictions; Metropolis-Hastings algorithm; Voting choice; Mortgage choice.

The probit model is a popular device for explaining binary choice decisions in econometrics. It has been used to describe choices such as labor force participation, travel mode, home ownership and type of education. These and many more examples can be found in Amemiya (1981) and Maddala (1983). Given the contribution of economics towards explaining such choices, and given the nature of data that are collected, prior information on the relationship between a choice probability and several explanatory variables frequently exists. In this paper we explore ways of incorporating prior information into estimation of the probit model. Our approach is Bayesian. We show how different prior probability density functions (pdfs) can be used to model different kinds of prior information that are likely to occur in practice. Two examples are considered. The first is concerned with the choice between fixed and variable interest-rate mortgages. In this example we illustrate how to include inequality information in the form of restrictions on the signs of the coefficients. The second example models voting choice. In this case prior information about probabilities in one election is used to formulate a prior pdf on parameters in a model constructed to explain voting choice in a subsequent election. In both examples the posterior pdfs on coefficients and on choice probabilities are not analytically tractable. Metropolis-Hastings algorithms are used to generate observations from posterior distributions on quantities of interest.

The first Bayesian analysis of binary choice models in the econometrics literature was that of Zellner and Rossi (1984). They derived a normal approximation to the posterior pdf of the coefficients, and, focusing mainly on the logit model, showed how importance sampling can be used to find posterior pdfs for coefficients, probabilities and elasticities. In line with the recent explosion of work using Markov Chain Monte Carlo (see, for example, Hill 1996), Albert and Chib (1993) show how data augmentation, in conjunction with the Gibbs sampler, can be used to estimate posterior pdfs of interest for the probit model. They generalize the analysis to a *t*-distribution, showing how a Gibbs sampler can be set up in the context of a scale mixture of normal distributions. Extensions to the multinomial probit model were also considered by Albert and Chib (1993) and later by Geweke, Keane and Runkle (1994, 1997). Our work differs because we focus on methods for including different types of prior information, and, because such prior information does not yield a

posterior pdf that is amenable to Gibbs sampling, we use a Metropolis-Hastings algorithm instead.

In Section 1 the model for explaining choice between fixed and variable rate mortgages is described. Unrestricted maximum likelihood estimates and standard errors for this model, as well as estimates and standard errors for some choice probabilities and elasticities, are discussed in Subsection 1.1. The scope for including sign restrictions on coefficients, and problems with choice-probability confidence intervals containing infeasible ranges, are noted. Bayesian methodology for imposing sign restrictions on coefficients is described in Subsection 1.2, and the consequent results are discussed in Subsection 1.3. We find that Bayesian and maximum likelihood estimation can lead to quite different results for the probabilities and elasticities. Section 2 is devoted to the model for voting choice, with the prior on probabilities being developed in Subsection 2.1, and the results being discussed in Subsection 2.2. Results from a diffuse prior on the coefficients are contrasted with those from an almost uniform prior on the probabilities.

1. MORTGAGE DATA AND INEQUALITY RESTRICTIONS

Dhillon, Shilling and Sirmans (1987) estimate a probit model designed to explain the choice by homebuyers of fixed versus adjustable rate mortgages. They use 78 observations from a bank in Baton Rouge, Louisiana, taken over the period January, 1983 to February, 1984. In this data set 46 fixed-rate and 38 adjustable-rate mortgages were chosen. Dhillon et al. used both financial measures and personal characteristics as explanatory variables in their model, and did not reject a hypothesis that the personal characteristics have no impact on the choice probability. We focus on the financial measures and introduce sign constraints in the form of inequality restrictions on the coefficients, using the signs implied by the discussion in Dhillon et al. The data are taken from Lott and Ray (1992).

The probit model can be written as

$$P_i = \Phi(x_i'\beta) \tag{1}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, β is a vector of unknown coefficients to be estimated, and, in the context of our example,

P_i = probability of choosing an adjustable rate mortgage.

The vector of explanatory variables x_i is of dimension 7. Its components and the expected signs of the corresponding coefficients are:

x_{1i} = 1;

x_{2i} = fixed interest rate ($\beta_2 > 0$);

x_{3i} = margin = the variable rate less the fixed rate ($\beta_3 < 0$);

x_{4i} = yield = the 10-year treasury rate less the one year treasury rate ($\beta_4 < 0$);

x_{5i} = points = ratio of points paid on adjustable rates to those paid on fixed rates ($\beta_5 < 0$);

x_{6i} = maturity = ratio of maturities on adjustable to fixed rates ($\beta_6 < 0$);

x_{7i} = net worth of borrower ($\beta_7 > 0$).

In general, the effect of a change in one of the explanatory variables (say the *k*-th variable) on the choice probability is given by the derivative

$$\frac{\partial P_i}{\partial x_{ki}} = \frac{\partial \Phi(x_i'\beta)}{\partial x_{ki}} = \beta_k \phi(x_i'\beta) \tag{2}$$

where $\phi(\cdot)$ is the standard normal probability density function. Since $\phi(\cdot)$ is always positive, whether an increase in x_k leads to an increase or a decrease in P_i is given by the sign of β_k. The fixed rate and the margin are designed to pick up cross-price and own-price effects, respectively, and hence their coefficients β_2 and β_3 are expected to be positive and negative, respectively. The yield variable represents a risk variable. The larger the yield the more likely it is that the adjustable rate will increase and hence the less attractive is the adjustable rate mortgage ($\beta_4 < 0$). Other things equal, the greater the relative points, the less likely an adjustable rate will be chosen ($\beta_5 < 0$). Assuming shorter maturities are more desirable than longer ones, we have $\beta_6 < 0$. Finally, the greater the net worth of the borrower, the more likely is the borrower to take the risk of an adjustable rate ($\beta_7 > 0$).

1.1 Unrestricted Maximum Likelihood Estimation

Unrestricted maximum likelihood estimates are given in Table 1. Note that all estimates have the expected signs. However, 95% confidence intervals for some of

the coefficients will include both positive and negative values and will hence have a region which is infeasible, in the sense that its values have the wrong sign. The coefficients where this happens are those with *p*-values greater than 0.05, namely, those for fixed rate, points and maturity. Thus, although Bayesian inequality-restricted estimation is unlikely to change the signs of estimated coefficients, it will have an impact on interval estimation, producing interval estimates without infeasible regions.

Table 1: Unrestricted Maximum Likelihood Estimates for Mortgage Data

Variable	Estimate	St. Error	*t*	p-value
constant	–1.877	4.225	–0.444	0.657
fixrate	0.499	0.277	1.799	0.072
margin	–0.431	0.174	–2.483	0.013
yield	–2.384	1.088	–2.191	0.028
points	–0.300	0.241	–1.242	0.214
maturity	–0.059	0.615	–0.096	0.923
networth	0.084	0.042	1.988	0.047

The coefficients are useful for examining the direction of a probability change that results from a change in an explanatory variable, but their magnitudes by themselves are not very informative. One is usually more interested in elasticities and probabilities evaluated at a particular point x_*. These quantities are given, respectively, by

$$E_{k^*} = \frac{\partial P_*}{\partial x_{k^*}} \frac{x_{k^*}}{P_*} = \beta_k x_{k^*} \frac{\phi(x_*'\beta)}{\Phi(x_*'\beta)} \tag{3}$$

and

$$P_* = \Phi(x_*'\beta) \tag{4}$$

To explore the differences between unrestricted maximum likelihood estimation and inequality restricted Bayesian estimation, two values for x_* were chosen, namely, observations 13 and 29 in the data set. Also, for the elasticities we focused on two of the more important explanatory variables, margin and yield. Observations 13 and 29 were chosen because they led to quite different estimated probabilities, one about 0.9 and the other about 0.05. Their characteristics and how they stand relative to the whole sample are given in Table 2. Note that the major difference between the two observations is in net worth of the borrower.

Table 2: Characteristics of Explanatory Variables for Mortgage Data

Variable	Mean	Minimum	Maximum	Obs 13	Obs 29
fixrate	13.25	11.76	14.50	13.5	12.13
margin	2.292	–0.90	5.50	2.5	3.36
yield	1.606	1.38	2.04	1.59	1.60
points	1.498	0.00	4.34	1.00	1.66
maturity	1.058	0.42	2.38	1.00	0.85
networth	3.504	–0.056	17.86	17.86	0.118

Maximum likelihood estimates for the probabilities, and the margin and yield elasticities, and corresponding standard errors in parentheses, for observations 13 and 29, are given in Table 3. The standard errors are obtained using the conventional first-order approximation for the asymptotic variance of a nonlinear function of the maximum likelihood estimator. See, for example, Judge et al. (1985, p.160). If $\hat{\beta}$ denotes the maximum likelihood estimator, then $\hat{P}_* = \Phi(x_*'\hat{\beta})$ and $\hat{E}_{k*} = \hat{\beta}_k x_{k*}\phi(x_*'\hat{\beta}) / \Phi(x_*'\hat{\beta})$, and the asymptotic variances can be derived as

$$\text{var}(\hat{P}_*) = [\phi(x_*'\hat{\beta})]^2 x_*'Vx_* \tag{5}$$

and

$$\text{var}(\hat{E}_{k*}) = x_{k*}^2 \text{ var}[\hat{\beta}_k \phi(x_*'\hat{\beta}) / \Phi(x_*'\hat{\beta})] \tag{6}$$

where V is the covariance matrix for $\hat{\beta}$, and the variance term in equation (6) is given by the k-th diagonal element of

$$\text{cov}\left[\hat{\beta}\,\frac{\phi(x_*'\hat{\beta})}{\Phi(x_*'\hat{\beta})}\right] = \left[\frac{\phi(x_*'\beta)}{\Phi(x_*'\beta)}\right]^2 QVQ' \tag{7}$$

with

$$Q = I - [x_*'\beta + \phi(x_*'\beta)/\Phi(x_*'\beta)]\beta x_*' \tag{8}$$

From Table 3, we see that 95% confidence intervals for both probabilities, and for both elasticities evaluated at observation 29, will contain infeasible regions. The interval for the probability at observation 13 will contain values greater than one, while that for observation 29 will contain values less than zero. We expect both elasticities to be negative, but the large standard errors for those at observation 29 will

lead to confidence intervals that contain a substantial region of positive values. When we talk of confidence intervals, we are assuming the usual large-sample practice of deriving intervals from the normal distribution is being employed. A possible reason for the standard errors for $\hat{E}_{29}$ being much larger than those of $\hat{E}_{13}$ is the appearance of $\Phi(x_*'\beta)$ in the denominator of equation (3). A value close to zero is likely to cause more instability.

Table 3: Estimated Probabilities and Elasticities for Mortgage Data

	Obs 13		Obs 29	
	ML (st. error)	Bayes (st. dev'n)	ML (st. error)	Bayes (st. dev'n)
Probability	0.879 (0.130)	0.865 (0.105)	0.052 (0.041)	0.058 (0.043)
Elasticity (Margin)	–0.237 (0.054)	–0.286 (0.230)	–2.966 (2.917)	–3.566 (1.578)
Elasticity (Yield)	–0.869 (0.206)	–1.070 (0.898)	–7.814 (8.183)	–9.387 (4.266)

1.2 Bayesian Estimation with Inequality Restrictions

For Bayesian estimation with inequality restrictions imposed on the signs of the coefficients, we begin with a uniform prior pdf over the feasible region. Let $I_\beta(R)$ denote an indicator function which is equal to 1 when β is such that all the sign constraints are satisfied (β belongs to the feasible region R) and zero otherwise. Then, the chosen prior pdf can be written as

$$f(\beta) \propto I_\beta(R) \tag{9}$$

The likelihood function is given by

$$f(y|\beta) = \prod_{i=1}^{N} [\Phi(x_i'\beta)]^{y_i} [1 - \Phi(x_i'\beta)]^{1-y_i} \tag{10}$$

where $y = (y_1, y_2, ..., y_N)'$ is a vector of binary variables with $y_i = 1$ if the *i*-th observation is a variable-rate mortgage and $y_i = 0$ if the *i*-th observation is a fixed rate mortgage. Using Bayes' Theorem, the posterior pdf for β can be written as

$$f(\beta|y) \propto f(\beta) f(y|\beta)$$

$$\propto I_{\beta}(R) \prod_{i=1}^{N} [\Phi(x_i'\beta)]^{y_i} [1 - \Phi(x_i'\beta)]^{1-y_i} \qquad (11)$$

1.1a Gibbs Sampling

Before describing the Metropolis-Hastings algorithm that we used to draw observations from this pdf, it is instructive to ask what would happen if we included the inequality restrictions into the framework of data augmentation and Gibbs sampling suggested by Albert and Chib (1993). In this framework latent variables y_i^* are introduced and, before observing the sample, they are assumed to be independent $N(x_i'\beta, 1)$. Then, Gibbs sampling is used to draw successively from the conditional posterior pdfs $f(y_i^*|\beta, y)$ and $f(\beta|y^*, y)$. When a uniform prior is used for β, the $f(y_i^*|\beta, y)$ are independent normal distributions, truncated at zero. The truncation is from below when $y_i = 1$ and from above when $y_i = 0$. These conditional posterior pdfs for the y_i^* do not change when inequality restrictions on β are introduced. The conditional posterior pdf $f(\beta|y^*, y)$ does change, however. Without inequality restrictions it is $N[b, (X'X)^{-1}]$ where $b = (X'X)^{-1} X'y^*$ and X is an ($N \times K$) matrix containing all the x_i in the sample. With inequality restrictions it becomes a truncated multivariate normal distribution with the same location vector and precision matrix, but truncated to the region *R*. Thus, instead of using a Metropolis-Hastings algorithm to draw observations from equation (11), one can proceed by drawing observations successively from truncated normal distributions. However, drawing observations from the *K*-dimensional truncated normal distribution for β is not always straightforward. A simple algorithm which draws from a normal distribution and that accepts feasible draws and rejects infeasible draws is likely to fail because of the low probability of making a feasible draw. One alternative is to transform the *K*-

dimensional truncated normal distribution into a series of univariate truncated normal distributions (Geweke 1991), but this approach is somewhat cumbersome.

1.2b Metropolis-Hastings Algorithm

As an alternative, we use the following random-walk Metropolis-Hastings algorithm to draw observations from the posterior pdf in equation (11). The maximum likelihood estimate for β was chosen as an initial value β_0 for the Markov chain. A scalar multiple of the maximum likelihood covariance matrix estimate V was used as a covariance matrix for the random-walk generator function. The steps for drawing the $(m+1)th$ observation $\beta_{(m+1)}$ were:

1. Draw a candidate value β^* from a $N(\beta_{(m)}, cV)$ distribution where c is a scalar set such that β^* is accepted approximately 50% of the time.

2. Compute $$r = \frac{f(\beta^* | y)}{f(\beta_{(m)} | y)}.$$

   Note that this ratio can be computed without knowledge of the normalising constant for $f(\beta | y)$. Also, if any of the elements of β^* fall outside the feasible parameter region, then $f(\beta^* | y) = 0$.

3. Draw a value u for a uniform random variable on the interval (0,1).

4. If $u \le r$, set $\beta_{(m+1)} = \beta^*$.

   If $u > r$, set $\beta_{(m+1)} = \beta_{(m)}$.

5. Return to step 1, with m set to $m + 1$.

A total of 200,000 draws were made, with the first 40,000 being discarded for a "burn in". Various tests for convergence were carried out; there was no evidence to suggest the chain had not achieved stationarity.

1.3 Results from Bayesian Estimation

The posterior means and standard deviations for the coefficients appear in Table 4, alongside their maximum likelihood counterparts. The means and standard deviations for the probabilities, and the selected elasticities, for observations 13 and 29, appear in Table 3. Plots of selected posteriors are given in Figure 1 (coefficients),

Figure 2 (probabilities) and Figure 3 (elasticities). With the exception of the coefficient on maturity, imposing the sign constraints has had only a small effect on the coefficient estimates. As expected, this effect is to increase the absolute value of the point estimates and to reduce estimation uncertainty, measured by the standard error in the case of maximum likelihood estimation, and the posterior standard deviation in the case of Bayesian estimation. The dotted pdfs in Figure 1 are normal pdfs centered at the maximum likelihood estimates and with the corresponding standard errors as standard deviations. Given these pdfs are used for sampling-theory interval estimation, they can be viewed as the sampling theorists' posterior pdfs. Since the unrestricted coefficient for maturity is almost zero, with a large standard error, truncation has a big impact in this case; the mode of the Bayesian posterior pdf is close to zero and the mean is almost 10 times larger than the maximum likelihood estimate. Interestingly, the other truncations have little effect, even for the coefficient of "points" where the maximum likelihood pdf has noticeable probability above zero.

Table 4: Maximum Likelihood and Bayesian Estimates for Mortgage Data

	ML (st. error)	Bayes (st. dev'n)
constant	–1.877 (4.225)	–1.323 (3.996)
fixrate	0.499 (0.277)	0.561 (0.253)
margin	–0.431 (0.174)	–0.494 (0.176)
yield	–2.384 (1.088)	–2.767 (1.078)
points	–0.300 (0.241)	–0.379 (0.216)
maturity	–0.059 (0.615)	–0.557 (0.421)
networth	0.084 (0.042)	0.085 (0.037)

A comparison of the Bayesian and maximum likelihood estimates of the probabilities in Table 4 suggests little difference between the results. However, the corresponding pdfs in Figure 2 show that there can be a considerable difference. If one blindly uses the normal distribution to construct interval estimates for P_{13} and P_{29} on the basis of the maximum likelihood results, the interval estimates will include negative probabilities and probabilities that exceed one. Bayesian estimation overcomes this problem.

Inference about elasticities is also very sensitive to whether one opts for large-sample maximum-likelihood inference or finite-sample Bayesian inference. Note the dramatic differences in the standard errors (deviations) in Table 3 and the differences in spread in Figure 3. It would appear that, when P_i is close to one, maximum likelihood estimation overstates the precision with which the elasticities are estimated; when P_i is close to zero, maximum likelihood understates this precision. We conjecture that the term $\phi(x_i'\beta)/\Phi(x_i'\beta)$ in equations (7) and (8) helps explain this phenomenon. It could be too small when evaluated at a point estimate where its denominator P_i is close to one, and too large when evaluated at a point estimate where its denominator P_i is close to zero.

Overall, we conclude that inequality restrictions have the expected effect on coefficient estimation. Also, large-sample (maximum-likelihood) estimation can produce misleading inferences about probabilities and elasticities. Moving to Bayesian inference appears to overcome these problems. Because we did not obtain posterior pdfs for the probabilities and elasticities without imposition of the inequality constraints, it is difficult to know whether the different inferences are a consequence of Bayesian estimation, or the inequality constraints. We suspect, however, that it is largely a consequence of using Bayesian inference.

2. VOTE DATA AND A PRIOR ON PROBABILITIES

In the first edition of his textbook, Greene (1990, p.673) estimated a probit model where the probability of a state voting democrat in the 1976 U.S. Presidential election was a function of median family income in that state, the median number of years of school completed by persons 18 years of age or older, the percentage of population living in a metropolitan area, and the region to which the state belongs.

Four regions were specified. This example was dropped from subsequent editions of his book, probably because a singular information matrix led to nonidentifiability of some of the regional parameters. We use the same example, modified in a number of ways. Our sample of data is for the 1996 Presidential election, with information from the 1992 election used to formulate a prior pdf. For variables to represent level of education, we use the percentage of population who are high school graduates and the percentage of population with bachelor's degrees. For regional effects two dummy variables are used, one that is equal to one for Southern states and one that is equal to one for Western states. The allocation of states to regions is somewhat arbitrary and slightly different from the equally arbitrary allocation made by Greene. The data are given in Table 5. The binary variable *Y* is equal to 1 if the number of voters voting Democrat exceeded the number voting Republican, and zero otherwise. Income is median household income in constant (1995) dollars, expressed in thousands of dollars. The next two columns are the percentages of the population with high school diplomas and bachelor's degrees, respectively. The column headed "Metrop" gives the percentage of population living in a metropolitan area or consolidated metropolitan area, as defined by the U.S. Office of Management and Budget, June 30, 1996. The code for region is 2 for Southern states, 3 for Western states and 1 for the others. The final column relates to the prior pdf and is discussed below.

2.1 The Prior Density Function

With these new definitions, we have a new probit model, designed to explain the probability of voting Democrat. Like before, it can be written as

$$P_i = \Phi(x_i'\beta) \tag{12}$$

with likelihood function

$$f(y|\beta) = \prod_{i=1}^{N} [\Phi(x_i'\beta)]^{y_i} [1 - \Phi(x_i'\beta)]^{1-y_i} \tag{13}$$

In the previous example we considered prior information in the form of inequality restrictions on the elements of β. We now wish to investigate how to include prior information on the P_i. Previous studies that employ an informative prior for the probit model (for example, Geweke 1999) usually do so in terms of β, presumably

Table 5: Data for Voting Example

State	Y	Income	HSchool	Bach	Metrop	Region	Prior Mean p
Alabama	0	25.991	77.6	19.3	67.7	2	0.46185
Alaska	0	47.954	92.1	27.5	41.3	3	0.43333
Arizona	1	30.863	82.6	19.5	87.6	3	0.48700
Arkansas	1	25.814	76.9	14.6	48.3	2	0.60024
California	1	37.009	80.7	27.5	96.6	3	0.58512
Colorado	0	40.706	87.6	28.9	84.0	3	0.52808
Connecticut	1	40.243	84.0	30.0	95.6	1	0.54127
Delaware	1	34.928	84.4	26.8	81.9	1	0.55263
DC	1	30.748	80.3	33.7	100.0	1	0.90187
Florida	1	29.745	81.4	21.7	92.9	2	0.48822
Georgia	0	34.099	78.8	22.3	68.5	2	0.50349
Hawaii	1	42.851	83.7	22.5	73.6	3	0.56646
Idaho	0	32.676	85.7	19.4	37.5	3	0.40294
Illinois	1	38.071	84.4	25.0	84.1	1	0.58586
Indiana	0	33.385	81.9	16.2	71.7	1	0.46162
Iowa	1	35.519	86.7	21.7	44.3	1	0.53712
Kansas	0	30.341	88.1	27.5	55.4	1	0.46429
Kentucky	1	29.810	75.4	17.6	48.2	2	0.51872
Louisiana	1	27.949	75.7	18.1	75.2	2	0.52679
Maine	1	33.858	85.8	20.0	35.8	1	0.55957
Maryland	1	41.041	84.7	32.2	92.8	1	0.58314
Massachusetts	1	38.574	85.9	33.5	96.1	1	0.62100
Michigan	1	36.426	86.0	21.0	82.4	1	0.54612
Minnesota	1	37.933	87.9	28.3	69.7	1	0.57716
Mississippi	0	26.538	77.5	20.9	35.3	2	0.45045
Missouri	1	34.825	80.1	22.9	68.0	1	0.56515
Montana	0	27.757	88.6	25.2	23.5	3	0.51839
Nebraska	0	32.929	86.0	21.3	51.3	1	0.38681
Nevada	1	36.084	85.4	19.9	85.7	3	0.51781
New Hampshire	1	39.171	85.1	27.0	59.8	1	0.50852
New Jersey	1	43.924	84.8	28.5	100.0	1	0.51414
New Mexico	1	25.991	78.0	23.6	56.7	3	0.55158
New York	1	33.028	80.0	25.8	91.8	1	0.59472
North Carolina	0	31.979	78.4	22.6	66.8	2	0.49533
North Dakota	0	29.089	82.6	20.5	42.7	3	0.42128
Ohio	1	34.941	86.2	21.5	81.1	1	0.51173
Oklahoma	0	26.311	85.2	20.5	60.2	3	0.44371
Oregon	1	36.374	84.7	24.3	70.2	3	0.56609
Pennsylvania	1	34.524	82.4	22.9	84.6	1	0.55545
Rhode Island	1	35.359	77.5	25.7	93.8	1	0.61739
South Carolina	0	29.071	77.3	19.2	69.6	2	0.45369
South Dakota	0	29.578	85.6	20.1	33.3	3	0.47710
Tennessee	1	29.015	76.1	17.1	68.0	2	0.52620
Texas	0	32.039	78.5	22.4	84.2	2	0.47761
Utah	0	36.480	89.5	26.7	77.1	3	0.36166
Vermont	1	33.824	84.4	23.7	27.7	1	0.60360
Virginia	0	36.222	81.3	28.0	77.9	2	0.47443
Washington	1	35.568	88.8	26.1	82.8	3	0.57599
West Virginia	1	24.880	77.3	14.7	41.8	2	0.57766
Wisconsin	1	40.955	87.1	22.4	67.7	1	0.52789
Wyoming	0	31.529	91.3	22.2	29.7	3	0.46259

because a normal prior on β works well within the framework of data augmentation and Gibbs sampling. However, because no direct intuitive interpretation of the elements of β exists, it is likely that most researchers would feel more comfortable placing a prior on the probabilities P_i. In this regard a family of pdfs which is defined over the interval (0, 1), and which is very flexible in terms of the alternative shapes that it can accommodate, is the beta family. Accordingly, we will assume that we have prior information on each of the P_i that can be expressed in terms of the independent beta pdfs

$$f(P_i) \propto P_i^{\alpha_i - 1}(1 - P_i)^{\gamma_i - 1} \qquad i = 1, 2, \ldots, N \tag{14}$$

Setting of the prior parameters α_i and γ_i is dependent on the prior information available. The joint prior pdf for all the probabilities in the sample $P = (P_1, P_2, \ldots, P_N)'$ is

$$f(P) \propto \prod_{i=1}^{N} P_i^{\alpha_i - 1}(1 - P_i)^{\gamma_i - 1} \tag{15}$$

Since the unknown parameters in the model are the elements in β, and Bayes' Theorem is formulated in terms of β, we are faced with the problem of transforming the *N*-dimensional pdf in (15) to a *K*-dimensional prior pdf for β. In our application $N = 51$ and $K = 7$; a transformation with such a reduction in dimension must clearly involve some restrictions.

Working in this direction, we first define $z_i = x_i'\beta$ and $z = (z_1, z_2, \ldots, z_N)'$. Then,

$$f(z) = f(P)\left|\frac{\partial P}{\partial z}\right|$$

$$\propto \prod_{i=1}^{N} [\Phi(z_i)]^{\alpha_i - 1}[1 - \Phi(z_i)]^{\gamma_i - 1}\phi(z_i) \tag{16}$$

To derive $f(\beta)$ from $f(z)$ we begin by setting $\theta = (\beta', \lambda')'$ where λ is an $(N–K)$ dimensional vector such that the restrictions necessary for transforming the *N*-dimensional pdf $f(z)$ to the *K*-dimensional pdf $f(\beta)$ can be written as $\lambda = 0$. Given this framework, it is reasonable to define the prior pdf for β as

$$f(\beta) = f(\beta|\lambda = 0)$$

that can be conveniently found as

$$\begin{aligned} f(\beta) &= f(\beta|\lambda = 0) \\ &= \left. \frac{f(\beta,\lambda)}{f(\lambda)} \right|_{\lambda=0} \\ &\propto \left. f(\beta,\lambda) \right|_{\lambda=0} \end{aligned} \tag{17}$$

This approach to reducing the dimension of a prior pdf was inspired by Kleibergen and van Dijk (1998), although their motivation and simultaneous-equations context are quite different from our motivation and context.

To define a suitable λ, we partition z as $z = (z_*', z_0')'$ where z_* is (K×1) and z_0 is ((N–K)×1). The matrix of explanatory variables X is similarly partitioned into a (K×K) non-singular matrix X_* and a ((N–K)×K) matrix X_0. Then, we can write

$$\begin{bmatrix} z_* \\ z_0 \end{bmatrix} = \begin{bmatrix} X_* & 0 \\ X_0 & I \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} \tag{18}$$

Now,

$$f(\theta) = f(z) \left| \frac{\partial z}{\partial \theta'} \right| \tag{19}$$

where

$$\left| \frac{\partial z}{\partial \theta'} \right| = \begin{vmatrix} \frac{\partial z_*}{\partial \beta'} & \frac{\partial z_*}{\partial \lambda'} \\ \frac{\partial z_0}{\partial \beta'} & \frac{\partial z_0}{\partial \lambda'} \end{vmatrix} = |X_*| \tag{20}$$

that can be absorbed into a proportionality constant. Thus,

$$\begin{aligned} f(\theta) &= f(\beta,\lambda) \\ &\propto f(z_*, z_0) \\ &= \prod_{i \in *} [\Phi(x_i'\beta)]^{\alpha_i - 1} [\Phi(x_i'\beta)]^{\lambda_i - 1} \phi(x_i'\beta) \times \\ &\quad \prod_{i \in 0} [\Phi(x_i'\beta + \lambda_i)]^{\alpha_i - 1} [1 - \Phi(x_i'\beta + \lambda_i)]^{\gamma_i - 1} \phi(x_i'\beta + \lambda_i) \end{aligned} \tag{21}$$

For the prior pdf for β, we have, therefore

$$f(\beta) \propto f(\beta, \lambda)\Big|_{\lambda=0}$$

$$\propto \prod_{i=1}^{N} [\Phi(x_i'\beta)]^{\alpha_i - 1} [1 - \Phi(x_i'\beta)]^{\gamma_i - 1} \phi(x_i'\beta) \qquad (22)$$

It is interesting that this prior is invariant with respect to the partitioning in equation (18). Before discussing the nature of the posterior pdf, and proceeding with estimation for the voting data, we need to set values for the prior parameters α_i and γ_i. We chose to make the priors relatively uniform, so that the results would not be unduly influenced by the 1992 results for which the same model may not be relevant. However, to use some of the information from 1992 we set the prior modes for each of the P_i equal to the proportion of voters who voted Democrat. This proportion was calculated from Democrat and Republican voters only, not those who voted for Perot. The prior modes are given in the last column of Table 5. Values of α_i and γ_i which produce a relatively flat pdf and which give the desired mode are given by the solution to the following two equations (see, for example, Evans, et al., 1993, p.31):

$$\alpha_i + \gamma_i = 2.5$$

$$\frac{\alpha_i - 1}{\alpha_i + \gamma_i - 2} = \text{prior mode}$$

2.2 Posterior Density Function and Results

Multiplying the prior pdf in (22) by the likelihood function yields the posterior pdf

$$f(\beta|y) \propto \prod_{i=1}^{N} [\Phi(x_i'\beta)]^{y_i + \alpha_i - 1} [1 - \Phi(x_i'\beta)]^{\gamma_i - y_i} \phi(x_i'\beta) \qquad (23)$$

Relative to the posterior pdf derived from a noninformative prior on β, the beta prior on the P_i changes the powers attached to each observation. In this sense its effect is similar to that from the inclusion of additional observations. Note, however, that α_i and γ_i can be fractions. The prior also introduces the additional term $\phi(x_i'\beta)$; it has the effect of making values of β that lead to probabilities close to zero or one less likely.

Table 6: Maximum Likelihood and Bayesian Estimates for β

	ML Estimate (Stand. Error)	Posterior Mean (Posterior St. Dev.)
constant	17.77 (8.14)	4.00 (3.34)
income	0.073 (0.068)	0.017 (0.030)
H. School	–0.230 (0.106)	–0.050 (0.043)
Bach.	–0.047 (0.078)	–0.015 (0.032)
Metrop.	0.021 (0.014)	0.0048 (0.0064)
South	–2.535 (0.884)	–0.617 (0.352)
West	–0.989 (0.551)	–0.252 (0.254)

The same Metropolis-Hastings algorithm that was used for the mortgage-choice model was used to draw observations from the posterior pdf in (23). A total of 50,000 draws were made using the Metropolis-Hastings algorithm, 10,000 of these were discarded for a burn in, and checks for convergence were made. The estimated posterior means and standard deviations from these draws appear in Table 6, along with the maximum likelihood estimates. The maximum likelihood estimates will be similar to the Bayesian estimates obtained from a uniform prior on β. Under these circumstances the likelihood function is identical to the posterior pdf for β, and hence the maximum likelihood estimates are identical to the posterior mode. With quadratic loss, the Bayesian estimates are taken as the posterior mean, which will be the same as the posterior mode if the posterior pdf is symmetric. As we shall see, the posterior pdfs are approximately symmetric, making it reasonable to take the maximum likelihood estimates as the Bayesian estimates under a uniform prior on β. Returning to Table 6, it can be seen that the two sets of estimates are remarkably different, given the seemingly mild prior information included on P_i to obtain the estimates in the

third column. The change in the results that follows the inclusion of this prior information follows a consistent pattern. All coefficient estimates become closer to zero and have a reduced standard deviation. This effect is confirmed by examining the four sets of pdfs in Figure 4. There are three pdfs in each set. One is the posterior pdf with the prior information on the P_i included, estimated from the Metropolis-Hastings draws. A second is obtained in the same way, but it is the posterior pdf from a diffuse uniform prior on β. The third is a normal distribution centered at the maximum likelihood estimate, and with standard deviation equal to its standard error. One can argue that the sampling theorist acts as if this last pdf represents his or her subjective knowledge about an element in β. The last two pdfs are very similar, as one might expect. The modes do not correspond exactly because the mode of the ML pdf is taken as the mode of the joint posterior pdf whereas the mode of the posterior that came from a diffuse prior is the mode of a marginal posterior pdf. Comparing the first pdf with the last two, we see that, in all cases, including the prior information has led to a shift towards zero and more precise estimation.

The reason for the shift towards zero becomes clear after some reflection. A uniform prior on β places a proportionately large weight on very big negative and very big positive values of these coefficients. Consequently, it implies a prior on the P_i which is *U*-shaped, going off to infinity at zero and one. Our almost uniform prior on the P_i reduces the large weight placed on extreme values of β and hence moves the posterior pdfs towards zero. Other results are also consistent with this observation.

In Table 7 the posterior means and standard deviations for some probabilities and elasticities, for the diffuse prior case and the proper prior case, are given for some selected states. The selected states are the seven with the largest population. Corresponding posterior pdfs for some of these cases are given in Figures 5 and 6. The elasticities are those for the percentage of the population with Bachelor's degrees and the percentage of the population that lives in a metropolitan region. Overall, the greater the level of education, the higher the probability of voting Republican, and, the greater the degree of urbanization, the higher the probability of voting Democrat. However, the relatively high standard errors do imply considerable uncertainty about the values. Consistent with our observations about β, the introduction of prior information on the P_i shifts the posterior pdfs for the P_i towards 0.5. For the states

where the pdfs were initially concentrated towards one (California, Illinois, New York and Pennsylvania), the shift is dramatic. The pdfs change from having a sharp mode at one to being more spread out and centered around a much lower value. In the other cases, where the initial pdfs had an internal mode, the shifts are less dramatic, but still considerable, and the posterior standard deviations decline rather than increase.

Table 7: Probability and Elasticity Estimates for Selected States

	Probs (P_i)		Elasticity (Bach)		Elasticity (Metrop)	
	Diffuse Prior	Proper Prior	Diffuse Prior	Proper Prior	Diffuse Prior	Proper Prior
California	0.94	0.63	–0.21	–0.28	0.24	0.27
	(0.09)	(0.11)	(0.47)	(0.56)	(0.36)	(0.39)
Florida	0.36	0.46	–0.71	–0.26	2.21	0.31
	(0.21)	(0.12)	(2.13)	(0.61)	(1.32)	(0.49)
Illinois	0.96	0.66	–0.08	–0.19	0.15	0.21
	(0.04)	(0.06)	(0.20)	(0.44)	(0.15)	(0.30)
New York	0.99	0.71	–0.04	–0.20	0.04	0.20
	(0.02)	(0.08)	(0.11)	(0.41)	(0.08)	(0.28)
Ohio	0.87	0.62	–0.06	–0.14	0.36	0.20
	(0.11)	(0.09)	(0.48)	(0.42)	(0.29)	(0.31)
Pennsylvania	0.98	0.68	–0.04	–0.15	0.07	0.19
	(0.03)	(0.07)	(0.11)	(0.38)	(0.09)	(0.27)
Texas	0.61	0.51	–0.55	–0.26	1.24	0.29
	(0.15)	(0.09)	(1.22)	(0.57)	(0.79)	(0.42)

The behaviour of the elasticities can also be related to whether or not the posterior pdf for the P_i (obtained from the diffuse prior) is concentrated towards one with a mode at one. Recall that the expression for an elasticity is

$$E_{ki} = \beta_k x_{ki} \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)} \ .$$

When P_i is concentrated at one, $\phi(x_i'\beta)$ will be concentrated towards zero and $\Phi(x_i'\beta)$ will be concentrated towards one. Accordingly, for the case of diffuse prior information, the posterior pdfs for the E_{ki} for California, Illinois, New York and Pennsylvania have sharp peaks at zero. Introducing the prior information leads to a

posterior pdf with a more regular shape, but brings with it an increase in uncertainty. For Florida and Texas, and to some extent Ohio, there is a different effect. In these cases the posterior pdf from the proper prior is centred more around zero and has a lower posterior standard deviation. The centering close to zero is likely to be a consequence of the location of the posterior pdf for β_k rather than the values of $\phi(x_i'\beta)$ and $\Phi(x_i'\beta)$ which will no longer be concentrated at zero and one, respectively.

We have demonstrated that the introduction of seemingly mild prior information about the P_i can have a considerable impact on the posterior pdfs for β, P_i and E_{ki}. When a uniform diffuse prior on β is specified, most of the prior weight for the P_i is at one or zero. Such a prior is specified not because we believe that one or zero are the most likely values, but because we want the information contained in the data, about the relationship between P_i and x_i, to dominate the posterior pdf. We are asking the data to convince us that P_i is not one or zero. A beta prior pdf like those we have specified seems a more reasonable representation of likely prior information, but has a large impact on the final posterior pdfs.

3. CONCLUSION

The probit model is a common one for modeling binary choice decisions in economics. Although Bayesian estimation of this model has been addressed in the statistics and econometrics literature, estimation using alternative kinds of prior information has received little attention. Our contribution and findings can be summarized as follows:

1. We have illustrated how prior information in the form of inequality constraints on the coefficients can be included in the estimation procedure.
2. Methodology for putting prior information on choice probabilities rather than coefficients has been described and applied.
3. For estimation, we focused not just on the coefficients, but also on choice probabilities and elasticities for a given set of explanatory variables.
4. One advantage of Bayesian inference is that the support for posterior pdfs on choice probabilities is the interval (0,1). Asymptotic confidence

intervals obtained from maximum likelihood estimation can include negative ranges or values where a probability exceeds unity.

5. Bayesian and maximum likelihood inference can lead to very different conclusions about the reliability of estimation of elasticities. Like in the previous point, most of the difference appears to be attributable to using finite sample inference rather than an asymptotic approximation. However, using Bayesian inference to impose what seems to be a relatively mild sign restriction on a coefficient can create a more noticeable truncation of the posterior pdf of the elasticity.

6. Placing prior information on the choice probabilities, rather than the coefficients, can have a dramatic impact on the posterior pdfs for the coefficients, the choice probabilities and the elasticities.

ACKNOWLEDGEMENT

Valuable comments were given by Denzil Fiebig and other contributors to the Econometrics Workshop at the University of Sydney.

REFERENCES

Albert, J., and S. Chib (1993), "Bayesian Analysis of Binary and Polychotomous Response Data", *Journal of the American Statistical Association*, 88, 669-679.

Amemiya, T. (1981), "Qualitative Response Models: A Survey", *Journal of Economic Liberature*, 19, 1483-1536.

Dhillon, U.S., J.D. Shilling, and C.F. Sirmans (1987), "Choosing Between Fixed and Adjustable Rate Mortgages", *Journal of Money, Credit and Banking*,19, 260-267.

Evans, M., N. Hastings and B. Peacock (1993), *Statistical Distributions*, 2nd edition, New York: John Wiley and Sons.

Geweke, J. (1999), "Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication", *Econometric Reviews*, 18, 1-74.

Geweke, J., M. Keane, and D. Runkle (1994), "Alternative Computational Approaches to Inference in the Multinomial Probit Model," *Review of Economics and Statistics*, 76, 609-632.

Geweke, J., M. Keane, and D. Runkle (1997), "Statistical Inference in the Multinomial Multiperiod Probit Model," *Journal of Econometrics*, 80, 125-166.

Gewke, J. (1991), "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints", in E.M. Keramidas and S.M. Kaufman, editors, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Seattle: Interface Foundation of America, 571-578.

Greene, W. (1990), *Econometric Analysis*, New York: Macmillan.

Hill, R.C. (1996), editor, Advances in *Econometrics Volumn 11A: Bayesian Computational Methods and Applications*, Greenwich, JAI Press.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.-C. Lee (1985), *The Theory and Practice of Econometrics*, second edition, New York: John Wiley and Sons.

Kleibergen, F., and H.K. van Dijk (1988), "Bayesian Simultaneous Equations Analysis using Reduced Rank Structures", *Econometric Theory*, 14, 701-743.

Lott, W.F., and S.C. Ray (1992), *Applied Econometrics: Problems with Data Sets*, Orlando: Dryden Press.

Maddala, G.S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, London: Cambridge University Press.

Zellner, A., and P.E. Rossi (1984), "Bayesian Analysis of Dichotomous Quantal Response Models", *Journal of Econometrics*, 25, 365-393.

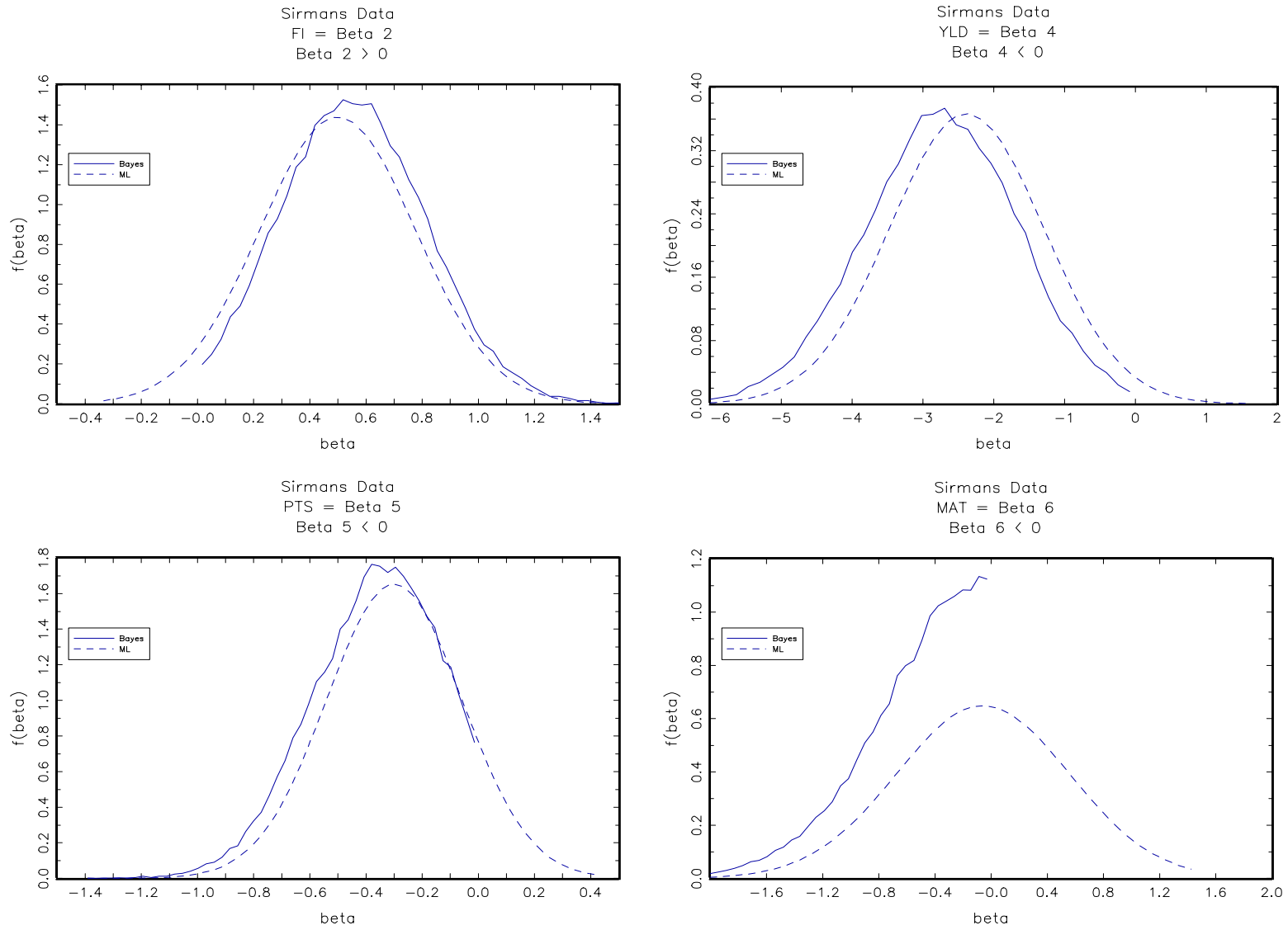

Figure 1 Posterior pdfs for Coefficients for Mortgage Data

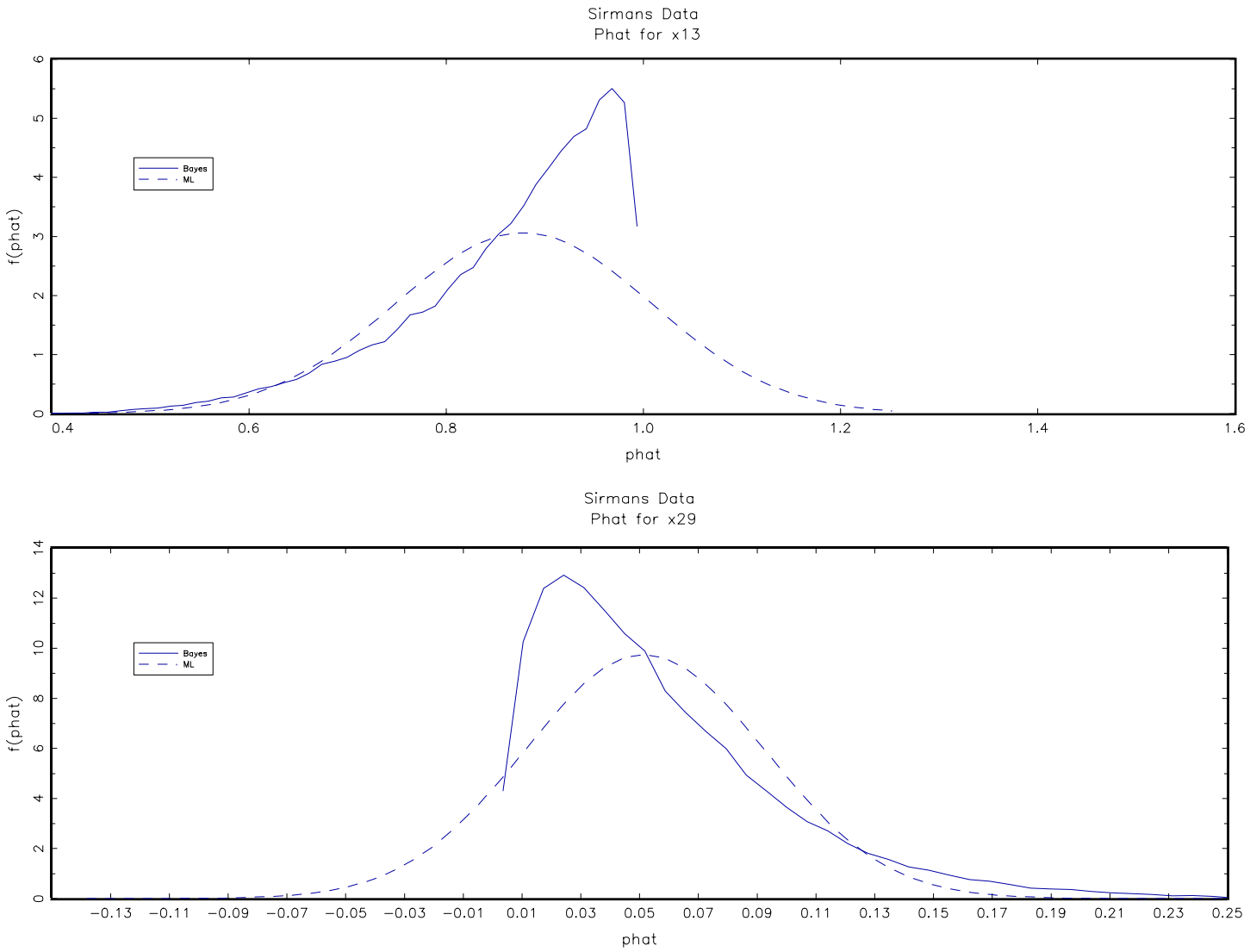

Figure 2 Posterior pdfs for Probabilities for Mortgage Data

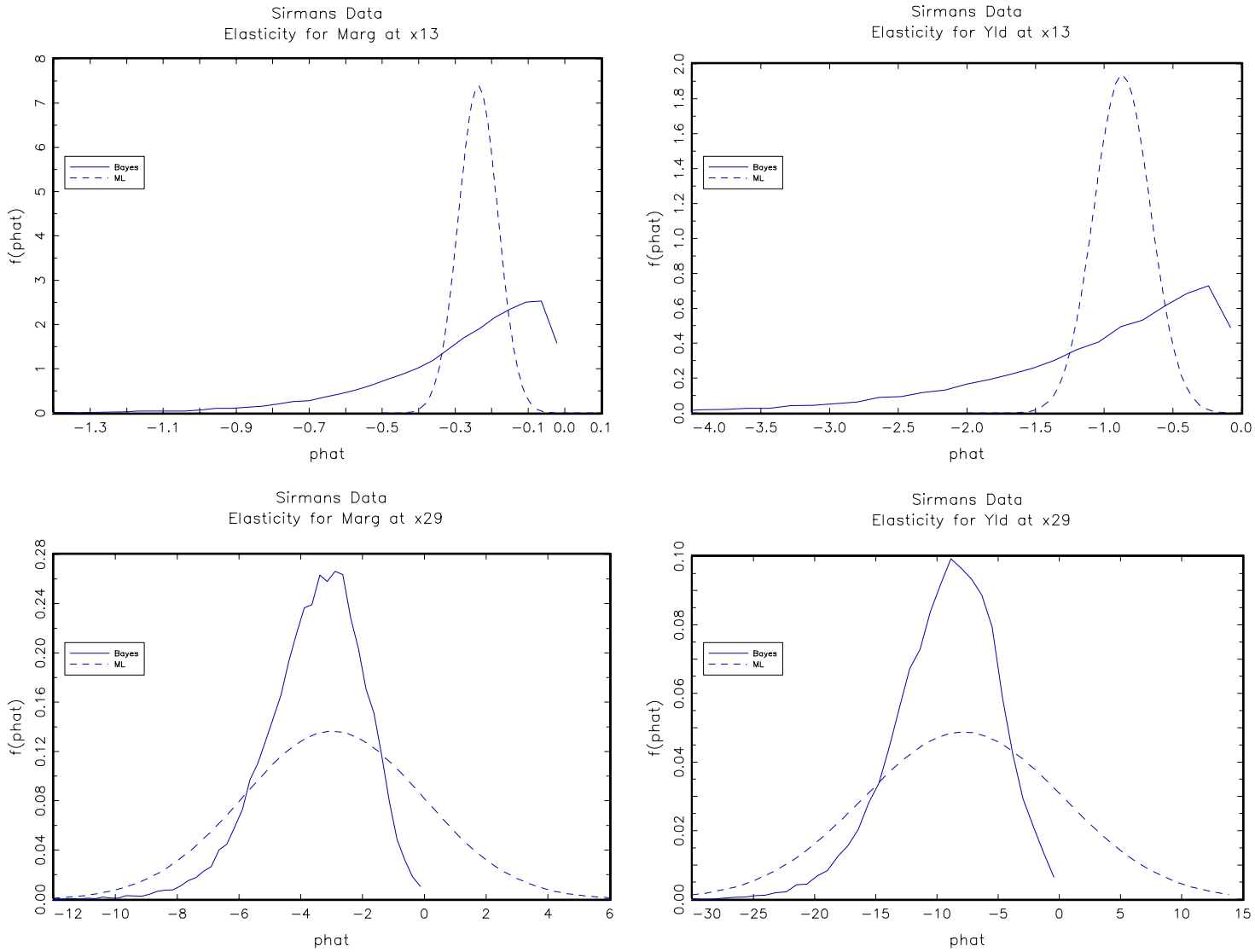

Figure 3 Posterior pdfs for Elasticities for Mortgage Data

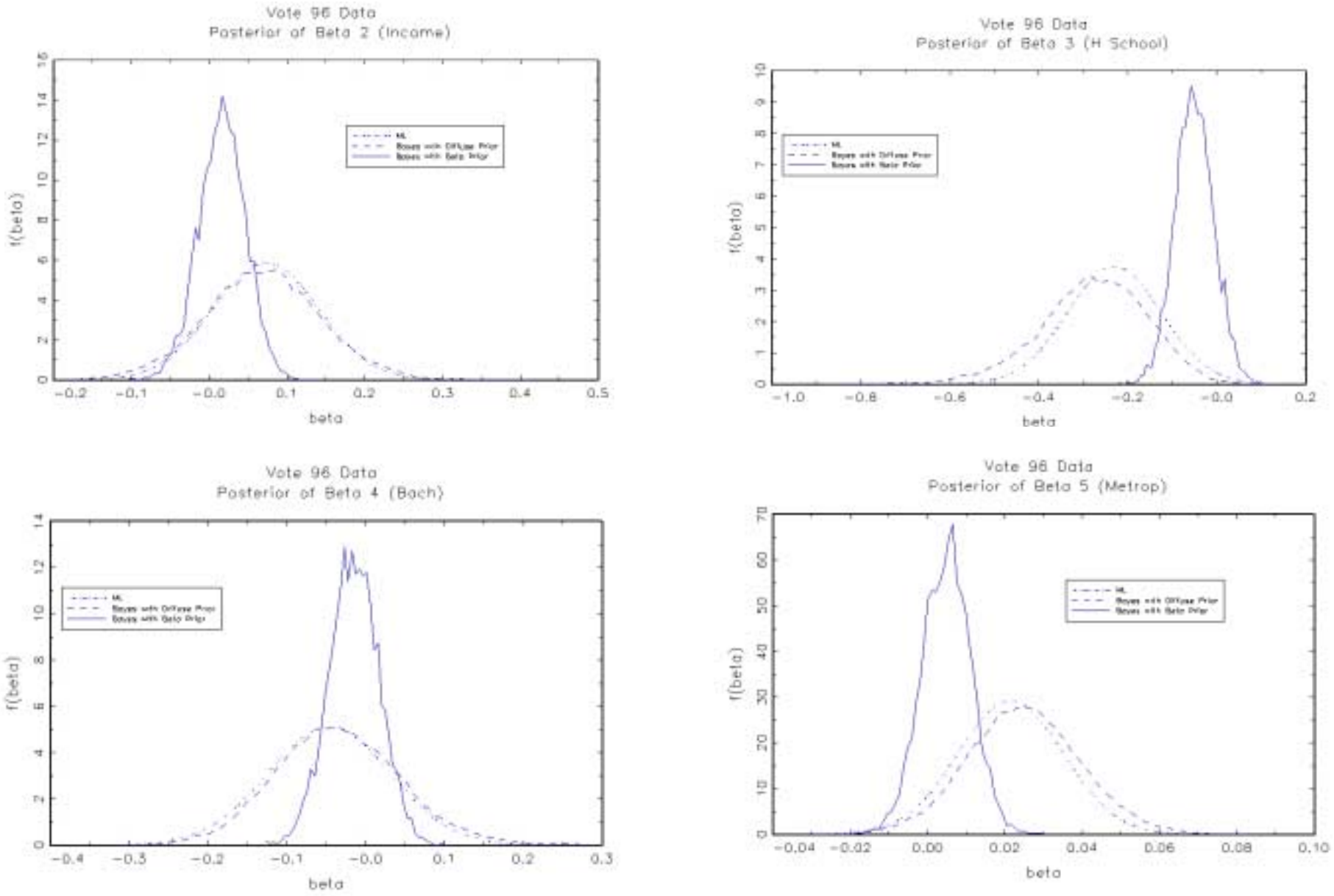

Figure 4 Posterior pdfs for Coefficients for Vote Data

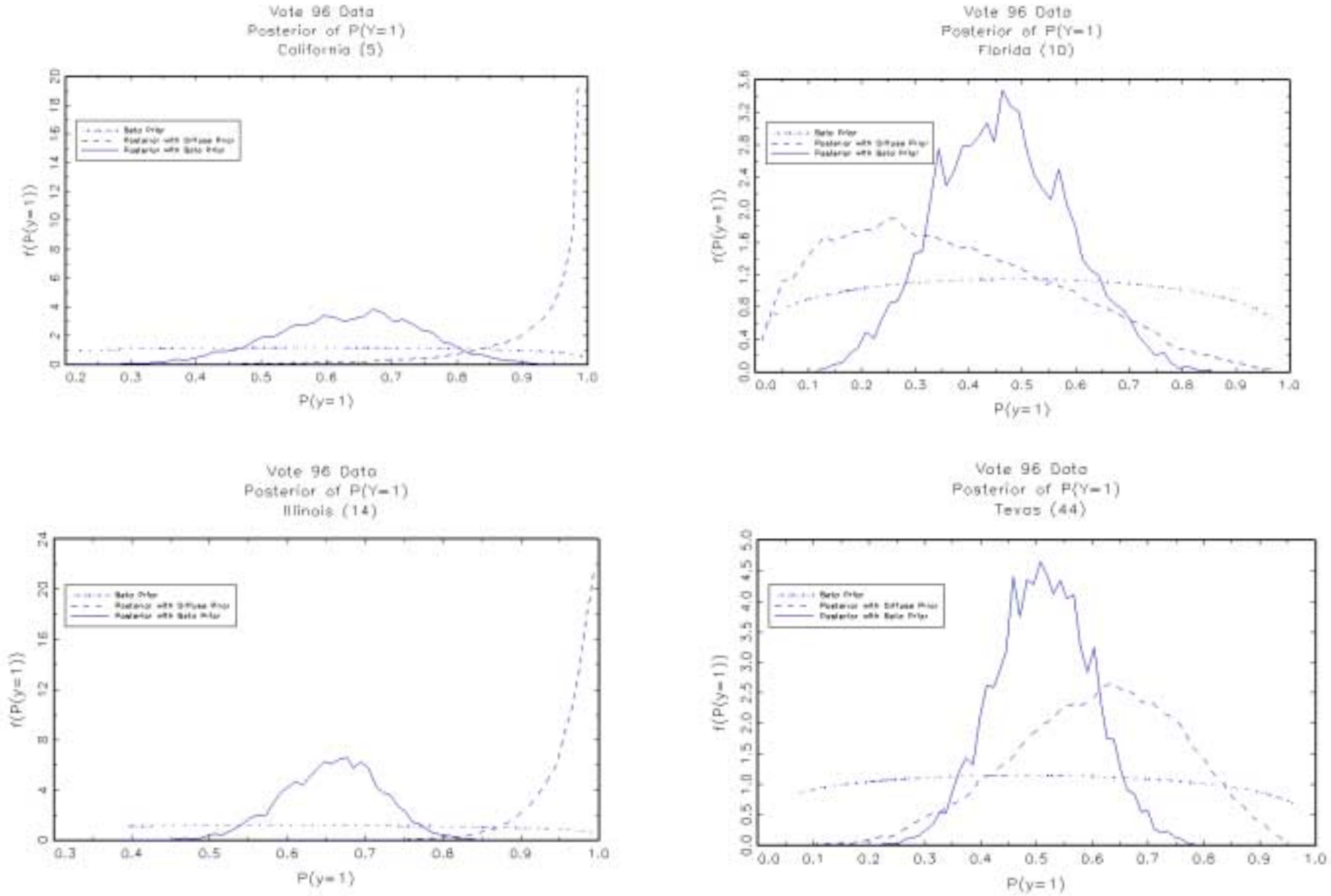

Figure 5 Posterior pdfs for Probabilities for Vote Data

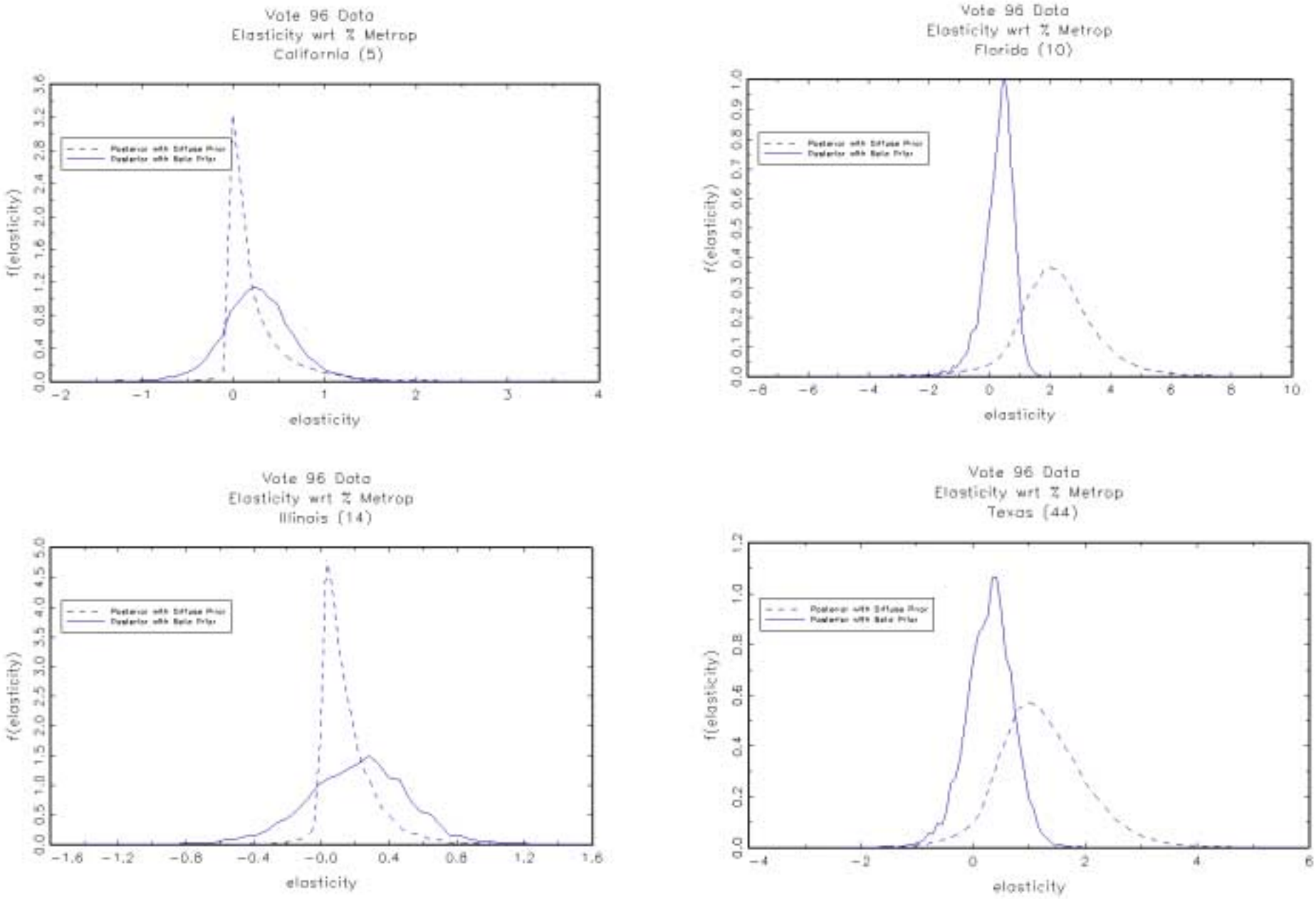

Figure 6 Posterior pdfs for Elasticities for Vote Data