EMPIRICAL BAYES FOR THE RELUCTANT FREQUENTIST

ROGER KOENKER AND JIAYING GU

Abstract. Empirical Bayes methods offer valuable tools for a large class of compound decision problems. In this tutorial we describe some basic principles of the empirical Bayes paradigm stressing their frequentist interpretation. Emphasis is placed on recent developments of nonparametric maximum likelihood methods for estimating mixture models. A more extensive introductory treatment will eventually be available in Koenker and Gu (2024). The methods are illustrated with an extended application to models of heterogeneous income dynamics based on PSID data.

1. Introduction

Empirical Bayes decision theory as introduced by Robbins (1951, 1956) represented a challenge to both the Wald (1950) and Savage (1954) strands of classical decision theory. Together with the revelations of Stein (1956) on the inadmissibility of the sample mean of a multivariate Gaussian vector in dimensions greater than two, Robbins' results showed that compound decision problems, that is, ensembles of exchangeable decision problems could be fruitfully combined to yield improved decisions for the entire ensemble. In effect, prior information could be extracted from the ensemble yielding decision rules that performed better than classical procedures that treated each problem in isolation.

A simple example illustrating the benefit of this "borrowing of strength" from several related problems appears in Robbins (1951). Suppose we observe independent realizations, $Y_1, \cdots, Y_n$ with each $Y_i \sim \mathcal{N}(\theta_i, 1)$ and $\theta_i \in \{-1, 1\}$. Our objective is to choose $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_n)$ to minimize the aggregate loss,

$$L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (2n)^{-1} \sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|.$$

The minimax rule for any single component of this problem is to choose $\hat{\theta}_i = \operatorname{sgn}(y_i)$. Robbins shows that this rule is also minimax for every component of the entire vector. The least favorable configuration of the θ_i is one for which each component is ± 1 with equal probability as if decided by a coin flip. This minimax rule yields an expected loss of $\Phi(-1) \approx 0.159$; it is also the minimax regret procedure. Can we do any better

Version: 18 September, 2024. This paper was commissioned for a special issue of JPE Micro titled "A Survey of Some Modern Econometric Methods" and will appear in due course.

than this? Suppose, having seen the entire sample, $y_1, \cdots, y_n$, we find that most of observations are positive. Should we really stick stubbornly with the minimax view that the θ_i take values, ± 1 with equal probability? It seems rather perverse. Suppose we knew that the probability that $\theta = 1$ was p, then we could easily compute the conditional probability that $\theta = 1$ given any particular y as,

$$\mathbb{P}_p(\theta = 1|y) = \frac{p\varphi(y-1)}{p\varphi(y-1) + (1-p)\varphi(y+1)}.$$

Evaluating the standard Gaussian density, φ, and simplifying we obtain the alternative decision rule,

$$\hat{\theta}_i = \text{sgn}(y + 0.5 \log(p/(1-p))).$$

Robbins suggests that we might *estimate* the probability p by $\hat{p} = (\bar{y} + 1)/2$, and consider a plug-in version of this revised decision rule. The logit adjustment has the desirable effect of biasing our estimates toward the more likely of the two alternatives, for example if $\hat{p} = 3/4$ we would use $\hat{\theta}_i = \text{sgn}(y_i + .549)$, reflecting our evidence that $Y_i = 1$ was more likely than its alternative. This is an empirical Bayes procedure. There has been no reliance on subjective prior information, nor on the pessimistic principle of minimaxity, only the empirical evidence of the observed sample has been relied upon. Robbins' method of moments estimator of p, could be replaced by other approaches such as maximum likelihood or even a formal conjugate Bayes procedure with a Beta prior, but ignoring the information contained in the ensemble of problems can be extremely costly in this setting. With our hypothetical $p = 3/4$ the asymptotic risk of using $\hat{p}$ is reduced by about 20 percent. This is a leading example of Robbins' claim for the "asymptotic sub-minimaxity" of empirical Bayes procedures relative to the minimax procedures of Wald.

For the Bayesophobic the foregoing example should raise no anxieties, they need only consider how to go about constructing a reasonable estimate of the probability p. Likewise, Bayesians should be entirely comfortable computing a full posterior for p that could be used to inform their decisions about the θ_i's. More complicated compound decision problems will require more complicated estimation of the mixture structure of the problem, and therein lies the charm of the empirical Bayes approach. Bayesian language and the machinery of Bayesian computation will prove to be convenient, but no further commitment to the catechism of the Reverend Bayes is required. Indeed, we will argue that empirical Bayes methodology represents an alloy of the best features of the frequentist and Bayesian traditions.

2. The Compound Decision Paradigm

Suppose that we are faced with an exchangeable, i.e. permutation invariant, ensemble of related problems:

$$Y_i \sim \varphi(y|\theta_i), \quad i = 1, 2, \cdots, n,$$

where φ denotes some familiar – usually exponential family – density, indexed by parameters, θ_i, that express the underlying heterogeneity of the problems. Our task is to choose a decision rule, $\boldsymbol{\delta} : \mathcal{Y} \to \mathcal{A}$, mapping realizations $Y = (Y_1, \cdots, Y_n)$ in the sample space, $\mathcal{Y}$, to actions, $a = \delta(Y)$, in an action space $\mathcal{A}$ that minimize the compound risk,

$$R_n(\boldsymbol{\delta}, \boldsymbol{\theta}) = n^{-1}\mathbb{E}\sum_{i=1}^{n} L(\boldsymbol{\delta}_i(Y), \theta_i).$$

Following Robbins (1956) we will restrict attention to simple, symmetric decision rules for which $\boldsymbol{\delta}_i(y_1, \cdots, y_n) = \delta(y_i)$. Such rules are now often referred to as separable rules. The function, δ may depend upon the entire sample, but given our objective and the exchangeable probabilistic structure of the problems there seems to be little merit in venturing beyond this class. With this restriction we can write,

$$\begin{aligned} R_n(\delta, \boldsymbol{\theta}) &= n^{-1}\mathbb{E}\sum_{i=1}^{n} L(\delta(Y_i), \theta_i) \\ &= n^{-1}\sum_{i=1}^{n}\int L(\delta(y), \theta_i)\varphi(y|\theta_i)dy \\ &= \int\int L(\delta(y), \theta)\varphi(y|\theta)dG_n(\theta)dy \end{aligned}$$

where $G_n(\theta) = n^{-1}\sum_{i=1}^{n} \mathbb{1}(\theta_i \leq \theta)$ denotes the empirical distribution function of the θ_i's. This is obviously an oracle risk function since we do not know G_n, but it provides a natural benchmark for evaluating the performance of various feasible decision rules.

How related do the problems need to be? Robbins (1951) infamously opined,

> No relation whatever is assumed to hold among the unknown parameters θ_i. To emphasize this point, Y_1 could be an observation on a butterfly in Ecuador, Y_2 on an oyster in Maryland, Y_3 the temperature of a star, and so on, all observations being taken at different times.

Of course, this is a bit disingenuous since we have already asserted that observations share a common conditional density, and that they are exchangeable. Ultimately, this issue of relatedness, or what Efron (2010) refers to as relevance, is bound up with the form of the mixing distribution G_n.

Theorem 1 (Fundamental Theorem of Compound Decisions)**.** *For separable decision rules compound risk is equal to the Bayes risk of a single copy of the compound decision problem with respect to the "prior" G_n.*

Thus, an optimal decision rule for our compound decision problem can be expressed as a Bayes rule minimizing posterior loss with respect to the "prior" G_n:

$$\begin{aligned} B_n(\delta) &= \int \left\{ \int_\Theta L(\delta(y), \theta) \varphi(y|\theta) dG_n(\theta) \right\} dy. \\ &= \int \left\{ \int_\Theta L(\delta(y), \theta) h(\theta|y) d\theta \right\} f(y) dy. \end{aligned}$$

where $h(\theta|y) = \varphi(y|\theta)dG_n(\theta)/f(y)$, is the posterior distribution of θ given $Y = y$, and $f(y) = \int \varphi(y|\theta)dG_n(\theta)$ is the marginal density of the Y_i.

There are many ways to break the appealing simplicity of the Fundamental Theorem. Exchangeability of our latent parameters, θ_i, is a powerful condition. We might suppose instead that they followed some non-degenerate stochastic process; designing simple decision rules would become much more difficult. A common feature of most empirical Bayes applications in economics is that observations, Y_i arrive with some associated measure of precision, say σ_i. Separable rules, $\delta(Y_i)$ no longer suffice, but if we consider expanding the class of decision rules to take the form $\delta(Y_i, \sigma_i)$ then we can write,

$$\begin{aligned} R_n(\delta, \boldsymbol{\theta}) &:= n^{-1} \mathbb{E} \sum_{i=1} L(\delta(Y_i, \sigma_i), \theta_i) \\ &= n^{-1} \sum_{i=1} \int \int L(\delta(y, \sigma), \theta_i) f(y, \sigma|\theta_i) dy d\sigma. \end{aligned}$$

In so doing we have not made any assumption about the nature of the dependence between θ_i and σ_i. We will discuss several variants of such models in Section 6. The presence of covariates is another source of potential jeopardy for the Fundamental Theorem. For panel data the familiar Mundlak (1978) critique argues that fixed effects estimation is preferable to random effects due to presumed dependence between the latent "individual specific" effects and the observed covariates. This view seems overly dogmatic in settings where the latent structure is of primary interest. Indeed, much of the recent applied literature using empirical Bayes methods opts instead for a preliminary projection step to remove covariate effects. When the covariate dimension is low then it is feasible to embed profile likelihood methods as we shall see in Section 6. Greenshtein and Ritov (2019) offer some other innovative ideas for handling covariates in the empirical Bayes framework.

From a formal Bayesian perspective exchangeability of the Y_i's yields via de Finetti's theorem a mixture representation of the Bayes risk with a generic prior, G, replacing the frequentist, G_n. At this point the two warring perspectives are essentially indistinguishable; controversy arises only when we begin to ask where can we find a viable G that will enable us to make good decisions. In the absence of a priori knowledge of G, this apparently requires some way to estimate the empirical distribution function G_n of the latent parameters θ_i. In our introductory example with $\theta_i \in \{-1, 1\}$ this

entailed estimating a scalar probability, more general settings demand more. We will consider parametric models for G_n in the next section, and then proceed to consider nonparamtric methods in the following section. Robbins with characteristic wit referred to this task as "Estimating the Inestimable" in a lecture series presented at Berkeley in the late 1980s.

3. Parametric Priors

When in doubt about the form of an unknown distribution, the Gaussian shape of Napoleon's hat naturally springs to mind. Suppose, instead of restricting the θ_i's to take only the values ± 1, we assume instead that they are Gaussian with mean, zero, and variance, σ_0^2, i.e. $\theta_i \sim \mathcal{N}(0, \sigma_0^2)$, and retain the assumption that they are embedded in standard Gaussian noise. It follows that the marginal distribution of the Y_i's is Gaussian with mean, zero, and variance, $1 + \sigma_0^2$. Under quadratic loss, $L(\delta(y), \theta) = (\delta(y) - \theta)^2$, the posterior mean,

$$\delta(y) = \mathbb{E}(\theta | Y = y) = \left(1 - \frac{1}{1 + \sigma_0^2}\right) y,$$

is the optimal Bayes rule, provided that σ_0^2 is known. If not, we can rely instead on an estimate based on $S \equiv \sum_{i=1}^n Y_i^2 \sim (1 + \sigma_0^2)\chi_n^2$. Recalling that an inverse χ_n^2 random variable has expectation, $(n-2)^{-1}$, we obtain the method of moments, empirical Bayes estimator,

$$\hat{\delta}(y) = \left(1 - \frac{n-2}{S}\right) y.$$

This is the classical James-Stein estimator, James and Stein (1961). It has strictly smaller compound risk than the inadmissible maximum likelihood estimator, $\bar{\delta}(y) = y$ when $n \geq 3$ for any sequence of θ_i's. Linear shrinkage of each coordinate toward zero may not result in a more accurate estimate for any one coordinate, but the compound risk of the entire vector is reduced. See the Appendix for a formal exposition.

There are many variations on the basic James-Stein rule: if we allow the Gaussian prior to have (unknown) mean, θ_0, estimable by the sample mean, $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$, we obtain the Efron-Morris rule,

$$\tilde{\delta}(y) = \bar{Y}_n + \left(1 - \frac{n-3}{\tilde{S}}\right)(y - \bar{Y}_n),$$

with $\tilde{S} \equiv \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$, so shrinkage occurs toward the sample mean rather than zero. It can happen that $\tilde{S} < n - 3$ in which case the shrinkage factor would flip the signs of the coordinate effects. This motivates consideration of "positive part" variants of the foregoing rules that restrict shrinkage to be non-negative.

Stein's revelation that the maximum likelihood estimator for the mean of a multivariate Gaussian random vector was inadmissible under quadratic loss in dimensions

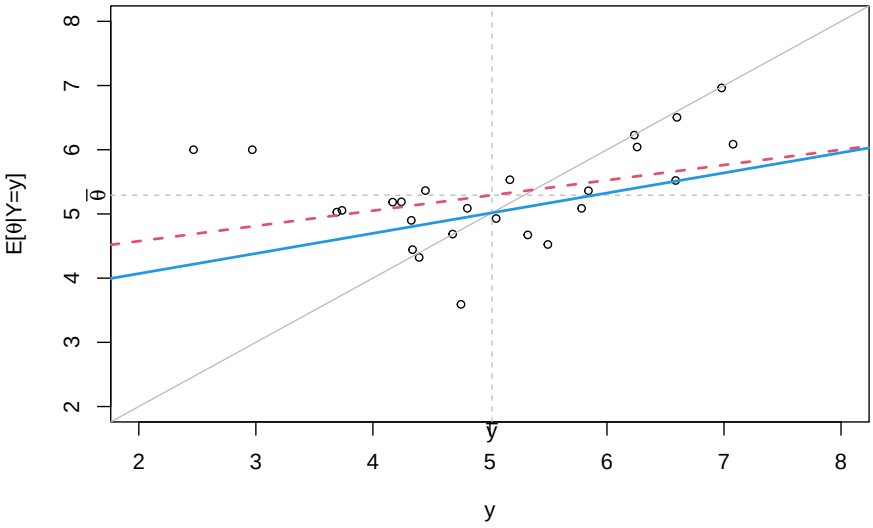

FIGURE 1. Stigler's Galtonian perspective on Stein shrinkage views it as an attempt to mimic the regression of the (latent) θ_i's on the observed y_i's, noting that the coefficients of this regression can be estimated without full knowledge of the θ_i's using only moment informations from the y_i's. In the figure, the solid line represents an oracle version of the Stein rule, while the dashed line is an estimate thereof; both shrink observed y_i's toward their collective sample mean in an effort to replace the naive maximum likelihood estimates, $\delta(y_i) = y_i$ (on the 45 degree line) with a more precise estimate of $\delta(y) = \mathbb{E}(\theta|Y = y)$.

greater than two came as a surprise to many practicing statisticians and is still perhaps not fully appreciated in the econometrics literature.

Stigler's "Galtonian perspective on shrinkage estimators," Stigler (1990), offers a novel interpretation of Stein shrinkage that may aid intuition. Imagine a thought experiment in which we observe not only the Y_i's, but also their associated θ_i's, which we can then plot as the points appearing in Figure 1. We are interested in the conditional expectation, $\mathbb{E}(\theta|Y = y)$, which is nothing but the regression line in this figure. Of course we don't have access to the actual θ_i's, so this might seem a bit fanciful, but the coefficients of the regression line require only the mean of the θ_i's and the $\mathbf{C}\text{ov}(Y, \theta)$, which can be consistently estimated by $\bar{Y}_n$ and $\tilde{S}/(n-2) - 1$, respectively. In the Figure the population regression line is depicted as the dashed red line, while the sample regression is dotted and blue. The grey 45 degree line contrasts the unshrunken, $\delta(y) = y$ estimator with the two variants of the Stein rule.

Relaxing the assumption of a homoscedastic prior we may incorporate heterogeneity in the precision of the observed Y_i's. Consider the model,

$$y_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \ldots, n, \ j = 1, \ldots, J.$$

The investigator observes pairs, $(\hat{\mu}_i, \hat{\sigma}_i^2)$ with $\hat{\mu}_i = J^{-1} \sum_j y_{ij}$ and $\hat{\sigma}_i^2 = (J-1)^{-1} \sum_j (y_{ij} - \hat{\mu}_i)^2$; the objective is often to test the hypotheses: $H_0 : \mu_i = 0$, or to select a subset that violate these hypotheses. In genomics this is often accomplished with a parametric empirical Bayes procedure implemented in the R package **limma**. Rather than positing a prior on the full parameter space, limma assumes that the σ_i^2 are independent of the μ_i and are drawn iidly from the inverse chi-squared distribution,

$$\sigma_i^{-2} \sim (v_0 s_0^2)^{-1} \chi_{v_0}^2 \quad i = 1, \ldots, n.$$

The hyperparameters, v_0 and s_0 can be estimated by maximum likelihood. The null hypotheses $H_0 : \mu_i = 0$ have p-values,

$$p_i = 2(1 - F_{t,v_0+v}(|\tilde{t}_i|)),$$

where $\tilde{t}_i = \hat{\mu}_i / \tilde{s}_i$, $\tilde{s}_i^2 = (v_0 s_0^2 + v \hat{\sigma}_i^2)/(v_0 + v)$, $v = J - 1$ and $F_{t,v}$ is the distribution function of a Student t random variable with v degrees of freedom. The parametric prior for the σ_i^2 shrinks the $\hat{\sigma}_i^2$ toward a common value and is intended to improve precision. The imposition of prior information on nuisance parameters while leaving the parameters of primary interest, in this case the μ_i, alone is referred to as "partially Bayes" by Cox (1975). See Ignatiadis and Sen (2023). More flexible nonparametric procedures will be considered in the next section.

Lindley and Smith (1972), appealing to results of de Finetti (1964) greatly elaborated the paradigm of Gaussian linear models with parametric Gaussian priors, and thus initiated the modern development of hierarchical models. They start from the compound decision premise that observations arise in an exchangeable fashion, so the Y_i's come from a mixture density. When both the conditional density, $\varphi(y \mid \theta)$, and the mixing distribution, G, are Gaussian, this leads to some elegant linear algebra.

Proposition 1 (Lindley and Smith)**.** *Suppose, for $\theta_1 \in \mathbb{R}^{p_1}$, our observed response $y \in \mathbb{R}^n$ is multivariate Gaussian with mean vector $A_1\theta_1$ and covariance matrix C_1, i.e. $y \sim \mathcal{N}(A_1\theta_1, C_1)$. Then, if θ_1 is also Gaussian, $\theta_1 \sim \mathcal{N}(A_2\theta_2, C_2)$, the marginal distribution of y is $\mathcal{N}(A_1A_2\theta_2, C_1 + A_1C_2A_1^\top)$ and $\theta_1|Y \sim \mathcal{N}(Bb, B)$ where $B^{-1} = A_1^\top C_1^{-1} A_1 + C_2^{-1}$, and $b = A_1^\top C_1^{-1} y + C_2^{-1} A_2\theta_2$.*

The linear model structure makes these shrinkage formulae more complicated, but it is still possible to recognize that the posterior mean of θ_1 is a matrix weighted average of the response vector, y, and the prior mean $A_2\theta_2$. The closely related papers of Chamberlain and Leamer (1976); Leamer and Chamberlain (1976) offer further insight into this phenomenon. Estimation of such hierarchical models generally requires some form of MCMC procedure, although when the covariance matrices C_1 and C_2 are spherical the formulae reduce to classical variance component analysis that goes

back to work by Balestra and Nerlove (1966) on the demand for natural gas in the econometrics literature.

Parametric non-Gaussian priors for linear regression models have emerged as a familiar device in modern high dimensional statistical modeling. When we consider penalized regression estimators that optimize,

$$\min \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda P(\beta),$$

we have available a large menu of choices for the penalty function P. The lasso penalty, $P(\beta) = \|\beta\|_1$, employing the ℓ_1-norm, is most familiar once we depart from the Gaussian "ridge" penalty, $P(\beta) = \|\beta\|_2^2$. It corresponds to imposing independent Laplace priors on β. Choice of the tuning parameter, λ controls how strongly we believe in the prior. As soon as we choose λ by cross-validation or some other form of sorcery we have ventured into the realm of empirical Bayes.

Another prominent option for the penalty, P, is treat the coordinates of β as if they were drawn iidly from a distribution with Cauchy tail behavior, as considered by Johnstone and Silverman (2004) and Castillo and van der Vaart (2012). Although such priors are generally structured to shrink coefficients toward zero, this is typically rationalized by some form of prior standardization of the design matrix. The empirical aspect of these procedures is generally restricted to choice of the tuning parameter λ representing the scale of the prior density. However, more flexibility can be achieved by permitting larger parametric families, for example Azevedo et al (2020) a class of Student priors for large-scale A/B testing settings with location, scale and degrees of freedom of the prior estimated by maximum likelihood.

Returning to the simple Gaussian sequence model with scalar parameters θ_i, in Figure 2 we contrast several forms of shrinkage: linear shrinkage with the classical Stein rule, the Laplace (lasso) procedure that shrinks a fixed amount in the tails but only moderately when y is near zero, and the Cauchy penalty that shrinks very aggressively near zero while large departures from zero are shrunken very little. It should be stressed that tuning the location and scale of these penalties offers some flexibility, but the selection of a functional form for such parametric priors involves a leap of Bayesian faith that may trouble some researchers.

Thus far we have focused entirely on settings in which our base model, $\varphi(y|\theta)$ is Gaussian. Parametric mixture priors play an important role in many other corners of statistics. Poisson models are often paired with gamma mixing, and the modern literature on survival analysis, is permeated by parametric models of "frailty." As anticipated by the pioneering critique of Heckman and Singer (1984), choosing a specific parametric model for frailty can be difficult so it is natural to turn to nonparametric methods for guidance.

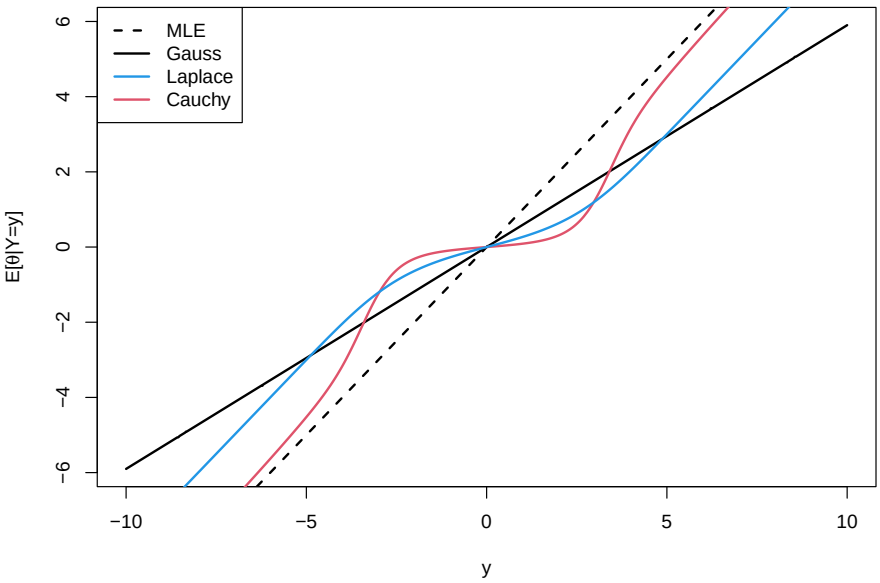

FIGURE 2. Comparison of several parametric shrinkage methods.

4. NONPARAMETRIC PRIOR ESTIMATION

Parametric models for the prior, G, can be difficult to choose, so it is natural to ask whether some nonparametric procedure can be used to estimate G. An affirmative answer to this question can be traced back to an abstract of Robbins (1950), much more fully elaborated in Kiefer and Wolfowitz (1956). Some further development of the idea of nonparametric maximum likelihood estimation of mixture models was made by Pfanzagl (1988), but practical implementation of these methods was delayed until Laird (1978) showed how the nascent EM algorithm, Dempster et al (1977), could be used to compute it. Heckman and Singer (1984) pioneered the EM approach to explore the sensitivity to various parametric frailty models of duration models in econometrics. Lindsay (1981, 1995) further clarified many aspects of the NPMLE, but computation remained a bottleneck due to slow convergence of the EM algorithm. Fortunately, modern developments in convex optimization have substantially improved computational prospects for the NPMLE.

It is easy to see that the NPMLE problem for mixtures is convex. We have a strictly convex objective subject to linear constraints over the convex set, $\mathcal{G}$, of distribution

functions:

$$\min_{G\in\mathcal{G}}\left\{-\sum_{i=1}^{n}\log f(y_i)\mid f(y_i)=\int\varphi(y_i|\theta)dG(\theta),\ i=1,\cdots,n\right\}$$

The problem is infinite dimensional, but it can be solved to desirable precision by restricting the support of solutions to a finite dimensional grid $(t_1 < t_2 < \cdots < t_m)$ contained in an interval between the minimum and maximum of the observations. Finite dimensional formulations can be expressed given this restriction. In primal form with A denoting an n by m matrix with typical element $\varphi(y_i|t_j)$, and $\boldsymbol{g}\in\mathcal{S}_m$ denoting the masses associated with each of the grid points and $\mathcal{S}_m$ the m-dimensional unit simplex, a primal formulation is given by:

$$\text{(P)}\qquad \min_{\boldsymbol{g}\in\mathcal{S}_m}\left\{-\sum_{i=1}^{n}\log f_i\mid \boldsymbol{f}=A\boldsymbol{g},\ \right\}$$

and in dual form as:

$$\text{(D)}\qquad \max_{\boldsymbol{\nu}\in\mathbb{R}^n}\left\{\sum_{i=1}^{n}\log\nu_i\mid A^\top\boldsymbol{\nu}\le n\mathbb{1}_m\ \right\}.$$

The dual form is usually somewhat more convenient for computation, and the primal solution can be easily recovered from the dual solution by solving for the coordinates of $\boldsymbol{g}$ from the linear system,

$$\sum_j\varphi(y_i|t_j)g_j=\frac{1}{\hat{\nu}_i},$$

restricted to the set $\{i:\hat{\nu}_i>0\}$, that is to those observations whose dual constraint is active at the dual solution. The number of these active constraints, m^*, is typically far fewer than its obvious upper bound of n, indeed it has been recently shown by Polyanskiy and Wu (2020) that as $n\to\infty$, $m^*=\mathcal{O}(\log n)$ when G has sub-Gaussian tails. This self-regularizing feature of the NPMLE may come as a surprise since many infinite dimensional inverse problems are ill-posed and do require some form of regularization, and consequent tuning parameter selection. In contrast, the NPMLE determines of the number, location, and mass of the atoms of the $\hat{G}$ solution from the data without any interference by the analyst.

Identifiability of G in mixture models is thoroughly treated by Teicher (1961, 1967) for a scalar mixing parameter.

Definition 1. *Let $\Phi(y|\theta)$ be a distribution function defined for all $\theta\in\Theta\subset\mathbb{R}$ and G be a distribution function defined on Θ, the mixture,*

$$F(y)=\int_\Theta\Phi(y|\theta)dG(\theta)$$

is identifiable if and only if there is a unique G yielding F.

For location, $\Phi(y-\theta)$, and scale, $\Phi(y\theta)$, mixtures this follows from the uniqueness of the characteristic function provided that the Fourier transforms of $\Phi(y)$ and $\Phi(e^y)$ respectively are not identically zero on some non-degenerate real interval, a condition that is trivially satisfied in most of the conventional empirical Bayes settings in which $\varphi(y|\theta)$ is a continuous (Lebesgue) density. In multivariate settings and many discrete data settings identification conditions are more delicate and partial identification is not uncommon. See Koenker and Gu (2024) for further details and references.

Given an estimate, $\hat{G}$, what should we do with it? Minimizing posterior compound loss is simplest with quadratic loss since it requires only computing posterior means. This is particularly convenient when the base distribution is of the exponential family form.

Proposition 2. *(Tweedie) For φ of the (natural) exponential family form,*

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

the posterior mean is,

$$\delta(y) \equiv \mathbb{E}(\eta|Y=y) = \frac{d}{dy}\log(f_G(y)/m(y)),$$

where $f_G(y) = \int \varphi(y|\eta)dG(\eta)$, is the marginal distribution of Y. And $\delta(y)$ is non-decreasing in y.

See Appendix A for a proof and some further details. When φ is standard Gaussian, so $m(y) = e^{-y^2/2}$, the posterior mean becomes,

$$\delta(y) \equiv \mathbb{E}(\eta|Y=y) = y + f_G'(y)/f_G(y).$$

In this Gaussian case, $f_G'(y)/f_G(y)$ can be interpreted as a shrinkage term that pulls the naive estimator, $\hat{\eta} = y$ to its posterior mean.

It is tempting to interpret the Tweedie formula as an invitation to estimate the marginal density f_G and construct estimates of the posterior mean accordingly. In the terminology of Efron (2014), this would be f-modeling in contrast to g-modeling which relies on a preliminary estimator of G. Although estimation of f_G by conventional kernel methods is easy, estimation of its log derivative is more challenging. More seriously, when φ is of the exponential family form Proposition 2 asserts that the posterior mean function is monotone in y, a condition that is awkward to impose for standard density estimators. See Koenker and Mizera (2014) for a proposal for a shape constrained nonparametric maximum likelihood estimator of f_G that imposes such a monotonicity constraint.

Example. To illustrate the NPMLE in a simple Gaussian location mixture setting, suppose that we observe a random sample of size $n = 1000$, with $Y_i \sim \mathcal{N}(\theta_i, 1)$ and $(\theta_i - 3)$ distributed as standard lognormal. The left panel of Figure 3 shows a histogram of the observed, Y_i's, and superimposed in red are the mass points of the estimated, $\hat{G}$. Of 1000 potential grid point locations for the illustrated NPMLE

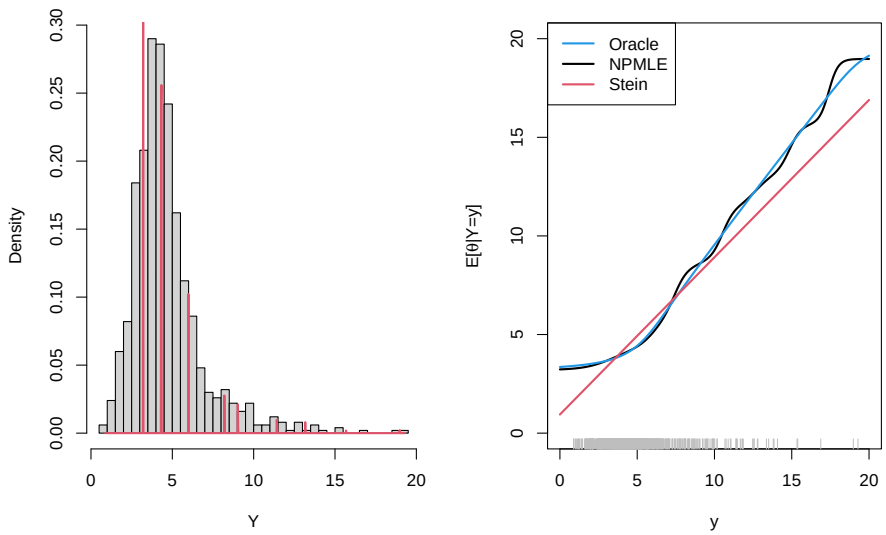

FIGURE 3. Left panel: Histogram of $n = 1000$ observations from a standard Gaussian location mixture with 3-parameter lognormal mixing distribution. An NPMLE estimate of the mixing distribution is superimposed on the histogram as the red mass points. Right panel: Three estimates of the posterior mean of θ.

solution only 13 have associated mass in excess of 0.001. The general shape of the discrete mass points does mimic the shape of the lognormal density that was used to generate the sample y_i's. Viewed as a density estimate this $\hat{G}$ is quite terrible, as an estimate of a lognormal distribution function it is somewhat better. But from a practical decision making perspective, $\hat{G}$ is useful primarily as a device for evaluation of smooth linear functionals like conditional means, and for this it is quite excellent as noted in the Appendix.

The right panel of Figure 3 plots posterior mean functions for three choices of $\hat{G}$. The red line represents the Stein rule. The blue curve depicts the oracle rule based on full knowledge of the lognormal form of G. And the black curve shows the estimated posterior mean corresponding to the NPMLE of G depicted in the left panel of the figure. The linear Stein rule clearly shrinks too much in both tails. The NPMLE estimate sticks quite closely to the oracle estimate except in the far right tail where data is sparse as indicated by the "rug plot" along the horizontal axis, and its discrete form inevitably produces some oscillation around the oracle estimate. ■

Although the posterior mean function of the preceding example is necessarily smoothed by the convolution producing f_G, and thus its log derivative, there are good reasons to consider smoother estimates of G itself. This is especially apparent

once we begin to consider inference for empirical Bayes point estimates or the ranking and selection problems considered in Gu and Koenker (2023). A smoother alternative to the Kiefer-Wolfowitz NPMLE of G, proposed by Efron (2016), expresses the log density of G in the exponential family form,

$$\log g(\theta|\alpha) = \sum_{k=1}^{p} z_k(\theta)\alpha_k - \gamma(\alpha)$$

where $z_k(\theta) \in \mathbb{R},\ k = 1, \cdots, p$ are elements of a natural spline basis expansion, and γ is the usual constant of integration for the one-parameter exponential family density. The penalized Efron log likelihood is then,

$$\ell(\alpha) = \sum_{i=1}^{n} \log f_\alpha(y_i) + c_0\|\alpha\|,$$

where the n vector $f_\alpha = (f_\alpha(y_1), \cdots, f_\alpha(y_n))$ can be written as, $f_\alpha = AZ\alpha$ with Z an m by p matrix with typical element $Z_{jk} = z_k(t_j)$, for a grid of values, $t_1, \cdots, t_m$. Choice of the basis Z, its dimension p, and the penalty parameter c_0 all call for some degree of expert judgement that could be guided by the relatively automatic NPMLE. The penalty term is not strictly necessary, but it allows one to specify a somewhat larger spline dimension p. Letting $p \to \infty$ and $c_0 \to 0$ recovers in the limit the NPMLE, although algorithms that fail to exploit convexity may struggle with the optimization. An alternative way to achieve smoothness of an estimator of G is simply to convolve the NPMLE $\hat{G}$ with some smooth density. This obviously would involve choosing a kernel and bandwidth, choices that might be usefully informed by a careful examination of the discrete form of the NPMLE.

When the latent parameter θ is of dimension two interior point methods for computing the NPMLE are still feasible using gridding as illustrated in Gu and Koenker (2017a,b). However, beyond dimension two such methods become unwieldy and alternative first-order methods are probably required. Recent progress in this direction can be found in Soloff et al (2021), Zhang et al (2022).

5. Empirical Bayes Methods for Discrete Data

The range of empirical Bayes methods extends far beyond the Gaussian mixture settings that we have emphasized thus far. Parametric Poisson mixture models have a long history in actuarial risk analysis and ecology. See, for example Bühlmann and Gisler (2005) and Fisher et al (1943), respectively. Given observations $y_1, \cdots, y_n$ with marginal density,

$$f_G(y) = \int \varphi(y|\theta) dG(\theta),$$

where $\mathbb{P}(Y_i = y|\theta_i) = \varphi(y|\theta_i) = e^{-\theta_i}\theta_i^y/y!$, Robbins (1956) proposed a nonparametric estimator of the posterior mean $\mathbb{E}[\theta|Y_i = y]$,

$$\delta(y) = \frac{\int \theta\varphi(y|\theta)dG(\theta)}{\int \varphi(y|\theta)dG(\theta)} = \frac{(y+1)f_G(y+1)}{f_G(y)}.$$

Since the quantities $f_G(y)$ can be easily estimated by the observed frequencies this is an extremely convenient f-modeling strategy. However, like other f-modeling methods it has the disadvantage that it may fail to respect the monotonicity of the Bayes rule as proscribed by Proposition 2. This is particularly problematic when G is heavy tailed since the tail frequencies of the mixture can be highly variable. A preferable G-modeling strategy is to employ the NPMLE, $\hat{G}$, for G in the Tweedie formula for the posterior mean. Although Polyanskiy and Wu (2021) have recently shown that the original Robbins estimator achieves a sharp asymptotic regret bound, the NPMLE G-modeling approach exhibits significantly improved performance in simulations reported in Koenker and Gu (2024).

Binary response data give rise to a wide variety of mixture models that can be analysed with empirical Bayes methods. In the simplest case, suppose that we have a sample $y_1, \cdots, y_n$ of outcomes from binomial experiments $B(m, p_i)$ as in the tack-flipping experiment of Beckett and Diaconis (1994), who describe the protocol of the experiment as follows.

> The example involves repeated rolls of a common thumbtack. A one was recorded if the tack landed point up and a zero was recorded if the tack landed point down. All tacks started point down. Each tack was flicked or hit with the fingers from where it last rested. A fixed tack was flicked 9 times. The data are recorded in Table 1. There are 320 9-tuples. These arose from 16 different tacks, 2 "flickers," and 10 surfaces. The tacks vary considerably in shape and in proportion of ones. The surfaces varied from rugs through tablecloths through bathroom floors.

Unconditionally on the type of tack and surface the experimental outcomes have marginal mixture density.

$$f_G(y) = \mathbb{P}(Y = y) = \int \binom{m}{y} p^y(1-p)^{m-y}dG(p).$$

Again, the mixing distribution G can be estimated by maximum likelihood. This yields a 3-point mixture for $\hat{G}$ that can be reproduced by running `demo(Bmix1)` in R from the package REBayes. This solution is essentially identical to that reported by Liu (1996) using the EM algorithm although interior point convex optimization methods are considerably quicker.

The binomial mixture model is easily adapted to situations with varying numbers of trials m, but a cautionary note is required regarding identification in such models.

Only $m+1$ distinct frequencies can be observed for $B(m,p)$ binomials and this implies that only m moments of G are identifiable. The consequences of partial identification in discrete response models is discussed in more detail in Koenker and Gu (2024) in the context of the Kline and Walters (2021) model of employment discrimination.

In binomial mixture models with a large number of trials it is often convenient to transform to the Gaussian model as for example in the extensive literature on baseball batting averages, see e.g. Gu and Koenker (2017a), or large correspondence experiments as in Kline et al (2024). In other settings it is more convenient to consider logistic models as for the Rasch model commonly used in educational testing or the Bradley-Terry model for rating participants in pairwise competition. See Gu and Koenker (2022).

6. Empirical Bayes Methods for Panel Data

Longitudinal data poses many new challenges and opportunities for empirical Bayes methods. In this section we will reprise some prior work in Gu and Koenker (2017b) on models of income dynamics and describe some extensions that broaden applicability of such models. The vast econometric literature on panel data methods has gradually embraced a wider variety of latent variable formulations designed to accommodate more general forms of heterogeneity. The quantile autoregression framework of Arellano et al (2017) is notable in this regard. Empirical Bayes methods have a complementary role to play in this literature and also provide a flexible approach to modeling heterogeneity in panel data.

We will begin by considering a simple Gaussian location-scale model,

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

with $u_{it} \sim \mathcal{N}(0,1)$. We need not interpret the t subscript temporally, but it is frequently natural to do so. We will provisionally assume that $\alpha_i \sim G_\alpha$ and $\theta_i \sim G_\theta$ are independent. We then have sufficient statistics:

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

and

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

where $r_i = (m_i - 1)/2$, $S_i = (m_i - 1)^{-1} \sum_{t=1}^{m_i} (y_{it} - \bar{y}_i)^2$, and $\gamma(s|a,b)$ is the density of the gamma distribution with parameters, (a,b). The log likelihood becomes,

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

Since the gamma component of the log likelihood is additively separable from the location component, we can solve for $\hat{G}_\theta$ in a preliminary step, and then solve for the $\hat{G}_\alpha$ 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 NPMLE problem for estimating G_α. As noted by an astute referee, this conditioning is analogous to that described in Reid (1995, equation(3.6)) in a parametric setting and may entail a loss of information. An implementation of this estimation strategy is available with the function `WTLVmix` in the R package REBayes.

It is also possible to relax the independence assumption on the location and scale effects completely. In Gu and Koenker (2017b) we use longitudinal data individuals from the Panel Study on Income Dynamics (PSID) to explore models of income dynamics with an arbitrary joint distribution of location and scale heterogeneity. We follow the sample selection of Meghir and Pistaferri (2004) to focus on male head of households aged 25 - 55 with at least 9 years of consecutive earnings data. Like much of the prior literature including Meghir and Pistaferri (2004) the effects of various individual specific covariates are removed in a preliminary projection step. We further restrict our attention to those whose earning starts from age 25 onwards. This leaves us with 938 individuals for whom we observe at least the early portion of their life cycle earnings. Among the 938 individuals, 50% of those we observe have reported earnings of 15 years starting from age 25, The longest span of recorded earnings in the sample is 26 years.

The implementation employs the function `WGLVmix` from the REBayes package. In the income dynamics application we find an apparent *negative* dependence between the α (location) and θ scale effects indicating that low "ability" individuals also tend to have higher income risk. In our prior work temporal dependence in the income process was specified as a simple AR(1) process whose coefficient, ρ was estimated by profile likelihood.

To make the AR(1) specification more explicit consider the model,

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

Assuming that initial conditions, y_{i0}, are 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 simpler to consider Chamberlainian dependence of the latent effects on covariates, while trying to maintain a nonparametric perspective, a topic we defer to future research. As we will see, more complex short run dynamics can be introduced via state space representations and Kalman filtering formulations of the likelihood.

Fixing ρ, and setting $\tilde{y}_{it} = y_{it} - \rho y_{it-1}$, our model can be expressed as,

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

And for Gaussian ϵ_{it}, sufficient statistics for α_i and θ_i are respectively the sample mean and sample variance:

$$\begin{aligned} \bar{y}_i &= \tfrac{1}{m_i}\textstyle\sum_{t=1}^{m_i} \tilde{y}_{it} \\ S_i &= \tfrac{1}{m_i-1}\textstyle\sum_{t=1}^{m_i}(\tilde{y}_{it}-\bar{y}_i)^2. \end{aligned}$$

Furthermore, we have, $\bar{y}_i \mid \alpha_i, \theta_i \sim \mathcal{N}((1-\rho)\alpha_i, \theta_i/m_i)$ and $(m_i-1)S_i/\theta_i \mid \theta_i \sim \chi^2_{m_i-1}$. Assuming the pairs (α_i, θ_i) are iid with joint distribution function H, we can discretize H on a two dimensional grid and write the likelihood of observing the sample paths $(\tilde{y}_{i1}, \ldots, \tilde{y}_{i,m_i}),\ i = 1, \cdots, n$ as a function of H, ρ, and compute the NPMLE for the distribution H. Profile likelihood can then be employed to estimate the parameter ρ.

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 bivariate 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, a dependence that we have suppressed in the notation, but can be estimated by maximizing the profile log likelihood,

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

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 the innovation distribution. They use a parametric three-component mixture formulation. Hirano (2002) adopts a more flexible Dirichlet prior specification for similar reasons. Browning et al (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 hope of identifying the joint distribution non-parametrically." In contrast, our approach allows only two latent factors corresponding to location and scale of the income process, but has the advantage that it does permit non-parametric estimation of their joint distribution.

The left panel of Figure 4 plots the NPMLE profile likelihood for ρ, which 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), Fan et al (2001), and Chen and Liao (2014). Our estimate of ρ is close to the estimate of Hospido (2012) who also allows an 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

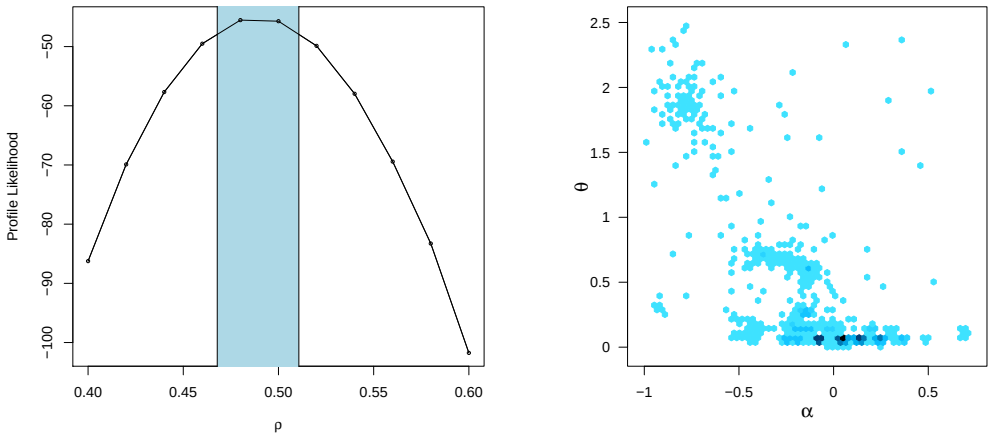

FIGURE 4. Profile Likelihood for the ρ Parameter and Heterogeneity Distribution $H(\alpha, \theta)$: In the left panel we plot the Kiefer-Wolfowitz profile likelihood as a function of ρ. The shaded region represents a 0.95 confidence interval for ρ based on a NPMLE version of the classical Wilks inversion 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.

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 variability is mistaken for AR persistence in innovations.

The right panel of Figure 4 plots the non-parametric estimate of the joint distribution 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 low draws for α are 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.

6.1. **Prediction of Income Trajectories.** We would like to adapt the univariate empirical Bayes rules for prediction 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 may appear quite different. 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 exercise as one facing a typical financial advisor. Similar problems present themselves in many biomedical

settings where diagnosis is based on reference growth charts or some other measures of the progression of disease.

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. These NPMLE posteriors are necessarily discrete, but we are entitled to sample from them for simulation purposes. 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, $y_{T_0+s} = \alpha + \hat{\rho} y_{T_0+s-1} + \sqrt{\theta} u_s$, for $s = 1, \cdots, T - T_0$, 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 may seem rather heretical. We began with a perfectly legitimate likelihood formulation: data was assumed to be generated from a very conventional Gaussian model, except that individuals had idiosyncratic (α, θ) parameters whose joint distribution, H, could be viewed as a prior. If this H were delivered on a silver platter by some local oracle we would be justified in proceeding 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. However, lacking a local oracle, we have relied instead on the NPMLE and our sample from PSID data 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 structural, i.e. "homogeneous" parameters. Admittedly, we have "sinned" – we've peeked at the data when we shouldn't have peeked, but our peeking has revealed a much more plausible H than we might have otherwise been 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 two pairs of individuals to illustrate the variety of earnings predictions generated by our model. In Figure 5 we contrast predictions for an individual with relatively large mean, i.e. 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. Realized trajectories are depicted by the solid black lines. For the high mean individual in the left panel of Figure 5, the bands are relatively narrow reflecting the fact that his "posterior" assigns little mass to high θ's. In contrast, for the lower mean individual in the right

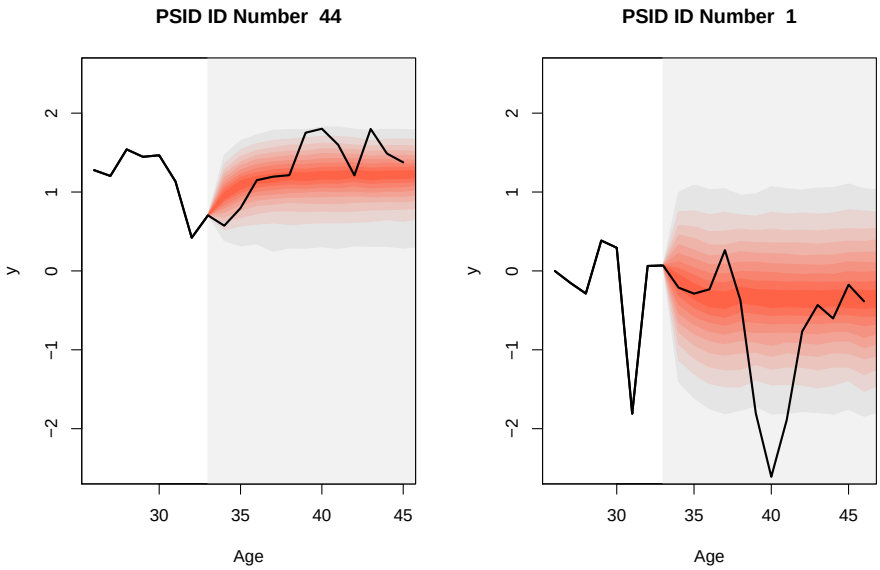

FIGURE 5. Fan Plot of Earnings Forecasts for Two Individuals: Based on the initial 9 years earnings, pointwise prediction bands are shown with graduated shading indicating bands from the 0.05 to 0.95 quantiles with the actual realizations superimposed as the black lines.

panel 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 90% uniform band (not shown) for this individual just barely covers this excursion.

In Figure 6 we contrast high mean, low variance individual with low mean, high variance one. The prediction band is very narrow for the former individual in the left panel, and much wider for the latter in the right panel. 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 toward 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.

6.2. **Inequality and the Distribution of Annual Income Increments.** Inequality of incomes, wealth and other indicators of social welfare have taken on an increasing salience as documented in the ongoing Deaton Review, Deaton (2018–). Using a 10% sample of U.S. Social Security records Guvenen et al (2022) have shown that the distribution of annual increments in log earnings has Pareto tail behavior. The left panel of Figure 7 reproduces a log-density plot appearing as their Figure 6 showing

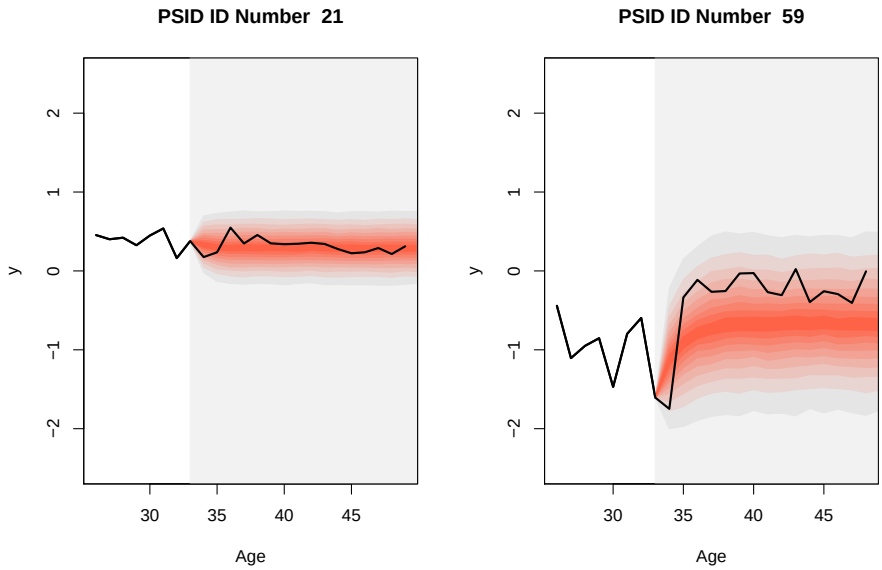

FIGURE 6. Fan Plot of Earnings Forecasts for Two Individuals: Based on the initial 9 years earnings, pointwise prediction bands are shown with graduated shading indicating bands from the 0.05 to 0.95 quantiles.

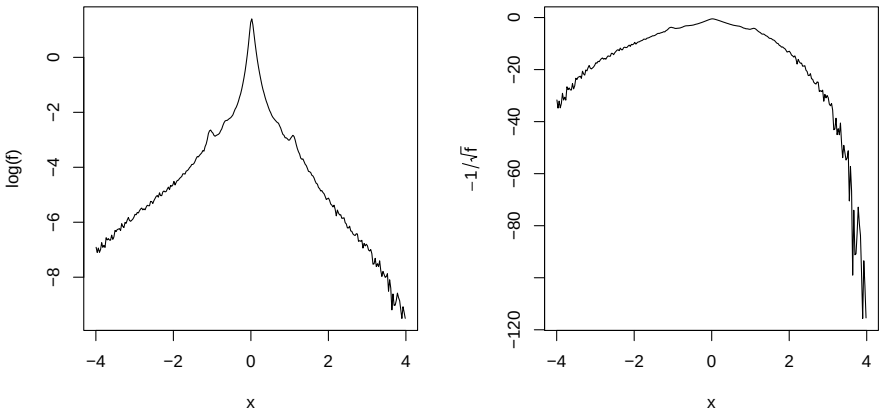

FIGURE 7. Guvenen et al plots of annual increments of log earnings: The left panel shows the log density plot reproduced from Figure 6 of Guvenen et al (2022), the right panel plots $-1/\sqrt{f(x)}$ yielding a much nicer concave shape.

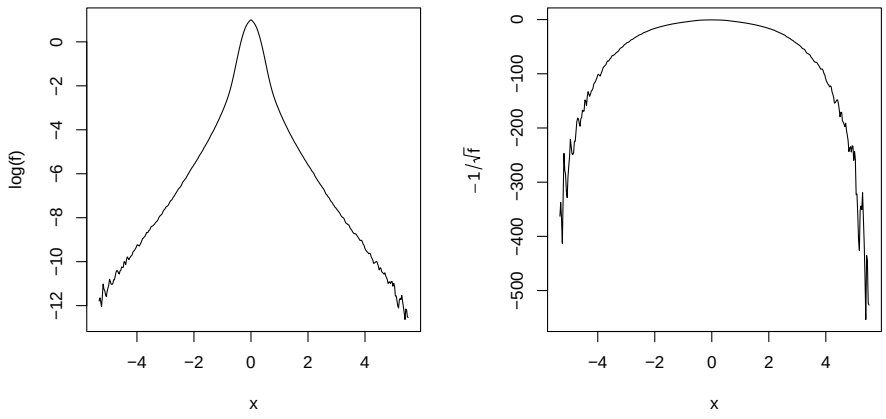

FIGURE 8. Marginal density of annual increments of log earnings based on simulated data from the estimated mixture model and PSID data. Ths simulation is based on 2500 sample paths for each of the 938 PSID sampled. As in the previous figure, the left panel is the log transformed density and the right panel is the $-1/\sqrt{f(x)}$ transformed density.

characteristic linear (Pareto) tail behavior with a tail exponent of about 0.40 in the left tail and 1.18 in the right tail. Although such densities fall outside the familiar class of log-concaves, they are nicely accommodated by the larger class of s-concave densities described in Koenker and Mizera (2018). The right panel of Figure 7 depicts the same density, now plotted not as as $\log f(x)$ but as $-1/\sqrt{f(x)}$ revealing a nice concave shape. Such densities can be easily estimated by shape constrained non-parametric methods as described in Koenker and Mizera (2010, 2018) and Han and Wellner (2016) and are implemented in the function `medde` in the R package REBayes.

Given our estimates of the bivariate mixture model, it is of interest to see whether the estimated model can generate a similar marginal density for annual increments in log earnings. To investigate this we generate 2500, $m = 50$, $M = 50$, sample paths for each of the 938 PSID sampled individuals using their individual specific posterior distributions $\hat{H}_i$, and the profile likelihood point estimate of ρ. These sample paths in log levels are then transformed to annual increments and a marginal density for these increments is then estimated. The resulting log and Hellinger transformed densities are shown in Figure 8. Not only are the shapes of the transformed densities remarkably similar to those in the Guvenen figure, the support of the estimated density is also remarkably consistent. It may seem surprising that our relatively small sample of 938 individuals from the PSID can create enough dispersion to generate

this extreme tail behavior, however further reflection suggests that the estimated scale heterogeneity of the model is capable of generating some rather wild trajectories.

6.3. **Heterogeneous ARMA Income Dynamics.** The simple AR(1) dynamics of the preceding models is especially convenient since partial differencing yields sufficient statistics that make the likelihood easily computed. However, there is a long tradition going back to Friedman (1957) of considering more complex dynamics that decompose the income process into transitory and permanent components. To illustrate how such models can be accommodated within the empirical Bayes framework, we will consider the simple ARMA(1,1) specification adopted by Blundell (2014):

$$\begin{aligned} y_{it} &= \mu_i + u_{it} + v_{it} \\ u_{it} &= \rho u_{i,t-1} + \sigma_\nu \nu_{it} \\ v_{it} &= \sigma_\eta \eta_{it} + \sigma_\eta \theta \eta_{i,t-1}. \end{aligned}$$

As before, y_{it} denotes the residual log income process after removing covariates effects. For simplicity, we provisionally assume homogeneous scale parameters, σ_ν and σ_η for the AR and MA components, respectively. The innovations, $\nu_{it} \sim N(0,1)$ and $\eta_{it} \sim N(0,1)$ are taken as iid and independent of one another. Substituting, we have the model,

$$y_{it} = \rho y_{i,t-1} + (1-\rho)\mu_i + \sigma_\nu \nu_{it} + \sigma_\eta \eta_{it} + (\theta - \rho)\sigma_\eta \eta_{i,t-1} - \rho\theta\sigma_\eta \eta_{i,t-2} \tag{1}$$

In state-space form the model can be expressed, following Harvey (1990) as,

$$\begin{aligned} y_{it} &= c_i + Z\alpha_{it} + G\xi_{it} \\ \alpha_{it} &= d_i + T\alpha_{i,t-1} + H\epsilon_{it} \end{aligned}$$

where $\alpha_{it} \in \mathbb{R}^m$, $T \in \mathbb{R}^{m\times m}$, $\xi_{it} \in \mathbb{R}$, $\epsilon_{it} \in \mathbb{R}^m$, $G \in \mathbb{R}$ and $Z \in \mathbb{R}^m$, $H \in \mathbb{R}^{m\times m}$. The dimension of α_{it} is 3 in our case, $d_i = G = 0$, $c_i = \mu_i$, $Z = (1,0,0)$ and

$$\begin{pmatrix} \alpha_{1it} \\ \alpha_{2it} \\ \alpha_{3it} \end{pmatrix} = \begin{pmatrix} \rho & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_{1i,t-1} \\ \alpha_{2i,t-1} \\ \alpha_{3i,t-1} \end{pmatrix} + \begin{pmatrix} 1 & 1 & 0 \\ \theta-\rho & 0 & 0 \\ -\theta\rho & 0 & 0 \end{pmatrix} \begin{pmatrix} \sigma_\eta & 0 & 0 \\ 0 & \sigma_\nu & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{it} \\ \nu_{it} \\ \epsilon_{3it} \end{pmatrix}.$$

More explicitly, we have,

$$\begin{aligned} \alpha_{1it} &= \rho\alpha_{1i,t-1} + \alpha_{2i,t-1} + \sigma_\eta \eta_{it} + \sigma_\nu \nu_{it} \\ \alpha_{2it} &= \alpha_{3i,t-1} + \sigma_\eta(\theta - \rho)\eta_{it} \\ \alpha_{3it} &= -\theta\rho\sigma_\eta \eta_{it}, \end{aligned}$$

and substituting the last two lines into the first and bring back to the first equation on y_{it}, we have,

$$y_{it} = (1-\rho)\mu_i + \rho y_{i,t-1} - \theta\rho\sigma_\eta \eta_{i,t-2} + \sigma_\eta(\theta-\rho)\eta_{i,t-1} + \sigma_\eta \eta_{it} + \sigma_\nu \nu_{it}$$

which is identical to (1).

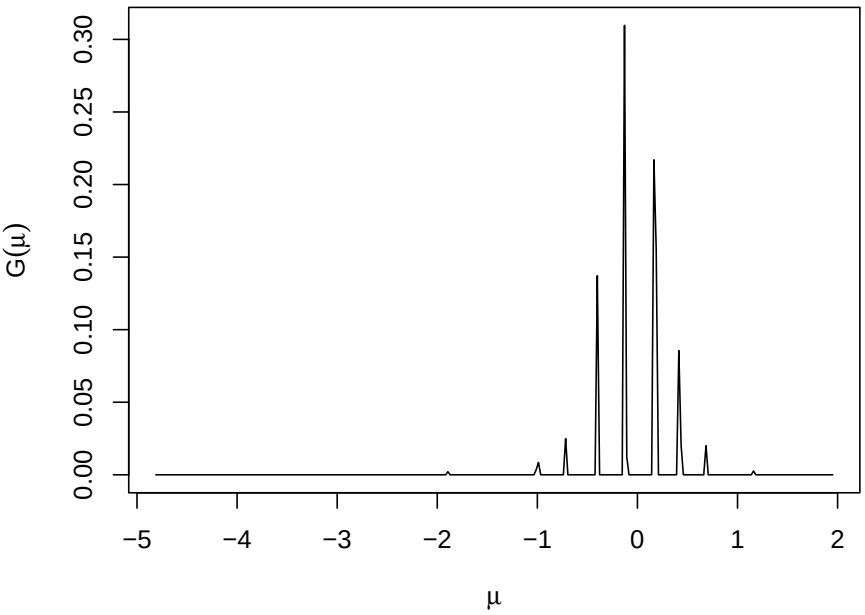

FIGURE 9. The estimated distribution of the individual fixed effect with the heterogeneous ARMA(1,1) model.

Our objective is now to estimate the latent distribution of the μ_i along with the structural parameters $(\rho, \theta, \sigma_\nu, \sigma_\eta)$. For a time series model like (1), the mixture likelihood of this model might appear intractable. However, given our Gaussian assumptions on the innovations, the likelihood – conditional on the structural parameters – for each trajectory $y_{i1}, \cdots y_{im_i}$ can be computed with the aid of the Kalman filter. Thus, with only μ_i (location) heterogeneity in the mixture model it is relatively straightforward to formulate the profile likelihood problem for the structural parameters; each entry in the NPMLE constraint matrix A is supplied by the Kalman filter, which recursively builds the likelihood evaluation for each parameter setting (see Harvey (1990) Section 3.4). It should perhaps be stressed that estimation of the mixing distribution for the μ_i, given the structural parameters is a convex optimization problem with a unique, quite parsimonious discrete solution. Profile likelihood for the remaining structural parameters is also straightforward once likelihood evaluations for the mixture problem are in place. In principle, there is no obstruction to reinstating scale heterogeneity into the model, however it seemed prudent to initially consider only location heterogeneity.

Preliminary estimation of this model employing a relatively coarse grid for profiling the structural parameters yields, $(\hat{\rho}, \hat{\theta}, \hat{\sigma}_\nu, \hat{\sigma}_\eta) = (0.49, 0.15, 0.17, 0.5)$. The corresponding estimated mixing distribution for the location parameters, μ_i is shown in Figure 9. It was surprising to us that the persistence of the income process in this version of the model was so weak. In Gu and Koenker (2017b) we asserted that the weak, $\rho = 0.48$, AR(1) persistence could be attributed to the inclusion of heterogeneous scale in the model. However, even with fixed scale, we find similar weak persistence in the ARMA(1,1) specification implying that the reliance on unit-root specifications of income processes may be questionable.

We should stress, however, that we still find the heterogeneous scale specification attractive because it enables one to make more reliable assessments of confidence bands for posterior mean predictions. In Appendix B we compare predictive fanplots for several representative individuals in our PSID sample. In the panels on the left side we have the predictions from the ARMA(1,1) model without any scale heterogeneity while in the right panels we have the predictions from the AR(1) model with both location and scale heterogeneity. Not unexpectedly, the ARMA(1,1) model prediction bands have the same width for all subjects, thereby over-covering for individuals with low variability in the initial period, and under-covering for those with high variability in the initial period. A secondary consequence of the scale homogeneity of the ARMA(1,1) model is that it fails to capture the extreme tail behavior illustrated in Figure 8 for the AR(1) model.

7. Conclusion

A survey of some recent developments in empirical Bayes methods focusing on nonparametric maximum likelihood estimation of mixture models for latent variables has been presented. A more extensive development will eventually be available in Koenker and Gu (2024). We believe that these methods offer valuable new tools for studying heterogeneity in its manifold forms in economics and we look forward to future developments.

APPENDIX A. TWEEDIE'S FORMULA

Robbins (1956) attributes Proposition 2 to Tweedie (1947). It follows by straightforward exponential family computations, as in van Houwelingen and Stijnen (1983),

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

Differentiating again,

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

implies that δ must be monotone.

Stein in his discussion of Efron and Morris (1973) observed that in the standard Gaussian case, $Y \sim \mathcal{N}(\theta, I_n)$ the oracle decision rule, $\delta(Y) = Y + \nabla \log f(Y)$, under quadratic loss has compound risk,

$$\begin{aligned}\mathbb{E}\|Y + \nabla\log f(Y) - \theta\|^2 &= \mathbb{E}\|Y-\theta\|^2 + \mathbb{E}\|\nabla\log f(Y)\|^2 + 2\mathbb{E}(Y-\theta)^\top \nabla \log f(Y) \\ &= n + \mathbb{E}\|\nabla\log f(Y)\|^2 + 2\mathbb{E}\left\{\frac{1}{f(Y)}\nabla^2 f(Y) - \|\nabla\log f(Y)\|^2\right\} \\ &= n - \mathbb{E}\left\{\|\nabla\log f(Y)\|^2 - \frac{2}{f(Y)}\nabla^2 f(Y)\right\} \\ &= n + 4\mathbb{E}\left\{\frac{\nabla^2\sqrt{f(Y)}}{\sqrt{f(Y)}}\right\}\end{aligned}$$

where ∇ is the vector of first partial derivatives, and ∇^2 is the Laplacian, $\sum \partial^2/\partial y_i^2$. The second equality follows from Stein's lemma, and the fourth from the identity,

$$\nabla^2\sqrt{f} = \nabla\cdot\nabla\sqrt{f} = \nabla\cdot\frac{\nabla f}{2\sqrt{f}} = \frac{1}{2\sqrt{f}}\nabla^2 f - \frac{1}{4f^{3/2}}\|\nabla f\|^2.$$

Note that the expression in the displayed equation provides an unbiased estimate of compound risk. Stein concludes that if $\sqrt{f}$ is superharmonic, that is, $\nabla^2\sqrt{f} \leq 0$, then the Tweedie oracle estimator, $\delta(Y) = Y + \nabla \log f(Y)$, is minimax.

Of course, practical implementation of such decision rules requires an estimator for f_G. To evaluate the cost of using an estimated decision rule, δ, instead of the Tweedie oracle rule, δ^*_G, we define regret as the difference in their risks,

$$\mathcal{R}_n(\delta, \mathcal{G}) = \sup_{G \in \mathcal{G}} \{R_n(\delta, G) - R_n(\delta^*_G, G)\}.$$

Regret depends upon the class, $\mathcal{G}$, of possible mixing distributions. Light tailed G, that is those with bounded support or sub-Gaussian tails, denoted $\mathcal{G}_\infty$, make it relatively easy to estimate the marginal density. Heavier tailed G, satisfying the moment condition, $\mathcal{G}_p = \{G : \int |u|^p dG_n(u) = \mathcal{O}(1)\}$, make it more difficult. For estimators based on the NPMLE, $f_{\hat{G}}$, in the Gaussian mixture setting, Jiang and Zhang (2009) have shown that,

$$\mathcal{R}_n(\delta_{\hat{G}_n}, \mathcal{G}) \lesssim \begin{cases} n^{-1}(\log n)^5 & \text{if } \mathcal{G} = \mathcal{G}_\infty. \\ n^{-p/(1+p)}(\log n)^{\frac{8+9p}{2+2p}} & \text{if } \mathcal{G} = \mathcal{G}_p \text{ for some fixed p,} \end{cases}$$

where $a_n \lesssim b_n$ denotes $a_n = \mathcal{O}(b_n)$. Theorem 1 of Polyanskiy and Wu (2021) establishes that these regret bounds are minimax rate optimal up to logarithmic factors. Thus, as long as G is light tailed, posterior mean rules based on the NPMLE and Tweedie's formula achieve essentially a parametric rate of convergence up to the log factor.

Appendix B. Predictive Distribution Comparison

In Figures 10 and 11 we compare predictive distributions for the scale homogeneous ARMA(1,1) model and the scale heterogeneous AR(1) model. It can be noted that the width of the ARMA(1,1) prediction bands are the same for both highly variable and very stable individuals in the pre-forecast period, while the AR(1) model that incorporates individual specific scale effects adapts to this difference.

References

Arellano M (2003) Panel Data Econometrics. Oxford U. Press

Arellano M, Blundell R, Bonhomme S (2017) Earnings and consumption dynamics: A nonlinear panel data framework. Econometrica 85:693–734

Azevedo EM, Deng A, Montiel Olea JL, Rao J, Weyl EG (2020) A/B testing with fat tails. Journal of Political Economy 128:4614–4672

Balestra P, Nerlove M (1966) Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas. Econometrica 34:585–612

Beckett L, Diaconis P (1994) Spectral analysis for discrete longitudinal data. Advances in Mathematics 103:107–128

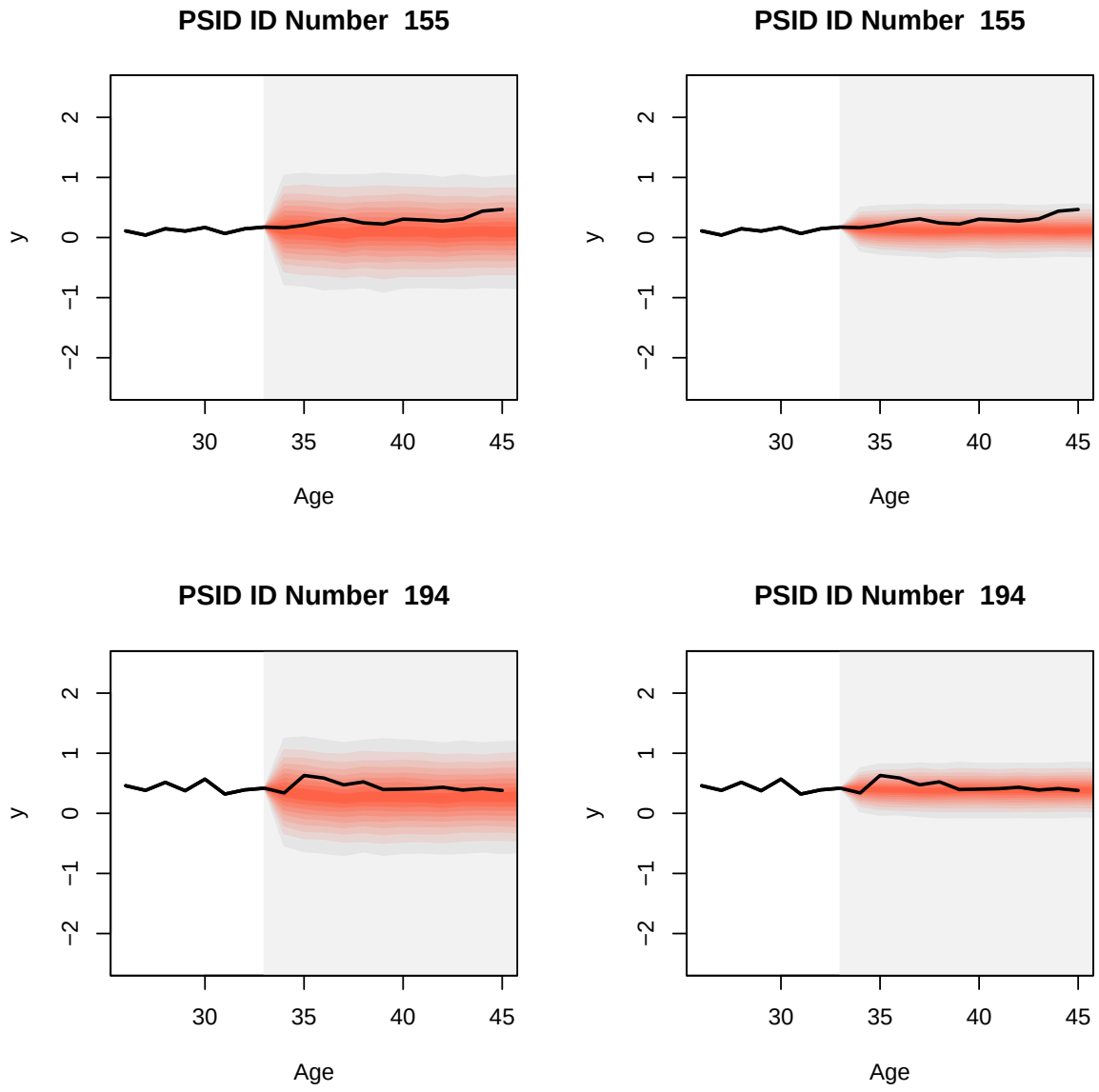

FIGURE 10. Left panels depict predictive bands for the ARMA(1,1) model, while right panels depict bands for the AR(1) heterogeneous scale model.

Blundell R (2014) Income dynamics and life-cycle inequality: Mechanisms and controversies. The Economic Journal 124:289–318

Browning M, Ejrnæs M, Alvarez J (2010) Modelling income processes with lots of heterogeneity. Review of Economic Studies 77:1353–1381

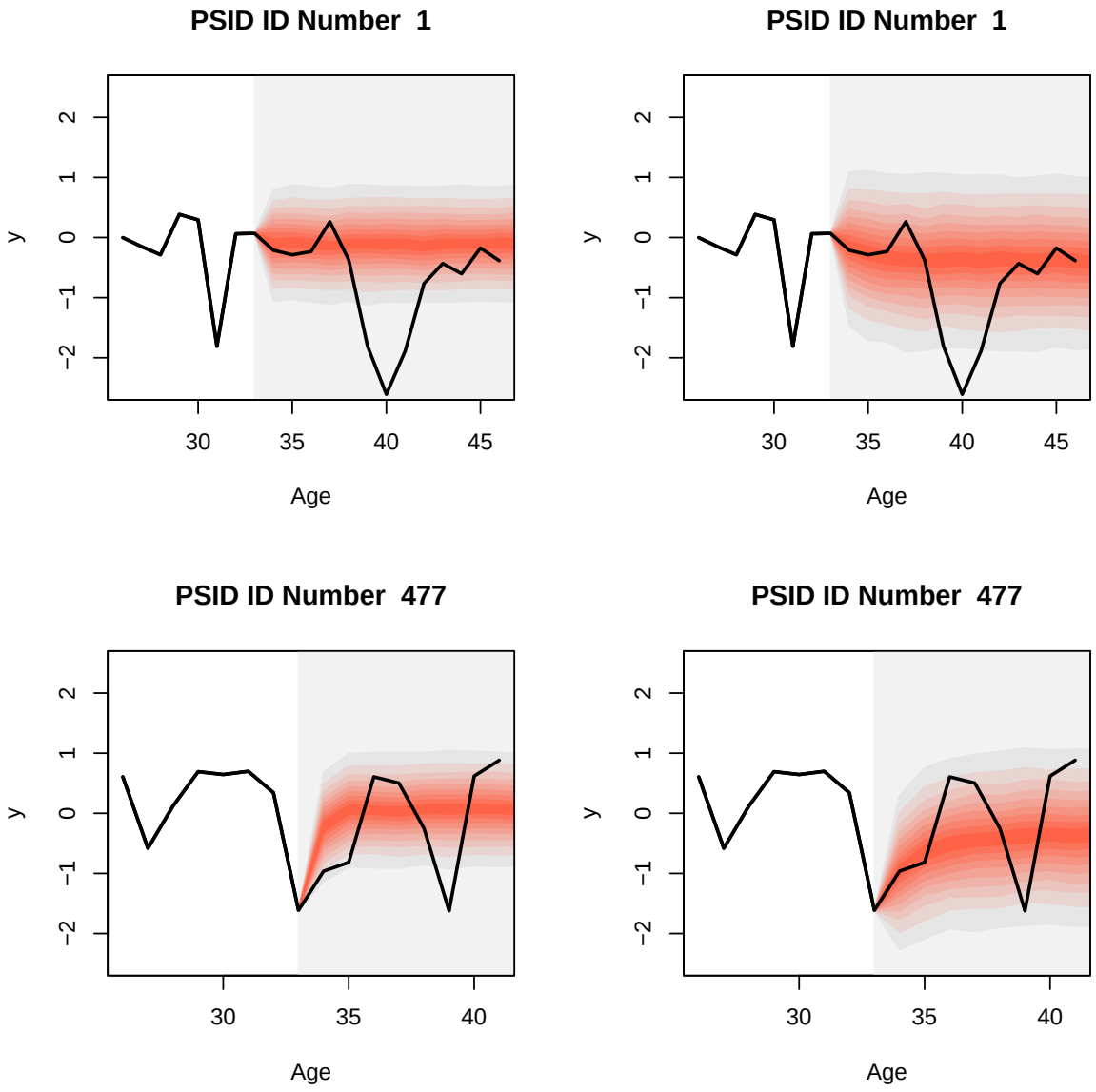

FIGURE 11. Left panels depict predictive bands for the ARMA(1,1) model, while right panels depict bands for the AR(1) heterogeneous scale model.

Bühlmann H, Gisler A (2005) A Course in Credibility Theory and its Applications. Springer-Verlag

Castillo I, van der Vaart A (2012) Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. Annals of Statistics 40:2069–2101

Chamberlain G, Hirano K (1999) Predictive distribution based on longitudinal earnings data. Annales d' Économie et de Statistique 55/56:211–242

Chamberlain G, Leamer EE (1976) Matrix weighted averages and posterior bounds. Journal of the Royal Statistical Society B 38:73–84

Chen X, Liao Z (2014) Sieve m inference on irregular parameters. Journal of Econometrics 182:70–86

Cox DR (1975) A note on partially Bayes inference and the linear model. Biometrika 62:651–654

Deaton A (2018–) Deaton review of inequalities, https://ifs.org.uk/inequality/

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society B 39:1–38

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

Efron B (2014) Two Modeling Strategies for Empirical Bayes Estimation. Statistical Science 29:285 – 301

Efron B (2016) Empirical Bayes deconvolution estimates. Biometrika 103:1–20

Efron B, Morris C (1973) Combining possibly related estimation problems. Journal of the Royal Statistical Society B 35:379–421

Fan J, Zhang C, Zhang J (2001) Generalized likelihood ratio statistics and Wilks phenomenon. Annals of Statistics 29:153–193

de Finetti B (1964) Foresight: Its logical laws and subjective sources. In: Kyburg HE, Smokler H (eds) Studies in Subjective Probability, Wiley, pp 93–158

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology 12:42 –58

Friedman M (1957) Theory of the Consumption Function. Princeton University Press

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

Greenshtein E, Ritov Y (2019) Comment: Empirical Bayes, compound decisions and exchangeability. Statistical Science 34:224–228

Gu J, Koenker R (2017a) Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data. Journal of Applied Econometrics 32:575–599

Gu J, Koenker R (2017b) Unobserved heterogeneity in income dynamics: An empirical Bayes perspective. Journal of Business and Economic Statistics 35:1–16

Gu J, Koenker R (2022) Ranking and selection from pairwise comparisons: Empirical bayes methods for citation analysis. American Economic Association Papers and Proceedings 112:624–629

Gu J, Koenker R (2023) Invidious comparisons: Ranking and selection as compound decisions. Econometrica 91:1–41

Guvenen F, Karahan F, Ozkan S, Son J (2022) What do data on millions of U.S. workers reveal about lifecycle earnings dynamics? Econometrica 89:2303–2339

Han Q, Wellner JA (2016) Approximation and estimation of s-concave densities via Rényi divergences. Annals of Statistics 44:1332 – 1359

Harvey AC (1990) Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press

Heckman J, Singer B (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

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

van Houwelingen J, Stijnen T (1983) Monotone Empirical Bayes Estimators for the Continuous One-parameter Exponential Family. Statistica Neerlandica 37:29–43

Ignatiadis N, Sen B (2023) Empirical partially Bayes multiple testing and compound χ^2 decisions. Available from: https://arxiv.org/abs/2303.02887

James W, Stein C (1961) Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, p 361

Jiang W, Zhang CH (2009) General maximum likelihood empirical Bayes estimation of normal means. Annals of Statistics 37:1647–1684

Johnstone I, Silverman B (2004) Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. Annals of Statistics 32:1594–1649

Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics 27:887–906

Kline P, Rose EK, Walters CR (2024) A discrimination report card. American Economic Review 114:2472–2525

Kline PM, Walters CR (2021) Reasonable doubt: Experimental detection of job-level employment discrimination. Econometrica 89:765–792

Koenker R, Gu J (2024) Empirical Bayes: Some Tools, Rules and Duals. Cambridge University Press

Koenker R, Mizera I (2010) Quasi-concave density estimation. Annals of Statistics 38(5):2998–3027

Koenker R, Mizera I (2014) Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. Journal of the American Statistical Association 109:674–685

Koenker R, Mizera I (2018) Shape Constrained Density Estimation Via Penalized Rényi Divergence. Statistical Science 33:510 – 526

Laird N (1978) Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association 73:805–811

Leamer EE, Chamberlain G (1976) A Bayesian interpretation of pretesting. Journal of the Royal Statistical Society B 38:85–94

Lindley DV, Smith AF (1972) Bayes estimates for the linear model. Journal of the Royal Statistical Society B pp 1–41

Lindsay B (1981) Properties of the maximum likelihood estimator of a mixing distribution. In: Taillie C, Patil GP, Baldessari BA (eds) Statistical Distributions in Scientific Work, D. Reidel, pp 95–109

Lindsay B (1995) Mixture models: theory, geometry and applications. In: NSF-CBMS regional conference series in probability and statistics

Liu J (1996) Nonparametric hierarchical Bayes via sequential imputations. Annals of Statistics 24(3):911–930

Meghir C, Pistaferri L (2004) Income variance dynamics and heterogeneity. Econometrica 72:1–32

Mundlak Y (1978) On the pooling of time series and cross section data. Econometrica 46:69–85

Murphy SA, van der Vaart AW (2000) On profile likelihood. Journal of the American Statistical Association 95:449–465

Pfanzagl J (1988) Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. Journal of Statistical Planning and Inference 19:137–158

Polyanskiy Y, Wu Y (2020) Self-regularizing property of nonparametric maximum likelihood estimator in mixture models, available from https://arxiv.org/abs/2008.08244

Polyanskiy Y, Wu Y (2021) Sharp regret bounds for empirical Bayes and compound decision problems, available from https://arxiv.org/abs/2109.03943

Reid N (1995) The Roles of Conditioning in Inference. Statistical Science 10:138 – 157

Robbins H (1950) A generalization of the method of maximum likelihood: Estimating a mixing distribution (abstract). Annals of Mathematical Statistics 21:314–315

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

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

Savage LJ (1954) Foundations of Statistics. Wiley

Soloff JA, Guntuboyina A, Sen B (2021) Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood, available from https://arxiv.org/abs/2109.03466

Stein C (1956) Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, University of California Press: Berkeley, vol I, pp 197–206

Stigler SM (1990) A Galtonian Perspective on Shrinkage Estimators. Statistical Science 5:147 – 155

Teicher H (1961) Identifiability of Mixtures. Annals of Mathematical Statistics 32:244–248

Teicher H (1967) Identifiability of mixtures of product measures. Annals of Mathematical Statistics 38:1300–1302

Tweedie MCK (1947) Functions of a statistical variate with given means, with special reference to Laplacian distributions. Mathematical Proceedings of the Cambridge Philosophical Society 43:41–49

Wald A (1950) Statistical Decision Functions. Wiley

Zhang Y, Cui Y, Sen B, Toh KC (2022) On efficient and scalable computation of the nonparametric maximum likelihood estimator in mixture models. Available from: https://arxiv.org/abs/2208.07514