UNOBSERVED HETEROGENEITY IN INCOME DYNAMICS: AN EMPIRICAL BAYES PERSPECTIVE

JIAYING GU AND ROGER KOENKER

ABSTRACT. Empirical Bayes methods for Gaussian compound decision problems involving longitudinal data are considered. The new convex optimization formulation of the nonparametric (Kiefer-Wolfowitz) maximum likelihood estimator for mixture models is employed to construct nonparametric Bayes rules for compound decisions. The methods are first illustrated with some simulation examples and then with an application to models of income dynamics. Using PSID data we estimate a simple dynamic model of earnings that incorporates bivariate heterogeneity in intercept and variance of the innovation process. Profile likelihood is employed to estimate an AR(1) parameter controlling the persistence of the innovations. We find that persistence is relatively modest, $\hat{\rho} \approx 0.48$, when we permit heterogeneity in variances. Evidence of negative dependence between individual intercepts and variances is revealed by the nonparametric estimation of the mixing distribution, and has important consequences for forecasting future income trajectories.

1. INTRODUCTION

Unobserved heterogeneity has become a pervasive concern throughout applied econometrics. Longitudinal data presents special opportunities and challenges for models of unobserved heterogeneity; in virtually all econometric applications involving panel data there will be some form of latent, i.e. unobserved, individual specific effects. Classical econometric methods adopt either a differencing strategy designed to purge these effects, or some form of shrinkage method to mitigate their undesirable "incidental parameter" effect. In this paper we will describe some new nonparametric empirical Bayes methods for estimation and prediction in panel data models with unobserved heterogeneity.

Version: May 14, 2015. This research was partially supported by NSF grant SES-11-53548. All computational results of this paper can be reproduced with the aid of the R packages REBayes, Koenker (2015), and Rmosek, Friberg (2012), and code available from the authors on request. The Rmosek package provides a convenient R interface to the Mosek optimization language of Andersen (2010). We would like to express our appreciation to referees, Associate Editor, Editor and participants in the CeMMaP-UCL, Harvard-MIT, Georgetown, Rutgers, Cornell and University of Maryland econometrics workshops for valuable comments while retaining full responsibility for all errors of omission and commission.

As stressed in recent work of Efron (2010, 2011), empirical Bayes methods pioneered by Robbins (1951, 1956) provide a statistical framework for many contemporary "big data" applications. Although they predate the development of hierarchical Bayes methods exemplified in the work of Lindley and Smith (1972) and Chamberlain and Leamer (1976), they share many common features. The transition from parametric to nonparametric empirical Bayes methods brings exciting new opportunities that greatly expand the flexibility of existing approaches to panel data modeling and its treatment of unobserved heterogeneity. Although we focus here on conditionally Gaussian models, we would like to stress at the outset that our methods, in particular our computational strategy for the Keifer-Wolfowitz nonparametric MLE, is applicable in a wide variety of other circumstances:

- A binomial example is discussed in Koenker and Mizera (2014), and some connections to nonparametric Bayes methods with Dirichlet process priors are developed in Gu and Koenker (2013),
- Poisson mixtures are considered in Brown, Greenshtein, and Ritov (2013),
- Weibull and Gompertz mixtures of the type introduced in Heckman and Singer (1984) are considered in Koenker and Gu (2013),
- Binary response with a nonparametric link function is considered in Cosslett (1983),
- Applications to multiple testing are discussed in Gu and Koenker (2015).

Other applications to mixtures of Pareto's, lognormals, Student t's, etc can be easily developed. Most of the foregoing methods are already implemented in the REBayes R package, Koenker (2015).

We will begin with a brief overview of empirical Bayes methods beginning with Robbins (1951). In Section 3 we extend the predominant Gaussian location mixture framework to accommodate nonparametric location *and scale* mixtures with covariates in the classical Gaussian panel data setting, including some simulation evidence to illustrate the performance of the new methods. Section 4 describes an extended application to models of heterogeneous income dynamics that illustrates both estimation and prediction aspects of the new methods including, notably, the introduction of a bivariate joint distribution of unobserved heterogeneity and covariate effects via profile likelihood methods.

In sharp contrast to the classical Gaussian hierarchical Bayes framework for panel data, or its frequentist analogues, the nonparametric mixture formulation of our proposed methods offers a much more flexible view of unobserved heterogeneity while preserving most of the virtues of the likelihood formalism.

2. EMPIRICAL BAYES: A BRIEF OVERVIEW

Given a simple parametric model, there is a natural temptation to complicate it by admitting that those immutable natural constants that constitute the model's original parameters might instead be random. One of the earliest examples of this type is the classical Gaussian random effects, compound decision problem introduced by Robbins (1951). We observe independent $Y_1, \cdots, Y_n$ each Gaussian with known, common variance, θ but individual specific means, $Y_i \sim \mathcal{N}(\alpha_i, \theta)$. Our objective is to estimate all the α_i's subject to squared error loss,

$$\mathcal{L}_2(\hat{\alpha}, \alpha) = \|\hat{\alpha} - \alpha\|_2^2 = \sum_{i=1}^{n} (\hat{\alpha}_i - \alpha_i)^2.$$

The naive (unbiased) solution would simply set $\hat{\alpha}_i = Y_i$, but the usual presumption in such circumstances would be that the observations have some common genesis, and consequently we may be able to "borrow strength" from the full sample to improve upon these myopic predictions based on a single observation.

Suppose we believed that the α_i were drawn iid-ly from the distribution, F, so the Y_i's would have convolution density $g(y) = \int \phi((y - \alpha)/\sqrt{\theta})/\sqrt{\theta} dF(\alpha)$: What would the Reverend Bayes advise? Elementary exponential family theory yields the following proposition. Concise proofs of our propositions appear in the Appendix.

Proposition 1. *For* $Y_i \sim \mathcal{N}(\alpha_i, \theta)$ *and* $\{\alpha_i\}$ *iid* F, *the Bayes rule under* $\mathcal{L}_2$ *loss is:*

$$\delta(y) = y + \theta g'(y)/g(y) \tag{1}$$

and $\delta(y)$ *is non-decreasing in* y.

Efron (2011) refers to this expression for $\delta(y)$ as Tweedie's formula, citing Robbins's (1956) attribution of it to M.C.K. Tweedie. Tukey (1974) provides an earlier attribution to Arthur Eddington appearing in Dyson (1926). A major objective of the present paper is to explore the consequences of extending this result to longitudinal settings in which we can estimate heterogeneity of scale as well as location.

Of course one may well ask: Where did this F come from? And this question leads us inevitably toward estimation of the density, g, and hence to the empirical Bayes paradigm. When F comes from a finite dimensional parametric family there are several prominent special cases, including the family of Stein rule methods. See Gu and Koenker (2013) for further details on this linkage for parametric settings.

2.1. Non-parametric Estimation of the Gaussian Mixture Model.

When we lack confidence in a particular parametric specification of the mixing distribution, F, we are faced with a more serious quandary. It is apparent that we need a nonparametric estimate of F, and in our Gaussian location mixture setting this is tantamount to solving a deconvolution problem: Find F such that the density,

$$g(y) = \int \varphi(y - \alpha) dF(\alpha)$$

matches that of the observed Y_i's. Deconvolution is notoriously difficult as shown by Carroll and Hall (1988) and Fan (1991), but before we despair a second look at

Tweedie's formula (1) reveals that we may not really *need* an estimate of F. We need only estimate the mixture density g, a task that can be accomplished at standard univariate non-parametric convergence rates for smooth densities, and smoothness is ensured by the Gaussian convolution whatever F might be.

Kernel density estimation of g as proposed by Brown and Greenshtein (2009) seems to be the natural approach, but in addition to the familiar, but still unsettling, requirement of choosing a bandwidth, kernel estimators of g have the drawback that they do not enforce the monotonicity of the Bayes rule. The latter failing can be addressed by a further monotonization step, or by a penalization approach as suggested in Koenker and Mizera (2014). However, a more direct approach is possible via the Kiefer-Wolfowitz non-parametric maximum likelihood estimator (NPMLE) for the mixture model. This approach was first proposed by Jiang and Zhang (2009) for the Gaussian compound decision problem, suggesting the EM algorithm as a computational strategy.

Although Kiefer and Wolfowitz (1956) established consistency of their MLE for the mixing distribution F, it was not until the appearance of Laird (1978) that a viable computational strategy for the estimator was available. The EM algorithm has remained the standard approach for its computation ever since. Heckman and Singer (1984) constitutes an influential econometric application. However, EM has notoriously slow convergence in such applications, and this fact has seriously inhibited the use of the NPMLE in applications. It introduces what is, in effect, a new smoothing parameter into the computational strategy controlled by the stopping criterion of the algorithm. Koenker and Mizera (2014) have recently proposed an alternative computational method for the NPMLE that circumvents these problems. For a broad class of mixture problems, the Kiefer-Wolfowitz estimator can be formulated as a convex optimization problem and solved efficiently by modern interior point methods. Quicker, more accurate computation of the NPMLE opens the way to a much wider range of applications of the method for models of heterogeneity. Considerable further detailed comparison of the EM and interior point comparison can be found in Koenker and Mizera (2014). We would like to stress here that the convexity of the Kiefer-Wolfowitz problem is a generic property of the mixture setting given the discretization represented by the grid. Regularization is provided by the positivity constraints on the mass associated with each of the grid locations, and optimization typically yields solutions with relatively few points of support for the mixing distribution. Thus, both the location of and mass associated with support points of the mixing distribution are encompassed by the solution. Attempts to restrict, a priori, the number of mass points or impose further constraints may imperil the convexity and create further convergence difficulties as has been observed in prior EM implementations. In the absence of further constraints EM as applied to the mixture model as originally proposed by Laird (1978) is convergent to the same global optimum as the interior point solution if one is willing to wait long enough, but fortunately this is no longer necessary.

In the next section we will describe how these methods can be adapted to longitudinal data, first for location and scale mixtures separately, then for location-scale mixtures and finally for location-scale mixtures with covariate effects. In contrast to compound decision problems with cross sectional data, richer longitudinal data offers new opportunities permitting more complex structures of unobserved heterogeneity.

3. ESTIMATING GAUSSIAN MIXTURE MODELS WITH LONGITUDINAL DATA

Extending the Gaussian compound decision problem with one location parameter per observation to unbalanced longitudinal observations in which we have m_i observations on each individual is quite straightforward. We will describe this relatively simple setting first, and then gradually introduce heterogeneous variance effects, first with independent prior assumptions and then with a general form of bivariate heterogeneity. Estimation of covariate effects via profile likelihood is then introduced. The section concludes with some simulation evidence intended to illustrate our estimation and prediction methods, leading to an extended application of the methods to models of earning dynamics.

Suppose for convenience that we have unit variance for the noise so $u_{it} \sim \mathcal{N}(0,1)$, and we have,

$$y_{it} = \alpha_i + u_{it}, \quad t = 1, \cdots, m_i, \quad i = 1, \cdots, n.$$

Sufficiency can be used to reduce the problem to the sample: $\bar{y}_i = m_i^{-1} \sum_{t=1}^{m_i} y_{it} \sim \mathcal{N}(\alpha_i, m_i^{-1})$. When the α_i's are iid from F, we can write the log likelihood of the observed y_{it}'s as,

$$\ell(F|y) = K(y) + \sum_{i=1}^{n} \log(\sqrt{m_i} \int \phi(\sqrt{m_i}(\bar{y}_i - \alpha))dF(\alpha))$$

Optimizing over an infinite dimensional F necessitates some form of discrete approximation. As in earlier EM implementations, such as that of Jiang and Zhang (2009), we take F to have a piecewise constant (Lebesgue) density on a relatively fine grid containing the empirical support of the observed $\bar{y}_i$'s. Maximizing the likelihood $\ell(F|y)$ generally yields a small number of discrete mass points whose location is determined obviously only up to the scale of the grid. With a few hundred grid intervals we can obtain a quite accurate estimate. Further refinement is always possible as discussed in Koenker and Mizera (2014), but already with a uniform grid with 300 points we have very precise positioning of the mass points of the mixing distribution, more precise than the statistical accuracy of the mass locations would justify. Letting $f_j : j = 1, \cdots, p$ denote the function values of dF

on this grid, we can express the constrained maximum likelihood problem as,

$$\max_f \{\sum_{i=1}^{n} \log(g_i) \mid g = Af,\ \sum_{j=1}^{p} f_j \Delta_j = 1,\ f \geqslant 0\}, \tag{2}$$

where $A = (A_{ij} = \sqrt{m_i} \int \phi(\sqrt{m_i}(\bar{y}_i - \alpha_j)))$ and Δ_j is the jth grid spacing. As posed, the problem is evidently convex, and therefore has a unique solution. It is well-known, going back to Kiefer and Wolfowitz (1956) and Laird (1978), that variational solutions to the original problem are discrete with fewer than n atoms. It is somewhat difficult to appreciate this result by viewing EM solutions, since the number of EM iterations required to obtain an accurate solution would test the patience of even the most diligent researchers. But interior point methods make this discreteness easily apparent. Since the number of non-negligible $\hat{f}_j > 0$ obtained is typically much smaller than n, often only a handful of points, even in large samples, this also guides our judgement regarding the adequacy of the original grid. As documented in Koenker and Mizera (2014) solving a small problem of this type with $n = 200$ and $p = 300$ grid points requires about 1 second for the Mosek optimizer and about 10 minutes to achieve a somewhat less precise solution via EM. Ten minutes may not seem prohibitive, but embedding larger problems of this type in profile likelihood settings where many such solutions are required is another story. Dicker and Zhao (2014) have recently shown that grids with $p = \sqrt{n}$ yield convergence in Hellinger distance of the *mixture* density at rate $\mathcal{O}_p(\log n/\sqrt{n})$, the parametric rate modulo the log term. Unfortunately, little is known at this stage about the convergence properties of the *mixing* distribution beyond the consistency result of Kiefer and Wolfowitz.

The dual formulation of primal problem (2) has proven to be somewhat more efficient from a computational viewpoint. The dual can be expressed as

$$\max_\nu \{\sum_{i=1}^{n} \log(\nu_i) | A^\top \nu \leqslant n 1_p, \nu \geqslant 0\}. \tag{3}$$

This formulation reveals that we are only required to solve for the n-dimension vector ν, albeit subject to an infinite dimensional constraint that we have discretized to an p dimensional grid, see Koenker and Mizera (2014) for further details.

3.1. **Estimating Gaussian Scale Mixtures.**

Gaussian scale mixtures can be estimated in much the same way that we have described for location mixtures. Suppose we now observe,

$$y_{it} = \sqrt{\theta_i} u_{it}, \quad t = 1, \cdots, m_i, \quad i = 1, \cdots, n$$

with $u_{it} \sim \mathcal{N}(0, 1)$. Sufficiency again reduces the sample to n observations on $S_i = m_i^{-1} \sum_{t=1}^{m_i} y_{it}^2$, and thus S_i has the gamma distribution with shape parameter,

$r_i = m_i/2$, and scale parameter θ_i/r_i, i.e.

$$\gamma(S_i|r_i, \theta_i/r_i) = \frac{1}{\Gamma(r_i)(\theta_i/r_i)^{r_i}} S_i^{r_i-1} \exp\{-S_i r_i/\theta_i\},$$

and the marginal density of S_i when the θ_i are iid from F is

$$g(S_i) = \int \gamma(S_i|r_i, \theta/r_i) dF(\theta).$$

To estimate F we can proceed exactly as before except that now the matrix A has typical element $\gamma(S_i|\theta_j)$ for θ_j on a fine grid covering the support of the sample S_i's.

3.2. **Estimating Gaussian Location-Scale Mixtures.** When both location and scale are heterogeneous we must combine the strategies already described. We should stress that modeling heterogeneity of scale parameters would not be possible with cross sectional data since individuals are then only measured once. The model is now,

$$y_{it} = \alpha_i + \sqrt{\theta_i} u_{it}, \quad t = 1, \cdots, m_i, \quad i = 1, \cdots, n$$

with $u_{it} \sim \mathcal{N}(0, 1)$. We will provisionally assume that $\alpha_i \sim F_\alpha$ and $\theta_i \sim F_\theta$ are independent. Again, we have sufficient statistics:

$$\bar{y}_i|\alpha_i, \theta_i \sim \mathcal{N}(\alpha_i, \theta_i/m_i)$$

and

$$S_i|r_i, \theta_i \sim \gamma(S_i|r_i, \theta_i/r_i),$$

where $r_i = (m_i - 1)/2$, and the log likelihood becomes,

$$\ell(F_\alpha, F_\theta|y) = K(y) + \sum_{i=1}^{n} \log \int\int \gamma(S_i|r_i, \theta/r_i)\sqrt{m_i}\phi(\sqrt{m_i}(\bar{y}_i - \alpha_i)/\sqrt{\theta})/\sqrt{\theta} dF_\alpha(\alpha) dF_\theta(\theta)$$

Since the scale component of the log likelihood is additively separable from the location component, we can solve for $\hat{F}_\theta$ in a preliminary step, as in the previous subsection, and then solve for the $\hat{F}_\alpha$ distribution. In fact, under the independent prior assumption, we can re-express the Gaussian component of the likelihood as Student-t and thereby eliminate the dependence on θ in the Kiefer-Wolfowitz problem for estimating F_α. This is highly convenient for estimation purposes, however it should be stressed that prediction restores the interdependence on both F_α and F_σ as we discuss in more detail below.

When the independent prior assumption is implausible, and this may be typical of many econometric applications like our income dynamics application, we can construct two dimensional grids. This makes the constraint matrix, A, a bit more unwieldy, but raises no new issues of principle. If, as in our empirical application to income dynamics, we permit a general bivariate prior for (α, θ), the Bayes rule for estimating α under $\mathcal{L}_2$ loss takes a considerably more complex form summarized in the following result.

Proposition 2. *Suppose that* $y_{it}|\alpha_i, \theta_i \sim \mathcal{N}(\alpha_i, \theta_i)$ *and* (α_i, θ_i) *are iid from* $H(\alpha, \theta)$, *then the Bayes rule for* α *conditional on the sufficient statistics* $\bar{y}_i$ *and* S_i *is*

$$\mathbb{E}(\alpha|\bar{y}_i, S_i) = \int \mathbb{E}(\alpha|\bar{y}_i, \theta) f(\theta|\bar{y}_i, S_i) d\theta$$

where $\mathbb{E}(\alpha|\bar{y}_i, \theta)$ *is the Bayes rule of Proposition 1, for fixed* θ, *and* $f(\theta|\bar{y}_i, S_i)$ *denotes the posterior density of* θ *for individual* i *under the prior* H. *The Bayes rule is monotone in* $\bar{y}_i$ *in the limit as* $S_i \to 0$ *and* $S_i \to \infty$, *however for intermediate values of* S_i *such monotonicity is no longer assured.*

Monotonicity rests upon the contribution of $\frac{d}{d\bar{y}} f(\theta|\bar{y}_i, S_i)$, since for fixed θ the contribution from inner expectation is monotone by Proposition 1. In the S_i limits the posterior $f(\theta|\bar{y}_i, S_i)$ puts all its mass on the most extreme points of the prior and consequently also produces a monotone Bayes rule for α as a function of $\bar{y}_i$. However, for more moderate values of S_i the situation is more complicated, and as we shall see in the empirical section, non-monotonicities can occur.

A natural question at this point might be: How do we know whether we *need* to bother with all this? Can we make some preliminary test for parameter heterogeneity? There is an extensive literature on this topic: Chesher (1984) and Cox (1983) constitute notable contributions drawing connections to the White (1982) information matrix test. Many of the proposals that have been made can be formulated as $C(\alpha)$ tests as in Neyman (1959) and Neyman and Scott (1966). The $C(\alpha)$ formulation can be viewed as an extended form of the Rao score test that among other things can accommodate general forms of nuisance parameter estimation. Gu (2013) provides a detailed theoretical treatment of $C(\alpha)$ tests for parameter heterogeneity, and Gu, Koenker, and Volgushev (2013) compares their performance to that of likelihood ratio tests.

3.3. **Covariate Effects.** Having seen how to estimate the Gaussian location-scale mixture model we will now briefly describe how to introduce covariate effects into the model, which now takes the form,

$$y_{it} = x_{it}\beta + \alpha_i + \sqrt{\theta_i} u_{it}.$$

Given a β it is easy to see that,

$$\bar{y}_i|\alpha_i, \beta, \theta_i \sim \mathcal{N}(\alpha_i + \bar{x}_i\beta, \theta_i/m_i)$$

so the sufficient statistic for α_i is $\bar{y}_i - \bar{x}_i\beta$. Similarly, the sufficient statistic for θ_i can be defined as,

$$S_i = \frac{1}{m_i - 1} \sum_{t=1}^{m_i} (y_{it} - x_{it}\beta - (\bar{y}_i - \bar{x}_i\beta))^2$$

and $S_i|\beta, \theta_i \sim \gamma(r_i, \theta_i/r_i)$, where as before, $r_i = (m_i - 1)/2$. Apparently, using the familiar panel data terminology, the sufficient statistic for α_i contains the between information, while the within information, deviations from the individual means, is contained in the S_i. A note of caution should be added however since the orthogonality of the within and between information enjoyed by the classical Gaussian panel data model no longer holds in this general mixture setting. This can be seen more clearly by examining the likelihood function,

$$\begin{aligned} L(\beta, h) &= \prod_{i=1}^{n} g((\alpha, \beta, \theta)|y_{i1}, \ldots, y_{im_i}) \\ &= \prod_{i=1}^{n} \int \int \prod_{t=1}^{m_i} \theta^{-1/2} \phi((y_{it} - x_{it}\beta - \alpha)/\sqrt{\theta}) h(\alpha, \theta) d\alpha d\theta \\ &= K \prod_{i=1}^{n} S_i^{1-r_i} \int \int (\theta/m_i)^{-1/2} \phi((\bar{y}_i - \bar{x}_i\beta - \alpha)/\sqrt{\theta/m_i}) \frac{e^{-R_i} R_i^{r_i}}{S_i \Gamma(r_i)} h(\alpha, \theta) d\alpha d\theta \end{aligned}$$

where $R_i = r_i s_i/\theta$ and $K = \prod_{i=1}^{n} \left(\frac{\Gamma(r_i)}{\sqrt{m_i} r_i^{r_i}} (1/\sqrt{2\pi})^{m_i - 1} \right)$.

Even with the independent prior assumption, $h(\alpha, \theta) = h_\alpha(\alpha) h_\theta(\theta)$, the likelihood does not factor because the Gaussian piece depends on both α and θ. However, the fact that S_i, hence the Gamma piece of the likelihood, does not depend on α provides a convenient estimation strategy by using the Gamma mixture to estimate h_θ, and a Studentized version of the Gaussian mixture, $(\bar{y}_i - \bar{x}_i\beta - \alpha_i)/\sqrt{S_i} \sim t_{m_i - 1}$, for estimating h_α. Including covariates adapts this estimation strategy: Given a β we can estimate the two mixing distributions and then evaluate the full profile likelihood. We will illustrate this approach in the empirical section, albeit with a more general mixture model that drops the independent prior assumption, and allows for covariates including lagged response. Our approach is related to recent work by Bonhomme and Manresa (2014) on grouped patterns of heterogeneity in panel data, in the sense that both approaches reduce the dimensionality of the heterogeneity distribution substantially, although the estimation methods employed are quite different. Convexity of our likelihood formulation ensures a unique solution and avoids the introduction of further tuning parameters, while the clustering algorithms employed by Bonhomme and Manresa require more delicate attention.

3.4. **Empirical Bayes Prediction: Some Simulation Evidence.** To develop some intuition about empirical Bayes methods we will consider some simple illustrative simulation examples in this section before turning to our main empirical application.

3.4.1. *Gaussian Location Mixtures.* Suppose that we have a random sample from the model: $y_i = \alpha_i + u_i$ with iid $u_i \sim \mathcal{N}(0, 1)$, and iid $\alpha_i \sim \frac{2}{3}\delta_{-h} + \frac{1}{3}\delta_{2h}$ as in

Chen (1995). Here, δ_a denotes the distribution with point mass one at the point a. If we were successful in estimating the distribution of α_i, we would expect that Tweedie's formula (1) should deliver predictions that correctly shrink the original observations toward their respective α_i's. Of course the nature of the shrinkage depends crucially on the loss function as well as the prior. Thus, $\mathcal{L}_1$ loss yields decisions that are closely related to classification, while $\mathcal{L}_2$ loss delivers a Bayes rule whose shrinkage is somewhat more mild.

In Figure 1 we illustrate the foregoing situation with $n = 400$ and $h = 0.5$, which represents a fairly challenging problem since the marginal density is still unimodal. In the left panel we have the estimated mixing distribution in red, with the target distribution represented in blue. The larger of the two actual mass points at $x = -0.5$ is quite accurately estimated, however the smaller mass point at $x = 1$ is split into two pieces by the Kiefer-Wolfowitz estimate. The true mixture distribution in blue in the middle panel appears to be reasonably accurately estimated by the red curve. The corresponding Bayes rules as derived in Proposition 1 in the right panel show that the empirical Bayes rule (in red) shrinks a little too aggressively in the left tail, and not quite aggressively enough in the right tail, compared to the omniscient Bayes rule in blue. But we should hasten to add that it represents an enormous improvement over the unbiased (naive) decision rule, $\hat{\alpha}_i = y_i$, depicted in grey.

Replacing the two mass point distribution by $\alpha_i \sim U[-h, 2h]$ yields an even more challenging problem. The Kiefer-Wolfowitz estimator tries valiantly to mimic the uniform mixture by a discrete mixture as illustrated in Figure 2. The two point mixing distribution appearing in the left panel does not seem to be a very satisfactory surrogate for the uniform, but as can be seen in the middle panel, it does a remarkably good job of imitating the correct mixture density. The Bayes rule comparison in the right panel again illustrates that the shrinkage in the tails is not ideal, but much preferable to the naive, unbiased rule.

3.4.2. *Gamma Scale Mixtures.* To explore the performance of empirical Bayes methods for gamma scale mixtures we illustrate a couple of similar cases to those appearing in the previous subsection. We first consider the longitudinal model,

$$y_{it} = \alpha_i + \sqrt{\theta_i} u_{it}, \quad t = 1, \cdots, m, \quad i = 1, \cdots, n,$$

with iid $u_{it} \sim \mathcal{N}(0, 1)$, and $\theta_i \sim F$. We take $m = 11$ and $n = 400$. We will provisionally ignore the heterogeneity in the α_i, or to be more explicit, adopt the naive practice of estimating them by $\bar{y}_i$. Denoting the individual specific variance estimates by $x_i = (m-1)^{-1} \sum_t (y_{it} - \bar{y}_i)^2$, the $\{x_i\}$ are then distributed as Gamma with shape parameter, $r = (m-1)/2$, scale parameter θ_i/r, and density,

$$\gamma(x_i|\theta_i) = \frac{1}{\Gamma(r)(\theta_i/r)^r} x_i^{r-1} \exp(-x_i r/\theta_i).$$

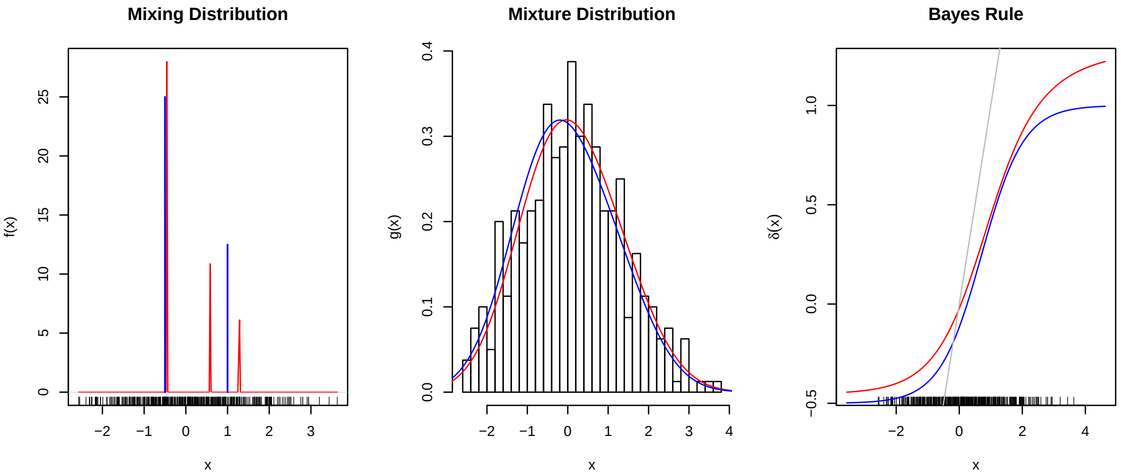

FIGURE 1. Empirical Bayes estimation for the Chen (1995) example: A sample of 400 observations from the model with $y_i = \alpha_i + u_i$ with iid $u_i \sim \mathcal{N}(0,1)$, and iid $\alpha_i \sim \frac{2}{3}\delta_{-h} + \frac{1}{3}\delta_{2h}$, is illustrated by the histogram in the middle panel and the "rug plots" in the adjacent panels. The Kiefer-Wolfowitz estimate of the mixing distribution is illustrated in the left panel in red, with the target distribution in blue. The corresponding estimate of the mixture density and Bayes rule appear in the other panels contrasted to their blue target functions. The unbiased, naive decision rule is depicted in the right panel in grey.

Thus, the marginal density of the sample variances is,

$$g(x) = \int \gamma(x|\theta)\,dF(\theta).$$

The Bayes rule under squared error loss for θ_i given x_i, originally derived by Robbins (1982), is given in the following proposition. Again, it should be stressed that the Bayes rule depends only on the mixture density, g, and not directly on the mixing distribution, F. Of course, indirectly the Bayes rule *does* depend on F and in particular the flat portions of the Bayes rule in the third panel of Figure 3 representing the points of attraction of Bayes shrinkage are essentially determined by the location of the estimated mass points of F. This is particularly crucial in the tails where even small mass points of $\hat{F}$ can exert a large influence on the shrinkage. Whether this sensitivity can be lessened by replacing the Gaussian mixture assumption by something heavier tailed constitutes an intriguing question for future research. The next proposition describes the Bayes rule for estimating the θ_i's under $\mathcal{L}_2$ loss for Gamma mixtures. Note that θ is not the natural parameter of

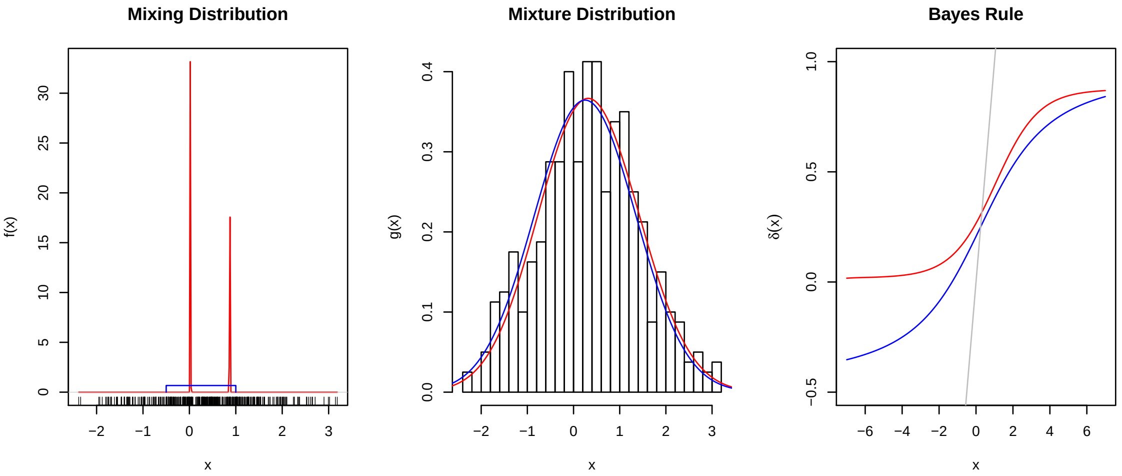
Mixing Distribution
Mixture Distribution
Bayes Rule

FIGURE 2. Empirical Bayes estimation for the Chen (1995) example: A sample of 400 observations from the model with $y_i = \alpha_i + u_i$ with iid $u_i \sim \mathcal{N}(0, 1)$, iid $\alpha_i \sim U[-h, 2h]$, and $h = 0.5$, is illustrated by the histogram in the middle panel and the "rug plots" in the adjacent panels. The Kiefer-Wolfowitz estimate of the mixing distribution is illustrated in the left panel in red, with the target distribution in blue. The corresponding estimate of the mixture density and Bayes rule appear in the other panels contrasted to their blue target functions. The unbiased, naive decision rule is depicted in the right panel in grey.

the exponential family in this case, so the monotonicity of the Bayes rule requires a brief additional argument in the Appendix.

Proposition 3. *For* $X_i \sim \Gamma(r, \theta_i/r)$ *and* $\{\theta_i\}$ *iid* F, *the Bayes rule under* $\mathcal{L}_2$ *loss is:*

$$\delta(x) = rx^{r-1}\int_x^\infty y^{1-r} g(y)dy/g(x) \tag{4}$$

and $\delta(x)$ *is non-decreasing in* x.

In Figure 3 we illustrate a typical outcome in a format like that of the previous figures. In this example we take F to be the two point distribution: $\frac{2}{3}\delta_{1.5} + \frac{1}{3}\delta_3$. The two point mixing distribution is quite well estimated by the Kiefer-Wolfowitz procedure, and the mixture density appears to be quite accurate as well. The empirical Bayes rule slightly over estimates the variances in the upper tail since it slightly overestimates the location of the upper mass point. But as for the previous

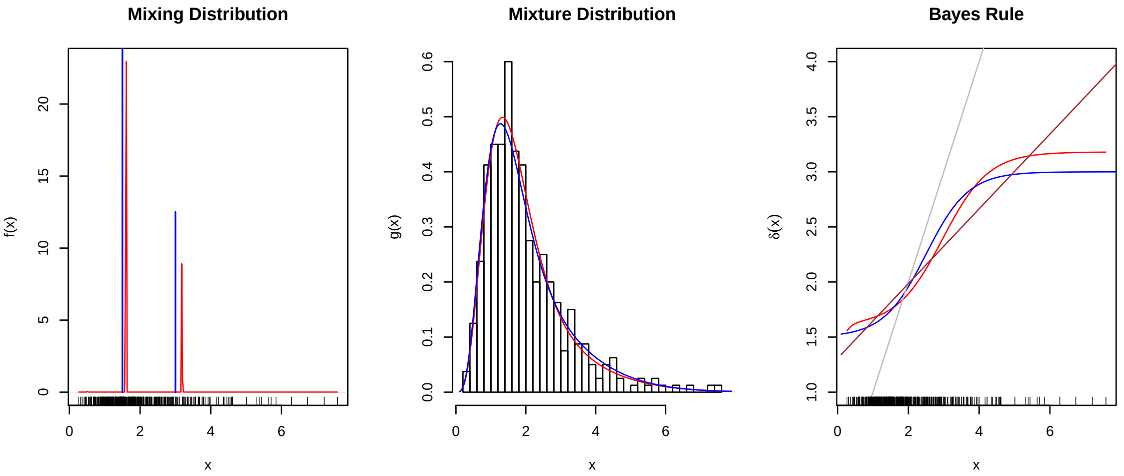

FIGURE 3. Empirical Bayes estimation for Gamma mixture example: A sample of $n = 400$ and $m = 11$ observations from the model $y_{it} = \sqrt{\theta_i} u_{it}$ with iid $u_{it} \sim \mathcal{N}(0, 1)$ and iid $\theta_i \sim \frac{2}{3}\delta_{1.5} + \frac{1}{3}\delta_3$ is illustrated by the histogram in the middle panel and the "rug plots" in the adjacent panels. The Kiefer-Wolfowitz estimate of the mixing distribution is illustrated in the left panel in red, with the target distribution in blue. The corresponding estimate of the mixture density and Bayes rule appear in the other panels contrasted to their blue target functions. The unbiased, naive decision rule is depicted in the right panel in grey. The brown line represents the linearized empirical Bayes rule proposed in Robbins (1982)

.

examples, there is an enormous improvement over the naive decision rule represented by the grey line. The brown line represents the linearized empirical Bayes rule proposed in Robbins (1982).

3.4.3. *Gaussian Location Scale Mixtures.* We now would like to consider joint estimation of location and scale mixtures in the context of our longitudinal model. We will maintain the assumption that α_i's and θ_i's are drawn independently, so we only have to estimate two univariate mixing densities rather than a general bivariate density. We illustrate the procedure with an example that combines a three point distribution for α and a three point distribution for θ: $y_{it} = \alpha_i + \sqrt{\theta_i} u_{it}$ with iid $u_{it} \sim \mathcal{N}(0, 1)$, iid $\alpha_i \sim \frac{1}{3}\delta_{-0.5} + \frac{1}{3}\delta_1 + \frac{1}{3}\delta_3$, and iid $\theta_i \sim \frac{1}{3}\delta_{0.5} + \frac{1}{3}\delta_2 + \frac{1}{3}\delta_4$, Maximizing the likelihood of Section 3.2, we obtain the estimates appearing in the first two panels of Figure 4 for the location and scale parameters respectively. As in the previous figures, estimates appear in red, and the true mixing distribution is represented by the blue lines. Focusing on the location parameter, the third panel of the figure depicts the histogram of the observed $\bar{y}_i$ with the estimated marginal density, by integrating out α and θ with respect to $\hat{F}_\alpha$ and $\hat{F}_\theta$, superimposed in red, and the true marginal density superimposed in blue. Finally, in the last panel of the figure we illustrate the empirical and idealized Bayes rule for the α_i's. This version of the Bayes rule presumes that prediction is based on knowledge of a location estimate, but nothing about the scale parameter beyond the distribution represented by the estimated mixing distribution. Even though the mixing distribution of the location parameter has a few extraneous mass points, the Bayes rule is remarkably accurate.

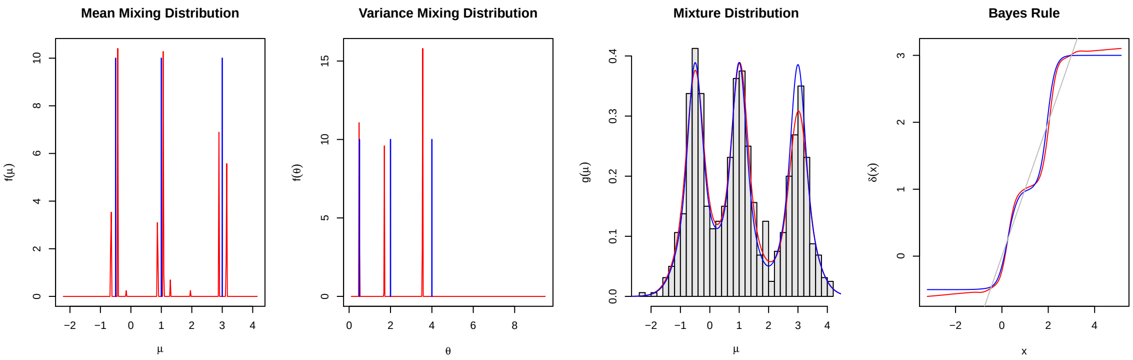

FIGURE 4. Empirical Bayes estimation for Gaussian location-scale mixture: A sample of $n = 800$ and $m = 11$ observations from the model $y_{it} = \alpha_i + \sqrt{\theta_i}u_{it}$ with iid $u_{it} \sim \mathcal{N}(0,1)$, iid $\alpha_i \sim \frac{1}{3}\delta_{-0.5} + \frac{1}{3}\delta_1 + \frac{1}{3}\delta_3$, and iid $\theta_i \sim \frac{1}{3}\delta_{0.5} + \frac{1}{3}\delta_2 + \frac{1}{3}\delta_4$, is illustrated by the histogram in the middle panel The Kiefer-Wolfowitz estimate of the mixing distributions is illustrated in the two left panels in red, with the target distribution in blue. The corresponding estimate of the mixture density and the Bayes rule appear in the other panels contrasted to their blue target functions.

How much can be gained by using an individual specific estimate of variance? The Bayes rules appearing in the last panel of Figure 4 are conditional only on the observed $\bar{y}$ with the variance effect integrated out. Thus, when we see a value of $\bar{y}$ near one of the mass points in $\{-0.5, 1, 3\}$, the Bayes rule shrinks aggressively toward the corresponding α. Between these values, the predicted α, being a conditional mean, takes intermediate values. Extreme values of $\bar{y}$ in either tail again get aggressively shrunk toward the extreme points of the estimated prior. The situation we have just described is artificial in the sense that we effectively are assuming that we have observed $\bar{y}_i$ for each cross sectional unit, but apparently have forgotten to compute the associated variance estimate. If we now rectify this oversight, we can consider a two dimensional Bayes rule that maps $(\bar{y}_i, S_i)$ pairs into predictions of the α_i's. Using the same data and the estimates underlying Figure 4 we illustrate a contour plot of this two dimensional Bayes rule (Proposition 2) in Figure 5. We see that for central values of $\bar{y}_i$ the contours are essentially vertical indicating the variance is uninformative about the mean, however for outlying values of $\bar{y}_i$ the nonlinearity of the Bayes rule is apparent with large observed variances making us more uncertain about the α_i's.

When $\bar{y}$ is in the extremes, the Bayes rule should shrink its estimate of α to the extreme mass points at -0.5 and 3, but since the estimated prior has smaller mass points nearby very extreme observations are attracted to these values. In both tails one can see the effect of the variance estimate on this shrinkage effect; when the estimated variance is small then there is more shrinkage to the nearest mass point of the α distribution. When the observed variance is large, then the posterior for α is more evenly divided among several mass points and consequently the posterior mean is more central. For example, when $\bar{y} = 1.5$ and the estimated variance is low, then we can be quite confident that the observation comes from the $\alpha = 1$ population. Similarly when $\bar{y} = -1.5$ and the estimated variance is low, we can be confident that this is a $\alpha = -0.5$ observation. However, in either of these cases as the variance increases our confidence ebbs, and the Bayes rule assigns more probability to the other nearby mass points. For central values of $\bar{y}$ the contours are nearly vertical indicating that the observed variance is not informative in this region. The observed pairs $(\bar{y}_i, S_i)$ are superimposed on the contours to give some sense of their dispersion.

This form of the Bayes rule clearly illustrates that variances are informative about the means in such circumstances, but the fact that we've imposed independence between α and θ may sacrifice valuable information in many applications. If we allow for dependence and estimate their joint distribution as in our empirical application, we will see that the sample variances provide crucial information for estimating α_i.

4. HETEROGENEOUS INCOME DYNAMICS

The vast literature on longitudinal models of income dynamics can be conveniently decomposed into two strands: one focusing on a permanent-transitory time-series structure that eschews individual specific sources of heterogeneity, exemplified by MaCurdy (1982), and going back at least to Friedman (1957), and another that relies on heterogeneity to account for observed persistence, as for example in Lillard and Weiss (1979), Baker (1997), Haider (2001), Guvenen (2009), Browning, Ejrnæs, and Alvarez (2010) and Hospido (2012). Considerable flexibility can be introduced into the former approach with the aid of age specific deterministic trends in mean and variances, as for example in Blundell, Graber, and Mogstad

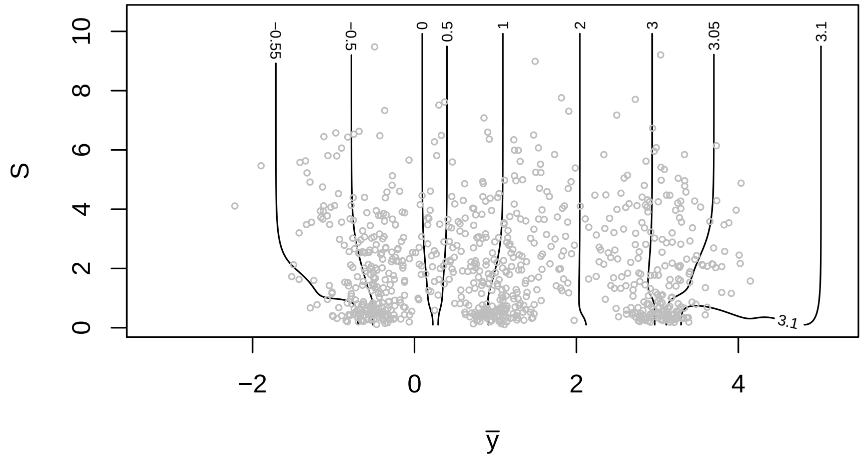

Figure 5. Bayes Rule for Gaussian location-scale mixture: Based on the observations from Figure 4 we illustrate the two dimensional Bayes rule for the mean parameter α as a contour plot.

(2014), or stochastic specifications of the variance process, as in Meghir and Pistaferri (2004). While most of the foregoing work relies on first and second order moment information and therefore, at least implicitly adopts a Gaussian framework, there is evidence that such assumptions may distort important features of the earnings process. Mixture models of individual heterogeneity introduce further flexibility: Horowitz and Markatou (1996) and Bonhomme and Robin (2010) explore semiparametric deconvolution, while Geweke and Keane (2000) and Hirano (2002) propose Bayesian MCMC methods for estimating semiparametric mixture models. Our nonparametric empirical Bayes approach maintains the mixture model formulation, but expands the nature of the heterogeneity to encompass both location and scale effects. In terms of estimation methods our approach is closest to that of Hirano since the NPMLE can be viewed as a limiting form of his Dirichlet process prior for the scale mixture setting. See Gu and Koenker (2013) for further details on this relationship, illustrated with an application to Gaussian location mixtures.

Our empirical analysis is based on the PSID sample used in Meghir and Pistaferri (2004), Browning, Ejrnæs, and Alvarez (2010) and Hospido (2012). The initial data consists of log real earnings of 2069 individuals between the ages of 25 and 55, with at least 9 consecutive records between 1968 and 1993. We further reduce the sample to 938 individuals who have continuous records from age 25 onwards.

We consider the model,

$$\begin{aligned} y_{it} &= \alpha_i + \beta_i x_{it} + \nu_{it} \\ \nu_{it} &= \rho \nu_{it-1} + \sqrt{\theta_i} \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(0, \sigma_\epsilon^2) \end{aligned}$$

Following standard practice in the literature, y_{it} denotes residuals from distinct annual regressions of log real earnings on a quadratic in age, and indicators for race, educational attainment, region and marital status. Heterogeneity around the mean earnings profile is captured by the random intercept and slope parameters; experience, x_{it}, is defined as age minus $\max\{\text{years of schooling}, 12\} - 6$. Heterogeneity in the variance of earnings is captured by the θ_i's. We do not model the initial observations, y_{i0}, and consequently the likelihood is conditional on the initial observation. One could introduce independent Gaussian y_{i0}'s. Assuming that each y_{i0} is drawn from the stationary distribution $\mathcal{N}(\alpha_i, \theta_i/(1 - \rho^2)$ is a convenient option and yields an efficient estimator for ρ provided that the assumption holds. One could also consider less restrictive specifications for y_{i0} at the cost of introducing additional parameters as in Arellano (2003). However it seems more propitious to consider Chamberlainian dependence on covariates, while trying to maintain a nonparametric perspective, a topic we defer to future research. More complex short run dynamics could be introduced via state space representations and Kalman filter formulations of the likelihood, but our strategy is to proceed parsimoniously trying to understand at each stage the consequences of expanding the flexibility of the model.

4.1. **Homogeneous Trend and Variance.** Under the restrictions that $\beta_i \equiv 0$ and $\theta_i \equiv 1$ we can rewrite the model as,

$$y_{it} = \rho y_{it-1} + (1 - \rho)\alpha_i + \epsilon_{it}.$$

This is a textbook dynamic panel model; in such models of earnings dynamics estimates of ρ are typically very close to one. These findings have led to considerable controversy over whether individual earnings processes "have a unit root." In contrast to Meghir and Pistaferri (2004), who postulate a permanent component of earnings with a unit-root, Browning, Ejrnæs, and Alvarez (2010) – using the same data – find no unit root after introducing further heterogeneity in covariance structure of the model. In Figure 6, we present some preliminary evidence that helps to explain why the persistence of innovations may be reduced by introducing heterogeneity, for example, by relaxing the restrictions of a homogeneous variance.

The QQ plots of Figure 6 confirm earlier evidence of Horowitz and Markatou (1996) and Guvenen, Karahan, Ozkan, and Song (2014) based on more extensive CPS and Social Security data respectively that earnings innovations are considerably heavier tailed than our usual Gaussian assumptions would imply. There are a variety of possible treatments for this disease: one option would be to abandon the Gaussian assumption entirely, but this would lead us into realm of choosing a non-Gaussian likelihood model that would, inevitably, be rather arbitrary. It is well

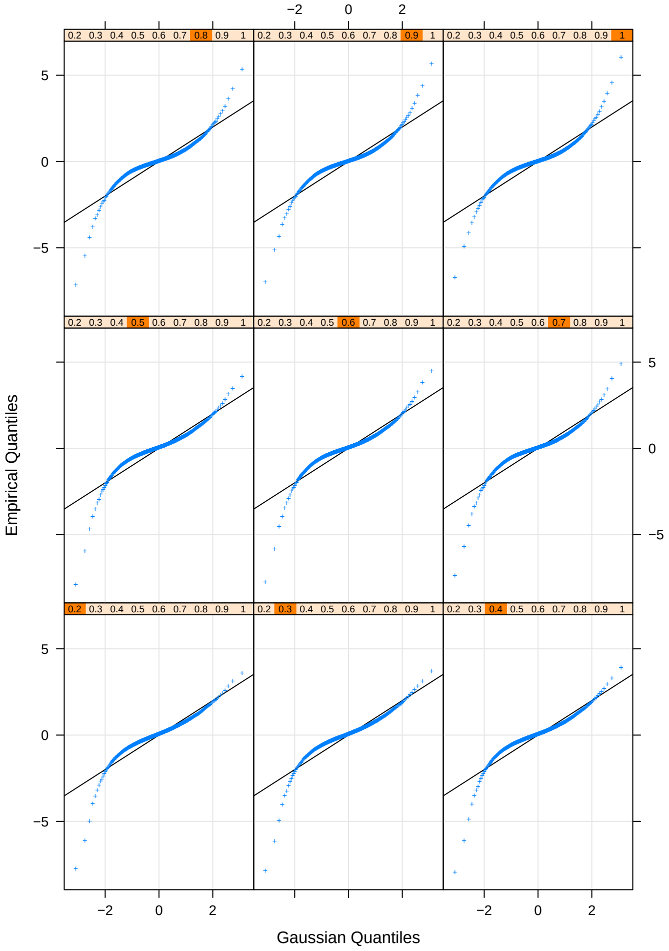

FIGURE 6. Normal QQ Plots of Partial Differenced Earnings for Various ρ: For $\rho \in \{0.2, 0.3, \cdots 1.0\}$ we plot empirical quantiles of the partial differences $y_{it} - \rho y_{it-1}$ standardized by their empirical standard deviation, against the corresponding Gaussian quantiles. The ρ's are indicated in the thin strip at the top of each panel, the solid line in each plot is the 45 degree line indicating conformity to the Gaussian hypothesis. It is apparent from the plot that the observed quantiles are far too leptokurtic, that is much too peaked near the median and exhibiting much heavier tails than the Gaussian. For small ρ there is also some left skewness in innovations that becomes less apparent for larger ρ.

known that heavy tailed distributions can be very flexibly modeled as scale mixtures of Gaussians, see for example the extensive discussion in Andrews, Bickel, Hampel, Huber, Rogers, and Tukey (1974), and we have already seen that it is relatively straightforward to estimate these mixture models; so this is approach we will adopt.

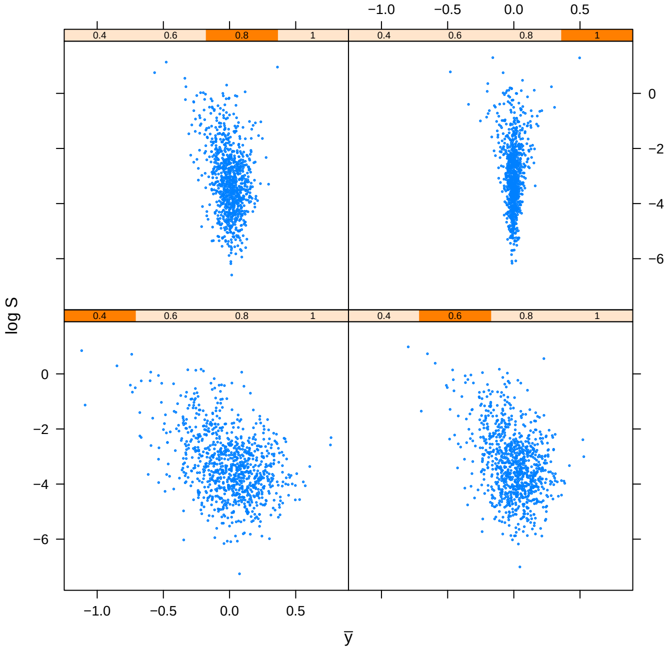

FIGURE 7. Scatterplot of Individual Specific Mean and Log Variance Effects for Various ρ: For $\rho \in \{0.4, 0.6, 0.8, 1.0\}$ we plot sample means, $\bar{y}$, and log variances, S, of the partial differences $y_{it} - \rho y_{it-1}$. The ρ's are indicated in the thin strip at the top of each panel. The more elliptical shape of the scatter for smaller ρ may suggest that it could be more parsimoniously fit by our Gaussian/Gamma location-scale mixture model.

To provide a further visual impression of the degree of individual heterogeneity we present in Figure 7 scatter plots of the individual specific sample means, $\bar{y}$, and log variances, S, for the partial differenced y_{it} data for several ρ's. In addition to confirming that there is substantial heterogeneity in these quantities the Figure also reveals that more moderate values of ρ yield a more elliptical scatter that seems to be favored by our Gaussian/Gamma location-scale mixture likelihood as we will see in the next subsection.

4.2. **Homogeneous Trend with Heterogeneous Variances.** If we fix ρ and σ_ϵ^2, and let $z_{it} = y_{it} - \rho y_{it-1}$, we can rewrite our model as,

$$z_{it} = (1-\rho)\alpha_i + \sqrt{\theta_i}\epsilon_{it}.$$

As in Section 3, under Gaussian conditions, sufficient statistics for α_i and θ_i are respectively the sample mean and sample variance:

$$\begin{aligned} \bar{y}_i &= \tfrac{1}{T_i} \textstyle\sum_{t=1}^{T_i} z_{it} \\ S_i &= \tfrac{1}{T_i - 1} \textstyle\sum_{t=1}^{T_i} (z_{it} - \bar{y}_i)/\sigma_\epsilon)^2. \end{aligned}$$

Furthermore, we have, $\bar{y}_i \mid \alpha_i, \theta_i \sim \mathcal{N}((1-\rho)\alpha_i, \theta_i\sigma_\epsilon^2/T_i)$ and $(T_i - 1)S_i/\theta_i \mid \theta_i \sim \chi^2_{T_i-1}$. Assuming the pairs (α_i, θ_i) are iid with distribution function H, we can discretize H on a two dimensional grid and write the likelihood of observing $(z_{i1}, \ldots, z_{iT_i})$ as a function of H, ρ and σ_ϵ^2, and apply the NPMLE.

Various special case of this model has been considered in the literature, for example the random effects model of Alvarez and Arellano (2003) assumes θ_i to be degenerate taking value 1 while $\alpha_i \sim \mathcal{N}(\psi y_{i0}, \sigma_h^2)$. This leads to a marginal density for the $\bar{y}_i$ conditional on y_{i0} as $\bar{y}_i \sim \mathcal{N}(\psi y_{i0}, \sigma_\alpha^2)$ with $\sigma_\alpha^2 = \sigma_\epsilon^2/T_i + \sigma_h^2$ as a free parameter. The parameters $(\rho, \psi, \sigma_h^2, \sigma_\alpha^2)$ can then be estimated by maximizing the likelihood conditional on y_{i0}. The Gaussian assumption on the α_i is very convenient and very commonly employed, notably in Chamberlain (1980), Chamberlain and Hirano (1999) among many others. However, the normality assumption on the α_i may be hard to justify. As we have seen in Figure 7, there is also considerable heterogeneity in the θ_i, and it seems plausible that there may be some dependence between α and θ. These considerations motivate us to consider a non-parametric maximum likelihood framework allowing us to estimate the non-parametric mixing distribution $H(\alpha, \theta)$ conditional on some structural parameters like ρ, that can, in turn, be estimated by maximizing a profile likelihood.

Without loss of generality, we can set $\sigma_\epsilon^2 = 1$, since it is not identified once we allow individual specific θ_i unless we make further moment restrictions on θ_i. We have the following NPMLE problem:

$$\hat{H}_\rho := \underset{H \in \mathcal{H}}{\operatorname{argmax}} \prod_{i=1}^{n} \int\int f(\bar{y}_i \mid \alpha, \theta) g(S_i \mid \theta) dH(\alpha, \theta)$$

where $\mathcal{H}$ is the space of all two dimensional distribution functions on the domain of $\mathbb{R} \times \mathbb{R}_+$. Here, f is the conditional normal density of $\bar{y}_i$ and g is the conditional gamma density for S_i. The NPMLE for H is indexed by ρ because both $\bar{y}_i$ and S_i involve ρ, which we have suppressed in the notation, but can be estimated by maximizing the profile log likelihood,

$$l(\rho, \hat{H}_\rho) = \sum_{i=1}^{n} K(\bar{y}_i, S_i) + \log \int\int f(\bar{y}_i \mid \alpha, \theta) g(S_i \mid \theta) d\hat{H}_\rho(\alpha, \theta).$$

Allowing heterogeneous individual variances in earnings innovations is not new. Geweke and Keane (2000) contend that variance heterogeneity is crucial to account for non-Gaussian features of innovation distribution and use a three-component mixture formulation. Hirano (2002) adopts a more flexible Dirichlet prior specification for similar reasons. Browning, Ejrnæs, and Alvarez (2010) also find significant evidence that the variance of innovations varies across individuals. Their model posits eight latent factors all of which are constrained to obey parametric marginals. They comment "Nowhere in the literature is there any indication of how to specify a general joint distribution for these parameters, nor is there any

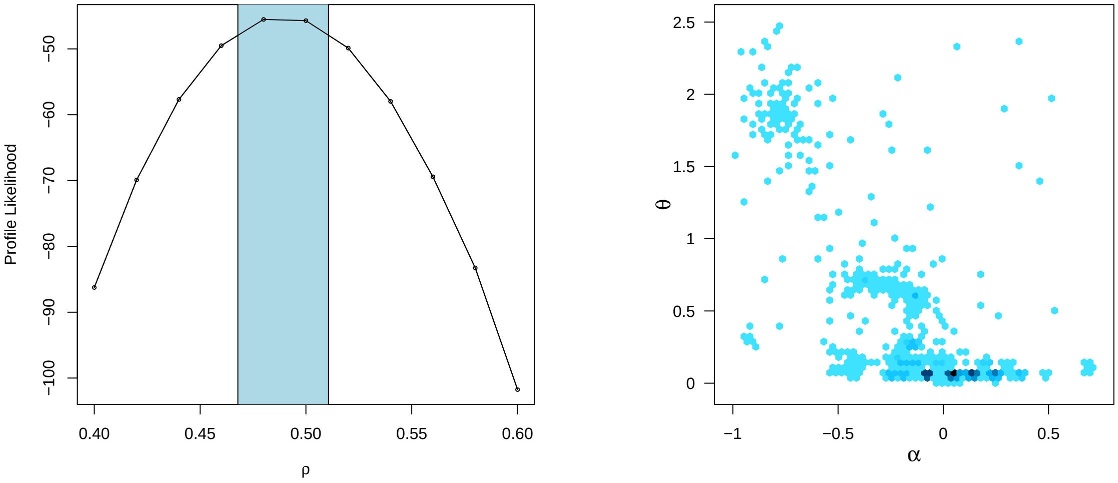

FIGURE 8. Profile Likelihood for the ρ Parameter and Heterogeneity Distribution $H(\alpha, \theta)$: In the left panel we plot the Kiefer-Wolfowitz profile likelihood as a function of ρ. The shaded region represents a 0.95 confidence interval for ρ based on the usual Wilks inversion procedure. In the right panel we plot the estimated joint heterogeneity distribution, evaluated at the optimal $\hat{\rho}$, $\hat{H}_{\hat{\rho}}(\alpha, \theta)$. Darker hexagons indicate greater mass, lighter ones less mass and white regions contain no mass.

hope of identifying the joint distribution non-parametrically." In contrast, Our approach allows only two latent factors, but has the advantage that it permits non-parametric estimation of their joint distributions.

What if $\hat{\rho} \approx 1$? Our joint distribution for (α_i, θ_i) would then be meaningless, since the α_i's would be annihilated. The left panel of Figure 8 plots the profile likelihood for ρ, which (fortunately) peaks at 0.48. The shaded region indicates a 0.95 confidence interval for ρ as determined by the classical Wilks inversion procedure, see e.g. Murphy and Van der Vaart (2000), and Fan, Zhang, and Zhang (2001). Our estimate of ρ is close to the estimate of Hospido (2012) who also allows a individual specific variance component in a ARCH effect variance. She adopts a fixed effect specification for (α_i, θ_i) and uses a bias corrected estimator for ρ to account for the asymptotic bias introduced by estimating all the incidental parameters $(\alpha_i, \theta_i), i = 1, \cdots, n$. A plausible explanation for why estimates of ρ tend to be close to one in models without heterogeneity in variances is that individual specific persistence is mistaken for AR persistence in innovations.

The right panel of Figure 8 plots the two-dimensional non-parametric estimate of $\hat{H}_{\hat{\rho}}(\alpha, \theta)$ on a 60×60 grid. Mass points of the estimated distribution are indicated by shaded hexagons with darker shading indicating more mass. The support of $\hat{H}$ is determined by the support of the observed $(\bar{y}_i, S_i)$. The mixing distribution shows some negative dependence between α and θ, especially for $\alpha < 0$. So a low draw for α is more likely to be accompanied by a more risky (higher) θ. Most of the mass of $\hat{H}$ is concentrated at very low levels of θ, but it is not at all obvious how one might represent this estimated heterogeneity by a conventional parametric model.

4.3. **Heterogeneous Trends and Variances.** Reintroducing trend heterogeneity to our model of earnings dynamics gives us,

$$y_{it} - \rho y_{it-1} = (1-\rho)\alpha_i + \beta_i \rho + (1-\rho)\beta_i x_{it} + \sqrt{\theta_i}\epsilon_{it},$$

and obviously brings a new layer of complexity to the estimation problem. Our framework is capable of incorporating this third dimension of heterogeneity and we have made some tentative estimation efforts for the full model. However, this is challenging not only due to the jump from 2d to 3d grids, but because the trend term invalidates our sufficient statistic dimension reduction device. Some preliminary testing for trend heterogeneity using the LM test recently proposed in Juhl and Lugovskyy (2014) produced very weak evidence against homogeneity. We have also considered a variety of other, more elaborate, modeling strategies for the variance effect including ARCH effects, and deterministic trends in the variance. These can be estimated by adding new parameters to the profile likelihood problem, but again we saw no compelling evidence that they were needed. However, further study, particularly with larger datasets like that of Guvenen, Karahan, Ozkan, and Song (2014) may reveal something different within our framework.

4.4. **Prediction.** We now return to our original objective: we would like to adapt the well-known univariate empirical Bayes rules described earlier to compound decision problems for longitudinal data models. This objective is closely aligned with the objectives of Chamberlain and Hirano (1999), although our computational methods, and perhaps our philosophical outlook, are quite distinct. Given an initial trajectory for an individual's earnings we would like to predict the remainder of the trajectory based not only on the prior history for the given individual, but also on the observed experience of a large sample of similar individuals. Chamberlain and Hirano motivate this prediction problem as one facing a typical financial advisor; similar problems present themselves in biomedical settings where diagnosis is based on reference growth charts.

Given a trajectory $\mathcal{Y}_0 = \{y_t : t = 1, \cdots, T_0\}$ for a hypothetical individual we can easily determine a posterior, $p(\alpha, \theta | \mathcal{Y}_0)$, based on our estimated mixture model. This NPMLE posterior is necessarily discrete, but one may feel entitled to draw uniformly from the grid rectangles of the estimated model for simulation purposes. In any case, the following simulation strategy can be employed to construct an ensemble of completed trajectories:

(1) Draw (α, θ) from $p(\alpha, \theta | \mathcal{Y}_0)$,
(2) Simulate $\mathcal{Y}_1 = \{y_t : t = T_0 + 1, \cdots, T\}$ as,

$$y_{T_0+s} = \alpha + \hat{\rho} y_{T_0+s-1} + \sqrt{\theta} u_s, \; s = 1, \cdots, T - T_0, \text{ and } u_s \sim \mathcal{N}(0,1),$$

to obtain m paths, $\mathcal{Y}_1$, then
(3) Repeat steps 1 and 2 M times.

This procedure yields mM trajectories from which it is easy to construct pointwise and/or uniform prediction bands.

From a formal Bayesian perspective the foregoing procedure is rather heretical. We began with a perfectly legitimate likelihood formulation: data was assumed to be generated from a very conventional Gaussian model, but individuals had idiosyncratic (α, θ) parameters whose distribution, H, could be viewed as a prior. If this H were delivered on a silver platter by some local oracle we could proceed just as we have described. Bayes rule would allow us to update H in the light of the observed initial trajectory, $\mathcal{Y}_0$ for each individual, and we would use these updated, individual specific, $\tilde{H}_i$'s to construct an ensemble of forecast paths. Various functionals of these forecast paths could then be presented. Lacking a local oracle, we have relied instead on the NPMLE and the largess of the PSID to produce an $\hat{H}$. Not only H, but also ρ and potentially other model parameters are estimated by maximum likelihood. Remarkably, no further regularization is required, and profile likelihood delivers an asymptotically efficient estimator of these "homogeneous" parameters. Admittedly, we have "sinned" – we've peeked when we shouldn't have peeked, but our peeking has revealed a much more plausible H than we could have otherwise expected to produce by pure introspection. This is the charm of the empirical Bayes approach.

Our prediction exercise takes $T_0 = 9$ so the first nine years of observed earnings have been used as $\mathcal{Y}_0$ to construct individual specific $\tilde{H}_i$ that are then used to construct pointwise confidence bands for earnings in subsequent years. We have selected a few pairs of individuals to illustrate the variety of earnings predictions generated by our model. In Figure 9 we contrast predictions for an individual with relatively large mean, high α, and large variance, high θ, with an individual with large variance, but lower mean. The "fan plot" depicts pointwise quantile prediction bands from 0.05 to 0.95 based on the simulated trajectories described above. For the high mean individual, the bands are relatively narrow reflecting the fact that the "prior" assigns little mass to high θ individuals. In contrast, for the lower mean individual the bands are much wider, indeed the upper portion of the band overlaps with the lower portion of the band for the higher α individual. Nevertheless, we see that the lower 0.05 quantile of the prediction band is exceeded. Our uniform band (not shown) for this individual just barely covers this excursion.

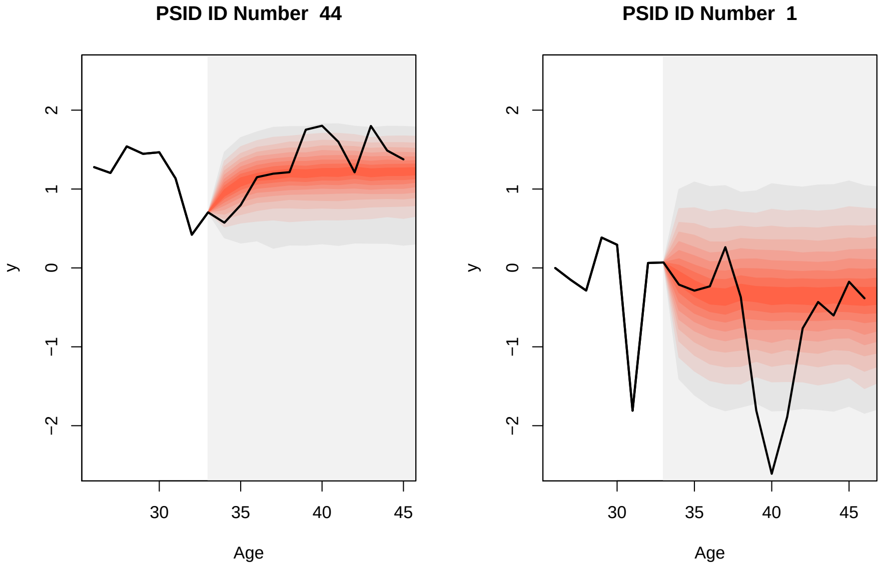

FIGURE 9. Fan Plot of Earnings Forecasts for Two Individuals: Based on the initial 9 years earnings, pointwise prediction bands are shown with graduated shading indicating bands from the 0.05 to 0.95 quantiles.

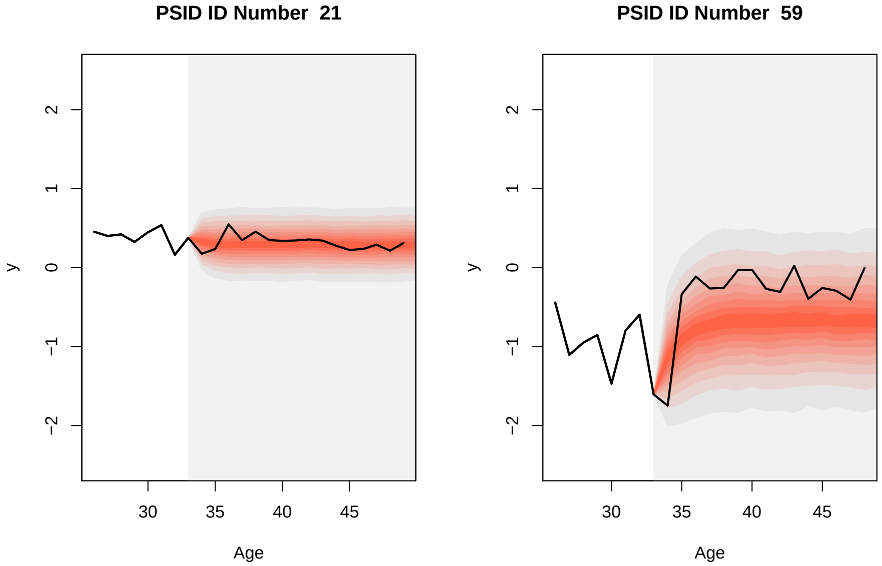

FIGURE 10. Fan Plot of Earnings Forecasts for Two Individuals: Based on the initial 9 years earnings, pointwise prediction bands are shown with graduated shading indicating bands from the 0.05 to 0.95 quantiles.

In Figure 10 we contrast high mean, low variance individual with low mean, high variance one. The prediction band is very narrow for the former individual, and much wider for the latter. Other features are also apparent from these figures: individuals who begin the forecast period below their pre-forecast mean, like PSID 59, are predicted to come back to their mean, and some asymmetry is visible, for example in PSID 44, whose lower tail is somewhat wider than the upper one. Note that asymmetry requires some asymmetry in the location component of the mixture distribution $\hat{H}$, since pure scale mixtures of Gaussians are necessarily symmetric.

4.5. **Estimation of Random Effects.** To conclude our discussion of earning dynamics we will briefly consider the problem of estimating random effects. Such problems have a long history; in econometrics they can be traced back to the seminal work of Goldberger (1962) on best linear unbiased prediction (BLUP). For a comprehensive survey of the early literature, see Robinson (1991). It may seem odd to consider estimation of random effects, but in many applications including our earning dynamics setting it is natural to ask: How would we estimate α_i's? The BLUP approach has a long history in animal breeding where α_i's are interpreted as a latent productivity variable. Our approach is considerably more flexible than earlier methods that assumed conjugate parametric priors for the mixing distributions.

In this section we illustrate the Bayes rule for estimating α_i's given the observed pair $(\bar{y}_i, S_i)$ for a given individual, and interpret the resulting shrinkage rules. Because of the general bivariate structure of estimated prior these shrinkage strategies can be considerably more complicated than those illustrated in the independent prior setting of the previous section. Figure 11 plots contours of the Bayes rule, $\hat{\alpha}_i = \mathbb{E}(\alpha|\bar{y}_i, S_i)$ in Proposition 2. This figure is analogous to Figure 5 except that the nature of the shrinkage for moderate S_i is more severe. If we first focus on the right side of the plot for positive $\bar{y}_i$'s we see that observations with moderate variances are shrunken quite substantially toward zero. So, for example, if we saw an observation with $\bar{y} = 0.5$ and $S = 0.25$ the Bayes rule estimates $\alpha = 0$. Why? The first thing to say is that we never saw points like this, the observed $(\bar{y}, S)$ pairs are depicted as the grey dots, so an S as big as 0.25 is much more likely to come from a low α individual and this accentuates the shrinkage. We should stress that the empirical distribution of the points appearing in the plot although they are a key ingredient in the construction of the estimated prior $\hat{H}$ illustrated in Figure 7, is only a starting point for building the Bayes rule underlying the contour plot. The Bayes rule requires updating individuals posterior for (α, θ) in the light of $\hat{H}$ and the observed $(\bar{y}, S)$ and then computing expectations as in Proposition 2. On the left side of the plot, for $\bar{y} < 0$ the situation is somewhat similar, but the shrinkage is less severe.

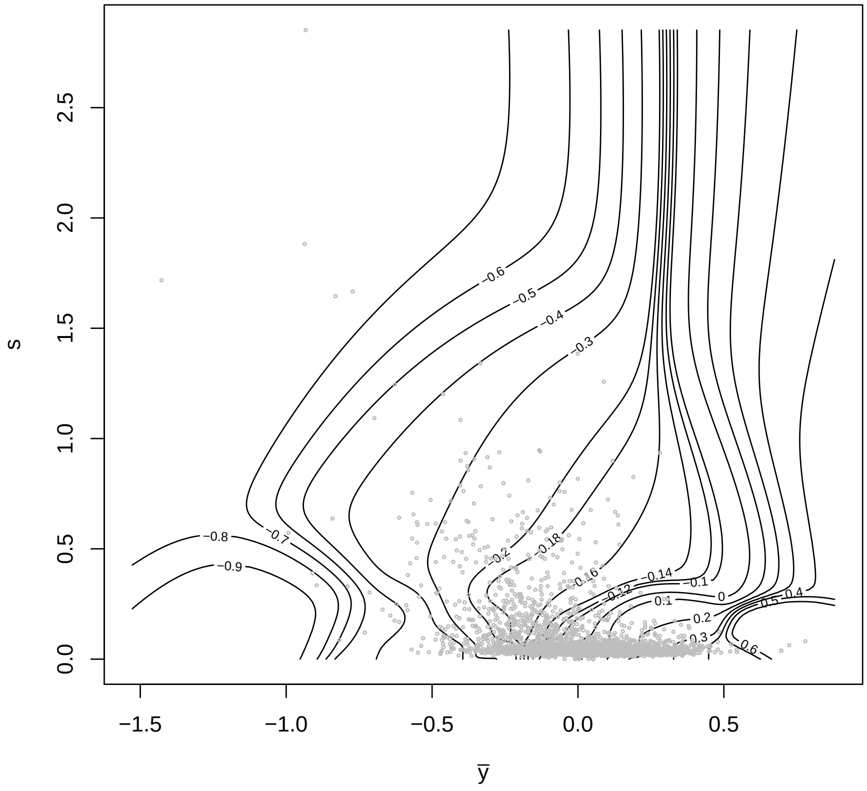

FIGURE 11. Contour Plot of the Bayes Rule $\mathbb{E}(\alpha|\bar{y}, S)$. The plot illustrates pairs $(\bar{y}_i, S_i)$ that produce the same posterior mean of α.

In Figure 12 we illustrate the Bayes rule for α as a function of $\bar{y}$ for several fixed values of S, essential plotting our contour values for horizontal cross-sections. The naive estimator, $\hat{\alpha} = \bar{y}$ is shown as the 45 degree line. For both low and high values of S we have monotone Bayes rules, so larger $\bar{y}$ implies larger $\hat{\alpha}$, however for the intermediate $S = 0.272$ value we see that the Bayes rule is clearly non-monotone. Similar calculations could be employed to estimate the variability parameter, θ, as a function of "observed" $(\bar{y}, S)$. (Recall that $(\bar{y}, S)$ implicitly depends upon an estimated ρ parameter.) Of course, there is nothing sacred about $\mathcal{L}_2$ loss, and it is entirely reasonable to consider other loss functions that would lead to alternative Bayes rules: posterior medians, posterior modes, etc.

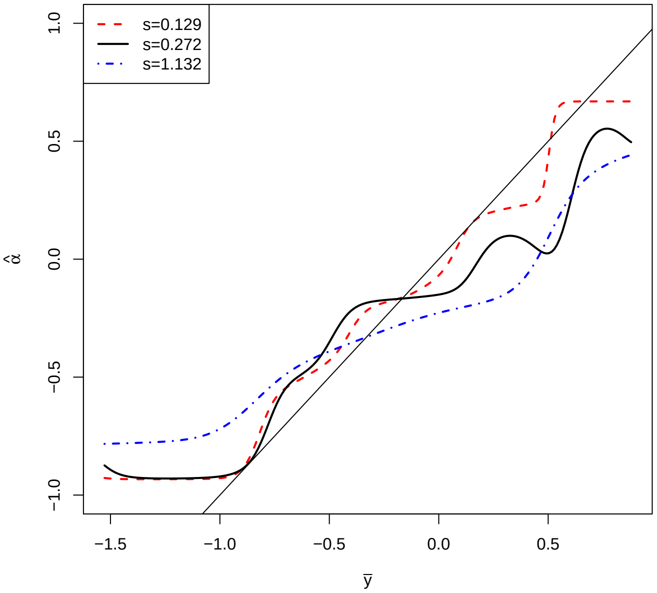

FIGURE 12. Bayes Rule $\hat{\alpha} = \mathbb{E}(\alpha|\bar{y}, S)$ for several (fixed) S. The plot depicts the posterior mean of α as a function of $\bar{y}$ for several values of S.

5. CONCLUSION

Models of unobserved heterogeneity for longitudinal data are common in applied econometrics. We have argued that empirical Bayes methods based on nonparametric maximum likelihood estimation of mixture models offer a natural formulation of these models. Recent developments in convex optimization greatly facilitate estimation of such models. Semiparametric versions of these models including covariate effects are shown to be effectively analyzed with profile likelihood. A potential criticism of the foregoing approach is that it requires us to assume a parametric form for the base distribution, in our setting the Gaussian. Of course, location-scale mixtures of Gaussians is quite a general class, so from a prediction perspective the normality assumption seems not to be terribly onerous.

Empirical Bayes applications have generally either assumed a parametric form for parameter heterogeneity as in the hierarchical Bayes literature or considered univariate parametric heterogeneity as in the more recent compound decision literature. We are not aware of any prior nonparametric bivariate heterogeneity specifications. Many econometric applications, however, involve mean-variance tradeoffs that naturally suggest more flexible bivariate specifications. As we have seen, modern optimization methods linked to the Kiefer-Wolfowitz MLE accommodate such models quite easily. Because the formulation is cast directly in terms of likelihood there are convenient methods of handling estimation and inference for other (global) parametric components via profiling. We would also like to stress that there is nothing crucial about the Gaussian framework that we have employed; other specifications of the base measure for the mixture can be easily accommodated. In addition to the normal-gamma mixtures explored here, we have also considered Weibull, Gompertz, Pareto, Binomial and Poisson mixtures in other work.

There are many possible extensions left to explore. More flexible treatment of the covariates in the initial stage of our procedure would be desirable; in larger datasets this could be easily handled with further stratification of the sample. More flexible treatment of the variance effects would also be desirable, either with deterministic age effects or some form of stochastic ARCH-type effects. Trend heterogeneity is also feasible, but perhaps only with larger scale data sources. We have tried to encourage further exploration of these methods by providing the R package *REBayes* that implements the methods we have described here as well as a variety of other model specifications.

REFERENCES

ALVAREZ, J., AND M. ARELLANO (2003): "The time series and cross-section asymptotics of dynamic panel data estimators," *Econometrica*, 71, 1121–1159.

ANDERSEN, E. D. (2010): "The MOSEK Optimization Tools Manual, Version 6.0," Available from `http://www.mosek.com`.

ANDREWS, D. F., P. J. BICKEL, F. R. HAMPEL, P. J. HUBER, W. H. ROGERS, AND J. W. TUKEY (1974): *Robust Estimates of Location: Survey and Advances*. Princeton.

ARELLANO, M. (2003): *Panel Data Econometrics*. Oxford U. Press.

BAKER, M. (1997): "Growth-rate heterogeneity and the covariance structure of life cycle earnings," *Journal of Labor Economics*, 15, 338–375.

BLUNDELL, R., M. GRABER, AND M. MOGSTAD (2014): "Labor Income Dynamics and the Insurance from Taxes, Transfers, and the Family," *Journal of Public Economics*, forthcoming.

BONHOMME, S., AND E. MANRESA (2014): "Grouped Patterns of Heterogeneity in Panel Data," preprint.

BONHOMME, S., AND J. ROBIN (2010): "Generalized Nonparametric Deconvolution with an application to Earnings Dynamics," *Review of Economic Studies*, 77, 491–533.

BROWN, L., AND E. GREENSHTEIN (2009): "Non parametric empirical Bayes and compound decision approaches to estimation of a high dimensional vector of normal means," *The Annals of Statistics*, 37, 1685–1704.

BROWN, L., E. GREENSHTEIN, AND Y. RITOV (2013): "The Poisson Compound Decision Problem Revisited," *J. Am. Stat. Assoc.*, 108, 741–749.

BROWNING, M., M. EJRNÆS, AND J. ALVAREZ (2010): "Modelling Income Proceses with Lots of Heterogeneity," *Review of Economic Studies*, 77, 1353–1381.

CARROLL, R., AND P. HALL (1988): "Optimal rates of convergence for deconvolving a density," *Journal of the American Statistical Association*, 83, 1184–1186.

CHAMBERLAIN, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 225–238.

CHAMBERLAIN, G., AND K. HIRANO (1999): "Predictive Distribution Based on Longitudinal Earnings Data," *Annales d' Économie et de Statistique*, 55/56, 211–242.

CHAMBERLAIN, G., AND E. E. LEAMER (1976): "Matrix weighted averages and posterior bounds," *Journal of the Royal Statistical Society. Series B*, 38, 73–84.

CHEN, J. (1995): "Optimal rate of convergence for finite mixture models," *The Annals of Statistics*, 23, 221–233.

CHESHER, A. (1984): "Testing for neglected heterogeneity," *Econometrica*, 54, 865–872.

COSSLETT, S. R. (1983): "Distribution-free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica*, 51, 765–782.

COX, D. (1983): "Some remarks on overdispersion," *Biometrika*, 70, 269–274.

DICKER, L., AND S. D. ZHAO (2014): "Nonparametric Empirical Bayes and Maximum Likelihood Estimation for High-Dimensional Data Analysis," preprint.

DYSON, F. (1926): "A Method for Correcting Series of Parallax Observations," *Monthly Notices of the Royal Astronomical Society*, 86, 686–706.

EFRON, B. (2010): *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge U. Press: Cambridge.

——— (2011): "Tweedie's Formula and Selection Bias," *Journal of the American Statistical Association*, 106, 1602–1614.

FAN, J. (1991): "On the optimal rates of convergence for nonparametric deconvolution problems," *The Annals of Statistics*, 19, 1257–1272.

FAN, J., C. ZHANG, AND J. ZHANG (2001): "Generalized likelihood ratio statistics and Wilks phenomenon," *The Annals of Statistics*, 29, 153–193.

FRIBERG, H. A. (2012): "Users Guide to the R-to-MOSEK Interface," Available from `http://rmosek.r-forge.r-project.org`.

FRIEDMAN, M. (1957): *A Theory of the Consumption Function*. Princeton U. Press.

GEWEKE, J., AND M. KEANE (2000): "An empirical analysis of earnings dynamics among men in the PSID: 1968 - 1989," *Journal of Econometrics*, 96, 293–356.

GOLDBERGER, A. S. (1962): "Best Linear Unbiased Prediction in the Generalized Linear Model," *Journal of the American Statistical Society*, 57, 369–375.

GU, J. (2013): "Neyman's $C(\alpha)$ Test for Unobserved Heterogeneity," `http://arxiv.org/abs/1302.0262`.

GU, J., AND R. KOENKER (2013): "Unobserved Heterogeneity in Longitudinal Data: An Empirical Bayes Perspective," preprint.

GU, J., AND R. KOENKER (2015): "On a Problem of Robbins," *International Statistical Review*, forthcoming.

GU, J., R. KOENKER, AND S. VOLGUSHEV (2013): "Testing for Homogeneity in Mixture Models," `http://arxiv.org/pdf/1302.1805`.

GUVENEN, F. (2009): "An Empirical Investigation of Labor Income Processes," *Review of Economic Dynamics*, 12, 58–79.

GUVENEN, F., F. KARAHAN, S. OZKAN, AND J. SONG (2014): "What Do Data on Millions of U.S. Workers Say About Life Cycle Earnings Risk," preprint.

HAIDER, S. (2001): "Earnings instability and earnings inequality of males in the United States: 1967-1991," *Journal of Labor Economics*, 19, 799–836.

HECKMAN, J., AND B. SINGER (1984): "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica*, 52, 63–132.

HIRANO, K. (2002): "Semiparametric Bayesian Inference in Autoregressive Panel data models," *Econometrica*, 70, 781–799.

HOROWITZ, J., AND M. MARKATOU (1996): "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145–168.

HOSPIDO, L. (2012): "Modelling heterogeneity and dynamics in the volatility of individual wages," *Journal of Applied Econometrics*, 27, 386–411.

JIANG, W., AND C.-H. ZHANG (2009): "General maximum likelihood empirical Bayes estimation of normal means," *Annals of Statistics*, 37, 1647–1684.

JUHL, T., AND O. LUGOVSKYY (2014): "A test for slope heterogeneity in fixed effect models," *Econometric Reviews*, 33, 906–935.

KIEFER, J., AND J. WOLFOWITZ (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27, 887–906.

KOENKER, R. (2015): "REBayes: An R package for empirical Bayes methods," Available from `http://cran.r-project.org`.

KOENKER, R., AND J. GU (2013): "Frailty, Profile Likelihood and Medfly Mortality," in *Contemporary Developments in Statistical Theory: A Festschrift for Hira Lal Koul*, ed. by S. Lahiri, A. Schick, A. Sengupta, and T. Sriram. Springer.

KOENKER, R., AND I. MIZERA (2014): "Convex Optimization, Shape Constraints, Compound Decisions and Empirical Bayes Rules," *J. of Am. Stat. Assoc.*, 109, 674–685.

LAIRD, N. (1978): "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, 73, 805–811.

LILLARD, L., AND Y. WEISS (1979): "Components of variation in panel earnings data: American scientists, 1960-1970," *Econometrica*, 47, 437–454.

LINDLEY, D. V., AND A. F. SMITH (1972): "Bayes estimates for the linear model," *Journal of the Royal Statistical Society. Series B*, 34, 1–41.

MACURDY, T. (1982): "The use of time series processes to model the error structure of earnings in a longitudinal data analysis," *Journal of Econometrics*, 18, 83–114.

MEGHIR, C., AND L. PISTAFERRI (2004): "Income Variance Dynamics and Heterogeneity," *Econometrica*, 72, 1–32.

MURPHY, S. A., AND A. W. VAN DER VAART (2000): "On profile likelihood," *Journal of the American Statistical Association*, 95, 449–465.

NEYMAN, J. (1959): "Optimal asymptotic tests of composite statistical hypotheses," in *Probability and Statistics: The Harald Cramer Volume*, ed. by U. Grenander. New York, John Wiley.

NEYMAN, J., AND E. SCOTT (1966): "On the use of $C(\alpha)$ tests of composite hypotheses," *Bull. Inst. Int. Statist.*, 41, 477–497.

ROBBINS, H. (1951): "Asymptotically subminimax solutions of compound statistical decision problems," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press: Berkeley.

——— (1956): "An empirical Bayes approach to statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press: Berkeley.

——— (1982): "Estimating Many Variances," in *Statistical Decision Theory and Related Topics III*, ed. by S. Gupta, and J. O. Berger, vol. 2. Academic Press: New York.

ROBINSON, G. K. (1991): "That BLUP is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6, 15–51.

TUKEY, J. (1974): "Named and Faceless Values: An Initial Exploration in Memory of Prasanta C. Mahalanobis," *Sankhyā*, 36, 125–176.

WHITE, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, pp. 1–25.

APPENDIX A. PROOF OF PROPOSITION 1

The simplest way to derive Tweedie's Formula in Proposition 1 for the Gaussian case seems to be to consider the more general exponential family compound decision problem in which

$$g(y) = \int \varphi(y, \eta) dF(\eta),$$

where φ is a known exponential family density with natural parameter η, so we may write,

$$\varphi(y, \eta) = m(y) e^{y\eta} h(\eta),$$

and F is again a mixing distribution over the parameter η. Quadratic loss implies that the Bayes rule is the conditional mean:

$$\begin{aligned} \delta(y) &= \mathbb{E}[\eta | Y = y] \\ &= \int \eta \varphi(y, \eta) dF / \int \varphi(y, \eta) dF \\ &= \int \eta e^{y\eta} h(\eta) dF / \int e^{y\eta} h(\eta) dF \\ &= \frac{d}{dy} \log(\int e^{y\eta} h(\eta) dF \\ &= \frac{d}{dy} \log(g(y)/m(y)) \end{aligned}$$

Differentiating again,

$$\begin{aligned} \delta'(y) = \frac{d}{dy} \left[\frac{\int \eta \varphi dF}{\int \varphi dF} \right] &= \frac{\int \eta^2 \varphi dF}{\int \varphi dF} - \left(\frac{\int \eta \varphi dF}{\int \varphi dF} \right)^2 \\ &= \mathbb{E}[\eta^2 | Y = y] - (\mathbb{E}[\eta | Y = y])^2 \\ &= \mathbb{V}[\eta | Y = y] \geqslant 0, \end{aligned}$$

implying that δ must be monotone. When φ is Gaussian with known variance θ we have natural parameter $\eta = \alpha/\theta$,

$$\varphi(y, \alpha/\theta) = \phi((y - \alpha)/\sqrt{\theta})/\sqrt{\theta} = K \exp\{-(y - \alpha)^2/2\theta\} = K e^{-y^2/2\theta} \cdot e^{y\alpha/\theta} \cdot e^{-\alpha^2/2\theta},$$

so $m(y) = e^{-y^2/2\theta}$ and the logarithmic derivative yields our Bayes rule in Proposition 1.

APPENDIX B. PROOF OF PROPOSITION 2

Suppose we have $y_{it} \mid \alpha_i, \theta_i \sim \mathcal{N}(\alpha_i, \theta_i)$. Let $\bar{y}_i$ and S_i be defined respectively as the sample mean and sample variance with conditional density $\phi(\bar{y}_i \mid \alpha_i, \theta_i)$ and $\gamma(S_i \mid \theta_i)$. Denote the marginal density for the vector $(y_{i1}, \dots, y_{im_i})$ as $g(\bar{y}_i, S_i)$. Under squared error loss, we wish to minimize the expected loss,

$$\min_{\hat{\alpha}} \; \mathbb{E}_{\alpha,\theta}[\|\hat{\alpha} - \alpha\|_2^2].$$

This leads to the Bayes rule:

$$\begin{aligned}
\hat{\alpha}_i = \mathbb{E}[\alpha \mid \bar{y}_i, S_i] &= \int_\alpha \alpha \int_\theta f(\alpha, \theta \mid \bar{y}_i, S_i) d\theta d\alpha \\
&= \int_\theta (\int_\alpha \alpha \phi(\bar{y}_i \mid \alpha, \theta) h(\alpha \mid \theta) d\alpha) \gamma(S_i \mid \theta) h(\theta) d\theta / g(\bar{y}_i, S_i) \\
&= \int_\theta \mathbb{E}[\alpha \mid \bar{y}_i, \theta] \tfrac{\gamma(S_i|\theta) \int_\alpha \phi(\bar{y}_i|\alpha,\theta) h(\alpha|\theta) d\alpha h(\theta)}{g(\bar{y}_i, S_i)} d\theta \\
&= \int_\theta \mathbb{E}[\alpha \mid \bar{y}_i, \theta] f(\theta \mid \bar{y}_i, S_i) d\theta
\end{aligned}$$

To check for monotonicity with respect to $\bar{y}$, we differentiate $\mathbb{E}[\alpha \mid \bar{y}, S]$ for some fix S, which leads to

$$\frac{d}{d\bar{y}} \mathbb{E}(\alpha \mid \bar{y}, S) = \int \frac{d}{d\bar{y}} \mathbb{E}(\alpha \mid \bar{y}, \theta) f(\theta \mid \bar{y}, S) d\theta + \int \mathbb{E}(\alpha \mid \bar{y}, \theta) \frac{d}{d\bar{y}} f(\theta \mid \bar{y}, S) d\theta$$

The first piece is non-negative due to Proposition 1, while the sign of the second piece is undetermined.

APPENDIX C. PROOF OF PROPOSITION 3

For the gamma mixture case there is an analogous formula as the Gaussian case in Proposition 1, for the natural parameter $-r/\theta$, proceeding as before,

$$\tilde{\delta}(x) = E[-r/\theta | X = x] = \frac{\int -\frac{r}{\theta} f(x|\theta) dG(\theta)}{\int f(x|\theta) dG(\theta)} = \frac{d}{dx} \log(\frac{g(x)}{x^{r-1}}) = \frac{g'(x)}{g(x)} - \frac{r-1}{x}.$$

If, on the other hand, we would, quite naturally, like to compute the expectation of the *unnatural* parameter θ, then we obtain instead,

$$\begin{aligned}
\delta(x) &= \mathbb{E}[\Theta | X = x] \\
&= \int \theta \gamma(x, \theta) dF(\theta) / \int \gamma(x, \theta) dF(\theta) \\
&= \int \frac{\theta}{\Gamma(r) (\theta/r)^r} x^{r-1} \exp(-xr/\theta) dF(\theta) / g(x) \\
&= r x^{r-1} \int \frac{(r/\theta)^r}{\Gamma(r)} \frac{\theta}{r}) \exp(-xr/\theta) dF(\theta) / g(x) \\
&= r x^{r-1} \int \frac{(r/\theta)^r}{\Gamma(r)} \int_x^\infty \exp(-xr/\theta) dF(\theta) / g(x) \\
&= r x^{r-1} \int_x^\infty y^{r-1} \int \gamma(y|\theta) dF(\theta) dy / g(x) \\
&= r x^{r-1} \int_x^\infty y^{1-r} g(y) dy / g(x).
\end{aligned}$$

It remains to show that this formulation of the Bayes rule is monotone:

$$\begin{aligned}
\delta'(x) &= \frac{r(r-1)x^{r-2}\int_x^\infty y^{1-r}g(y)dy + rx^{r-1}(-x^{1-r})g(x)}{g(x)} \\
&\quad - \frac{rx^{r-1}\int_x^\infty y^{1-r}g(y)dy g'(x)}{g^2(x)} \\
&= \frac{rx^{r-1}\int_x^\infty y^{1-r}g(y)dy}{g(x)}\left[\frac{r-1}{x} - \frac{g'(x)}{g(x)}\right] - r \\
&= -\delta(x)\tilde{\delta}(x) - r \\
&= \delta(x)\left[E[\frac{r}{\theta}|X=x] - \frac{r}{E[\theta|X=x]}\right] \\
&= \delta(x)r\left[E[\frac{1}{\theta}|X=x] - \frac{1}{E[\theta|X=x]}\right],
\end{aligned}$$

which is positive by Jensen's inequality.