Smoothly Mixing Regressions

John Geweke and Michael Keane
Departments of Economics and Statistics, University of Iowa
john-geweke@uiowa.edu
Department of Economics, Yale University
michael.keane@yale.edu

November 15, 2005

Abstract

This paper extends the conventional Bayesian mixture of normals model by permitting state probabilities to depend on observed covariates. The dependence is captured by a simple multinomial probit model. A conventional and rapidly mixing MCMC algorithm provides access to the posterior distribution at modest computational cost. This model is competitive with existing econometric models, as documented in the paper's illustrations. The first illustration studies quantiles of the distribution of earnings of men conditional on age and education, and shows that smoothly mixing regressions are an attractive alternative to non-Baeysian quantile regression. The second illustration models serial dependence in the S&P 500 return, and shows that the model compares favorably with ARCH models using out of sample likelihood criteria.

Acknowledgement 1 *Both authors acknowledge financial support from grant R01-HD37060-01, National Institutes of Health, and the first author acknowledges financial support from grants SBR-9819444 and SBR-0214303, National Science Foundation. We thank two referees, Giovanni Amisano, Justin Tobias, and participants in seminars at Brescia University, Iowa State University, University of Iowa, and University of Rome for helpful comments on previous versions of this work. Responsibility for any errors remains with the authors.*

The linear model has always held a special place at the core of econometric theory and practice. The founders of the discipline, in Rotterdam and at the Cowles Commission, attacked the problem of providing an empirical foundation for economic policy using models in which relationships between the essential variables were linear. As the profession has grown in sophistication and ambition models have become more elaborate, but to a striking degree this has been achieved by using transformations of linear structures. This central position of the linear model can be ascribed in part to its simplicity, familiarity and ease of interpretation, characteristics that most introductory texts and courses in econometrics seek to convey.

Much of econometrics, whether old or recent, is concerned with the relationship of a vector of relevant covariates $\mathbf{x}$ to a vector of outcomes $\mathbf{y}$, the nature of the relationship being mediated to lesser or greater degree by economic theory. Fifty years ago theory and practice addressed the regression $\mathrm{f}(\mathbf{x}, \boldsymbol{\theta}) = \mathrm{E}(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta}$ being a vector of structural parameters. In the simultaneous equation work of the Cowles Commission (Koopmans (1950); Hood and Koopmans (1953)) f is linear in $\mathbf{x}$ but not $\boldsymbol{\theta}$; in the Rotterdam model (Barten (1964); Theil (1967)) and its descendents f is often linear in both $\mathbf{x}$ and $\boldsymbol{\theta}$, with linear and nonlinear constraints on $\boldsymbol{\theta}$. While progress in the intervening fifty years can be measured in many dimensions, two are important to the topic of this study. First, when economic theory tightly mediates $\mathrm{f}(\mathbf{x}, \boldsymbol{\theta})$ the function is typically nonlinear in $\mathbf{x}$, $\boldsymbol{\theta}$ or both, and when it does not the assumption of linearity can often be overturned by the evidence. In both cases econometric theory and practice have risen to the challenge, typically through appropriate modification of the linear model at the core of the discipline. Second, answering interesting questions in economics often requires the entire conditional distribution $\mathrm{p}(\mathbf{y} \mid \mathbf{x})$: this is nearly always the case in any problem involving decision making with a relevant loss function, from macroeconomic policy to the pricing of options to the evaluation of economic welfare.

This study addresses the question of inference for $\mathrm{p}(y \mid \mathbf{x})$ when data are plentiful, economic theory does not constrain p, and y is univariate. It proposes a model for the full conditional density, taking linear models and latent variables structures well-established in econometrics and combining them in a new way. In so doing it is related to several branches of the econometric and statistical literature, including nonparametric and quantile regression as well as mixture models. Section 1 discusses these relations. Section 2 provides a full exposition of the model, with details of inference relegated to an appendix. The following two sections illustrate how the model answers the questions it is designed to address. Section 3 uses data from the panel survey of income dynamics to infer the distribution of earnings conditional on age and education, a relation that has been a focal point of the quantile regression literature. Section 4 applies the model to the distribution of asset returns conditional on the history of returns, using a decade of Standard and Poors 500 data. It applies the model to the assessment of value at risk and presents evidence that the model compares well with other approaches to asset return modeling. The concluding section

describes some of the key outstanding challenges in conditional distribution modeling.

1 Modeling conditional distributions

Conditional distributions arise at the heart of economic analysis. Formal decision-making typically incorporates the distribution of a random variable y that is unknown at the time the decision is made, conditional on the value of a vector of covariates $\mathbf{x}$ that are known. This distribution is conventionally represented by the conditional probability density function $\mathrm{p}(y \mid \mathbf{x})$. In a related but less formal context, the econometrician may reasonably be asked about many properties of this conditional distribution and expected to provide answers that are reliable and internally consistent. In both contexts the *entire conditional distribution* is required. We emphasize this point here because the model and methods in this study can be used for any functional of $\mathrm{p}(y \mid \mathbf{x})$, including expected utility and measures of inequality. Yet it is possible to present only a few such functionals in a single study and we choose to concentrate on the moments and quantiles that have been the focus of the econometric literature.

Much of this literature is concerned with particular properties of $\mathrm{p}(y \mid \mathbf{x})$, the leading example being the regression function $\mathrm{E}(y \mid x)$ for univariate x. Non-Bayesian nonparametric methods are now standard topics in graduate econometric and statistics courses and components of the econometrician's toolbox; standard references include Härdle (1990) and Green and Silverman (1994). Härdle and Tsybakov (1997) and Fan and Yao (1998) extend these methods to $\mathrm{var}(y \mid x)$, but we are not aware of any treatment of $\mathrm{p}(y \mid x)$ using the same methods. Bayesian foundations for smoothing splines originate with Wahba (1978) and Shiller (1984), with extensions to semiparametric approaches for multivariate $\mathbf{x}$ by Smith and Kohn (1996), Koop and Poirier (2004) and Geweke (2005, Section 5.4.1) among others. In both Bayesian and non-Bayesian approaches a fully nonparametric treatment of $\mathrm{E}(y \mid \mathbf{x})$ for multivariate $\mathbf{x}$ raises new but surmountable issues: in the Bayesian literature, for example, see the work of Ruggiero (1994) on hazard models. Of particular relevance for the work here are Voronoi tesellation methods for spline smoothing (Green and Sibson, 1978), for which Bayesian methods have recently been developed by Holmes et al. (2005); we return to these methods in Section 2.5. To the extent this literature addresses $\mathrm{p}(y \mid \mathbf{x})$ it typically does so using strongly parametric models for $\varepsilon = y - \mathrm{E}(y \mid \mathbf{x})$; a notable exception is Smith and Kohn (1996). Müller and Quintana (2004) provide an accessible review of the Bayesian nonparametric regression literature.

Koenker and Bassett (1978) note that if $P[y \leq \mathrm{f}_q(\mathbf{x})] = q$, then $\mathrm{f}_q(\mathbf{x})$ minimizes the expectation of

$$(1-q)\,|y - \mathrm{f}_q(\mathbf{x})|\, I_{(-\infty,\mathrm{f}_q(\mathbf{x}))}(y) + q\,|y - \mathrm{f}_q(\mathbf{x})|\, I_{(\mathrm{f}_q(\mathbf{x}),\infty)}(y).$$

If $\mathrm{f}_q(\mathbf{x}) = \mathrm{f}\left(\mathbf{x}, \boldsymbol{\theta}_q\right)$ then

$$\widehat{\boldsymbol{\theta}}_q = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \left[(1-q)\left|y - \mathrm{f}_q(\mathbf{x}, \boldsymbol{\theta})\right| I_{(-\infty, \mathrm{f}(\mathbf{x}, \boldsymbol{\theta}))}(y) + q\left|y - \mathrm{f}_q(\mathbf{x}, \boldsymbol{\theta})\right| I_{(\mathrm{f}(\mathbf{x}, \boldsymbol{\theta}), \infty)}(y)\right]$$

is a consistent estimator of $\boldsymbol{\theta}_q$ and $\mathrm{f}\left(\mathbf{x}, \widehat{\boldsymbol{\theta}}_q\right)$ is a consistent estimator of quantile q of $\mathrm{p}(y \mid \mathbf{x})$. Quantile regression, as this procedure has come to be known, accomplishes some of the objectives illustrated in Sections 3 and 4, but it is more limited in two respects. First, it seeks only to address (at most) a finite number of quantiles of $\mathrm{p}(y \mid \mathbf{x})$, rather than providing a complete model for the conditional distribution. Second and derivative from this objective, quantile regression does not make use of the restrictions $q < q^* \implies \mathrm{f}_q(\mathbf{x}) \leq \mathrm{f}_{q^*}(\mathbf{x})$, leading to loss of efficiency in finite sample and the accommodation of the contradiction that $\mathrm{f}\left(\mathbf{x}, \widehat{\boldsymbol{\theta}}_q\right) > \mathrm{f}\left(\mathbf{x}, \widehat{\boldsymbol{\theta}}_{q^*}\right)$ for some combinations of $\mathbf{x}$ and $q < q^*$. Buchinsky (1994) takes this approach in an application similar to the one in Section 3; see also Manning et al. (1995) and Angrist et al. (2004). Nonparametric methods for regression functions may also be applied to the functions $\mathrm{f}_q(\mathbf{x})$ themselves (Yu and Jones (1998)).

Over the past decade mixture models have emerged as a practical and theoretically appealing device for flexible specification of the entire conditional distribution, fueled in substantial part by Markov chain Monte Carlo (MCMC) methods for Bayesian inference; see Escobar and West (1995). There is a substantial body of recent work in Bayesian econometrics that concentrates on flexible modeling of $\varepsilon = y - \mathrm{E}(y \mid \mathbf{x})$ under the assumption that ε is i.i.d. and independent of $\mathbf{x}$. Geweke and Keane (2000) uses a finite mixture of normals model for transitory shocks to earnings in an otherwise conventional life-cycle model with panel data, and Hirano (2002) does much the same thing beginning with a Dirichlet process prior centered on a normal distribution. Griffin and Steel (2004) take a similar approach in a stochastic production frontier model, concentrating flexibility on the firm component. Smith and Kohn (1996) combine a mixture of normals specification for the disturbance with nonparametric treatment of the regression but their focus was on robust inference for the conditional means rather than inference for $\mathrm{p}(y \mid \mathbf{x})$.

The assumption that $\varepsilon = y - \mathrm{E}(y \mid \mathbf{x})$ is independent of $\mathbf{x}$ is clearly inappropriate in many applied econometric settings: two decades of work on asset returns (Bollerslev et al. (1992), Jacquier et al. (1994), Kim et al. (1998)) provide spectacular counterexamples, and the econometrics literature (White (1980)) has long emphasized robustness with respect to conditional heteroscedasticity. Mixture models have been extended to build dependence between ε and $\mathbf{x}$. Some of the first instances in econometrics were motivated by overdispersion relative to the canonical Poisson regression (Wedel et al. (1993)) and negative binomial (Deb and Trivedi (1997)) models. Morduch and Stern (1997) applied this idea to conditional distributions of continuous random variables in an approach related to the one taken here. Since the

pioneering work of Ferguson (1973, 1974) mixtures with Dirichlet process priors have been an important tool for modeling unconditional distributions. If this approach is taken to inference about the joint distribution of y and $\mathbf{x}$, one immediately has inference for $\mathrm{p}(y \mid \mathbf{x})$ as a logical byproduct. Examples include Müller et al. (1996) and Chamberlain and Imbens (2003), the latter with an application related to the example in Section 3. Including $\mathbf{x}$ in the modeling exercise substantially increases computational demands. It also requires that the investigator think about $\mathrm{p}(\mathbf{x})$ in situations where $\mathbf{x}$ is otherwise ancillary. Several investigators have recently focused directly on $\mathrm{p}(y \mid \mathbf{x})$ using Dirichlet process priors to enforce the idea that the mapping from $\mathbf{x}$ to $\mathrm{p}(y \mid \mathbf{x})$ is smooth. DeIorio et al. (2004) uses this strategy for categorical covariates, and Griffin and Steel (2005) for continuous covariates. The approach in this literature closest to the one taken here is Dunson and Pillai (2004), who use kernel smoothing of continuous covariates, to which we return in Section 2.5.

2 The model

We turn first to a full statement of the smoothly mixing regression model.

2.1 Normal mixture models

The point of departure for our work is a general statement of the discrete mixture of normals model for T observations $\mathbf{y} = (y_1, \ldots, y_T)'$ of a variable of interest and T corresponding observations of covariates $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]'$. Linear and nonlinear transformations of the $n \times 1$ vectors $\mathbf{x}_t$, for example those used to create polynomials, produce the $T \times k$ matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_T]'$ and $T \times p$ matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_T]'$. Corresponding to each observation there is a latent variable $\widetilde{s}_t$ taking on one of the values $1, \ldots, m$ $(m \geq 2)$ and then

$$y_t \mid (\mathbf{u}_t, \mathbf{v}_t, \widetilde{s}_t = j) \sim N\left(\boldsymbol{\beta}'\mathbf{u}_t + \boldsymbol{\alpha}_j'\mathbf{v}_t, \sigma_j^2\right) \quad (j = 1, \ldots, m). \tag{1}$$

This expression is a component of all the models considered in this article. In all cases $\mathbf{u}_t$ is a $k \times 1$ vector of observed covariates and $\mathbf{v}_t$ is a $p \times 1$ vector of observed covariates. The first component of these vectors is always the constant term $u_{t1} = 1$ or $v_{t1} = 1$ respectively. In the conventional mixture of normals model, $\widetilde{s}_t$ is independent and identically distributed and independent of $\mathbf{u}_t$ and $\mathbf{v}_t$, with

$$P(\widetilde{s}_t = j \mid \mathbf{u}_t, \mathbf{v}_t) = \pi_j \ \ (j = 1, \ldots, m). \tag{2}$$

If $p = 1$, then (1)-(2) amount to a linear regression model in which the disturbance term is i.i.d. and distributed as a normal mixture. (Note that β_1 and $\alpha_1, \ldots, \alpha_m$ are unidentified in this case. Section 2.3 returns to identification and related matters.) This accommodates a wide variety of distributions even when m is small, as illustrated

in Figure 1. If $p = 0$, then the disturbance distribution specializes to a scale mixture of normals, which is symmetric, as illustrated in panels (a) and (f) of Figure 1.

If $k = 1$ and $p > 1$, then the regression in (1)-(2) is $\mathrm{E}(y_t \mid \mathbf{v}_t) = \beta_1 + \sum_{j=1}^{m} \pi_j \boldsymbol{\alpha}_j' \mathbf{v}_t$. The distribution of the population residual $\varepsilon_t = y_t - \mathrm{E}(y_t \mid \mathbf{v}_t)$ is a normal mixture. The shape of the mixture distribution depends on the value of $\mathbf{v}_t$, but the dependence is restricted – it amounts to changes in the location of the component normal densities displayed in Figure 1. In this study, we describe a substantially richer dependence of the distribution on covariates, and illustrate its application in Sections 3 and 4. The case $\mathbf{u}_t = \mathbf{v}_t$, with $k = p > 1$, describes the same family of population models. We do not pursue this case in this study, but it is attractive in formulating a hierarchical prior distribution, as discussed in Section 2.3.

In time series applications model (2) for the latent states can be generalized to the first-order Markov process

$$P(\widetilde{s}_t = j \mid \mathbf{u}_t, \mathbf{v}_t, , \widetilde{s}_{t-1} = i, \widetilde{s}_{t-2}, \ldots, \widetilde{s}_1) = p_{ij}. \tag{3}$$

The combination of (1) and (3), for $\mathbf{u}_t = (1, y_{t-1}, \ldots, y_{t-k+1})$ and $p = 1$, is the Markov normal mixture model (Lindgren (1978); Albert and Chib (1993); Chib(1996)). In this model the distribution of y_t depends indirectly on $y_{t-1}, y_{t-2}, \ldots$ by means of the filtered probabilities $P(\widetilde{s}_t = j \mid y_{t-1}, y_{t-2}, \ldots)$; see Geweke (2005, Section 7.4) for discussion and illustration. In this study we describe an alternative structure for time series and illustrate its application in Section 4. The next section outlines the structure.

2.2 State probabilities

An important potential shortcoming of the mixture model (1)-(2) is that in the typical application with $p = 1$, only the location of the conditional distribution of y_t depends on $\mathbf{x}_t$. This excludes conditional heteroscedasticity, which has been documented in many settings, and other kinds of dependence of $p(y_t \mid \mathbf{x}_t)$ on $\mathbf{x}_t$. In the smoothly mixing regression model the state probabilities $P(\widetilde{s}_t = j)$ depend directly on a $\mathbf{x}_t$, introducing a flexible dependence of the conditional distribution on observables, including conditional heteroscedasticity, in ways to be described shortly.

In order to model dependence of $p(y_t \mid \mathbf{x}_t)$ on $\mathbf{x}_t$, let $\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \cdots & \mathbf{z}_T \end{bmatrix}'$ be a $T \times q$ matrix created from $\mathbf{X}$ by means of the same kinds of linear and nonlinear transformations used to create $\mathbf{U}$ and $\mathbf{V}$ from $\mathbf{X}$. The entire universe of models for a finite number of discrete outcomes provides candidates for the structure of $P(\widetilde{s}_t = j \mid \mathbf{z}_t)$. The subset that is tractable depends on the methods of inference and the nature of the application. We have found that a simplified multinomial probit model works effectively in conjunction with Bayesian MCMC and data sets characteristic of many applications in econometrics. The model can be described in terms of an $m \times 1$ vector of latent states $\widetilde{\mathbf{w}}_t$ corresponding to observation t

$$\widetilde{\mathbf{w}}_t = \boldsymbol{\Gamma} \mathbf{z}_t + \boldsymbol{\zeta}_t, \;\; \boldsymbol{\zeta}_t \overset{iid}{\sim} N(\mathbf{0}, \mathbf{I}_m), \tag{4}$$

and then

$$\widetilde{s}_t = j \text{ iff } \widetilde{w}_{tj} \geq \widetilde{w}_{ti} \;\; (i = 1, \ldots, m)\,. \tag{5}$$

The model (4)-(5) produces simple and tractable conditional posterior distributions in the context of a Gibbs sampling algorithm, as detailed in Section 2.4, and the resulting MCMC algorithm has good mixing properties. Other widely used models for multiple discrete outcomes are less tractable in this setting. One such alternative, a full multinomial probit specification for $\widetilde{s}_t$, would replace the variance matrix $\mathbf{I}_m$ with the $m \times m$ positive definite matrix $\mathbf{\Sigma}$ in (4) together with an appropriate constraint to identify the scale of the latent vector $\widetilde{\mathbf{w}}_t$. This leads to the multimodal likelihood function documented in Keane (1992), compounded here by the fact that the discrete outcome is itself unobserved. A second alternative to (4)-(5) is a multinomial logit model. For the same covariate vector $\mathbf{z}_t$ it has the same number of parameters and similar flexibility in mapping $\mathbf{z}_t$ into state probabilities. However the conditional posterior distributions for the logit model parameters are nonconjugate. In our experience the mixing properties of the resulting MCMC algorithms are inferior when (4)-(5) is replaced with a logit model.

The smoothly mixing regression model is the combination of (1), (4) and (5). The adverb "smoothly" derives from the fact that the conditional state probabilities implied by (4)-(5) are continuous functions of the covariates $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_T]'$. Covariates can enter the smoothly mixing regression model through the vector $\mathbf{u}_t$ if $k > 1$, the vector $\mathbf{v}_t$ if $p > 1$, and the vector $\mathbf{z}_t$ if $q > 1$. In our work we have considered five of the eight cases produced by these three binary contingencies. If $k = p = q = 1$ then y_t is distributed as a mixture of normals independent of covariates. This model is not interesting as a point of reference in our subsequent illustrations, and so we exclude it from further consideration in this study, but it might be useful in other work.

If $q = 1$ then (4)-(5) is a straightforward reparameterization of (2). As previously noted in Section 2.1, if $k > 1$ and $p = 1$ then the model reduces to a linear regression with disturbances distributed as a mixture of normals, while $k = 1$ and $p > 1$ produces a mixture of regressions. We denote these models A and B, respectively. The case $k = p > 1$ with $\mathbf{u}_t = \mathbf{v}_t$ describes the same model. The redundant parameterization can be useful in constructing hierarchical priors, an extension not pursued in this study.

When $q > 1$ then the distinctive features of the smoothly mixing regression model come into play. If, in addition, $k = 1$ and $p = 1$, then the model describes a mixture of normal distributions in which each of the m normal constituents is independent of covariates, but component probabilities move with changes in covariates. We denote this model C. It includes specifications in which distributions but not means depend on covariates, a plausible feature of asset returns in many settings. Section 4 reports evidence of just an instance of model C.

In model D, $k = q > 1$, $p = 1$ and $\mathbf{u}_t = \mathbf{z}_t$. This may be regarded as a textbook linear regression model $y_t = \boldsymbol{\beta}'\mathbf{z}_t + \varepsilon_t$ in which the distribution of the disturbance

term ε_t is a mixture of normals distribution with state probabilities depending on covariates. Note, however, that the regression is now in fact nonlinear, since

$$\mathrm{E}\left(y_t \mid \mathbf{x}_t\right) = \boldsymbol{\beta}'\mathbf{z}_t + \sum_{j=1}^{m} \alpha_j \cdot P\left(\widetilde{s}_t = j \mid \mathbf{z}_t\right).$$

The restriction $\mathbf{u}_t = \mathbf{z}_t$ stems from the nonparametric spirit of the model: the distinction between component conditional probabilities and the means of the component distributions is one of technical convenience rather than of substance.

The richest structure we consider is model E, with $k = 1$, $p = q > 1$ and $\mathbf{v}_t = \mathbf{z}_t$, a covariate-dependent mixture of normal linear regression models. Again the case $\mathbf{u}_t = \mathbf{v}_t = \mathbf{z}_t$ is formally identical and could be used to facilitate hierarchical priors. Table 1 summarizes the five structures we consider in subsequent examples in this study.

2.3 Identification, parameterization and priors

There are several equivalent parameterizations of the smoothly mixing regressions model, and equations (1), (4) and (5) present just one. In this study we use a parameterization that facilitates the expression of prior distributions and also leads to a MCMC posterior simulator with good mixing properties. To facilitate the discussion consider four groups of unobservables in (1) and (4)-(5): the variance parameters σ_j^2 $(j = 1, \ldots, m)$ in (1), the coefficient vector $\boldsymbol{\beta}$ and matrix $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m]$ in (1), the coefficient matrix $\boldsymbol{\Gamma}$ in (4), and the latent variables $\widetilde{\mathbf{s}} = (\widetilde{s}_1 \cdots \widetilde{s}_T)'$ in (1) and (5) and $\widetilde{\mathbf{W}}' = [\widetilde{\mathbf{w}}_1, \ldots, \widetilde{\mathbf{w}}_T]$ in (4)-(5).

It proves convenient to represent the variance by rewriting (1) as

$$y_t \mid \left(\mathbf{u}_t, \mathbf{v}_t, \widetilde{s}_t = j\right) \sim N\left(\boldsymbol{\beta}'\mathbf{u}_t + \boldsymbol{\alpha}_j'\mathbf{v}_t, \sigma^2 \cdot \sigma_j^2\right) \quad (j = 1, \ldots, m), \tag{6}$$

thus representing the variance in state j as $\sigma^2 \cdot \sigma_j^2$. This decomposition makes it possible to separate considerations of scale and shape in thinking about the normal distributions comprising the mixture, in much the same way as in Geweke and Keane (2000). Shape is governed by the state-specific components σ_j^2. The prior distributions of these parameters are independent inverted gamma, and identification is resolved by centering the distributions of $1/\sigma_j^2$ about 1:

$$\underline{\nu}^{*2}/\sigma_j^2 \mid \underline{\nu}^* \overset{iid}{\sim} \chi^2\left(\underline{\nu}^*\right). \tag{7}$$

The investigator chooses the hyperparameter $\underline{\nu}^*$, which governs the thickness of the tails of the conditional distribution of y_t: the smaller the value of $\underline{\nu}^*$ the greater the prior kurtosis of the distribution, whereas this distribution excludes the event of excess kurtosis as $\underline{\nu}^* \to \infty$. The scale of the disturbance is governed by the parameter σ^2. Its conditionally conjugate prior distribution is also inverted gamma,

$$\underline{s}^2/\sigma^2 \mid \left(\underline{s}^2, \underline{\nu}\right) \sim \chi^2\left(\underline{\nu}\right). \tag{8}$$

The hyperparameters $\underline{s}^2$ and $\underline{\nu}$ can be derived by thinking about plausible values for σ^2, for example a centered prior credible interval.

The coefficients in $\boldsymbol{\beta}$, $\mathbf{A}$ and $\boldsymbol{\Gamma}$ all mediate the impact of covariates on the conditional distribution, and therefore prior distributions must be chosen with regard to the specific composition of $\mathbf{u}_t$, $\mathbf{v}_t$ and $\mathbf{z}_t$ respectively. Consider first the case of $\boldsymbol{\beta}$. In both of the illustrations in this study the elements of the covariate vector $\mathbf{u}_t$ are functions of the covariate vector $\mathbf{x}_t = (a_t, b_t)'$ with u_{tj} of the form $a_t^{\ell_a} b_t^{\ell_b}$ ($\ell_a = 0, \ldots, L_a$; $\ell_b = 0, \ldots, L_b$), and so $k = (L_a + 1)(L_b + 1)$. These polynomial basis functions are attractive because they provide flexibility while maintaining the analytical convenience of a model that is linear in $\boldsymbol{\beta}$, conditional on all other parameters. There are other functional forms with the same attraction that also could have been used in this work; see Geweke (2005, section 5.4) for details. To express a prior distribution for $\boldsymbol{\beta}$, construct a grid of points of the form

$$\begin{aligned} G = \{(a_i, b_j) : a_i = a_1^*, a_1^* + \Delta_a, \ldots, a_1^* + (N_a - 1)\Delta_a, a_2^*, \\ b_i = b_1^*, b_1^* + \Delta_b, \ldots, b_1^* + (N_b - 1)\Delta_b, b_2^*\} \end{aligned} \tag{9}$$

where $\Delta_a = (a_2^* - a_1^*)/N_a$ and $\Delta_b = (b_2^* - b_1^*)/N_b$. Corresponding to each point (a_i, b_i) in the grid is a $k \times 1$ vector $\mathbf{c}_i$ with entries of the form $a_i^{\ell_a} b_i^{\ell_b}$; $(a_i, b_i)'$ is mapped into $\mathbf{c}_i$ in the same way that $\mathbf{x}_t$ is mapped into $\mathbf{u}_t$. Arrange these vectors in the $r \times k$ matrix $\mathbf{C} = \begin{bmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_r \end{bmatrix}$, where $r = (N_a + 1)(N_b + 1)$ The form of the prior distribution is then

$$\mathbf{C}\boldsymbol{\beta} \mid \left(\underline{\mu}, \underline{\tau}_{\boldsymbol{\beta}}^2\right) \sim N\left(\boldsymbol{\iota}_r \underline{\mu}, \underline{\tau}_{\boldsymbol{\beta}}^2 r \mathbf{I}_r\right), \tag{10}$$

where $\boldsymbol{\iota}_r$ denotes an $r \times 1$ vector of units. If $N_a \geq L_a$ and $N_b \geq L_b$ this provides a proper Gaussian prior for $\boldsymbol{\beta}$. The force of (10) is to provide a reasonable range of values of the polynomial $\boldsymbol{\beta}'\mathbf{u}$ over a domain (9) to which the model is meant to apply. The parameter $\underline{\tau}_{\boldsymbol{\beta}}^2$ provides the overall tightness of the prior, which is little affected by the number of grid points due to the presence of r in (10).

The prior distribution for $\mathbf{A}$ is constructed in similar fashion,

$$\mathbf{C}\boldsymbol{\alpha}_j \mid \left(\sigma^2, \underline{\tau}_{\boldsymbol{\alpha}}^2\right) \overset{iid}{\sim} N\left[\mathbf{0}_r, \underline{\tau}_{\boldsymbol{\alpha}}^2 \sigma^2 r \mathbf{I}_r\right] \quad (j = 1, \ldots, m). \tag{11}$$

The mean of $\mathbf{0}_r$ reflects the fact that location is handled by means of $\underline{\mu}$ in (10). The scale of the distribution is conditioned on σ^2 because $\mathbf{A}$ controls the shape of the conditional distribution, whereas $\boldsymbol{\beta}$ controls its location. For example, as $\underline{\tau}_{\boldsymbol{\alpha}}^2$ increases, so does the probability of multimodality of the conditional distribution, whereas the distribution approaches a scale mixture of normals as $\underline{\tau}_{\boldsymbol{\alpha}}^2 \to 0$. The prior distribution of the shape of the conditional distributions is also affected by the choice of the covariates $\mathbf{u}_t$, $\mathbf{v}_t$ and $\mathbf{z}_t$. An advantage of the polynomial basis functions used in the subsequent examples is that constraints on shape implied by polynomials of different order are well understood.

Identification issues well known in multinomial probit models also arise in (4)-(5). Scaling is resolved by the fixed variance matrix, but the problem of translation remains. Any proper prior distribution for $\mathbf{\Gamma}$ would resolve the identification question in the posterior distribution, and in the case of Gaussian priors this would be accomplished through the mean of the prior distribution. We have chosen instead to fully identify $\mathbf{\Gamma}$ by means of the restrictions $\boldsymbol{\iota}_m'\mathbf{\Gamma} = \mathbf{0}_q'$. Equivalently, define an $m \times m$ orthonormal matrix $\mathbf{P} = \begin{bmatrix} \mathbf{p}_1 & \mathbf{P}_2 \end{bmatrix}$ in which $\mathbf{p}_1 = \boldsymbol{\iota}_m m^{-1/2}$. Then define the $(m-1) \times p$ matrix $\mathbf{\Gamma}^* = \mathbf{P}_2'\mathbf{\Gamma}$ and work with the parameters in $\mathbf{\Gamma}^{*\prime} = \left[\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_{m-1}^*\right]$ rather than those in $\mathbf{\Gamma} = \mathbf{P}_2\mathbf{\Gamma}^*$. The prior distribution is

$$\mathbf{C}\boldsymbol{\gamma}_j^* \mid \left(\sigma^2, \underline{\tau}_\gamma^2\right) \overset{iid}{\sim} N\left(\mathbf{0}_r, \underline{\tau}_\gamma^2 \sigma^2 r \mathbf{I}_r\right) \quad (j = 1, \ldots, m-1), \tag{12}$$

which has only the single hyperparameter $\underline{\tau}_\gamma^2$. The appendix of the article shows that while the matrix $\mathbf{P}_2$ is not unique, the prior distribution of $\mathbf{\Gamma}$ implied by (12) is the same no matter what the choice of $\mathbf{P}_2$. For a given specification of covariates and m discrete outcomes the number of parameters in Γ^* is the same as the number of parameters in a conditional logit model with the same covariates and outcomes.

Both the model specification and the prior distribution are exchangeable with respect to the numbering of the states $\widetilde{s}_t$, and so there are $m!$ copies of the posterior distribution, each with a different permutation of the states. This exchangeability reflects the fact that states in this model do not have intrinsic properties – that is, we do not apply names like "graduate students" in the earnings example in Section 3 or "crash" in the asset returns example in Section 4. If this were the case exchangeable priors would be inappropriate, and there are other problems of interpretation in that case as well (see Celeux et al. (2000)). The only function of the states in the smoothly mixing regression model is to provide flexibility in the conditional distribution. The conditional distribution is invariant with respect to permutation of the states, and our functions of interest (discussed in the next section) depend only on the conditional distribution. Therefore formal lack of identification of the state numbers has no consequences for our work.

2.4 Inference and functions of interest

Conditional on the parameters $\boldsymbol{\beta}$, $\mathbf{A}$, $\mathbf{\Gamma}^*$, σ^2 and σ_j^2 $(j = 1, \ldots, m)$ and the covariate matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{Z}$ the smoothly mixing regression model consisting of (1), (4) and (5) completely specifies the distribution of the latent variables $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{s}}$ and the observables $\mathbf{y}$. The prior distribution of the parameters, in turn, does not involve the latent variables or observables and is completely specified by (7), (8), (10), (11) and (12). The kernel of the posterior density is the product of the corresponding prior density, the probability density of latent variables conditional on parameters, and the probability density of observables conditional on latent variables and parameters. Its arguments consist of the parameters and latent variables.

The posterior distribution is well suited to simulation by MCMC. The appendix of the article fully specifies the algorithm. With one exception it is a straightforward Gibbs sampler with five blocks: $\boldsymbol{\beta}$ and $\mathbf{A}$, σ^2, σ_j^2 $(j = 1, \ldots, m)$, $\boldsymbol{\Gamma}^*$ and $\widetilde{\mathbf{W}}$. For each observation t the parameters $\boldsymbol{\Gamma}^*$, vector of latent variables $\widetilde{\mathbf{w}}_t$ and (5) fully determine $\widetilde{s}_t$, which is therefore redundant in the MCMC algorithm. Conditional on $\widetilde{\mathbf{s}}$, the posterior distribution of $\boldsymbol{\beta}$, $\mathbf{A}$, σ^2 and σ_j^2 $(j = 1, \ldots, m)$ is nearly the same as that of the normal linear regression model, and these parameters can be simulated in much the same manner as in the Gibbs sampler for that model (see Geweke (2005, Examples 2.1.1 and 4.3.1)). The conditional posterior distribution of $\boldsymbol{\Gamma}^*$ involves only the latent variables $\widetilde{\mathbf{W}}$, and in this distribution the vectors $\boldsymbol{\gamma}_j^*$ $(j = 1, \ldots, m-1)$ are Gaussian and mutually independent.

The conditional posterior distribution of $\widetilde{\mathbf{W}}$ is complicated by the fact that its elements determine the states $\widetilde{\mathbf{s}}$. Each row $\widetilde{\mathbf{w}}_t$ of $\widetilde{\mathbf{W}}$ is conditionally independent of all the other rows, with conditional posterior density kernel

$$\exp\left[-\left(\widetilde{\mathbf{w}}_t - \boldsymbol{\Gamma}\mathbf{z}_t\right)'\left(\widetilde{\mathbf{w}}_t - \boldsymbol{\Gamma}\mathbf{z}_t\right)/2\right] \tag{13}$$

$$\cdot \sum_{j=1}^{m}\left[\prod_{i=1}^{m} I_{(-\infty,\widetilde{w}_{jt}]}\left(\widetilde{w}_{ti}\right)\right] \tag{14}$$

$$\cdot\, \sigma_j^{-1} \exp\left[-\left(y_t - \boldsymbol{\alpha}_j'\mathbf{v}_t - \boldsymbol{\beta}'\mathbf{u}_t\right)^2 / 2\sigma^2\sigma_j^2\right]. \tag{15}$$

This density is nonstandard but well suited to a Metropolis within Gibbs step. The source distribution is

$$\widetilde{\mathbf{w}}_t \sim N\left(\boldsymbol{\Gamma}\mathbf{z}_t, \mathbf{I}_m\right);$$

(13) is the corresponding kernel. The function (14) selects $j = j^* : \widetilde{w}_{tj^*} \geq \widetilde{w}_{ti}\ \forall$ $i = 1, \ldots, m$. Then the ratio of the target to source density is (15). We accept the draw with probability

$$\min\left\{\frac{\sigma_{j^*}^{-1} \exp\left[-\left(y_t - \boldsymbol{\alpha}_{j^*}'\mathbf{v}_t - \boldsymbol{\beta}'\mathbf{u}_t\right)^2 / 2\sigma^2\sigma_{j^*}^2\right]}{\sigma_j^{-1} \exp\left[-\left(y_t - \boldsymbol{\alpha}_j'\mathbf{v}_t - \boldsymbol{\beta}'\mathbf{u}_t\right)^2 / 2\sigma^2\sigma_j^2\right]}, 1\right\}$$

where j denotes the state assignment in the previous MCMC step for observation t, and j^* is the state assignment implied by the candidate draw. (Note $j^* = j$ implies acceptance.)

For the results reported in Sections 3 and 4 we executed $M = 12,000$ iterations of the Markov chain, with samples of size approximately $T = 2,500$. This requires between one and two minutes using state-of-the-art desktop hardware and fully compiled code; more generally, computing time is roughly proportional to the product of MCMC iterations M and samples size T. We discard the first 2,000 iterations and use every 100'th of the remaining 10,000 iterations for analysis. There is no detectable serial correlation in the 100 iterations used for analysis. The code as well

as the derivation of the algorithm detailed in the appendix were checked using the joint distribution test described in Geweke (2004) and Geweke (2005, Section 8.1.2).

The speed of the MCMC algorithm is due in substantial part to its simulation of latent states, an efficient alternative to the computation of state probabilities. This is an advantage of Bayesian MCMC approaches generally over methods like simulated maximum likelihood and simulated method of moments that require the computation of state probabilities (see Geweke et al. (1994) and Geweke and Keane (2001)). On the other hand many functions of interest, including the posterior quantiles illustrated in the next two sections, require evaluations of these state probabilities, as do evaluations of the likelihood function entailed in formal methods of model comparison like those discussed in the same two sections. Evaluation of these probabilities is simpler here than for multinomial probit models generally, because

$$\begin{aligned} P\left(\widetilde{s}_t = j \mid \mathbf{\Gamma}, \mathbf{z}_t\right) &= P\left[\widetilde{w}_{tj} \geq \widetilde{w}_{ti} \;\; (i = 1, \ldots, m) \mid \mathbf{\Gamma}, \mathbf{z}_t\right] \\ &= \int_{-\infty}^{\infty} \mathrm{p}\left(\widetilde{w}_{tj} = y \mid \mathbf{\Gamma}, \mathbf{z}_t\right) \cdot P\left[\widetilde{w}_{ti} \leq y \;\; (i = 1, \ldots, m) \mid \mathbf{\Gamma}, \mathbf{z}_t\right] dy \\ &= \int_{-\infty}^{\infty} \phi\left(y - \boldsymbol{\gamma}_j' \mathbf{z}_t\right) \prod_{i \neq j} \Phi\left(y - \boldsymbol{\gamma}_i' \mathbf{z}_t\right) dy. \end{aligned} \tag{16}$$

This integral has only a single dimension (regardless of value of m) and so it can be evaluated by conventional quadrature methods without the need to employ simulation approximations like the GHK probability simulator (Geweke and Keane (2001, Section 2.1)). The following two examples illustrate how the evaluation of (16) is embedded in the computation of functions of interest and model evaluation criterion functions.

Suppose, as in the examples we consider subsequently, that $\mathbf{x} = (a, b)'$, and the function of interest is the quantile

$$\begin{aligned} c\left(q, a, b\right) &= \left\{c : P\left(y \leq c \mid a, b, Data, SMR\right) = q\right\} \\ &= \left\{c : P\left(y \leq c \mid \mathbf{u}, \mathbf{v}, \mathbf{z}\,, Data, SMR\right) = q\right\} \end{aligned} \tag{17}$$

for specified $q \in (0, 1)$. This entails embedding the evaluation of the posterior c.d.f. $P\left(y \leq c \mid a, b, Data, SMR\right)$ in an iterative root-finding algorithm. The posterior distribution is conveyed through the iterations of the MCMC algorithm. If M such iterations are used in the approximation then the distribution of y conditional on a and b is a mixture of $M \cdot m$ normal distributions. Let $\boldsymbol{\beta}, \mathbf{A}, \mathbf{\Gamma}^*, \sigma^2, \sigma_1^2, \ldots, \sigma_m^2$ be the parameter values in a particular iteration, and let $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{z}$ be the covariate values corresponding to $\mathbf{x} = (a, b)'$. Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \mathbf{A}, \mathbf{\Gamma}^*, \sigma^2, \sigma_1^2, \ldots, \sigma_m^2\}$. Then

$$P\left(y \leq c \mid \mathbf{x}, \boldsymbol{\theta}\right) = \sum_{j=1}^{m} \Phi\left(\frac{c - \boldsymbol{\beta}'\mathbf{u} - \boldsymbol{\alpha}_j'\mathbf{v}}{\sigma \cdot \sigma_j}\right) \cdot P\left(\widetilde{s}_t = j \mid \mathbf{\Gamma}, \mathbf{z}\right)$$

and $P\left(y \leq c \mid a, b, Data, SMR\right)$ is the average of these probabilities over the iterations. In the work reported here we have used $M = 100$ equally spaced iterations out of 10,000 successive iterations of the MCMC algorithm.

Because there are two substantively distinct covariates a and b, we can communicate the surface (17) in a series of graphs, one for each of several values of q. Results in the next section use a grid of $P = 450$ combinations of a and b to evaluate $Q = 7$ quantiles. If the root-finding algorithm for c requires R iterations, and we report results for Q quantiles, then $MmPRQ$ evaluations of (16) are involved, where R is the average number of iterations required to solve for the inverse c.d.f. from the c.d.f. We did not track the value of R, but used a relative error criterion of 10^{-10} for convergence of the quadrature evaluation of (16). Beginning with the output of the MCMC posterior simulator and using state-of-the-art desktop hardware and fully compiled software the time required for these computations is about 3 minutes to produce a set of contours like those shown in Figures 3 and 4.

Formal methods of model evaluation also involve the computation of (16) embedded in

$$\mathrm{p}\left(y_t \mid \mathbf{u}_t, \mathbf{v}_t, \mathbf{z}_t, \boldsymbol{\theta}\right) = \sum_{j=1}^{m} \left(\sigma \cdot \sigma_j\right)^{-1} \phi\left(\frac{y_t - \boldsymbol{\beta}'\mathbf{u}_t - \boldsymbol{\alpha}_j \mathbf{v}_t}{\sigma \cdot \sigma_j}\right) \cdot P\left(\widetilde{s}_t = j \mid \boldsymbol{\Gamma}, \mathbf{z}_t\right). \quad (18)$$

For problems of the size considered in Sections 3 and 4, evaluation of (18) requires roughly 10^{-3} to 2×10^{-3} seconds. The modified cross-validated log scoring rules we use to compare variants of the SMR model, described in Section 3, evaluate (18) at about 600 observations for 100 alternative parameter values. The total computation time is one to two minutes beginning from the output of the MCMC posterior simulator, and the numerical standard error is 0.2 to 0.4. As described in Section 4 we also use predictive likelihoods for model comparison, again with 100 alternative parameter values. The evaluation of (18) adds negligibly to the time required by the posterior simulator, and numerical standard error is again 0.2 to 0.4. A number of simulation procedures could, in principle, be used to evaluate the marginal likelihood, all of which require repeated evaluation of the likelihood function at alternative parameter values. (See Geweke (2005, Section 8.2) for an overview of these methods.) It is an open question, left to future research, how many alternative parameter values would be required to achieve comparable numerical standard errors.

2.5 Comparison with related models

Express the smoothly mixing regression model in the compact form

$$\mathrm{p}\left(y \mid \mathbf{x}, m, k, p, q, \boldsymbol{\theta}\right) = \sum_{j=1}^{m} \mathrm{p}_1\left(y \mid \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\alpha}_j, \sigma^2, \sigma_j^2, \widetilde{s} = j; k, p\right) \mathrm{p}_2\left(\widetilde{s} = j \mid \mathbf{x}, q; \boldsymbol{\Gamma}^*\right), \quad (19)$$

plus the prior distribution of the parameters. In this expression $\mathbf{x}$ denotes the $n \times 1$ vector of substantive covariates, for example $\mathbf{x} = (a, b)'$ in the example just discussed, and the remaining notation is the same established at the start of this section with $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\Gamma}^*, \sigma^2\sigma_1^2, \ldots, \sigma_m^2\}$. The choice of k maps $\mathbf{x}$ into the covariate vector $\mathbf{u}$, p maps $\mathbf{x}$ into $\mathbf{v}$, and q maps $\mathbf{x}$ into $\mathbf{z}$, through a set of basis functions in each case. (These functions are polynomials in the examples in the next two sections.) Then p_1 is the conditional density of the continuous random variable y implied by (6) and p_2 is the probability distribution of the discrete random variable $\widetilde{s}_j$ implied by (4)-(5). Various configurations of k, p and q lead to the model variants A through E discussed at the end of Section 2.2 and detailed in Table 1. This expression facilitates comparison of the smoothly mixing regression model with closely related approaches to modeling $\mathrm{p}(y \mid \mathbf{x})$ found in the literature, which tend to be special or limiting cases of (19).

Morduch and Stern (1997) approach the problem of modeling the impact of a child's sex and other covariates on a continuous measure of health outcomes as a particular case of (19) except that p_2 maps $\mathbf{x}$ into $P(\widetilde{s} = j)$ using a logistic rather than a normal distribution. Since they used maximum likelihood rather than Bayesian methods, the logistic is a natural choice. That study set $m = 2$, and the rest of the model remained tightly parameterized; in particular there was no analogue of the expansions of basis functions implicit in k, p and q that are central in the examples in the rest of this study. Since that study found only marginal evidence in favor of $m = 2$ as opposed to $m = 1$ this was a reasonable strategy; by comparison, our examples finds strong evidence for $m > 2$. Wedel et al. (1993) and Deb and Trivedi (1997) use a similar approach but for discrete rather than continuous outcomes.

There are a number of related approaches in the literature than can be interpreted as limiting cases of (19). Perhaps the most important is $m = \infty$ in conjunction with a Dirichlet process prior for the mixture. Recent work of Dunson and Pillai (2004) appears the closest in this genre to our approach. It specifies $\mathrm{p}(y \mid \mathbf{x}) = \int \mathrm{p}(y \mid \mathbf{x}, \boldsymbol{\phi})\, dG_{\mathbf{x}}(\boldsymbol{\phi})$; $G_{\mathbf{x}}(\boldsymbol{\phi})$ is made smooth with respect to $\mathbf{x}$ through a conditional mixing structure. In application the number of states is finite, and a distribution over different values of m emerges as a byproduct. In the context of (19) Dunson and Pillai (2004) concentrate on expansion in m and q (though with a different functional form p_2), maintaining parsimonious specification of k and p. Our approach emphasizes richness in both p_1 and p_2; this requires considering several dimensions (m, k, p and q), and thereby imposes some costs on the investigator. The subsequent examples illustrate some systematic ways of managing these several dimensions simultaneously, and provide some evidence on the gains from pursuing flexibility in both p_1 and p_2.

If we write the multinomial probit state model (5) in the form

$$\widetilde{\mathbf{w}}_t = h \cdot \boldsymbol{\Gamma}\mathbf{z}_t + \boldsymbol{\zeta}_t;\ \boldsymbol{\Gamma} \text{ fixed}$$

then as $h \to \infty$

$$\widetilde{s}_t = j \text{ iff } \boldsymbol{\gamma}_j'\mathbf{z}_t \geq \boldsymbol{\gamma}_i'\mathbf{z}_t\,(i = 1, \ldots, m)$$

where $\boldsymbol{\gamma}_i$ denotes row i of $\boldsymbol{\Gamma}$. Thus models in which state classifications are deterministic functions of $\mathbf{x}$ emerge as limiting cases of the smoothly mixing regression models. These classifications amount to a partition of the domain of $\mathbf{x}$. Special instances of the limit include the Voronoi tessellation approach to multidimensional spline smoothing used in spatial statistics, engineering and computer science (Holmes et al., 2005), and threshold autoregression time series models (Geweke and Terui, 1993). Large but finite values of h have the effect of smoothing the splines at the join points.

3 Earnings

To gain some practical experience with the model and illustrate its properties, we turn first to a subset of the panel survey of income dynamics (PSID) sample of earnings for men used in our previous research (Geweke and Keane (2000)). It consists of the 1993 earnings of those 2,698 men in the PSID who were white, between the ages of 25 and 65 inclusive, and earned more than $1,000 in that year. (As in the previous work, the exact choice of the truncation point has no substantive impact on the results because very few men in the sample have positive earnings of less than $1,000.) We also know the age and education of these men, and we focus on the distribution of the logarithm of earnings (y_t) conditional on age (a_t) and education (b_t). This is also the focal point of much of the applied quantile regression literature noted in the introduction. Neither our treatment nor this literature (e.g. Buchinsky (1994)) addresses causal interpretation of this relation, which raises important questions about identification (e.g. Heckman and Robb (1985)). Since we use fully interacting polynomial functions of a and b results would not change were we to substitute experience, defined as age minus education minus six, as is sometimes reported. The objective is to characterize $\mathrm{p}\left(y \mid \mathbf{x}\right)$, $\mathbf{x} = (a, b)'$, using the smoothly mixing regression model.

Figure 2 indicates the distribution of the covariates in the sample, omitting those very few men with less than eight years of education. Subsequently we present conditional distributions for the same ranges of covariate values. The covariate distribution is important in appreciating the results. We expect to find greater certainty about the conditional distribution at those values for which the sample provides substantial information (for example 12 to 16 years of education and ages 30 to 45) than for those where there is little information (for example, less than 10 years of education and younger than 35). Given the flexibility of the smoothly mixing regression model, we expect to see the wide ranges of uncertainty suggested by the cell counts in Figure 2.

Applying the smoothly mixing regression model requires choice among the model specifications A through E, the number of mixture components m, and the polynomial orders L_a and L_b. We must also select the seven hyperparameters of the prior distribution detailed in Section 2.3: $\underline{\nu}^*$, $\underline{s}^2$, $\underline{\nu}$, $\underline{\mu}$, $\underline{\tau}_{\boldsymbol{\beta}}^2$, $\underline{\tau}_{\boldsymbol{\alpha}}^2$ and $\underline{\tau}_{\boldsymbol{\gamma}}^2$, as well as the grid G for a and b indicated in (9). The last choice is the simplest: it indicates the values of age and education to which the model applies, ages 25 through 65 and years of

education 0 through 17, the last number being the topcode for years of education in the PSID.

Our methods of analysis, detailed shortly, do not depend in any critical way on the choices of prior hyperparameters. (This is not the case for some alternative methods, in particular marginal likelihoods and Bayes factors for model comparison, which we do not use in this study.) To avoid rounding error and similar complications involving polynomials, we rescaled the covariates so that they each ranged from -1 to 1, but made no change to the natural log of earnings y_t. The prior distributions pertain to the rescaled data, but all the reported results are converted to the original units. The sample mean of y_t is about 10, and the sample standard deviation is about 0.8. For the coefficients $\boldsymbol{\beta}$ in (10) the mean is $\underline{\mu} = 10$ and the variance is $\underline{\tau}^2_{\boldsymbol{\beta}} = 1$. This applies a rough approximation of the sample distribution to each point on the grid, but recall that the variance in the normal prior distribution (10) is scaled by the number of grid points r, so that this prior has the weight of one observation. Our prior for $\mathbf{A}$ is even weaker, setting $\underline{\tau}^2_{\boldsymbol{\alpha}} = 4$, expressing the specification that there may be substantial differences across components. In the prior for $\boldsymbol{\Gamma}^*$, $\underline{\tau}^2_{\boldsymbol{\gamma}} = 16$. This provides enough flexibility that a particular mixture component might be nearly certain for some combinations of age and education and nearly impossible for others. For the parameter σ^2 we selected the prior hyperparameters $\underline{s}^2 = \underline{\nu} = 2$, corresponding to a centered 90% credible interval of $(0.57, 4.47)$ for σ. We set $\underline{\nu}^* = 2$, implying that ratios σ_j^2/σ_i^2 less than 1/9 or greater than 9 each have probability 0.10. We checked final results for sensitivity to choice of hyperparameters and, as expected, none is detectable.

We approach the problem of choosing among the model specifications A through E, the number of mixture components m, and the polynomial orders L_a and L_b using a modified cross-validated log scoring rule. A full cross-validated log scoring rule (Gelfand et al. (1992), Bernardo and Smith (1994, Section 6.1.6)) entails evaluation of

$$\sum_{t=1}^{T} \log \left[\mathrm{p} \left(y_t \mid \mathbf{Y}_{T/t}, a_t, b_t, SMR \right) \right],$$

where $\mathbf{Y}_{T/t}$ denotes the sample with observation t removed. This is computationally expensive, because it requires T posterior simulators (Draper and Krnjajic (2005)), and there are quite a few combinations of model specifications, m, L_a and L_b to sort through. In our modification of the cross-validated log scoring rule we randomly order the sample of 2698 observations and use the first $T_1 = 2153$ for inference. We then compute the individual log-score for each of the last 545 observations

$$\sum_{t=T_1+1}^{T} \log \left[\mathrm{p} \left(y_t \mid \mathbf{Y}_{T_1}, a_t, b_t, SMR \right) \right], \tag{20}$$

where $\mathbf{Y}_{T_1}$ denotes the first T_1 observations of the randomly ordered sample.

Table 2 conveys the results of this exercise. The numerical standard errors for the entries in the table range from 0.18 to 0.71, with the majority being between 0.20 and 0.40. Several findings about model performance are noteworthy.

1. Specification C, a smooth mixture of normal distributions each with a fixed mean and variance, performs the poorest. In view of the well-documented smooth relationship between the conditional expectation of earnings and the covariates age and education, this is not surprising. In principle, by mixing a large number of such distributions these smooth relationships could be well approximated. This accounts for the increases in log score moving from $m = 2$ to $m = 3$ to $m = 4$. Even with $m = 4$, however, the performance of this specification is inferior to that of specifications A, B, D or E.

2. For any combination of m, L_a and L_b, the log score for D exceeds that of A; A always exceeds C while D nearly always falls short of B or E. Model A is a regression model, conventional in every respect except that the disturbance $\varepsilon_t = y_t - \boldsymbol{\beta}'\mathbf{u}_t$ is a mixture of normals independent of $\mathbf{u}_t$. The superior log score of model D relative to A may be interpreted as evidence against the independence of ε_t and $\mathbf{u}_t$.

3. Specifications B and E are superior to specification D in nearly all cases. Unlike any of the other specifications, B and E incorporate mixtures of linear combinations of covariates, and the results in Table 1 may be taken as evidence that this is important in modeling the distribution of earnings conditional on age and education. There is no systematic tendency for one of these specifications to outperform the other.

Figures 3 and 4 provide quantiles of the posterior conditional distribution of earnings for the model with the highest log score in Table 2: specification E with a mixture of $m = 3$ normal distributions, the covariates being interactive polynomials of order $L_a = 4$ in age and $L_b = 2$ in education. The first of these two figures provides results for quantiles below and at the median and the second does so for quantiles at the median and above. Quantiles $c(q)$ are increasing functions of education for all ages and values of q, except in the lowest quantile $q = 0.05$ for highly educated men under the age of 30 (completing dissertations?) and over the age of 50. The age profile varies substantially by quantile and by level of education. For high school graduates ($b = 12$) the lower the quantile, the shallower the increase in earnings from ages 30 to 50, and the sharper the decline in earnings after age 50, with the decline beginning somewhat sooner in the lower quantiles. This phenomenon is somewhat more pronounced for college graduates ($b = 16$). The earnings of these men in quantiles below the median peak well before age 50 and the decline in earnings is precipitous beyond age 60. At quantiles $q = 0.75$ and above, there is no decline in the earnings of these men before age 65.

Since specification B is a special case of E, it is plausible that the close comparisons of these specifications in Table 1 indicate that the models have similar implications for the distribution of earnings conditional on age and education. We examine this implication by reproducing in Figure 5 the same quantiles as in Figure 3, except that these quantiles come from model B rather than from model E. The results are similar, in that the major patterns of age, education and quantile just noted are reproduced, but there are some minor differences. The decrease in earnings for the lowest quantile $q = 0.05$ at 17 years of education for very young and old men noted in Figure 3(a) is not exhibited in Figure 5(a), and the difference between high school and college for men age 40 is about 0.6 at the lowest quantile in model B, whereas it is about 0.75 in model E. Comparisons for quantiles above the median (not shown here) also show similar results.

We conclude that results do not depend much on the choice between models B and E, and continue the analysis with model E. Figure 6 reproduces the results of Figure 3, but with the substantially reduced polynomial orders $L_a = 2$ and $L_b = 1$. The diminished flexibility in the modeling of quantiles is immediately evident. For example, earnings of college graduates in quantiles $q = 0.05$ and $q = 0.10$ do not fall as precipitously beyond age 50 in Figure 6 as in Figure 3. For men in the lowest quantiles who did not graduate from high school (i.e., $b < 12$) this model shows a substantially smaller decline in income than do the models with higher orders of polynomials. Thus the deterioration in fit for $L_a = 2$ and $L_b = 1$ versus $L_a = 4$ and $L_b = 2$ noted above appears to be reflected in distortions for functions of interest of the kind one would expected from oversmoothing.

Dispersion in the conditional distribution can be measured by differences in quantiles as a function of age and education. Figure 7 does this using the conditional interquartile range, the length of the centered 80% conditional credible interval for log earnings, and the length of the centered 90% conditional credible interval. The alternative measures all indicate that dispersion is smallest and roughly constant for a wide range of the conditioning variables, with minimum dispersion for men under age 30 with 12 or 13 years of education, and not much increase until the age of 55 for education levels between 10 and 15 years. Outside this region dispersion increases rapidly. For men under 55 it is about the same for 10 and 16 years of education, and dispersion is greatest for older, highly educated men. As noted earlier from Figures 3 and 4, this is due to a combination of a sharp drop in the lowest earnings quantiles of highly educated men, combined with modest increases in the highest earnings quantiles, approaching age 65. Panel (d) of Figure 7 indicates that the conditional distribution of log earnings is skewed to the left at all ages, with skewness being most pronounced for older men and for younger poorly educated men.

As emphasized in Section 2.4, these quantiles pertain to the posterior distribution, which integrates uncertainty about parameters with the uncertainty conditional on parameter values. We can sort out these two sources of uncertainty by examining the posterior variation in the quantiles of the population distribution. Each drawing

from the posterior distribution in the MCMC algorithm implicitly generates a distribution of log earnings conditional on age and education. Figure 8 portrays the 10% quantile of these distributions explicitly for four widely spaced drawings from the Markov chain. (Recall that the chain uses 12,000 iterations, of which the first 2,000 are discarded and 100 equally spaced drawings are used for the analysis here; the numbering in Figure 8 is with respect to these 100 drawings.) The general pattern in the four figures is about the same. Closer examination bears out the properties relative to the sample distribution of covariates (Figure 2) previously anticipated. At points where the sample is most highly concentrated there is almost no variation in the quantile across the four panels. For example, at $a = 40$ and $b = 12$, the 10% quantile ranges from about 9.55 to about 9.625. On the other hand the combination $a = 30$, $b = 17$ produces quantiles between 9.5 (panel (b)) and 10.2 (panel (a)).

A wide range of exercises of this kind can be undertaken, depending on the aspect of the earnings distribution of interest. Since we have presented only quantiles, perhaps it bears emphasis that our method allows access to the entire distribution in distinction to the methods surveyed in the introduction that provide only estimates of quantiles. For example, if inequality is measured using a Gini coefficient, or if one is interested in the probability that earnings fall below a particular threshold, the analysis can proceed in similar fashion and results can be presented in the same way.

4 Stock returns

In our second illustration the variable of interest y_t is daily returns on the Standard and Poors (S&P) 500 index, measured as $y_t = 100 \log (p_t/p_{t-1})$, where p_t is closing S&P 500 index on day t and p_{t-1} is the closing S&P 500 index on the previous trading day $t-1$. The objective is to characterize

$$\mathrm{p}\left[y_t \mid y_{t-s} \ \ (s > 0)\right]$$

using the smoothly mixing regression model. To begin, we construct two functions a_t and b_t of y_{t-s} $(s > 0)$ that are likely to be important for the distribution of y_t, based on the substantial literature on this topic. The first function is the preceding day's return $a_t = y_{t-1}$. The second function incorporates the history of absolute returns,

$$b_t = (1-g) \sum_{s=0}^{\infty} g^s \left|y_{t-2-s}\right|^{\kappa}. \tag{21}$$

In (21) the summation begins with y_{t-2} because y_{t-1} already enters the model in flexible fashion through a_t. The parameters g and κ are unknown and could enter the model symmetrically with other unobservables, but in this illustration we fix them at reasonable values, in a manner to be described shortly. Our sample consists of every trading day in the 1990's. We utilize daily data from the 1980's, as well, in constructing (21), so the finite truncation of the lag has no real impact on the results.

The hyperparameters of the prior distribution are similar to those for the earnings illustration, because the scale of the data in this illustration is about the same as the scale of the data in that illustration. As before, $\underline{s}^2 = \underline{\nu} = \underline{\nu}^* = 2$. The grid G for the parameters of $\boldsymbol{\beta}$, $\mathbf{A}$ and $\boldsymbol{\Gamma}$ consists of 101 evenly spaced values of a, ranging from -10 to 10, and 51 values of $b = (1-g)\sum_{s=0}^{\infty} g^s \left|b^*\right|^{\kappa}$ where b^* ranges from 0 to 10 in increments of 0.2. As in the previous example, $\underline{\tau}^2_{\boldsymbol{\beta}} = 1$, $\underline{\tau}^2_{\boldsymbol{\alpha}} = 9$ and $\underline{\tau}^2_{\boldsymbol{\gamma}} = 16$. We set $\underline{\mu} = 0$, a reasonable approximation to the unconditional mean of returns. For the same reasons given in Section 3, our quantile functions of interest show no detectable sensitivity to changes in the hyperparameters of these very weak priors, nor do our model comparison exercises. The same is true of posterior moments of the conditional distribution, which we also examine briefly in this section.

We compare variants of the smoothly mixing regression model using a modified log-scoring rule. Given κ and g, we form the covariates a and b and then randomly order the 2628 observations (a_t, b_t, y_t). We use the first $T_1 = 2017$ for inference and then compute the modified log scoring rule (20) as in the previous example. Because $\{y_t\}$ is a time series and a_t and b_t are each functions of lagged values of y_t, the interpretation of the log scoring rule is not as natural here as it is with the cross-section data of Section 3, and we return to a more natural method of evaluation below. Table 3 conveys some of the results of this exercise; numerical standard errors are similar to those for Table 2, mostly between 0.2 and 0.4.

Based on the results of this exercise, we selected for further work specification C with $m = 3$ components in the mixture, polynomial orders $L_a = L_b = 2$, and the parameters $\kappa = 1.0$ and $g = 0.95$ for the construction of the volatility covariate b. There are several reasons for these choices, based on prior considerations and the results in Table 3.

1. We began the exercise with a strong prior in favor of specification C, which mixes distributions of the form $N\left(\mu_j, \sigma_j^2\right)$ with the probability weights for the distributions depending on (a_t, b_t). This model captures conditional heteroscedasticity, a critical and well-documented characteristic of asset return data. While other specifications, particularly D and E, also capture this phenomenon, they do so at the cost of introducing more parameters that link $\mathrm{E}\left(y_t \mid \widetilde{s}_t = j\right)$ systematically to a_t and b_t. If $\widetilde{s}_t$ were observed, all but the weakest such links would create arbitrage opportunities. In fact $\widetilde{s}_t$ is unobserved, but changing volatility will provide strong signals about $\widetilde{s}_t$ in many periods t.

2. The scores in Panel A support this reasoning. Specification A has mean effects but no conditional heteroscedasticity and fares most poorly. Specification B improves relative to C, D, and E as the number of components m increases, but always takes fourth place. Specifications D and E, which nest C, have roughly comparable scores. There is no notable improvement in scores beyond specification C with $m = 3$ components.

3. Panel B shows that the choice of polynomial orders has much less consequence for the scores. Our choice $L_a = L_b = 2$ is one of the two smallest scores, but given numerical standard errors of 0.2 to 0.4 there is not much discrimination among polynomial orders.

4. Panel C suggests that the score is maximized by (g, κ) near the chosen values of $g = 0.95$ and $k = 1.0$. While these could be treated as unknown parameters, our subsequent graphical interpretation of the posterior distribution benefits from keeping these values fixed.

We compared the chosen specification with some alternative models for asset returns on the basis of their out-of-sample forecasting records. We began by computing the predictive likelihood of the SMR model for the period 1995-1999. Let $T_1 = 1314$ denote the last trading day in 1994 and $T = 2628$ the last trading day in 1999. Then the predictive likelihood is

$$\log\left[\mathrm{p}\left(y_{T_1+1}, \ldots, y_T \mid \mathbf{Y}_{T_1}, SMR\right)\right] = \sum_{s=T_1+1}^{T} \log\left[\mathrm{p}\left(y_s \mid \mathbf{Y}_{s-1}, SMR\right)\right]. \tag{22}$$

where $\mathbf{Y}_t = \{y_1, \ldots, y_t\}$. Each of the $T - T_1$ terms on the right-hand side of (22) can be approximated from the output of a posterior simulator using observations $1, \ldots, s-1$. For each of $T - T_1$ samples the execution of the MCMC algorithm with 12,000 iterations requires one to two minutes using state-of-the-art software and compiled code, but the results are an order of magnitude more accurate than any other method for the approximation of log predictive likelihoods of which we are aware.

This exercise produces the log predictive likelihood -1602.0, the entry in the first line of Table 4; the numerical standard error is 0.45. The remaining lines provide comparisons with three models in the ARCH family, plus the i.i.d. normal model as benchmark. For the closest of the ARCH competitors, t-GARCH(1,1), Table 4 provides both the predictive likelihood and the cumulative one-step-ahead predictive densities

$$\sum_{s=T_1+1}^{T} \log\left[\mathrm{p}\left(y_s \mid \widehat{\boldsymbol{\theta}}_{s-1}, SMR\right)\right] \tag{23}$$

where $\widehat{\boldsymbol{\theta}}_{s-1}$ denotes the maximum likelihood estimates of the model parameters based on the sample $\mathbf{Y}_{s-1}$. The result is close to the predictive likelihood because the posterior distribution of the model parameters is tightly concentrated around the maximum likelihood estimate. (Given that sample sizes range from 1314 to 2627 and the fact that the model has only four parameters, this is not surprising.) In the interest of sparing computational expense Table 4 reports only (23) for the other models. The superior performance of the t-GARCH models within the ARCH family, using a likelihood criterion, is consistent with the literature (e.g. Dueker (1997), Yang

and Brorsen (1992)). However, the SMR model substantially outperforms t-GARCH on this same criterion.

Figure 9 shows most of the sample of the substantive variables a_t and b_t in our specification of the smoothly mixing regression model for the Standard and Poors 500 return. (Because of scaling considerations, Figure 9 excludes about 10% of the sample.) As in the illustration with the earnings data in the previous section, knowing the distribution of the covariates helps in the interpretation of the results. Given the flexibility of the model, we expect to find that the reliability of the results varies directly with the concentration of the data points in Figure 9.

As in Section 3, our primary functions of interest are the quantiles of the conditional distribution. In the case of asset returns conditional moments are also of interest, for several reasons: there is a well-founded presumption that the conditional mean of returns should not differ from a normal rate of return, for all reasonable values of the conditioning covariates a and b; there is a substantial literature concerned with the conditional variance of returns; and there is some interest in whether the distribution is skewed. Let $\boldsymbol{\theta}$ denote the vector of parameters, consisting of the elements of $\boldsymbol{\beta}$, $\mathbf{A}$, $\boldsymbol{\Gamma}^*$, σ^2 and σ_j^2 $(j = 1, \ldots, m)$, and let $m(\boldsymbol{\theta})$ be a corresponding function of population moments. Figure 10 reports posterior means for four choices of $m(\cdot)$: mean in panel (a), standard deviation in (b), coefficient of skewness in (c) and coefficient of kurtosis in (d). Figure 11 provides posterior standard deviations of these same four moments in the corresponding panels.

Figure 10, panel (a), shows that the posterior mean return is between 0.05% (corresponding to an annual return of 13.4%) and 0.055% (14.9%) for values of a and b corresponding to roughly three-quarters of the sample. The corresponding panel of Figure 11 shows that the posterior standard deviation of the conditional mean increases sharply for points (a, b) that are less typical of the sample, more so when the previous day's return a is negative than when it is positive. The two panels strongly suggest that the hypothesis that the conditional mean is constant, which we have not examined formally, would be sustained. Not surprisingly, the conditional standard deviation is larger in periods of greater volatility (higher values of b, see Figure 10 panel (b)). There is also evidence of a leverage effect, that is, greater volatility following a negative return than a positive one. Comparison of panel (b) in Figure 11 with panel (b) in Figure 10 strongly suggests that more formal analysis would sustain this hypothesis. The leverage effect and the sample distribution of a and b appear to account for the posterior standard deviation of the mean in Figure 10 panel (a).

There is no evidence of skewness in the conditional distribution. In panel (c) of Figure 10 the posterior means of the population skewness coefficient are quite small, and the corresponding panel in Figure 11 indicates that the posterior standard deviation of the conditional skewness coefficients always exceeds their posterior mean. Panel (d) of Figure 10 shows an interesting pattern of excess kurtosis, but these posterior means of the population moment must be interpreted in the context of

their substantial posterior standard deviation, shown in the corresponding panel of Figure 11. For most configurations of the covariates the posterior standard deviation of the coefficient of excess kurtosis varies between 1 and 1.5, and differences in the posterior mean are on the same order. Given a large return, a, in a period of low volatility, b, there may be very high kurtosis; but the data underlying these conditions are thin (Figure 9) and the evidence is correspondingly tenuous (Figure 11, panel (d), lower right corner).

Figures 12 and 13 show the quantiles of the conditional posterior distribution of returns. This display for quantile 0.50 is empty, because the value of this quantile is about 0.02 for all values of a and b; since it never falls below 0 or rises above 0.25, no lines are shown in panel (d) of Figures 12 or panel (a) of Figure 13. Comparison of Figures 12 and 13 clearly reveals that returns are symmetric, or nearly so, for all values of a and b, which in turn is consistent with the small absolute values of the coefficient of skewness noted in Figures 10 and 11.

Each of these figures, especially Figure 12(a), provide rich information on value at risk, expressed in units of percent return. Regardless of the current period return a, value at risk is an increasing function of b, which measures volatility in recent weeks. On the other hand, for most levels of volatility b, value at risk is a decreasing function of the current period return. Especially during periods of high volatility, value at risk declines sharply as a increases, including values of a that are unusually high, like 2% or 3%. This feature emerges in other variants of the model, including those with more components m and higher orders of polynomials L_a and L_b.

Figure 14 uses the quantiles to portray the dispersion and asymmetry of the posterior distribution conditional on a and b. The pattern of dispersion is broadly similar to that indicated by the standard deviation in Figure 10(b). Consistent with this portrayal and value at risk, dispersion always increases as a function of b and nearly always decreases as a function of a. Careful comparison of panels (a), (b) and (c) of Figure 14 with Figure 10(b) also reveals the non-normality of the distribution. For example, the ratio of the interquartile range to the standard deviation is 1.35 in the normal distribution, but for all values of a and b the ratio is smaller in the conditional posterior distribution of returns. When excess kurtosis is modest (see Figure 10(d)) the difference between quantiles 0.95 and 0.05 is longer, relative to the standard deviation, than is the case for the normal; but when excess kurtosis is highest, the difference is still shorter.

As emphasized in Section 2.4, the quantiles pertain to the posterior distribution, not the population distribution. We emphasize these quantiles here because they are the ones pertinent for decisionmaking in general and they address the question of value at risk directly. Just as in the case of moments, however, one can also pose a slightly different question: what is the degree of uncertainty about quantiles of the population conditional distribution? Figure 15 uses the same technique as Figure 8 in Section 3 to answer this question, showing the population 5% quantile from four widely separated draws in the MCMC sample. Comparison of Figure 15 with Figure 9

shows that the dispersion of the posterior distribution of this quantile moves inversely with the concentration of sample points. For $a = 0$ and $b = 0.5$, the 5% quantile is between -1.20 and -1.25 in all four panels; for $a = -3$ and $b = 1.5$ it exceeds -3.00 in panel (b) and is less than -4.00 in panel (a). All four panels display the feature of Figure 12(a) emphasized previously: value at risk not only increases with volatility b, but it also generally decreases as the algebraic value of the current period return increases.

5 Conclusions and further research

Inference about conditional distributions $\mathrm{p}(y \mid \mathbf{x})$ lies at the heart of many problems in econometrics, and the question has always figured prominently in the research agenda of the profession. This study has documented progress in eliminating restrictive assumptions about $\mathrm{p}(y \mid \mathbf{x})$ and increasing the scope of application of the resulting methods. It has proposed a new approach to inference for conditional distributions, the smoothly mixing regression (SMR) model, largely by combining models well-established in other contexts in econometrics. The result is a flexible and practical procedure that produces a full posterior distribution for the conditional distribution. This study provided two illustrative examples. For one example, comparison with leading alternative models showed that SMR is superior using out-of-sample likelihood criteria. Further such comparisons, using alternative applications and data and alternative metrics, should shed more light on the relative strengths and weaknesses of SMR, but are beyond the scope of this study. The illustrations here emphasize quantiles as functions of interest, and direct comparison with the application of quantile regression models would be interesting.

There are further immediate applications of SMR in investigative and decision-making contexts. Because the approach taken here models the full conditional distribution (unlike quantile regression) it provides a foundation for inference about functionals $\int \mathrm{f}(y)\,\mathrm{p}(y \mid \mathbf{x})\,dy$. For example, the model can be used to estimate the fraction of a population exceeding some threshold c as a function of the covariates $\mathbf{x}$, simply by taking $\mathrm{f}(y) = I_{(c,\infty)}(y)$. Questions of this form arise in spatial statistics, where y might measure toxicity and $\mathbf{x}$ is a 2×1 location coordinate. If y measures the price of an asset and c is the strike price of a call option then $\mathrm{f}(y) = (y - c)\, I_{(c,\infty)}(y)$ provides the option's expected revenue. Some settings combine such functionals with quantiles: for example, if $c(q)$ is quantile q and y is nonnegative then the Lorenz curve is $L(q) = \int_0^{c(q)} y\mathrm{p}(y \mid \mathbf{x})\,dy / \int_0^{\infty} y\mathrm{p}(y \mid \mathbf{x})\,dy$.

This study was confined to a univariate continuous outcome variable y, and the applications were confined to a small number of substantive covariates $\mathbf{x}$. Extension to outcomes that are not strictly continuous appears relatively straightforward, especially for models that modulate discrete outcomes through continuous random variables (e.g. probit and ordered probit models). In the earnings example the

original outcome variable is mixed continuous and discrete, but this feature disappeared when men with earnings of less than $1,000 were excluded from the sample. SMR could be extended to the original sample by labeling the first component in the mixture as the outcome $y_t = 0$. A multivariate generalization of SMR would have many applications, especially in portfolio problems. For an $N \times 1$ outcome vector $\mathbf{y}$, a natural extension is to apply SMR separately to the N components of $\mathrm{p}(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{N} \mathrm{p}(y_i \mid y_1, \ldots, y_{i-1}, \mathbf{x})$. For all but very small values of N, however, one encounters the curse of dimensionality in the specification of the list of covariates well-established in nonparametric regression. If one were to pursue exactly the same approach taken in this study, it would emerge in the very large number of terms of the interactive polynomials $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{z}$ constructed from $y_1, \ldots, y_{i-1}$ and $\mathbf{x}$.

The Dirichlet process prior specification discussed in Section 2.5 emphasizes a nonparametric approach to the mixture in combination with a (typically parsimonious) parametric formulation of the component p_1 in (19), whereas this study explores the impact of increasingly flexible parameterizations of p_1 and p_2 as well as increasing the number of mixture components m. These approaches are complementary, and promising avenues for future research are to either increase the flexibility of p_1 in approaches like that of Dunson and Pillai (2004) and Griffin and Steel (2005), or to take up an explicitly nonparametric specification of the mixture components in SMR. The examples in this study provide only some hints of the practical returns to these extensions. The results in Tables 2 and 3 seem to suggest roughly the same upper bound on the cross-validated log score by expanding in the numbers of components in the covariates $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{z}$, as opposed to the number of mixture components m. Our finding that the posterior quantiles are similar in these cases (recall the comparison of Figures 3 and 5) also suggests that results may not be very sensitive to just how the model is expanded. Much more experience with these kinds of comparisons is required before it is possible to make general recommendations for applied work.

Beyond these practical questions, there is a substantial agenda of work in the econometric theory of inference for $\mathrm{p}(y \mid \mathbf{x})$. For the specific case of $\mathrm{E}(y \mid x)$, x univariate, there are well established results for several approaches to nonparametric regression; see, for example, Härdle (1990) and Green and Silverman (1994), on asymptotic distribution theory. The problem for conditional distributions is more difficult and to our knowledge there are no comparable results. We therefore simply indicate the kinds of questions for which answers would be interesting and useful in the context of the SMR model, while noting that essentially the same issues come up in all approaches. For a suitable distance measure D from one density to another, in the context of (19) what conditions are sufficient for

$$\min_{\boldsymbol{\theta}} D\left[\mathrm{p}(y \mid \mathbf{x}), \mathrm{p}(y \mid \mathbf{x}, m, k, p, q, \boldsymbol{\theta})\right] \to 0$$

with suitable expansion of m, k, p, and q? And under what conditions will the posterior distribution of this measure become concentrated in a neighborhood of 0?

Alternatively, what conditions would guarantee that the posterior distribution of

$$\int \mathrm{f}\left(y\right) \mathrm{p}\left(y \mid \mathbf{x}, m, k, p, q, \boldsymbol{\theta}\right) dy$$

is asymptotically concentrated around $\int \mathrm{f}\left(y\right) \mathrm{p}\left(y \mid \mathbf{x}\right) dy$? Theoretical underpinnings of this kind would be welcome future developments.

References

Albert, J. and S. Chib (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts, Journal of Business and Economic Statistics 11: 1-15.

Angrist, J., V. Chernozhukov and I. Fernandez-Val (2004). Quantile regression under misspecification with an application to the U.S. wage structure. MIT Department of Economics working paper; Econometrica, forthcoming.

Barten, A.P. (1964). Consumer demand functions under conditions of almost additive preferences. Econometrica 32: 1-38.

Bollerslev, T., R.Y. Chou and K.F. Kroner (1992). ARCH modeling in finance: a review of the theory and empirical evidence. Journal of Econometrics 52: 5-59.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: applications of quantile regression. Econometrica 62: 405-458

Celeux, G., M. Hurn and C.P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association 95: 957-970.

Chamberlain, G. and G. Imbens (2003). Nonparametric applications of Bayesian inference. Journal of Business and Economic Statistics 21: 12-18.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. Journal of Econometrics 95: 79-97.

Deb, P, and P.K. Trivedi (1997). Demand for medical care by the elderly in the United States: a finite mixture approach. Journal of Applied Econometrics 12: 313-336.

DeIorio, M., P. Müller, G.L. Rosner and S.N. MacEachern (2004). An ANOVA model for dependent random measures. Journal of the American Statistical Association 99: 205-215.

Draper and Krnjajic (2005). Bayesian model specification. In preparation.

Dueker, M.J. (1997). Markov switching in GARCH processes and mean-reverting stock-market volatility. Journal of Business and Economic Statistics 15: 26-34.

Dunson, D.B. and N. Pillai (2004). Bayesian density regression. Duke University Institute of Statistics and Decision Sciences working paper 2004-33.

Escobar, M.D. and M. West (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90: 577-588.

Fan, J.Q. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85: 645-660.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1: 209-230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. The Annals of Statistics 2: 615-629.

Gelfand, A.E., D.K. Dey and H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods. J.M. Bernardo,

J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), Bayesian Statistics 4. Oxford: Oxford University Press.

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. Journal of the American Statistical Association 99: 799-804.

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

Geweke, J., M. Keane and D. Runkle (1994). Alternative computational approaches to inference in the multinomial probit model. Review of Economics and Statistics 76: 609-632.

Geweke, J. and M. Keane (2000). An empirical analysis of earnings dynamics among men in the PSID: 1968-1989. Journal of Econometrics 92: 293-356.

Geweke, J. and M. Keane (2001). Computationally intensive methods for integration in econometrics. J. Heckman and E.E. Leamer (eds.), Handbook of Econometrics volume 5, 3463-3568. Amsterdam: North-Holland.

Geweke, J. and N. Terui (1993). Bayesian threshold autoregressive models for nonlinear time series. Journal of Time Series Analysis 14: 441-455.

Green, P.J. and R. Sibson (1978). Computing Dirichlet tessellations in the plane. The Computer Journal 21: 268-273.

Green, P.J. and Silverman, B.W. (1994). Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. London: Chapman and Hall.

Griffin, J.E. and M.F.J. Steel (2004). Semiparametric Bayesian inference for stochastic frontier models. Journal of Econometrics 123: 121-152.

Griffin, J.E. and M.F. J. Steel (2005). Order-based dependent Dirichlet processes. Journal of the American Statistical Association, forthcoming.

Härdle, W. (1990). Applied Nonparametric Regression, Cambridge University Press.

Härdle, W. and Tsybakov (1997). Local polynomial estimators of the volatility function in nonparametric regression. Journal of Econometrics 81: 223-242.

Heckman, J.J. and R. Robb (1985). Using longitudinal data to estimate age, period and cohort effects in earnings equations. W.M. Mason and S.E. Fienberg, Cohort Analysis in Social Research. New York: Springer-Verlag.

Hirano, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models. Econometrica 70: 781-799.

Holmes, C.C., D.G.T. Denison and B.K. Mallick (2005). Bayesian prediction via partitioning. Journal of Computational and Graphical Statistics, forthcoming.

Hood, W.C. and T.C. Koopmans (1953). Studies in Econometric Method. Chicago: Cowles Foundation for Research in Economics.

Jacquier, E., N.G. Polson and P.E. Rossi (1994). Bayesian analysis of stochastic volatility models. Journal of Business and Economic Statistics 12: 371-389.

Keane, M. (1992). A note on identification in the multinomial probit model. Journal of Business and Economic Statistics 10: 192-200.

Kim, S., N. Shephard and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. Review of Economic Studies 65: 361-393.

Koenker, R. and G. Bassett (1978). Regression quantiles. Econometrica 46: 33-50.

Koop, G. and D.J. Poirier (2004). Bayesian variants of some classical semiparametric regression techniques. Journal of Econometrics 123: 259-282.

Koopmans, T.C. (1950). Statistical Inference in Dynamic Economic Models. Chicago: Cowles Foundation for Research in Economics.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. Scandinavian Journal of Statistics 5: 81-91.

Manning, W.G., L. Blumber and L.H. Moulton (1995). The demand for alcohol – the differential response to price. Journal of Health Economics 14: 123-148.

Morduch, J.J. and H.S. Stern (1997). Using mixture models to detect sex bias in health outcomes in Bangladseh. Journal of Econometrics 77: 259-276.

Müller, P., A. Erkanli and M. West (1996). Bayesian curve fitting using multivariate normal mixtures. Biometrika 83: 67-79.

Müller, P. and F.A. Quintana (2004). Nonparametric Bayesian analysis. Statistical Science 19: 95-110.

Ruggiero, M. (1994). Bayesian semiparametric estimation of proportional hazards models. Journal of Econometrics 62: 277-300.

Shiller, R.J. (1984). Smoothness priors and nonlinear regression. Journal of the American Statistical Association 79: 609-615.

Smith, M. and R. Kohn (1996). Nonparametric regression using Bayesian variable selection. Journal of Econometrics 75: 317-344.

Theil, H. (1967). Economics and Information Theory. Amsterdam: North-Holland

Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. Journal of the Royal Statistical Society, Series B 40: 364-372.

Wedel, M., W.S. Desarbo, J.R. Bult and V. Ramaswamy, (1993). A latent class Poissson regression model for heterogeneous count data. Journal of Applied Eocnometrics 8: 397-411.

White, H. (1980), A heteroscedsticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrica 48: 817-838.

Yang, S.R. and B.W. Brorsen (1992). Nonlinear dynamics of daily cash prices. American Journal of Agricultural Economics 74: 706-715.

Yu, K.M. and M.C. Jones (1998). Local linear quantile regression. Journal of the American Statistical Association 93: 228-237.

Technical Appendix

This appendix provides the joint distribution of the parameters, latent variables and observables in the model. It derives and presents the conditional posterior distributions that comprise the MCMC posterior simulator described in the text. The Appendix uses precisions rather than variances: $h = 1/\sigma^2$ and $h_j = 1/\sigma_j^2$ $(j = 1, \ldots, m)$.

1. Fixed hyperparameters of the prior distribution

   (a) $\underline{s}^2$, $\underline{\nu}$: Prior distribution of h

   (b) $\underline{\nu}^*$: Prior distribution of h_j $(j = 1, \ldots, m)$

   (c) $\underline{\mu}$, $\underline{\tau}^2_{\boldsymbol{\beta}}$: Prior distribution of $\boldsymbol{\beta}$

   (d) $\underline{\tau}^2_{\boldsymbol{\alpha}}$: Prior distribution of $\boldsymbol{\alpha}_j$ $(j = 1, \ldots, m)$

   (e) $\underline{\tau}^2_{\boldsymbol{\gamma}}$: Prior distribution of $\boldsymbol{\gamma}^*_j$ $(j = 1, \ldots, m-1)$

2. Prior distributions of parameters

   (a) $\underline{s}^2 h \mid (\underline{s}^2, \underline{\nu}) \sim \chi^2(\underline{\nu})$

   (b) $\underline{\nu}^* h_j \overset{iid}{\sim} \chi^2(\underline{\nu}^*)$ $(j = 1, \ldots, m)$

   (c) $\mathbf{C}\boldsymbol{\beta} \mid \left(\underline{\mu}, \underline{\tau}^2_{\boldsymbol{\beta}}\right) \sim N\left[\boldsymbol{\iota}_r \underline{\mu}, \underline{\tau}^2_{\boldsymbol{\beta}} r \mathbf{I}_r\right] \implies \boldsymbol{\beta} \sim N\left(\underline{\boldsymbol{\beta}}, \underline{\mathbf{H}}^{-1}_{\boldsymbol{\beta}}\right)$

   (d) $\mathbf{C}\boldsymbol{\alpha}_j \mid (\sigma^2, \underline{\tau}^2_{\boldsymbol{\alpha}}) \overset{iid}{\sim} N\left[\mathbf{0}_r, \underline{\tau}^2_{\boldsymbol{\alpha}} \sigma^2 r \mathbf{I}_r\right]$ $(j = 1, \ldots, m)$
   $\implies \boldsymbol{\alpha}_j \overset{iid}{\sim} N\left(\mathbf{0}_r, \sigma^2 \underline{\mathbf{H}}^{-1}_{\boldsymbol{\alpha}}\right)$ $(j = 1, \ldots, m)$

   (e) $\mathbf{C}\boldsymbol{\gamma}^*_j \overset{iid}{\sim} N\left[\mathbf{0}_r, \underline{\tau}^2_{\boldsymbol{\gamma}} r \mathbf{I}_r\right]$ $(j = 1, \ldots, m-1)$
   $\implies \boldsymbol{\gamma}^*_j \overset{iid}{\sim} N\left(\mathbf{0}_r, \underline{\mathbf{H}}^{-1}_{\boldsymbol{\gamma}^*}\right)$ $(j = 1, \ldots, m-1)$

3. Groupings and transformations of parameters

   (a) $\mathbf{h} = (h_1, \ldots, h_m)'$

   (b) $\mathbf{A} = \left[\begin{array}{ccc} \boldsymbol{\alpha}_1 & \cdots & \boldsymbol{\alpha}_m \end{array}\right]$, $\boldsymbol{\alpha} = \text{vec}(\mathbf{A})$

   (c) $\boldsymbol{\delta}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$

   (d) $\boldsymbol{\Gamma}^{*\prime} = \left[\begin{array}{ccc} \boldsymbol{\gamma}^*_1 & \cdots & \boldsymbol{\gamma}^*_{m-1} \end{array}\right]$, $\boldsymbol{\gamma}^* = \text{vec}(\boldsymbol{\Gamma}^*)$

   (e) Define $\mathbf{p}_1 = \boldsymbol{\iota}_m m^{-1/2}$, $\underset{m \times m}{\mathbf{P}} = \left[\begin{array}{cc} \mathbf{p}_1 & \mathbf{P}_2 \end{array}\right] : \mathbf{P}\mathbf{P}' = \mathbf{P}'\mathbf{P} = \mathbf{I}_m$; $\boldsymbol{\Gamma} = \mathbf{P}_2 \boldsymbol{\Gamma}^*$

   Remark. This induces a degenerate normal prior on $\boldsymbol{\Gamma}$.

$$\boldsymbol{\Gamma} = \mathbf{P} \cdot \left[\begin{array}{c} \mathbf{0}' \\ \boldsymbol{\Gamma}^* \end{array}\right] \implies \boldsymbol{\Gamma}' = \left[\begin{array}{cc} \mathbf{0} & \boldsymbol{\Gamma}^{*\prime} \end{array}\right] \cdot \left[\begin{array}{c} \mathbf{p}_1' \\ \mathbf{P}_2' \end{array}\right] = \boldsymbol{\Gamma}^{*\prime} \mathbf{P}_2'.$$

Then

$$vec\left(\boldsymbol{\Gamma}'\right) = vec\left(\boldsymbol{\Gamma}^{*\prime}\mathbf{P}_2'\right) = \left(\mathbf{P}_2 \otimes \mathbf{I}_q\right) vec\left(\boldsymbol{\Gamma}^{*\prime}\right)$$

has mean $\mathbf{0}_{(m-1)q}$ and variance

$$\left(\mathbf{P}_2 \otimes \mathbf{I}_q\right)\left(\mathbf{I}_{m-1} \otimes \underset{\sim}{\mathbf{H}}_{\boldsymbol{\gamma}^*}^{-1}\right)\left(\mathbf{P}_2' \otimes \mathbf{I}_q\right) = \mathbf{P}_2\mathbf{P}_2' \otimes \underset{\sim}{\mathbf{H}}_{\boldsymbol{\gamma}^*}^{-1} = \left[\mathbf{I}_m - \mathbf{p}_1\mathbf{p}_1'\right] \otimes \underset{\sim}{\mathbf{H}}_{\boldsymbol{\gamma}^*}^{-1}.$$

Hence the choice of $\mathbf{P}_2$ is of no consequence.

4. Distributions of latent variables

   (a) $\widetilde{\mathbf{w}}_t \mid (\boldsymbol{\Gamma}, \mathbf{z}_t) \sim N\left(\boldsymbol{\Gamma}\mathbf{z}_t, \mathbf{I}_m\right)$ $(t = 1, \ldots, T)$

   (b) $\widetilde{s}_t \mid \widetilde{\mathbf{w}}_t = j$ iff $\widetilde{w}_{tj} \geq \widetilde{w}_{ti}$ $(i = 1, \ldots, m; j = 1, \ldots, m)$

5. Groupings and transformations of latent variables

   (a) $\widetilde{\mathbf{s}}' = (\widetilde{s}_1, \ldots, \widetilde{s}_T)$ and $\underset{T\times m}{\widetilde{\mathbf{S}}} = [d_{tj}]$, $d_{tj} = \delta\left(\widetilde{s}_t, j\right)$

   (b) $T_j = \sum_{t=1}^{T} \delta\left(\widetilde{s}_t, j\right)$

   (c) $\widetilde{\mathbf{H}} = \mathrm{diag}\left(h_{\widetilde{s}_1}, \ldots, h_{\widetilde{s}_T}\right)$

   (d) $\widetilde{\mathbf{w}}_t^* = \mathbf{P}'\widetilde{\mathbf{w}}_t$, $\widetilde{\mathbf{W}}^{*\prime} = \left[\begin{array}{ccc} \widetilde{\mathbf{w}}_1^* & \cdots & \widetilde{\mathbf{w}}_T^* \end{array}\right]$, $\widetilde{\mathbf{W}}^* = \left[\begin{array}{ccc} \widetilde{\mathbf{w}}_{(1)}^* & \cdots & \widetilde{\mathbf{w}}_{(m)}^* \end{array}\right]$

6. Prior density kernels

   (a) $\mathrm{p}\left(h \mid \underline{s}^2, \underline{\nu}\right) \propto h^{(\underline{\nu}-2)/2} \exp\left(-\underline{s}^2 h/2\right)$

   (b) $\mathrm{p}\left(\mathbf{h} \mid \underline{\nu}^*\right) \propto \prod_{j=1}^{m} h_j^{(\underline{\nu}^*-2)/2} \exp\left(-\underline{\nu}^* h_j/2\right)$

   (c) $\mathrm{p}\left(\boldsymbol{\beta} \mid \underline{\boldsymbol{\beta}}, \underset{\sim}{\mathbf{H}}_{\boldsymbol{\beta}}\right) \propto \exp\left[-\left(\boldsymbol{\beta}-\underline{\boldsymbol{\beta}}\right)' \underset{\sim}{\mathbf{H}}_{\boldsymbol{\beta}} \left(\boldsymbol{\beta}-\underline{\boldsymbol{\beta}}\right)/2\right]$

   (d) $\mathrm{p}\left(\boldsymbol{\alpha} \mid h, \underset{\sim}{\mathbf{H}}_{\boldsymbol{\alpha}}\right) \propto h^{mT/2} \exp\left\{-\boldsymbol{\alpha}'\left[\mathbf{I}_m \otimes \left(h \cdot \underset{\sim}{\mathbf{H}}_{\boldsymbol{\alpha}}\right)\right]\boldsymbol{\alpha}/2\right\}$

   $$= h^{mT/2} \exp\left[-\sum_{j=1}^{m} \boldsymbol{\alpha}_j' \left(h \cdot \underset{\sim}{\mathbf{H}}_{\boldsymbol{\alpha}}\right)\boldsymbol{\alpha}_j/2\right]$$

   (e) $\mathrm{p}\left(\boldsymbol{\gamma}^* \mid \underset{\sim}{\mathbf{H}}_{\boldsymbol{\gamma}^*}\right) \propto \exp\left(-\sum_{j=1}^{m-1} \boldsymbol{\gamma}_j^{*\prime} \underset{\sim}{\mathbf{H}}_{\boldsymbol{\gamma}^*} \boldsymbol{\gamma}_j^*/2\right)$

7. Latent vector density kernels

   (a) $\mathrm{p}\left(\widetilde{\mathbf{W}} \mid \boldsymbol{\Gamma}, \mathbf{z}_t\right) \propto \exp\left[-\sum_{t=1}^{T} \left(\widetilde{\mathbf{w}}_t - \boldsymbol{\Gamma}\mathbf{z}_t\right)'\left(\widetilde{\mathbf{w}}_t - \boldsymbol{\Gamma}\mathbf{z}_t\right)/2\right]$

   $$\Longrightarrow \mathrm{p}\left(\widetilde{\mathbf{W}}^* \mid \boldsymbol{\Gamma}^*, \mathbf{Z}\right) \propto$$

   $$\exp\left(-\sum_{t=1}^{T}\left(w_{1t}^*/2\right)\right) \exp\left[-\sum_{j=1}^{m-1}\left(\widetilde{\mathbf{w}}_{(j+1)}^* - \mathbf{Z}\boldsymbol{\gamma}_j^*\right)'\left(\widetilde{\mathbf{w}}_{(j+1)}^* - \mathbf{Z}\boldsymbol{\gamma}_j^*\right)/2\right]$$

(b) $\mathrm{p}\left(\widetilde{s}_t = j \mid \mathbf{w}_t\right) = \prod_{i=1}^{m} I_{(-\infty, w_{tj}]}\left(w_{ti}\right)$ $(j = 1, \ldots, m; t = 1, \ldots, T)$

8. Observables

   (a) Covariates $\mathbf{u}_t$ $(t = 1, \ldots, T)$; $\mathbf{U}' = \left[\begin{array}{ccc} \mathbf{u}_1 & \cdots & \mathbf{u}_T \end{array}\right]$
   (b) Covariates $\mathbf{v}_t$ $(t = 1, \ldots, T)$; $\mathbf{V}' = \left[\begin{array}{ccc} \mathbf{v}_1 & \cdots & \mathbf{v}_T \end{array}\right]$
   (c) Covariates $\mathbf{z}_t$ $(t = 1, \ldots, T)$; $\mathbf{Z}' = \left[\begin{array}{ccc} \mathbf{z}_1 & \cdots & \mathbf{z}_T \end{array}\right]$
   (d) Outcomes y_t $(t = 1, \ldots, T)$; $\mathbf{y}' = (y_1, \ldots, y_T)$

9. Conditional density of outcomes $\mathbf{y}$ (equivalent expressions)

   (a) $\mathrm{p}\left(\mathbf{y} \mid \widetilde{\mathbf{s}}, \mathbf{A}, \boldsymbol{\beta}, h, \mathbf{h}, \mathbf{U}, \mathbf{V}\right)$

$$\propto h^{T/2} \left(\prod_{j=1}^{m} h_j^{T_j/2} \right) \exp\left[-h \sum_{t=1}^{T} h_{\widetilde{s}_t} \left(y_t - \boldsymbol{\alpha}_{\widetilde{s}_t}' \mathbf{v}_t - \boldsymbol{\beta}' \mathbf{u}_t \right)^2 / 2 \right]$$

   (b) $\mathrm{p}\left(\mathbf{y} \mid \widetilde{\mathbf{s}}, \mathbf{A}, \boldsymbol{\beta}, h, \mathbf{h}, \mathbf{U}, \mathbf{V}\right)$

$$\propto h^{T/2} \left| \widetilde{\mathbf{H}} \right|^{1/2} \exp\left\langle -h \left\{ \mathbf{y} - \left[\left(\boldsymbol{\iota}_m' \otimes \mathbf{V} \right) \circ \left(\widetilde{\mathbf{S}} \otimes \boldsymbol{\iota}_p' \right) \right] \boldsymbol{\alpha} - \mathbf{U}\boldsymbol{\beta} \right\}' \widetilde{\mathbf{H}} \left\{ \mathbf{y} - \left[\left(\boldsymbol{\iota}_m' \otimes \mathbf{V} \right) \circ \left(\widetilde{\mathbf{S}} \otimes \boldsymbol{\iota}_p' \right) \right] \boldsymbol{\alpha} - \mathbf{U}\boldsymbol{\beta} \right\} / 2 \right\rangle .$$

10. Conditional posterior distributions of parameter blocks

   (a) From (6a), (6d), (9a),

$$\mathrm{p}\left(h \mid \underline{s}^2, \underline{\nu}, \mathbf{h}, \boldsymbol{\beta}, \mathbf{A}, \widetilde{\mathbf{s}}, \mathbf{y}, \mathbf{U}, \mathbf{V}\right) \propto h^{(\underline{\nu} + mp + T - 2)/2}$$
$$\cdot \exp\left\{ - \left[\underline{s}^2 + \sum_{j=1}^{m} \boldsymbol{\alpha}_j' \underline{\mathbf{H}}_{\boldsymbol{\alpha}} \boldsymbol{\alpha}_j + h \sum_{t=1}^{T} h_{\widetilde{s}_t} \left(y_t - \boldsymbol{\alpha}_{\widetilde{s}_t}' \mathbf{v}_t - \boldsymbol{\beta}' \mathbf{u}_t \right)^2 \right] h/2 \right\} .$$

   Thus $\overline{s}^2 h \sim \chi^2\left(\overline{\nu}\right)$ where $\overline{\nu} = \underline{\nu} + mp + T$ and

$$\overline{s}^2 = \underline{s}^2 + \sum_{j=1}^{m} \boldsymbol{\alpha}_j' \underline{\mathbf{H}}_{\boldsymbol{\alpha}} \boldsymbol{\alpha}_j + \sum_{t=1}^{T} h_{\widetilde{s}_t} \left(y_t - \boldsymbol{\alpha}_{\widetilde{s}_t}' \mathbf{v}_t - \boldsymbol{\beta}' \mathbf{u}_t \right)^2 .$$

   (b) From (6b) and (9a), $\mathrm{p}\left(\mathbf{h} \mid \underline{\nu}^*, h, \boldsymbol{\beta}, \mathbf{A}, \widetilde{\mathbf{s}}, \mathbf{y}, \mathbf{U}, \mathbf{V}\right) \propto$

$$\prod_{j=1}^{m} h_j^{(\underline{\nu}^* + T_j - 2)/2} \exp\left\{ - \left[\underline{\nu}^* + h \sum_{t=1}^{T} \delta\left(\widetilde{s}_t, j\right) \left(y_t - \boldsymbol{\alpha}_{\widetilde{s}_t}' \mathbf{v}_t - \boldsymbol{\beta}' \mathbf{u}_t \right)^2 \right] h_j/2 \right\} .$$

Thus $\{h_1, \ldots, h_m\}$ are conditionally independent: $\overline{s}_j^2 h_j \sim \chi^2(\overline{\nu}_j)$ with $\overline{\nu}_j = \underline{\nu}^* + T_j$ and

$$\overline{s}_j^2 = \underline{\nu}^* + h \sum_{t=1}^{T} \delta(\widetilde{s}_t, j) \left(y_t - \boldsymbol{\alpha}'_{\widetilde{s}_t} \mathbf{v}_t - \boldsymbol{\beta}' \mathbf{u}_t\right)^2.$$

(c) From (6c), (6d) and (9b),

$$\mathrm{p}\left(\boldsymbol{\delta} \mid \underline{\boldsymbol{\beta}}, \underline{\mathbf{H}}_{\boldsymbol{\beta}}, \underline{\mathbf{H}}_{\boldsymbol{\alpha}}, h, \mathbf{h}, \widetilde{\mathbf{s}}, \mathbf{y}, \mathbf{U}, \mathbf{V}\right) \propto \exp\left[-\left(\boldsymbol{\delta} - \overline{\boldsymbol{\delta}}\right)' \overline{\mathbf{H}}_{\boldsymbol{\delta}} \left(\boldsymbol{\delta} - \overline{\boldsymbol{\delta}}\right)\right],$$

where

$$\overline{\mathbf{H}}_{\boldsymbol{\delta}} = \begin{bmatrix} \overline{\mathbf{H}}_{11} & \left[(\boldsymbol{\iota}'_m \otimes \mathbf{V}) \circ \left(\widetilde{\mathbf{S}} \otimes \boldsymbol{\iota}'_p\right)\right]' \widetilde{\mathbf{H}}\mathbf{U} \\ \mathbf{U}'\widetilde{\mathbf{H}} \left[(\boldsymbol{\iota}'_m \otimes \mathbf{V}) \circ \left(\widetilde{\mathbf{S}} \otimes \boldsymbol{\iota}'_p\right)\right] & \underline{\mathbf{H}}_{\boldsymbol{\beta}} + \mathbf{U}'\widetilde{\mathbf{H}}\mathbf{U} \end{bmatrix}$$

with

$$\overline{\mathbf{H}}_{11} = \mathbf{I}_m \otimes (h \cdot \underline{\mathbf{H}}_{\boldsymbol{\alpha}}) + h \cdot \left[(\boldsymbol{\iota}'_m \otimes \mathbf{V}) \circ \left(\widetilde{\mathbf{S}} \otimes \boldsymbol{\iota}'_p\right)\right]' \widetilde{\mathbf{H}} \left[(\boldsymbol{\iota}'_m \otimes \mathbf{V}) \circ \left(\widetilde{\mathbf{S}} \otimes \boldsymbol{\iota}'_p\right)\right],$$

and $\overline{\boldsymbol{\delta}} = \overline{\mathbf{H}}_{\boldsymbol{\delta}}^{-1} \overline{\mathbf{v}}_{\boldsymbol{\delta}}$ with

$$\overline{\mathbf{v}}_{\boldsymbol{\delta}} = \begin{bmatrix} \left[(\boldsymbol{\iota}'_m \otimes \mathbf{V}) \circ \left(\widetilde{\mathbf{S}} \otimes \boldsymbol{\iota}'_p\right)\right]' \widetilde{\mathbf{H}}\mathbf{y} \\ \underline{\mathbf{H}}_{\boldsymbol{\beta}} \underline{\boldsymbol{\beta}} + \mathbf{U}'\widetilde{\mathbf{H}}\mathbf{y} \end{bmatrix}.$$

(d) From (6e) and (7a),

$$\mathrm{p}\left(\boldsymbol{\gamma}^* \mid \underline{\mathbf{H}}_{\boldsymbol{\gamma}}, \widetilde{\mathbf{W}}^*, \mathbf{Z}\right) \propto \prod_{j=1}^{m-1} \exp\left[-\left(\boldsymbol{\gamma}_j^* - \overline{\boldsymbol{\gamma}}_j^*\right)' \overline{\mathbf{H}}_{\boldsymbol{\gamma}^*} \left(\boldsymbol{\gamma}_j^* - \overline{\boldsymbol{\gamma}}_j^*\right)\right],$$

where $\overline{\mathbf{H}}_{\boldsymbol{\gamma}^*} = \underline{\mathbf{H}}_{\boldsymbol{\gamma}^*} + \mathbf{Z}'\mathbf{Z}$ and $\overline{\boldsymbol{\gamma}}_j^* = \overline{\mathbf{H}}_{\boldsymbol{\gamma}^*}^{-1} \mathbf{Z}' \widetilde{\mathbf{w}}_{(j+1)}^*$ $(j = 1, \ldots, m-1)$. This implies the conditionally independent distributions

$$\gamma_j^* \sim N\left(\overline{\boldsymbol{\gamma}}_j^*, \underline{\mathbf{H}}_{\boldsymbol{\gamma}} + \mathbf{Z}'\mathbf{Z}\right) \quad (j = 1, \ldots, m-1).$$

11. Conditional posterior distributions of latent variables

(a) From (7a), (7b) and (9a), $\mathrm{p}\left(\widetilde{\mathbf{W}} \mid h, \mathbf{h}, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\Gamma}, \mathbf{y}, \mathbf{U}, \mathbf{V}, \mathbf{Z}\right) \propto$

$$\prod_{t=1}^{T} \left\{ \exp\left[-\left(\widetilde{\mathbf{w}}_t - \boldsymbol{\Gamma}\mathbf{z}_t\right)' \left(\widetilde{\mathbf{w}}_t - \boldsymbol{\Gamma}\mathbf{z}_t\right)/2\right] \right.$$
$$\cdot \sum_{j=1}^{m} \left[\prod_{i=1}^{m} I_{(-\infty, \widetilde{w}_{jt}]}(\widetilde{w}_{ti})\right]$$
$$\left. \cdot h_j^{1/2} \exp\left[-h \cdot h_j \left(y_t - \boldsymbol{\alpha}'_j \mathbf{v}_t - \boldsymbol{\beta}' \mathbf{u}_t\right)^2/2\right] \right\}.$$

Thus the latent vectors $\widetilde{\mathbf{w}}_t$ are conditionally independent. This conditional distribution is managed using the Hastings-Metropolis step described in the text.

(b) From (7b), $\widetilde{s}_t \mid \mathbf{w}_t = \{j : w_{tj} \geq w_{ti}\,(i = 1, \ldots, m)\}$.

Table 1
Summary of Model Structures

Model	Parameter configuration	Description
A	$q = 1$, $k > 1$, $p = 1$	Fixed mixture of disturbances in regression
B	$q = 1$, $k = 1$, $p > 1$	Fixed mixture of linear regressions
C	$q > 1$, $k = 1$, $p = 1$	Smooth mixture of normal distributions
D	$q > 1$, $k > 1$, $p = 1$, $\mathbf{u}_t = \mathbf{z}_t$	Smooth mixture of disturbances in regression
E	$q > 1$, $k = 1$, $p > 1$, $\mathbf{v}_t = \mathbf{z}_t$	Smooth mixture of linear regressions

Table 2
Modified Cross-Validated Log Score
Models for Earnings

			Model specification				
m	L_a	L_b	A	B	C	D	E
2	2	1	-527.9	-525.8	-587.5	-524.7	-526.6
2	3	1	-528.0	-523.3	-586.9	-527.2	-522.6
2	4	2	-525.3	-521.7	-586.2	-523.6	-521.1
3	2	1	-526.6	-520.2	-546.9	-525.3	-519.1
3	3	1	-525.9	-518.3	-547.3	-525.0	-520.2
3	4	2	-523.9	-516.3	-540.7	-521.1	-516.1
3	5	3	-524.4	-519.5	-540.6	-523.3	-523.5
4	2	1	-526.0	-516.4	-534.8	-523.3	-517.2
4	3	1	-525.5	-520.8	-534.5	-524.2	-518.3
4	4	2	-523.7	-517.6	-534.8	-521.0	-519.1
4	5	3	-526.3	-518.0	-534.0	-524.6	-518.0

Table 3
Modified Cross-Validated Log Score
Models for S&P 500 Returns

Panel A: Model specification and m ($L_a = L_b = 2$, $g = 0.95$, $\kappa = 1.0$)			
	$m = 2$	$m = 3$	$m = 4$
A	-647.1	-645.1	-645.1
B	-650.4	-625.9	-611.1
C	-618.9	-609.2	-609.4
D	-618.8	-610.5	-608.6
E	-617.2	-613.2	-607.1
Panel B: Polynomial orders (Model C, $m = 3$, $g = 0.95$, $\kappa = 1.0$)			
	$L_b = 1$	$L_b = 2$	$L_b = 3$
$L_a = 1$	-612.9	-610.7	-611.1
$L_a = 2$	-610.9	-609.2	-609.2
$L_a = 3$	-610.5	-610.4	-611.3
Panel C: Volatility specification ($L_a = L_b = 2$, Model C, $m = 3$)			
	$k = 0.7$	$k = 1.0$	$k = 1.5$
$g = 0.90$	-619.7	-612.2	-613.4
$g = 0.95$	-609.4	-609.2	-610.0
$g = 0.98$	-609.6	-610.4	-610.4

Table 4
Out of sample model comparisons

Model	Predictive likelihood	Recursive ML
SMR	-1602.0	
t-GARCH(1,1)	-1625.5	-1624.7
Threshold EGARCH(1,1)		-1637.5
GARCH(1,1)		-1660.5
Normal iid		-1848.5

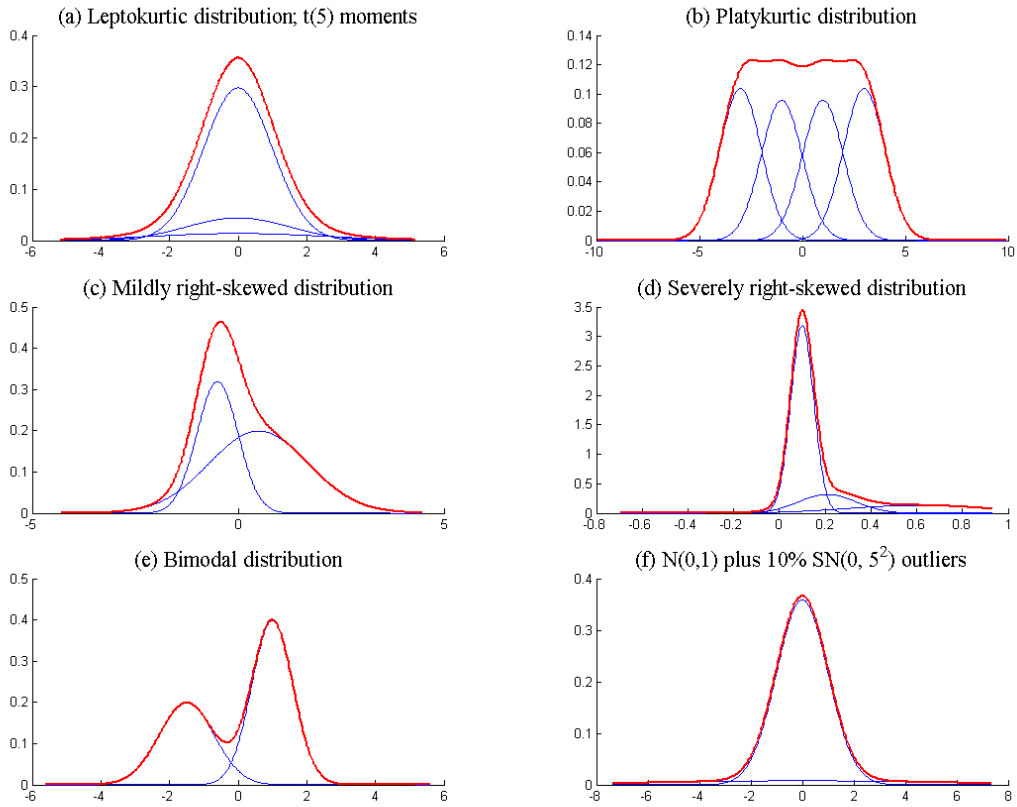

Figure 1: Some normal mixture densities (red lines), with component normal densities multiplied by probabilities p_i (blue lines).

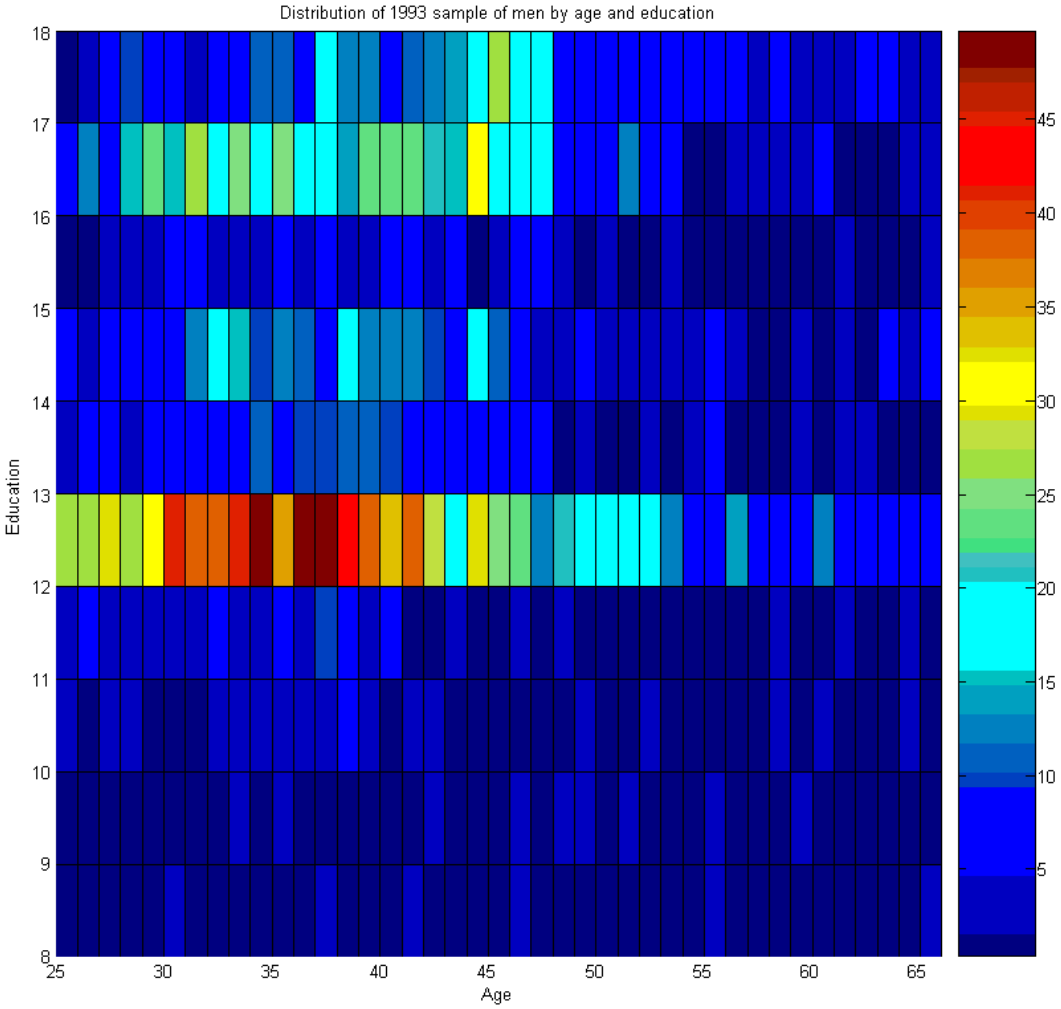

Figure 2: Distribution of the covariates in the earnings data

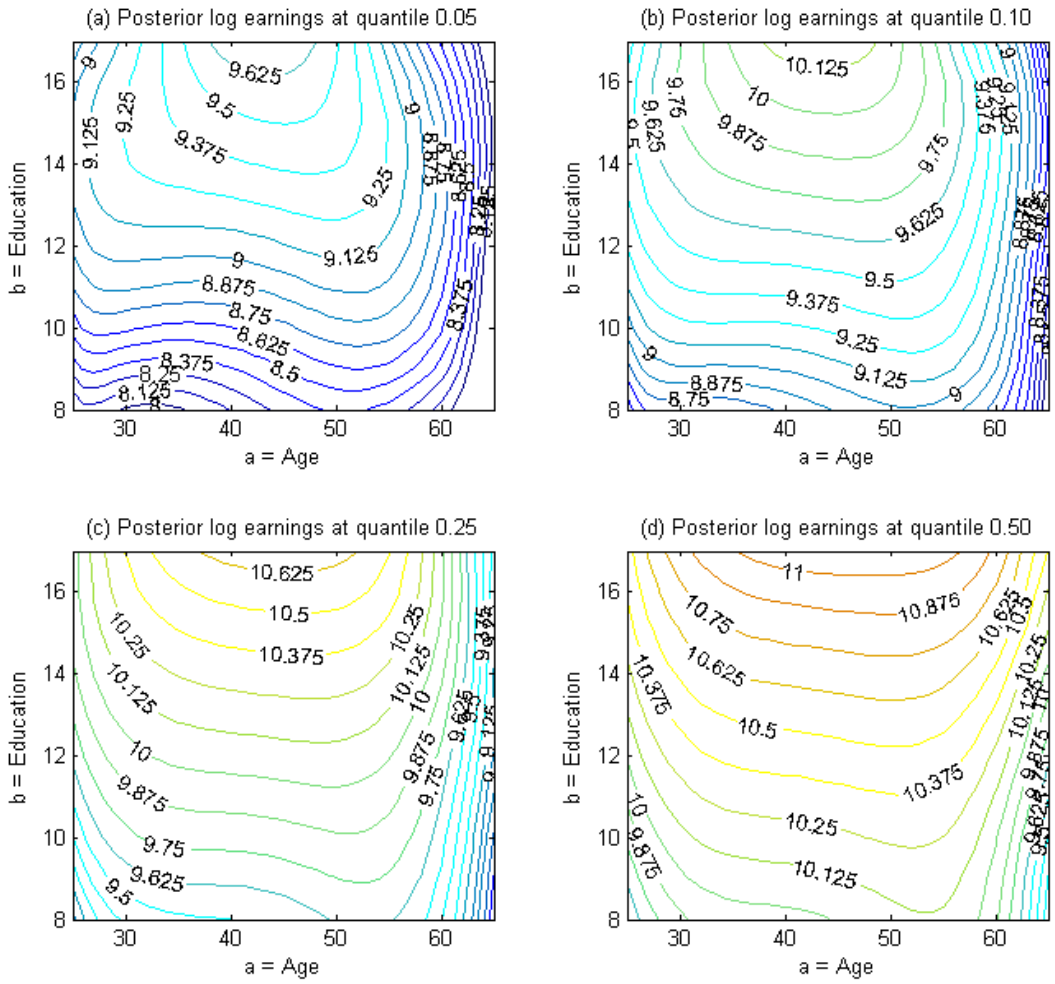

Figure 3: Quantiles of the posterior conditional distribution of earnings for model E, $m = 3$ components, polynomial orders $L_a = 4$ and $L_b = 2$, PSID illustration.

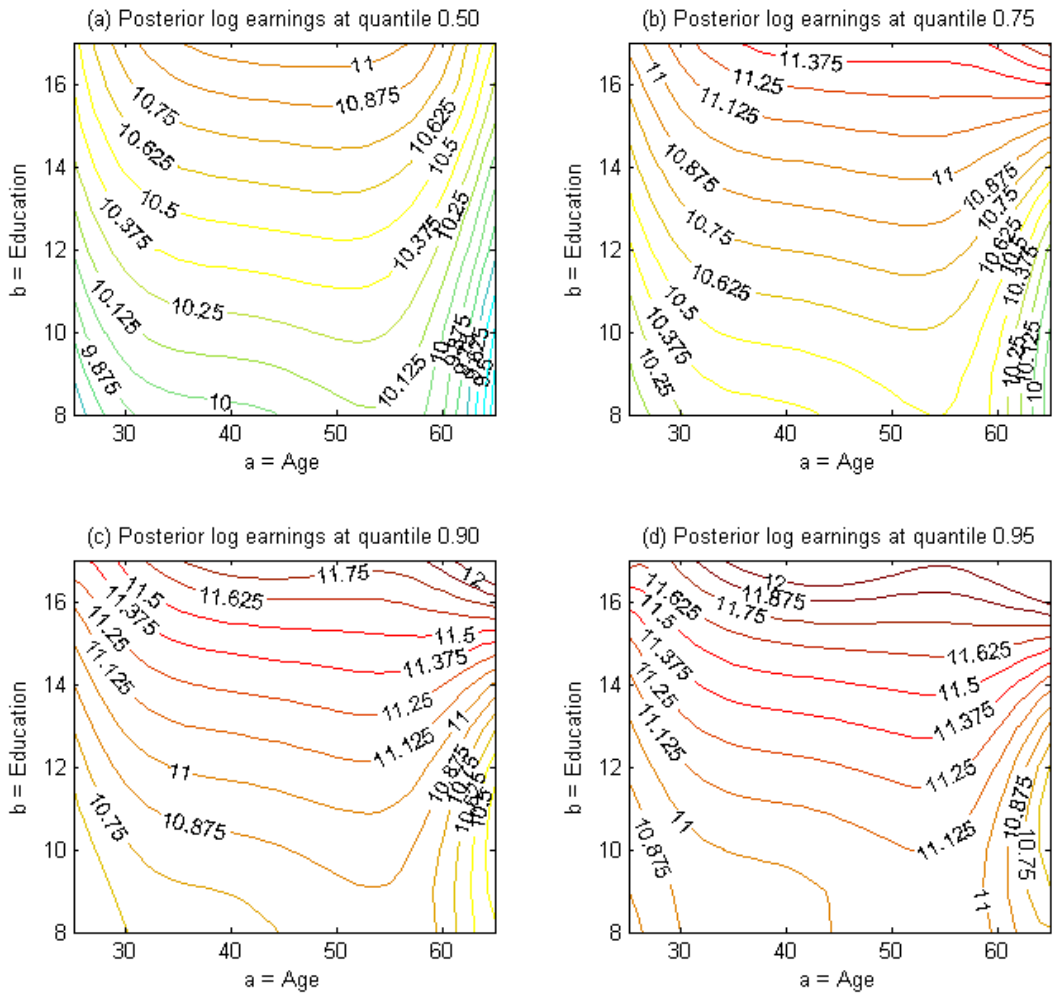

Figure 4: Quantiles of the posterior conditional distribution of earnings for model E, $m = 3$ components, polynomial orders $L_a = 4$ and $L_b = 2$, PSID illustration.

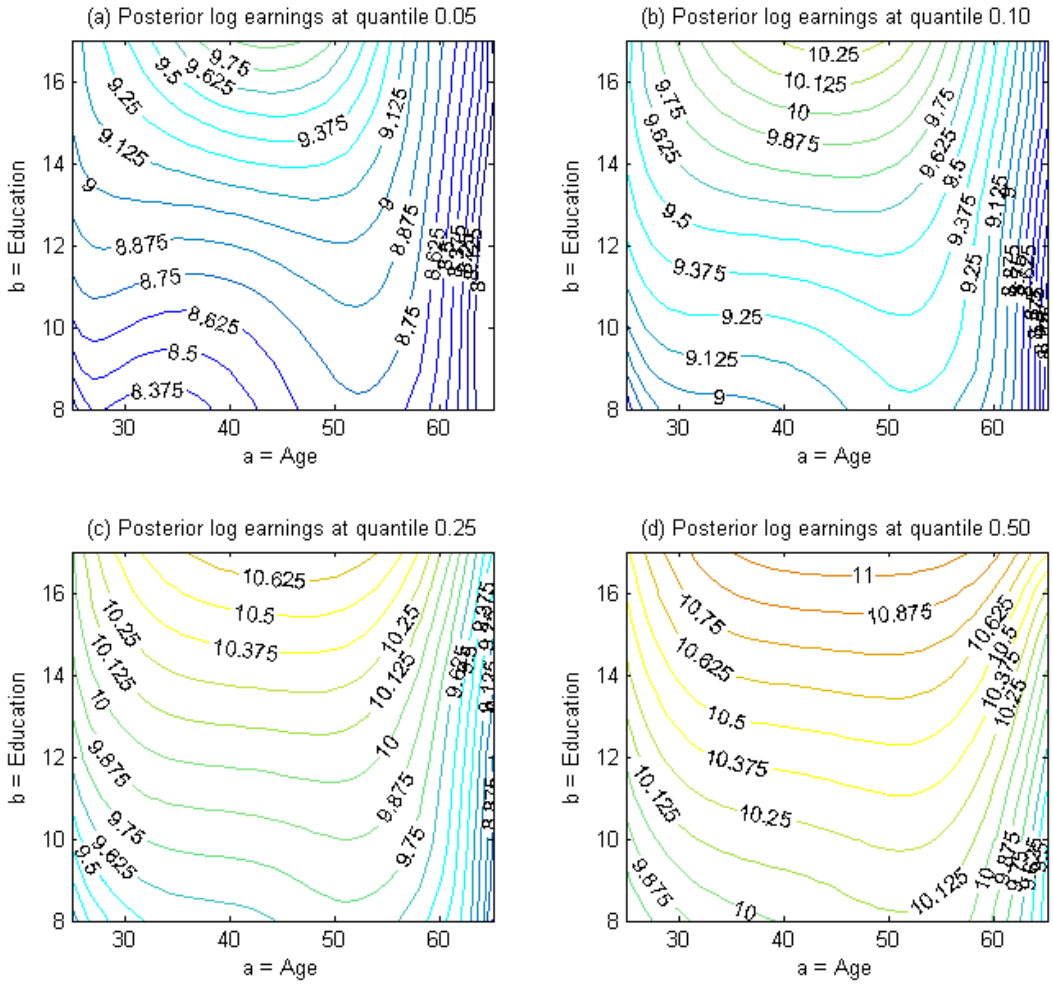

Figure 5: Quantiles of the posterior conditional distribution of earnings for model B, $m = 3$ components, polynomial orders $L_a = 4$ and $L_b = 2$, PSID illustration.

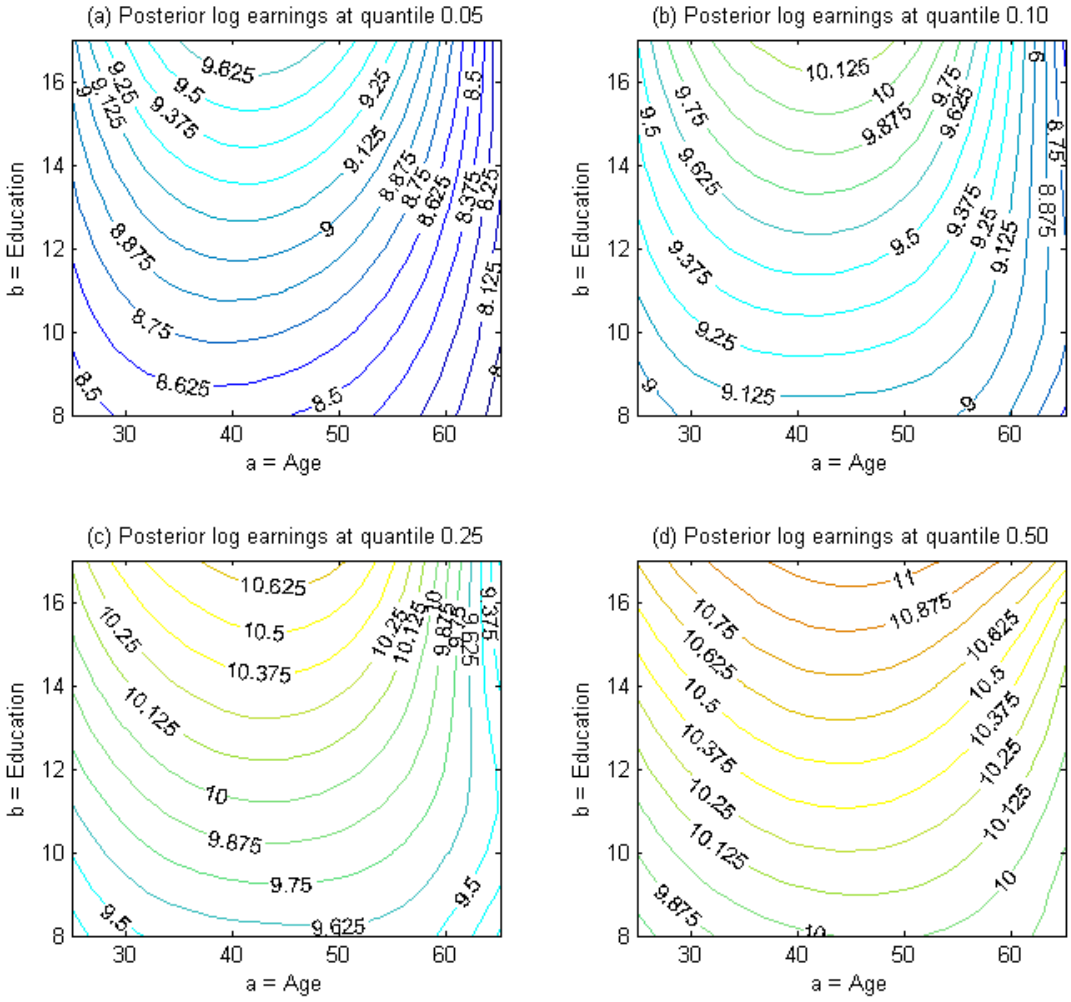

Figure 6: Quantiles of the posterior conditional distribution of earnings for model E, $m = 3$ components, polynomial orders $L_a = 2$ and $L_b = 1$, PSID illustration.

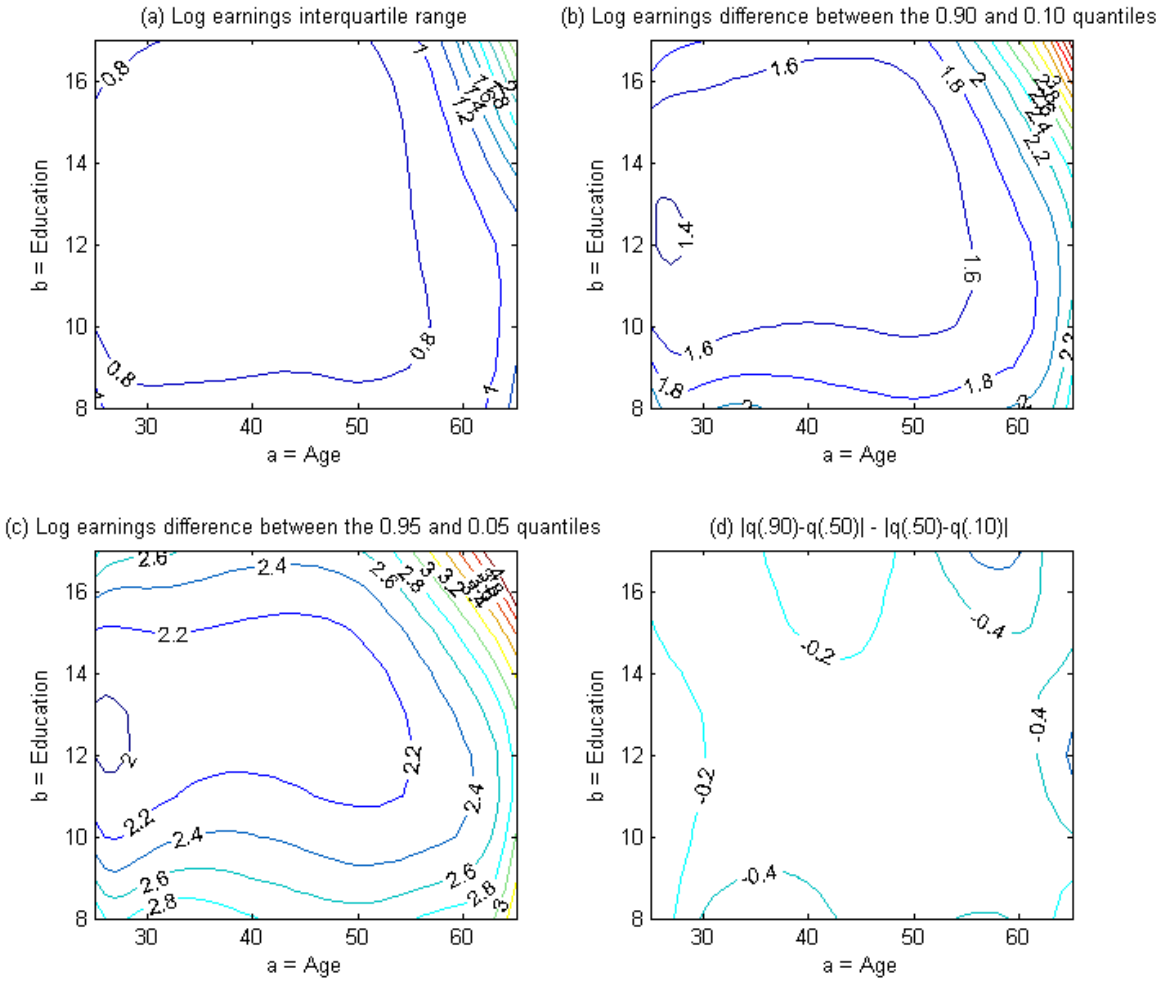

Figure 7: Aspects of the dispersion of the posterior conditional distribution of earnings, PSID illustration

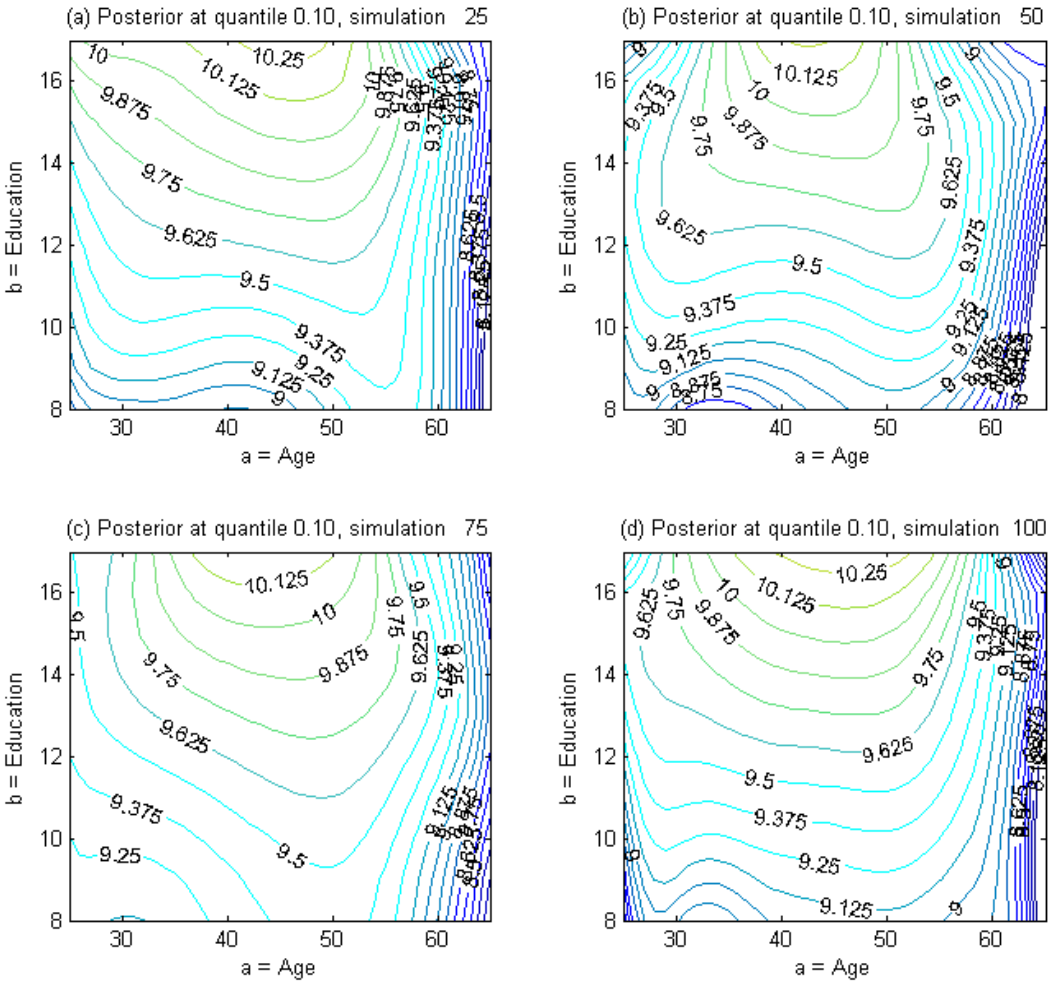

Figure 8: Four random draws from the posterior distribution of the population 10% quantile of the conditional distribution of earnings, PSID illustration

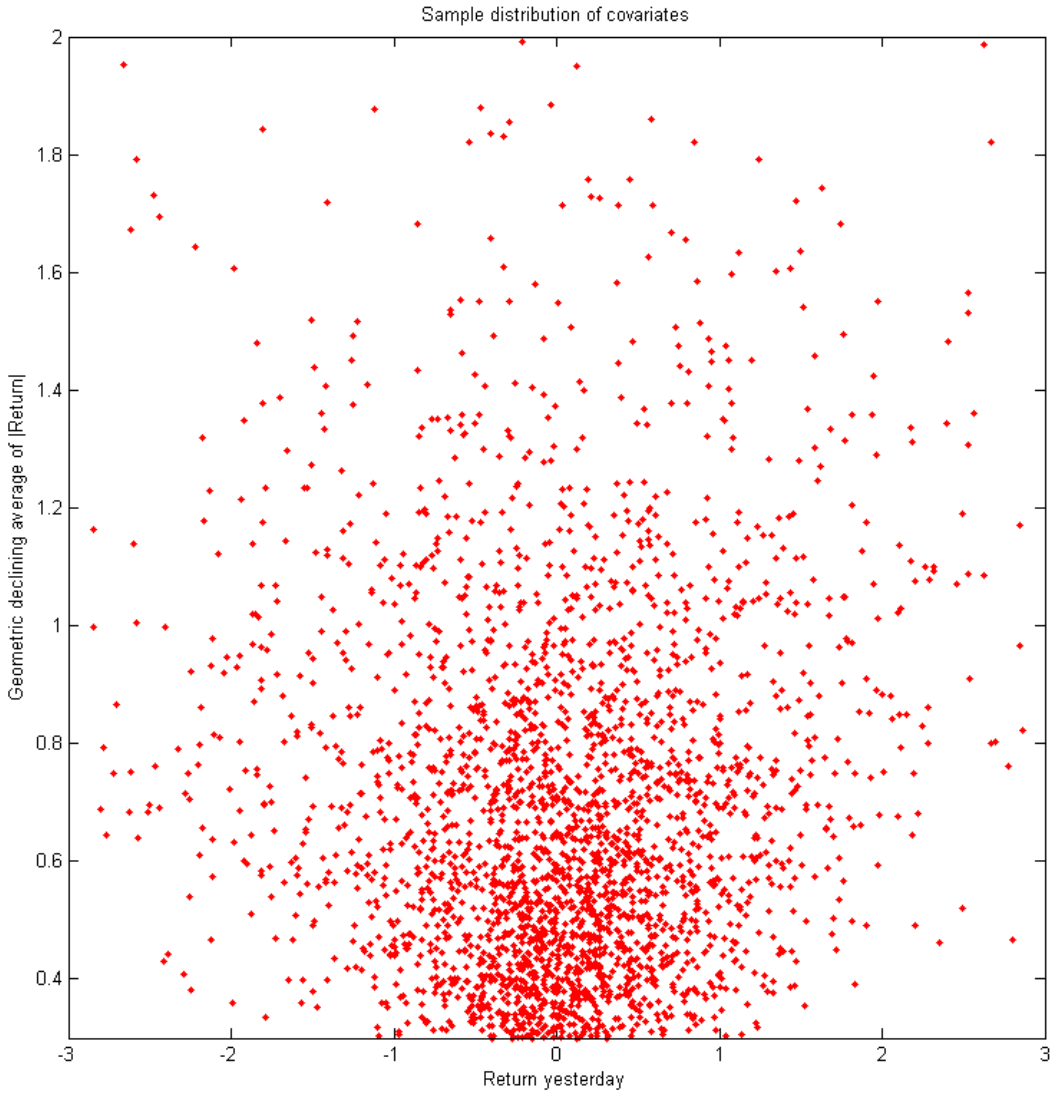

Figure 9: Sample distribution of a and b, S&P 500 returns illustration

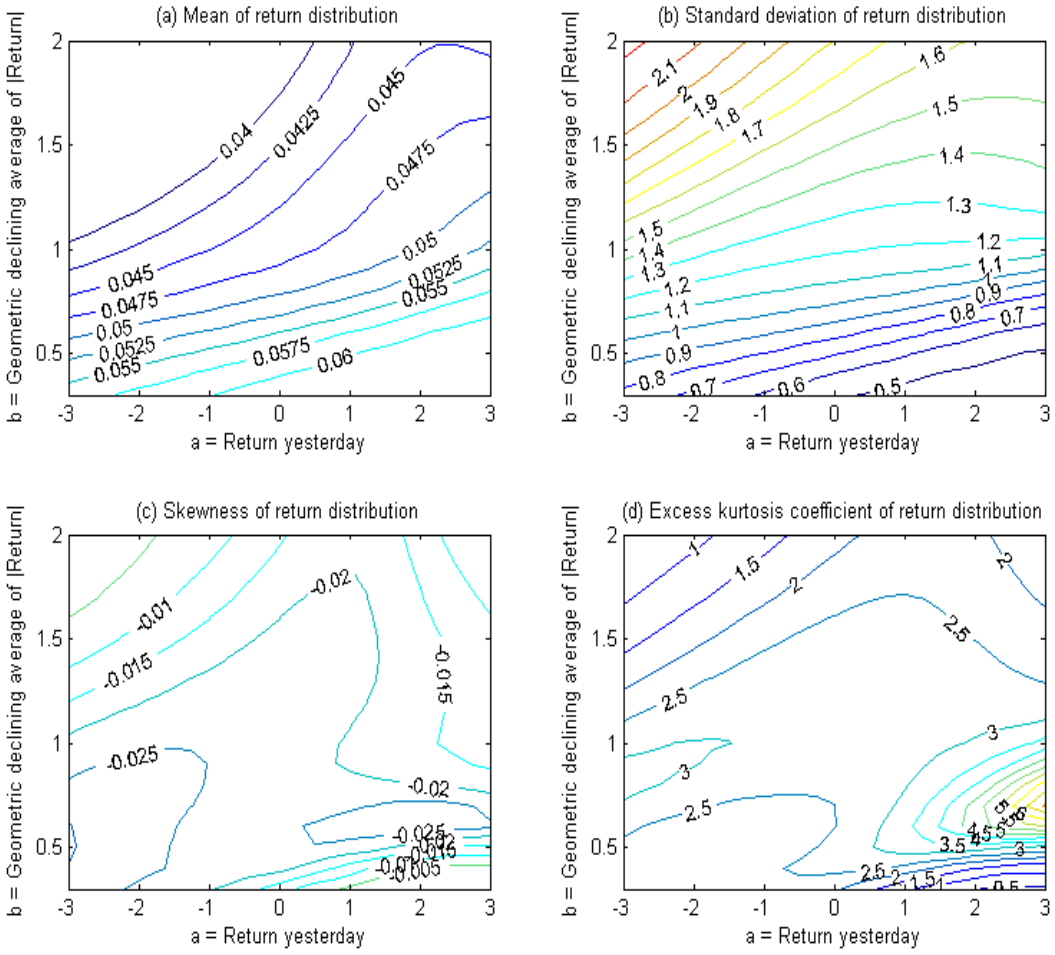

Figure 10: Posterior means of four population conditional moments, S&P 500 example

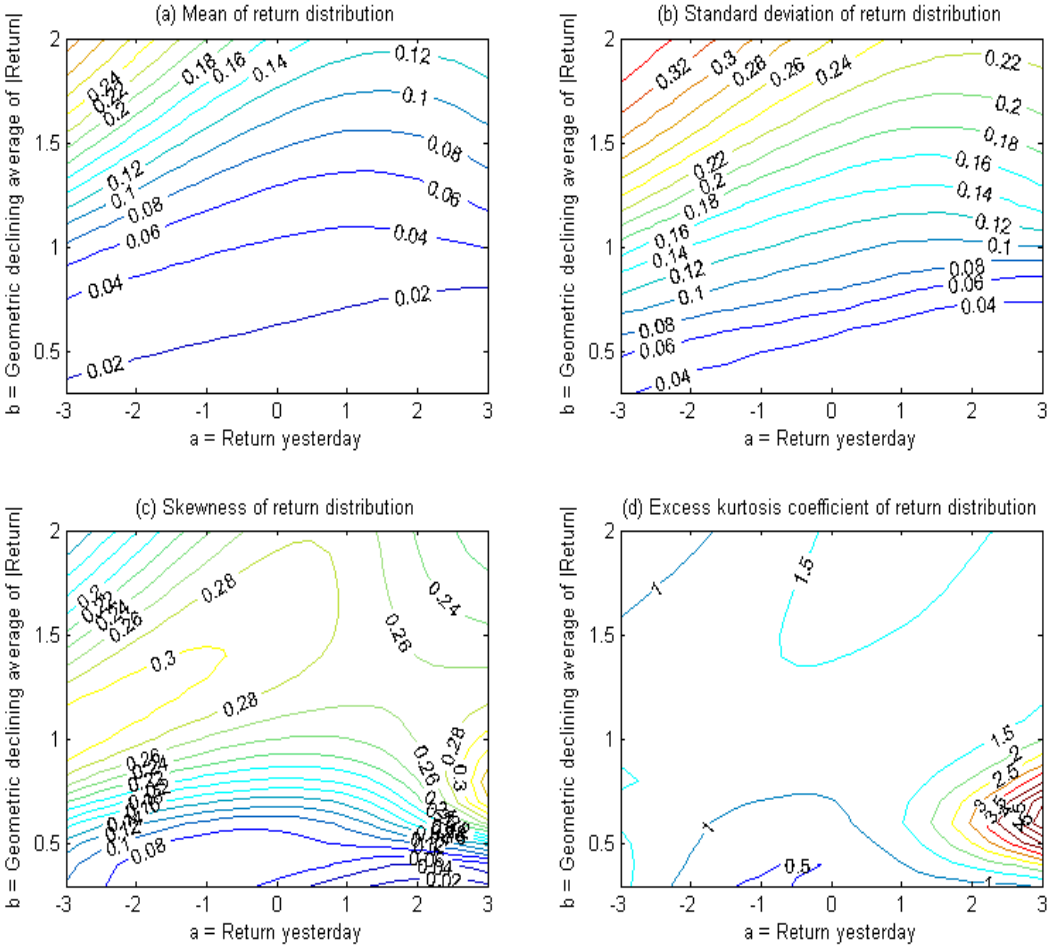

Figure 11: Posterior standard deviations of four population conditional moments, S&P 500 illustration

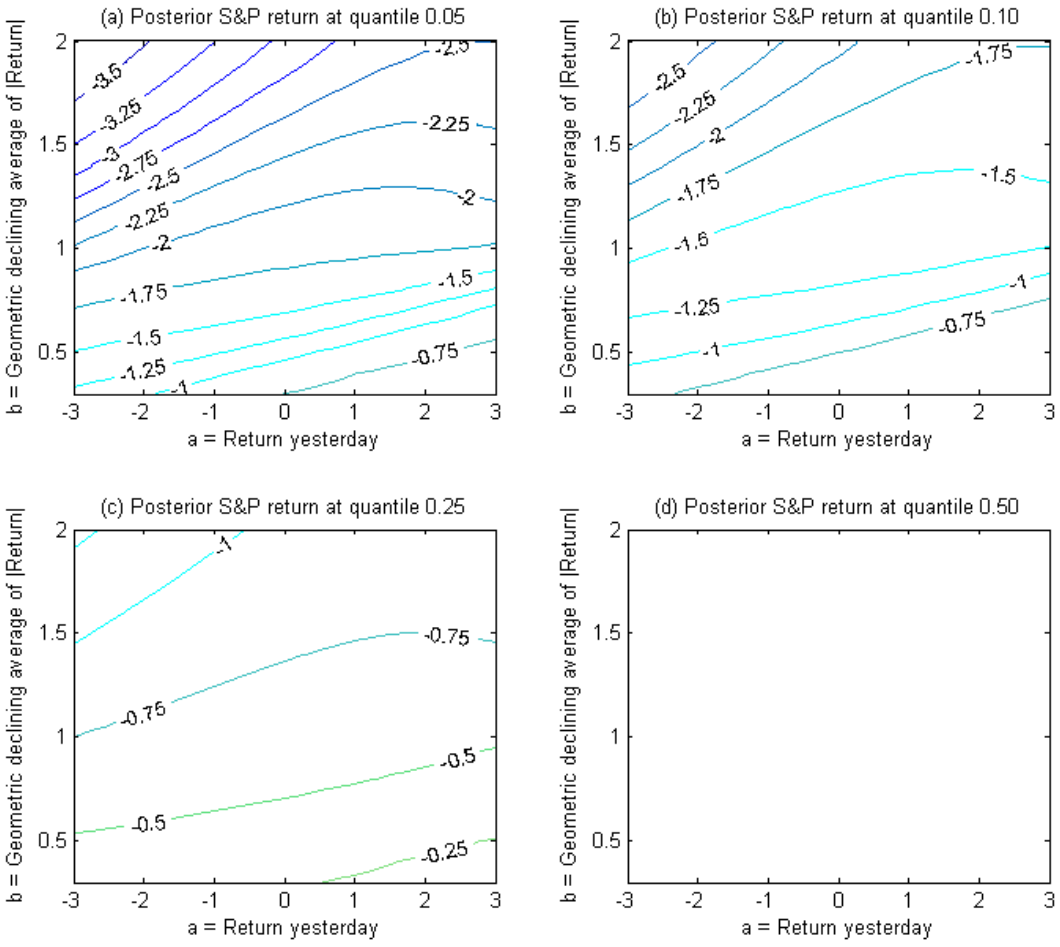

Figure 12: Quantiles of the posterior conditional distribution of returns, S&P 500 illustration

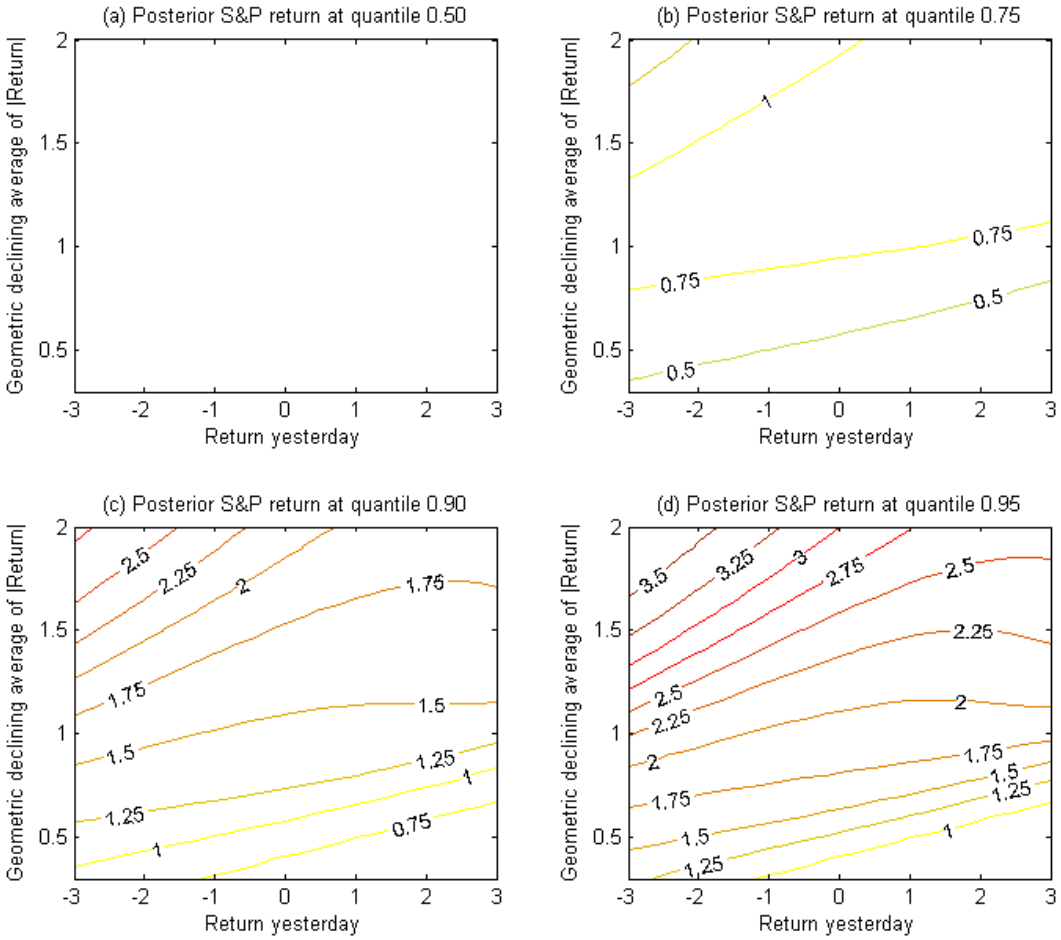

Figure 13: Quantiles of the posterior conditional distribution of returns, S&P 500 illustration

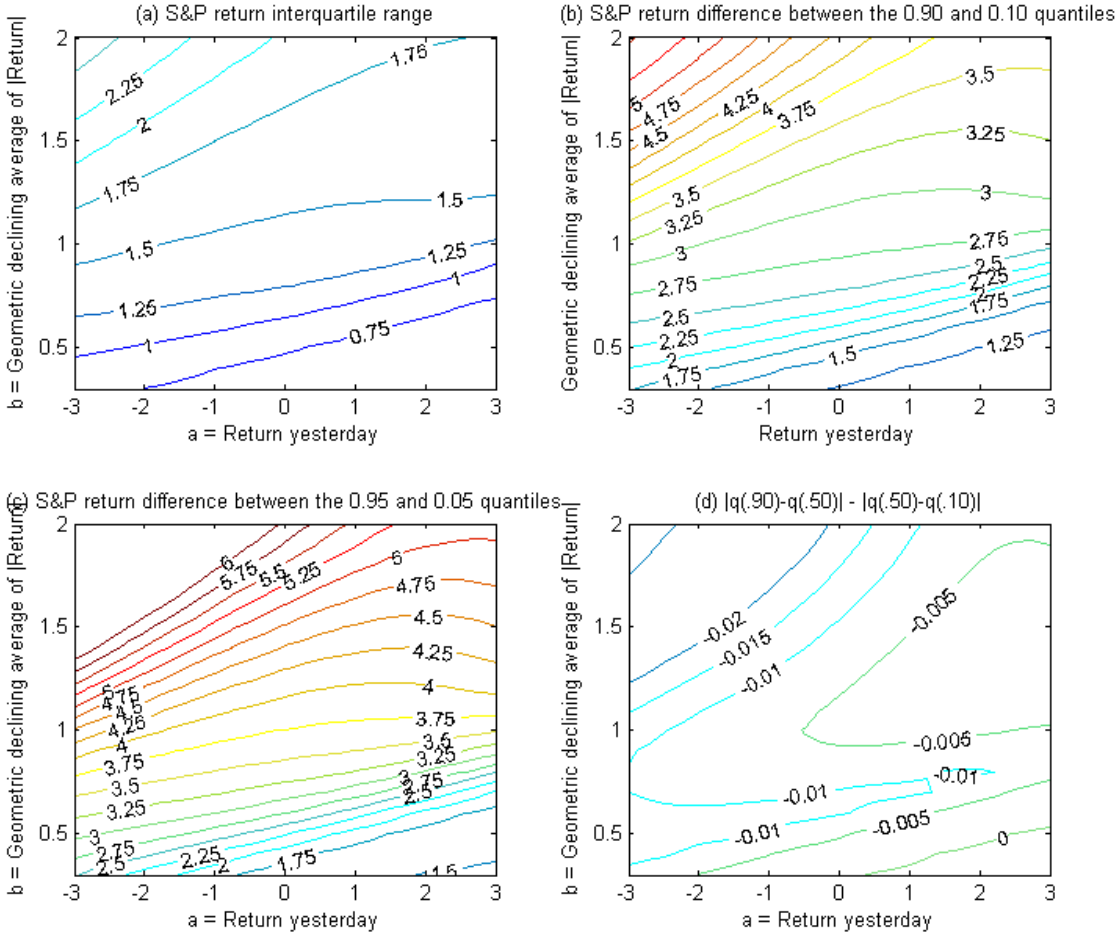

Figure 14: Aspects of the dispersion of the posterior conditional distribution of returns, S&P 500 illustration

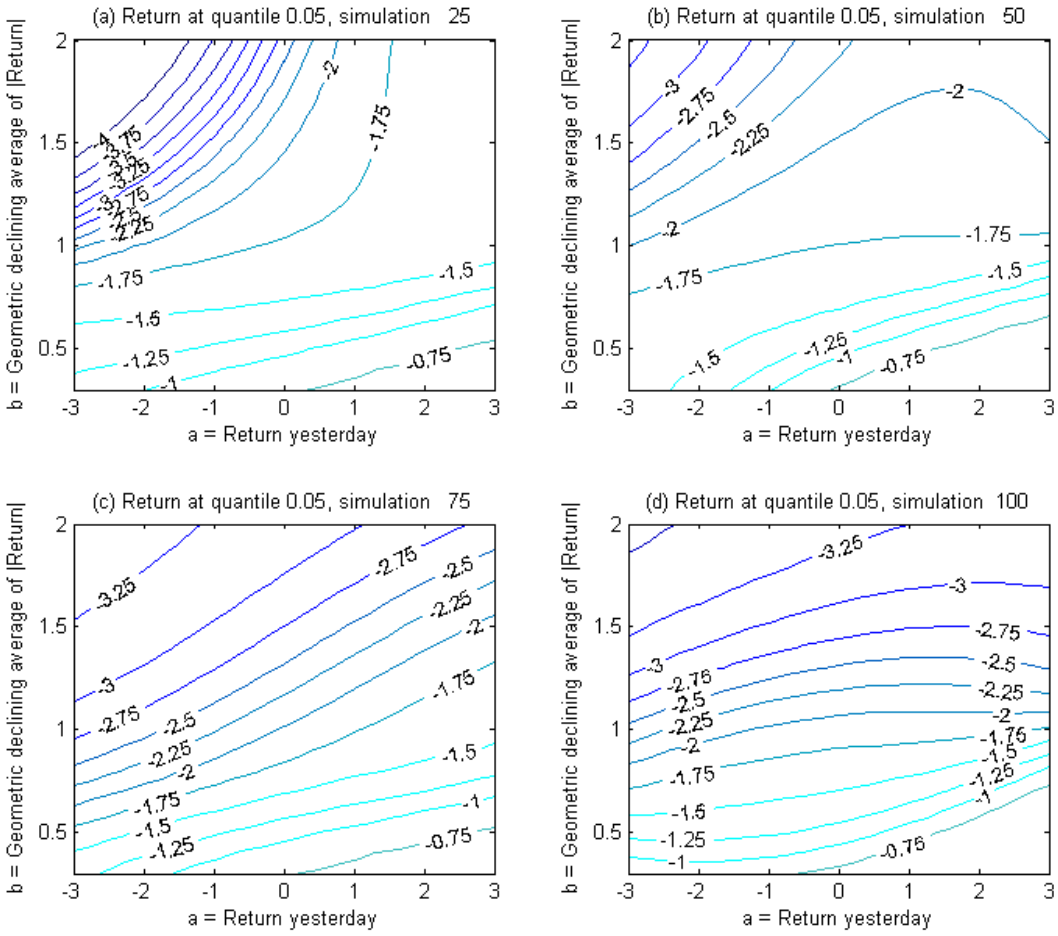

Figure 15: Four random draws from the posterior distribution of the population 5% quantile of the conditional distribution of returns, S&P 500 illustration