Federal Reserve Bank of Minneapolis
Research Department Staff Report 237

August 1997

Mixture of Normals Probit Models

John Geweke*

University of Minnesota
and Federal Reserve Bank of Minneapolis

Michael Keane*

University of Minnesota
and Federal Reserve Bank of Minneapolis

ABSTRACT

This paper generalizes the normal probit model of dichotomous choice by introducing mixtures of normals distributions for the disturbance term. By mixing on both the mean and variance parameters and by increasing the number of distributions in the mixture these models effectively remove the normality assumption and are much closer to semiparametric models. When a Bayesian approach is taken, there is an exact finite-sample distribution theory for the choice probability conditional on the covariates. The paper uses artificial data to show how posterior odds ratios can discriminate between normal and nonnormal distributions in probit models. The method is also applied to female labor force participation decisions in a sample with 1,555 observations from the PSID. In this application, Bayes factors strongly favor mixture of normals probit models over the conventional probit model, and the most favored models have mixtures of four normal distributions for the disturbance term.

JEL classification: Primary, C25; secondary, C11
Keywords: Discrete choice, Markov chain Monte Carlo, Normal mixture

*Geweke wishes to acknowledge financial support from National Science Foundation grant SBR-9514865, Keane from NSF grant SBR-9511186. We thank the editor, Robin Carter, Daniel Houser, and seminar participants at the University of Western Ontario for comments on an earlier draft, but blame only each other for remaining errors. We thank John Landon-Lane and Lance Schibilla for research assistance. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

In econometric specifications of dichotomous choice models, the probit and logit specifications are commonly used. Other specifications have been suggested (Maddala 1983, pp. 27–32; Aldrich and Nelson 1984), but in econometric applications, the probit and logit specifications have been used almost exclusively. The probit model is easy to use, and the logit specification is even more tractable. Because the probit specification can be made free of the problem of independence of irrelevant alternatives (Hausman and McFadden 1984) when moving from dichotomous to polytomous choice, whereas the logit specification cannot, the probit model has become a mainstay in econometrics, and it is likely to remain one.

It is widely appreciated that any misspecification of functional form in a dichotomous choice model will lead to inconsistent estimates of conditional choice probabilities. In particular, if a probit specification is maintained when a different specification is true, misleading inferences about conditional choice probabilities and the effects of changes in covariates on these probabilities may result. In this paper, we consider strict generalizations of the probit specification that remain within the class of linear dichotomous choice models:

$$\mathrm{P}\left(d_t = 1\right) = \mathrm{P}\left(\beta'\mathbf{x}_t + \varepsilon_t > 0\right)$$

where d_t is the choice indicator, $\mathbf{x}_t$ is a vector of covariates, and ε_t is an i.i.d. disturbance. Our approach is fully parametric and Bayesian. This approach permits us to obtain explicit evaluations of $\mathrm{P}\left(d_t = 1 \middle| \mathbf{x}_t = \mathbf{x}^*\right)$ for any $\mathbf{x}^*$, which are ultimately the focus of any application, and makes it possible to compare alternative generalizations with each other as well as with the conventional probit model. These evaluations and comparisons can, in principle, be accomplished with all Bayesian approaches to dichotomous choice. (See, for example, Zellner and Rossi 1984; Albert and Chib 1993; Koop and Poirier 1993.) However, recent developments in numerical methods have greatly simplified the computation of posterior moments and Bayes factors. The class of specifications taken up here is the mixture of normals distribution, in which

$$\mathrm{p}(\varepsilon) = (2\pi)^{-1/2} \sum\nolimits_{j=1}^{m} p_j h_j^{1/2} \exp\left[-.5 h_j \left(\varepsilon - \alpha_j\right)^2\right].$$

The generation of the shock may be described by first drawing from one of m specified normal distributions, with probabilities $p_1, \ldots, p_m$, and then drawing the shock ε from that distribution.

There is a large literature in econometric theory that has taken a semiparametric approach to this problem by dealing with consistent estimation of β and making regularity assumptions about the distribution of the shock rather than specifying it. The development of this approach includes Cosslett (1983), Manski (1985), Gallant and Nychka (1987), Powell, Stock, and Stoker (1989), Horowitz (1992), Ichimura (1993), and Klein and Spady (1993). Lewbel (1997) extends this approach to consistent estimation of moments of ε. This approach and the one taken here are complementary. Both break free of the normality assumption of the probit model. On the one hand, the semiparametric approach introduces a series of approximations that yields consistent estimates of covariate coefficients given weak assumptions (for example, differentiability of the density function) about the shock distribution, whereas we make no claim that the mixture of normals family will similarly accommodate all such distributions. On the other hand, our method leads directly to exact inference for $\mathrm{P}\left(d_t = 1 \middle| \mathbf{x}_t = \mathbf{x}^*\right)$, which the semiparametric approach does not do even asymptotically.

The organization of the paper is simple. The next section describes the mixture of normals probit model, beginning by extending the treatment of the probit model by Albert and Chib (1993) and moving through to the development of a Markov chain Monte Carlo posterior simulator and the evaluation of the marginal likelihood. Experiments with artificial data provide some evidence on the ability of different models within the class to cope with alternative shock distributions. These are reported in Section 3. A substantive example pertaining to women's labor force participation, which uses a subset of the Panel Study of Income Dynamics (PSID) with 1,555 observations, is presented in Section 4. In this example, the conventional probit model is overwhelmingly rejected in favor of the mixture of normals probit model; several members of the mixture family have low Bayes factors relative to one another, including mixtures of as many as 5 normals with 13 free parameters describing the shape of the shock distribution. Some fairly obvious extensions of this work are mentioned in the concluding section.

2. Bayesian inference for probit models

In the probit model, the observables are $\underset{k\times T}{\mathbf{X}'} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_T\right]$ and $\mathbf{d}' = \left(d_1, \ldots, d_T\right)$, $d_t = 0$ or 1. The relationships of interest are

(2.1) $$\tilde{y}_t = \beta'\mathbf{x}_t + \varepsilon_t, \quad d_t = \chi_{[0,\infty)}\left(\tilde{y}_t\right) \quad (t = 1, \ldots, T)$$

where the indicator function $\chi_S(z) = 1$ if $z \in S$ and $\chi_S(z) = 0$ if $z \notin S$. Let $\tilde{\mathbf{y}}' = \left(\tilde{y}_1, \ldots, \tilde{y}_T\right)$.

2.1 Conventional probit model

In the conventional probit model,

$$(2.2) \qquad \varepsilon_t | \mathbf{X} \overset{IID}{\sim} \mathrm{N}(0,1)$$

and it is convenient to complete this model with the prior distribution $\beta \sim \mathrm{N}\left(\underline{\beta}, \underline{\mathbf{H}}_\beta^{-1}\right)$, where $\underline{\beta} \in \mathrm{R}^k$ and $\underline{\mathbf{H}}_\beta$ is a $k \times k$ positive definite precision matrix:

$$(2.3) \qquad \mathrm{p}(\beta) = (2\pi)^{-k/2} \left|\underline{\mathbf{H}}_\beta\right|^{1/2} \exp\left[-.5\left(\beta - \underline{\beta}\right)' \underline{\mathrm{H}}_\beta \left(\beta - \underline{\beta}\right)\right].$$

Albert and Chib (1993) develop a Gibbs sampling algorithm for Bayesian inference in this model. From (2.1) and (2.2),

$$\mathrm{p}\left(\mathbf{d}, \tilde{\mathbf{y}} | \beta, \mathbf{X}\right) = \mathrm{p}\left(\tilde{\mathbf{y}} | \beta, \mathbf{X}\right) \mathrm{p}\left(\mathbf{d} | \tilde{\mathbf{y}}\right)$$

$$(2.4) \qquad = (2\pi)^{-T/2} \exp\left[-.5(\tilde{\mathbf{y}} - \mathbf{X}\beta)(\tilde{\mathbf{y}} - \mathbf{X}\beta)\right] \prod\nolimits_{t=1}^{T} \left[d_t \chi_{[0,\infty)}(\tilde{y}_t) + (1 - d_t)\chi_{(-\infty,0)}(\tilde{y}_t)\right]$$

$$(2.5) \qquad = \prod\nolimits_{t=1}^{T} (2\pi)^{-1/2} \exp\left[-.5\left(\tilde{y}_t - \beta' \mathbf{x}_t\right)^2\right] \left[d_t \chi_{[0,\infty)}(\tilde{y}_t) + (1 - d_t)\chi_{(-\infty,0)}(\tilde{y}_t)\right].$$

The joint posterior density for β and $\tilde{\mathbf{y}}$ is $\mathrm{p}(\beta, \tilde{\mathbf{y}} | \mathbf{d}, \mathbf{X}) \propto \mathrm{p}\left(\mathbf{d}, \tilde{\mathbf{y}} | \beta, \mathbf{X}\right) \mathrm{p}(\beta)$. When we take the product of the joint density (2.4) and the prior density (2.3) and examine the kernel of this expression in β,

$$(2.6) \qquad \beta | (\tilde{\mathbf{y}}, \mathbf{X}) \sim \mathrm{N}\left(\overline{\beta}, \overline{\mathbf{H}}_\beta^{-1}\right), \quad \overline{\mathbf{H}}_\beta = \underline{\mathbf{H}}_\beta + \mathbf{X}'\mathbf{X}, \quad \overline{\beta} = \overline{\mathbf{H}}_\beta^{-1}\left(\underline{\mathbf{H}}_\beta \underline{\beta} + \mathbf{X}'\tilde{\mathbf{y}}\right).$$

Taking $\mathrm{p}\left(\mathbf{d}, \tilde{\mathbf{y}} | \beta, \mathbf{X}\right)$ in the form (2.5), we have the T conditionally independent distributions

$$(2.7) \quad \tilde{y}_t | (\mathbf{d}, \beta, \mathbf{X}) = \tilde{y}_t | (d_t, \beta, \mathbf{x}_t) \sim \mathrm{N}(\beta'\mathbf{x}_t, 1) \text{ subject to } \tilde{y}_t \geq 0 \text{ if } d_t = 1 \text{ and } \tilde{y}_t < 0 \text{ if } d_t = 0.$$

Beginning from an arbitrary point $\beta^{(0)} \in \mathrm{R}^k$, we construct the sequence $\left\{\beta^{(r)}, \tilde{\mathbf{y}}^{(r)}\right\}$ by drawing sequentially from the T distributions in (2.7) and the distribution (2.6). (In all results reported subsequently for this model, $\beta^{(0)} \sim \mathrm{N}\left(\underline{\beta}, \underline{\mathbf{H}}^{-1}\right)$.) Given any point $\left(\beta^*, \tilde{\mathbf{y}}^*\right)$ in the support of $\mathrm{p}(\beta, \tilde{\mathbf{y}} | \mathbf{d}, \mathbf{X})$ and any subset A of the support with positive Lebesgue measure, the probability of moving from $\left(\beta^*, \tilde{\mathbf{y}}^*\right)$ into A in one iteration of this algorithm is strictly positive. Therefore, the process $\left\{\beta^{(r)}, \tilde{\mathbf{y}}^{(r)}\right\}$ is ergodic (Tierney 1994; Geweke 1997a, Section 3.3), which implies that if $\mathrm{E}\left[\mathrm{g}(\beta) | \mathbf{d}, \mathbf{X}\right]$ exists, then $R^{-1} \sum_{m=1}^{M} \mathrm{g}\left(\beta^{(m)}\right) \xrightarrow{a.s.} \mathrm{E}\left[\mathrm{g}(\beta) | \mathbf{d}, \mathbf{X}\right]$.

2.2 Mixture of normals probit model

In the mixture of normals probit model,

$$(2.8) \quad \varepsilon_t = \sum\nolimits_{j=1}^{m} e_{tj}\left(\alpha_j + h_j^{-1/2}\eta_t\right), \quad \alpha' = (\alpha_1, \ldots, \alpha_m) \in \mathrm{R}^m, \quad \mathbf{h}' = (h_1, \ldots, h_m) \in \mathrm{R}_+^m$$

$$(2.9) \qquad \eta_t | \mathbf{X} \overset{IID}{\sim} \mathrm{N}(0,1).$$

The random vectors $\mathbf{e}_t' = (e_{t1},\ldots,e_{tm})$ are i.i.d., each with a multinomial distribution with parameters $p_J = \mathrm{P}(e_{tj} = 1)$ $(j = 1,\ldots,m)$:

$$(2.10) \qquad \mathbf{e}_t \overset{IID}{\sim} \mathrm{MN}(p_1,\ldots,p_m), \quad \mathbf{p}' = (p_1,\ldots,p_m) \in S_m$$

where S_m is the unit simplex in R^m.

Without further restrictions, the model is clearly unidentified in the sense that more than one set of values of the parameters in (2.1) and (2.8)–(2.10) imply the same $\mathrm{p}(\mathbf{d}|\mathbf{X})$. Three specific identified versions of the model are of interest, each consisting of a set of further restrictions on (2.1) and (2.8)–(2.10).

In the *full mixture of normals model*,

(i) rank$(\mathbf{X}) = k$ and $\mathbf{a}'\mathbf{X}' \neq (1,\ldots,1)$ for any $k \times 1$ vector **a**;

(ii) $p_j > 0 \; \forall \; j$;

(iii) the support of $\beta'\mathbf{x}_t$ is a set of positive Lebesgue measure;

(iv) either

(a) $\alpha_{j-1} < \alpha_j$ $(j = 2,\ldots,m)$ or

(b) $h_{j-1} < h_j$ $(j = 2,\ldots,m)$; and

(v) $h_j^* = 1$ for some j^*.

In the *scale mixture of normals model*, $\alpha_j = 0$ $(j = 1,\ldots,m)$, **X** may (and generally does) include an intercept, and (iv-b) obtains. In the *mean mixture of normals model*, $h_j = 1$ $(j = 1,\ldots,m)$ and (iv-a) obtains. The orderings in (iv) are labeling restrictions that prevent interchanging the components of the mixture; obviously, other labeling restrictions are possible.

For Bayesian inference, it is convenient to complete the model with independent prior distributions for β, α, **h**, and **p**. The prior distribution for β is (2.3). Except in the scale mixture of normals model, $\alpha \sim \mathrm{N}(\underline{\alpha}, \underline{\mathbf{H}}_\alpha^{-1})$, where $\underline{\alpha} \in \mathrm{R}^m$ and $\underline{\mathbf{H}}_\alpha$ is an $m \times m$ positive definite matrix:

$$(2.11) \qquad \mathrm{p}(\alpha) = (2\pi)^{-m/2} |\underline{\mathbf{H}}_\alpha|^{1/2} \exp\left[-.5(\alpha - \underline{\alpha})' \underline{\mathbf{H}}_\alpha (\alpha - \underline{\alpha})\right]$$

subject to (iv-a) in the mean mixture of normals model and in the full mixture of normals model if (iv-b) is not invoked.

Except in the mean mixture of normals model, $\underline{s}_j^2 h_j \sim \chi^2(\underline{\nu}_j)$ $(j \neq j^*)$, where $\underline{s}_j^2 > 0$ and $\underline{\nu}_j > 0$. If (iv-b) is not imposed, then the prior density of **h** is

$$(2.12) \qquad \mathrm{p}(\mathbf{h}) = \prod\nolimits_{j=1, j \neq j^*}^{m} \left[2^{\underline{\nu}_j/2} \Gamma(\underline{\nu}_j/2)\right]^{-1} (\underline{s}_j^2)^{\underline{\nu}_j/2} h_j^{(\underline{\nu}_j - 2)/2} \exp(-.5\underline{s}_j^2 h_j).$$

If (iv-b) is imposed, then the support is truncated accordingly, and the normalization constant must be adjusted.

Finally, $\mathrm{p}(\mathbf{p}) \sim \mathrm{Beta}(\mathbf{r})$, $\mathbf{r} \in \mathrm{R}_+^m$:

$$\text{(2.13)} \qquad \mathrm{p}(\mathbf{p}) = \left[\Gamma\left(\textstyle\sum_{j=1}^{m} r_j\right) \Big/ \textstyle\prod_{j=1}^{m} \Gamma\left(r_j\right)\right] \textstyle\prod_{j=1}^{m} p_j^{\left(r_j-1\right)}.$$

Since the likelihood function

$$\text{(2.14)} \quad \mathrm{p}\left(\mathbf{d} \middle| \beta, \alpha, \mathbf{h}, \mathbf{p}, \mathbf{X}\right) = \textstyle\prod_{t=1}^{T} \Big\langle d_t \textstyle\sum_{j=1}^{m} p_j \Phi\left[h_j^{1/2}\left(\alpha_j + \beta' \mathbf{x}_t\right)\right] \\ + \left(1 - d_t\right)\Big\{1 - \textstyle\sum_{j=1}^{m} p_j \Phi\left[h_j^{1/2}\left(\alpha_j + \beta' \mathbf{x}_t\right)\right]\Big\}\Big\rangle$$

is bounded between zero and one,

$$\text{(2.15)} \qquad \mathrm{p}\left(\mathbf{d} \middle| \beta, \alpha, \mathbf{h}, \mathbf{p}, \mathbf{X}\right) \mathrm{p}(\beta)\, \mathrm{p}(\alpha)\, \mathrm{p}(\mathbf{h})\, \mathrm{p}(\mathbf{p})$$

is finitely integrable over its support, and, consequently, the posterior distribution $\mathrm{p}(\beta, \alpha, \mathbf{h}, \mathbf{p} | \mathbf{d}, \mathbf{X})$, proportional to (2.15), exists. Since β, α, $\mathbf{h}$, and $\mathbf{p}$ have prior moments of all orders, they also have posterior moments of all orders. And since for any specified $\mathbf{x}_{T+s}\ (s > 0)$, $\mathrm{p}\left(d_{T+s} = 1 \middle| \beta, \alpha, \mathbf{h}, \mathbf{p}, \mathbf{x}_{T+s}\right)$ is bounded between zero and one, $\mathrm{p}\left(d_{T+s} = 1 \middle| \mathbf{x}_{T+s}, \mathbf{X}\right)$ has posterior moments of all orders.

2.3 A posterior simulator

In the mixture of normals probit model,

$$\mathrm{p}\left(\mathbf{d}, \tilde{\mathbf{y}}, \mathbf{e} \middle| \beta, \alpha, \mathbf{h}, \mathbf{p}, \mathbf{X}\right) = \mathrm{p}\left(\mathbf{e} \middle| \mathbf{p}\right) \mathrm{p}\left(\tilde{\mathbf{y}} \middle| \mathbf{e}, \beta, \alpha, \mathbf{h}, \mathbf{X}\right) \mathrm{p}\left(\mathbf{d} \middle| \tilde{\mathbf{y}}\right).$$

When we define $L_t = \left(j\!:\!e_{tj} = 1\right)$ and $T_j = \sum_{t=1}^{T} e_{tj}$

$$\text{(2.16)} \qquad \mathrm{p}\left(\mathbf{e} \middle| \mathbf{p}\right) = \textstyle\prod_{t=1}^{T} \prod_{j=1}^{m} p_j^{e_{tj}} = \prod_{j=1}^{m} p_j^{T_j}$$

$$\text{(2.17)} \qquad \mathrm{p}\left(\tilde{\mathbf{y}} \middle| \mathbf{e}, \beta, \alpha, \mathbf{h}, \mathbf{X}\right) = (2\pi)^{-T/2} \textstyle\prod_{j=1}^{m} h_j^{T_j/2} \exp\left[-.5 \sum_{t=1}^{T} h_{L_t}^{1/2} \left(\tilde{y}_t - \alpha' \mathbf{e}_t - \beta' \mathbf{x}_t\right)^2\right]$$

$$\text{(2.18)} \qquad \mathrm{p}\left(\mathbf{d} \middle| \tilde{\mathbf{y}}\right) = \textstyle\prod_{t=1}^{T} \left[d_t \chi_{[0,\infty)}\left(\tilde{y}_t\right) + \left(1 - d_t\right) \chi_{(-\infty,0)}\left(\tilde{y}_t\right)\right].$$

The product of (2.16), (2.17), and (2.18) and the prior density kernels (2.3), (2.11), (2.12), and (2.13), is a kernel of the posterior distribution of the latent variables $\mathbf{e}$ (equivalently, $\left\{L_t\right\}_{t=1}^{T}$) and $\tilde{\mathbf{y}}$ and the parameter vectors β, α, $\mathbf{h}$, and $\mathbf{p}$. Posterior distributions for individual groups of latent variables and parameters, conditional on all the other latent variables and parameters and the data, are easily derived from these expressions as follows.

The kernel in $\tilde{\mathbf{y}}$ is the product of (2.17) and (2.18), from which the $\tilde{y}_t$ are conditionally independent with

$$\tilde{y}_t \sim \mathrm{N}\left(\beta' \mathbf{x}_t + \alpha' \mathbf{e}_t,\, h_{L_t}^{-1}\right) \text{ subject to } \tilde{y}_t \geq 0 \text{ if } d_t = 1 \text{ and } \tilde{y}_t < 0 \text{ if } d_t = 0.$$

The kernel in $\left\{L_t\right\}_{t=1}^{T}$ is the product of (2.16) and (2.17), which shows that the L_t are conditionally independent with

$$\mathrm{P}\left(L_t = j\right) \propto p_j \exp\left[-.5 h_j^{1/2} \left(\tilde{y}_t - \alpha_j - \beta' \mathbf{x}_t\right)^2\right].$$

The kernel in β is the product of (2.3) and (2.17), from which

$$\beta \sim \mathrm{N}\left(\overline{\beta}, \overline{\mathbf{H}}_{\beta}^{-1}\right), \quad \overline{\mathbf{H}}_{\beta} = \underline{\mathbf{H}}_{\beta} + \sum_{t=1}^{T} h_{L_t} \mathbf{x}_t \mathbf{x}_t', \quad \overline{\beta} = \overline{\mathbf{H}}_{\beta}^{-1}\left(\underline{\mathbf{H}}_{\beta}\underline{\beta} + \sum_{t=1}^{T} h_{L_t} \mathbf{x}_t \tilde{y}_t\right).$$

The kernel in α is the product of (2.11) and (2.17), which yields

$$\alpha \sim \mathrm{N}\left(\overline{\alpha}, \overline{\mathbf{H}}_{\alpha}^{-1}\right), \quad \overline{\mathbf{H}}_{\alpha} = \underline{\mathbf{H}}_{\alpha} + \sum_{t=1}^{T} \mathbf{e}_t \mathbf{e}_t', \quad \overline{\alpha} = \overline{\mathbf{H}}_{\alpha}^{-1}\left[\underline{\mathbf{H}}_{\alpha}\underline{\alpha} + \sum_{t=1}^{T} \mathbf{e}_t\left(\tilde{y}_t - \beta' \mathbf{x}_t\right)\right]$$

subject to $\alpha_1 < \ldots < \alpha_m$ if this labeling restriction has been invoked. (The algorithm of Geweke (1991) provides for efficient imposition of the inequality constraints.)

The kernel in **h** is the product of (2.12) and (2.17), which indicates

$$\overline{s}_j^2 h_j \sim \chi^2\left(\overline{\nu}_j\right), \quad \overline{s}_j^2 = \underline{s}_j^2 + \sum e_{tj}\left(\tilde{y}_t - \alpha_j - \beta' \mathbf{x}_t\right)^2, \quad \overline{\nu}_j = \underline{\nu}_j + T_j$$

for $j \neq j^*$ and subject to $h_1 < \ldots < h_m$ if the labeling restriction on **h** has been invoked. Whether or not this restriction applies, it is straightforward to draw the h_j sequentially.

Finally, the posterior kernel in **p** is the product of (2.13) and (2.16),

$$\mathbf{p} \sim \mathrm{Beta}\left(r_1 + T_1, \ldots, r_m + T_m\right).$$

It is straightforward to verify that the lower semicontinuity and boundedness conditions of Roberts and Smith (1994) for ergodicity of the Gibbs samplers are satisfied by the posterior distribution. Therefore any starting value $\left(\beta^{(0)}, \alpha^{(0)}, \mathbf{h}^{(0)}, \mathbf{p}^{(0)}\right)$ may be used. As a practical matter, however, we have found that unless the dimension *k* of β is small, convergence is slow if the initial values are drawn from the respective prior distributions (2.3), (2.11), (2.12), and (2.13). In the case of a labeling restriction on **h**, the difficulty is that since $\beta^{(0)}$ is quite unrepresentative of the posterior distribution with very high probability, initial values of $h_j^{(r)}$ tend to be quite small for $j < j^*$, or, if $h_1 = 1$, initial values of $\beta_j^{(r)}$ become quite small, and many thousands of iterations are required before convergence to the posterior distribution. (A similar problem arises with a labeling restriction on α.) This difficulty is avoided by drawing $\beta^{(0)}$ from (2.3), sampling in the conventional probit model for a few hundred iterations, and then beginning a full set of draws in the mixture of normals probit model.

2.4 Comparison of models

To this point, the number of mixtures *m* has been taken as given. In fact, it will be of some interest to compare the plausibility of alternative values of *m*, and, in particular, to compare mixture of normals models (*m*>1) with the conventional probit model. It may also be of interest to compare alternative specifications of prior distributions. We can do this formally by means of Bayes factors using the extensions of the method of Gelfand and Dey (1994) outlined in Geweke (1997b).

Generically, for a model *j* with data density $\mathrm{p}_j\left(\mathbf{y} \middle| \theta_j\right)$, completed with a prior density $\mathrm{p}_j\left(\theta_j\right)$, $\theta_j \in \Theta_j$, the marginal likelihood may be defined

(2.19) $$M_j = \int_{\Theta_j} \mathrm{p}_j\left(\mathbf{y}\middle|\theta_j\right)\mathrm{p}_j\left(\theta_j\right)d\theta_j.$$

(It is important that the data and prior densities be properly normalized, that is, that $\int_Y \mathrm{p}\left(\mathbf{y}\middle|\theta_j\right)d\mathbf{y} = 1 \;\forall\; \theta_j \in \Theta_j$ and $\int_{\Theta_j} \mathrm{p}\left(\theta_j\right)d\theta_j = 1$.) The Bayes factor in favor of model *j* versus model *k* is M_j / M_k. Gelfand and Dey point out that if $\mathrm{f}_j\left(\theta_j\right)$ is any function with the property $\int_{\Theta_j} \mathrm{f}_j\left(\theta_j\right)d\theta_j = 1$, then the posterior expectation of $\mathrm{f}_j\left(\theta_j\right) \big/ \mathrm{p}_j\left(\mathbf{y}\middle|\theta_j\right)\mathrm{p}_j\left(\theta_j\right)$ in model *j* is M_j^{-1}:

$$\int_{\Theta_j} \frac{\mathrm{f}_j\left(\theta_j\right)}{\mathrm{p}_j\left(\mathbf{y}\middle|\theta_j\right)\mathrm{p}_j\left(\theta_j\right)}\mathrm{p}\left(\theta_j|\mathbf{y}\right)d\theta_j = \int_{\Theta_j} \frac{\mathrm{f}_j\left(\theta_j\right)}{\mathrm{p}_j\left(\mathbf{y}\middle|\theta_j\right)\mathrm{p}_j\left(\theta_j\right)} \cdot \frac{\mathrm{p}_j\left(\mathbf{y}\middle|\theta_j\right)\mathrm{p}_j\left(\theta_j\right)}{\int_{\Theta_j} \mathrm{p}_j\left(\mathbf{y}\middle|\theta_j\right)\mathrm{p}_j\left(\theta_j\right)}d\theta_j$$

$$= \frac{1}{\int_{\Theta_j} \mathrm{p}_j\left(\mathbf{y}\middle|\theta_j\right)\mathrm{p}_j\left(\theta_j\right)}.$$

Thus, if $\left\{\theta_j^{(r)}\right\}$ is the output of an ergodic posterior simulator,

$$R^{-1}\sum_{r=1}^{R} \mathrm{f}_j\left(\theta_j^{(r)}\right) \Big/ \mathrm{p}\left(\mathbf{y}\middle|\theta_j^{(r)}\right)\mathrm{p}\left(\theta_j^{(r)}\right) \xrightarrow{a.s.} M_j^{-1}.$$

Gelfand and Dey also observe that for convergence to occur at a practical rate, it is quite helpful if $\mathrm{f}_j\left(\theta_j\right) \big/ \mathrm{p}_j\left(\mathbf{y}\middle|\theta_j\right)\mathrm{p}_j\left(\theta_j\right)$ is bounded above. If $\mathrm{f}_j\left(\theta_j\right)$ has "thick tails" relative to $\mathrm{p}_j\left(\mathbf{y}\middle|\theta_j\right)\mathrm{p}_j\left(\theta_j\right)$, then this will not be the case. For the case of continuous θ_j, Geweke (1997b) avoids this problem by taking $\mathrm{f}_j\left(\theta_j\right)$ to be the density of a multivariate normal distribution centered at $\hat{\theta}_j = R^{-1}\sum_{r=1}^{R}\theta_j^{(r)}$ with variance $R^{-1}\sum_{r=1}^{R}\left(\theta_j^{(r)} - \hat{\theta}_j\right)\left(\theta_j^{(r)} - \hat{\theta}_j\right)'$, truncated to a highest density region of the multivariate normal distribution.

In the mixture of normals probit model, this procedure is not practical if applied to the augmented parameter vector $\left(\beta, \alpha, \mathbf{h}, \mathbf{p}, \tilde{\mathbf{y}}, \mathbf{e}\right)$ used in the posterior simulator, because the length of this vector is more than twice the sample size. But since the probability function (2.15) for **d** is available in essentially closed form, with $\tilde{\mathbf{y}}$ and **e** marginalized analytically, the procedure can be applied by using the parameter vector $\left(\beta, \alpha, \mathbf{h}, \mathbf{p}\right)$ and the product of (2.15) with the prior densities (2.3), (2.12), (2.13), and (2.14). (In accounting for the labeling restrictions, the normalization factor for the prior density can be found by independence Monte Carlo; accuracy on the order of 10^{-3} in the logarithm of the normalizing constant can typically be achieved in a few seconds. This means that this source of approximation error will contribute only about 0.1% to the evaluation of the marginal likelihood (2.19).) To increase the accuracy of the approximation, it is helpful to reparameterize the h_j by $\log\left(h_j^{1/2}\right)$ and **p** by $\log\left(p_j / p_m\right)$ $\left(j = 1, \ldots, m-1\right)$. Since the truncated normal density $\mathrm{f}_j\left(\theta_j\right)$ may still not be contained in the support of the parameter space because of labeling restrictions, the normalizing constant for this distribution must be systematically adjusted by Monte Carlo as well.

3. Some results with artificial data

Before proceeding to substantive applications, we conducted some experiments with artificial data. The main purpose of these experiments was to check software and gain some appreciation of how large a sample might be required to produce posterior moments that differ in an interesting way from prior moments, how much computing time would be required, and — by implication — what might be the practical scope for application of the methods described in Section 2. As a byproduct, the experiments provide some indication of the ability of these methods to detect departures from the conventional probit model specification and of the mixture of normals probit model to approximate other distributions of the disturbance. The latter questions are of no interest to a purely subjective Bayesian, but are probably of considerable concern to non-Bayesians

We used five data generating processes, shown in Table 1. In each process, there is a single explanatory variable with mean zero and standard deviation five, and a coefficient of one. The first three processes are special cases of the mixture of normals probit specification. The first is a conventional probit model. The second and third are mixtures of two normals, the third having a bimodal distribution of the shock ε_t. In the fourth data generating process, the shock is Cauchy, and in the fifth, it is logit. The sample size is 2,000 in every process.

We used three model specifications, shown in Table 2. In each specification, the prior distribution of the slope and intercept coefficients is independent standard normal. In the specifications with two or more mixtures, the prior correlation of the intercept coefficients is .8, which indicates that the coefficients are believed close together relative to their distance from zero. The first model is the conventional probit model. The second model has two mixtures: a labeling restriction on the precisions and a setting of the larger precision to unity. The parameters of the prior distribution are chosen so that the upper 5% of the distribution for h_1 is truncated by the labeling restriction $h_1 < 1$ and so that the .05 quantile is $h_1 = .01$. Thus, this prior distribution allows considerable scope for a leptokurtic distribution. The third model has three mixtures, again using a labeling restriction on the precisions, with $h_2 = 1$. The smaller precision has the same prior distribution as in the preceding model. For the larger prior precision , the lower 5% of the distribution is truncated by the labeling restriction $h_3 > 1$, and the .95 quantile is $h_3 = 10$.

We carried out computations on a Sun Ultra 200 Sparcstation using software written in Fortran 77. In each computation 10,000 iterations of the Gibbs sampler were employed, and the last 8,000 were used in the approximation of posterior moments. Computation time was 14 minutes for the conventional probit model, 28 minutes for the mixture of two normals, and 39 minutes for the mixture of three normals. Since the draws exhibit positive serial correlation, there is less information in 8,000 iterations than there would be in 8,000 hypothetical i.i.d. drawings from the posterior distribution. The ratio of the number of such i.i.d. draws to the number required in any given posterior simulator is the relative numerical efficiency of that posterior simulator. In the conventional probit model relative numerical efficiency for the Gibbs sampler was between .1 and .2 for most moments. In the mixture of two normals, it was between .01 and 06. In the mixture of three normals, it was between .005 and .05.

Discrimination between models using Bayes factors is of considerable interest. Table 3 shows marginal likelihoods for all data sets and models. Consider first the data generated from the conventional probit model (set 1). The Bayes factor in favor of the conventional probit model is 16 over the two-mixture model and 2,440 over the three-mixture model. The intuitive interpretation of this result is that while the mixture models are correctly specified, they spread prior probability over a larger space than does the conventional probit model, which leaves less probability for the data generating process. This is an example of how Bayes factors penalize excessive parameterization.

For the data generated from the two-mixture specification, the Bayes factor against the incorrectly specified conventional probit model is overwhelming (about 1.45×10^{12}). The three-mixture model is correctly specified, but, again, the penalty for excessive parameterization is exhibited in a Bayes factor of 3.7 in favor of the two-mixture model. Results for the full-mixture data (set 3) are similar. For the case in which the distribution of the disturbance is Cauchy, the Bayes factor between the two mixture models is one, but the Bayes factor against the conventional probit specification is huge: 6.78×10^{33}. Neither the two- nor the three-mixture model provides an improvement on the conventional probit model in the case of the logistic distribution, with the Bayes factor in favor of the conventional probit model being 5.5 against the two-mixture model and 6.7 against the three-mixture model. In view of the well-documented great similarity of the probit and logit specifications (Maddala 1983), these findings for the logit data are not surprising.

The predictive distributions, $\mathrm{P}(d = 1|x)$, are the main focus of interest in any application of these models. Figures 1 through 5 show some aspects of these distributions for the five data sets, respectively. Each figure corresponds to one of the artificial data generating processes shown in Table 1, and each contains six panels. Each panel shows a relevant range of *x* values on the horizontal axis. The three panels on the left indicate the value of $\mathrm{P}(d = 1|x)$ in the data generating process with a solid line and the .25 and .75 posterior quantiles for this population value with dotted and dashed lines, respectively. The three panels on the right indicate the derivative of $\mathrm{P}(d = 1|x)$ with respect to *x* in the data generating process with a solid line, and .25 and .75 posterior quantiles for this population value with dotted and dashed lines. The upper pair of panels shows results for the conventional probit model, the middle pair for the two full mixture model, and the third pair for the three full mixture model.

Results for the conventional probit data generating process are shown in Figure 1. For a well-specified model and large sample sizes, we expect that the population value of $\mathrm{P}(d = 1|x)$ and its first derivative would be bracketed by the posterior .25 and .75 quantiles about half the time. On the one hand, the conventional probit data generating process is a special case of all three model probability density functions (p.d.f.'s), but the posterior interquartile values match the population $\mathrm{P}(d = 1|x)$ better for the conventional probit specification than for either of the mixture models. On the other hand, there are not many values of *x* for which there is a serious discrepancy between posterior and data generating process (DGP) in the sense that the true value lies more than an interquartile range's width from the boundary of the interquartile range itself. The obvious deterioration in the visual match between distributions evident in Figure 1, in moving from the conventional probit to the more complicated mixture of normals models, is reflected in the marginal likelihoods in the first row of Table 3.

In the case of the scale mixture of normals data generating process (Figure 2), the conventional probit model correctly specifies a symmetric distribution of the disturbance, but cannot capture the thick tails of the population density. The most obvious feature of Figure 2 is the distance of the posterior interquartile values from the population values of $\mathrm{P}(d = 1|x)$ for the conventional probit model and the corresponding closeness for the mixture models. This is reflected in the marginal likelihood values in the second row of Table 3. Interquartile ranges are larger for the two-mixture model than for the conventional probit model and slightly larger yet for the three-mixture model. But the fit of the conventional probit model is so bad that its probability is very low relative to the other two (as indicated in the second row of Table 3). The three-mixture model looks a little closer to the DGP than does the two-mixture model in the interquartile range metric, but the interquartile range is sufficiently larger that it receives a little lower marginal likelihood.

The full mixture of normals DGP implies a bimodal p.d.f. for the shock. None of the models captures the bimodality, as indicated in Figure 3, but the mixture models come much closer than does the conventional probit model. As was the case for the scale mixture of normals DGP, interquartile ranges increase with model complexity. The conventional model has difficulty with the tails, complicated by the asymmetry of the distribution. Although the population distribution of ε_t has mean zero, its median is positive. The distribution in the conventional probit specification is symmetric, and this causes overprediction rather than underprediction of $\mathrm{P}(d=1|x)$ for small values of *x*. The mixture models exhibit difficulties that are similar qualitatively but of considerably less importance quantitatively. This behavior is due to the prior distribution of α, in which the standard deviation of $\alpha_1-\alpha_2$ is .63, whereas the population value of $\alpha_1-\alpha_2$ is over seven times this value. The prior is centered at a symmetric distribution, and so under- and over-prediction similar to the conventional probit model results.

Results using the Cauchy DGP are shown in Figure 4. The difficulty of the conventional probit model in coping with the very thick tails of the Cauchy distribution is obvious. By contrast, both mixture models provide good approximations to the true predictives $\mathrm{P}(d=1|x)$ over a very wide range.

Finally, Figure 5 compares posterior quartiles with DGP values in the case of a logistic DGP. The three models approximate the population value in similar ways, and interquartile ranges are about the same for all three. None of the models matches the thick tails of the logistic, but the discrepancy is not large in comparison with interquartile ranges. The visual similarity of the three models, evident in Figure 5, is reflected in the closeness of the marginal likelihoods shown in the fifth row of Table 3.

4. An example: Labor force participation of women

To provide a substantive application of the mixture of normals probit model, we use an otherwise standard regression for women's labor force participation, employing data from the PSID. The data pertain to those women observed in 1988, observed since the time they became a household head or spouse, for whom an uninterrupted history of annual hours worked is available, whose parents' educational attainment is available, and for whom spouse income and number of children are known. The sample size is 1,555. From the data, we construct 17 covariates, shown in Table 4. If a woman reports positive hours of work for 1988, she is a labor force participant; 80% of the women in the sample are labor force participants.

For the purposes of this illustration, independent Gaussian priors were constructed for each of the 17 covariate coefficients. In each case, the prior distribution has mean zero. The standard deviation is chosen by considering a large but reasonable effect of a change in the corresponding covariate on the probability of labor force participation, given that the probability of labor force participation is about one-half. This construction of the prior distribution for β is shown in Table 5, which provides the prior standard deviation for each coefficient along with the reasoning about effects on probability of labor force participation that led to the choice.

The prior distribution for the mean vector α of the normal mixture is the same as that used for the artificial data. The full specification of the normal mixture models is based on combinations of four prior distributions for precisions, indicated at the bottom of Table 5. Priors A and D were introduced in the experiments with artificial data. Prior B constrains the precision to be less than one, but places much smaller probability on very low precisions than does prior A. Prior C constrains the precision to be greater than one, but places smaller probability on high precisions than does prior D.

We report results for 13 models, shown in Table 6. Besides the conventional probit models, there are two groups of mixture of normals probit models. The first group uses scale mixtures of normals, thereby imposing symmetry of the shock distribution, and the second group uses full mixtures of normals. The idea is to vary the number of mixtures by using combinations of the prior distributions for precisions shown in Table 5 and labeling restrictions on precisions to identify the mixtures. In the case of mixtures of two and three normals, some alternative prior distributions are used to gauge the effect of alternative prior distributions for the mixtures on marginal likelihood.

Formal model comparison is straightforward based on the marginal likelihoods shown in Table 6. The most striking feature of the results is the poor performance of the conventional probit model relative to the mixture of normals probit models. Relative to the conventional probit model, the Bayes factor in favor of the mixture of normals model with the smallest marginal likelihood (the full mixture of two normals employing $h_A \leq 1$) is 200,000. Relative to this latter model, the Bayes factor in favor of the mixture of normals model with the largest marginal likelihood (the scale mixture of four normals) is 445. A second regular feature of the results is that for both the full mixture and the scale mixture groups, marginal likelihood increases as the number of mixtures is increased from two to three to four and then decreases slightly for the mixture of five normals. The mixtures of four and five normals provide an extremely rich family for a univariate distribution, with anywhere from 6 (scale mixture of four normals) to 13 (full mixture of five normals) parameters. Nevertheless, these models more than carry their own weight in the sense that this increased uncertainty about the form of the distribution is more than compensated by the better explanation of the data. A final regular feature of the results is evidence in favor of scale mixture models as opposed to full mixture models. However, this evidence is not strong. In some cases, the difference is on the order of numerical approximation error, and in one case, the full mixture model is preferred.

Posterior moments of the covariate coefficients in the conventional probit model, and in the scale mixture of four normals model, are shown in Table 7. The coefficients in the two models have different meanings, since the distributions of the shocks in the models are not the same. Nevertheless, the posterior means and standard deviations are quite similar. Seven of the seventeen covariates are important in the sense that their posterior means are on the order of the large but reasonable effects used to choose the prior standard deviations. (See Table 5.) These posterior means are also several posterior standard deviations from zero in each case. Probability of labor force participation declines with age for both single and married women. As expected, the effect is greater for married than single women, which corresponds to a change in probability of about 12% over 30 years for married women and 35% over 30 years for single women if labor force participation probabilities are around 50%. Labor force participation probability declines by about 20% with the first child and by about 6% for each child thereafter (again, beginning from participation probabilities of around 50%). Spouse income, Aid to Families with Dependent Children (AFDC), food stamp benefits, and cumulative work experience all have strong effects on the probability of labor force participation. Further details on these effects are presented below.

From the sampled values for the parameters of the distributions, it is straightforward to construct the posterior distribution of the probability density of the shock term ε_t. These densities are shown in Figures 6 and 7 for the scale mixture of four normals (the model with the highest marginal likelihood of those considered) and, for comparison, the full mixture of four normals (in which the Bayes factor relative to the scale mixture is a respectable .45). In each figure, the posterior median and interquartile ranges are shown for each ordinate. The leptokurtosis in the scale mixture of four normals is immediately evident in Figure 6. When the ratio of the third to the first quartile is considered, this figure suggests that the posterior uncertainty in the seven parameters of this distribution is most manifest near the origin and in the tails of the probability density. The full mixture of four normals density, which depends on 10 parameters, is shown in Figure 7. This distribution is strongly negatively skewed, a mode that is near zero but distinctly negative, and a thick positive tail.

Some predictive distributions of interest are shown in Table 8. Two examples are studied. In each example, results are shown for the conventional probit model and for the mixture of normals probit model with the highest marginal likelihood. In the first example, the probability of labor force participation for a 30-year-old woman with two children and no spouse present is examined, as AFDC and food stamp benefit levels (given no labor force participation) are changed from no benefits to the sample average for a woman in this situation in 1988 to the sample maximum. In the conventional probit model, the increase of benefits from zero to the maximum increases the probability of labor force nonparticipation from .042 to .174, while in the scale mixture of four normals model, the increase is only from .049 to .131. Notice that the posterior standard deviation of the labor force participation probabilities is lower in the scale mixture model than in the conventional probit model in every case: uncertainty about the individual parameters in this model, taken in isolation, does not imply greater uncertainty about the predictive probabilities than in the case of the conventional probit model where there is no uncertainty about the shape of the shock distribution.

In the second example shown in Table 8, cumulative labor market experience and spouse's income are varied for a married woman aged 30 with two children so as to vary the probability of labor force participation between roughly .01 and .99. Differences in the posterior means of participation probabilities can be appreciable, reaching .078 (about one posterior standard deviation) for a woman with no labor market experience and a spouse earning $75,000 per year. Ratios for participation or nonparticipation approach two when these probabilities are small. When probabilities of both participation and non-participation are substantial, the posterior standard deviations in the conventional probit model tend to be smaller, whereas when one is small and one is large, there is less uncertainty in the scale mixture of normals model.

A graphical presentation of the second example is given in Figures 8 and 9. Each figure shows posterior .25 and .75 quantiles from the conventional probit model (dotted lines) and the scale mixture of four normals model (dashed lines). Figure 8 contrasts the predicted probability of labor market participation as hours of experience are varied. The difference is greatest at around 7,000 hours (3.5 years) of experience, where the mixture models implies a substantially higher probability of participation. The .25 quantile of the mixture posterior is about the same as the .75 quantile of the conventional probit posterior at this value. As hours of experience increase, however, probability of participation in the conventional probit model approaches one much faster than in the mixture model. Both effects are the consequence of the thin tails of the normal relative to the scale mixture of four normals. Qualitatively similar characteristics are seen in the comparison for labor market participation probability as a function of spouse income in Figure 9. As income rises, participation probability drops. Over a long range, the participation probability drops faster in the mixture model, but eventually it is overtaken by the probability in the conventional probit model because of the thinner tails of the shock density in that model.

At least in this example, the mixture of normals probit model provides a marked improvement over the conventional probit model. The data are sufficiently informative about the nature of the distribution of the shock that the most preferred model has four mixtures with six free parameters in the distribution, and the most flexible model — five mixtures with thirteen free parameters — is still in the running, with posterior probability about 4.5% that of the most preferred model. Moreover, distributions of predictive probabilities of the kind usually studied with this model differ substantially between the conventional probit and mixture models. The mixture models imply a weaker effect of welfare benefit levels, and a stronger effect of spouse income on labor force participation than does the conventional probit model.

5. Conclusion

In the conventional probit model, there are no unknown parameters in the distribution of the shock term. We have generalized this specification by adopting mixtures of normals distributions in which the probability density of the shock term is governed by many unknown parameters. In the example pursued in this paper, Bayes factors strongly favored mixture of normals distributions over the conventional probit specification. The most favored model was a scale mixture of four normals with six free parameters for the shock density, but a full mixture of five normals with thirteen free parameters for the shock probability density was closely competitive with a Bayes factor of .45. We were pleasantly surprised that a sample with about 1,500 observations would be so informative for the distribution of the shock in a linear latent variable model.

Against this background, it is reasonable to contemplate several extensions of this work. The generalization of the mixture of normals model to the multivariate case is straightforward as is its incorporation into Markov chain Monte Carlo methods for the multinomial probit model that we have described elsewhere (Geweke, Keane, and Runkle 1994, 1997). In view of the irregular likelihood surface in the dimensions of the variance matrix in this model (Keane 1992), whether full or scale mixture of models will succeed as generalizations of the multinomial probit model seems to us very much an open question. An alternative extension is to mix distributions other than the normal or to mix the normal and/or several other families. The Student's *t* family is an obvious candidate, itself a continuous scale mixture of normal distributions to which methods similar to those used here can be applied (Geweke 1993). Finally, we note that the assumption of linearity in the covariates $\mathbf{x}_t$ is as much a convenient assumption, in most applications, as is normalilty in the conventional probit model. Relaxing this assumption in favor of the obvious expansion families (for example, Taylor or Laurent series) is straightforward and lends itself well to incorporation of subjective priors in the way we have done here. Clearly, there are interactions between relaxing both linearity and normality; in the limit, one cannot do both. We plan to explore these issues in future work.

References

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

Aldrich, J., and F. Nelson, 1984, *Linear Probability, Logit, and Probit Models.* Beverly Hills: Sage Publications.

Cosslett, S.R., 1983, "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica* **51**: 765–782.

Gallant, A.R., and D.W. Nychka, 1987, "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica* **55**: 363–390.

Gelfand, A.E., and D.K. Dey, 1994, "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society Series B* **56**: 501–514.

Geweke, J., 1991, "Efficient Simulation from the Multivariate Normal and Student-*t* Distributions Subject to Linear Constraints," in E. M. Keramidas (ed.), *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, 571–578. Fairfax: Interface Foundation of North America, Inc.

Geweke, J., 1993, "Bayesian Treatment of the Independent Student-*t* Linear Model," *Journal of Applied Econometrics* **8**: S19–S40.

Geweke, J., 1997a, "Posterior Simulators in Econometrics," in D. Kreps and K.F Wallis (eds.), *Advances in Economics and Econometrics: Theory and Applications*, vol. III, 128–165. Cambridge: Cambridge University Press.

Geweke, J., 1997b, "Simulation-Based Bayesian Inference for Economic Time Series," in R.S. Mariano, T. Schuermann and M. Weeks (eds.), *Simulation-Based Inference in Econometrics: Methods and Applications.* Cambridge: Cambridge University Press, forthcoming.

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.

Hausman, J., and D. McFadden, 1984, "Specification Tests for the Multinomial Logit Model," *Econometrica* **52**: 1219–1240.

Horowitz, J.L., 1992, "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica* **60**: 505–531.

Ichimura, H., 1993, "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics* **58**: 71–120.

Keane, M.P., 1992, "A Note on Identification in the Multinomial Probit Model," *Journal of Business and Economic Statistics* **10**: 193–200.

Klein, R.W. and R.H. Spady, 1993, “An Efficient Semiparametric Estimator for Binary Response Models,” *Econometrica* **61**: 387–421.

Koop, G., and D. J. Poirier, 1993, “Bayesian Analysis of Logit Models Using Natural Conjugate Priors,” *Journal of Econometrics* **56:** 323–340.

Lewbel, A., 1997, “Semiparametric Estimation of Location and Other Discrete Choice Moments,” *Econometric Theory* **13**: 32–51.

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

Manski, C.F., 1985, “Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator,” *Journal of Econometrics* **27**: 313–333.

Powell, J.L., J.H. Stock, and T.M. Stoker, 1989, “Semiparametric Estimation of Index Coefficients,” *Econometrica* **57**: 1403–1430.

Roberts, G.O., and A.F.M. Smith, 1994, “Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms,” *Stochastic Processes and Their Applications* **49**: 207–216.

Tierney, L., 1994, “Markov Chains for Exploring Posterior Distributions” (with discussion and rejoinder), *Annals of Statistics* **22**: 1701–1762.

Zellner, A., and P.E. Rossi, 1984, “Bayesian Analysis of Dichotomous Quantal Response Models,” *Journal of Econometrics* **25**: 365–393.

Table 1

Artificial data generating processes

All data sets: $x_t \overset{IID}{\sim} \mathrm{N}(0, 25)$

Data sets 1-4: $\tilde{y}_t = \beta x_t + \varepsilon_t, \quad \beta = 1$

$$(x_t, \varepsilon_t) \quad \text{i.i.d.}$$

$$d_t = \begin{cases} 1 \text{ if } \tilde{y}_t > 0 \\ 0 \text{ if } \tilde{y}_t \leq 0 \end{cases}$$

Data Set 1: Normal (conventional probit specification)

$$\varepsilon_t \sim \mathrm{N}(0, 1)$$

Data Set 2: Scale mixture of two normals

$$\varepsilon_t \sim \begin{cases} \mathrm{N}(0, 1) \ (p_1 = .5) \\ \mathrm{N}(0, 25) \ (p_2 = .5) \end{cases}$$

Data Set 3: Full mixture of two normals

$$\varepsilon_t \sim \begin{cases} \mathrm{N}(1.5, 1.0) \ (p_1 = .667) \\ \mathrm{N}(-3.0, 4.0) \ (p_2 = .333) \end{cases}$$

Data Set 4: Cauchy distribution

$$\varepsilon_t \sim \mathrm{Cauchy}(0, 1)$$

Data Set 5: Logit distribution

$$\mathrm{P}(\mathrm{y}_\mathrm{t} = 1 | x_t) = \exp(x_t) / [1 + \exp(x_t)]$$

Table 2

Model specifications
Artificial data

All models:

$$\begin{aligned} \tilde{y}_t &= \beta x_t + \varepsilon_t \\ \varepsilon_t \big| (x_1, \ldots, x_T) &\quad \text{i.i.d.} \\ d_t &= \begin{cases} 1 \text{ if } \tilde{y}_t 0 \\ 0 \text{ if } \tilde{y}_t \le 0 \end{cases} \\ \beta &\sim \mathrm{N}(0, 1) \end{aligned}$$

Conventional probit model:

$$\begin{aligned} \varepsilon_t \big| (x_1, \ldots, x_T) &\sim \mathrm{N}(\alpha, 1) \\ \alpha &\sim \mathrm{N}(0, 1) \end{aligned}$$

Two full mixtures:

$$\mathrm{P}\left[\varepsilon_t \big| (x_1, \ldots, x_T) \sim \mathrm{N}\left(\alpha_j, h_j^{-1}\right)\right] = p_j$$

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1. & .8 \\ .8 & 1. \end{bmatrix} \right)$$

$$3.832 h_1 \sim \chi^2(.996), \quad h_2 = 1, \quad h_1 < h_2$$

$$(p_1, p_2) \sim \mathrm{Beta}(5, 5)$$

Three full mixtures:

$$\mathrm{P}\left[\varepsilon_t \big| (x_1, \ldots, x_T) \sim \mathrm{N}\left(\alpha_j, h_j^{-1}\right)\right] = p_j$$

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1. & .8 & .8 \\ .8 & 1. & .8 \\ .8 & .8 & 1. \end{bmatrix} \right)$$

$$3.832 h_1 \sim \chi^2(.996), \quad h_2 = 1, \quad 1.076 h_3 \sim \chi^2(4.85); \quad h_1 < 1 < h_3$$

$$(p_1, p_2, p_3) \sim \mathrm{Beta}(3.5, 3.5, 1)$$

Table 3

Marginal likelihoods[1]

Data set	Conventional probit	Two full mixtures	Three full mixtures
1 (Normal)	-261.7 (.08)	-263.5 (.16)	-269.5 (.35)
2 (Scale mixture)	-771.1 (.02)	-743.1 (.08)	-744.4 (.15)
3 (Full mixture)	-621.5 (.03)	-601.6 (.07)	-602.8 (.43)
4 (Cauchy)	-772.9 (.02)	-695.0 (.19)	-695.0 (.14)
5 (Logit)	-246.5 (.07)	-248.2 (.14)	-248.4 (.18)

[1] Numerical standard errors of numerical approximations are shown in parentheses.

Table 4

Covariate definitions

PSID labor force participation

1.	Black	1 if black; else 0
2.	Age (Single)	Age if not married; else 0
3.	Age (Married)	Age if married; else 0
4.	Education	Years of schooling
5.	Married	1 if married; else 0
6.	Kids	1 if children present; else 0
7.	#Kids	Number of children present
8.	Spouse$	Spouse’s income in current (1988) dollars; 0 if no spouse
9.	Spouse$0	1 if spouse present with no income; else 0
10.	Family$	Household unearned income
11.	F-HS	1 if father graduated from high school but not college; else 0
12.	F-Coll	1 if father graduated from college; else 0
13.	M-HS	1 if mother graduated from high school but not college; else 0
14.	M-Coll	1 if mother graduated from college; else 0
15.	AFDC$	Monthly AFDC income if woman does not work (1988 $)
16.	Food$	Monthly food stamp eligibility if woman does not work (1988 $)
17.	WorkExp	Cumulative number of hours worked since becoming a household head

Table 5

Model specifications

PSID labor force participation

Covariate priors (Gaussian, mean zero)

	Description	Prior standard deviation	Derivation of prior standard deviation
1.	Black	.125	$\Delta p = .05$ at $p = .5$
2.	Age (Single)	.00417	$\Delta p = .05$ for 55 vs. 25 at $p = .5$
3.	Age (Married)	.03333	$\Delta p = .40$ for 55 vs. 25 at $p = .5$
4.	Education	.00417	$\Delta p = .10$ for 10 vs. 16 at $p = .5$
5.	Married	.125	$\Delta p = .05$ at $p = .5$
6.	Kids	.250	$\Delta p = .10$ at $p = .5$
7.	#Kids	.125	$\Delta p = .05$ at $p = .5$
8.	Spouse$	3.57×10^{-6}	$\Delta p = .05$ per \$35,000 at $p = .5$
9.	Spouse$0	.125	$\Delta p = .05$ at $p = .5$
10.	Family$	3.57×10^{-6}	$\Delta p = .05$ per \$35,000 at $p = .5$
11.	F-HS	.05	$\Delta p = .02$ at $p = .5$
12.	F-Coll	.05	$\Delta p = .02$ at $p = .5$
13.	M-HS	.10	$\Delta p = .04$ at $p = .5$
14.	M-Coll	.10	$\Delta p = .04$ at $p = .5$
15.	AFDC$	6.25×10^{-4}	$\Delta p = .25$ per \$1,000 at $p = .5$
16.	Food$	6.25×10^{-4}	$\Delta p = .25$ per \$1,000 at $p = .5$
17.	WorkExp	6.25×10^{-5}	$\Delta p = .05$ per year (2,000 hours)

Normal mixture means

All means 0, all variances 4, all covariances 3.2

Normal mixture precisions: $\underline{s}^2 h \sim \chi^2(\underline{v})$

	Values of $\underline{s}^2$ and $\underline{v}$	Derivation of $\underline{s}^2$ and $\underline{v}$
A.	$\underline{s}^2 = 3.832$, $\underline{v} = .996$	$\mathrm{P}(h \leq .01) = \mathrm{P}(h > 1) = .05$
B.	$\underline{s}^2 = 21.16$, $\underline{v} = 12.1$	$\mathrm{P}(h \leq .25) = \mathrm{P}(h > 1) = .05$
C.	$\underline{s}^2 = 31.36$, $\underline{v} = 45.9$	$\mathrm{P}(h \leq 1) = \mathrm{P}(h > 2) = .05$
D.	$\underline{s}^2 = 1.076$, $\underline{v} = 4.85$	$\mathrm{P}(h \leq 1) = \mathrm{P}(h > 10) = .05$

Table 6

Summary of prior specifications and marginal likelihoods

PSID labor force participation

Model	Beta prior for **p**	Precisions mixed	Log marginal likelihood	
			Approximation	Numerical Standard Error
Conventional probit	-----	------	-566.61	.012
Full mixture, 2 normals	5.0, 5.0	$h_A \le 1$	-554.4	.19
Full mixture, 2 normals	5.0, 5.0	$h_B \le 1$	-554.2	.15
Full mixture, 3 normals	3.5, 3.5, 1.0	$h_A \le 1 \le h_D$	-550.6	.34
Full mixture, 3 normals	3.5, 3.5, 1.0	$h_B \le 1 \le h_C$	-553.0	.34
Full mixture, 4 normals	1.5, 2.0, 1.5, 1.0	$h_A \le h_B \le 1 \le h_D$	-549.1	.19
Full mixture, 5 normals	1.5, 2.0, 3.5, 0.5, 0.5	$h_A \le h_B \le 1 \le h_C \le h_D$	-551.4	.22
Scale mixture, 2 normals	5.0, 5.0	$h_A \le 1$	-553.9	.33
Scale mixture, 2 normals	5.0, 5.0	$h_B \le 1$	-553.7	.12
Scale mixture, 3 normals	3.5, 3.5, 1.0	$h_A \le 1 \le h_D$	-551.9	.39
Scale mixture, 3 normals	3.5, 3.5, 1.0	$h_B \le 1 \le h_C$	-552.3	.12
Scale mixture, 4 normals	1.5, 2.0, 1.5, 1.0	$h_A \le h_B \le 1 \le h_D$	-548.3	.15
Scale mixture, 5 normals	1.5, 2.0, 3.5, 0.5, 0.5	$h_A \le h_B \le 1 \le h_C \le h_D$	-548.5	.19

Table 7

Covariate coefficient posterior moments

PSID labor force participation

		Normal prior	Posterior			
			Conventional probit		Scale mixture of four normals	
	Intercept	$(0,\ 4^2)$	1.213	(.183)	1.019	(.222)
1.	Black	$(0,\ .125^2)$	.0158	(.0781)	.037	(.0708)
2.	Age (Single)	$(0,\ .00417^2)$	-.0104	(.0038)	-.0099	(.0038)
3.	Age (Married)	$(0,\ .03333^2)$	-.0288	(.0066)	-.0274	(.0066)
4.	Education	$(0,\ .00417^2)$	.0023	(.0042)	.0021	(.0041)
5.	Married	$(0,\ .125^2)$	.189	(.117)	.206	(.117)
6.	Kids	$(0,\ .250^2)$	-.361	(.135)	-.383	(.144)
7.	#Kids	$(0,\ .125^2)$	-.151	(.045)	-.129	(.056)
8.	Spouse$	$(0,\ (3.57\times10^{-6})^2)$	-7.29×10^{-6}	(2.29×10^{-6})	-6.89×10^{-6}	(2.27×10^{-6})
9.	Spouse$0	$(0,\ .125^2)$	-.00650	(.117)	.00186	(.116)
10.	Family$	$(0,\ 3.57\times10^{-6})^2)$	-1.07×10^{-6}	(2.92×10^{-6})	$-.77\times10^{-6}$	(2.83×10^{-6})
11.	F-HS	$(0,\ .05^2)$	.0300	(.0445)	.0238	(.0429)
12.	F-Coll	$(0,\ .05^2)$	-.0074	(.0474)	-.0180	(.0473)
13.	M-HS	$(0,\ .10^2)$	-.0223	(.0675)	-.0265	(.0628)
14.	M-Coll	$(0,\ .10^2)$	-.0113	(.0863)	-.0102	(.0855)
15.	AFDC$	$(0,\ (6.25\times10^{-4})^2)$	-5.88×10^{-4}	(3.06×10^{-4})	-5.07×10^{-4}	(2.82×10^{-4})
16.	Food$	$(0,\ (6.25\times10^{-4})^2)$	-12.18×10^{-4}	(4.39×10^{-4})	-10.77×10^{-4}	(4.26×10^{-4})
17.	WorkExp	$(0,\ (6.25\times10^{-5})^2)$	11.80×10^{-5}	$(.83\times10^{-5})$	12.36×10^{-5}	(2.24×10^{-5})

Table 8

Effects of some covariates on labor force participation probability

AFDC and food stamp benefits

Base case: Unmarried, not black, 2 kids, age 30, 12 years education, 6.25 years work experience

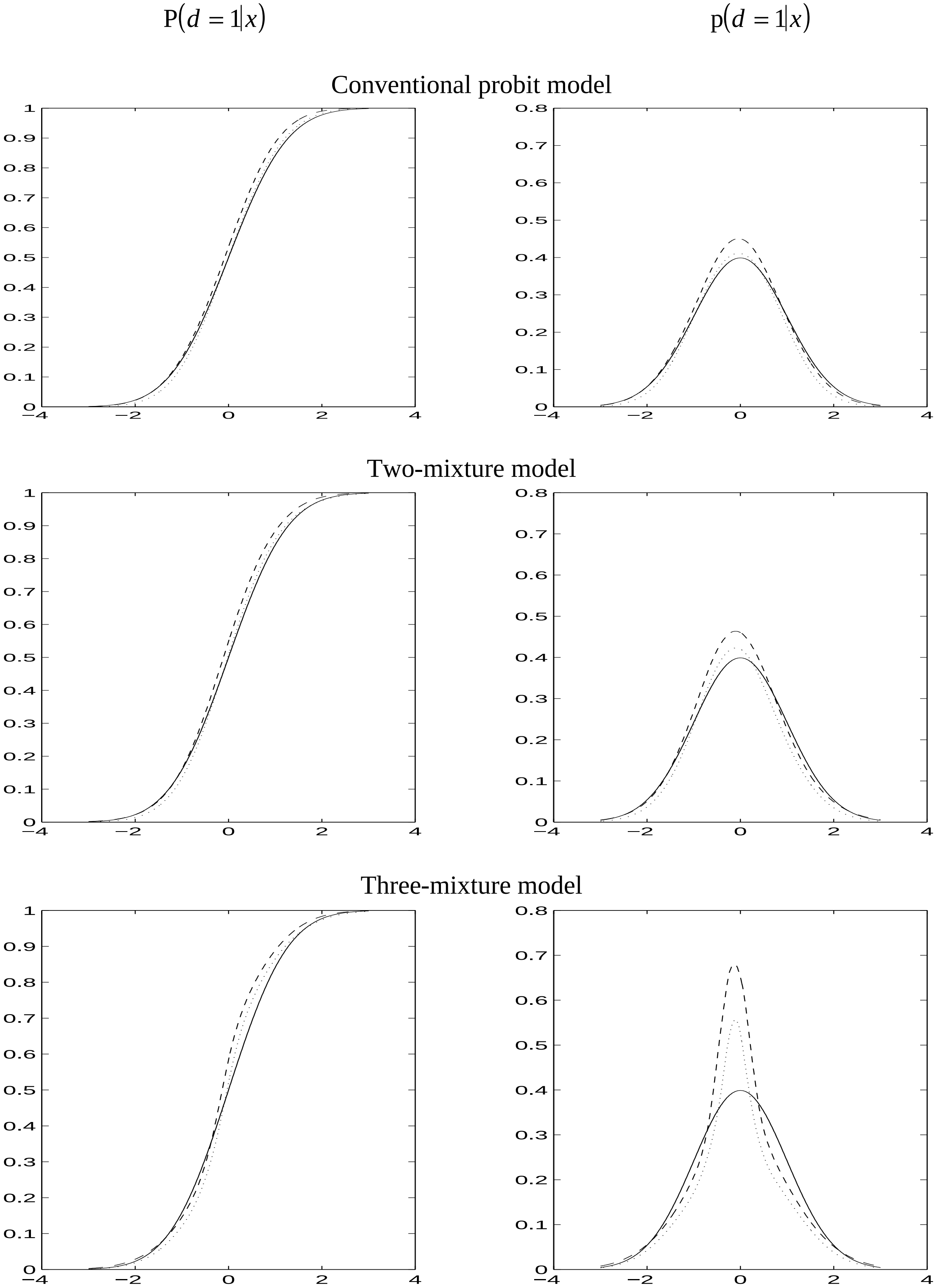
	Posterior moments of probability			
	Conventional probit		Scale mixture of four normals	
AFDC = $0, Food Stamps = $0	.958	(.015)	.951	(.012)
AFDC= $289, Food Stamps= $197	.908	(.019)	.921	(.014)
AFDC= $633, Food Stamps= $344	.826	(.047)	.869	(.036)
AFDC and Food Stamps 0 vs. max	-.131	(.052)	-.083	(.039)

Some combinations of work experience and spouse's income

Base case: married, not black, 2 kids, age 30, 12 years education

	Posterior moments of probability			
	Conventional probit		Scale mixture of four normals	
Experience 10 years; income $25,000	.992	(.004)	.983	(.006)
Experience 8 years; income $25,000	.974	(.009)	.969	(.008)
Experience 6 years; income $25,000	.932	(.019)	.939	(.014)
Experience 4 years; income $25,000	.846	(.033)	.877	(.027)
Experience 2 years; income $25,000	.710	(.049)	.744	(.058)
Experience 0 years; income $25,000	.534	(.061)	.497	(.089)
Experience 0 years; income $50,000	.463	(.063)	.399	(.089)
Experience 0 years; income $75,000	.393	(.070)	.315	(.090)
Experience 0 years; income $100,000	.326	(.079)	.249	(.090)
Experience 0 years; income $150,000	.215	(.089)	.159	(.083)
Experience 0 years; income $200,000	.136	(.088)	.106	(.074)
Experience 0 years; income $500,000	.011	(.042)	.018	(.039)

Figure 1: Population values and posterior quartile values: Normal data generating process



$\mathrm{P}(d=1|x)$

$\mathrm{p}(d=1|x)$

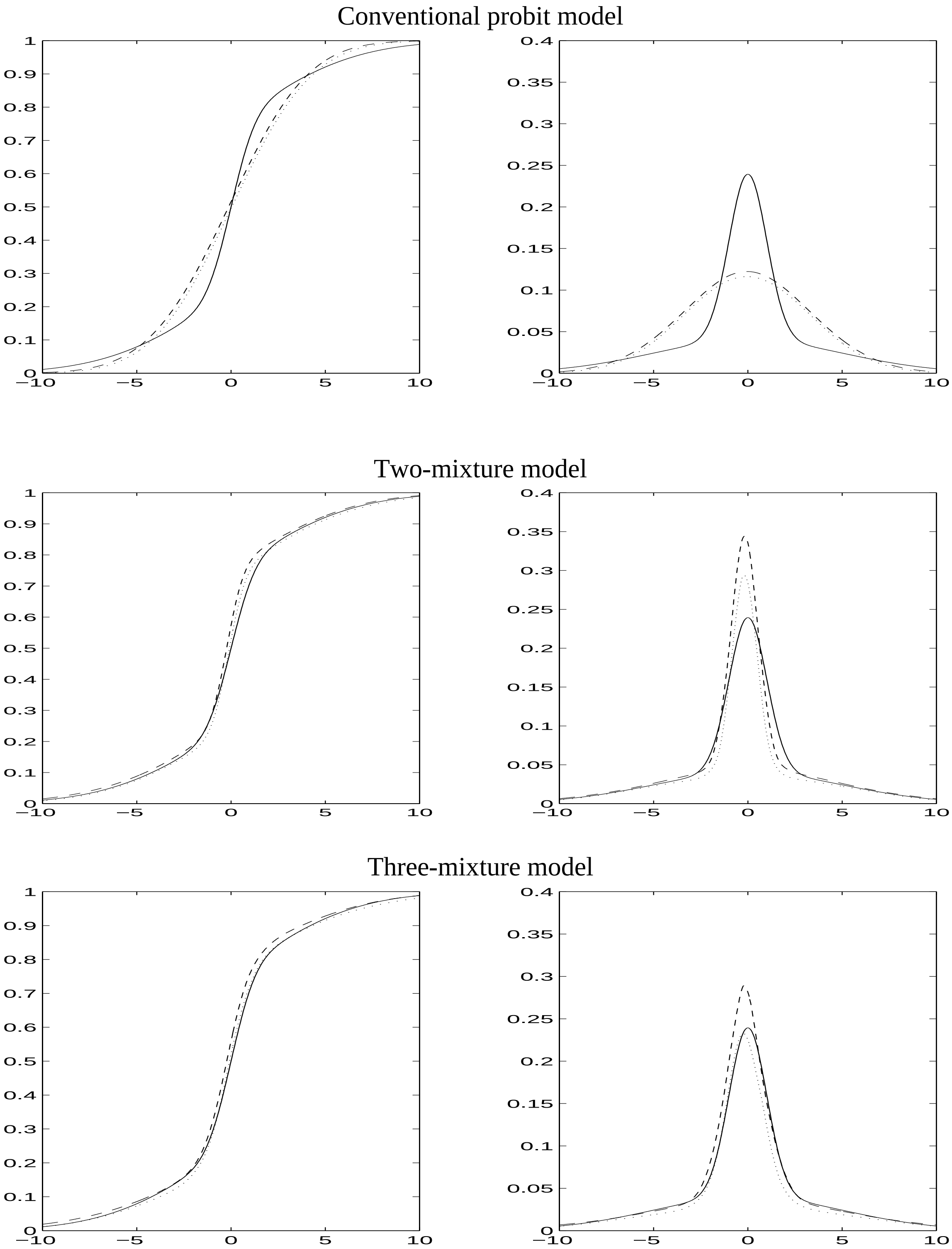
Conventional probit model

Two-mixture model

Three-mixture model

Figure 2: Population values and posterior quartile values: Scale mixture data generating process

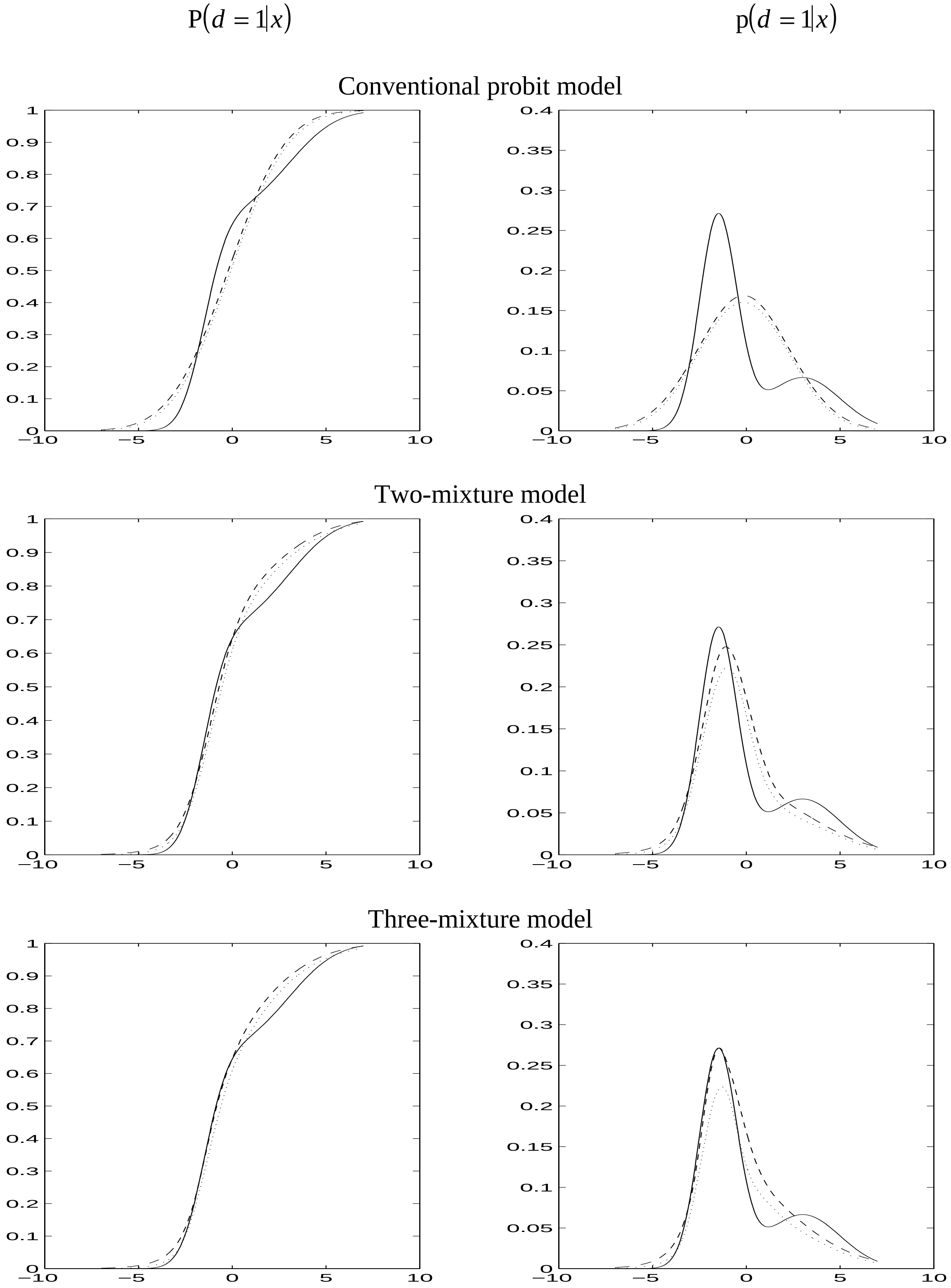
$\mathrm{P}(d=1|x)$ $\mathrm{p}(d=1|x)$


Conventional probit model

Two-mixture model

Three-mixture model

Figure 3: Population values and posterior quartile values: Full mixture data generating process



$P(d=1|x)$

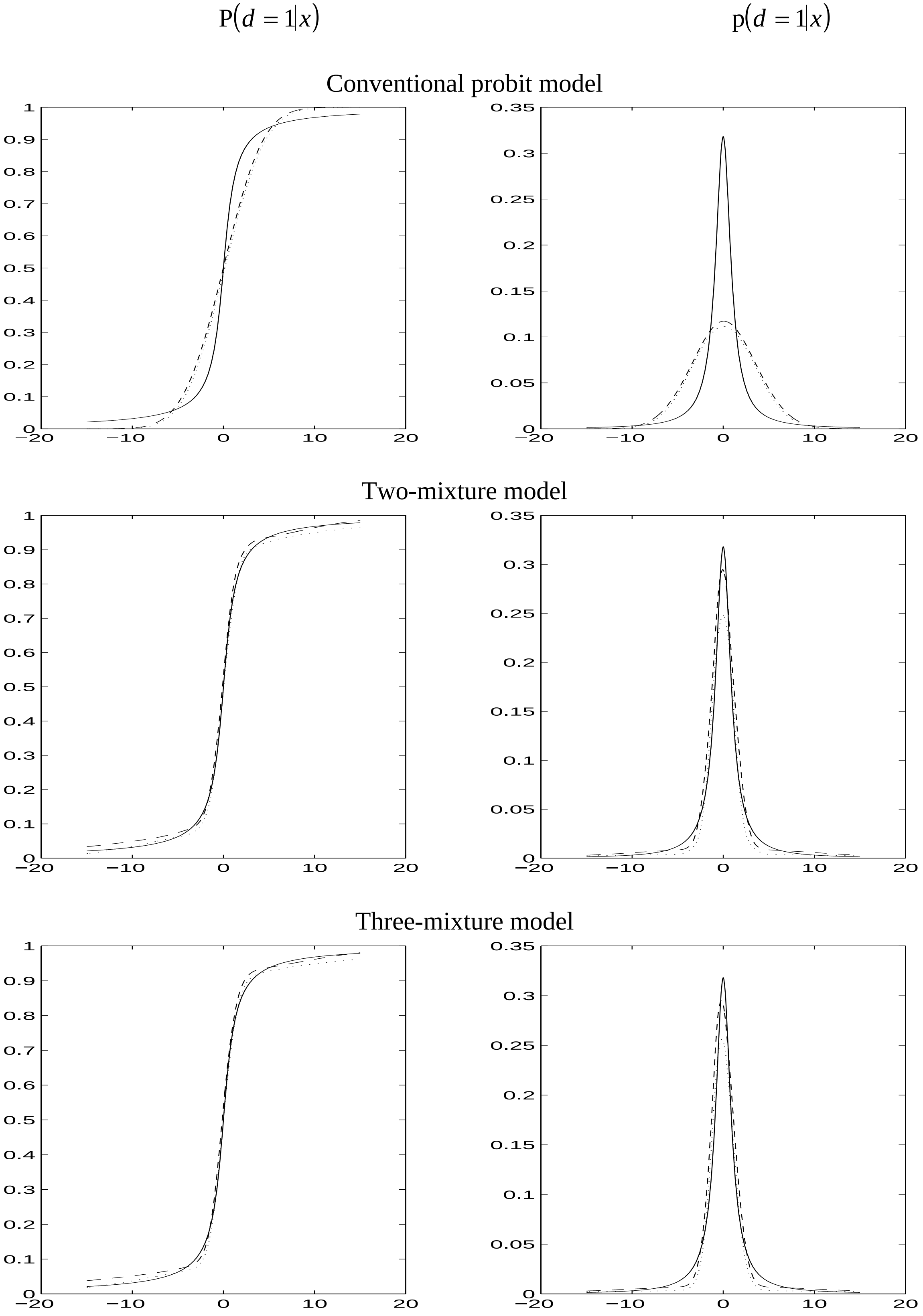
$p(d=1|x)$

Conventional probit model

Two-mixture model

Three-mixture model

Figure 4: Population values and posterior quartile values: Cauchy data generating process



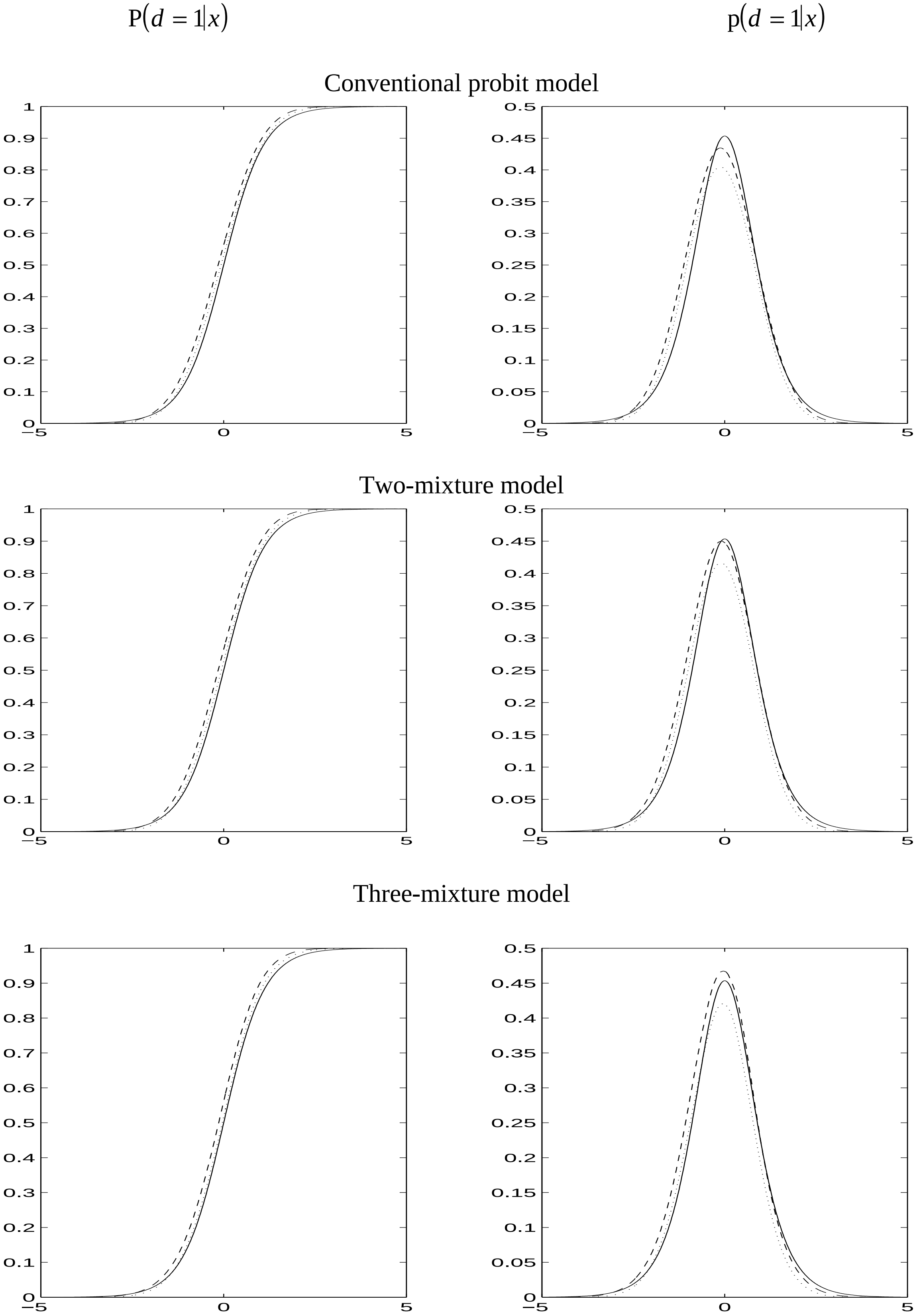
$\mathrm{P}(d=1|x)$ $\mathrm{p}(d=1|x)$

Conventional probit model

Two-mixture model

Three-mixture model

Figure 5: Population values and posterior quartile values: Logit data generating process



$\mathrm{P}(d = 1|x)$

$\mathrm{p}(d = 1|x)$

Conventional probit model

Two-mixture model

Three-mixture model

Figure 6: Disturbance probability density function, scale mixture of four normals

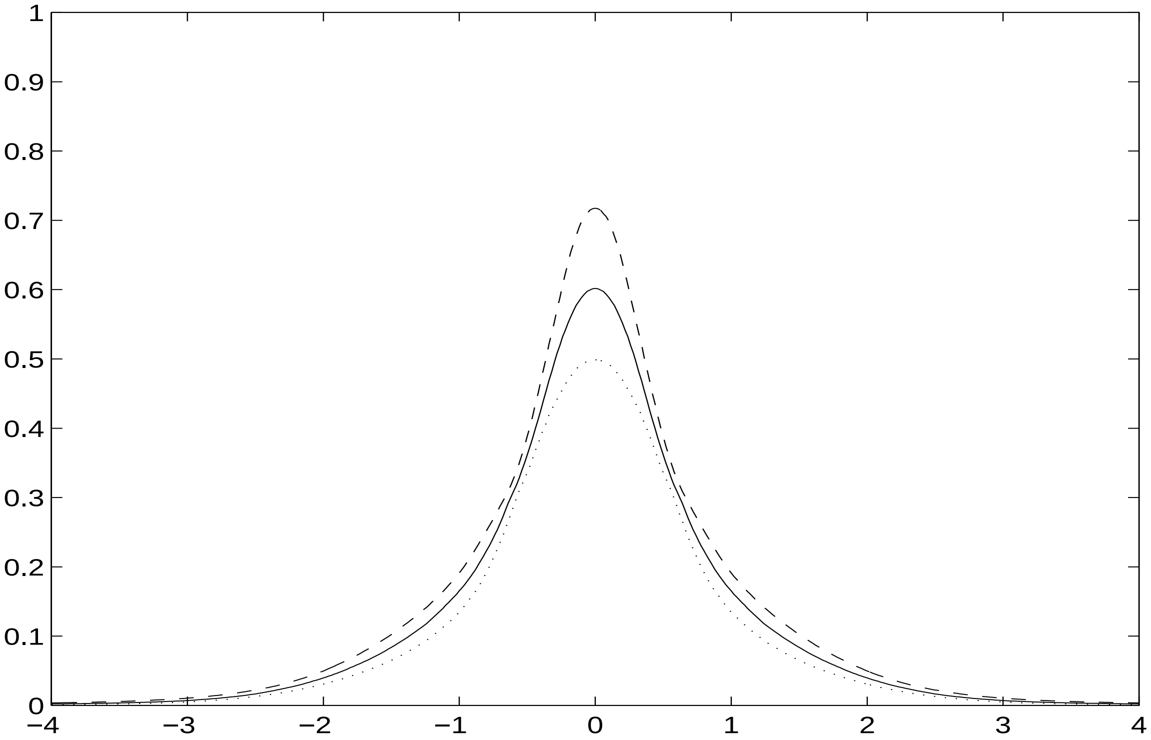


Figure 7: Disturbance probability density function, full mixture of four normals

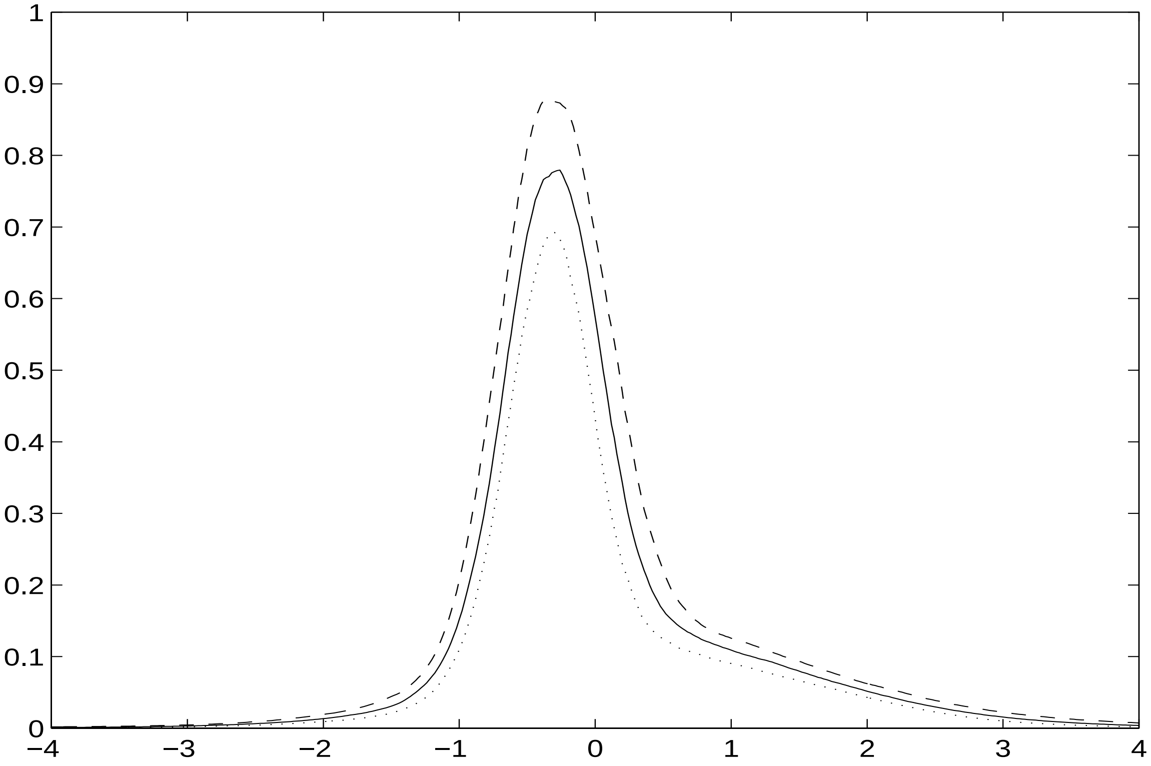


Figure 8: Labor force participation probability as a function of experience

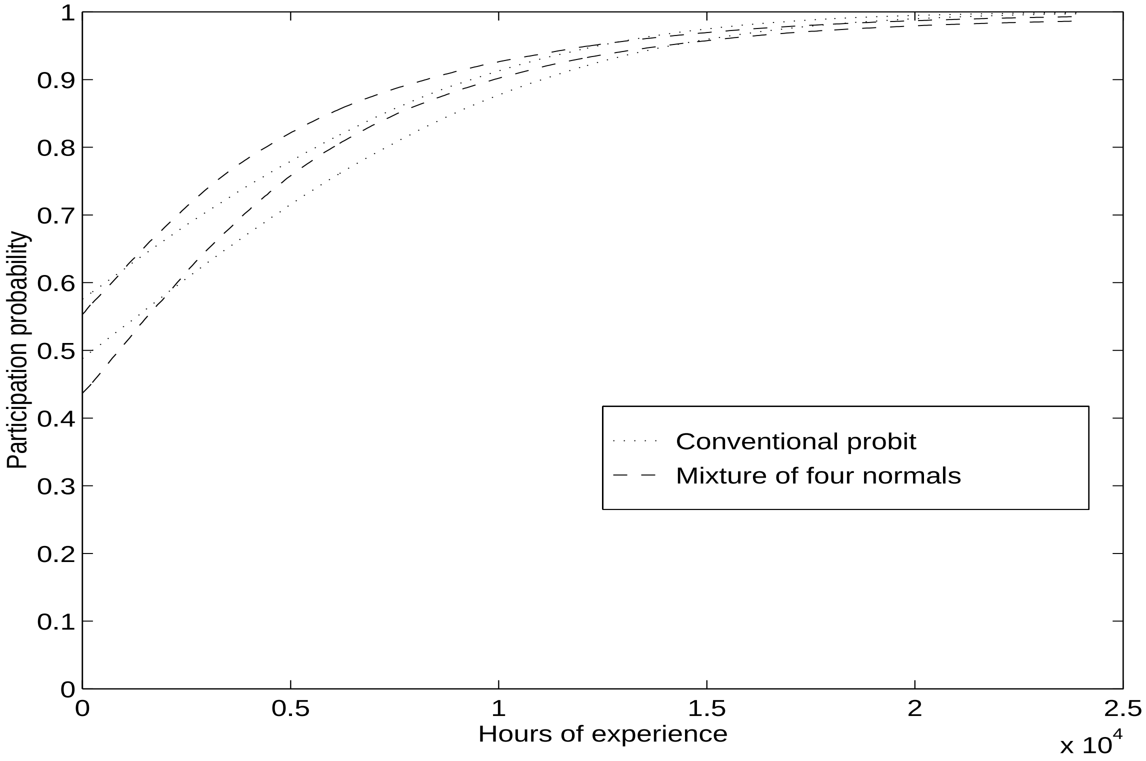
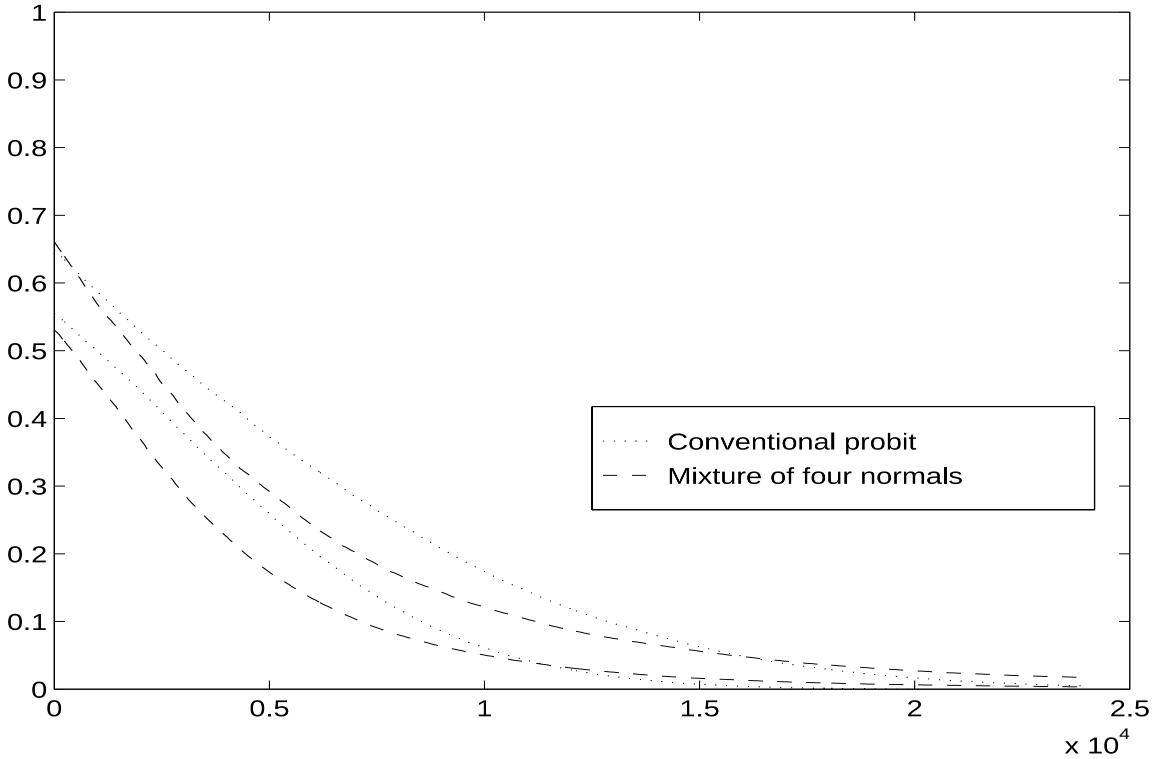


Figure 9: Labor force participation probability as a function of spouse's income


Conventional probit
Mixture of four normals
x 10^4