Dynamic Panel Probit with Flexible Correlated Effects*

Martin Burda[†] Matthew Harding[‡]

December 2008

Abstract

In this paper, we analyze a dynamic panel probit model with two flexible latent effects: first, unobserved individual heterogeneity that is allowed to vary in the population according to an assumption-free nonparametric distribution, and second, with a latent serially correlated common error component. In doing so, we extend the approach developed in Albert and Chib (1993), Albert and Chib (1996), and Chib and Carlin (1999) by releasing restrictive parametric assumptions on the latent individual effect and eliminating potential spurious state dependence with latent time effects. Due to a data augmentation strategy, virtually all model parameters are sampled directly without resorting to Metropolis-Hastings steps that would require additional tuning. The estimation problem is re-cast in the form of an iterative scheme of latent regressions whereby the latent utility is explicitly sampled along with other model parameters. This feature avoids the curse of dimensionality arising in likelihoods specified as high-dimensional integrals. We show in simulations that capturing nonparametrically unobserved heterogeneity of irregular form greatly reduces the bias of estimated average partial effects relative to the random effects probit model. Moreover, the bias reduction is achieved even when the unobserved individual effects are correlated with the observed regressors. We apply our method to the estimation of a patent equation on firm level data. We find statistically significant technology spillover effects as well as clusters in unobserved firm level heterogeneity.

JEL: C11, C13, C15, C23, C25

Keywords: Dynamic latent variables, Markov Chain Monte Carlo, Dirichlet Process prior

*We are grateful to Nick Bloom for sharing the data for the empirical application with us. We also thank Siddhartha Chib, David Dahl, Christian Gourieroux, Ivan Jeliazkov, Andriy Norets, and seminar participants at UC Berkeley and UC Irvine for useful comments. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca).

[†]Department of Economics, University of Toronto, 150 St. George St., Toronto, ON M5S 3G7, Canada; Phone: (416) 978-4479; Email: `martin.burda@utoronto.ca`

[‡]Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305; Phone: (650) 723-4116; Fax: (650) 725-5702; Email: `mch@stanford.edu`

1. **Introduction**

There is broad agreement that individual heterogeneity plays a crucial function in many economic models. In linear models, panel data is used to identify the effects of interest while at the same time controlling for unobserved individual heterogeneity (Hausman and Taylor 1981). In spite of the voluminous research on modeling heterogeneity, we know relatively little about how to estimate nonlinear models with unobserved heterogeneity of an unknown form. Many models suffer from substantial theoretical and computational challenges (Arellano and Hahn 2006).

It has long been recognized that maximum likelihood analysis of nonlinear limited dependent variable (LDV) models with panel data is feasible only under relatively restrictive assumptions (Butler and Moffitt 1982). Thus, typically we need to make very specific distributional assumptions on the unobserved parameters. These assumptions usually involve the parametric specification of the distribution of the unobserved effects and the imposition of restrictions on the type of dependence between right hand side variables and the unobserved heterogeneity in the data. The difficulty that such models pose in general lies in the likelihood function containing multivariate integrals that are often analytically intractable, but which are necessary as we integrate out the unobserved heterogeneity. Simulation based methods that seek to approximate such integrals have been reported as highly unreliable due to multiple local modes in the objective function and the lack of robustness to various maximization routines and starting values (Knittel and Metaxoglou 2008).

Relaxing limiting assumptions on the distribution of unobserved heterogeneity is important as in many economic models such restrictions cannot be justified by economic theory. Thus, an important generalization of these panel data models allows the individual heterogeneity to be correlated with the other regressors. Unfortunately, in the case of nonlinear panel data models it is not possible to remove the unobserved effects by differencing as is commonly done in linear models. One possibility is to treat the unobserved effects as nuisance parameters to be estimated along with the parameters of interest. Unfortunately, this approach requires large amounts of data, as consistency is guaranteed only in the large N and large T limit. In most microeconomic applications, the econometrician only has a small number of repeated cross-sections to work with and the estimation of the individual fixed effects as incidental parameters induces bias. In the logit case Abrevaya (1999) shows that the model with fixed effects and only two time periods leads to severe bias and the estimated coefficients are twice their true value. Note however that it is possible to sometimes circumnavigate this problem by redefining the quantity of interest. Fernandez-Val (2007) shows that under certain assumptions the inclusion of fixed effects does not affect the consistency of the marginal effects.

Removing the parametric assumptions on the distribution of unobserved heterogeneity is also a major concern since economic models are usually silent on how to formally describe individual heterogeneity. Furthermore, recent attempts at estimating nonlinear models nonparametrically require involved identification discussions and are often rather difficult to implement in the data (Berry and Haile 2008).

A third and very important weakness of existing models is that they typically require homoskedasticity over time (Arellano and Honore 2001). Economic models however often imply the existence of persistent shocks over time. Applied researchers routinely include time trends or time fixed effects to account for time dependence.

Fuelled by advances in computation, as well as their flexibility and conceptual simplicity, Bayesian methods provide a powerful alternative to the more traditional approaches to solving these problems. In particular, Bayesian hierarchical models can be readily extended to incorporate inference on latent classes of similar individuals or mixtures of distributions for various objects of interest. This makes Bayesian modeling an extremely flexible tool and a promising avenue to explore relaxing the assumptions discussed. Recently, Burda, Harding, and Hausman (2008) introduced a flexible mixed model for multinomial discrete choice where the key parameters of interest are allowed to follow an assumption-free nonparametric density specification.

In some special cases such as the probit model, Bayesian data augmentation completely avoids the need to specify the likelihood in the form of a multivariate integral. This feature was introduced for the probit model in a seminal paper by Albert and Chib (1993). Instead of formulating the likelihood by integrating out the latent utility, the estimation problem is re-cast in the form of an iterative scheme of linear regressions where the latent utility is explicitly sampled along with other model parameters. Thus, the estimation is free from the curse of dimensionality that plagues inference with integral-based likelihoods. The approach was further developed for LDV models to include parametric random latent effects in Albert and Chib (1996) and Chib and Carlin (1999).

In this paper, we further extend this line of research by introducing a model with two latent variables: first, we introduce unobserved individual heterogeneity that is allowed to vary in the population according to an assumption-free nonparametric distribution, and second a latent error component that is serially correlated over time. Due to the data augmentation strategy, virtually all model parameters are sampled directly without resorting to Metropolis-Hastings steps that would require additional tuning. The unobserved individual effects are allowed to be correlated with the observed regressors. Our model thus extends beyond the class of traditional random effects models (for a discussion on this issue, see e.g. Wooldridge 2001). We specify the Dirichlet Process (DP) prior for the distribution of the unobserved heterogeneity component, while parametrizing the latent time process. Our method of analysis is shown to perform favorably in simulations compared to existing benchmark models.

Our work complements a number of recent parametric attempts at modeling panel data with limited dependent variables, such as the panel probit model with AR(1) error terms of Burda, Liesenfeld, and Richard (2008) and Liesenfeld, Moura, and Richard (2008). A recent state-of-the-art Bayesian nonparametric analysis was introduced in Chib and Jeliazkov (2006) who study a binary dependent variable model with AR(p) errors and normally distributed unobserved individual heterogeneity. These authors focus on a non-parametric estimation of an unknown function of the model covariates, while we model nonparametrically the distribution of the unobserved individual heterogeneity not directly entering the model.

Another major contribution of our paper is the efficient computation of the posteriors using a new split-merge algorithm (Dahl 2005) that is substantially more efficient that samplers used previously in similar contexts. The SAMS sampler can update in one move large blocks of elements involved in implementation of the Dirichlet Process sampling scheme. It thus avoids a shortcoming of sequential samplers, such as the Polya urn scheme, that can get stuck in particular clustering configurations due to the one-at-a-time nature of their updates. Moreover, the SAMS algorithm is applicable to both conjugate and non-conjugate DP mixture model.

The use of panel probit models has increased in popularity with the availability of large N, large T data sets. There are many interesting applications in economics ranging from labor supply to the economic determinants of intra-state war. In this paper we chose to focus on a stylized application involving the determinants of firm innovation as measured by the patenting of new products. The study relies on a recent dataset collected by Bloom, Schankerman and Van Reenen (2007) which captures a majority of the patents granted between 1980 and 2001 in the US. We investigate the effect of technology spillovers and product market rivalry on patenting activity. We document the presence of both statistically significant technology spillover effects and firm level heterogeneity. The estimated distribution of firm level heterogeneity shows many interesting features and its multimodality suggests the clustering of heterogeneity across different firms.

The remainder of the paper is organized as follows. Section 2 introduces our model and discusses the assumptions and sampling procedures. Section 3 presents a series of Monte-Carlo studies comparing the performance of our new method with other traditional approaches, focusing on those that are easily available from commercial software such as STATA. Section 4 presents an application of the method to the estimation of the effect of technological spillovers and product market competition on innovation. Section 5 concludes.

2. Model

Consider a sample of binary responses y_{it}, for N individuals indexed by i, and T time periods indexed by t. We assume that the data is drawn from the following data-generating process:

$$\widetilde{y}_{it} = \mathbf{x}_{it}\beta + \tau_i + \lambda_t + \epsilon_{it} \tag{2.1}$$

$$y_{it} = \mathbf{1}\left(\widetilde{y}_{it} \geq 0\right) \tag{2.2}$$

where $\mathbf{1}\left(\mathcal{C}\right)$ denotes the indicator function which takes the value one if the condition $\mathcal{C}$ is satisfied and zero otherwise, x_{it} is a $(1 \times K)$ vector of explanatory variables, τ_i represents unobserved individual heterogeneity, and λ_t captures latent time effects. The term $\widetilde{y}_{it}$ can be thought of as a latent utility of individual i at time t. This is an error-components model where the unobserved error is decomposed into three parts, an individual specific error τ_i, a time specific component λ_t and an idiosyncratic and transitory shock ϵ_{it}. In this model we observe the covariates x_{it}, but not τ_i, λ_t or ϵ_{it}. We write the model in terms of the latent variable $\widetilde{y}_{it}$, which is not observed by the econometrician, who only observes the binary variable y_{it}.

Let $\widetilde{\mathbf{y}}_i = (\widetilde{y}_{i1}, \ldots, \widetilde{y}_{iT})'$, $\widetilde{\mathbf{y}} = (\widetilde{\mathbf{y}}'_1, \ldots, \widetilde{\mathbf{y}}'_N)'$, $\mathbf{X}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iT})'$, and $\mathbf{X} = (\mathbf{X}'_1, \ldots, \mathbf{X}'_N)'$, $\lambda = (\lambda_1, \ldots, \lambda_T)'$, $\tau = (\tau_1\iota', \ldots, \tau_N\iota')'$, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})'$, $\epsilon = (\epsilon'_1, \ldots, \epsilon'_T)'$ and let ι denote a $(T \times 1)$ vector of ones. Model (2.1) can thus be re-written more compactly as $\widetilde{\mathbf{y}}_i = \mathbf{X}_i\beta + \tau_i\iota + \lambda + \epsilon_i$ for $i = 1, \ldots, N$ or simply $\widetilde{\mathbf{y}} = \mathbf{X}\beta + \tau + \lambda + \epsilon$. Following the notation in Geweke (2005), let the set-valued function $C_{it} = c_{it}(\widetilde{y}_{it})$ with

$C_{it} = (-\infty, 0]$ if $y_{it} = 0$ and $C_{it} = (0, \infty)$ if $y_{it} = 1$. Denote the collection $\mathbf{C} = \{C_{it} : i = 1, \ldots, N; t = 1, \ldots, T\}$.

The hierarchical structure of our model alows us to distinguish four different layers of parameters. The first layer corresponds to the structure of the error components τ_i, λ_t and ϵ_{it}. Its properties are given by the following Assumption:

ASSUMPTION 1. *The error components τ_i, λ_t, and ϵ_{ti}, for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, are mutually independent conditionally on the X.*

The second parameter layer characterizes the distributional properties of the first-level parameters in Assumptions 2-4 specify the model for this level.

ASSUMPTION 2. *The variables $\tau_1, \ldots, \tau_N$ are independent, identically distributed conditionally on X*

$$\tau_i \sim F_\tau$$

where F_τ is a continuous smooth unknown distribution.

Instead of imposing a parametric family assumption, F_τ will be estimated as an infinite mixture of distributions using a Dirichlet Process prior which we shall introduce below. Alternative models that rely on Gaussianity of τ_i are bound to be misspecified with a detrimental effect for inference on model coefficients and average partial effects. Moreover, our sampling mechanism allows for joint posterior correlation of τ_i with other regressors even though we do not model this feature explicitly with prior assumptions in the absence of any initial information on this property. Since τ_i is sampled conditional on X_i, such potential relationship is entirely data-driven. The next key assumption concerns the nature of the latent time effects:

ASSUMPTION 3. *λ_t is assumed to follow a stationary Gaussian autoregressive process*

$$\lambda_t = \rho_1 \lambda_{t-1} + \ldots + \rho_s \lambda_{t-s} + \eta_t,$$

with $\eta_t \sim N(0, \sigma_\eta^2)$. Furthermore, η_t is independent of ϵ_{ti} for each $t = 1, \ldots, T$.

Failure to account for serial correlation of the error term has potential negative consequences. The fixed effects approach of including time dummies to capture the time effect will increase the number of parameters to be estimated linearly with T which may not be desirable for moderate or larger datasets in the time dimension. Our model avoids these shortcomings by its flexible specification, at the cost of increased complexity of the sampling scheme.

The following assumption defines the probit structure of the model:

ASSUMPTION 4. *$\epsilon_{it} \sim N(0, 1)$ is a stochastic error component uncorrelated with any other regressor.*

Our proposed model builds on the traditional error-components framework due to its popularity in applied work. The random error to an observation, $\tau_i + \lambda_t + \epsilon_{it}$ is given by the sum of an individual effect τ_i, a time effect λ_t and an idiosyncratic shock ϵ_{it} independently distributed across time and cross-sectional units. Variations on this framework are however possible and easily absorbed by our methods.

One possible alternative is to allow for the individual effects to depend on some observables w_i, where the first row of w_i is a vector of ones. The binary response model now takes the form:

$$\begin{aligned} \widetilde{y}_{it} &= \mathbf{x}_{it}\beta + w_i\tau_i + \lambda_t + \epsilon_{it} \\ y_{it} &= \mathbf{1}\left(\widetilde{y}_{it} \geq 0\right). \end{aligned}$$

This model can be estimated in a very similar fashion to our discussion below. The only exception being that τ_i is now a vector to be sampled from a multivariate Dirichlet process. This extends the current framework to a setup with both a random intercept and random coefficients and may potentially add extra flexibility. Estimating such a model if feasible when the number of variables captured by w_i is small. The estimation details for a multivariate Dirichlet process are very similar to those discussed in Burda, Harding and Hausman (2008) for the multinomial case and will not be repeated here.

It is also possible to make alternative assumptions on the time component which can be easily captured by our modeling approach. Let's assume for the moment that the time component takes the simple AR(1) form, $\lambda_t = \rho\lambda_{t-1} + \eta_t$. By repeated substitution for the case of large T we find that our model can be re-writen as:

$$\begin{aligned} \widetilde{y}_{it} &= \mathbf{x}_{it}\beta + \tau_i + \eta_t + \rho\eta_{t-1} + \rho^2\eta_{t-2} + \cdots + \epsilon_{it} \\ y_{it} &= \mathbf{1}\left(\widetilde{y}_{it} \geq 0\right). \end{aligned}$$

The model allows outcomes to depend on a weighted average of past realizations of a latent time process η_t in addition to the idiosyncratic innovations ϵ_{it}. This has the attractive feature of allowing us to interpret the unobserved time shocks as consisting of a permanent shock η_t with exponentially declining weights ρ^t and a transitory component ϵ_{it}. While many economic models share this feature, other alternatives are also possible such as imposing an AR(p) assumption on the error term ϵ_{it} directly. The use of different time correlation assumptions and the resulting estimation challenges in the probit model are discussed in more detail in Chib and Jeliazkov (2006).

Notice that while we allow for the individual effects τ_i to follow an unknown distribution, we maintain the parametric assumption on remainder of the error term. This is done for identification purposes in order to avoid the convolution of two unknown distributions. A possible alternative might be to specify the distribution of unoberved individual effects parametrically and then let the observational error ϵ_{it} be assumption free. This however produces further complications since it requires a scale restriction on the link function in order to guarantee that the observed outcome is bounded between 0 and 1. To our knowledge the imposition of such scale restrictions in Dirichlet Process Mixture models has not yet been attempted but it may provide an interesting avenue for future research. The model introduced in this section is however closer to current economic practice which relies on structural assumptions to identify the phenomenon of interest while allowing for the "deeper" parameters such as preferences or individual heterogeneity to remain assumption free. We therefore find the current setup more appealing from an economic point of view.

The third parameter layer is formed by parameters of primary economic interest captured in the vector $\theta = (\beta', \sigma_\eta, \rho')'$. The prior assumptions on this layer are as follows:

ASSUMPTION 5.

$$\beta \sim N(\underline{\beta}, \underline{\Sigma}_\beta) \tag{2.3}$$

$$\sigma_\eta^2 \sim IG(v_0, s_0) \tag{2.4}$$

$$\rho \sim Uniform(\Omega) \tag{2.5}$$

where $\Omega \subseteq \mathbb{R}^s$ is the stationarity region of the autoregressive process.

The fourth parameter layer is comprised of all hyperparameters introduced in Assumptions 2-5. In order to fully characterize this layer, we will elaborate on the model specified for the distribution of the unobserved heterogeneity component. Assumption 2 implies the following model based on Neal (2000):

$$\tau_i|\psi_i \sim F_\tau(\psi_i) \tag{2.6}$$

$$\psi_i|G \sim G \tag{2.7}$$

$$G \sim DP(\alpha, G_0) \tag{2.8}$$

Thus, F_τ is specified as an infinite mixture of distributions $F_\tau(\psi)$ with the mixing distribution over ψ being G. Here, ψ_i are hyperparameters of the distribution $F_\tau(\psi_i)$ of τ_i drawn from a random probability measure G which itself is distributed according to a Dirichlet Process (DP) prior. The DP prior for G is indexed by two hyperparameters: a distribution G_0 that defines the "location" of the DP prior, and a positive scalar precision parameter α. The distribution G_0 may be viewed as a baseline prior that would be used in a typical parametric analysis. The flexibility of the DP prior model environment stems from allowing G to stochastically deviate from G_0. The precision parameter α determines the concentration of the prior for G around the DP prior location G_0 and thus measures the strength of belief in G_0. For large values of α, a sampled G is very likely to be close to G_0, and vice versa.

The fourth parameter layer is thus formed by the hyperparameters $\{\psi_i\}_{i=1}^N$, G, α, G_0, $\underline{\beta}$, $\underline{\Sigma}_\beta$, v_0, and s_0. In our implementation, G_0, $\underline{\beta}$, $\underline{\Sigma}_\beta$, v_0, and s_0 are fixed, $\{\psi_i\}_{i=1}^N$, and α are sampled, while bypassing explicit sampling of G. The details are given further below.

2.1. Gibbs Sampling

Under the Model 2.1 and Assumptions 1-5, the joint posterior density can be decomposed into the following Gibbs blocks:

(1) $\beta|\tau, \lambda, \psi, \theta_{/\beta}, \widetilde{\mathbf{y}}, \mathbf{y}, \mathbf{X}$
(2) $\widetilde{\mathbf{y}}|\tau, \lambda, \psi, \theta, \mathbf{y}, \mathbf{X}$
(3) Update the assignments of τ_i to latent classes by alternating between the SAMS (Dahl 2005) and Algorithm 7 (Neal 2000), which includes sampling $\{\psi_i\}_{i=1}^N$
(4) $\tau_i|\psi, \theta, \lambda, \widetilde{\mathbf{y}}, \mathbf{y}, \mathbf{X}$ for each i

(5) $\lambda|\tau, \psi, \theta, \widetilde{\mathbf{y}}, \mathbf{y}, \mathbf{X}$
(6) $\sigma_\eta^2|\tau, \lambda, \psi, \theta_{/\sigma_\eta^2}, \widetilde{\mathbf{y}}, \mathbf{y}, \mathbf{X}$
(7) $\rho|\tau, \lambda, \psi, \theta_{/\rho}, \widetilde{\mathbf{y}}, \mathbf{y}, \mathbf{X}$

We will now elaborate on each of these blocks individually.

2.2. **Sampling** β

In this block we apply the method of Albert and Chib (1993) to the recentered latent variable

$$\widetilde{y}_{it}^* = \widetilde{y}_{it} - \tau_i - \lambda_t.$$

Thus, the joint conditional density of $(\beta, \widetilde{\mathbf{y}}^*)$ is given by

$$p(\beta, \widetilde{\mathbf{y}}^*|\tau, \lambda, \psi, \theta_{/\beta}, \widetilde{\mathbf{y}}, \mathbf{y}, \mathbf{X}) \propto \exp\left[-\frac{1}{2}\left(\beta - \underline{\beta}\right)' \underline{\Sigma}_\beta^{-1}\left(\beta - \underline{\beta}\right)\right] \exp\left[-\frac{1}{2}\left(\widetilde{\mathbf{y}}^* - \mathbf{X}\beta\right)'\left(\widetilde{\mathbf{y}}^* - \mathbf{X}\beta\right)\right]$$

yielding a closed form of the conditional posterior for β which facilitates direct sampling

$$\beta|\cdot \sim N\left(\overline{\beta}, \overline{\Sigma}\right)$$

where

$$\begin{aligned} \overline{\beta} &= \overline{\Sigma}\left(\underline{\Sigma}^{-1}\underline{\beta} + \mathbf{X}'\widetilde{\mathbf{y}}^*\right) \\ \overline{\Sigma} &= \left(\underline{\Sigma}^{-1} + (\mathbf{X}'\mathbf{X})\right)^{-1} \end{aligned}$$

In the application, we specify the hyperparameter values $\underline{\beta} = 0$ and $\underline{\Sigma}_\beta = 10I$ where I is the identity matrix. This specification is aimed at making the prior for β sufficienty diffuse.

2.3. **Sampling** $\widetilde{y}_{it}$

Here we benefit from the second step of the Albert and Chib (1993) procedure, augmented by τ_i and λ_t. Thus, sample directly

$$\begin{aligned} \widetilde{y}_{it}|\cdot &\sim N(v_{it}, 1) \\ v_{it} &= \mathbf{x}_{it}\beta + \tau_i + \lambda_t \end{aligned}$$

truncated by 0 from the left if $y_{it} = 1$ and from the right if $y_{it} = 0$.

2.4. **Updating Latent Class Assignments**

For this block we utilize a hybrid sampler that alternates between the non-conjugate version of the Sequentially Allocated Split-Merge (SAMS) sampler (Dahl, 2005), and Algorithm 7 of Neal (2000). This approach is suggested by Dahl (2005) as optimally combining the virtues of each method: the ability to move large blocks of elements among latent classes in one step for the former, and one-at-a-time allocations of individual elements among latent classes for the latter.

The SAMS sampler is based on an alternative expression of the model (2.6)-(2.8) in terms of a set partition $\pi = \{S_1, \ldots, S_q\}$ for $S_0 = \{1, \ldots, n\}$ in addition to the latent class parameters $\phi = \{\phi_{S1}, \ldots, \phi_{Sq}\}$ where ϕ_S is associated with component S. The set partition π for S_0 is a set of subsets $S_1, \ldots, S_q$ such that (1) $\cup_{S\in\pi} S = S_0$, (2) $S \cap S^* = \emptyset$ for all $S \neq S^*$, and (3) $S \neq \emptyset$ for all $S \in \pi$. Using this notation, the model (2.6)-(2.8) can be recast as (Dahl, 2005):

$$\tau_i|\pi,\phi \sim F_\tau(\phi_S^i) \tag{2.9}$$

$$\psi|\pi \sim \prod_{S\in\pi} G_0(\phi_S) \tag{2.10}$$

$$\pi \sim b\prod_{S\in\pi} \eta_0\Gamma(|S|) \tag{2.11}$$

where $|S|$ is the number of elements of the component S. The sampling scheme works as follows: In each MC iteration, uniformly select a pair of distinct indeces i and j. If i and j belong to the same component in π, say S, propose π^* by splitting S. Otherwise, i and j belong to different components in π, say S^i and S^j. Propose π^* by merging S^i and S^j. In each case, compute the Metropolis-Hastings (MH) ratio $a(\pi^*, \phi^*|\pi, \phi)$ and accept the new latent class configuration π^* with probability given by this ratio. We derive the MH ratio for our model in the following Section.

Algorithm 7 of Neal (2000), which we utilize in every alternate MC step, is based on limiting probabilities of a latent class finite mixture model with the number of classes tending to infinity. The sampling procedure itself is built around drawing with a stochastic number of mixture components or classes whose number and size varies at each MC iteration. Denote by c a label of a generic latent class with membership count N_c. Given the current state of the system, τ_i are first re-assigned into latent classes with labels c_i whereby new classes can be created and old ones may vanish. The probabilities of class assignment for the τ_i are proportional to the likelihood of τ_i conditional on the current draw of the class parameters ψ_c. Second, the class parameters ψ_c are updated in a standard way for each class separately. If we specify F_τ as an infinite mixture of Normals, then $\psi_c = (\mu_{\tau c}, \sigma^2_{\tau c})$ are the moments of the Normal density.

For updating ψ in the Algorithm 7 scan, we specify F_τ as a mixture of Normals with $\psi = (\mu_\tau, \sigma^2_\tau)$. Since for all τ_i that fall into one latent class it holds that $\tau_i \sim N(\mu_{\tau c}, \sigma^2_{\tau c})$ we can apply result B (p. 300) of Train (2003) to each latent class separately: for a $IG(s_0, v_0)$ prior, the posterior of $\sigma^2_{\tau c}$ is given by $IG(s_1, v_1)$ with $v_1 = v_0 + N_c$ and $s_1 = (v_0 s_0 + N_c \bar{s}_{c_i})/(v_0 + N_c)$ where $\bar{s}_c = N_c^{-1}\sum_{i=1}^{N_c} \tau_i^2$. We utilize a diffuse IG prior. Analogously, to sample $\mu_{\tau c}$ we use result A of Train (2003) applied to each latent class. The hyperparameter of the DP prior α is sampled according to the scheme of Escobar and West (1995).

2.5. **Sampling τ_i**

Let $\widetilde{\mathbf{y}}_i^{**} = \widetilde{\mathbf{y}}_i - \mathbf{X}_i\beta - \lambda$. Then

$$\widetilde{\mathbf{y}}_i^{**} = \tau_i\iota + \epsilon_i$$

Consider for the moment the case $\tau_i \sim N(\underline{\tau}, \sigma^2_\tau)$; it will be used as a building block in the DP prior sampling. In this case, for every i we have one latent regression with one parameter τ_i and a $(T \times 1)$ vector of ones as explanatory variables in place of a hypothetical $\mathbf{X}_i$. Using standard latent regression results (see e.g. Lancaster, 1997),

$$\begin{aligned}
&(2.12) \qquad & p(\tau_i|\cdot) &= \phi\left(\overline{\tau}_i, \overline{\sigma}^2_{\tau i}\right) \\
&(2.13) \qquad & \overline{\tau}_i &= \overline{\sigma}^2_{\tau i}\left(\sigma_\tau^{-2}\underline{\tau} + \sum_{i=1}^{T}\widetilde{\mathbf{y}}_i^{**}\right) \\
&(2.14) \qquad & \overline{\sigma}^2_{\tau i} &= \left(\sigma_\tau^{-2} + T\right)^{-1}
\end{aligned}$$

Since $\tau_i \sim N(\mu_{\tau c}, \sigma^2_{\tau c})$ given a previous assignment to the latent class c, let $\underline{\tau} = \mu_{\tau c}$, $\sigma^2_\tau = \sigma^2_{\tau c}$ and sample τ_i directly from (2.12).

2.6. Sampling λ

This is a variant of the same scheme as for direct sampling of β, with a special structure of the X variables and recentered y. Let $\widetilde{\mathbf{y}}_{i\lambda} = \widetilde{\mathbf{y}}_i - \mathbf{X}_i\beta - \tau_i\iota$ and $\mathbf{X}_\lambda = \iota_N \otimes \mathbf{I}_T$. Then

$$\widetilde{\mathbf{y}}_\lambda = \mathbf{X}_\lambda\lambda + \varepsilon$$

and hence we can sample directly from $N\left(\overline{\lambda}, \overline{\Sigma}_\lambda\right)$ where

$$\begin{aligned}
\overline{\lambda} &= \overline{\Sigma}_\lambda\left(\underline{\Sigma}_\lambda^{-1}\underline{\lambda} + \mathbf{X}_\lambda'\widetilde{\mathbf{y}}_\lambda\right) \\
\overline{\Sigma}_\lambda &= \left(\underline{\Sigma}_\lambda^{-1} + (\mathbf{X}_\lambda'\mathbf{X}_\lambda)\right)^{-1}
\end{aligned}$$

Since λ follows a mean-zero process, we set the prior $\underline{\lambda} = 0$. $\underline{\Sigma}_\lambda$ is the prior covariance matrix of the autoregressive process. For ease of implementation we restrict ourselves to the $AR(1)$ specification.

2.7. Sampling ρ

Note that for the $AR(1)$ process,

$$p(\lambda_t|\lambda_{t-1},\cdot) \propto \begin{cases} \exp\left(-\frac{(1-\rho^2)}{2\sigma^2_\eta}\lambda_1^2\right), & t = 1 \\ \exp\left(-\frac{1}{2\sigma^2_\eta}(\lambda_t - \rho\lambda_{t-1})^2\right), & t = 2, ..., T \end{cases}$$

and hence

$$\begin{aligned}
p(\rho|\lambda) &= \exp\left(-\frac{1}{2\sigma^2_\eta}\left[(1-\rho^2)\lambda_1^2 + \sum_{t=2}^{T}(\lambda_t - \rho\lambda_{t-1})^2\right]\right) \\
&= \exp\left(\frac{1}{2\sigma^2_\eta}\left[\rho^2\left(\sum_{t=2}^{T}\lambda_{t-1}^2 - \lambda_1^2\right) - 2\rho\sum_{t=2}^{T}\lambda_t\lambda_{t-1} + \sum_{t=1}^{T}\lambda_t^2)\right]\right)
\end{aligned}$$

Matching this expression with a Gaussian kernel $\exp\left(-\frac{1}{2\sigma^2}\left[\rho^2 - 2\rho\mu + \mu^2\right]\right)$ yields

$$
\begin{aligned}
\overline{\sigma}_\rho^2 &= \sigma_\eta^2 \left(\sum_{t=2}^{T-1} \lambda_t^2\right)^{-1} \\
\overline{\mu}_\rho &= \frac{\overline{\sigma}_\rho^2}{\sigma_\eta^2} \sum_{t=2}^{T} \lambda_t \lambda_{t-1} \\
&= \left(\sum_{t=2}^{T-1} \lambda_t^2\right)^{-1} \sum_{t=2}^{T} \lambda_t \lambda_{t-1}
\end{aligned}
$$

We can therefore sample ρ directly from $N(\overline{\mu}_\rho, \overline{\sigma}_\rho^2)$ truncated at -1 and 1 to preserve stationarity. Extension to $AR(p)$ will amend the likelihood function $p(\lambda_t|\lambda_{t-1}, \cdot)$ but the approach would be quite similar to the current case.

2.8. Sampling σ_η^2

We use the result derived in Burda, Liesenfeld, and Richard (2008) which adapts the standard result on sampling univariate variances (given e.g. by result B, p. 300, of Train 2003) to the likelihood of the variance of the AR process. Conditional on λ and ρ, the likelihood function of σ_η^2 takes the form

$$
L(\sigma_\eta^2|\lambda, \theta_{/\sigma_\eta^2}) \propto \frac{\sqrt{1-\rho^2}}{\sigma_\eta \sqrt{2\pi}} \exp\left[-\frac{1-\rho^2}{2\sigma_\eta^2}\lambda_1^2\right] \prod_{t=2}^{T} \frac{1}{\sigma_\eta\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_\eta^2}\left(\lambda_t - \rho\lambda_{t-1}\right)^2\right]
$$

An $IG(v_0, s_0)$ prior has density

$$
k(\sigma_\eta^2) = \frac{1}{m_0 \sigma_\eta^{(v_0+1)/2}} \exp\left[-\frac{v_0 s_0}{2\sigma_\eta}\right]
$$

where m_0 is a normalizing constant. We can then sample directly from the posterior

$$
\begin{aligned}
L(\sigma_\eta^2|\lambda, \theta_{/\sigma_\eta^2}) &\propto L(\sigma_\eta^2|\lambda, \theta_{/\sigma_\eta^2}) k(\sigma_\eta^2) \\
&\propto \frac{1}{\sigma_\eta^{(T+v_0+1)/2}} \exp\left[-\frac{\left(1-\rho^2\right)\lambda_1^2 + \sum_{t=2}^{T}\left(\lambda_t - \rho\lambda_{t-1}\right)^2 + v_0 s_0}{2\sigma_\eta^2}\right] \\
&= IG(v_1, s_1)
\end{aligned}
$$

where

$$
\begin{aligned}
v_1 &= v_0 + T \\
s_1 &= \frac{v_0 s_0 + \left(1-\rho^2\right)\lambda_1^2 + \sum_{t=2}^{T}\left(\lambda_t - \rho\lambda_{t-1}\right)^2}{v_0 + T}
\end{aligned}
$$

In the application, the prior for σ_η^2 will be specified as diffuse with $s_0 \to 0$ and $v_0 = 0$.

2.9. The SAMS sampler

In this Section, we explicitly derive the form of the MH ratio for our case. For a general description, see Dahl (2005). Let k be the successive values in random permutations of the indeces in S. In our model, the

MH ratio is given by

$$a(\pi^*, \phi^*|\pi, \phi) = \min\left[1, \frac{p(\pi^*, \phi^*|y)}{p(\pi, \phi|y)} \frac{q(\pi, \phi|\pi^*, \phi^*)}{q(\pi^*, \phi^*|\pi, \phi)}\right]$$

If the proposal involves a split, $q(\pi^*, \phi^*|\pi, \phi)$ is the split probability and $q(\pi, \phi|\pi^*, \phi^*) = 1$ is the merge probability. If the proposal involves a merge, the roles of $q(\pi^*, \phi^*|\pi, \phi)$ and $q(\pi, \phi|\pi^*, \phi^*)$ are reversed. Consider e.g. proposal for a split:

$$q(\pi^*, \phi^*|\pi, \phi) = \prod_{k=1}^{N} P\left(k \in S^i|S^i, S^j, \phi, y\right) P(\phi_{S^i})$$

The first term is given in equation (13) in Dahl (2005). The second term $P(\phi_{S^i})$ is the proposal density of the new ϕ_{S^i}. The merge probability is

$$q(\pi, \phi|\pi^*, \phi^*) = 1$$

By Bayes theorem,

$$p(\pi, \phi|y) \propto p(y|\pi, \phi)p(\pi, \phi) \tag{2.15}$$

where $p(y|\pi, \phi)$ is the likelihood

$$p(y|\pi, \phi) = \prod_{i=1}^{n} p(y_i|\phi_{S^i})$$

and $p(\pi, \phi)$ is the prior

$$p(\pi, \phi) = p(\phi|\pi)p(\pi) \tag{2.16}$$

where

$$\begin{aligned} p(\phi|\pi) &= {}_{S\in\pi} F_0(\phi_S) \\ p(\pi) &= b_{S\in\pi}\eta_0\Gamma(|S|) \\ b^{-1} &= \prod_{i=1}^{n} \Gamma\left(\eta_0 + i - 1\right) \end{aligned}$$

Note that for a split of a class S^s into S^i and S^j,

$$\frac{p(y|\pi^*, \phi^*)}{p(y|\pi, \phi)} = \frac{\prod_{t=1}^{|S^i|} p(y_t|\phi_{S^i}) \prod_{t=1}^{|S^j|} p(y_t|\phi_{S^j})}{\prod_{t=1}^{|S^s|} p(y_t|\phi_{S^s})} \tag{2.17}$$

where the index t in $p(y_t|\phi_{S^i})$ refers to elements of the class S^i. Similarly, for a merge of classes S^i and S^j into S^s,

$$\frac{p(y|\pi^*, \phi^*)}{p(y|\pi, \phi)} = \frac{\prod_{t=1}^{|S^s|} p(y_t|\phi_{S^s})}{\prod_{t=1}^{|S^i|} p(y_t|\phi_{S^i}) \prod_{t=1}^{|S^j|} p(y_t|\phi_{S^j})}$$

i.e. the inverse of the ratio of split probabilities. Note that for a split we can use the stored values of the likelihood evaluations from the allocation of k into S^i and S^j. Hence only two additional likelihood evaluations $p(y_i|\phi_{S^i})$ and $p(y_j|\phi_{S^j})$ that initiated the split need to be performed for obtaining the ratio $\frac{p(y|\pi,\phi)}{p(y|\pi^*,\phi^*)}$. For a merge, only $2|S^i| + |S^j|$ likelihood evaluations need to be performed, which for small classes can be substantially less than the sample size n.

In the same spirit, for computing prior components for a split

$$\frac{p(\phi^*|\pi^*)}{p(\phi|\pi)} = \frac{F_0(\phi_{S^i})F_0(\phi_{S^j})}{F_0(\phi_{S^s})} \tag{2.18}$$

and

$$\frac{p(\pi^*)}{p(\pi)} = \frac{\Gamma(|S^i|)\Gamma(|S^j|)}{\Gamma(|S^s|)} \tag{2.19}$$

while for a merge

$$\frac{p(\phi^*|\pi^*)}{p(\phi|\pi)} = \frac{F_0(\phi_{S^s})}{F_0(\phi_{S^i})F_0(\phi_{S^j})}$$

and

$$\frac{p(\pi^*)}{p(\pi)} = \frac{\Gamma(|S^s|)}{\Gamma(|S^i|)\Gamma(|S^j|)}$$

Thus, using (2.15) - (2.19), the ratio of the p-terms for a split becomes

$$\frac{p(\pi^*, \phi^*|y)}{p(\pi, \phi|y)} = \frac{\prod_{t=1}^{|S^i|} p(y_t|\phi_{S^i}) \prod_{t=1}^{|S^j|} p(y_t|\phi_{S^j})}{\prod_{t=1}^{|S^s|} p(y_t|\phi_{S^s})} \frac{F_0(\phi_{S^i})F_0(\phi_{S^j})}{F_0(\phi_{S^s})} \frac{\Gamma(|S^i|)\Gamma(|S^j|)}{\Gamma(|S^s|)} \tag{2.20}$$

while for a merge this ratio is given by the inverse of the expression in (2.20).

3. Monte-Carlo Evidence

We test the performance of our approach on a series of Monte Carlo studies based on simulated datasets. Our goal is to evaluate the robustness of our approach to relaxing the three main assumptions which are the main focus of our study. We wish to estimate a discrete valued probit model in the presence of individual heterogeneity, which is potentially correlated with other right hand side variables and autoregressive time effects.

While establishing good numerical properties of this new method is essential, it is also important to check the extent to which the estimation is actually improved in a number of samples over what one would have done if this new approach were not available. Choosing a benchmark estimation procedure is more of an art than a science and thus we proceed with the assumption that most economists are inclined to estimate such models using one of the available models in the popular STATA econometrics package. Clearly, other alternatives abound, but it is beyond the scope of this paper to attempt such a detailed comparison.

An economist approaching the econometric model in equation 2.1 for the first time has a simple fall-back option, to ignore the error components structure and estimate the model using a simple pooled probit regression without considering the impact of individual heterogeneity or correlated time effects. Note that this effectively treats all observations is iid draws from the underlying data generating process and ignores the panel data structure completely.

Most likely however the standard approach will involve the estimation of a panel data probit model with random effects and a full set of time dummies to capture any existing stochastic time trends. Note that the presence of the additional incidental parameters may incur non-negligible computational costs as we shall discuss below.

We will refer to these basic models as our benchmark cases as they can be easily approached using existing STATA procedures. The first econometric model is estimated using the `probit` command, while the second is estimated using `xtprobit,re` with a full set of time dummies accounting for the latent time effects.

3.1. **Average Partial Effects**

Since in nonlinear models the estimated coefficients are only of limited interest by themselves, it is often convenient when evaluating the performance of a model to also consider other estimated quantities of interest. Average partial effects (APEs) are particularly useful for computing economic counterfactuals and widely used in applied work. We describe how they can be computed within the setup of our model and actual simulation results will be covered in the remainder of this section.

In nonlinear models, the APEs are functions of the coefficient estimates and hence need to be computed separately. In our case, the APEs can be calculated directly based on their definition. We utilize the classical concept of the APEs augmented with the latent variables. Let

$$\begin{aligned} m_{itk} &= \frac{\partial E[y_{it}|\mathbf{x}_{it}\beta, \tau_i, \lambda_t]}{\partial \mathbf{x}_{itk}} \\ &= \phi\left(\mathbf{x}_{it}\beta + \tau_i + \lambda_t\right)\beta_k \end{aligned}$$

denote the marginal effect of a change in $\mathbf{x}_k$, where $\phi(\cdot)$ denotes the standard normal density function. Define

$$\widetilde{\gamma} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\phi\left(\mathbf{x}_{it}\beta + \tau_i + \lambda_t\right) \quad (3.1)$$

The APE of $\mathbf{x}_k$ on $\mathbf{y}$ is then given by

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m_{itk} = \widetilde{\gamma}\beta_k \quad (3.2)$$

Due to data augmentation, we sample explicitly τ_i and λ_t throughout the MC iterations and hence can compute the APEs directly as from the equivalent computation of $\widetilde{\gamma}$,

$$\gamma = \frac{1}{NTS}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{S}\phi\left(\mathbf{x}_{it}\beta_s + \tau_{is} + \lambda_{ts}\right) \quad (3.3)$$

where s is the index over MC steps. This result is standard in the Bayesian literature on data augmentation. In both the application and the simulation study, we also report the mean bias and means squared error of the estimated "APE scale" coefficient γ defined in (3.3). To obtain the APEs, γ is simply multiplied by each respective β_k.

3.2. **Simulation Results**

All simulation results are reported in Tables 1-4. We refer to our model as the "Flexible Latent Effects Probit" (FLE) model. Keeping the same simulation design we vary the amount of data available at each run by varying N and T. We consider simulations where $N \in \{100, 300, 700, 1000\}$ and $T \in \{10, 20, 50\}$.

We then report the mean bias and root mean square error for each coefficient β_k in our simulation design. Additionally, for our model we also report these statistics for the latent time process parameters ρ and σ_η.

One hundred replications of the pooled probit for the average sample size ran virtually instantly on a 3.0 GHz dual-core Windows computer in Stata. However, the same set of results for the RE probit with time dummies took almost a day to calculate the same computer. By contrast our FLE probit took less than half hour on a 2.0 GHz unix machine using our Fortran code. We have found our approach to be computationally very efficient. All chains mix very well and appear to have converged within the burn-in section. The autocorrelation of β_k become statistically insignificant after 10-15 lags while the ones of σ_η and ρ after 1-2 lags.

In spite of the fairly involved sampling procedure required by our method, it does have a very important feature. The iteration between the samplers of Dahl (2005) and Neal (2000) removes the need for starting values. The SAMS sampler is capable of re-allocating large blocks of data to one of the latent classes while the Neal algorithm addresses the individual by individual allocation to latent classes. This allows us to initialize the procedure with a unique parametric component. This is then rapidly split into classes by the SAMS sampler before the Neal procedure continues to fine-tune the posterior draws. As the Markov chain evolves only individual changes are made and the Neal procedure dominates in terms of MH step acceptance. Only very rarely the data will warrant a large re-allocation of the observations between latent classes. After a few hundred MC steps, the average acceptance rate of the SAMS MH step falls to around 2%.

In Tables 1-2 we present the mean absolute deviation and mean square errors for a simple classical data generating process where τ_i is uncorrelated with any of the included right hand side variables $\mathbf{x_k}$. We generate the data as follows. All $\mathbf{x_k}$ observations are drawn as $N(0, 1/9)$. The true parameter values are $\beta_0 = \{0, 1, 1, -1\}$, $\sigma_{\eta 0} = 0.5$, $\rho_0 = 0.5$. We draw the unobserved heterogeneity as an equal mixture between two Normal components, with means -2 and 2 and variance 1/5. The length of the MCMC chain is 5000 and we discard the first 4000 steps as the burn-in period. We initialize the algorithm by letting all observations be part of the same latent class, and allow the sampler to do further splits as required.

The pooled probit model performs badly in terms of mean bias independently of the sample size. Typically we find mean bias in excess of 75%. The bias in the APE scale is also very large. Recall that the APE scale is bounded between 0 and (approximately 0.4), thus a mean bias of 0.3 is extremely large. If we now estimate the same models using the random effects probit with time effects, the performance will improve, but only slightly. This model controls for the unobserved heterogeneity using Normally distributed individual effects. Notice that since the underlying data generating process is bi-modal and the modes are away from zero, the Normal assumption on the individual effects will perform badly. In particular, the APE scale remains severely biased throughout. Furthermore, the computation is extremely slow for large T. The case with $N = 1000$ and $T = 50$ took several days to run.

The proposed Bayesian model by contrast performs very well. Even in samples as small as $N = 100$ and $T = 10$ the bias in the coefficients is lass than 5%. The bias in the APE scale is very small throughout. The coefficient estimates have small mean squared errors and we find no evidence of overdispersion due to the nonparametric component of our model. The flexible latent effects probit model additionally estimated the autoregressive coefficient ρ and the standard deviation of the time process innovations σ_η. The estimation of these parameters is rather difficult if the time dimension is very short. If $T = 50$ however the estimator performs very well. Notice however that the estimation of the model parameters β and of the marginal effects does not depend on these time parameters, which explains why the β are sharply estimated even in panel data with only a few of repeated observations.

In Tables 3-4 we report results for the simulation design where the individual effects are correlated with the right hand side variables. In this case we augment both the pooled probit and the random effects probit in STATA with time means of the $\mathbf{x}$ variables which alleviates potential biases associated with such correlation (Chamberlain 1982). We employ the following simple classical simulation design. The variables $\mathbf{x_2}$ and $\mathbf{x_3}$ are drawn uniformly on $[-1, 1]$. We generate $\mathbf{x_1}$ to be correlated with τ_i employing the following stylized approach to induce correlations. For a draw $u_i \sim U[0, 1]$, we generate τ_i and $\mathbf{x_1}$ as:

- 1st quarter of individuals: $\tau_i = 3 + u_i$ and $\mathbf{x_{1i}} = -4 - 0.2u_i$
- 2nd quarter of individuals: $\tau_i = -3 - u_i$ and $\mathbf{x_{1i}} = -1 - 0.2u_i$
- 3rd quarter of individuals: $\tau_i = 3 + u_i$ and $\mathbf{x_{1i}} = 1 - 0.2u_i$
- 4th quarter of individuals: $\tau_i = -3 - u_i$ and $\mathbf{x_{1i}} = 4 - 0.2u_i$

This yields a correlation coefficient between τ_i and $\mathbf{x_{1i}}$ close to -0.5 in each case. In some cases, inclusion of the time means of $\mathbf{x}$ may capture some of the effect of unobserved heterogeneity in computing the APEs. However, in our simulation the mass of τ_i is distributed in each quadrant with respect to $\mathbf{x_{1i}}$ and hence the role of the time-means is limited. Omission of the time means resulted in larger biases on all estimated parameters though. The true parameter values are $\beta_0 = \{0, 0.5, 0.5, 0.5\}$, $\sigma_{\eta 0} = 0.75$, $\rho_0 = 0.5$. The length of the MCMC chain is 5000 and we discard the first 4000 steps as the burn-in period. We initialize the algorithm by letting all observations be part of the same latent class, and allow the sampler to do further splits as required.

The coefficients estimates derived from running the pooled probit model are severely downward biased. The bias appears to be very similar at different sample sizes. In fact it seems that a sample with $N = 1000$, $T = 50$ leads to results which are just as biased as those derived from pooling a sample with $N = 100$, $T = 10$ observations. The estimate of the APE scale is also severely biased since it ignores both the individual effects and the latent time effects. The parameter estimates of the random effects probit are also biased and some coefficients are estimated with a large mean square error. The the APE scale bias does not appear to diminish with increasing N. Controlling for time effects with time dummies is computationally challenging and may provide unreliable inference on the other coefficient estimates when T is large.

By contrast, the Bayesian model proposed in this paper performs very well both in terms of bias and mean square error. The bias in the APE scale is also very small, indicating that the marginal effects will be nearly unbiased in almost all samples considered in this exercise. The autocorrelation parameter ρ is estimated with bias in small samples but the estimation bias drops below 3% as we increase the sample size to $T = 50$. Once again notice that the unbiased estimation of this parameter is not required to achieve unbiased marginal effects.

To summarize our simulation results, we have found that pooling the data and ignoring individual heterogeneity and latent time processes leads to extremely biased results. The common practice of estimating a panel probit model with random effects augmented with time dummies and time means of $\mathbf{x}$ improves estimation but the biases persist. Moreover, the estimation of a large number of incidental parameters poses computational challenges and leads to inconsistent estimates. The flexible latent effects probit model introduced in this paper, performs very well and leads to nearly unbiased results for both the model parameters and the estimated average effects. We caution though against placing too much reliance on the estimated auto-correlation parameters in the effects, since it is less reliable in small samples when there is a limited number of observations across time.

4. Application: Innovation and R&D Spillovers

We apply the method developed in this paper to the estimation of a patent equation on firm level data and compare the estimation results with those obtained by more traditional approaches. We employ data from a recent study of firm level research and development (R&D) by Bloom, Schankerman and Van Reenen (2007) (denoted by BSV for the rest of this section).

The authors of the above study collected firm level accounting data, such as sales, from the US Compustat database. This data was then matched to the NBER US Patent and Trademark Office data containing detailed information on granted US patents. This resulted in an unbalanced panel of 715 firms with observations recorded between 1980 and 2001.

BSV argue that R&D leads to two major externalities. One the one hand, R&D may increase the productivity of firms using similar technology. A firm can benefit from the R&D conducted by another firm in the same technology area. On the other hand, it can have a product market rivalry effect, which is detrimental to social welfare. Using the firm level information available, the authors map the location of each firm in both the technology and product space, by comparing information on patents and information on sales across firms.

In order to measure the distance between firms, BSV allocate the available patents into 425 different classes. The distance between firms is measured as the uncentered correlation between the allocation into patent classes of two distinct firms. The degree of technological spillover (SpillTech) is then measured as the technology distance weighted average of the R&D stock of all other firms.

BSV measure the distance between firms in the product market by decomposing each firm's sales by four digit industry code. Most firms are multiproduct firms and span a total of 762 different industries. The distance between firms in the product market is then measured as the uncentered correlation between the allocation of sales activity of firms into industries. The degree of product market spillovers (SpillSIC) is then computed as the product market distance weighted average of the R&D stock of all other firms.

In a simple model of R&D it is possible to derive a number of theoretical implications of these two spillover effects. In particular the marginal effect of technology spillovers on patent counts is positive while the marginal effect of product market spillovers on patent counts is zero. We can test these predictions in a binary dependent variable setting. First we can define the dependent variable to be one if the patent count for that firm is greater than zero. We can think of this case as an indicator of innovation for a given firm-year dyad. We can also construct an alternative definition of the dependent variable as being one if the patent count exceeds five. This indicator is designed to approximately capture firms with intense R&D activity.

We can now regress both of these indicators on the measures of technological and product market spillovers discussed above. It is reasonable to assume however that there may be unobserved firm level heterogeneity in innovation. Corporate culture between different companies is likely to encourage different degrees of intensity of the innovation activity in firms . Furthermore, the companies may be subject to systemic time effects across the US which will affect all companies reflecting technological trends. In order to account for some of the unobserved heterogeneity, we augment the specification above with two additional controls. One corresponds to firm sales, while the other corresponds to the pre-existing stock of R&D available within the firm. Furthermore, we lag all right hand side variables by one period so as to remove the possibility of contemporaneous effects.

We report the estimation results for the two patent models in Table 5. We estimate each model using the pooled probit model, the random effects probit model with time dummies and the flexible latent effects probit model introduced in this paper. Posterior mens are reported for the latter, obtained from chains of total length 10,000 MC steps with a 5,000 burn-in section. The pooled probit model appears to produce results which are inconsistent with economic theory, while both the random effects probit and the flexible latent effects probit model produce comparable coefficient estimates. Larger firms and those with a larger R&D stock are more likely to engage in innovation. Technology spillovers have a large, positive and statistically significant effect on the probability of patenting innovations. Product market spillovers by contrast only have a small and statistically negligible effect. These results are also consistent with the count model reported in BSV.

While both the random effects probit and the flexible latent effects probit models imply very similar parameter estimates, it is important to recall that the flexible latent effects probit model does not impose any assumption on the unobserved heterogeneity. The estimated distributions of heterogeneity are shown in Figures 1-2 for the two different models. Both distributions have strong multi-modal features. These may reflect the presence of missing variables important for characterizing innovation. Notice in particular that both models seem to imply the clustering of the unobserved effects into groups of firms. While most unobserved firm effects are clustered around zero, it also seems that there are firms with individual effects which are fairly large on either side of zero. The model of intense activity has a much greater dispersion of individual heterogeneity. Our model also indicates the existence of a time factor reflecting trends in technology. This factor is measured as having moderate persistence over time, with an autocorrelation coefficient of approximately 0.5.

Recall that even though the estimated coefficients are similar in the random coefficients probit model and in the new flexible latent effects probit model, other quantities of interest such as marginal effects may be different since they also depend on the estimate of the unobserved heterogeneity. In equation 3.1 it can be seen that the APE scale depends not only on the estimated β coefficients but also on the estimates of the latent individual and time variables and thus may lead to quantitatively different answers. A more precise estimate of the distribution of unobserved heterogeneity will thus improve our estimates of the marginal effects. Using the results in Table 5 for the model of a positive patent count, we find that the flexible latent effects probit model estimates a marginal effect for technology spillovers that is 20% smaller than the marginal effect estimated by the random effects probit model. Such differences in the estimated quantities of interest may produce very different policy recommendations.

A particularly attractive feature of our econometric method is its ability to explore the distribution of unobserved heterogeneity in greater detail than is possible in a classical random effects model. The distribution of unobserved heterogeneity becomes an object of interest in itself, rather than just a mysterious quantity that has to be removed from the data in order to obtain consistent parameter estimates on the observable variables.

As noted above, in our application, the distribution of unobserved heterogeneity has strong multi-modal features which may provide policy relevant conclusions on the unobserved drivers of firm level innovation. While a Normal approximation to this distribution may often be adequate if we are interested only in the estimated coefficients on the observable variables, it also misses the valuable additional information obtained by using a nonparametric estimate of the unobserved heterogeneity.

In order to explore the distribution of unobserved heterogeneity, τ_i, in our two patent models discussed above we need to make sure that its behavior is not implicitly restricted by the estimation procedure or some other deep model parameters. One parameter that is of concern to us is the α smoothing parameter that controls the extent to which the Dirichlet Process draws mixture distributions that are more or less "similar" to the baseline parametric distribution (Normal). Previous studies, such as Burda, Harding and Hausman (2008), find that that the estimated distribution is relatively insensitive to the choice of the α parameter and the statistical literature recommends a small number close to 1 as the optimal choice of α. In order to evaluate this claim we also implemented the sampling of the α parameter following the procedure described in Escobar and West (1995) as part of the estimation. The resulting distribution for the first patent model is plotted in Figure 3. It is concentrated around a mode of 2 thus giving further credence to the traditional choice of α as a small number.

For each model we can explore the evolution of the draws of τ_i parameters for specific companies in order to investigate if the draws are concentrated or show large variances as the Markov Chain progresses. Overall, we have found the draws to be remarkably stable indicating a clear tendency of the model to associate each firm with a narrow range of draws of τ_i. This is an important indicator that the clustering of τ_i values observed in the estimated distribution of unobserved heterogeneity may contain relevant information since it establishes a fairly tight link between firms and different modes of the distribution of heterogeneity. To exemplify we plot the draws of τ_i for the first five firms for the model of positive patent count in Figure 4. As we can see, all the draws lie in a narrow range associated with a specific cluster and the differences between the estimated unobserved heterogeneity across firms appear to be persistent and quite stable. If a draw of τ_i jumps to a different cluster, it does not stay there long and returns shortly back to its long term average. Very similar results were also obtained for the model of patent count > 5 and are not reproduced here.

It is now possible to construct a series of hypotheses about the nature of the unobserved heterogeneity and possible explanations of the nature of its clustering. One immediate such hypothesis is that there are unobserved industry level determinants of innovation which are responsible for the observed clustering.

In order to investigate this, we plot the average value of τ_i by each SIC code for the two estimated models of innovation in Figures 5 and 6. Unfortunately, the resulting graphs don't appear to allow for a particularly intuitive explanation of unobserved heterogeneity as driven by industry level factors as no clear patterns emerge across SIC codes. Moreover, we find large differences in company types even within SIC categories. It appears that associating the unobserved heterogeneity with industry categories may obscure important differences between firms as far as their innovation activity is concerned.

Nevertheless, it we can use these plots to identify clear outliers, firms with extreme levels of unobserved heterogeneity, both positive and negative, that is either promoting or prohibiting innovation. For some of these outlier we were able to extract a fairly stable industry pattern which appears to indicate that these firms are concentrated in certain industries. We capture some of the industries corresponding to these outliers in Table 6 by focusing our attention on industries with strong negative effects (defined as SIC codes with an average $\tau_i < -1.5$) and strong positive effects (defined as SIC codes with an average $\tau_i > 1.5$)and evaluate both models of patent activity.

The list of industries that are identified as containing outliers in the distribution of unobserved heterogeneity appears to make intuitive sense. Firms with a particularly strong negative unobserved effect are drawn from industries such as meat packing, retail stores, motion picture theaters and printing. Firms identified as having a strong positive unobserved effect are drawn by contrast from industries such as beverages, military equipment or household appliances.

In order to evaluate the extent to which industry level heterogeneity is responsible for the clustering observed in Figures 5 and 6 we re-estimate the models using industry dummies in addition to the variables introduced above. The nonparametric density estimates of the unobserved heterogeneity are almost identical to the ones previously discussed and are not included in the paper. The inclusion of industry dummies appears to have only a very small effect in controlling for unobserved heterogeneity.

We can therefore conclude that while industry level heterogeneity may explain some of the outliers (both positive and negative) in terms of R&D activity, it does not fully explain the complicated multimodal distribution of firm level unobserved heterogeneity estimated nonparametrically by our model. Substantial firm level heterogeneity persists even within SIC codes and additional variables would be required to capture and provide a detailed economic explanation of the heterogeneity measured by our econometric model.

5. Conclusion

This paper introduced a new Bayesian semi-parametric approach to the estimation of the probit model in panel data with unobserved heterogeneity. This new model substantially improves on current methods by relaxing three assumptions that are often either ignored or treated in an ad-hoc fashion in empirical work.

First, we allow for the presence of unobserved individual effects of unknown functional form. Second, we allow for the unobserved heterogeneity to be correlated with the observables. Finally, our model allows for the presence of dynamics through a latent autoregressive factor.

We employ a combination of recent sampling algorithms in order to draw from a Dirichlet Process Mixture. These sampling procedures are shown to be computationally efficient in a number of Monte Carlo simulations. The underlying parameters are shown to be estimated with high precision. When the underlying heterogeneity is not well approximated by a Normal distribution then the gains from using this new model are substantial relative to a traditional random effects probit approach.

We also apply the methods to the estimation of a patent equation in the presence of both technological and product market spillover effects. We show that technological innovation is subject to substantial firm level heterogeneity. Some of the outliers in the distribution of heterogeneity can be explained as being driven by heterogeneity at the level of specific industries. The distribution of firm level heterogeneity has however complex multimodal features and the clustering cannot fully be explained as occurring at industry level.

References

Abrevaya, J. (1999): "Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable," *Journal of Econometrics*, 93(2), 203–228.

Albert, J., and S. Chib (1993): "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88(422), 669–679.

Albert, J., and S. Chib (1996): "Bayesian modeling of binary repeated measures data with application to crossover trials," in *Bayesian Biostatistics*, ed. by D. A. Berry, and D. K. Stangl. Marcel Dekker, New York.

Arellano, M., and J. Hahn (2006): "A Likelihood-Based Approximate Solution To The Incidental Parameter Problem In Dynamic Nonlinear Models With Multiple Effects," Working paper, CEMFI.

Arelllano, M., and B. Honore (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics, Volume 5*, ed. by J. Heckman, and E. Leamer. Elsevier Science.

Berry, S., and P. Haile (2008): "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," Cowles Foundation.

Bloom, N., M. Schankerman, and J. Van Reenen (2007): "Identifying Technology Spillovers and Product Market Rivalry," NBER Working Paper 13060.

Burda, M., M. C. Harding, and J. A. Hausman (2008): "A Bayesian Mixed Logit-Probit Model for Multinomial Choice," *Journal of Econometrics*, 147(2), 232–246.

Burda, M., R. Liesenfeld, and J.-F. Richard (2008): "Bayesian Analysis of a Probit Panel Data Model with Unobserved Individual Heterogeneity and Autocorrelated Errors," Working paper.

Butler, J. S., and R. Moffitt (1982): "A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model," *Econometrica*, 50(3), 761–764.

Chamberlain, G. (1982): "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5–46.

Chib, S., and B. Carlin (1999): "On MCMC Sampling in Hierarchical Longitudinal Models," *Statistics and Computing*, 9, 17–26.

Chib, S., and I. Jeliazkov (2006): "Inference in Semiparametric Dynamic Models for Binary Longitudinal Data," *Journal of the American Statistical Association*, 101(474), 685–700.

Dahl, D. B. (2005): "Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models," Technical report, Texas A&M University.

Fernandez-Val, I. (2007): "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models," working paper, Boston University.

Geweke, J. (2005): *Contemporary Bayesian Econometrics and Statistics*. Wiley.

Hausman, J. A., and W. E. Taylor (1981): "Panel Data and Unobservable Individual Effects," *Econometrica*, 49(6), 1377–98.

Knittel, C. R., and K. Metaxoglou (2008): "Estimation of random coefficient demand models: challenges, difficulties and warnings," working paper series, NBER.

Liesenfeld, R., G. V. Moura, and J.-F. Richard (2008): "Dynamic Panel Probit Models for Current Account Reversals and their Efficient Estimation," Working paper.

Neal, R. (2000): "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9(2), 249–265.

Train, K. (2003): *Discrete Choice Methods with Simulation*. Cambridge University Press.

Wooldridge, J. M. (2001): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

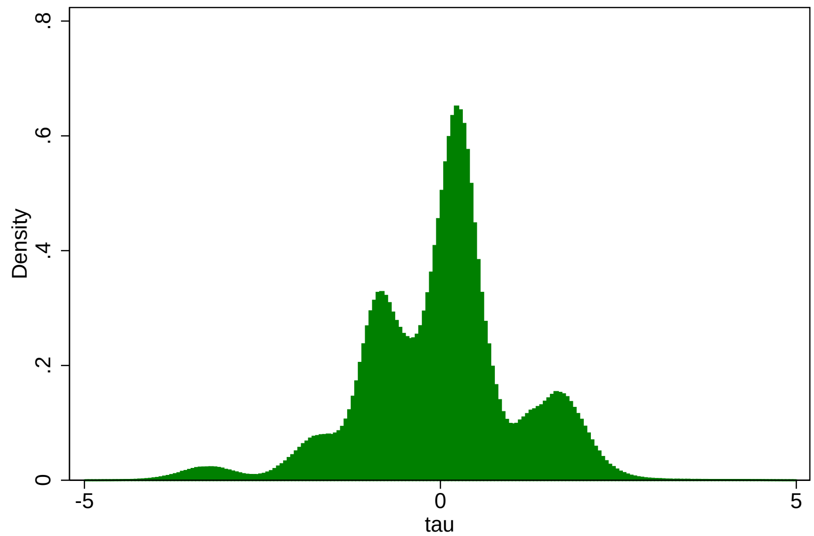

FIGURE 1. Distribution of unobserved heterogeneity. FLE Probit for Patent Count>0.

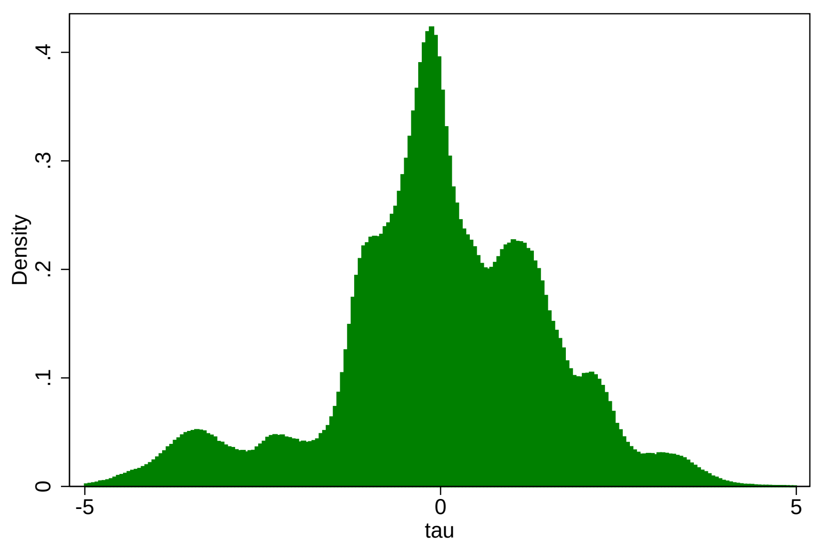

FIGURE 2. Distribution of unobserved heterogeneity. FLE Probit for Patent Count>5.

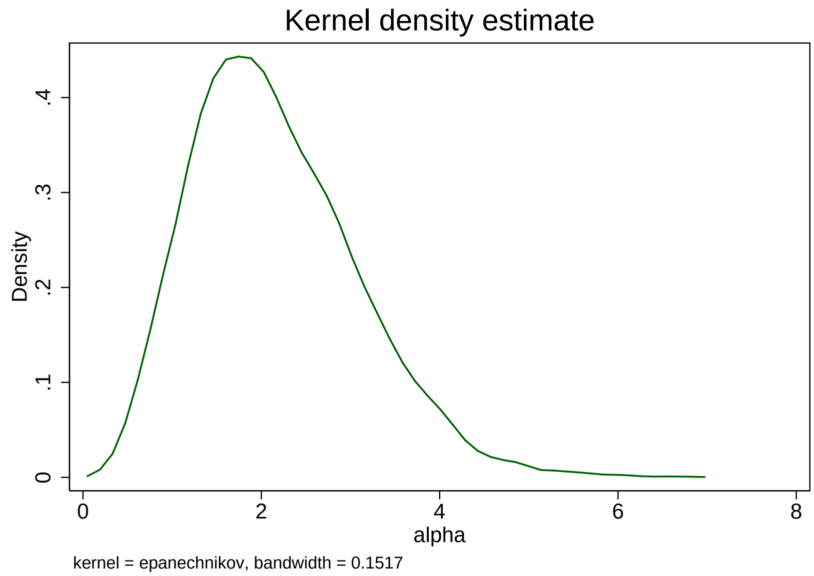

FIGURE 3. Density of draws of α. Model for Patent Count > 0.

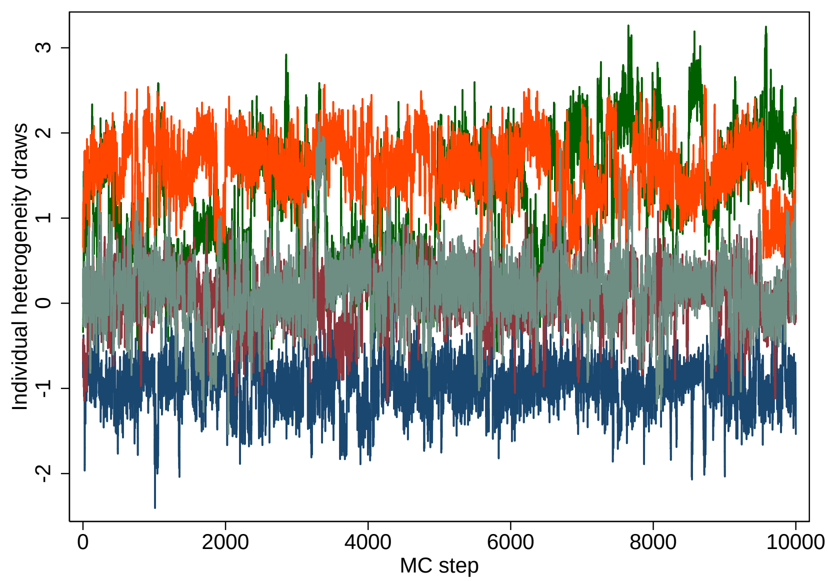

FIGURE 4. Draws of τ_i for the first five companies. Model for Patent Count > 0.

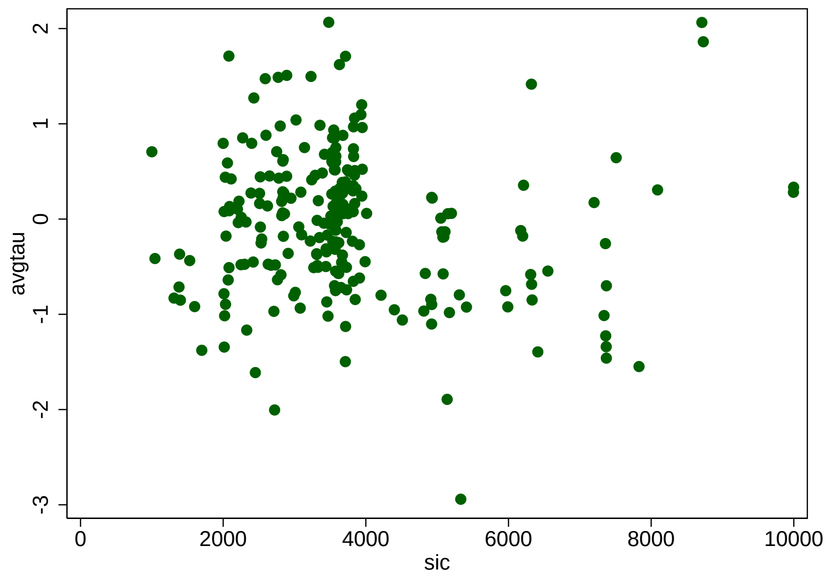

FIGURE 5. Average τ_i of companies for each SIC. Model for Patent Count > 0.

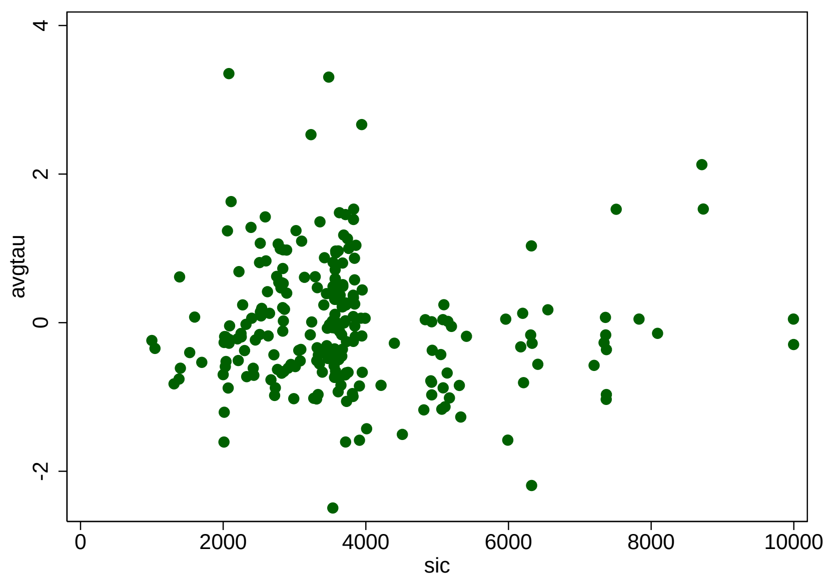

FIGURE 6. Average τ_i of companies for each SIC. Model for Patent Count > 5.

N	T	APE scale	β_0	β_1	β_2	β_3	σ_η	ρ
Pooled Probit								
100	10	0.289	-0.000	-0.752	-0.763	0.736	-	-
300	10	0.293	-0.000	-0.745	-0.751	0.756	-	-
700	10	0.294	0.001	-0.745	-0.750	0.749	-	-
1000	10	0.295	-0.000	-0.749	-0.748	0.745	-	-
100	20	0.286	0.001	-0.736	-0.745	0.734	-	-
300	20	0.293	-0.001	-0.743	-0.746	0.747	-	-
700	20	0.295	0.001	-0.753	-0.750	0.751	-	-
1000	20	0.295	0.000	-0.749	-0.747	0.751	-	-
100	50	0.290	-0.001	-0.757	-0.747	0.743	-	-
300	50	0.293	0.001	-0.743	-0.748	0.745	-	-
700	50	0.294	0.000	-0.745	-0.748	0.745	-	-
1000	50	0.295	-0.000	-0.746	-0.748	0.748	-	-
Random Effects Probit with Serial Correlation								
100	10	0.183	-0.205	0.115	0.055	-0.133	-	-
300	10	0.200	0.055	0.086	0.050	-0.045	-	-
700	10	0.200	-0.081	0.065	0.055	-0.050	-	-
1000	10	0.203	-0.018	0.069	0.063	-0.086	-	-
100	20	0.194	0.173	0.099	0.049	-0.084	-	-
300	20	0.203	0.000	0.040	0.025	-0.017	-	-
700	20	0.207	0.105	0.021	0.034	-0.028	-	-
1000	20	0.208	-0.052	0.027	0.030	-0.026	-	-
100	50	0.200	-0.022	0.018	0.049	-0.058	-	-
300	50	0.205	-0.019	0.025	0.014	-0.021	-	-
700	50	0.209	-0.235	0.019	0.004	-0.015	-	-
1000	50	0.209	-0.117	0.021	0.012	-0.010	-	-
Flexible Latent Effects Probit								
100	10	0.001	0.004	0.025	-0.046	-0.060	0.114	-0.309
300	10	0.001	0.014	0.024	-0.015	0.031	0.067	-0.231
700	10	0.001	0.010	-0.001	-0.016	0.023	0.074	-0.223
1000	10	0.000	-0.007	0.007	0.013	-0.017	0.102	-0.263
100	20	-0.003	0.016	0.084	0.025	-0.070	0.162	-0.185
300	20	-0.000	-0.003	0.021	0.002	0.003	0.104	-0.089
700	20	-0.001	0.002	-0.007	0.006	0.001	0.065	-0.096
1000	20	0.001	-0.001	-0.010	-0.003	0.007	0.034	-0.062
100	50	0.001	0.012	-0.054	0.108	-0.150	0.110	-0.018
300	50	-0.001	0.001	0.022	0.010	-0.016	0.062	-0.075
700	50	-0.000	0.000	0.011	-0.004	-0.008	0.030	-0.049
1000	50	-0.000	0.000	0.011	0.001	0.001	0.029	-0.030

TABLE 1. Mean Absolute Deviations, X and τ_i uncorrelated

N	T	APE scale	β_0	β_1	β_2	β_3	σ_η	ρ
			Pooled Probit					
100	10	0.289	0.029	0.755	0.768	0.739	-	-
300	10	0.293	0.013	0.746	0.752	0.757	-	-
700	10	0.294	0.009	0.745	0.751	0.750	-	-
1000	10	0.295	0.007	0.749	0.748	0.745	-	-
100	20	0.286	0.030	0.738	0.746	0.736	-	-
300	20	0.293	0.009	0.744	0.746	0.748	-	-
700	20	0.295	0.006	0.753	0.751	0.751	-	-
1000	20	0.295	0.004	0.749	0.747	0.751	-	-
100	50	0.290	0.022	0.757	0.748	0.744	-	-
300	50	0.293	0.008	0.743	0.748	0.745	-	-
700	50	0.294	0.006	0.745	0.748	0.745	-	-
1000	50	0.295	0.004	0.746	0.748	0.748	-	-
			Random Effects Probit with Serial Correlation					
100	10	0.187	1.159	0.365	0.357	0.276	-	-
300	10	0.202	1.097	0.157	0.170	0.164	-	-
700	10	0.202	1.114	0.120	0.113	0.105	-	-
1000	10	0.204	0.905	0.100	0.098	0.110	-	-
100	20	0.195	1.157	0.200	0.159	0.214	-	-
300	20	0.204	0.907	0.106	0.107	0.109	-	-
700	20	0.208	0.931	0.076	0.084	0.067	-	-
1000	20	0.209	1.002	0.066	0.056	0.055	-	-
100	50	0.201	0.877	0.111	0.112	0.113	-	-
300	50	0.206	1.205	0.071	0.052	0.061	-	-
700	50	0.209	0.943	0.046	0.039	0.040	-	-
1000	50	0.209	0.970	0.044	0.038	0.031	-	-
			Flexible Latent Effects Probit					
100	10	0.013	0.163	0.370	0.386	0.334	-0.010	0.581
300	10	0.008	0.079	0.191	0.193	0.196	-0.006	0.540
700	10	0.005	0.055	0.121	0.127	0.126	-0.005	0.526
1000	10	0.004	0.048	0.101	0.106	0.104	-0.005	0.554
100	20	0.010	0.127	0.257	0.229	0.271	0.046	0.382
300	20	0.005	0.059	0.137	0.143	0.139	-0.003	0.328
700	20	0.004	0.041	0.092	0.097	0.085	-0.002	0.343
1000	20	0.006	0.031	0.090	0.077	0.083	-0.002	0.332
100	50	0.006	0.044	0.123	0.159	0.212	0.004	0.185
300	50	0.003	0.034	0.090	0.078	0.083	-0.001	0.214
700	50	0.002	0.024	0.058	0.056	0.055	-0.001	0.198
1000	50	0.003	0.020	0.054	0.053	0.049	-0.000	0.203

TABLE 2. Root Mean Square Errors, X and τ_i uncorrelated

N	T	APE scale	β_0	β_1	β_2	β_3	σ_η	ρ
			Pooled Probit					
100	10	0.273	-0.036	-0.380	-0.381	-0.385	-	-
300	10	0.274	-0.025	-0.389	-0.389	-0.381	-	-
700	10	0.275	-0.028	-0.388	-0.388	-0.387	-	-
1000	10	0.274	-0.026	-0.390	-0.393	-0.387	-	-
100	20	0.271	-0.026	-0.384	-0.380	-0.378	-	-
300	20	0.274	-0.025	-0.385	-0.387	-0.384	-	-
700	20	0.275	-0.027	-0.385	-0.387	-0.385	-	-
1000	20	0.275	-0.026	-0.387	-0.386	-0.387	-	-
100	50	0.271	-0.025	-0.381	-0.383	-0.385	-	-
300	50	0.273	-0.026	-0.384	-0.386	-0.386	-	-
700	50	0.275	-0.026	-0.386	-0.386	-0.384	-	-
1000	50	0.275	-0.027	-0.385	-0.385	-0.386	-	-
			Random Effects Probit with Serial Correlation					
100	10	0.121	-0.060	0.065	0.087	0.037	-	-
300	10	0.148	-0.046	0.007	-0.001	0.022	-	-
700	10	0.147	-0.173	0.013	0.016	-0.001	-	-
1000	10	0.149	-0.349	0.008	-0.010	0.019	-	-
100	20	0.155	-0.116	0.028	0.032	0.034	-	-
300	20	0.180	-0.326	-0.002	0.004	-0.003	-	-
700	20	0.189	-0.138	-0.004	-0.010	-0.008	-	-
1000	20	0.190	-0.600	-0.013	-0.010	-0.008	-	-
100	50	0.152	-0.026	0.037	0.022	0.022	-	-
300	50	0.171	-0.004	0.013	0.003	0.005	-	-
700	50	0.179	-0.127	0.004	0.001	0.011	-	-
1000	50	0.178	-0.194	0.003	0.004	0.003	-	-
			Flexible Latent Effects Probit					
100	10	-0.001	-0.005	-0.034	0.049	0.013	0.063	-0.246
300	10	0.000	-0.001	-0.030	-0.011	0.018	-0.038	-0.230
700	10	-0.001	0.004	-0.003	0.015	-0.003	-0.008	-0.224
1000	10	-0.001	0.005	-0.004	-0.016	0.012	-0.049	-0.270
100	20	-0.001	0.004	-0.015	0.037	0.040	0.118	-0.166
300	20	-0.001	0.002	-0.021	0.016	0.009	0.025	-0.063
700	20	0.000	0.006	-0.016	-0.000	0.001	0.019	-0.064
1000	20	0.001	0.006	-0.011	-0.001	0.001	-0.020	-0.091
100	50	-0.003	0.003	0.021	0.028	0.028	0.148	-0.113
300	50	-0.001	0.005	0.006	0.005	0.007	0.033	-0.069
700	50	-0.001	0.004	0.003	0.002	0.010	-0.009	-0.011
1000	50	-0.000	0.002	0.002	0.004	0.002	-0.006	-0.030

TABLE 3. Mean Absolute Deviations, X and τ_i correlated

N	T	APE scale	β_0	β_1	β_2	β_3	σ_η	ρ
			Pooled Probit					
100	10	0.273	0.044	0.382	0.383	0.388	-	-
300	10	0.274	0.029	0.389	0.389	0.382	-	-
700	10	0.275	0.029	0.388	0.388	0.387	-	-
1000	10	0.274	0.028	0.390	0.393	0.388	-	-
100	20	0.271	0.034	0.385	0.381	0.380	-	-
300	20	0.274	0.027	0.386	0.387	0.384	-	-
700	20	0.275	0.028	0.385	0.387	0.386	-	-
1000	20	0.275	0.026	0.387	0.386	0.387	-	-
100	50	0.271	0.031	0.381	0.384	0.385	-	-
300	50	0.273	0.027	0.384	0.386	0.387	-	-
700	50	0.275	0.026	0.386	0.386	0.384	-	-
1000	50	0.275	0.027	0.385	0.385	0.386	-	-
			Random Effects Probit with Serial Correlation					
100	10	0.124	1.973	0.182	0.203	0.184	-	-
300	10	0.150	1.639	0.094	0.090	0.091	-	-
700	10	0.149	1.487	0.064	0.073	0.051	-	-
1000	10	0.150	1.621	0.047	0.049	0.072	-	-
100	20	0.158	1.679	0.093	0.122	0.117	-	-
300	20	0.181	1.897	0.064	0.064	0.059	-	-
700	20	0.190	1.385	0.039	0.048	0.041	-	-
1000	20	0.191	1.282	0.041	0.039	0.038	-	-
100	50	0.154	1.638	0.083	0.065	0.064	-	-
300	50	0.171	1.463	0.037	0.037	0.035	-	-
700	50	0.180	1.655	0.027	0.026	0.028	-	-
1000	50	0.179	1.399	0.020	0.024	0.020	-	-
			Flexible Latent Effects Probit					
100	10	0.012	0.147	0.125	0.226	0.215	-0.011	0.543
300	10	0.008	0.084	0.123	0.124	0.126	-0.005	0.514
700	10	0.006	0.056	0.061	0.085	0.075	-0.005	0.528
1000	10	0.006	0.045	0.059	0.067	0.080	-0.004	0.544
100	20	0.009	0.106	0.119	0.161	0.162	-0.005	0.386
300	20	0.006	0.059	0.129	0.088	0.090	-0.003	0.328
700	20	0.005	0.036	0.125	0.062	0.058	-0.002	0.321
1000	20	0.006	0.031	0.059	0.054	0.056	-0.002	0.315
100	50	0.006	0.073	0.071	0.098	0.096	0.017	0.227
300	50	0.004	0.036	0.040	0.054	0.055	-0.001	0.213
700	50	0.002	0.025	0.027	0.036	0.036	-0.001	0.191
1000	50	0.002	0.022	0.022	0.031	0.029	-0.001	0.185

TABLE 4. Root Mean Square Errors, X and τ_i correlated

	Patent Count > 0			Patent Count > 5		
	probit	RE probit	FLE	probit	RE probit	FLE
Ln(SpillTech)	-0.376	0.389	0.364	-0.225	0.535	0.553
	(0.551)*	(0.045)*	(0.046)*	(0.055)*	(0.089)*	(0.095)*
Ln(SpillSIC)	-0.116	0.023	0.037	-0.068	0.075	0.072
	(0.227)*	(0.021)	(0.019)	(0.025)*	(0.035)*	(0.040)
Ln(R&D Stock)	0.265	0.303	0.313	0.443	0.482	0.517
	(0.242)*	(0.027)*	(0.027)*	(0.033)*	(0.047)*	(0.052)*
Ln(Sales)	-0.236	0.202	0.204	-0.041	0.386	0.356
	(0.225)*	(0.025)*	(0.022)*	(0.269)	(0.043)*	(0.049)*
APE scale	0.339*	0.203	0.170	0.168	0.100	0.096
ρ			0.645			0.722
			(0.145)*			(0.118)*

Table 5. Estimation of Patent Equation

Notes: Estimation results are reported for the pooled probit model, the random effects probit and the flexible latent effects probit. Standard errors are reported in brackets. All independent variables are lagged by one period. All regressions include a constant and a dummy for observations where the lagged RD stock is zero.

Model 1: Patent Count > 0

Strong Negative Effects

2451 - Mobile Homes
2721 - Periodicals Publishing Printing
5140 - Wholesale: Groceries And Reld Products
5331 - Variety Stores
7830 - Motion Picture Theaters

Strong Positive Effects

2080 - Beverages
2891 - Adhesives Sealants
3480 - Military Equipment: Ordnance And Accessories, Except Vehicles And Guided Missiles
3630 - Household Appliances
3715 - Transportation Equipment: Truck Trailers
8711 - Engineering Services
8731 - Commercial Physical/Biological Research

Model 2: Patent Count > 5

Strong Negative Effects

2011 - Meat Packing Plants
3537 - Industrial Trucks Tractors Trailers
3716 - Motor Homes
3911 - Jewelry-Precious Metal
4512 - Air Transportation-Scheduled
5990 - Retail Stores, Not Elsewhere Classified
6324 - Insurance carriers: Hospital Medical Service Plans

Strong Positive Effects

2080 - Beverages
2111 - Cigarettes
3231 - Glass Products Made Of Purchased Glass
3480 - Military equipment: Ordnance And Accessories, Except Vehicles And Guided Missiles
3829 - Measuring Controlling Devices, Nec
3942 - Dolls Stuffed Toys
7510 - Automotive Rental And Leasing
8711 - Engineering Services
8731 - Commercial Physical/Biological Research

TABLE 6. Industries identified as having extreme individual effects on firm innovation.