Chapter 2

Semiparametric Bayesian Models for Dynamic Earnings Data

2.1 Introduction: Full Probability Models for Earnings Dynamics

This paper develops new models for dynamic panel data and applies them to study longitudinal data on earnings from the Panel Study of Income Dynamics (PSID). We propose Bayesian semiparametric versions of commonly used random effects autoregressive models and error components models. To model the unknown distributions without having to make strong parametric assumptions, we draw heavily on recent advances in the theory and computation of nonparametric Bayesian models using Dirichlet process priors.

One important application of these earnings dynamics models is to optimal consumption problems. Suppose for example that we are interested in solving the following stylized problem, based on work by Deaton (1991). At time T, the agent chooses a contingent strategy for consumption $(c_T, c_{T+1}, \ldots)$ to maximize expected utility

$$E\left[\sum_{t=T}^{\infty} \beta^{t-T} u(c_t)\right]$$

subject to the budget constraint

$$w_t = (1+r)(w_{t-1} - c_{t-1}) + Y_t,$$

and a no–borrowing constraint

$$w_t - c_t \geq 0.$$

Here w_t denotes the agent's wealth at time t, after receiving labor income Y_t. To close the model one needs to specify the initial wealth w_T, the discount factor β and the interest rate r, and specify a joint probability distribution for earnings $(Y_T, Y_{T+1}, \ldots)$. One could assume that the logarithm of earnings $y_t \equiv \log Y_t$ follows a simple autoregression:

$$y_t = \gamma + \rho y_{t-1} + \epsilon_t. \tag{2.1}$$

Alternatively, an error components model might be used:

$$y_t = v_t + \epsilon_t, \tag{2.2}$$

where the ϵ_t are independent over time, and v_t follows a low-order autoregressive process, possibly a random walk. In general the optimal consumption policy will depend on the income process.

It is usually not reasonable to suppose that such complex structural models are literally true descriptions of the decision-making of individuals, but one may ask what an optimum based on available information should look like, and whether individuals appear to be acting nearly optimally. Consumption models with similar features have been proposed and studied in Deaton (1991), Carroll (1997), Engen and Gruber (1995), Hubbard, Skinner, and Zeldes (1994), Gourichas and Parker (1996), and Viceira (1997). This recent work emphasizes how a precautionary motive for saving (convex marginal utility of money), impatience, and limits on borrowing can generate "buffer-stock" savings, in which it is optimal to keep a small amount of savings to guard against low income realizations.

These papers have had to make strong parametric assumptions in order to obtain complete probability distributions for the earnings process. For example, in the autoregressive model, classical semiparametric methods could be used to estimate γ, ρ, and the variance of ϵ_t from longitudinal data on labor earnings, but conventional methods (e.g. MaCurdy (1982), Chamberlain (1984), and Holtz-Eakin, Newey, and Rosen (1988)) do not provide estimates of the distribution of ϵ_t. The buffer-stock models therefore assume normality for any unknown distributions, largely erasing the potential benefits of semiparametric inference.

Our proposed semiparametric Bayesian models deliver inference for the entire earnings process, while allowing us to remove the restriction that all components of the model belong to known parametric families. There turns out to be strong evidence against normality in the earnings data, suggesting that consumption models which rely on this assumption should be viewed with caution. Another important feature of this approach is that uncertainty about parameters is carefully accounted for at every stage in the analysis, and we report *distributions* on unknown parameters rather than point estimates. The approach developed below has a modular nature, and can be readily extended to even richer models, the primary cost being computational difficulty rather than analytic intractability or concerns about the adequacy of asymptotic approximations.

Two key difficulties encountered in nonparametric and semiparametric Bayesian analysis are the need to specify prior distributions appropriately, and the complexity of some of the necessary calculations. In our application, we will see that the first issue corresponds very closely to the problem of choosing "smoothing parameters" in classical nonparametric and semiparametric methods. We regard the device of a prior distribution as a rich, flexible way to make such a priori judgments of smoothness, and will examine this issue in detail when we study the earnings data. While the computations remain difficult, they are now feasible by taking advantage of Markov chain Monte Carlo (MCMC) integration methods.

The semiparametric panel data estimator of Horowitz and Markatou (1996) also estimates underlying distributions nonparametrically, and in this sense is complementary to the approach we will develop below. Their method requires the data analyst to make judgments about smoothness in choosing bandwidth parameters in a generally ad hoc manner, as conventional optimal smoothing results do not apply in this context. The approach taken here allows the amount of smoothing to depend partly on the data, and also has the advantage that one can be fairly explicit about the assumptions about the shape of the density are being imposed, even before seeing the data. In addition, we are able to provide interval estimates and various other measures of uncertainty about any relevant quantity, and incorporate parameter uncertainty into predictive distributions. For the relatively small sample sizes in the data examined here, relying exclusively on point estimates would be ill-advised.

Semiparametric Bayesian dynamic panel data models should also be useful in other contexts besides the consumption problem outlined above. Much of the econometric work on longitudinal earnings data has focused on forecasting transitions in and out of poverty (see for example Lillard and Willis (1978), Horowitz and Markatou (1996), and Geweke and Keane (1997)); our Bayesian methods provide predictive distributions which are naturally suited to answering such questions. Recent work by Moffitt and Gottschalk (1995) examines how income dynamics have changed over time in the U.S.; these studies use classical minimum distance methods to decompose the growth in earnings variance into permanent and transitory components. Difficulties with minimum distance, especially in the choice of the weighting matrix, are reported by Abowd and Card (1989) and Altonji and Segal (1996), so alternative methods may be useful. Other economic models of choice under uncertainty, such as portfolio choice and dynamic labor supply models, could be examined with this methodology. In addition, predictive distributions can be useful as normative tools for policy decisions.

An alternative way to realize many of the advantages of the semiparametric Bayesian approach is to use parametric but highly complex models. Recent work on longitudinal data in this vein includes Geweke and Keane (1997), Carlin (1996), Chamberlain and Hirano (1997), and Rossi, McCulloch, and Allenby (1995); in particular Geweke and Keane use a finite mixture model which is similar to the infinite mixture model used here.[1] The infinite mixture approach is a useful way to avoid having to specify the number of mixture components in advance, and we will see that it can do a good deal of smoothing if the data warrant it. More generally, models which do not need to make very strong distributional assumptions should be useful in many economic applications, where a priori reasoning typically provides little information on the form of underlying distributions.

[1] The earnings model developed by Geweke and Keane (1997) incorporates many features of the PSID data, such as the dynamics of marital status, which are not addressed here.

2.2 A First Look at the Data, using a Bayesian Density Model

2.2.1 PSID Data

Our data is drawn from the PSID and records annual labor earnings for males who were heads of households between the ages of 24 and 33, during the period from 1967 to 1991. Only males with complete data on race, education, and earnings, who were heads of household with positive earnings at least one year prior to age 24, and with positive earnings for the ten consecutive years, were included. This resulted in 10 years of earnings data for 516 individuals. The construction of this sample is similar to the "young men" sample of Geweke and Keane (1996), and is designed to catch individuals early in their careers when uncertainty about future earnings might be rather high. Note that our "time" dimension corresponds to age: $t = 1$ means the year the individual was 24, and this can correspond to any calendar year betwen 1967 and 1982 (since an individual who was 24 in 1982 would be 33 in 1991, when our sample ends). However, we do retain and use calendar year information in the "first-stage" regressions described next.

Define z_{it} to be the logarithm of earnings for individual $i = 1, \ldots, n$ in year $t = 1, \ldots, T$, in 1991 dollars based on the CPI. We regress z_{it} on a constant, an indicator for race, indicators for education equal to some high school, high school, some college, college, and beyond college, and calendar year. Separate regressions are run for each age t, and the residuals from the least–squares regressions are kept and used as our "data." We denote these residuals by y_{it}. Working with residuals simplifies the analysis and lets us focus attention on earnings dynamics, but it would be possible to directly incorporate the covariates in the models developed below. Extensions of the earnings dynamics models to incorporate the covariates are discussed in the appendix, in section 2.6.4.

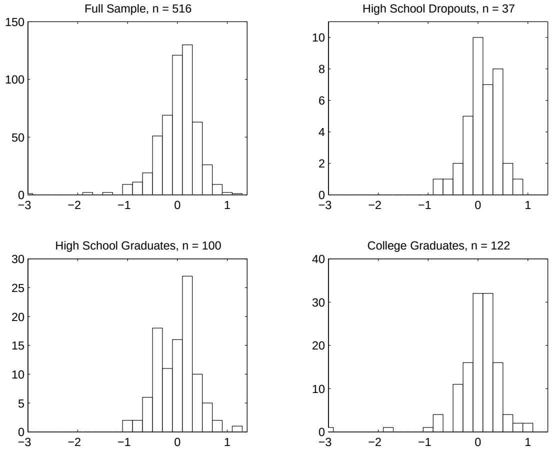
Before considering earnings dynamics, we begin with the simpler problem of characterizing the distribution of the earnings residuals at age 24 (y_{i1} in our notation). We work with the full sample ($n = 516$), but also draw three subsamples by selecting only white heads of households that were not part of the Survey of Economic Opportunity (SEO) sample and considering separately high school dropouts, high school graduates who did not receive further education, and college graduates who did not receive further education. There are 37 high school dropouts, 100 high school graduates, and 122 college graduates who meet this further sample restriction. Figure 2.1 displays histograms of earnings residuals at age 24 for the full sample and the three subsamples.[2] In the full sample and the college graduates subsample, there is some visual evidence of skewness and heavy tails relative to normality. In the high school graduates subsample there is some weak evidence of multimodality which we will examine in more detail below.

2.2.2 A Nonparametric Bayesian Model for Random Densities

Our model for the distribution of earnings follows Lo (1984) and Ferguson (1983); to deal with computational issues we rely heavily on recent work by Escobar (1994), Escobar and West (1995), and West, Müller, and Escobar (1994). In addition to being of interest in its own right, the model for densities reviewed and applied here will form a critical component

[2] Of course, a histogram is itself a nonparametric density estimate, in which the width of the bins is the critical smoothing parameter.

Figure 2.1: Earnings Residuals, Age 24



in the semiparametric models for earnings dynamics developed below.

The log earnings residuals at age 24, $y_{i1}, \ldots, y_{n1}$, are assumed to have an unknown probability density function q. Let $\phi(\cdot|\mu, \sigma^2)$ denote the density function of a normal random variable with mean μ and variance σ^2. Ferguson (1983) notes that any probability density function can be approximated by a countable mixture of normal densities[3]:

$$q(\cdot) = \sum_{j=1}^{\infty} p_j \phi(\cdot|\mu_j, \sigma_j^2), \tag{2.3}$$

where $\sum_{j=1}^{\infty} p_j = 1$.

Since this model has an infinite-dimensional parameter $\theta \equiv \{(p_j, \mu_j, \sigma_j^2)_{j=1}^{\infty}\}$, more structure is needed for this approach to give sensible results with a finite amount of data. One possibility is to put a prior distribution $p(\theta)$ on the unknown parameter, regard equation 2.3 as specifying a conditional density for y_i given θ—a likelihood function—and use the laws of conditional probability to calculate $P(\theta|y_1, \ldots, y_n)$, the posterior

[3] More precisely, the set of countable normal mixtures is dense in L^1.

distribution for θ given the data. One could use fairly sharp restrictions: for example, the restriction that $p_1 = 1$ results in the normal model; letting the p_j be arbitrary but restricting μ_j to be 0 for all j limits attention to symmetric densities. But it seems difficult to know in advance whether such restrictions are reasonable. Instead we try to tailor our prior distribution to reflect a judgment that the density should be "smooth." Start with the normal location and scale parameters, and use the conjugate prior:

$$\frac{1}{\sigma_j^2} \sim \frac{\chi^2(s)}{sQ}; \qquad \mu_j | \sigma_j^2 \sim \mathcal{N}(m, b \cdot \sigma_j^2),$$

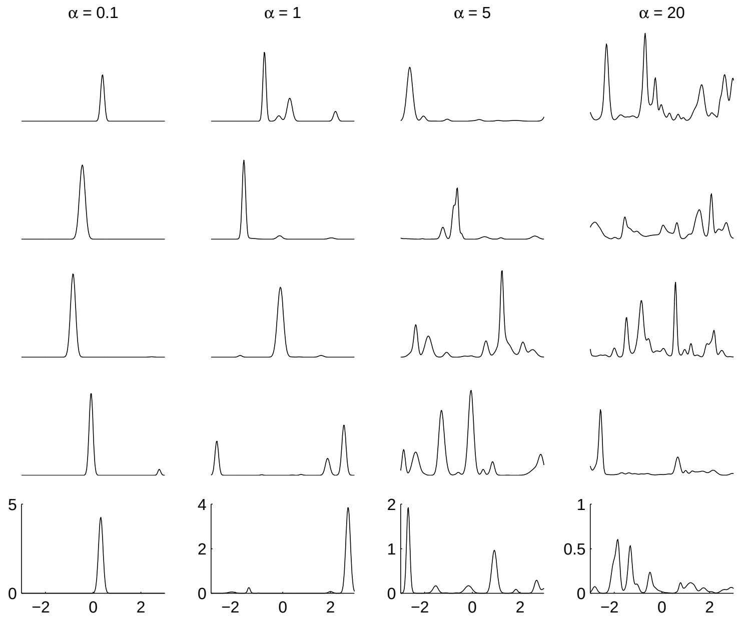
independently across the j. Q can be regarded as a prior guess for the variance term σ_j^2, with larger values of s making the distribution relatively tighter around Q. (We could allow s to be noninteger by using the fact that the scaled chi-square distribution is a special case of the gamma distribution.) Similarly, m is a prior guess for the component means μ_j, with b controlling whether the prior is relatively tight or diffuse.

Next consider the weights $(p_1, p_2, \ldots)$. We will put a distribution on this sequence by a "stick-breaking" construction. Let $r_1, r_2, \ldots$ be a sequence of i.i.d. random variables with a Beta$(1, \alpha)$ distribution. (Recall that the Beta(b_1, b_2) distribution is supported on the unit interval, and has mean $b_1/(b_1 + b_2)$.) Form $p_1 = r_1$ and for $j = 2, \ldots$ set

$$p_j = r_j \prod_{l=1}^{j-1} (1 - r_l).$$

The weights will be random, but will satisfy $\sum_{j=1}^{\infty} p_j = 1$ almost surely. A physical analogy is to start with a stick of length 1, break off a piece of length $p_1 = r_1$, then break off another piece from the remaining portion by taking r_2 times its length, and so on. If α is close to zero, so that the Beta$(1, \alpha)$ distribution puts most of its mass near 1, the

Figure 2.2: Draws from the Prior for $q(\cdot)$



process will tend to have p_1 large and the remaining probabilities small, whereas if α is large, many components will tend to have similar weights.

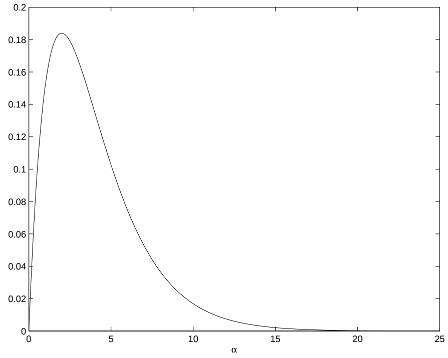
To examine the implications of this prior in more detail, we set $s = 3$, $Q = .01$, $m = 0$, and $b = 10/Q$, and we generate (approximate) draws for the mixture density $q(\cdot)$ for a range of values for α.[4] We have chosen a very diffuse prior for the location parameters μ_j, which leads the prior draws to have components whose centers are spaced relatively far apart; however, we will see below that the corresponding posteriors tend to be very localized around the support of the data, so this defect of the prior is arguably minor in practice. The prior for the variance component is centered at .01 (recall that we are modeling log earnings residuals), but setting $s = 3$ implies a relatively diffuse prior around Q.

[4] This procedure is described in the appendix, section 2.6.

Five draws for $q(\cdot)$ for each of four values for α are plotted in Figure 2.2; the plots have been truncated to lie within $[-3, 3]$. Setting $\alpha = 0.1$ in the first column leads to a distribution for q that puts most of its weight on essentially one- or two-component representations (at least within the range of the plots). For intermediate values for α, a range of multimodal and asymmetric densities can be accomodated, but setting $\alpha = 20$ in the last column leads to extremely "bumpy" densities which do not seem to accord with our intuition that the density of earnings residuals should be fairly smooth. Clearly, in conjunction with choices for s, Q, m, and b, the specification of α will play a major role in imposing prior smoothness on the unknown density. This raises the possibility that we could place a prior on α rather than regard it as fixed, to permit the posterior to adapt to the degree of smoothness in the data.[5] Our prior has α distributed as $\mathcal{G}(2, .5)$, a gamma distribution with mean $2/.5$ and variance $2/(.5^2)$. Its density is plotted in Figure 2.3. This places low weight on excessively jittery densities of the kind seen in the last column of Figure 2.2, but does allow a wide range of smoothness so that we can examine how informative the data can be about this issue. In addition, it is desirable to permit relatively high values of α for some purposes; for example, one way that the posterior can capture heavy-tailed distributions is to use a large number of overlapping components in the tails. Thus, the prior for α was chosen to incorporate some notion of smoothness, without ruling out interesting and potentially relevant densities.

[5] The other hyperparameters, especially s and Q, also play an important role in generating smooth densities, and these too could be given prior distributions. The direct role of α in controlling the component weights makes it somewhat easier to examine the departure from a one-component normal representation, so in this sense it is more convenient to focus on α rather than s or Q.

Figure 2.3: Gamma(2,.5) Prior for α



2.2.3 The Dirichlet Process

There is an equivalent representation of this model that will be useful. We suppose that each y_i is drawn from a normal distribution with mean μ_i and variance σ_i^2:

$$y_i|\mu_i, \sigma_i^2 \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

Define $\theta_i = (\mu_i, \sigma_i^2)$. Suppose these are themselves drawn from a discrete distribution P:

$$\theta_i|P \sim P.$$

It can be shown that the density model outlined earlier is equivalent to this hierarchical model if the prior distribution on P is a Dirichlet process (Ferguson (1973)), with base measure αP_0, where α is as defined above and P_0 is a probability measure for θ that specifies that

$$\frac{1}{\sigma^2} \sim \frac{\chi^2(s)}{sQ}; \qquad \mu|\sigma^2 \sim \mathcal{N}(m, b \cdot \sigma^2).$$

We will use the notation $P \sim \mathcal{D}(\alpha P_0)$ to indicate this. The Dirichlet process is described in further detail in the appendix, section 2.6, as it is useful for many of the computations.

Since we are regarding α as random with a gamma distribution, our complete prior can be viewed as a mixture of Dirichlet processes (Antoniak (1974)). The equivalence of these two representations is a consequence of the Sethuraman–Tiwari construction of the Dirichlet process (Sethuraman and Tiwari (1982), Sethuraman (1994)).

This representation in terms of the latent variables θ_i is similar in structure to duration models studied by Heckman and Singer (1984), who call the latent variables "unobserved heterogeneity," and focus on estimating the distribution of the latent variables (P in our notation) via maximum likelihood. Here, however, there is no compelling economic reason to give the θ_i a special interpretation, so they should be viewed solely as a modeling device.

To calculate posterior densities one typically has to perform an integration over the parameter space, either analytically or numerically. For this model, a useful analytic result is not available; on the other hand, numerically integrating over an infinite-dimensional space is infeasible. While various approximation methods have been developed, we use a method due to Escobar (1994) and Escobar and West (1995), further refined in West, Müller, and Escobar (1994). This approach introduces the θ_i as latent variables, conditional on which one can analytically integrate out the unknown P. Thus posterior inference only requires a numeric integration over a finite–dimensional space (here, the dimension is $2n+1$), which can be carried out via MCMC. It turns out that the West, Müller, and Escobar method can be readily extended to the semiparametric models we use below to study earnings dynamics. (Further computational details are given in the appendix, section 2.6.)

2.2.4 Posterior Inference for the Earnings Data

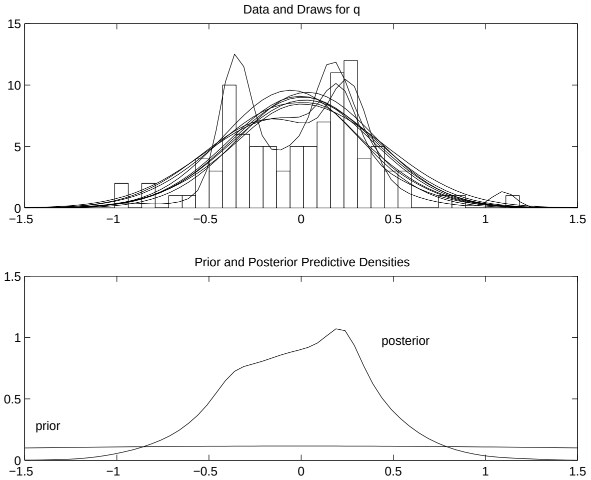
Figure 2.4 displays posterior inference for the high school graduates subsample; corresponding figures for the other subsamples are given in the appendix, section 2.7. The top panel gives a histogram of the data with a sample of draws for the countable mixture density q (for ease of viewing only 10 draws for q are displayed, even though many more were generated and used for subsequent inference).

The bottom panel gives *predictive densities* for a new data point. We extend the model of equation 2.3 to a hypothetical new individual $y_{n+1,1}$. Its density conditional on θ would be q; however, since θ is not known, we wish to provide a density that is unconditional on θ. If θ is given a distribution, say $P(\theta)$, then this can be done by integrating with respect to that distribution:

$$p(y_{n+1}) = \int q(y_{n+1}|\theta)dP(\theta).$$

Plotted in the bottom panel are two predictive densities: one curve is the density integrating over the posterior for θ, and the other is the corresponding predictive distribution based only on the prior–a nearly flat line. This panel indicates that by itself the prior leads to an uninformative predictive density. But the prior is informative, in that it governs how the posterior predictive density will respond to new data. In particular, our prior smooths the data a good deal but does not completely ignore the weak evidence of bimodality in the data, While most of the draws for q in the top panel are essentially unimodal, ignoring the two "bumps" in the data, some posterior weight is given to a bimodal representation; a different prior which concentrated most of its mass on unimodal densities would likely

Figure 2.4: Posterior Inference, High School Graduates



Data and Draws for q

Prior and Posterior Predictive Densities

downweight the evidence of bimodality much more. Averaging over the draws for q, as the posterior predictive density in the bottom panel does, results, in an asymmetric predictive density that would be difficult to approximate by any standard parametric distribution.

This model for the unknown density of the earnings residuals results in reasonably tight, plausible posterior distributions despite the rather small sample sizes. The posterior predictive distribution resembles the results one might obtain from kernel smoothers, and placing a prior distribution on α has allowed the predictive distribution to adapt to the degree of smoothness in the data, much in the way that modern kernel smoothing uses data-based choices for the bandwidth. This method has an important advantage over kernel and related nonparametric methods in that there is a well–defined posterior distribution for the object of interest, the unknown density. So we can characterize and examine the remaining uncertainty about q in many different ways. For example, the draws for q in Figure 2.19 (in the appendix, section 2.7) suggest much greater posterior uncertainty using the extremely small subsample of high school dropouts than with the larger subsamples as in Figures 2.20 and 2.21, as one would expect. If interest centered on a particular functional of q, there would be a well-defined posterior distribution for this quantity, and it would be straightforward to provide interval estimates and compare them to the corresponding intervals based only on the prior.

Another notable feature of the posterior distribution is that the posterior draws for the unknown density q tend to place most of their mass near the observed support of the data, even though the prior, which was quite diffuse for the location parameters μ_j, did not rule out much more extreme distributions. In this sense the posterior seems to be fairly localized around the data. At the same time, a small number of draws for q do place some weight at extreme positive and negative values, because the data has not completely dominated the prior. So in some applications (such as with unbounded loss or utility functions) some care must be taken with this particular choice of prior distribution; in the further analysis below we will make some modifications to the prior which will help remedy this problem.

2.3 Modeling Earnings Dynamics: Autoregressions with Random Effects

The density model examined in the previous section indicates that normality may not be a reasonable assumption for the marginal distribution of log earnings residuals, so parametric models with normal disturbances may fail to capture interesting aspects of the earnings process. One possibility would be to try to build a completely general, nonparametric model of the joint distribution of $(Y_{i1}, \ldots, Y_{iT})$, conditional on covariates, by using a multivariate extension of the density model. But as the dimensionality of the problem increases, the prior distribution will play an increasingly important role relative to the data; this is an example of the so-called curse of dimensionality. So there is a greater danger that a poor choice for the prior may have serious consequences for the resulting inference. Moreover, it would not be clear how to extend the model to future time periods, which is essential for the forecasting and consumption applications which motivated the analysis. It seems sensible therefore to try to impose some additional structure, via restrictions on the form of the likelihood function, while still allowing some parts of the model to be nonparametric.[6] We will therefore focus on developing *semiparametric* models, in which the likelihood function is not completely general even though its parameter is infinite-dimensional.

To get some idea of the kinds of restrictions that might be appropriate, we begin by first examining empirical autocovariance matrices of the earnings residuals. Recall that our data records earnings (residuals) for n individuals over 10 years. The covariance matrices are displayed in the appendix, section 2.7, and show positive autocovariances which decay slightly at longer lags. The autocovariances do remain fairly high even over 10 years, suggesting that between-individual variation plays an important role. So a low-order linear model might be an appropriate restriction, provided that we allowed some permanent (or nearly permanent) additive component to differ across individuals, in order to induce strong between-individual variation. In addition to displaying the covariance

[6] Such restrictions are unlikely to be literally true, of course.

matrix for the full data set, we calculate covariance matrices for the three education-based subsamples. The autocovariance structures appear to be quite different, a feature which we will examine in more detail using the full probability models.

2.3.1 Parametric Random Effects Model

The form of the covariance matrices suggests that we should look for a model with some "permanent" form of heterogeneity across individuals, and a relatively simple form of dependence over time for a given individual. A convenient model, widely used in econometric analyses of dynamic panel data, is the first-order autoregression with random effects. A parametric version of this model is:

$$y_{it} = \gamma_i + \rho \cdot y_{i,t-1} + \epsilon_{it}, \qquad i = 1, \ldots, n, \;\; t = 2, \ldots, T, \tag{2.4}$$

where we assume normality for both the innovation term ϵ_{it} and the random effect γ_i:

$$\epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^{-1}), \qquad \gamma_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\psi, \Omega);$$

and the ϵ_{it} and γ_i are assumed to be mutually independent. We use the improper flat prior for ρ and ψ and independent scaled chi-square priors for τ and Ω^{-1}:

$$\tau \sim \frac{\chi^2(1)}{.01}, \quad \Omega^{-1} \sim \frac{\chi^2(1)}{.01}.$$

This prior is very close to the conventional vague prior for τ, which would be uniform for $\log(\tau)$.

Posterior distributions for the entire sample and the three subsamples, obtained by simulation, are summarized in Table 2.1. The MCMC algorithm is described in more detail in section 2.6. The autoregressive parameter ρ, which is one of the key determinants of the persistence of earnings innovations, varies dramatically by educational level; ρ increases from about 0.35 for the high school dropouts to 0.60 for the college graduates; even accounting for uncertainty about ρ in the various models, there is strong evidence of systematic differences in persistence.[7] Pooling the groups, as is commonly done in studies of longitudinal earnings dynamics, could therefore be misleading, especially if the analysis focuses on forming predictive distributions for future earnings. In what follows, we will only work with the three subsamples, rather than pooling individuals with different levels of education.

Treating each subsample separately means that the sample sizes become quite small (37 by 10 for the high school dropouts), and a natural concern is that statistical inference may be seriously misleading. This suggests that we should focus on entire posterior distributions, or at least provide interval estimates for interesting quantities, rather than relying on point estimates.

2.3.2 Nonparametric Random Effects

One way to generalize the parametric random effects model is to relax the assumption of normality for the distribution of the individual components γ_i. We continue to use the random effects model of equation 2.4, and the assumption that the innovations ϵ_{it} are

[7] Bertrand (1997) finds similar results and interprets this as evidence that for higher education groups, wage setting is more likely to take place in the context of internal labor markets, rather than in spot markets for labor.

Table 2.1: Posterior Inference in the Normal Random Effects Model

	ρ	$\tau^{-1/2}$	ψ	$\Omega^{1/2}$
Full Sample	0.51	0.27	0.00	0.16
	(0.02)	(0.00)	(0.01)	(0.01)
High School Dropouts	0.35	0.32	0.03	0.20
	(0.07)	(0.01)	(0.04)	(0.04)
High School Graduates	0.45	0.30	0.00	0.18
	(0.05)	(0.01)	(0.02)	(0.03)
College Graduates	0.60	0.25	0.00	0.13
	(0.05)	(0.01)	(0.01)	(0.03)

Posterior means (standard deviations) obtained by MCMC simulation.

normal. The γ_i have an arbitrary density q with the prior used in the previous section.[8] This model is closely related to the model of Rosner and Müller (1996).

The estimated distributions of the random effects are nonnormal; so this model would have different implications than the fully parametric model for predicting γ_i for some new individual not in the sample. However, this kind of predictive distribution is not our primary focus. We are mainly interested in predicting future earnings given some earnings history, so it is more relevant to ask whether the two models give different results for predictive distributions of future earnings, for a given earnings history. One can consider any earnings history, but it is convenient to focus on an observed earnings history, so we would focus on forecasting future earnings, for an individual in the data sample. (This notion of a predictive distribution is discussed at length below.)

We have found that inference for the key parameters, and predictive distributions, are virtually identical to results using the normal random effects prior (see for example Table 2.2, where the "Normal" and "NP Random Effects" estimates are nearly indistinguishable), although the estimated distribution for γ_i in the more general model is clearly nonnormal. We therefore consider an alternative, and potentially more important, generalization of the parametric random effects model.

2.3.3 Nonparametric Errors

One interpretation of the results using the model with nonparametric random effects is that the data are sufficiently informative about the individual parameters γ_i that little

[8] Computational details for posterior inference in this model are available on request from the author.

is gained by using the nonparametric formulation, provided that the parametric prior is not too far away from the "true" distribution. Therefore, we continue to use the random effects model in equation 2.4, but now consider a general form for the innovations, and a normal prior for the random effects. The random effects prior takes the form

$$\gamma_i \sim \mathcal{N}(0, \Omega), \qquad \frac{1}{\Omega} \sim \chi^2(1)/(.01).$$

We have modeled the random effect as having zero mean, because it would be difficult to impose the same condition on the innovations. The innovations are modeled as arising from the countable mixture of normals model, represented in the latent variable notation as:

$$\epsilon_{it} \sim \mathcal{N}(\mu_{it}, \sigma_{it}^2), \qquad \theta_{it} \equiv (\mu_{it}, \sigma_{it}^2) \sim P,$$

$$P \sim \mathcal{D}(\alpha P_0), \qquad \alpha \sim \mathcal{G}(2, .5),$$

$$P_0(\mu, \sigma): \quad \frac{1}{\sigma^2} \sim \frac{\chi^2(3)}{3 \cdot 0.01}, \quad \mu|\sigma^2 \sim \mathcal{N}(0, (4/.01) \cdot \sigma^2).$$

This differs slightly from the prior distribution used in the earlier density estimation, in that the prior variance of the location parameters μ_j is smaller. If the prior variance of μ_j is too large, the resulting predictive distributions will give some weight to very extreme, clearly implausible values. However, even this new prior variance for μ_j is still quite large, so it may be useful to examine even more restrictive prior specifications in future work. Results for AR parameters are summarized in Table 2.2, in rows indicated by "NP Errors." Estimates are similar to the previous two models for the high school dropouts subsample, but quite different for the high school and college graduates subsamples. In particular, the nonparametric errors model finds much greater persistence as measured by the autoregressive parameter ρ, but this is accompanied by a smaller variance for γ_i. We will examine this finding in greater detail below, using the predictive distribution.

2.3.4 Predictive Distributions

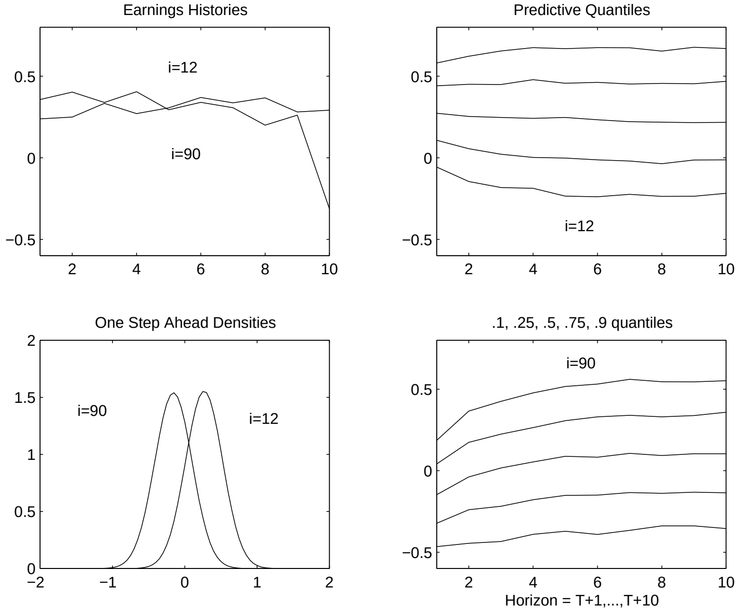
The tables provide an incomplete view of the implications of the models. We could provide plots for the density of ϵ_{it}, but it seems more fruitful to go back to our initial motivation. For both the optimal consumption problem, and for forecasting transitions in and out of poverty, we want to provide distributions for future earnings conditional on an earnings history. This means extending the models to future, hypothetically observable periods $t = T+1, \ldots, T+H$, and examining the implications of our models for the joint distribution of $y_{i,T+1}, \ldots, y_{T+H}$. Conditional on the entire set of parameters, the model specifies a conditional distribution for future earnings given any earnings history. However, since the parameters are not known, we need to decide how to deal with this source of uncertainty. For the forecasting application it seems sensible to form forecasts that take into account our uncertainty about the parameters of the model. In the optimal consumption problem, one could take the position that the agent making consumption decisions somehow knows the stochastic process for income. On the other hand, if the purpose of solving the optimal consumption problem is to obtain a best consumption policy based on the available information, it would make more sense not to condition on information we do not actually have. So we will focus on predictive distributions that are distributions for future data

Table 2.2: Comparing Posterior Inference in the Random Effects Models

Sample/Model	ρ	$\tau^{-1/2}$	$\Omega^{1/2}$
High School Dropouts			
Normal	0.35 (0.07)	0.32 (0.01)	0.20 (0.04)
NP Random Effects	0.36 (0.07)	0.32 (0.01)	– –
NP Errors	0.34 (0.04)	– –	0.23 (0.03)
High School Graduates			
Normal	0.45 (0.05)	0.30 (0.01)	0.18 (0.03)
NP Random Effects	0.45 (0.04)	0.30 (0.01)	– –
NP Errors	0.59 (0.06)	– –	0.14 (0.02)
College Graduates			
Normal	0.60 (0.05)	0.25 (0.01)	0.13 (0.03)
NP Random Effects	0.61 (0.05)	0.25 (0.01)	– –
NP Errors	0.83 (0.02)	– –	0.05 (0.01)

Posterior means (standard deviations) obtained by MCMC simulation.

Figure 2.5: Predictive Inference, Parametric Model



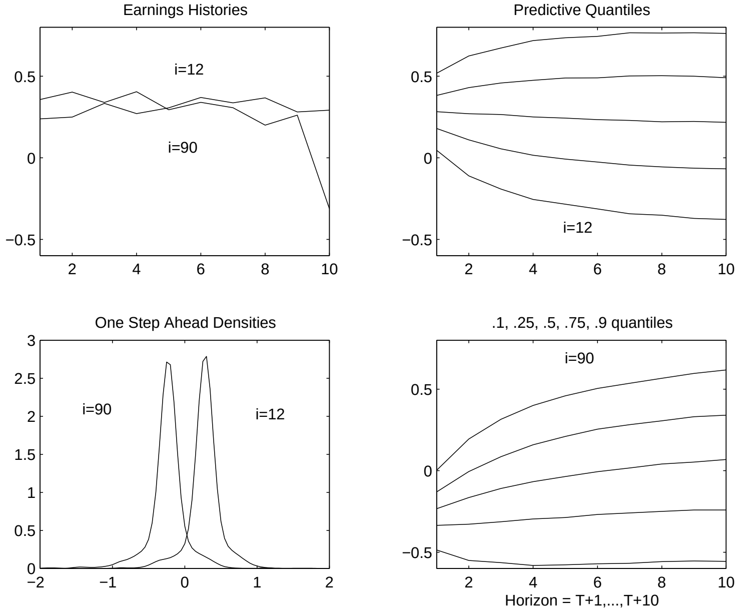
conditional on an earnings history, but not conditional on any unknown parameters. This can be done by starting with the distribution for future data conditional on past data and parameters and integrating over the posterior distribution for the parameters: let $z \equiv \{y_{jt} : j = 1, \ldots, n, t = 1, \ldots, T\}$ denote past data, and define

$$p(y_{i,T+1}, \ldots, y_{i,T+H}|z) = \int p(y_{i,T+1}, \ldots, y_{i,T+H}|z, \theta) dP(\theta|z).$$

2.3.5 Comparing Predictive Distributions

Each model implies a mapping from the set of possible earnings histories to a joint distribution for future earnings. We can examine the predictive distribution for any individual (as well as for any hypothetical earnings history); to display clearly some of the distinctions between the parametric and semiparametric models, we will begin by focusing on two college graduates with similar earnings histories except in year 10, when one of the

Figure 2.6: Predictive Inference, Semiparametric Model



individuals experienced an abrupt drop in earnings. Figure 2.5 shows the earnings histories in the top left plot, predictive densities for $y_{i,T+1}$ in the lower left plot, and in the two plots on the right, quantiles (.1, .25, .5, .75, .9) of the predictive distributions of future earnings for $t = T+1, \ldots, T+10$ under the parametric random effects model. Each of the quantile plots corresponds to one of the two individuals. Figure 2.6 shows corresponding plots under the model with nonparametric disturbances.

The predictive densities are strikingly different; compared to the normal-based predictive densities, more of their mass is concentrated within about .25 of their modes, but the tail probabilities clearly decay to zero at a much slower rate, so that these distributions are heavy-tailed. This is consistent with other studies which have found direct and indirect evidence of nonnormality and heavy tails in data on individual earnings.[9] It may also

[9] For example, MaCurdy (1982) estimates a number of panel data models on earnings data using normal-

help to explain the common complaint (see e.g. Deaton (1991)) that estimated variances in studies of earnings dynamics seem implausibly large. Also, there is weak evidence of skewness in that the density to the left of the mode appears to fall off slower than on the right. The predictive quantiles under the model with nonparametric errors start out more concentrated near $y_{i,T}$, but fan out rather quickly so that by ten years there is not much noticeable difference between the two models. However, the semiparametric model places ρ much higher than the parametric model, so that despite the smaller variance for γ_i, the mean-reversion occurs more slowly in the intermediate years.

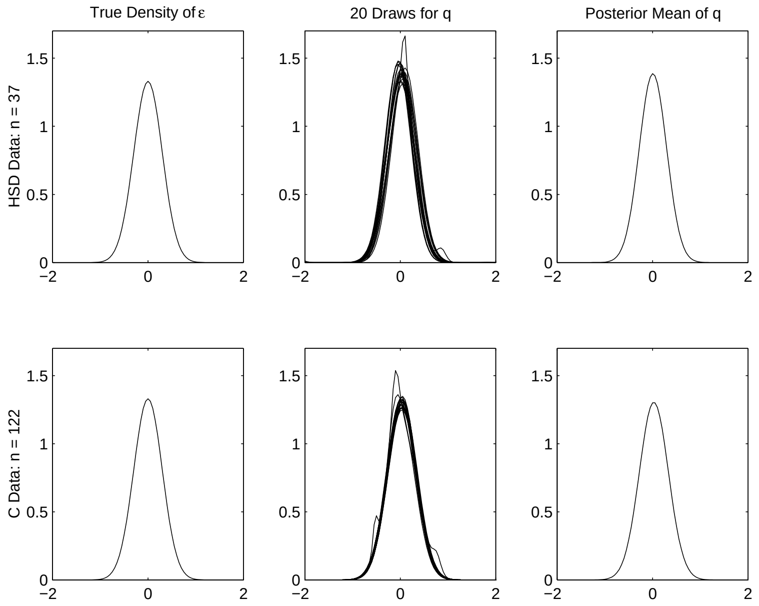
2.3.6 Analysis of Generated Normal Data

Our model seems to find strong evidence against the assumption of normality for ϵ_{it}. It could be that some unintended feature of our prior, rather than the evidence in the data, is driving this result. To try to assess whether the prior is adequate in this sense, we have generated artificial data under the assumption that the random effects autoregressive model holds and the disturbances ϵ_{it} are normally distributed; we generate an artificial data set of high school dropouts, and an artificial data set of college graduates.

Figure 2.7 shows results using this generated data set. Inference for the generated high school dropouts sample (with $n = 37$) is displayed in the first row, and the second row shows inference for the generated college sample (with $n = 122$). The first column shows the probability density function for the true normal distribution used to generate the ϵ_{it}. The second column shows twenty draws for the density of ϵ_{it} from its posterior

based MLE, but corrects the standard errors to reflect possible nonnormality. The corrections lead to standard errors twice as large as under the assumption of normality.

Figure 2.7: Inference using Generated Data

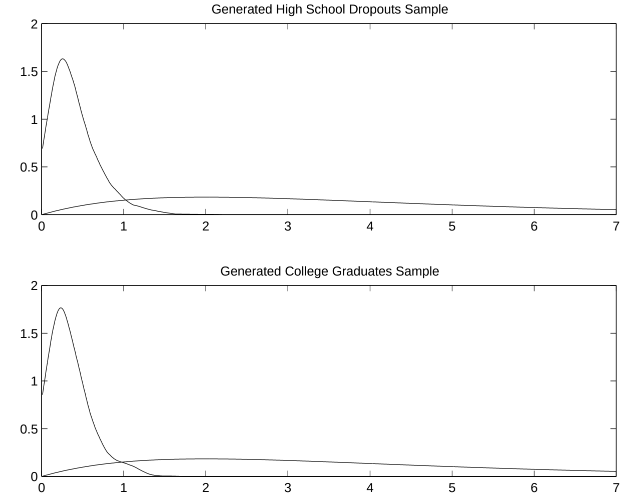


distribution. Since this is a posterior distribution on the space of density functions, each draw is a distinct density function. The posterior mean for this unknown density is given in the third column.

In both cases, the posterior means in the third column look very much like the true distribution. Moreover, the sample of draws for the density generally look reasonable. As expected, there is more variability in the smaller high school dropouts data; nevertheless, most of the draws are fairly close to the true density. The draws for the density in the college subsample are generally very close to the true density. This suggests that the evidence against normality in the previous analysis should be taken seriously.

More generally, it is encouraging to note that the posterior seems to be responding appropriately to the smoothness of the underlying data, a fact that can also be seen by examining the posterior for the parameter α. Figure 2.8 shows smoothed MCMC ap-

Figure 2.8: Posterior Distributions for α, Generated Data



proximations to the posterior distribution for α with the two generated data sets. Each posterior plot also shows the $\mathcal{G}(2, .5)$ prior for α, which is the nearly flat curve. In comparison to the prior, the two posterior distributions place most of their mass on relatively small values for α. A different prior for α, that put more mass in the range $[0, 1]$, might allow the posterior distribution to be even more concentrated near 0, resulting in a posterior distribution for q that would perform even better if the true density of the disturbances were normal. But this would necessarily come at the expense of adaptability to departures from normality, so it is reassuring that the current choice of prior does quite well under normality.

2.3.7 Correlated Random Effects

The random effects models examined up to now have an undesirable, and easily relaxed, restriction. The model in equation 2.4 is a model for $y_{i2}, \ldots, y_{iT}$ conditional on y_{i1}. This means that we have specified that the random effect γ_i is independent of y_{i1}, a restriction that seems unreasonable. For example, suppose y_{i1} were drawn from its stationary distribution given γ_i (and the other parameters); there would then be a strong link between γ_i and y_{i1}, rather than independence. A better prior distribution for γ_i would build in correlation with the initial condition:

$$\gamma_i | y_{i1} \sim \mathcal{N}(\psi y_{i1}, \Omega).$$

Table 2.3 summarizes inference for some of the key parameters in the correlated random effects models. In general the posterior means for the autoregressive coefficients ρ are lower than under the more restrictive independent random effects models; compare Table 2.2 with Table 2.3. There is still a marked contrast between the parametric and semiparametric estimates, especially for the higher education groups. For both the high school graduates and the college graduates subsamples, the semiparametric estimates for the autoregressive coefficient ρ are higher than the parametric estimates, but this is accompanied by lower values for ψ, the coefficient relating the individual effect γ_i to the initial condition y_{i1}. So the implications for predictive distributions are not entirely clear from this table alone.

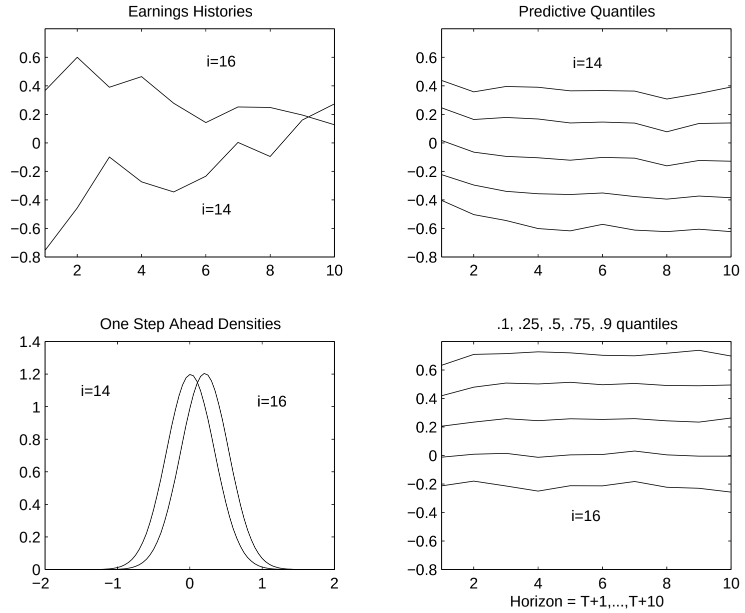
In Figures 2.9 and 2.10 we examine predictive distributions for two high school dropouts who have quite different average earnings, but similar earnings in the last period. In the model with normal disturbances, the predictive quantiles over the ten year horizon show some mean reversion, which is more noticeable for the inividual with rising income over his earnings history. However, the predictive quantiles do not fan out very much over the horizon. In comparison, the semiparametric model estimates the density of the disturbances to be extremely skewed, with a long left tail; this can be seen in the predictive densities

Table 2.3: Correlated Random Effects Models

Sample/Model	ρ	$\tau^{-1/2}$	ψ	$\Omega^{1/2}$
High School Dropouts				
Normal	0.34	0.32	0.27	0.19
	(0.06)	(0.01)	(0.11)	(0.04)
NP Errors	0.33	–	0.27	0.22
	(0.04)	–	(0.11)	(0.03)
High School Graduates				
Normal	0.42	0.29	0.32	0.15
	(0.04)	(0.01)	(0.05)	(0.02)
NP Errors	0.53	–	0.26	0.13
	(0.05)	–	(0.05)	(0.02)
College Graduates				
Normal	0.51	0.24	0.23	0.13
	(0.04)	(0.01)	(0.04)	(0.02)
NP Errors	0.75	–	0.10	0.07
	(0.03)	–	(0.03)	(0.01)

Posterior means (standard deviations) obtained by MCMC simulation.

Figure 2.9: Predictive Inference, Parametric CRE Model, High School Dropouts



for $y_{i,T+1}$. It seems that in order to capture the heavy tails, the parametric normal model has had to set the estimated variance rather high. So in general the quantiles are much more compressed even at long horizons in the semiparametric model.

Next, consider two high school graduates, chosen to have similar earnings histories, except in the last period. Figures 2.11 and 2.12 display predictive distributions for these individuals. Generally, the predictive quantiles at long horizons are similar, but at short horizons the quantiles in the semiparametric model are more compressed. The predictive densities for period $T+1$ are noticeably different, with the semiparametric model displaying tails that decay more slowly. Unlike with the high school dropouts subsample, here the density of the disturbances seems to be more nearly symmetric.

Figures 2.13 and 2.14 display predictive inference for two college graduates who have similar values for y_{iT} but different average earnings over their histories. The higher esti-

Figure 2.10: Predictive Inference, Semiparametric CRE Model, High School Dropouts

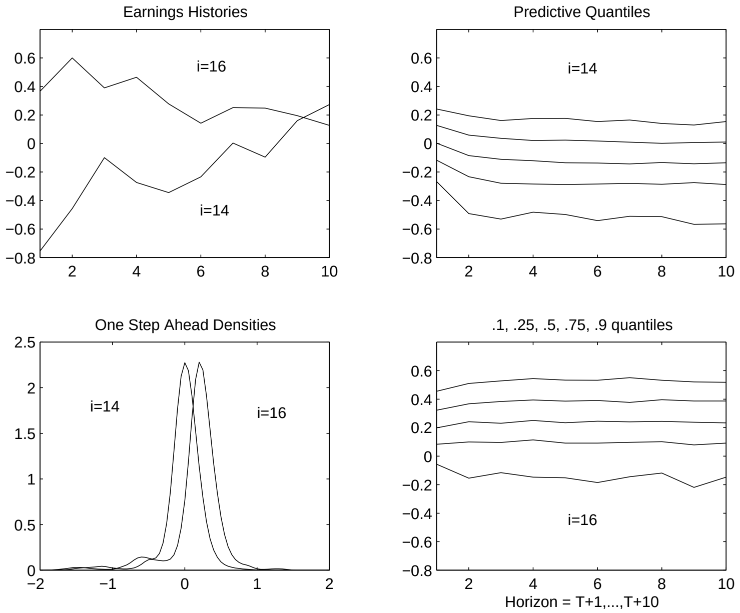


Figure 2.11: Predictive Inference, Parametric CRE Model, High School Graduates

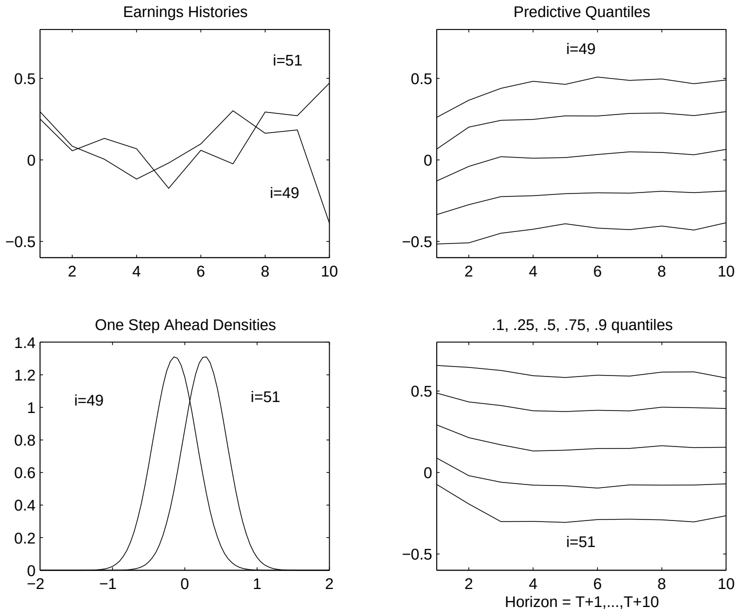
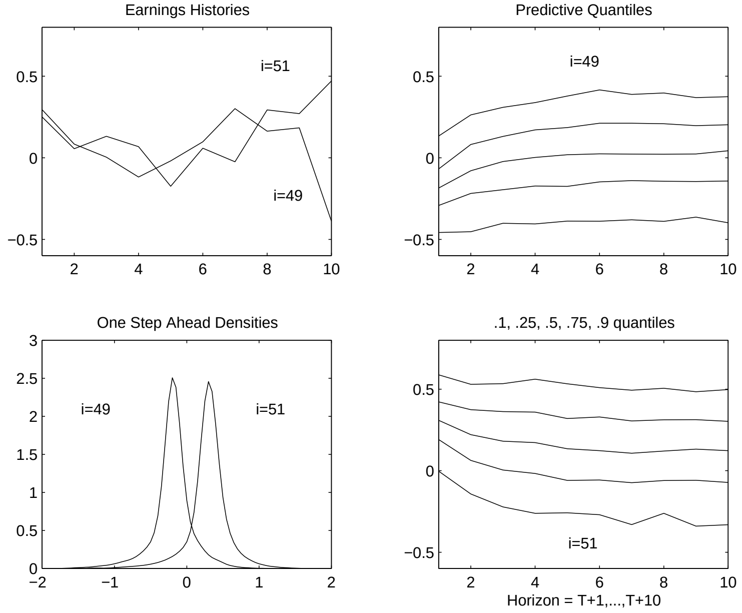


Figure 2.12: Predictive Inference, Semiparametric CRE Model, High School Graduates



mated value for ρ in the semiparametric model is reflected in the predictive densities for $y_{i,T+1}$, which are almost indistinguishable. In contrast, in the parametric model there is less overlap, because reversion to the mean happens more rapidly. In the short term, the predictive quantiles shift more slowly in the semiparametric model, but by ten years the quantiles are fairly similar. However, in contrast to the other subsamples, the quantiles under the semiparametric model are wider than with the parametric model.

We also show, in Figures 2.15 and 2.16, predictive distributions for the same two individuals examined under the (uncorrelated) random effects model. These look similar to Figures 2.5 and 2.6; in both sets of figures the mean-reversion occurs much more slowly under the semiparametric model.

One potential problem with the AR(1) specification we have been using up to now is that it forces predictive distributions to be quite sensitive to earnings in the final period.

Figure 2.13: Predictive Inference, Parametric CRE Model, College Graduates

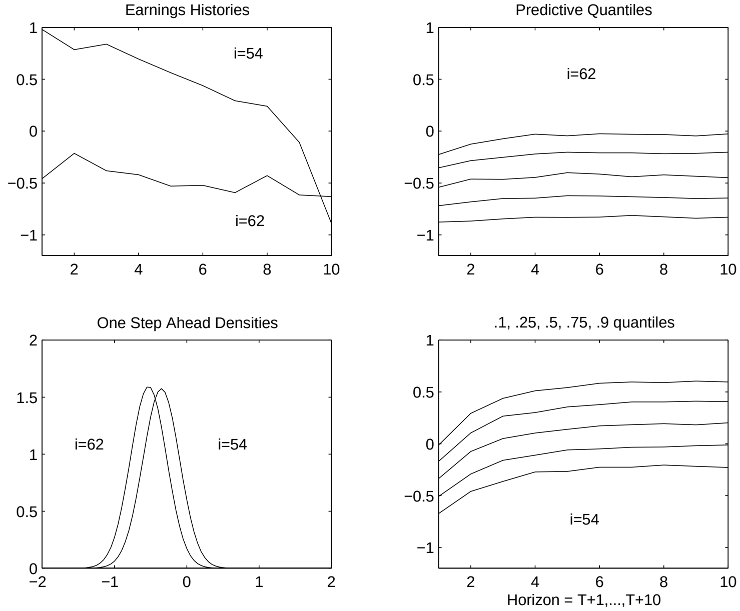


Figure 2.14: Predictive Inference, Semiparametric CRE Model, College Graduates

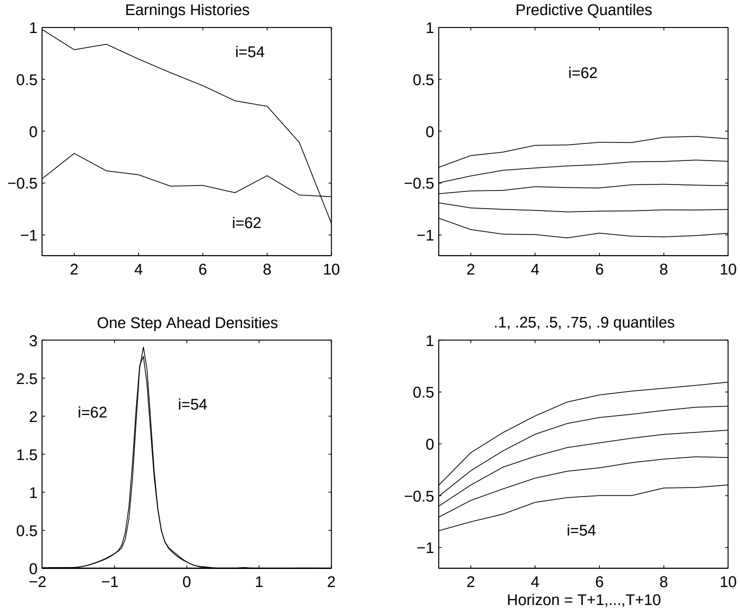


Figure 2.15: Predictive Inference, Parametric CRE Model, College Graduates

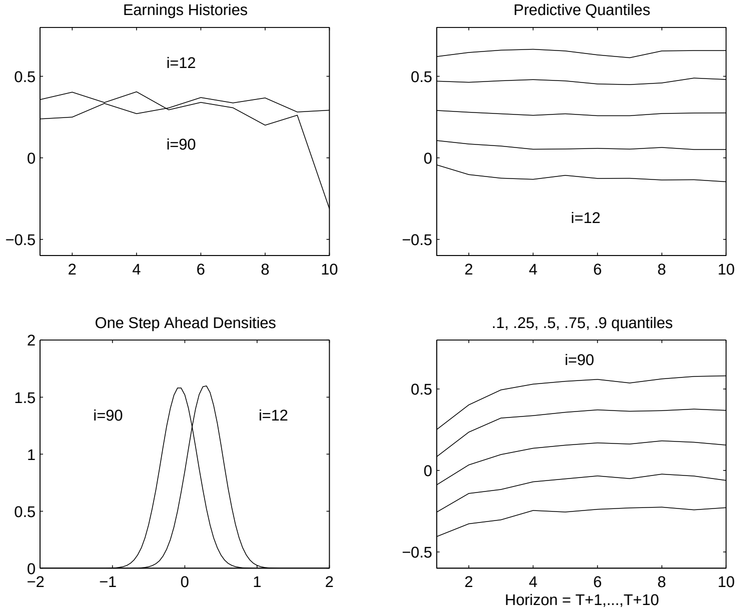
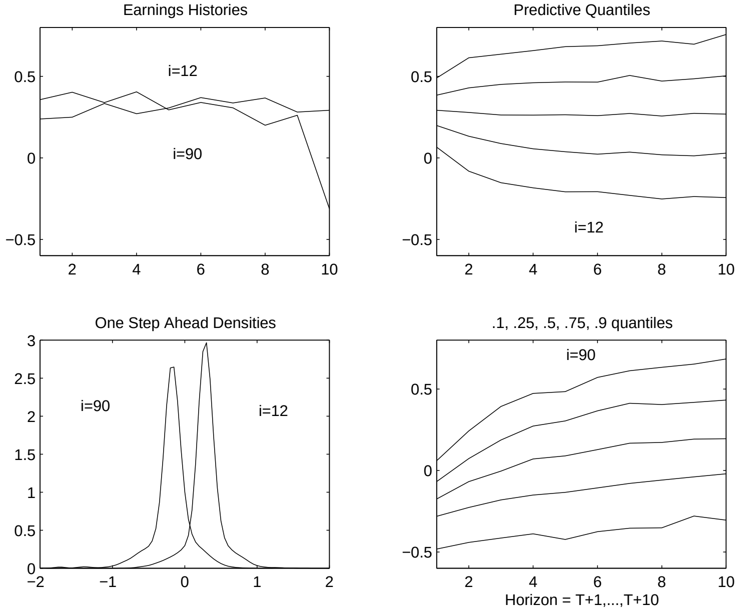


Figure 2.16: Predictive Inference, Semiparametric CRE Model, College Graduates



This can be seen in Figures 2.15 and 2.16, in which a one-period decline in earnings has a pronounced effect on predictive distributions at short to moderate horizons. It could be that a better predictive distribution would attribute some probability that such an abrupt earnings decline was completely temporary. One way to examine this possibility is to extend the model directly, by adding MA and higher-order AR terms; in the next section we will pursue a slightly different strategy that is widely used in optimal consumption models.

2.4 Error Components Models

Many different generalizations of the AR(1) model of the previous section are possible; one that has been studied and used quite extensively has the form:

$$y_{it} = \gamma_i + v_{it} + \epsilon_{it}, \qquad i = 1, \ldots, n, \;\; t = 1, \ldots, T;$$

$$v_{it} = \rho v_{i,t-1} + w_{it} \qquad i = 1, \ldots, n, \;\; t = 2, \ldots, T;$$

$$v_{i1} \sim \mathcal{N}(0, \sigma_v^2), \;\; \epsilon_{it} \sim \mathcal{N}(0, \sigma^2), \;\; w_{it} \sim \mathcal{N}(0, \sigma_w^2), \;\; \gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2).$$

A semiparametric extension would let the distribution for ϵ_{it} be general:

$$\epsilon_{it} \sim \mathcal{N}(\mu_{it}, \sigma_{it}^2), \qquad \theta_{it} \equiv (\mu_{it}, \sigma_{it}^2) \sim P, \qquad P \sim \mathcal{D}(\alpha P_0), \qquad \alpha \sim \mathcal{G}(a_1, a_2).$$

This error components model is equivalent to an ARMA(1,1) up to second moments. A problem with this formulation is that the individual components γ_i become unidentified when $\rho = 1$. The posterior distribution is still well-defined if a proper prior distribution is used for all the parameters, but for diffuse priors the posterior (and some of the computations to obtain the posterior) can be ill-behaved. We will drop the individual effect, so that our model is:

$$y_{it} = v_{it} + \epsilon_{it},$$

$$v_{it} = \rho v_{i,t-1} + w_{it},$$

$$v_{i1} \sim \mathcal{N}(0, \sigma_v^2), \;\; w_{it} \sim \mathcal{N}(0, \sigma_w^2), \;\; \gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2).$$

and either

$$\epsilon_{it} \sim \mathcal{N}(0, \sigma^2)$$

or

$$\epsilon_{it} \sim \mathcal{N}(\mu_{it}, \sigma_{it}^2), \quad \theta_{it} \equiv (\mu_{it}, \sigma_{it}^2) \sim P, \quad P \sim \mathcal{D}(\alpha P_0), \quad \alpha \sim \mathcal{G}(a_1, a_2).$$

There is still "heterogeneity" because v_{i1} will vary across individuals, but if ρ is much smaller than 1 in absolute value, this heterogeneity will shrink over time. We use independent $\chi^2(1)/(.01)$ priors for σ^{-2}, σ_v^{-2}, and σ_w^{-2}, and a uniform prior for ρ.

Table 2.4 shows posterior means and standard deviations, obtained by simulation, for the parametric and semiparametric versions of the error components model. The autoregressive coeffiecient ρ is estimated to be close to, but slightly less than, 1. Under the parametric model the estimates for ρ are generally lower than under the semiparametric models. But in most of the models ρ seems close enough to 1 that there is not a substantial loss of generality in dropping the individual effect γ_i. Comparing different educational subsamples, there is some weak evidence that ρ increases with education, but there is a stronger contrast in some of the variance parameters. As education increases, σ is

Table 2.4: Error Components Models

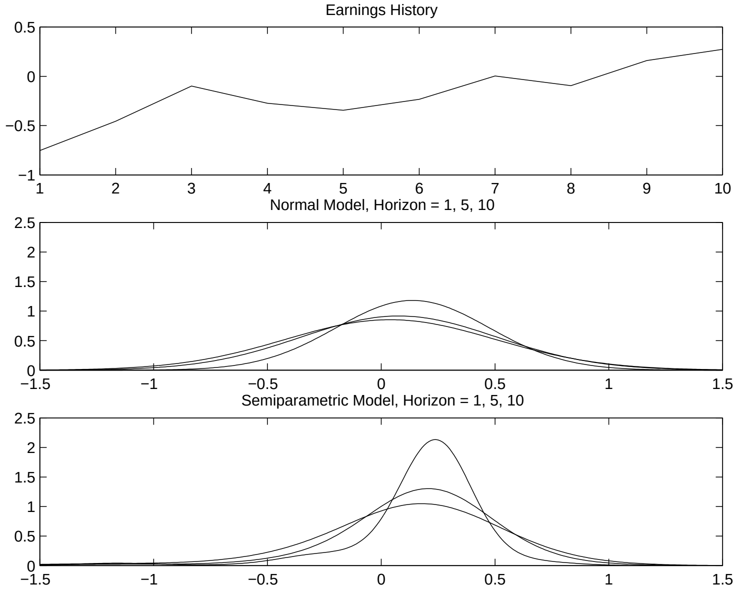
Sample/Model	ρ	σ_w	σ_v	σ
High School Dropouts				
Parametric	0.87	0.21	0.27	0.23
	(0.05)	(0.03)	(0.05)	(0.02)
Semiparametric	0.95	0.12	0.28	–
	(0.03)	(0.01)	(0.04)	–
High School Graduates				
Parametric	0.89	0.20	0.33	0.20
	(0.03)	(0.02)	(0.03)	(0.01)
Semiparametric	0.96	0.13	0.32	–
	(0.01)	(0.01)	(0.03)	–
College Graduates				
Parametric	0.92	0.17	0.42	0.16
	(0.02)	(0.01)	(0.03)	(0.01)
Semiparametric	0.97	0.12	0.33	–
	(0.01)	(0.01)	(0.02)	–

Posterior means (standard deviations) obtained by MCMC simulation.

estimated to be smaller, while σ_v is estimated to be larger. So more of the residual variance in earnings is explained by (nearly) permanent differences across individuals, rather than by transitory shocks.

Figure 2.17 compares predictive inference based on the parametric and semiparametric error components models for one high school dropout, whose earnings history is plotted in the top panel. The middle panel shows predictive densities for $y_{i,T+1}$, $y_{i,T+5}$, and $y_{i,T+10}$, earnings at one, five, and ten year horizons, based on the parametric model. These become dispersed at the longer horizons, though the difference between the five and ten year horizons is very small. The bottom panel shows the same predictive densities,

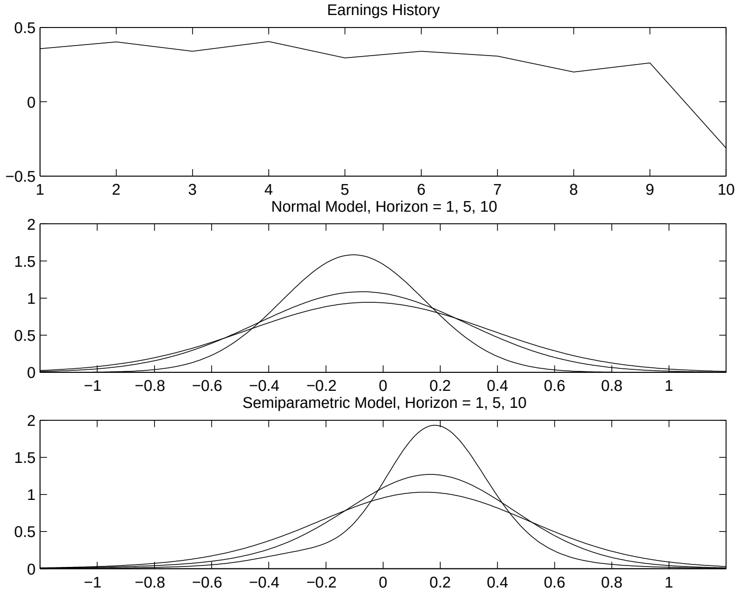
Figure 2.17: EC Model, High School Dropouts



but calculated using the semiparametric model. These are very different, especially at the one year horizon. The predictive density one period ahead has a long left tail, but also appears to be more compressed near its mode. For example, much more of the mass of the predictive density appears to fall between 0 and 0.5 under the semiparametric model than under the parametric model.

Figure 2.18 shows a similar set of plots for a college graduate who experienced a drop in earnings in the last period. (This is the same individual who appears in the first set of plots for the random effects autoregressive models.) The predictive densities under the semiparametric model exhibit skewness, but there is also a remarkable difference in the location of the predictive densities. Under the semiparametric model, the predictive densities have their mode near 0.2, while under the parametric model the mode is near -0.1. Allowing the distribution of ϵ_{it} to be general instead of restricting it to be normal

Figure 2.18: EC Model, College Graduates



leads to very different filtering rules for the (nearly) permanent component v_{it}, so that a large, abrupt decline in earnings is regarded as more likely to be temporary under the semiparametric model.

2.5 Conclusion

In this paper we have applied recent advances in nonparametric Bayesian modeling to develop a simple, but flexible, class of models for earnings dynamics. The first step was to build a model for density estimation, by constructing a nonparametric prior out of elementary probability distributions (normal, chi-square, gamma, and beta). Our goal was to avoid making sharp restrictions but apply a prior judgment that (log) earnings distributions should be smooth and possibly quite close to normal. Application of this model to the PSID data on earnings at age 24 led to distinctly nonnormal density estimates.

Then we turned to the more difficult problem of modeling the joint distribution of earnings from age 24 to age 33. We used restrictions on the form of the dependence over time to put some structure on what would otherwise be an intractable problem, but allowed the disturbances in the resulting low-order linear models to be completely general, by incorporating our nonparametric density model. These generalizations of conventional parametric models yielded markedly different predictive distributions, especially at short horizons. In comparison to normal-based models, the semiparametric predictive distributions appeared to be heavy-tailed and skewed to the left. At the same time, predictive distribution were more compressed around their mode, because the normal-based models had to inflate the estimated variance of earnings to cope with the small amount of mass at extreme values. The semiparametric models also showed quite different responses to large earnings shocks. In the AR(1) models with random effects, the semiparametric models tended to have higher estimates for the autoregressive coefficient ρ than the normal-based models, so that the effect of a large shock would not fade as rapidly. In the error components model (which is almost a strict generalization of the random effects AR(1) model), there was also greater persistence in the semiparametric versions. But this effect was overshadowed by the tendency of the semiparametric model to attribute more of a large earnings shock to temporary rather than permanent shifts in future earnings. So the predictive distributions were relatively insensitive to a large earnings shock in a single period; however, they could respond quite dramatically to a sustained downturn in earnings.

We conclude that economic models that assume conditional log normality of the earnings process should be viewed with caution, and that incorporating more realistic distributional assumptions can have consequences for forecasting and optimal consumption policies, although this needs to be examined more carefully in future work. The ability of this semiparametric approach to cope with heavy-tailed and skewed distributions suggests that it may be usefully adapted to other economic contexts, such as modeling financial asset returns, where predictive distributions are likely to be useful. More generally, this paper has tried to illustrate how two strategies for dealing with difficult inference problems—using general likelihood functions but taking advantage of smoothness, and imposing restrictions on the form of the likelihood function—can be fruitfully combined in practice.

2.6 Computational Appendix

We review a number of results on posterior inference using the Dirichlet process prior, beginning with its definition and basic properties, and moving on to a discussion of methods for posterior simulation using MCMC. Then we briefly present the algorithms used for posterior simulation in the parametric and semiparametric dynamic panel data models.

2.6.1 Definition of the Dirichlet Process

Let Θ be a space and $\mathcal{A}$ a σ-field of subsets of Θ, let ν be a finite non–null measure on $(\Theta, \mathcal{A})$. A stochastic process P indexed by $A \in \mathcal{A}$ is said to be a *Dirichlet process* on $(\Theta, \mathcal{A})$ with base measure ν if for any measurable partition $(A_1, \dots, A_k)$ of Θ, the random vector $(P(A_1), \dots, P(A_k))$ is distributed as $D(\nu(A_1), \dots, \nu(A_k))$, where D denotes the Dirichlet distribution. We denote the Dirichlet process by $P \sim \mathcal{D}(\nu)$.

Ferguson (1973) (see also Antoniak (1974) and Sethuraman (1994)) defined the Dirichlet process and showed the following useful properties:

1. The Dirichlet process is a probability distribution on the space of probability distributions on $(\Theta, \mathcal{A})$, and selects a discrete distribution with probability one.

2. Let θ be a random variable with distribution P, where P has a Dirichlet process distribution. Then the marginal distribution of θ is $\nu/\nu(\Theta)$, that is, the Dirichlet base measure normalized to be a probability measure.

3. Let θ be a random variable with distribution P, where P has a Dirichlet process prior distribution with base measure ν. Then the posterior distribution of P given θ

is also a Dirichlet process with base measure $\nu + \delta(\theta)$. Here $\delta(\theta)$ denotes unit point mass at θ.

In the development of our prior we have used an alternative construction of the Dirichlet process due to Sethuraman and Tiwari (1982) (see also Sethuraman (1994)). Let $\lambda_1, \lambda_2, \ldots$ be independent draws from the normalized Dirichlet base measure $\nu/\nu(\Theta)$. Let $r_1, r_2, \ldots$ be independent draws from $\text{Beta}(1, \nu(\Theta))$, a beta distribution with parameters 1 and $\nu(\Theta)$. Form $p_j = r_j \prod_{l=1}^{j-1}(1-r_l)$. This constructs a sequence $\{p_1, p_2, \ldots\}$ as a beta "stick–breaking" process. Then define

$$P = \sum_{j=1}^{\infty} p_j \delta(\lambda_j).$$

It is clear that the sum of the weights must be one, and that the parameter α will control whether the process gives most of its weight to a few components or will spread its weight over many components.

This construction of a random probability measure P is shown in Sethuraman (1994) to be equivalent to the Dirichlet process. One cannot use this construction to obtain exact draws from a Dirichlet process; however, by taking the first J terms in the sequences $\{\lambda_j\}$ and $\{p_j\}$, where J is suitably large, one can obtain an approximate draw for P.

2.6.2 Computations using the Dirichlet Process

Recall that the density model specifies: for $i = 1, \ldots, n$,

$$y_i|\theta_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad \theta_i \equiv (\mu_i, \sigma_i^2)|P \sim P, \tag{2.5}$$

$$P \sim \mathcal{D}(\alpha P_0). \tag{2.6}$$

Here $\alpha > 0$ is a scalar and P_0 is a probability measure. Thus, αP_0 is a finite non–null measure and can serve as the Dirichlet process base measure, taking the place of ν in the original definition of the Dirichlet process. $P_0(\theta) = P_0(\mu, \sigma^2)$ specifies:

$$\frac{1}{\sigma^2} \sim \frac{\chi^2(s)}{sQ}, \qquad \mu|\sigma^2 \sim \mathcal{N}(m, b \cdot \sigma^2),$$

and (α, s, Q, m, b) are assumed given for now. Alternatively, one can write the model as:

$$y_i|P \sim \sum_{j=1}^{\infty} p_j \phi(\mu_j, \sigma_j^2), \tag{2.7}$$

where p_j, and $\theta_j \equiv (\mu_j, \sigma_j^2)$ are as given in the Sethuraman–Tiwari construction. So conditional on P, the distribution of an observation is a countably infinite normal mixture.

Escobar (1994) showed that using the Dirichlet process prior leads to a useful set of conditional distributions. For now, suppose α is fixed. Let $\theta^{(i)} \equiv (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ be the set of values of θ for units other than i. Then,

$$P|\theta^{(i)} \sim \mathcal{D}(\alpha P_0 + \sum_{j \neq i} \delta(\theta_j)).$$

Hence

$$\theta_i|\theta^{(i)} \sim E(P|\theta^{(i)}) \sim \frac{\alpha}{\alpha + n - 1} P_0 + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta(\theta_j).$$

Recall that y_i is independent of P and $\theta_j, j \neq i$ conditional on θ_i. Thus the conditional distribution of θ_i given $\theta^{(i)}$ and y can be obtained by taking the preceding expression as a prior distribution and applying Bayes' Theorem. This gives

$$\theta_i|\theta^{(i)}, Y \sim q_{i0}P_{i0} + \sum_{j \neq i} q_{ij}\delta(\theta_j), \tag{2.8}$$

where the probabilities q_{ij} are given by:

$$q_{i0} = c \cdot \alpha \cdot \int \phi(y_i|\mu, \sigma^2) dP_0(\mu, \sigma^2),$$

$$q_{ij} = c \cdot \phi(y_i|\mu_j, \sigma_j^2).$$

Here c is a normalizing constant, and $\phi(\cdot|\mu, \sigma^2)$ continues to indicate a normal density. P_{i0} can be interpreted as the posterior distribution for θ_i under the prior P_0. Since the prior was chosen to have a conjugate form, it is convenient to sample from P_{i0}. Note that the integral in the expression for q_{i0} is the marginal density of the data under the prior P_0. Given our choice for P_0 this can be expressed as a Student–t density and can be calculated directly.

Escobar (1994) and Escobar and West (1995) simulate the posterior distribution by iteratively drawing from the distribution in (2.8) for $i = 1, \dots, n$. This defines a Gibbs sampler on the space of values for $(\mu_i, \sigma_i^2)_{i=1}^n$ which converges to a stationary distribution equal to the posterior distribution. (Convergence issues are treated in Escobar (1994) and Escobar and West (1995); for a more general introduction to the theory underlying MCMC computations see Tierney (1994).) Notice that the nuisance parameter P has been essentially integrated out in the construction of the Markov chain.

By the discreteness of the Dirichlet process, for any draw $\theta = (\theta_1, \dots, \theta_n)$, there will only be $k \leq n$ distinct values of θ_i, and in some applications the posterior for k may concentrate most of its mass on $k = 1$.[10] Recent work has made further use of this clustering structure to improve the performance of simulation algorithms and to extend them to more complicated versions of this model.

Let $\xi_j, j = 1, \ldots, k$ denote the k distinct values of the θ_i. Define the *configuration* S_i by $S_i = j$ if $\theta_i = \xi_j$, and define $n_j = \#\{S_i = j\}$, $Y_j = \{y_i : S_i = j\}$. Since k will often be quite small, rather than work with θ it is often convenient to work with the equivalent variables $S = (S_1, \ldots, S_n)$ and $\xi = (\xi_1, \ldots, \xi_k)$. More importantly, working with the configurations suggests alternative simulation strategies.

West, Müller, and Escobar (1994), following work by MacEachern (1994) and Bush and MacEachern (1996), suggest constructing a Gibbs sampler by an alternative sequence of conditional draws. First, note that we can rewrite (2.8) as

$$\theta_i|\theta^{(i)}, Y \sim q_{i0}P_{i0} + \sum_{j=1}^{k^{(i)}} n_j^{(i)} q_{ij} \delta(\xi_j^{(i)}), \tag{2.9}$$

where the (i) superscript refers to variables defined on all units other than i. That is, $\xi^{(i)} = (\xi_1^{(i)}, \ldots, \xi_{k^{(i)}}^{(i)})$ are the distinct values among $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$. After cycling through these conditional distributions, we append a set of draws for the ξ_j given the configuration vector S. Note that conditional on the configuration vector S, the distribution for the ξ given the data Y is straightforward. Each ξ_j is independent of the other elements of ξ and has a distribution given by

$$p(\xi_j|\xi_1, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_k, S, Y, P_0) \propto \prod_{Y_j} \phi(y_i|\xi_j) dP_0, \tag{2.10}$$

[10] However, k will increase with n, as emphasized by the infinite normal mixture representation.

which is the posterior for ξ_j given the data in Y_j under the prior P_0. Given the conjugate form of P_0, this is straightforward. These additional steps have the effect of shifting the cluster locations at each Gibbs iteration, which helps keep the Markov chain from "getting stuck" for many periods at a particular value for θ. (More precisely, this extra step leads the Markov chain to have a transition kernel which is absolutely continuous with respect to the posterior measure.)

If P_0 had a nonconjugate form, it would be difficult to directly evaluate q_{i0}. MacEachern and Müller (1994) and Walker and Damien (1997) devise alternative MCMC strategies for nonconjugate specifications of P_0.

West (1992) and Escobar and West (1995) show that instead of fixing α, the total mass of the Dirichlet process base measure, at some constant value, one can place a prior distribution on α that has a tractable form. We outline the steps needed for computation; a derivation of these steps can be found in those articles.

Let the prior for α be $\mathcal{G}(a_1, a_2)$, with $a_1 > 0$. Then we can draw for α given k, the number of distinct values of θ_i, by appending the following steps to the Gibbs sampler.

1. Given the current value of α, draw η from $\text{Beta}(\alpha + 1, n)$, a beta distribution with mean $(\alpha + 1)/(\alpha + 1 + n)$.

2. Given η, draw a new value for α from a two–point mixture of gamma random variables:

$$\pi_\eta \mathcal{G}(a_1 + k, a_2 - \log(\eta)) + (1 - \pi_\eta)\mathcal{G}(a_1 + k - 1, a_2 - \log(\eta))$$

   where π_η is defined by $\pi_\eta/(1 - \pi_\eta) = (a_1 + k - 1)/[n(a_2 - \log(\eta))]$.

It is also possible to place prior distributions on the parameters that characterize P_0. If these are chosen to have a conjugate form, then inference would be straightforward by extending the Gibbs sampler. However, in what follows we will simply keep the hyperparameters for P_0 fixed.

2.6.3 Predictive Densities

There are different ways to summarize the implications of this model. One particularly useful approach is to study the predictive distribution of a new observation under the model. Note that for a new unit $i = n + 1$, we have

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta(\theta_i), \tag{2.11}$$

and therefore the distribution of y_{n+1} conditional on $\theta_1, \ldots, \theta_n$ is a mixture of k normal distributions and a Student–t distribution:

$$Y_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} T_s(m, Q(b + 1)) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \mathcal{N}(\mu_i, \sigma_i^2), \tag{2.12}$$

where the t distribution arises from marginalization of θ_{n+1} over the distribution P_0. We can then average over the posterior distribution of $(\theta_1, \ldots, \theta_n)$ to get the (marginal) predictive density $p(y_{n+1}|y_1, \ldots, y_n)$. Since the Gibbs sampler provides draws for the θ_i's, this averaging can be done by Monte Carlo.

We would also like to have a way to show the posterior uncertainty about the distribution of future observations. One possibility is to plot samples of the mixture density corresponding to (2.12). However, this corresponds to a draw for $E(q|\theta_1, \ldots, \theta_n)$, instead of a draw for $q(\cdot|P)$, and so samples of this density would understate the uncertainty about $q(\cdot|P)$.[11] Instead, following a suggestion of Gary Chamberlain, we construct approximate draws for $P|\theta_1, \ldots, \theta_n$ using the Sethuraman–Tiwari construction, and plot the correponding draws for $q(\cdot|P)$ as a useful indication of posterior uncertainty about model parameters. Conditional on a draw for $\theta_1, \ldots, \theta_n$ produced by the Gibbs sampler, we can draw, for $j = 1, \ldots, J$,

$$\lambda_j \overset{\text{i.i.d.}}{\sim} \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta(\theta_i),$$

$$r_j \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha + n),$$

where the λ_j and $r_{j'}$ are mutually independent. We then form

$$p_j = r_j \prod_{l=1}^{j-1} (1 - r_l).$$

We can then approximate a draw for q by

$$q_J(\cdot|P) = \sum_{j=1}^{J} p_j \phi(\cdot|\lambda_j) \approx q(\cdot|P).$$

In the applications developed below, the p_j decay to zero quite rapidly, so setting J around 50 or 100 seems to be more than adequate to visualize q. We also note that these approximations for P and q do not feed back into the Gibbs sampler, so that the truncation at a finite value for J does not affect convergence to the posterior distribution for $\theta_1, \ldots, \theta_n$, or the Monte Carlo evaluation of the predictive density $p(Y_{n+1}|Y)$.

[11] Nevertheless, retaining these draws can be useful in order to assess convergence of the Gibbs sampler.

2.6.4 MCMC Algorithm for the Random Effects Autoregressions

Parametric Random Effects

The Gibbs sampler successively samples for blocks of parameters according to the following set of distributions:

Blocks: (ρ, τ), $(\gamma_1, \ldots, \gamma_n)$, (ψ, Ω^{-1}).

(ρ, τ): Define

$$\hat{\rho} = \sum_{i=1}^{n} \sum_{t=2}^{T} y_{i,t-1}(y_{it} - \gamma_i) / \sum_{i=1}^{n} \sum_{t=2}^{T} y_{i,t-1}^2.$$

Then

$$\tau \sim \chi^2(n(T-1))/[.01 + \sum_{i=1}^{n} \sum_{t=2}^{T} (y_{it} - \gamma_i - \hat{\rho} y_{i,t-1})^2],$$

$$\rho \sim \mathcal{N}(\hat{\rho}, (\tau \sum_{i=1}^{n} \sum_{t=2}^{T} y_{i,t-1}^2)^{-1}).$$

(γ_i): Define

$$H_i = (\Omega^{-1} + (T-1)\tau),$$

$$\gamma_i^* = H_i^{-1}[\Omega^{-1}\psi + \tau \sum_{t=2}^{T} (y_{it} - \rho y_{i,t-1})].$$

Then

$$\gamma_i \sim \mathcal{N}(\gamma_i^*, H_i^{-1}).$$

(ψ, Ω): Define

$$\hat{\psi} = \sum_{i=1}^{n} \gamma_i / n.$$

Then

$$\Omega^{-1} \sim \chi^2(n)/[.01 + \sum_{i=1}^{n} (\gamma_i - \hat{\psi})^2],$$

$$\psi \sim \mathcal{N}(\hat{\psi}, \Omega/n).$$

Incorporating Covariates: The algorithm presented above can be modified to directly model the conditional mean of log earnings (z_{it}) as a function of covariates. The following is a sketch of how to modify the MCMC algorithms to deal with these model extensions. Let

$$z_{it} = g(x_{it}; t, \omega) + y_{it}, \tag{2.13}$$

where $g(x_{it}; t, \omega)$ denotes the conditional mean of z_{it} given covariates x_{it}, age t, and parameters ω. By having the covariates include indicators for calendar year this could also be made to incorporate aggregate shocks to earnings. Suppose also that y_{i1} has a $\mathcal{N}(0, \sigma_y^2)$ distribution.

Note that given ω, y_{it} can be formed by taking difference between z_{it} and $g(x_{it}; t, \omega)$. After forming the y_{it} in this way, the Gibbs sampler steps outlined above would continue to be appropriate, given the current draw for ω. The Gibbs sampler would be augmented at every iteration with two additional draws: a draw for σ_y, and a draw for ω. The step to draw for σ_y would only need to use y_{i1}, which again can be formed from equation 2.13. The step to draw for ω could be based on exisiting results for multivariate normal densities, since conditional on the other model parameters, equation 2.13 has the form of a multivariate regression model with known variance matrix. A similar argument could be used to incorporate covariates into the other Gibbs samplers described below.

Semiparametric Random Effects: Nonparametric Errors

Blocks: (ρ), (Ω), $(\gamma_1, \ldots, \gamma_n)$, (θ), (α).

(ρ): Define

$$\hat{\rho} = \sum_{i=1}^{n} \sum_{t=2}^{T} \sigma_{it}^{-2} y_{i,t-1}(y_{it} - \gamma_i - \mu_{it}) / \sum_{i=1}^{n} \sum_{t=2}^{T} \sigma_{it}^{-2} y_{i,t-1}^2.$$

Then

$$\rho \sim \mathcal{N}(\hat{\rho}, (\sum_{i=1}^{n} \sum_{t=2}^{T} \sigma_{it}^{-2} y_{i,t-1}^2)^{-1}).$$

(Ω):

$$\Omega^{-1} \sim \chi^2(1+n)/[.01 + \sum_{i=1}^{n} \gamma_i^2].$$

(γ_i): Define

$$H_i = \Omega^{-1} + \sum_{t=2}^{T} \sigma_{it}^{-2},$$

$$\gamma_i^* = H_i^{-1} \sum_{t=2}^{T} \sigma_{it}^{-2}(y_{it} - \rho y_{i,t-1} - \mu_{it}).$$

Then

$$\gamma_i \sim \mathcal{N}(\gamma_i^*, H_i^{-1}).$$

(θ_{it}): Define $\epsilon_{it} = y_{it} - \gamma_i - \rho y_{i,t-1}$, for $i = 1, \ldots, n$, $t = 2, \ldots, T$. For each pair (i,t), apply the following steps: Let $\xi_j^{(it)}$, $j = 1, \ldots, k^{(it)}$, denote the $k^{(it)}$ distinct values of $\{\theta_{ls} : (l,s) \neq (i,t)\}$, and define $n_j^{(it)} = \#\{\theta_{ls} : \theta_{ls} = \xi_j^{(it)}, (l,s) \neq (i,t)\}$. Calculate

$$\tilde{q}_0 = \alpha \cdot t_s(\epsilon_{it} | m, Q(b+1)),$$

where $t_s(\cdot|\mu, \sigma^2)$ denotes the density of a t distribution with s degrees of freedom, location parameter μ, and scale parameter σ. Also define

$$\tilde{q}_j = n_j^{(it)} \phi(\epsilon_{it} | \xi_j^{(it)}),$$

and

$$c = \tilde{q}_0 + \sum_{j=1}^{k^{(it)}} \tilde{q}_j.$$

Then form

$$q_0 = \tilde{q}_0/c, \qquad q_j = \tilde{q}_j/c.$$

Choose $s_{it} \in \{0, 1, \ldots, k^{(it)}\}$ according to

$$\Pr(s_{it} = \iota) = q_\iota.$$

If $s_{it} = 0$, draw $\theta_{it} = (\mu_{it}, \sigma_{it}^2)$ as follows:

$$\sigma_{it}^{-2} \sim \chi^2(4)/[3(.01) + \left(\frac{.01/4}{.01/4 + 1}\right) \epsilon_{it}^2],$$

$$\mu_{it} \sim \mathcal{N}((.01/4 + 1)^{-1}\epsilon_{it}, \sigma_{it}^2/(.01/4 + 1)).$$

If $s_{it} \neq 0$, set

$$\theta_{it} = \xi_{s_{it}}^{(it)}.$$

After cycling through all (i, t) pairs in this fashion, redraw for the cluster locations. Use $\xi_j \equiv (\mu_j, \sigma_j^2)$, $j = 1, \ldots, k$, to denote the k distinct values for θ_{it} (now ranging over all i, t). Define $S_{it} = j$ if $\theta_{it} = \xi_j$. Let $n_j = \#\{S_{it} = j\}$. For each $j = 1, \ldots, k$, draw for σ_j^{-2} and μ_j according to:

$$\sigma_j^{-2} \sim \chi^2(3 + n_j)/[3(.01) + \sum_{(i,t):S_{it}=j} (\epsilon_{it} - \bar{\epsilon}_j)^2 + \frac{(.01/4)n_j}{(.01/4) + n_j}\bar{\epsilon}_j^2],$$

where

$$\bar{\epsilon}_j = n_j^{-1} \sum_{(i,t):S_{it}=j} \epsilon_{it},$$

and draw

$$\mu_j \sim \mathcal{N}(\mu_j^*, H^{-1}),$$

where

$$H = \sigma_j^{-2}(.01/4 + n_j)$$

$$\mu_j^* = H^{-1}\sigma_j^{-2} \sum_{(i,t):S_{it}=j} \epsilon_{it}.$$

Then set $\theta_{it} = \xi_{S_{it}}$ for all i, t.

(α): Define k to be the number of distinct values for θ_{it} as in the previous block. Draw

$$\eta \sim \text{Beta}(\alpha + 1, n(T-1)),$$

using the current value for α. Then form

$$\pi_x = (2 + k - 1)/[n(T-1)(.5 - \log(\eta))],$$

$$\pi_\eta = \pi_x/[1 + \pi_x].$$

With probability π_η, draw

$$\alpha \sim \mathcal{G}(2 + k, .5 - \log(\eta));$$

otherwise draw

$$\alpha \sim \mathcal{G}(2 + k - 1, .5 - \log(\eta)).$$

Parametric Correlated Random Effects

Blocks: (ρ, τ), (γ), (ψ, Ω^{-1}).

(ρ, τ): Define

$$\hat{\rho} = \sum_{i=1}^{n}\sum_{t=2}^{T} y_{i,t-1}(y_{it} - \gamma_i) / \sum_{i=1}^{n}\sum_{t=2}^{T} y_{i,t-1}^2.$$

Then draw

$$\tau \sim \chi^2(n(T-1))/[.01 + \sum_{i=1}^{n}\sum_{t=2}^{T}(y_{it} - \gamma_i - \hat{\rho} y_{i,t-1})^2],$$

$$\rho \sim \mathcal{N}(\hat{\rho}, (\tau \sum_{i=1}^{n}\sum_{t=2}^{T} y_{i,t-1}^2)^{-1}).$$

(γ_i): Define

$$H = (\Omega^{-1} + (T-1)\tau)$$

and

$$\gamma_i^* = H^{-1}[\Omega^{-1}\psi y_{i1} + \tau \sum_{t=2}^{T}(y_{it} - \rho y_{i,t-1})].$$

Then

$$\gamma_i \sim \mathcal{N}(\gamma_i^*, H^{-1}).$$

(ψ, Ω^{-1}): Define

$$\hat{\psi} = \sum_{i=1}^{n} \gamma_i y_{i1} / [\sum_{i=1}^{n} y_{i1}^2].$$

Then

$$\Omega^{-1} \sim \chi^2(n)/[.01 + \sum_{i=1}^{n}(\gamma_i - \hat{\psi} y_{i1})^2],$$

$$\psi \sim \mathcal{N}(\hat{\psi}, \Omega(\sum_{i=1}^{n} y_{i1}^2)^{-1}).$$

Correlated Random Effects with Nonparametric Errors

Blocks: (ρ), $(\gamma_1, \ldots, \gamma_n)$, (ψ, Ω), (θ), (α).

(ρ): Define

$$\hat{\rho} = [\sum_{i=1}^{n}\sum_{t=2}^{T} \sigma_{it}^{-2} y_{i,t-1}(y_{it} - \gamma_i - \mu_{it})]/[\sum_{i=1}^{n}\sum_{t=2}^{T} \sigma_{it}^{-2} y_{i,t-1}^2].$$

Then

$$\rho \sim \mathcal{N}(\hat{\rho}, (\sum_{i=1}^{n}\sum_{t=2}^{T} \sigma_{it}^{-2} y_{i,t-1}^2)^{-1}).$$

(γ_i):

$$H_i = (\Omega^{-1} + \sum_{t=2}^{T} \sigma_{it}^{-2})$$

$$\gamma_i^* = H_i^{-1}[\Omega^{-1}\psi y_{i1} + \sum_{t=2}^{T} \sigma_{it}^{-2}(y_{it} - \rho y_{i,t-1} - \mu_{it})].$$

Then

$$\gamma_i \sim \mathcal{N}(\gamma_i^*, H_i^{-1}).$$

(ψ, Ω^{-1}): Define

$$\hat{\psi} = \sum_{i=1}^{n} \gamma_i y_{i1} / [\sum_{i=1}^{n} y_{i1}^2].$$

Then

$$\Omega^{-1} \sim \chi^2(n)/[.01 + \sum_{i=1}^{n}(\gamma_i - \hat{\psi} y_{i1})^2],$$

$$\psi \sim \mathcal{N}(\hat{\psi}, \Omega(\sum_{i=1}^{n} y_{i1}^2)^{-1}).$$

(θ): same as in semiparametric model with uncorrelated random effects.

(α): same as in semiparametric model with uncorrelated random effects.

2.6.5 Simulating Predictive Distributions in the Autoregressive Models

We are interested in simulating the predictive distribution of $(y_{i,T+1}, \ldots, y_{i,T+H})$. By recursive substitution,

$$y_{i,T+h} = (1 + \rho + \cdots + \rho^{h-1})\gamma_i + \rho^h y_{iT} + \epsilon_{i,T+h} + \rho\epsilon_{i,T+h-1} + \cdots + \rho^{h-1}\epsilon_{i,T+1}.$$

So in the models with $\epsilon_{it} \sim \mathcal{N}(0, \tau^{-1})$, we can write

$$\begin{pmatrix} y_{i,T+1} \\ \vdots \\ y_{i,T+H} \end{pmatrix} \sim \mathcal{N}(\mu_i, \Sigma_i),$$

where

$$\mu_i = \begin{pmatrix} 1 \\ \vdots \\ \sum_{h=0}^{H-1} \rho^h \end{pmatrix} \gamma_i + \begin{pmatrix} \rho \\ \vdots \\ \rho^H \end{pmatrix} y_{iT},$$

$$A = \begin{bmatrix} 1 & & & & 0 \\ \rho & 1 & & & \\ \rho^2 & \rho & 1 & & \\ \vdots & & & \ddots & \\ & & & & 1 \end{bmatrix},$$

and

$$\Sigma_i = AA'/\tau.$$

This suggests the following method for evaluating the predictive distributions using Monte Carlo. Let $(\rho^{(j)}, \tau^{(j)}, \gamma_i^{(j)})$ denote draws for the parameters from the jth iteration of the Gibbs sampling algorithm, after discarding some initial set of iterations to allow for burn-in. Form $\mu_i^{(j)}$ and $\Sigma_i^{(j)}$ according to the previous expressions, and construct the predictive distribution as:

$$F(y_{i,T+1}, \dots, y_{i,T+H}|z) \approx \frac{1}{J} \sum_{j=1}^{J} \Phi(y_{i,T+1}, \dots, y_{i,T+H}|\mu_i^{(j)}, \Sigma_i^{(j)}).$$

where $\Phi(\cdot|\mu, \Sigma)$ denotes the distribution function of a multivariate normal random variable with mean vector μ and variance matrix Σ.

In the model with nonparametric errors, some additional work is needed in order to efficiently simulate from the distribution of ϵ_{it}. Conditional on $\theta_{i,T+1}, \dots, \theta_{i,T+H}$, we can write

$$\begin{pmatrix} y_{i,T+1} \\ \vdots \\ y_{i,T+H} \end{pmatrix} \sim \mathcal{N}(\mu_i, \Sigma_i),$$

where

$$\mu_i = \begin{pmatrix} 1 \\ \vdots \\ \sum_{l=0}^{H-1} \rho^l \end{pmatrix} \gamma_i + \begin{pmatrix} \rho \\ \vdots \\ \rho^H \end{pmatrix} y_{iT} + A \begin{pmatrix} \mu_{i,T+1} \\ \vdots \\ \mu_{i,T+H} \end{pmatrix},$$

$$\Sigma_i = A \begin{bmatrix} \sigma^2_{i,T+1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{i,T+H} \end{bmatrix} A',$$

and A is as in the the parametric case.

At each step j of the Gibbs sampler, recursively draw for $\theta_{i,T+1}, \ldots, \theta_{i,T+H}$, conditional on the current draw for α and $\{\theta_{it} : i = 1, \ldots, n, t = 2, \ldots, T\}$, according to

$$\theta^{(j)}_{i,T+1} \sim \frac{\alpha^{(j)}}{\alpha^{(j)} + n(T-1)} P_0 + \frac{1}{\alpha^{(j)} + n(T-1)} \sum_{l=1}^{n} \sum_{s=2}^{T} \delta(\theta^{(j)}_{ls}),$$

and for $h = 2, \ldots, H$,

$$\begin{aligned} \theta^{(j)}_{i,T+h} \sim & \frac{\alpha^{(j)}}{\alpha^{(j)} + n(T-1) + h - 1} P_0 \\ & + \frac{1}{\alpha^{(j)} + n(T-1) + h - 1} \sum_{l=1}^{n} \sum_{s=2}^{T} \delta(\theta^{(j)}_{ls}) \\ & + \frac{1}{\alpha^{(j)} + n(T-1) + h - 1} \sum_{s=T+1}^{h-1} \delta(\theta^{(j)}_{is}). \end{aligned}$$

So there will be a positive probability that $\theta^{(j)}_{i,T+1} = \theta^{(j)}_{i,T+2}$, for example. Then form $\mu^{(j)}_i$ and $\Sigma^{(j)}_i$ according to the expressions above, and approximate the predictive distribution by

$$F(y_{i,T+1}, \ldots, y_{i,T+H}|z) \approx \frac{1}{J}\sum_{j=1}^{J} \Phi(y_{i,T+1}, \ldots, y_{i,T+H}|\mu_i^{(j)}, \Sigma_i^{(j)}).$$

2.6.6 MCMC Algorithms for the Error Components Models

Parametric Error Components Model

Define $v_i \equiv (v_{i1}, \ldots, v_{iT})'$, and $v \equiv (v_1, \ldots, v_n)$.

Blocks: (σ_v^{-2}), (ρ, σ_w^{-2}), (σ^{-2}), $(v_1, \ldots, v_n)$.

(σ_v):

$$\sigma_v^{-2} \sim \chi^2(1+n)/[.01 + \sum_{i=1}^{n} v_{i1}^2]$$

$(\rho,\ \sigma_w)$: Define

$$\hat{\rho} = \sum_{i=1}^{n}\sum_{t=2}^{T} v_{it}v_{i,t-1}/[\sum_{i=1}^{n}\sum_{t=2}^{T} v_{i,t-1}^2]$$

Then

$$\sigma_w^{-2} \sim \chi^2(n(T-1))/[.01 + \sum_{i=1}^{n}\sum_{t=2}^{T}(v_{it} - \hat{\rho}v_{i,t-1})^2]$$

$$\rho \sim \mathcal{N}(\hat{\rho}, \sigma_w^2(\sum_{i=1}^{n}\sum_{t=2}^{T} v_{i,t-1}^2)^{-1})$$

(σ):

$$\sigma^{-2} \sim \chi^2(nT+1)/\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - v_{it})^2$$

(v_i): Define $y_i \equiv (y_{i1}, \ldots, y_{iT})'$. By examining the kernel of the posterior distribution for terms involving v_i and simplifying the resulting expression to get the kernel of a multivariate normal density:

$$v_i \sim \mathcal{N}(v_i^*, H^{-1}),$$

where

$$H = \sigma^{-2} I_T + \Lambda^{-1}$$

$$\Lambda = A \text{ diag}\{\sigma_v^2, \sigma_w^2, \ldots, \sigma_w^2\} A'$$

$$A = \begin{bmatrix} 1 & & & & 0 \\ \rho & 1 & & & \\ \rho^2 & \rho & 1 & & \\ \vdots & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

and

$$v_i^* = H^{-1} \sigma^{-2} y_i.$$

Semiparametric Error Components Model

Blocks: σ_v^{-2}, (ρ, σ_w^{-2}), $(v_1, \ldots, v_n)$, (θ), (α).

(σ_v^{-2}):

$$\sigma_v^{-2} \sim \chi^2(1+n)/[.01 + \sum_{i=1}^n v_{i1}^2]$$

(ρ, σ_w^{-2}): Define

$$\hat{\rho} = \sum_{i=1}^n \sum_{t=2}^T v_{it} v_{i,t-1} / [\sum_{i=1}^n \sum_{t=2}^T v_{i,t-1}^2]$$

Then

$$\sigma_w^{-2} \sim \chi^2(n(T-1))/[.01 + \sum_{i=1}^n \sum_{t=2}^T (v_{it} - \hat{\rho} v_{i,t-1})^2]$$

$$\rho \sim \mathcal{N}(\hat{\rho}, \sigma_w^2 (\sum_{i=1}^n \sum_{t=2}^T v_{i,t-1}^2)^{-1})$$

(v_i):

$$v_i \sim \mathcal{N}(v_i^*, H^{-1}),$$

where

$$\Sigma_i = \ \text{diag}\{\sigma_{i1}^2, \ldots, \sigma_{iT}^2\},$$

Λ is defined as in the parametric error components model,

$$H = \Sigma^{-1} + \Lambda^{-1},$$

and

$$v_i^* = H^{-1}\Sigma_i^{-1}(y_{i1} - \mu_{i1}, \ldots, y_{iT} - \mu_{iT})'.$$

(θ_{it}): Define $\epsilon_{it} = y_{it} - v_{it}$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$. For each pair (i, t), apply the following steps: Let $\xi_j^{(it)}$, $j = 1, \ldots, k^{(it)}$, denote the $k^{(it)}$ distinct values of $\{\theta_{ls} : (l, s) \neq (i, t)\}$, and define $n_j^{(it)} = \#\{\theta_{ls} : \theta_{ls} = \xi_j^{(it)}, (l, s) \neq (i, t)\}$. Calculate

$$\tilde{q}_0 = \alpha \cdot t_s(\epsilon_{it}|m, Q(b+1)),$$

$$\tilde{q}_j = n_j^{(it)}\phi(\epsilon_{it}|\xi_j^{(it)}),$$

and

$$c = \tilde{q}_0 + \sum_{j=1}^{k^{(it)}} \tilde{q}_j.$$

Then form

$$q_0 = \tilde{q}_0/c, \qquad q_j = \tilde{q}_j/c.$$

Choose $s_{it} \in \{0, 1, \ldots, k^{(it)}\}$ according to

$$\Pr(s_{it} = \iota) = q_\iota.$$

If $s_{it} = 0$, draw $\theta_{it} = (\mu_{it}, \sigma^2_{it})$ as follows:

$$\sigma_{it}^{-2} \sim \chi^2(4)/[3(.01) + \left(\frac{.01/4}{.01/4+1}\right)\epsilon_{it}^2],$$

$$\mu_{it} \sim \mathcal{N}((.01/4+1)\epsilon_{it}, \sigma^2_{it}/(.01/4+1)).$$

If $s_{it} \neq 0$, set

$$\theta_{it} = \xi^{(it)}_{s_{it}}.$$

After cycling through all (i,t) pairs in this fashion, redraw for the cluster locations. Use $\xi_j \equiv (\mu_j, \sigma_j^2)$, $j = 1, \ldots, k$, to denote the k distinct values for θ_{it} (now ranging over all i,t). Define $S_{it} = j$ if $\theta_{it} = \xi_j$. Let $n_j = \#\{S_{it} = j\}$. For each $j = 1, \ldots, k$, draw for σ_j^{-2} and μ_j according to:

$$\sigma_j^{-2} \sim \chi^2(3+n_j)/[3(.01) + \sum_{(i,t):S_{it}=j} (\epsilon_{it} - \bar{\epsilon}_j)^2 + \frac{(.01/4)n_j}{(.01/4)+n_j}\bar{\epsilon}_j^2],$$

where

$$\bar{\epsilon}_j = n_j^{-1} \sum_{(i,t):S_{it}=j} \epsilon_{it},$$

and

$$\mu_j \sim \mathcal{N}(\mu_j^*, H^{-1}),$$

where

$$H = \sigma_j^{-2}(.01/4 + n_j)$$

$$\mu_j^* = H^{-1}\sigma_j^{-2} \sum_{(i,t):S_{it}=j} \epsilon_{it}.$$

Then set $\theta_{it} = \xi_{S_{it}}$ for all i,t.

(α): Define k to be the number of distinct values for θ_{it} as in the previous block. Draw

$$\eta \sim \text{Beta}(\alpha+1, n \cdot T),$$

using the current value for α. Then form

$$\pi_x = (2+k-1)/[n \cdot T(.5 - \log(\eta))],$$

$$\pi_\eta = \pi_x/[1+\pi_x].$$

With probability π_η, draw

$$\alpha \sim \mathcal{G}(2+k, .5 - \log(\eta));$$

otherwise draw

$$\alpha \sim \mathcal{G}(2+k-1, .5 - \log(\eta)).$$

2.6.7 Predictive Distributions in the Error Components Models

In the parametric version of the error components model, use recursive substitution to write

$$y_{i,T+h} = \rho^h v_{iT} + w_{i,T+h} + \rho w_{i,T+h-1} + \cdots + \rho^{h-1} w_{i,T+1} + \epsilon_{i,T+h}.$$

So we can write

$$\begin{pmatrix} y_{i,T+1} \\ \vdots \\ y_{i,T+H} \end{pmatrix} \sim \mathcal{N}(\mu_i, \Sigma_i),$$

where

$$\mu_i = \begin{pmatrix} \rho \\ \vdots \\ \rho^H \end{pmatrix} v_{iT},$$

and

$$\Sigma_i = \sigma_w^2 AA' + \sigma^2 I,$$

and A is as defined for the autoregressive random effects models. To construct the Monte Carlo approximation to the predictive distribution, let $(\rho^{(j)}, v_{i,T}^{(j)}, \sigma_w^{(j)2}, \sigma^{(j)2})$ denote the jth draw from the Gibbs sampling algorithm, after discarding some initial set of iterations to allow for burn-in. For each draw, form $\mu_i^{(j)}$, $\Sigma_i^{(j)}$, $A^{(j)}$, and construct the predictive distribution as

$$F(y_{i,T+1}, \ldots, y_{i,T+H}|z) \approx \frac{1}{J}\sum_{j=1}^{J} \Phi(y_{i,T+1}, \ldots, y_{i,T+H}|\mu_i^{(j)}, \Sigma_i^{(j)}).$$

In the semiparametric case the distribution of future earnings, conditional on $\theta_{i,T+1}, \ldots,$ $\theta_{i,T+H}$, v_{iT}, ρ, and σ_w, is

$$\begin{pmatrix} y_{i,T+1} \\ \vdots \\ y_{i,T+H} \end{pmatrix} \sim \mathcal{N}(\mu_i, \Sigma_i),$$

where

$$\mu_i = \begin{pmatrix} \rho \\ \vdots \\ \rho^H \end{pmatrix} v_{iT} + \begin{pmatrix} \mu_{i,T+1} \\ \vdots \\ \mu_{i,T+H} \end{pmatrix},$$

$$\Sigma_i = AA'\sigma_w^2 + \begin{bmatrix} \sigma_{i,T+1}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{i,T+H}^2 \end{bmatrix}.$$

and A is as defined earlier. So to form a Monte Carlo approximation to the predictive density, recursively draw for $\theta_{i,T+1}, \ldots, \theta_{i,T+H}$, conditional on the current draw for α and $\{\theta_{it} : i = 1, \ldots, n, t = 2, \ldots, T\}$, according to

$$\theta_{i,T+1}^{(j)} \sim \frac{\alpha^{(j)}}{\alpha^{(j)} + nT} P_0 + \frac{1}{\alpha^{(j)} + nT} \sum_{l=1}^{n}\sum_{s=1}^{T} \delta(\theta_{ls}^{(j)}),$$

and for $h = 2, \ldots, H$,

$$\begin{aligned}\theta_{i,T+h}^{(j)} \sim \; & \frac{\alpha^{(j)}}{\alpha^{(j)} + nT + h - 1} P_0 \\ & + \frac{1}{\alpha^{(j)} + nT + h - 1} \sum_{l=1}^{n} \sum_{s=1}^{T} \delta(\theta_{ls}^{(j)}) \\ & + \frac{1}{\alpha^{(j)} + nT + h - 1} \sum_{s=T+1}^{h-1} \delta(\theta_{is}^{(j)}).\end{aligned}$$

Then form $\mu_i^{(j)}$ and $\Sigma_i^{(j)}$, and approximate the predictive distribution by

$$F(y_{i,T+1}, \ldots, y_{i,T+H}|z) \approx \frac{1}{J} \sum_{j=1}^{J} \Phi(y_{i,T+1}, \ldots, y_{i,T+H}|\mu_i^{(j)}, \Sigma_i^{(j)}).$$

2.7 Additional Tables and Figures

2.7.1 Density Estimates

Figure 2.19: High School Dropouts

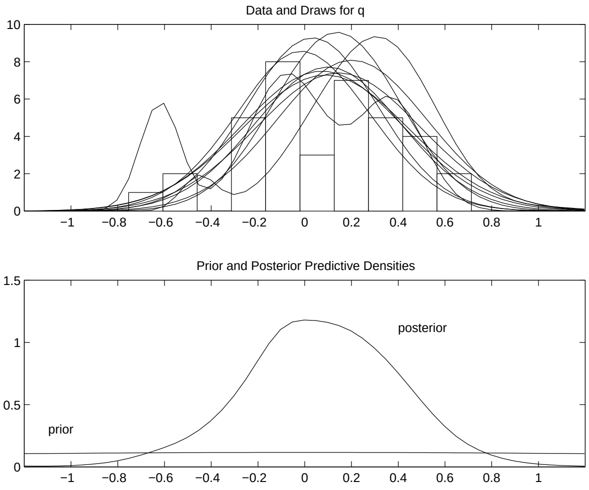
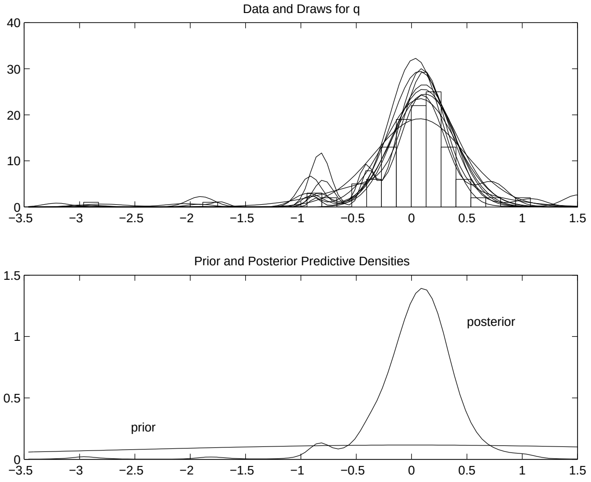
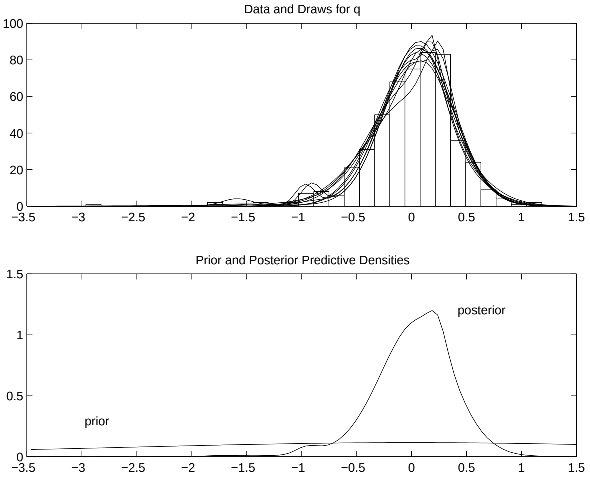


Figure 2.20: College Graduates


Data and Draws for q

Prior and Posterior Predictive Densities

Figure 2.21: Full Sample


Data and Draws for q

Prior and Posterior Predictive Densities

2.7.2 Covariance Matrices

Table 2.5: Covariance Matrix, All Residuals

	$T = 1$	2	3	4	5	6	7	8	9	10
1	0.17	0.10	0.10	0.09	0.10	0.10	0.10	0.09	0.07	0.07
2	0.10	0.14	0.11	0.10	0.09	0.10	0.10	0.08	0.08	0.08
3	0.10	0.11	0.16	0.13	0.11	0.12	0.11	0.11	0.10	0.10
4	0.09	0.10	0.13	0.18	0.13	0.13	0.12	0.12	0.11	0.10
5	0.10	0.09	0.11	0.13	0.20	0.14	0.14	0.13	0.12	0.10
6	0.10	0.10	0.12	0.13	0.14	0.23	0.17	0.14	0.13	0.11
7	0.10	0.10	0.11	0.12	0.14	0.17	0.21	0.17	0.15	0.12
8	0.09	0.08	0.11	0.12	0.13	0.14	0.17	0.22	0.17	0.15
9	0.07	0.08	0.10	0.11	0.12	0.13	0.15	0.17	0.21	0.17
10	0.07	0.08	0.10	0.10	0.10	0.11	0.12	0.15	0.17	0.23

Table 2.6: Covariance Matrix, High School Dropouts

	$T = 1$	2	3	4	5	6	7	8	9	10
1	0.11	0.06	0.05	0.05	0.05	0.07	0.04	0.05	0.02	0.04
2	0.06	0.15	0.09	0.04	0.06	0.08	0.07	0.08	0.06	0.05
3	0.05	0.09	0.09	0.06	0.04	0.06	0.04	0.05	0.06	0.05
4	0.05	0.04	0.06	0.49	0.20	0.15	0.12	0.15	0.11	0.15
5	0.05	0.06	0.04	0.20	0.22	0.14	0.16	0.17	0.11	0.12
6	0.07	0.08	0.06	0.15	0.14	0.21	0.14	0.13	0.07	0.08
7	0.04	0.07	0.04	0.12	0.16	0.14	0.17	0.17	0.10	0.11
8	0.05	0.08	0.05	0.15	0.17	0.13	0.17	0.20	0.14	0.14
9	0.02	0.06	0.06	0.11	0.11	0.07	0.10	0.14	0.16	0.14
10	0.04	0.05	0.05	0.15	0.12	0.08	0.11	0.14	0.14	0.16

Table 2.7: Covariance Matrix, High School Graduates

	$T = 1$	2	3	4	5	6	7	8	9	10
1	0.16	0.09	0.09	0.08	0.09	0.10	0.11	0.10	0.09	0.08
2	0.09	0.13	0.09	0.10	0.08	0.08	0.09	0.08	0.08	0.08
3	0.09	0.09	0.15	0.12	0.09	0.09	0.11	0.12	0.10	0.10
4	0.08	0.10	0.12	0.20	0.14	0.11	0.13	0.14	0.13	0.12
5	0.09	0.08	0.09	0.14	0.19	0.12	0.15	0.15	0.14	0.11
6	0.10	0.08	0.09	0.11	0.12	0.32	0.20	0.14	0.11	0.09
7	0.11	0.09	0.11	0.13	0.15	0.20	0.25	0.20	0.16	0.13
8	0.10	0.08	0.12	0.14	0.15	0.14	0.20	0.25	0.20	0.15
9	0.09	0.08	0.10	0.13	0.14	0.11	0.16	0.20	0.23	0.17
10	0.08	0.08	0.10	0.12	0.11	0.09	0.13	0.15	0.17	0.19

Table 2.8: Covariance Matrix, College Graduates

	$T = 1$	2	3	4	5	6	7	8	9	10
1	0.21	0.14	0.14	0.13	0.12	0.14	0.11	0.10	0.08	0.09
2	0.14	0.16	0.13	0.12	0.10	0.13	0.11	0.10	0.09	0.09
3	0.14	0.13	0.20	0.15	0.13	0.17	0.14	0.14	0.13	0.12
4	0.13	0.12	0.15	0.16	0.13	0.15	0.13	0.12	0.11	0.11
5	0.12	0.10	0.13	0.13	0.15	0.17	0.14	0.12	0.10	0.10
6	0.14	0.13	0.17	0.15	0.17	0.26	0.18	0.16	0.14	0.14
7	0.11	0.11	0.14	0.13	0.14	0.18	0.20	0.16	0.15	0.14
8	0.10	0.10	0.14	0.12	0.12	0.16	0.16	0.20	0.17	0.17
9	0.08	0.09	0.13	0.11	0.10	0.14	0.15	0.17	0.20	0.19
10	0.09	0.09	0.12	0.11	0.10	0.14	0.14	0.17	0.19	0.23