EMPIRICAL BAYESBALL REMIXED: EMPIRICAL BAYES METHODS FOR LONGITUDINAL DATA

JIAYING GU AND ROGER KOENKER

Abstract. Empirical Bayes methods for Gaussian and binomial compound decision problems involving longitudinal data are considered. A recent convex optimization reformulation of the nonparametric maximum likelihood estimator of Kiefer and Wolfowitz (1956) is employed to construct nonparametric Bayes rules for compound decisions. The methods are illustrated with an application to predicting baseball batting averages, and the age profile of batting performance. An important aspect of the empirical application is the general bivariate specification of the distribution of heterogeneous location and scale effects for players that exhibits a weak positive association between location and scale attributes. Prediction of players' batting averages for 2012 based on performance in the prior decade using the proposed methods shows substantially improved performance over more naive methods with more restrictive treatment of unobserved heterogeneity. Comparisons are also made with nonparametric Bayesian methods based on Dirichlet process priors, which can be viewed as a regularized, or smoothed, version of the Kiefer-Wolfowitz method.

1. Introduction

Unobserved heterogeneity has become a pervasive concern throughout applied econometrics, and there has been a resurgence of interest in empirical Bayes methods for estimating hierarchical models with random parameters. Much of this literature has focused on the parametric Gaussian random effects model developed by Lindley and Smith (1972). One prominent source of such applications is the literature on teacher evaluation. Guarino, Maxfield, Reckase, Thompson, and Wooldridge (2015) have recently argued that empirical Bayes methods may be misguided when teacher assignment is closely tied to student performance, but as expected they show that these methods perform well under random assignment. Prediction of insurance liability claims also relies heavily on the linear shrinkage rules arising from the Gaussian random effects paradigm as demonstrated in Bühlmann and Gisler (2005) and the extensive related literature in actuarial science. Less attention has been paid to nonparametric mixture models in econometric applications with the notable exception of the seminal paper of Heckman and Singer (1984), who advocated use of the Kiefer and Wolfowitz (1956) nonparametric maximum likelihood estimator in a Weibull mixture model of unemployment durations. In this paper we will describe some new, Kiefer-Wolfowitz based, nonparametric empirical Bayes methods for estimation and prediction in longitudinal data models with unobserved heterogeneity, and compare and contrast them with some existing nonparametric Bayesian proposals.

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.

Key words and phrases. Empirical Bayes, mixture models, nonparametric maximum likelihood.

Version: April 6, 2016. This research was partially supported by NSF grant SES-11-53548. All computational results of this paper can be reproduced with the R packages REBayes, Koenker and Gu (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 the referees for extremely constructive comments on earlier versions. We would also like to express our appreciation to participants in the CeMMaP-UCL, Harvard-MIT, Georgetown, Rutgers, Cornell and University of Maryland seminars for valuable comments while retaining full responsibility for any other errors of omission and commission.

Although they predate the development of hierarchical Bayes methods exemplified in the work of Lindley and Smith (1972), 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 longitudinal data modeling and its treatment of unobserved heterogeneity.

We will begin with a brief review of some developments in empirical Bayes methods beginning with Robbins (1951), touching on the connections to Stein rule methods and finally evolving into modern nonparametric mixture variants. In Section 3 we extend the predominant Gaussian location mixture framework to accommodate nonparametric location *and scale* mixtures in the classical Gaussian longitudinal data setting. Section 4 describes an extended application on baseball batting averages that illustrates both estimation and prediction aspects of the new methods including, notably, the introduction of covariate effects via profile likelihood methods. This paper complements parallel work on related methods for models of income dynamics Gu and Koenker (2015), expanding it to models for discrete (binomial) data and providing an explicit comparison with more formal nonparametric Bayesian methods based on Dirichlet process priors.

In sharp contrast to the classical Gaussian hierarchical Bayes framework for longitudinal data with parametric mixing distributions, or its frequentist counterparts, the nonparametric mixture formulation of our proposed methods offers a more flexible view of unobserved heterogeneity while preserving most of the virtues of the Bayesian formalism. In particular, more flexible nonparametric modeling of unobserved heterogeneity leads to improved predictive performance.

2. Empirical Bayes: A Brief Review

Given a simple parametric statistical model, there is a natural Bayesian temptation to complicate it by building a hierarchical structure on top of it. As surveyed by Good (1979) one of the earliest examples of this type was the (classified) work of Turing in 1941, elaborated in Good (1953). Another prominent strand of this literature was the Gaussian random effects compound decision problem introduced by Robbins (1951). In Robbins's setting we observe independent $Y_1, \cdots, Y_n$ each Gaussian with common variance, σ^2 but individual specific means, $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$. Our objective is to estimate all the μ_i's subject to squared error loss,

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

The naive (unbiased) solution would simply set $\hat{\mu}_i = Y_i$, but a natural presumption in such circumstances would be that the observations have some common genesis, and consequently that 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 μ_i were drawn iid-ly from the distribution, F, so the Y_i's would have convolution density $g(y) = \int \phi((y-\mu)/\sigma)/\sigma dF(\mu)$. Then the Bayes rule would take the form

$$\delta(y) = y + \sigma^2 g'(y)/g(y) \tag{1}$$

Efron (2011) refers to (1) as Tweedie's formula citing Robbins's (1956) attribution of it to M.C.K. Tweedie. Of course one might 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 familiar special cases.

2.1. **Some Parametric Examples.**

(1) Suppose $\sigma^2 = 1$ and we believed that the μ_i's were iid $\mathcal{N}(0, \sigma_0^2)$, so the Y_i's are iid $\mathcal{N}(0, 1+\sigma_0^2)$, the Bayes rule would be,

$$\delta(y) = \left(1 - \frac{1}{1+\sigma_0^2}\right) y.$$

Thus, we shrink our naive estimates all toward zero. When σ_0^2 is unknown, $S = \sum Y_i^2 \sim (1+\sigma_0^2)\chi_n^2$, and recalling that an inverse χ_n^2 random variable has expectation, $(n-2)^{-1}$, we obtain the Stein rule in its simplest form:

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

(2) When, slightly more generally, $\mu_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$ we shrink instead toward the prior mean,

$$\delta(y) = \mu_0 + \left(1 - \frac{1}{1+\sigma_0^2}\right)(y - \mu_0),$$

and estimating the prior parameters costs us one degree of freedom, so we obtain the celebrated James-Stein estimator,

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

for $\bar{Y}_n = n^{-1}\sum Y_i$ and $S = \sum (Y_i - \bar{Y}_n)^2$.

(3) If each observation has its own *known* variance: $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and $\mu_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$, as might be plausible in the case that each Y_i is from a different measuring device each with known precision, or as in the ubiquitous baseball batting average examples, as in Brown (2008) and Jiang and Zhang (2010), in which binomial variances depend upon a known number of "at bats" in the initial period. In such cases we have the Bayes rule,

$$\delta(y_i) = \mu_0 + \left(1 - \frac{\sigma_i^2}{\sigma_0^2 + \sigma_i^2}\right)(y_i - \mu_0)$$

(4) Further generalizing, we may wish to replace μ_0 by a function of observable covariates, say $z_i^\top \beta_0$. Then, as in Jiang and Zhang (2010), we obtain a positive-part James-Stein estimator,

$$\hat{\delta}(y_i) = \left(1 - \frac{p-2}{\sum (z_i^\top \hat{\beta}/\sigma_i)^2}\right)_+ z_i^\top \hat{\beta} + \left(1 - \frac{n-p-2}{\sum (y_i - z_i^\top \hat{\beta})^2/\sigma_i^2}\right)_+ (y_i - z_i^\top \hat{\beta})$$

where p denotes the dimension of β and $(u)_+ = uI(u > 0)$.

(5) Another important class of examples arises from the assumption of sparsity, that is, an assertion that most of the μ_i are probably zero. Johnstone and Silverman (2004) consider a model in which,

$$dF(\mu) = (1-w)\delta_0(\mu) + w\varphi_\nu(\mu)$$

where with probability $(1-w)$, $\mu = 0$, while with probability w it is drawn from a density, φ, with scale parameter, ν. They compare performance of several hard and soft threshholding rules in addition to empirical Bayes procedures that estimate the parameters w and ν. This is closely related to an extensive recent literature on more formal Bayesian methods for the Gaussian sequence model, e.g. Castillo and van der Vaart (2012).

The simulation designs of Johnstone and Silverman (2004) have served as a benchmark for several more recent studies of empirical Bayes methods including Brown and Greenshtein (2009), Jiang and Zhang (2009), and Koenker and Mizera (2014), all of which explore non-parametric estimation of the Gaussian location mixture model.

2.2. **Non-parametric Estimation of the Gaussian Mixture Model.** Lacking confidence in any particular parametric specification that would allow us to estimate g parametrically, we are apparently led into the quagmire of Gaussian deconvolution. Before mobilizing any heavy empirical characteristic function artillery it is worth considering what might be accomplished with more conventional statistical machinery. Noting that Tweedie's formula requires only knowledge of the marginal density, g, of the observed Y_i's, Brown and Greenshtein (2009) propose simply estimating g by conventional kernel methods, thereby circumventing entirely the problem of estimating the mixing distribution, F. As they point out, however, a potential drawback of the kernel approach is that it fails to impose the constraint, implied by the Gaussian noise assumption, that the Bayes rule, $\delta(y)$, is monotone in y. Koenker and Mizera (2014) describe a maximum likelihood method of estimating g subject to this monotonicity constraint, or equivalently a convexity constraint on $K(y) = \frac{1}{2}y^2 + \log g(y)$. This approach improves predictive performance somewhat relative to the unconstrained kernel estimator, but it still fails to fully exploit the structures of the Gaussian compound decision model.

Jiang and Zhang (2009) propose a more direct attack on the Gaussian compound decision problem. Reviving the proposal of Kiefer and Wolfowitz (1956) for nonparametric maximum likelihood estimation of the general mixture model, they show that a fixed point variant of the EM algorithm suggested by Laird (1978) has excellent predictive performance. The only downside of their approach is that EM can be computationally extremely burdensome. Recently, Koenker and Mizera (2014) have suggested replacing the EM fixed point iteration,

$$\hat{f}_j^{(k+1)} = n^{-1} \sum_{i=1}^{n} \frac{\hat{f}_j^{(k)} \phi(y_i - u_j)}{\sum_{\ell=1}^{m} \hat{f}_\ell^{(k)} \phi(y_i - u_\ell)},$$

defined on a grid $\{u_1, u_2, \cdots, u_m\}$, by an interior point solution of the convex program:

$$\min\{-\sum_{i=1}^{n} \log(g_i) \mid Af = g,\ f \in S\},$$

where $A = (\phi(y_i - u_j))$ and $S = \{s \in \mathbb{R}^m | 1^\top s = 1,\ s \geqslant 0\}$. So f_j denotes the estimated mixing density estimate $\hat{f}$ at the grid point u_j, and g_i denotes the estimated mixture density estimate, $\hat{g}$, at Y_i. It is well-known from Laird (1978) and Lindsay (1995) that the Kiefer-Wolfowitz estimator produces a discrete $\hat{F}$, typically with only a few mass points. The fineness of the grid values controls the accuracy of the location of these mass points.

We have generally found a few hundred equally spaced grid points adequate, but further accuracy is always available by refinement of the grid. On relatively small test problems with sample size $n = 200$ and $m = 300$ equally spaced grid points, interior point methods achieve considerably more accurate solutions than EM, and require less than one second while EM requires 10 minutes. Our implementation of interior point methods is based on the Mosek implementation of Andersen (2010) as linked to R language R Core Team (2014) via the packages RMosek Friberg (2012) and REBayes Koenker and Gu (2015). Further details on the capabilities of the REBayes computing environment for empirical Bayes methods is provided in Koenker and Gu (2016), including models for survival and count data as well as those discussed here.

In the next section we will describe how these methods can be extended 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 structure of unobserved heterogeneity.

3. Estimating Gaussian Mixtures 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 straightforward. Assuming for convenience that we have unit variance for the noise so $u_{it} \sim \mathcal{N}(0,1)$, we have,

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

but sufficiency reduces the problem to our third example with $\hat{\mu}_i = m_i^{-1} \sum_{i=1}^{m_i} y_{it} \sim \mathcal{N}(\mu_i, m_i^{-1})$. When the μ_i's are iid from F, we can write the log likelihood of the observed y_{it}'s as,

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

Optimizing over the 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 discrete measure on a relatively fine grid containing the empirical support of the observed $\hat{\mu}_i$'s. 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 really 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 $A = (A_{ij} = (\sqrt{m_i}\phi(\sqrt{m_i}(\hat{\mu}_i - \mu)))$ and Δ_j is the jth grid spacing. As posed, the problem is evidently convex, having a strictly convex objective function subject to linear equality and inequality constraints, 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 the most diligent researchers. But interior point methods make this discreteness easily apparent. Since the number of non-negligible $\hat{f}_i > 0$ 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. Again, larger n might justify a refinement of the grid at very modest increase in computational effort.

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

$$\max_v \{ \sum_{i=1}^{n} \log(v_i) | A^\top v \leqslant n 1_p, v \geqslant 0 \}, \tag{3}$$

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 $2r_i s_i/\theta_i$ with $r_i = m_i/2$ has the gamma distribution with shape parameter, r_i and scale

parameter θ_i/r_i, i.e.

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

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

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

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 emphasize here that the scope for modeling heterogeneity for the scale parameter would not be possible with cross sectional data since individuals are then only measured once. The model is now,

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

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

$$\hat{\mu}_i|(\mu_i,\theta_i) \sim \mathcal{N}(\mu_i, \theta_i/m_i)$$

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,

$$\ell(F_\mu, F_\theta|y) = \sum_{i=1}^n \log \int\int \gamma(s_i|r_i,\theta/r_i)\sqrt{m_i}\phi(\sqrt{m_i}(\hat{\mu}_i-\mu)/\sqrt{\theta})/\sqrt{\theta}dF_\mu(\mu)dF_\theta(\theta)$$

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 $\hat{F}_\mu$ 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_μ. This is highly convenient for estimation purposes, however it should be stressed that prediction restores the interdependence on both F_μ and F_θ. In this independent prior setting we can also iterate back and forth between the gamma component of the mixture likelihood, and the Gaussian component likelihood, we explore both of these computational methods in the application section below.

When the independence assumption is implausible, and this may be typical of many applications where there is some aspect of the problem that suggests that μ_i's and θ_i's are positively (or negatively) correlated, we can construct two dimensional grids. This makes the constraint matrix, A, somewhat larger, but raises no new issues in principle. We discuss this briefly in the next subsection which also describes how covariate effects can be introduced.

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

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

Given a β it is easy to see that,

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

so the sufficient statistic for μ_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 longitudinal data terminology, the sufficient statistic for μ_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,

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

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

Even with the independent prior assumption, $h(\mu, \theta) = h_\mu(\mu)h_\theta(\theta)$, the likelihood does not factor because the Gaussian piece depends on both μ_i and θ_i. However, the fact that S_i, hence the Gamma piece of the likelihood, does not depend on μ_i provides a convenient estimation strategy by using the Gamma mixture to estimate h_θ, and a Studentized version of the Gaussian mixture, $(\bar{y}_i - \bar{x}_i\beta - \mu_i)/\sqrt{S_i/m_i} \sim t_{m_i-1}$, for estimating h_μ. We will compare this estimation strategy with an iterative method that employs a Gaussian mixture form of the likelihood to obtain $\hat{F}_\mu$.

Given our initial solution $\hat{F}_\theta^{(0)}$ based only S_i, we can maximize the log likelihood,

$$\sum_{i=1}^{n}\log\left\{\int\int\gamma(s_i \mid r_i, \theta/r_i)\frac{1}{\sqrt{\theta/m_i}}\phi(\frac{\hat{\mu}_i - \mu}{\sqrt{\theta/m_i}})d\hat{F}_\theta^{(0)}(\theta)dF_\mu(\mu)\right\}$$

to obtain $\hat{F}_\mu^{(0)}$. Continuing the iteration by maximizing the log-likelihood

$$\sum_{i=1}^{n}\log\left\{\int\int\gamma(s_i \mid r_i, \theta/r_i)\frac{1}{\sqrt{\theta/m_i}}\phi(\frac{\hat{\mu}_i - \mu}{\sqrt{\theta/m_i}})d\hat{F}_\mu^{(0)}(\mu)dF_\theta(\theta)\right\}$$

we obtain $\hat{F}_\theta^{(1)}$ and then solve for $\hat{F}_\mu^{(1)}$ by maximizing the log-likelihood

$$\sum_{i=1}^{n}\log\left\{\int\int\gamma(s_i \mid r_i, \theta/r_i)\frac{1}{\sqrt{\theta/m_i}}\phi(\frac{\hat{\mu}_i - \mu}{\sqrt{\theta/m_i}})dF_\mu(\mu)d\hat{F}_\theta^{(1)}(\theta)\right\}.$$

Iteration continues until the likelihood fails to improve by more than a specified tolerance. Note that once we integrate out the hatted distribution in each of the two likelihood expressions we have our standard convex Kiefer-Wolfowitz problem with a strictly convex objective subject to linear constraints. The likelihood at each step is increasing, so convergence to a local maximum is guaranteed. To see this consider the first step: given the initial $\hat{F}_\theta^{(0)}$, and a fixed grid for μ, $\hat{F}_\mu^{(0)}$ is

the unique maximizer, so we have,

$$\sum_{i=1}^{n} \log \left\{ \int\int \gamma(s_i \mid r_i, \theta/r_i) \frac{1}{\sqrt{\theta/m_i}} \phi\left(\frac{\hat{\mu}_i - \mu}{\sqrt{\theta/m_i}}\right) d\hat{F}_\theta^{(0)}(\theta) dF_\mu(\mu) \right\}$$
$$\leqslant \sum_{i=1}^{n} \log \left\{ \int\int \gamma(s_i \mid r_i, \theta/r_i) \frac{1}{\sqrt{\theta/m_i}} \phi\left(\frac{\hat{\mu}_i - \mu}{\sqrt{\theta/m_i}}\right) d\hat{F}_\theta^{(0)}(\theta) d\hat{F}_\mu^{(0)}(\mu) \right\}.$$

In the next step solving for $\hat{F}_\theta^{(1)}$, using the same grid for θ as was used to solve for $\hat{F}_\theta^{(0)}$, we have for all F_θ on this specified grid,

$$\sum_{i=1}^{n} \log \left\{ \int\int \gamma(s_i \mid r_i, \theta/r_i) \frac{1}{\sqrt{\theta/m_i}} \phi\left(\frac{\hat{\mu}_i - \mu}{\sqrt{\theta/m_i}}\right) d\hat{F}_\mu^{(0)}(\mu) dF_\theta(\theta) \right\}$$
$$\leqslant \sum_{i=1}^{n} \log \left\{ \int\int \gamma(s_i \mid r_i, \theta/r_i) \frac{1}{\sqrt{\theta/m_i}} \phi\left(\frac{\hat{\mu}_i - \mu}{\sqrt{\theta/m_i}}\right) d\hat{F}_\mu^{(0)}(\mu) d\hat{F}_\theta^{(1)}(\theta) \right\},$$

so, in particular, this holds for $F_\theta = \hat{F}_\theta^{(0)}$, and the same holds for $\hat{F}_\mu^{(k)}$ and $\hat{F}_\theta^{(k)}$ for $k = 1, 2, \cdots$, as we continue the iteration. Despite the convenient bi-convexity of the problem, there is no guarantee that we obtain joint optimality from this iteration scheme. Ironically, if we relax the independence condition on the "prior" mixing distribution and estimate the general bivariate mixing distribution all these caveats vanish and we have an unambiguous convex optimization problem, albeit with a somewhat more elaborate gridding strategy.

Including covariates adapts these estimation strategies: Given a β we estimate the mixing distribution and then evaluate the full profile likelihood. We will illustrate this approach for the general bivariate heterogeneity distribution in the next section. Our approach is related to recent work by Weinstein, Ma, Brown, and Zhang (2015) on grouped patterns of heterogeneity in the normal mean model, although the longitudinal data structure here permits estimation methods that are considerably more general.

4. Empirical Bayesball

Following a long tradition in the empirical Bayes literature, we now describe our experience with the methods described above deployed to predict U.S. Major League baseball batting averages.

4.1. **The Data.** From ESPN (2012) we have collected monthly data on the number of at bats and hits for all U.S Major League baseball players from the regular seasons of 2002-2011, as well as an indicator of whether the player is a pitcher. These ten prior seasons are employed to fit our mixture model and then used to predict performance of players in the 2012 season. We have aggregated this annual data into half seasons to produce an unbalanced panel, with observations on players with more than 10 at bats in any half season, and players with no less than 3 half-seasons, leaving 1072 players and a total of 10,570 observations. Since it is reasonable to assume that the batting performance for pitchers and non-pitchers are sufficiently different, we only focus on non-pitchers in our data analysis. Using the same selection criteria, this leaves us with 898 players and a total of 9,199 observations. For the final subsection on age effects we have also collected information on birth year of each player from the ESPN website.

4.2. **The Model.** Following Brown (2008), we consider transformed batting averages

$$y_{it} = \arcsin\left(\sqrt{\frac{H_{it} + 0.25}{N_{it} + 0.5}}\right)$$

where H_{it} denotes the number of "hits" of player i in period t and N_{it} denotes his number of "at bats" in this period. We will assume, that the y_{it}'s are Gaussian with means, $\mu_i = \arcsin(\sqrt{p_i})$, where p_i is the individual specific batting success probability and variances, $\theta_i \nu_{it}^2 = \theta_i/(4N_{it})$. The additional individual specific scale parameter θ_i allows us to consider deviations from the variance dictated by the binomial-Gaussian transformation. Given an unbalanced panel with $t = 1, \cdots, m_i$ for n players, we can define the sufficient statistics:

$$\hat{\mu}_i = (\sum_{t=1}^{m_i} \nu_{it}^{-2})^{-1} \sum_{t=1}^{m_i} y_{it}/\nu_{it}^2$$

and

$$S_i = \frac{1}{m_i - 1} \sum_{t=1}^{m_i} (y_{it} - \hat{\mu}_i)^2/\nu_{it}^2.$$

with Gaussian and Gamma distributions, respectively: $\hat{\mu}_i|\mu_i, \theta_i \sim \mathcal{N}(\mu_i, \theta_i \nu_i^2)$ and $S_i|\theta_i \sim \gamma(r_i, \theta_i/r_i)$, where we set $\nu_i^2 = (4 \sum_{t=1}^{m_i} N_{it})^{-1}$ and $r_i = (m_i - 1)/2$.

For the sake of clarity, we defer introducing covariates until the final subsection. In effect we assume that players draw a μ_i and a θ_i at random from a distribution with density $h(\mu, \theta)$. Sufficiency then implies that we can write the likelihood of the sample, as a function of $(\mu, \theta) \in \mathbb{R}^n \times \mathbb{R}_+^n$,

$$\begin{aligned} L(h) &= \prod_{i=1}^{n} g((\mu, \theta) \mid y_{i1}, \cdots, y_{im_i}) \\ &= \prod_{i=1}^{n} \int\int \prod_{t=1}^{m_i} \phi((y_{it} - \mu)/\sqrt{\theta}\nu_{it})/(\sqrt{\theta}\nu_{it}) h(\mu, \theta) d\mu d\theta \\ &= K \prod_{i=1}^{n} \int\int \phi((\hat{\mu}_i - \mu)/\sqrt{\theta}\nu_i)/(\sqrt{\theta}\nu_i) \frac{e^{-R_i} R_i^{r_i}}{S_i \Gamma(r_i)} h(\mu, \theta) d\mu d\theta \end{aligned}$$

where $R_i = r_i S_i/\theta_i$.

As we have already noted, it is convenient at this point to make the further assumption that the mixing density h factors into $h = h_\mu h_\theta$. In this case, since the likelihood contribution of the S_i is independent of the μ_i's, we can solve the resulting (Kiefer-Wolfowitz) maximum likelihood problem by first estimating the mixing density, h_θ and then estimating h_μ. For the independent prior case this is especially convenient since the likelihood can be decomposed into gamma and Student components, or somewhat more generally into gamma and Gaussian components, as described above. For the more general dependent prior case, we no longer have the Gamma separability, but this imposes no inherent technical difficulty except that it involves two dimensional gridding. The possibility of nonparametrically estimating the joint prior allows arbitrary dependence between the (μ, θ) that provides an interesting interaction between hitting ability (measured by μ) and hitting consistency (measured by θ) and leads to a more sophisticated Bayes rule for both μ and θ.

4.3. **Estimation Results**

4.3.1. *Independent Prior.* In Figure 1 we depict the estimated mixing densities for the means and variances, assuming independence between μ's and θ's, for the model described in the previous subsection. In the transformed scale of the μ's we see one large peak, and several smaller ones, with slight upper and lower "foothills." For the variance parameter θ we see only one very pronounced peak slightly above one, and one smaller peak below one. Recall that the variance scale is relative to the binomial model, which implies that variance ν_{it} is completely determined by the observed data on the number of at bats for each player (recall $\nu_{it}^2 = 1/4N_{it}$). Thus, a single peak at exactly one would imply exact adherence to the binomial model. Instead, we see a modest over-dispersion effect from

the large peak, and a much smaller peak corresponding to players exhibiting under-dispersion. For the latter, more consistent, players prediction of future performance would presumably be slightly easier.

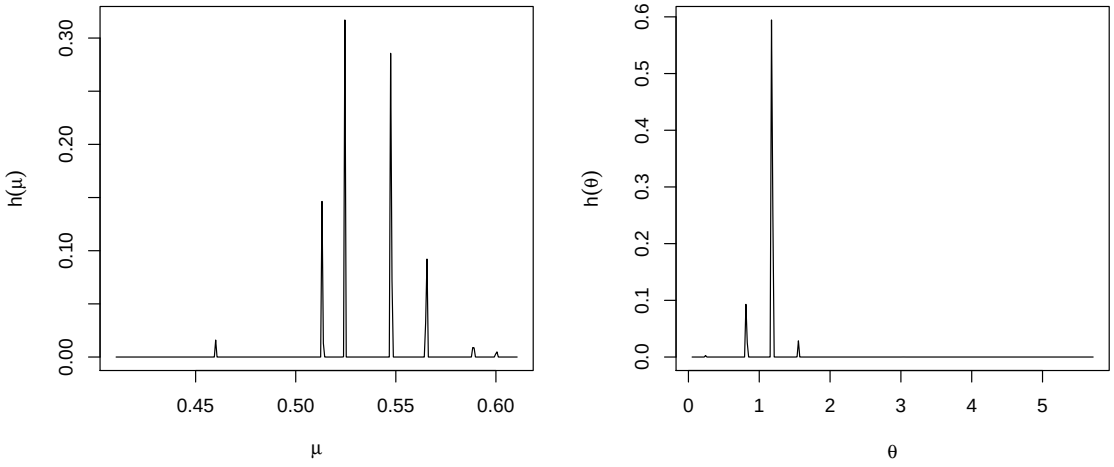

FIGURE 1. Estimated Mixing Distributions for μ and θ based on 2002-11 longitudinal data for non-pitchers. Note that these distributions corresponds to the transformed, approximately Gaussian, observations. See Figure 2 for the untransformed estimated batting average distribution implied by this $h_\mu(\mu)$ distribution. The $h_\theta(\theta)$ distribution depicted in the right panel can be interpreted as an estimate of a mixture of under and over dispersion components of the observed variances; the (larger) mass point with $\theta > 1$ shows that most players exhibit overdispersion relative to the binomial model, while the (smaller) mass point with $\theta < 1$ corresponds to a group of players that are less variable more consistent than predicted by the binomial model.

Transforming the estimated mixing density, or prior, for the μ's back to the natural scale of batting averages yields the distribution shown in left panel of Figure 2. Again, we see a similar configuration of peaks, but now located at more familiar places on the $[0, 1]$ interval – at least from baseball perspective. In the right panel of Figure 2 we illustrate the estimated mixing density from the Kiefer-Wolfowitz MLE of a binomial model based on the aggregated baseball data over the period 2002-11. This estimate has the advantage of simplicity, but it neglects the potential for over or under dispersion of the data.

4.3.2. *Joint Prior.* More generally as discussed earlier, we can relax the independence assumption and estimate the joint prior distribution of $h(\mu, \theta)$. These estimation results are presented in Figure 3, which shows both the two dimensional and three dimensional plot of the Kiefer-Wolfowitz estimates of the joint distribution $\hat{h}(\mu, \theta)$ on a 100 by 100 grid. The joint distribution shows some positive dependence between the mean and the variance with a Spearman's rho measure of rank correlation of 0.4 based on a 1000 simulated samples from $\hat{h}(\mu, \theta)$ with standard error 0.004. The positive dependence suggests that players with a better batting ability also tend to have more variability in their hitting performance (after accounting for the differences in the number of at bats).

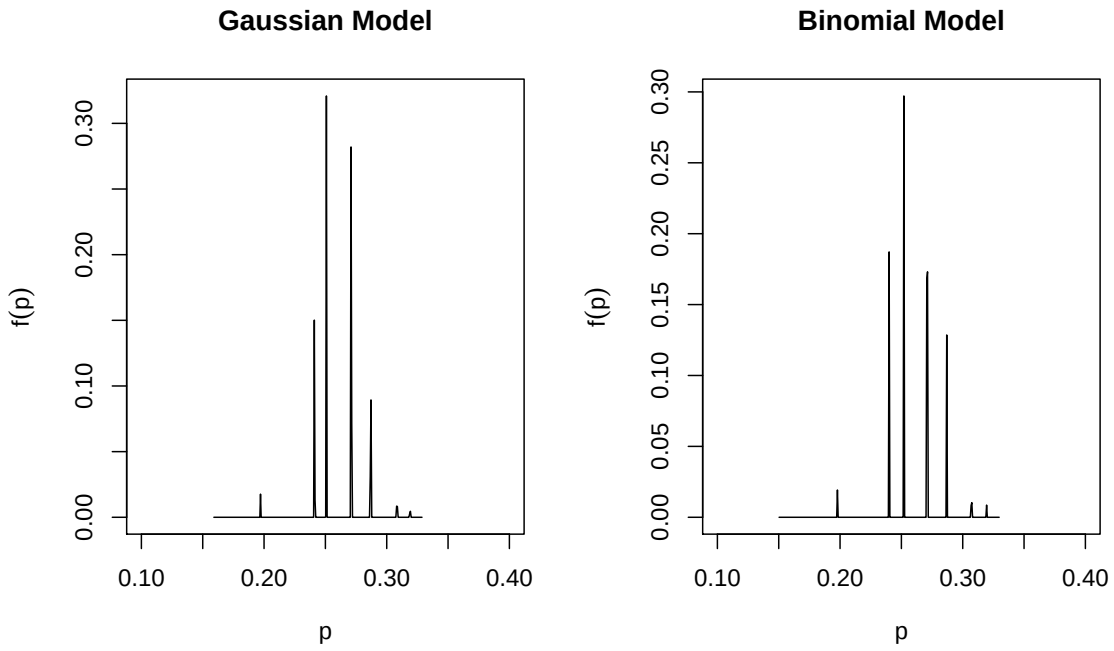

FIGURE 2. Estimated Mixing Distributions for Batting Averages for non-pitchers: In the left panel we present the transformed density based on the Gaussian model described above. And in the right panel we show the estimated mixing distribution for the MLE of the binomial model based on the aggregated performance of the players from 2002-11.

If we adopt the managerial perspective of players as assets, this resembles the familiar phenomenon that high return is associated with higher risk.

The estimated joint prior distribution $\hat{h}(\mu, \theta)$ presents us a special opportunity to identify players that stand out as having large estimates for μ but small estimates for θ. For each data point $(\hat{\mu}_i, S_i)$, Bayes theorem leads to a player specific posterior distribution $h(\mu, \theta \mid \hat{\mu}_i, S_i)$. Under the squared error loss, the optimal estimator for μ that minimizes $\mathbb{E}(\delta_\mu - \mu)^2$ then leads to the Bayes rule $\delta_\mu = \mathbb{E}(\mu|\hat{\mu}, S)$. Similarly, the Bayes rule for estimating θ under squared error loss is $\delta_\theta = \mathbb{E}(\theta|\hat{\mu}, S)$. Figure 4 presents the scatter plot of $(\delta_\mu, \delta_\theta)$ for all the players given their batting history over 2002 - 2011. Table 1 lists players that have the largest δ_μ (right panel) and the smallest δ_θ (left panel) estimates based on these point estimates. Most of the top hitters (right panel players) have relatively large posterior expectation of θ except for Albert Pujols who has the highest estimated μ (converting back to the probability scale, Pujols's predicted batting average is 0.322) but also the lowest θ among the top 10 players. All players in the left panel have under-dispersion relative to the binomial model. We may highlight a few players: Cuddyer, Nady, Blake and Hernandez, who have relatively high estimates for their batting abilities among those players with significant underdispersion. Among these, Michael Cuddyer has the highest estimates for δ_μ, and also exhibits the lowest variability, δ_θ by a substantial margin among all 898 players.

4.4. **Out-of-Sample Model Checking.** Before proceeding to the more conventional baseball prediction exercise, we first conduct some out-of-sample model checks. Out of the 898 non-pitchers,

Name	$\hat{\mu}$	S	δ_μ	δ_θ
Michael Cuddyer	0.549	0.215	0.549	0.294
Paul Bako	0.489	0.434	0.499	0.686
Xavier Nady	0.551	0.324	0.547	0.756
Jeff Mathis	0.453	0.688	0.458	0.847
Kevin Cash	0.444	0.661	0.461	0.855
Koyie Hill	0.477	0.607	0.497	0.867
Casey Blake	0.540	0.462	0.542	0.879
Dewayne Wise	0.490	0.515	0.510	0.884
Ramon Hernandez	0.545	0.466	0.545	0.889
Yorvit Torrealba	0.537	0.444	0.537	0.906

Name	$\hat{\mu}$	S	δ_μ	δ_θ
Albert Pujols	0.611	0.904	0.604	1.090
Ichiro Suzuki	0.604	2.316	0.597	1.633
Joe Mauer	0.605	1.398	0.593	1.391
Barry Bonds	0.607	2.307	0.593	1.454
Miguel Cabrera	0.599	0.926	0.592	1.333
Todd Helton	0.598	1.604	0.592	1.406
Vladimir Guerrero	0.598	1.031	0.591	1.349
Matt Holliday	0.596	0.810	0.591	1.336
Magglio Ordonez	0.592	1.794	0.591	1.380
Manny Ramirez	0.593	1.451	0.590	1.363

TABLE 1. The left panel consists of players with the lowest posterior expectation of θ and the right panel consists players with the highest posterior expectation of μ.

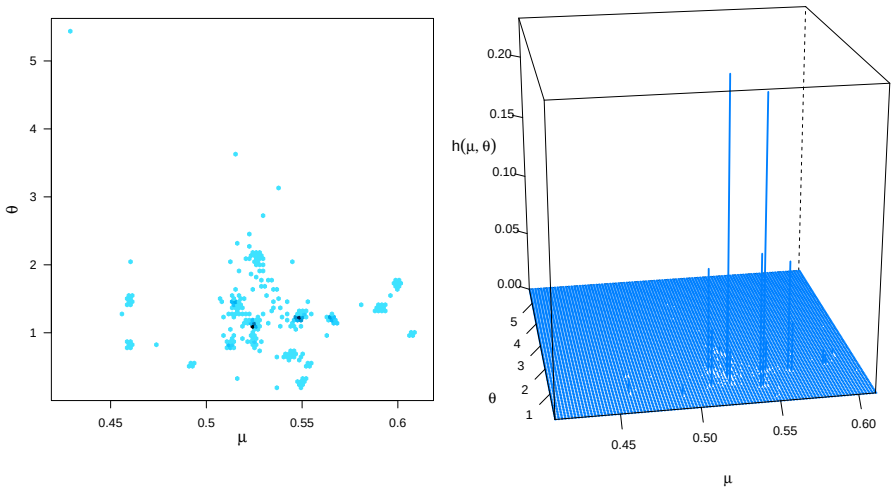

FIGURE 3. Estimated Joint Mixing Distributions for for Non-pitchers: In the left panel we present the two dimensional plots allowing a clear visualization of the location of the support points for (μ, θ). Darker color represents the support points having higher probability mass. In the right panel we show the three dimensional plot which gives a better visualization of the magnitude of the probability weights and illustrate the discreteness feature of the estimated joint distribution.

there remains 344 players with more than 40 at bats in 2012 and had 3 or more half seasons prior to 2012 and were therefore qualified subjects for our out-of-sample model evaluation.

We conduct the following posterior predictive checks based on the estimated models introduced above, and also compare their performance with more formal nonparametric Bayesian methods based on Dirichlet process priors. First, we evaluate the likelihood of observing the 2012 batting outcome

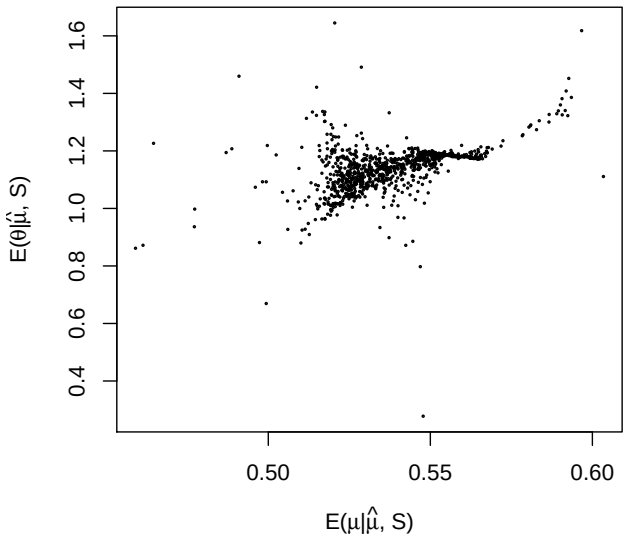

FIGURE 4. Bayes rule of (μ, θ) for all non-pitchers based on the Kiefer-Wolfowitz nonparametric estimates of the joint prior distribution $h(\mu, \theta)$ and their 2002-2011 batting history.

based on the posterior predictive distribution. Given the Gaussian location-scale mixture model, the posterior predictive density for the 2012 batting average for individual i who has N_i at bats can be found as

$$\hat{f}_i(y|\hat{\mu}_i, S_i, N_i) = \int\int \phi((y-\mu)/\sqrt{\theta/4N_i})/\sqrt{\theta/4N_i}\, d\hat{H}_i(\mu, \theta|\hat{\mu}_i, S_i) \tag{4}$$

where $\hat{H}_i$ is the individual specific posterior distribution of (μ, θ) updated based on his own 2002-2011 history. The posterior predictive density can be used to assess model fit. We look at both the log likelihood value and the tail probability evaluated at the 2012 batting outcomes. The log likelihoods are based on a particular model assumption (e.g. different assumptions on the distribution of (μ, θ)), indexed by m, is computed as $\ell_m \equiv \sum_{i=1}^{344} \log \hat{f}_i(y_i|\hat{\mu}_i, S_i, N_i)$, which can be interpreted as the log likelihood of observing the new data using the predictive model. Counting the number of predictive tail events provides another measure for evaluating model adequacy, we compute

$$T_{ff} = \sharp\{i \mid Y_i \geqslant \hat{Q}_i(1-\tau), \text{or } Y_i \leqslant \hat{Q}_i(\tau)\},$$

where $\hat{Q}_i(\cdot)$ is the empirical quantile function of the posterior predictive distribution for the ith player and Y_i is his realized 2012 (transformed) batting average. We simply count the number of players whose 2012 realizations occur below the τ or above the $1-\tau$ tail of the posterior predictive distribution based on the various methods. These values, with $\tau = 0.025$, are reported along with in-sample log likelihoods for three Gaussian models in Table 2.

	LV-Dep	LV-ISG	LV-ING	G-L
Tail Index	25	28	27	35
Log-Likelihood	605.7	607.1	605.2	595.7

TABLE 2. Posterior log likelihood and T_{ff} performance for various Gaussian transformation models: LV-Dep refers to the location-scale Gaussian mixture model with the joint nonparametric prior, LV-ISG is the location-scale mixture model with independent prior on (μ, θ) estimated by the Student/Gamma procedure, and LV-ING is the independent prior model estimated with the iterative Normal/Gamma method. G-L is location Gaussian mixture model that ignores over/under dispersion, but accounts for the differences in number of at bats in each period.

All the above posterior predictive checks can also be conducted for the binomial model discussed earlier in Section 4.3.1 with the posterior predictive distribution becoming,

$$\hat{f}_i(h|\hat{H}_i, \hat{N}_i, N_i) = \int \binom{N_i}{h} p^h (1-p)^{N_i - h} d\hat{F}_i(p|\hat{H}_i, \hat{N}_i),$$

and the tail index defined in the same way. Given the binomial model, the predictive distribution is the posterior distribution of the number of hits in 2012 for player i given his number of at bats N_i and his previous batting performance.

In Table 3 we present for several binomial models the forecast tail index for extreme events, T_{ff} with $\tau = 0.025$. The smaller this number, the more confidence we have in the respective forecasting model. To compare the performance of the foregoing predictions with more classical Bayesian methods we have also considered an alternative nonparametric formulation of the binomial model employing a Dirichlet process prior as introduced by Ferguson (1973), Antoniak (1974) and Ferguson (1983). Deferring a discussion of the details to Appendix B, we consider two versions of the Dirichlet model both with Dirichlet prior $\mathcal{D}(\alpha G_0)$ and base distribution G_0 taken as Beta$(1, 1)$, i.e. uniform on $[0, 1]$. The scaling parameter, α is either 0.01 or 10, both reflecting relative ignorance about the prior. Markov chain Monte-Carlo methods produce a posterior of the predictive density for the mixing distribution corresponding to the point estimate produced by the Kiefer-Wolfowitz estimator. The mean of this posterior, as plotted in Figure 5 serves as a predictive density for the mixing distribution of the binomial parameter can thus be used to produce a predictive distribution on hits for our out-of-sample 2012 players as for the binomial model described above. The prediction performance of the binomial mixture model using the Kiefer-Wolfowitz MLE or the nonparametric Bayesian method with Dirichlet process prior leads to very similar results. Table 2 also suggests that under the Gaussian model, allowing over or under dispersion of the data is an important consideration. The likelihood is smaller for the Gaussian location mixture model that forces $\theta_i = 1$ (Model G-L in Table 2) and it has considerably more tail events than the two other models that allow for variance heterogeneity.

4.5. **Age and the Skill of Batting.** Finally, to illustrate the role that covariates can play in the empirical Bayes approach we have developed, we consider a model in which a player's age influences his batting average. Following the approach outlined in the previous section we compute the profile likelihood on a grid of β parameters. The only required modification is the reweighting by the ν_{it}'s. Given the conventional wisdom that the batting performance has a hump shaped age profile, we consider a simple quadratic age effect. A contour plot of the profile likelihood is shown as Figure 6 and we find, after unrescaling, that the age effect appears as shown in Figure 7. The peak occurs at 27 years, by age 39 performance has declined by about 0.02 units in the transformed batting average, which corresponds to a decline of a typical batting average of 0.320 at the peak to 0.302.

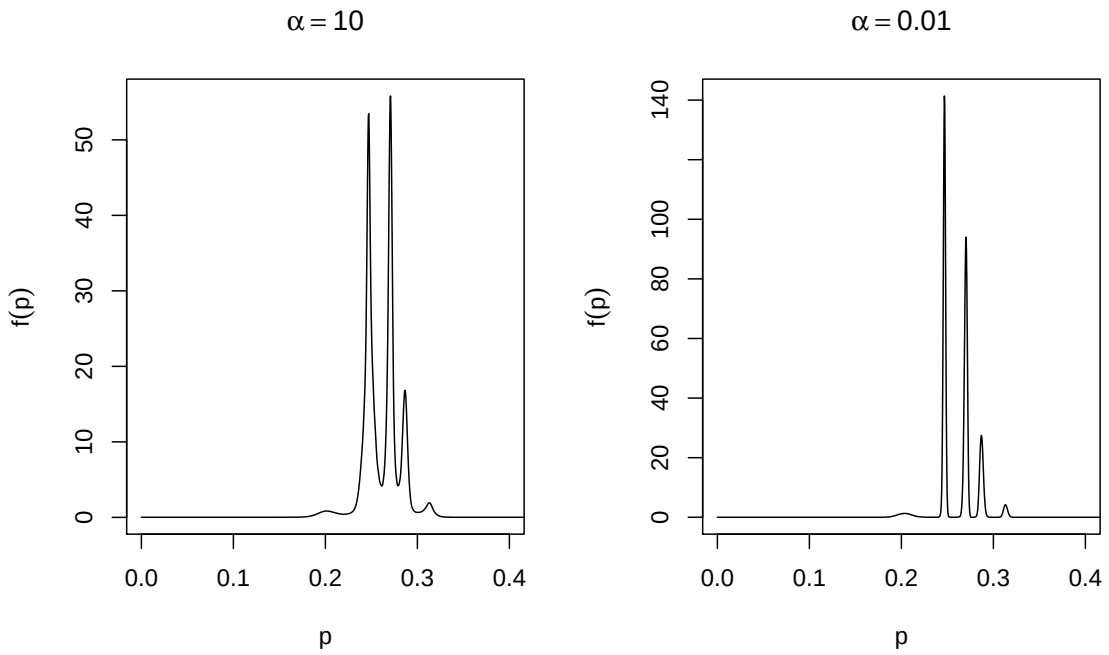

Figure 5. Dirichlet Estimates of the Mixing Distribution: In the left panel we plot the estimated mixing distribution based on the Dirichlet prior with precision parameter $\alpha = 10$. In the right panel prior precision has been reduced to $\alpha = 0.01$. These estimates may be interpreted as regularized versions of those produced by the Kiefer-Wolfowitz MLE appearing in Figure 2.

	B-KW	DP-0.01	DP-10
Tail Index	30	30	31
Log-Likelihood	−1286.5	−1288.6	−1286.9

Table 3. Posterior log likelihood and T_{ff} performance for various binomial mixture models: B-KW is the Binomial mixture model with aggregated data based on the Kiefer-Wolfowitz estimator for the mixing distribution. DP-0.01 is the binomial model with a Dirichlet Process prior and prior precision parameter $\alpha = 0.01$. DP-10 is the same Dirichlet model with $\alpha = 10$.

The well-known baseball guru Bill James is on record as asserting that batting ability peaks at age 27, so our estimates are at least consistent with his not-so-casual empiricism. Our estimated age profile also bears an uncanny resemblance to the age profile of Ty Cobb over the period 1905 to 1928 analyzed in Morris (1983). Incorporating the quadratic age effects improves the predictive performance slightly, so that the likelihood is increased to 620.4 while the number of tail events is reduced to 26 of the 344 players predicted.

4.6. **Prediction.** An inherent difficulty for prediction with baseball data is that the number of at bats for the 344 out-of sample players in 2012 varies from 41 to 670. Since the number of at bats N_{it}

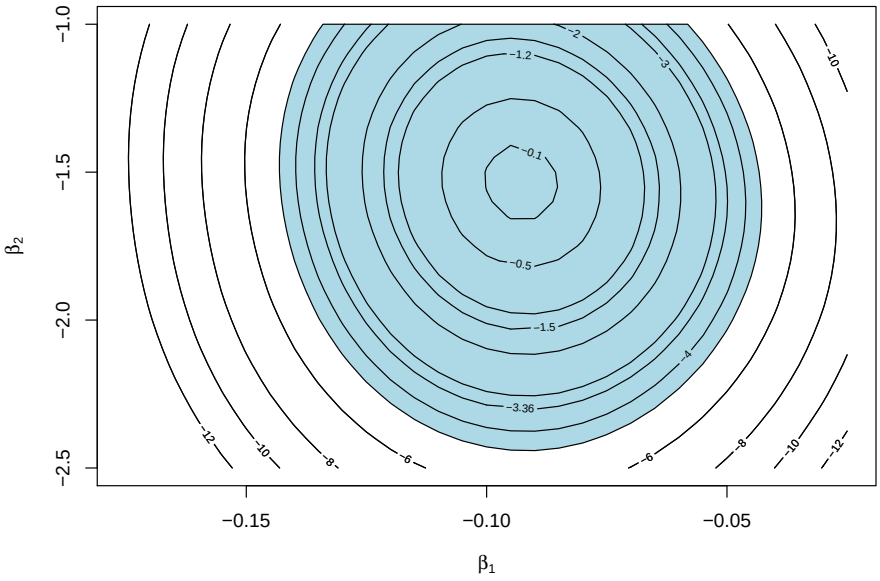

FIGURE 6. Contours of Profile Likelihood for a Quadratic Age Effect on Batting Average: The Wilks confidence region for the age effect parameters is represented by the shaded region using 90% critical values from the χ^2_2 distribution.

affects the variability of the transformed batting averages, $y_{it} \sim \mathcal{N}(\mu_i, \theta_i/4N_{it})$, even if we were to get perfect estimates for each (μ_i, θ_i), players with larger N_{it} are subject to less variability for y_{it}, making them (presumably) easier to predict. To unify the comparison, we adopted the following measure proposed in Brown (2008)

$$\text{TSE} = \sum_i \left((Y_{i,2012} - \delta_{\mu,i})^2 - \frac{1}{4N_{i,2012}} \right)$$

where $Y_{i,2012}$ is the transformed hitting averages in 2012 for all qualified out-of-sample players and $\delta_{\mu,i}$ is a prediction from various models. Without the $1/4N_{i,2012}$ term, this is the usual sum of squared error, and the additional term accounts for the variance effect due to different number of at bats in 2012. An alternative, more straight-forward measure that we will also report is the normalized sum of squared error,

$$\text{NSE} = \sum_i \left(4N_{i,2012}(Y_{i,2012} - \delta_{\mu,i})^2 \right).$$

This measure weights players who have more at bats more heavily.

Earlier literature has illustrated that empirical Bayes methods can improve predictive performance substantially over conventional linear regression models. Most this work employs a single cross section. For example, both Brown (2008) and Jiang and Zhang (2010) predict batting averages for the second half season of 2005 using data from the first half. Jiang and Zhang (2010) concludes that a linear model with the number of at bats, an indicator of whether the player was a pitcher and their interaction, together with an individual specific latent location shift effect yielded the best

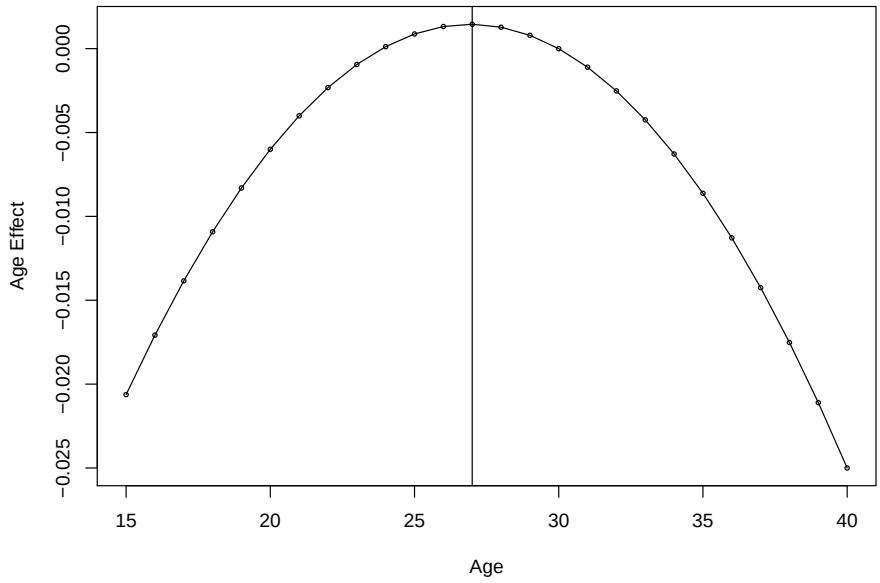

FIGURE 7. The Estimated Quadratic Age Effect: The vertical axis is in units of the transformed batting average.

predictive performance. More explicitly their model takes the form,

$$y_i = X_i\beta + \mu_i + \sigma_i\epsilon_i$$

with $\epsilon_i \sim \mathcal{N}(0,1)$ and $\sigma_i^2 = 1/4N_i$ is assumed to be a known quantity. Pitchers bat less frequently and are generally weaker hitters than other players, so it might be preferable to treat them as a separate sub-population. One might also argue that conditioning on the number of player at bats violates the causal perspective of the predictive model: baseball managers decide how many at bats players have based on their prior performance. For these reasons we restrict attention to non-pitchers in our prediction exercise, and condition only on players' age.

Conditional on a $\hat{\beta}$, Jiang and Zhang (2010) proposed estimation of G by the nonparametric maximum likelihood method of Kiefer and Wolfowitz by via the EM algorithm. Estimation of β can then be carried out by profile likelihood. Prediction for each individual player under $\mathcal{L}_2$ loss becomes,

$$\delta_{\mu,i} = x_i\hat{\beta} + \frac{\int \mu\varphi(y_i - x_i\hat{\beta} - \mu)/\sigma_i)/\sigma_i d\hat{G}(\mu)}{\int \varphi(y_i - x_i\hat{\beta} - \mu)/\sigma_i)/\sigma_i d\hat{G}(\mu)}$$

Rather than conditioning on only the most recent half-season performance, one might want to consider longitudinal models that utilize a full record of past performance. Lai, Su, and Sun (2014) have recently considered a parametric empirical Bayes model of the general form,

$$y_{it} = \sum_{j=1}^{p} \rho_j \bar{y}_{t-j} + x_{it}\beta + \mu_i + \epsilon_{it}.$$

Batting averages in the second half of the 2010 season are predicted using training data from the prior half seasons starting in 2006. BIC model selection simplifies their predictive model by setting $\beta = 0$ and $p = 1$. Here $\bar{y}_{t-j}$ is defined as $\sum_{i \in \mathcal{S}_{t-j}} y_{i,t-j}/\#|\mathcal{S}_{t-j}|$, the average of the transformed batting average among the set of players $\mathcal{S}_{t-j}$ that play in half-season $t-j$, so dynamics are restricted to an average half-season effect. In contrast to the nonparametric approach of Jiang and Zhang (2010), Lai, Su, and Sun (2014) adopt a parametric formulation of the latent ability effect, assuming $\mu_i \sim \mathcal{N}(0, \tau^2)$. This assumption has the advantage that it leads to linear shrinkage rules, but otherwise seems hard to justify.

The nonparametric empirical Bayes methods we have described in earlier sections encompass both the dynamic longitudinal features of the Lai, Su, and Sun (2014) approach as well as an expanded version of the nonparametric heterogeneity of the Jiang and Zhang (2010) models with both location and scale heterogeneity. Crucially, interior point methods make the computation of the Kiefer Wolfowitz estimator much more efficient and therefore greatly facilitates the requisite profile likelihood optimization.

In Table 4 we compare the predictive performance of several variants of these models; batting averages for the 2012 season are predicted based on data for the 2002-11 seasons. When location-scale model is used which accounts for possible additional unobserved heterogeneous θ, the prediction for each individual player, with past history summarized by the pair $(\hat{\mu}_i(x_i, \hat{\beta}), S_i(x_i, \hat{\beta}))$ defined in Section 4.2 is given by

$$\delta_{\mu,i} = x_i\hat{\beta} + \frac{\int\int \mu\varphi((\hat{\mu}_i - x_i\hat{\beta} - \mu)/\sqrt{\theta}v_i)/(\sqrt{\theta}v_i)\gamma(S_i|r_i, \theta)d\hat{H}(\mu, \theta)}{\int\int \varphi((\hat{\mu}_i - x_i\hat{\beta} - \mu)/\sqrt{\theta}v_i)/(\sqrt{\theta}v_i)\gamma(S_i|r_i, \theta)d\hat{H}(\mu, \theta)}$$

where the covariate vector x_i consists a quadratic in player age, and $\bar{y}_{2011}$, mean batting average for 2011. Rows labeled "LV" in the table refer to these models with both location and scale heterogeneity. Imposing independence of the location and scale effects in these models actually improves predictive performance without covariates, so we have focused primarily on the models that impose independence on $H(\mu, \theta)$. In the income dynamics application of Gu and Koenker (2015) dependence in the estimated mixture distribution plays an important role. To illustrate the potential importance of accounting for the additional over/under dispersion in θ, we also consider a location mixture model which only admits the heterogeneous μ_i for individual players (for example, force $\theta_i = 1$ for all i). The prediction, with past history summarized by $\hat{\mu}_i(x_i, \hat{\beta})$ defined in Section 4.2 is then

$$\delta_{\mu,i} = x_i\hat{\beta} + \frac{\int\int \mu\varphi((\hat{\mu}_i - x_i\hat{\beta} - \mu)/v_i)/v_i d\hat{H}(\mu)}{\int\int \varphi((\hat{\mu}_i - x_i\hat{\beta} - \mu)/v_i)/v_i d\hat{H}(\mu)}$$

Row labeled "L" in the table refer to predictions based on these models. The column labeled as RTSE and RNSE in Table 4 presents the relative performance of all models relative to the "Naive" prediction that uses $Y_{i,2011}$ as a prediction for $y_{i,2012}$ based on TSE or NSE measure respectively. All of the empirical Bayes procedures improve substantially on this naive forecast. The simple "Lag" model that uses only $\bar{Y}_{2011}$ as a linear predictor for $Y_{i,2012}$ and allows no heterogeneity suggests that the past year average batting performance is a useful predictor for the linear part of the model. Adding covariates effects to the LV models substantially improves predictions. When we restrict to only location heterogeneity predictions, the model without covariates remain quite good, but those with covariates not as strong as their corresponding more flexible location-scale models.

We also compare the prediction in terms of the probability scale as in Muralidharan (2010). In this setting, we can also consider models based on binomial model assumptions. Following Jiang and Zhang (2010), we convert the prediction back to probability scale via $\hat{p}_i = \sin^2(\delta_{\mu,i})$ and the

Models	TSE	RTSE	NSE	RNSE	TSEp	RTSEp
LV-ISG	0.323	0.499	594.418	0.596	0.241	0.622
LV-ING	0.331	0.512	598.742	0.600	0.247	0.635
LV-Dep	0.328	0.507	598.059	0.599	0.245	0.631
LV-ISG-Age	0.276	0.427	563.406	0.565	0.206	0.530
LV-ING-Age	0.284	0.439	568.327	0.570	0.211	0.544
LV-ISG-Age-Lag	0.218	0.337	540.281	0.542	0.163	0.421
LV-ING-Age-Lag	0.225	0.348	544.004	0.545	0.168	0.432
Naive	0.647	1.000	997.683	1.000	0.388	1.000
Lag	0.382	0.590	913.163	0.915	0.312	0.803
L	0.328	0.507	597.626	0.599	0.245	0.631
L-AGE	0.361	0.558	624.307	0.626	0.271	0.697
L-AGE-Lag	0.511	0.790	753.566	0.755	0.386	0.993
Bmix					0.246	0.633
DP-10					0.246	0.632
DP-0.01					0.248	0.638

Table 4. Predictive performance for various models. Predictions based on models with heterogeneity in both location and scale are identified by the row label LV. Three variants of the LV models are compared: LV-Dep refers to the unrestricted joint distribution model, LV-ISG refers to the independent prior model estimated with the Student/Gamma procedure, and LV-ING to the independent prior model estimated with the Normal/Gamma iterative procedure. Models with only location heterogeneity have prefix L. Naive prediction refers to using only $Y_{i,2011}$ as a prediction for $Y_{i,2012}$, Lag prediction refers to using only $\bar{Y}_{2011}$ as a linear predictor for $Y_{i,2012}$. Models with additional covariate effects are identified by appended Age and/or Lag suffixes Binomial mixture models are labeled Bmix for predictions based on the NPMLE, and DP$-\alpha$ for predictions based on a binomial mixture model with Dirichlet process prior and precision parameter α. The TSE, NSE and TSEp columns evaluate predictive performance on the transformed and probability scales, respectively. RTSE, RNSE and RTSEp evaluate performance relative to the Naive model.

total sum of square on the probability scale is defined as

$$\text{TSEp} = \sum_i \left((p_{i,2012} - \hat{p}_i)^2 - p_{i,2012}(1 - p_{i,2012})/N_{i,2012} \right)$$

where $p_{i,2012}$ is the ratio of number of hits to the number of at bats for qualified out-of-sample player i in 2012 and $\hat{p}_i$ is the prediction for the success probability based on various models. The last three rows In Table 4 correspond to a Binomial mixture model and two Dirichlet Process prior models with different precision parameters α (DP-10 for $\alpha = 10$ and DP-0.01 for $\alpha = 0.01$). The probability scale prediction from the transformed data with location or location-scale mixture model dominates those using the Binomial model. The column labeled as RTSEp is again the relative performance of all models relative to the "Naive" prediction that uses $H_{i,2011}/N_{i,2011}$ as a prediction for $p_{i,2012}$ based on TSEp.

5. Conclusion

Models of unobserved heterogeneity for longitudinal data are common in applied statistics. 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 their estimation. 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 is not especially onerous. Moreover, in our baseball application normality is easily justified since the transformed batting averages have a strong claim to approximate normality. In some special cases it may be possible to *estimate* the base distribution nonparametrically from extraneous sample information. Recent work of Bonhomme and Sauder (2011) illustrates an educational treatment effect application of this type for which a deconvolution strategy based on empirical characteristic functions is employed. We hope to explore the extension of Kiefer-Wolfowitz mixture methods to such settings in future work.

REFERENCES

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

ANTONIAK, C. (1974): "Mixtures of Dirichlet Process with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2(6), 1152–1174.

BLACKWELL, D., AND J. MACQUEEN (1973): "Ferguson Distribution via Pólya Urn Schemes," *The Annals of Statistics*, 1(2), 353–355.

BONHOMME, S., AND U. SAUDER (2011): "Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling," *Review of Economics and Statistics*, 93, 479–494.

BROWN, L. (2008): "In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies," *The Annals of Applied Statistics*, 2, 113–152.

BROWN, L., AND E. GREENSHTEIN (2009): "Nonparametric empirical Bayes and compound decision approaches to estimation of a high dimensional vector of normal means," *The Annals of Statistics*, 37(4), 1685–1704.

BÜHLMANN, H., AND A. GISLER (2005): *A course in credibility theory and its applications*. Springer.

BUSH, C., AND S. MACEACHERN (1996): "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, pp. 275–285.

CASTILLO, I., AND A. VAN DER VAART (2012): "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences," *Annals of Statistics*, 40, 2069–2101.

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.

ESCOBAR, M., AND M. WEST (1994): "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90(430), 577–588.

ESPN (2012): "Major League Baseball Statistics," `http://espn.go.com/mlb/statistics`.

FERGUSON, T. (1973): "A Bayesian analysis of some no parametric problems," *The Annals of Statistics*, 1, 209–230.

——— (1983): "Bayesian Density Estimation by Mixtures of Normal Distributions," in *Recent Advances in Statistics*, ed. by H. Rizvi, and J. Rustagi, pp. 287–302. Academic Press: New York.

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

GELFAND, A., AND A. SMITH (1990): "Sampling-based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

GOOD, I. (1953): "On the population frequencies of species and the estimation of population parameters," *Biometrika*, 40, 237–264.

——— (1979): "Some History of the Hierarchical Bayesian Methodology (with discussion)," in *Bayesian Statistics: Proceedings of the First International Meeting in Valencia*, ed. by J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, pp. 489–510. University of Valencia.

GU, J., AND R. KOENKER (2015): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," *J. of Business and Economic Statistics*, forthcoming.

GUARINO, C., M. MAXFIELD, M. D. RECKASE, P. N. THOMPSON, AND J. M. WOOLDRIDGE (2015): "An Evaluation of Empirical Bayess Estimation of Value-Added Teacher Performance Measures," *J. of Educational and Behavioral Statistics*, 40, 190–222.

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.

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

Jiang, W., and C.-H. Zhang (2010): "Empirical Bayes in-season prediction of baseball batting averages," in *Borrowing Strength: Theory Powering Applications: A Festschrift for Lawrence D. Brown*, vol. 6, pp. 263–273. Institute for Mathematical Statistics.

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

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

——— (2016): "REBayes: A Vignette on Empirical Bayes Methods," Available from `http://cran.r-project.org`.

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

Lai, T., Y. Su, and K. Sun (2014): "Dynamic Empirical Bayes Models and Their applications to longitudinal data anlaysis and prediction," *Statistica Sinica*, 24, 1505 – 1528.

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

Lindley, D. V., and A. F. Smith (1972): "Bayes estimates for the linear model," *Journal of the Royal Statistical Society. Series B*, pp. 1–41.

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," *The Annals of Statistics*, 24(3), 911–930.

Lo, A. (1984): "On a Class of Bayesian Nonparametric Estimates: 1. Density Estimates," *The Annals of Statistics*, 12, 351–357.

Morris, C. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–55.

Muralidharan, O. (2010): "An Empirical Bayes Mixture Method for Effect Size and False Discovery Rate Estimation," *The Annals of Applied Statistics*, 4, 422–438.

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

R Core Team (2014): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria.

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.

Weinstein, A., Z. Ma, L. D. Brown, and C.-H. Zhang (2015): "Group-Linear Empirical Bayes Estimates for a Heteroscedastic Normal Mean," `http://arxiv.org/pdf/1503.08503`.

West, M. (1990): "Bayesian Kernal Density Estimation," *ISDS Discussion Paper #90-A02*.

West, M., P. Müller, and M. Escobar (1994): "Hierarchical Priors and Mixture Models, With Application in Regression and Density Estimation," in *Aspects of Uncertainty*, ed. by P. Freeman, and A. Smith, pp. 363–386. Wiley: New York.

Appendix A. Dirichlet Process Methods for Compound Decisions

Ferguson (1973) first proposed using the Dirichlet process as a prior distribution for estimating an unknown probability measure P. Antoniak (1974) and Ferguson (1983) extend the Dirichlet process to a mixture of Dirichlet processes, suited to compound decision problems like the binomial mixture model we have considered above. We do not observe samples on $\mu_1, \ldots, \mu_n$ directly, instead we would like to find the posterior distribution of μ conditional on observing $y_1, \ldots, y_n$. Antoniak (1974) shows that if $\mu_1, \ldots, \mu_n$ are drawn from a distribution F, with Dirichlet process prior $\mathcal{D}(\alpha G_0)$, and if $Y_1, \ldots, Y_n$ are i.i.d. random variable with density $g(y) = \int f(y|\mu) dF(\mu)$, then the posterior distribution of F given $y_1, \ldots, y_n$ is a mixture of Dirichlet processes:

$$F \mid y_1, \ldots, y_n \sim \int \ldots \int \mathcal{D}(\alpha G_0 + \sum_{i=1}^{n} \delta_{\mu_i}) dH(\mu_1, \ldots, \mu_n | y_1, \ldots, y_n)$$

The above posterior seems rather intractible, however with the help of the Pólya urn representation of the Dirichlet process by Blackwell and MacQueen (1973), one obtains a simplified form of the posterior distribution, (see, for example, Lo (1984)). We will see that this representation opens the way for MCMC sampling from the posterior distribution.

The Pólya urn representation provides the following conditional distribution,

$$\mu_i|\mu_1,\ldots,\mu_{i-1} \sim \frac{\alpha}{\alpha+i-1}G_0(\mu_i) + \frac{1}{\alpha+i-1}\sum_{j=1}^{i-1}\delta_{\mu_j}(\mu_i) \tag{5}$$

where δ_{μ_j} denotes the (Dirac) distribution with mass one at point μ_j. In the Pólya urn scheme each new draw μ_i is a random draw from the base distribution G_0 with probability $\frac{\alpha}{\alpha+i-1}$ and otherwise take the value of the previous $i-1$ μ's, each with probability $1/(\alpha+i-1)$. The joint distribution is therefore,

$$p(\mu_1,\ldots,\mu_n) = \prod_{i=1}^{n}\frac{\alpha G_0(\mu_i)+\sum_{j=1}^{i-1}\delta_{\mu_j}(\mu_i)}{\alpha+i-1},$$

and the posterior distribution of $(\mu_1,\ldots,\mu_n|y)$ can thus be written as,

$$p(\mu_1,\ldots,\mu_n|y) \propto \prod_{i=1}^{n} f(y_i|\mu_i)\frac{\alpha G_0(\mu_i)+\sum_{j=1}^{i-1}\delta_{\mu_j}(\mu_i)}{\alpha+i-1}. \tag{6}$$

Note that we have integrated out the infinite dimensional object F, the distribution for μ,

A.1. **Gibbs Sampling I.** Our aim is to sample from the above posterior distribution. This can be accomplished by simulating a Markov chain that has the posterior as its stationary equilibrium distribution. The Gibbs sampling algorithm in Escobar and West (1994) provides a simple way to do this. It relies on the simple form of the posterior distribution of μ_i conditional on the remaining parameters $\mu_{(-i)} = (\mu_1,\ldots,\mu_{i-1},\mu_{i+1},\ldots,\mu_n)$ and data y such that we only need to track one variable at each step instead of all n of them. The easiest implementation chooses G_0 as the conjugate prior. For the binomial setting, this leads us to choose G_0 as a beta distribution with density $\mathcal{B}(a,b)$.

We can write the posterior distribution as,

$$\begin{aligned}
p(\mu_i|\mu_{(-i)},y) &= \frac{f(y_i|\mu_1,\ldots,\mu_n)\cdot p(\mu_i|\mu_{(-i)})}{\int f(y_i|\mu_1,\ldots,\mu_n)\cdot p(\mu_i|\mu_{(-i)})\,d\mu_i} \\
&\propto \frac{\alpha}{\alpha+n-1}\mathcal{B}(a,b)f(y_i|\mu_i) + \frac{1}{\alpha+n-1}\sum_{j\neq i} f(y_i|\mu_j)\delta_{\mu_j} \\
&\propto \frac{\alpha}{\alpha+n-1}\frac{B(y_i+a,l_i-y_i+b)}{B(a,b)}\mathcal{B}(y_i+a,l_i-y_i+b) \\
&+ \frac{1}{\alpha+n-1}\sum_{j\neq i}\mu_j^{y_i}(1-\mu_j)^{l_i-y_i}\delta_{\mu_j}
\end{aligned} \tag{7}$$

where $B(\cdot,\cdot)$ denotes the Beta function, $\mathcal{B}(\cdot,\cdot)$ denotes the beta density and $f(y_i|\mu_i)$ is the binomial density. Note that the posterior distribution only depends on y_i because for $j \neq i$, y_j is conditionally independent of μ_i given μ_j's.

Gibbs sampling initiates the Markov chain by sampling $(\mu_1^{(0)},\ldots,\mu_n^{(0)})$ from $\mathcal{B}(y_i+a,l_i-y_i+b)$ (the posterior obtained by updating the prior G_0 via the likelihood $f(y_i|\mu_i)$ up to a normalization constant). The first step of the Markov Chain is defined as:

Sample $\mu_1^{(1)}$ from $p(\mu_1|\mu_2^{(0)},\ldots,\mu_n^{(0)},y)$.

Sample $\mu_2^{(1)}$ from $p(\mu_2|\mu_1^{(1)},\mu_3^{(0)},\ldots,\mu_n^{(0)},y)$.

...

Sample $\mu_n^{(1)}$ from $p(\mu_n|\mu_1^{(1)},\ldots,\mu_{n-1}^{(1)},y)$.

Continuing the iteration, the chain stabilizes at its equilibrium distribution and the resulting sample constitute draws from the distribution with density function $p(\mu_1,\ldots,\mu_n|y)$ as in (6).

A.2. **Gibbs Sampling II.** Gibbs sampling as above may have rather slow convergence because of the clustering behavior of the μ_i's (as already noted by Antoniak (1974)). As discussed in Neal (2000), since the Gibbs sampling algorithm implemented above can not change μ for more than one observation at a time, changes of μ values for observations in the same cluster occur rather rarely, leading to rigidity in the chain and hence slow convergence. This can be avoided by introducing a latent class variable as implemented in Bush and MacEachern (1996) and West, Müller, and Escobar (1994). It is also the approach adopted by the DPbetabinom function in the R package DPpackage that we employ.

The latent class model is equivalent to the model introduced above, except that we specify a configuration variable $\mathcal{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_n)$ that classifies the μ_i's into k distinct clusters. Denoting $n_j = \#\{\mathcal{S}_i = j\}$ for $j = 1, \ldots, k$, the distinct values of μ form the set $\{\theta_1, \ldots, \theta_k\}$ and we set $I_j = \{i : \mathcal{S}_i = j, i = 1, \ldots, n\}$. As discussed in Antoniak (1974) and West (1990), the θ_i are random samples from G_0 and k is related to the sample size n and the precision parameter α. Given k, the μ_i's are selected from the set of θ according to a multinomial distribution. With the above described structure, we have the distribution for μ_i conditional on $\mu_{(-i)}$ as,

$$p(\mu_i \mid \mu_{(-i)}, \mathcal{S}_{(-i)}, k_{(-i)}) \sim \frac{\alpha}{\alpha + n - 1} G_0 + \frac{1}{\alpha + n - 1} \sum_{j=1}^{k_{(-i)}} n_j^{(-i)} \delta_{\theta_j^{(-i)}}$$

where $\mathcal{S}_{(-i)}$ is the configuration corresponding to all entries in $\mu_{(-i)}$ and $k_{(-i)}$ is the number of distinctive values of $\mu_{(-i)}$, contained in the set $\theta^{(-i)}$ and $n_j^{(-i)} = \#\{\mathcal{S}_{(-i)} = j\}$ for $j = 1, \ldots, k_{(-i)}$. We still have a Pólya urn scheme: with probability $\alpha/(\alpha + n - 1)$, μ_i is a random draw from G_0, and otherwise takes a value from the set $\theta^{(-i)}$ with probability proportional to the multinomial counts $n_j^{(-i)}$.

Choosing G_0 to be the conjugate prior $\mathcal{B}(a, b)$, the posterior distribution of μ_i conditional on data y and $\mu_{(-i)}$ is thus,

$$\begin{aligned} p(\mu_i \mid \mu_{(-i)}, \mathcal{S}_{(-i)}, k_{(-i)}, y) &\propto \tfrac{\alpha}{\alpha+n-1} f(y_i|\mu_i)\mathcal{B}(a, b) + \tfrac{1}{\alpha+n-1} \textstyle\sum_{j=1}^{k_{(-i)}} n_j^{(-i)} f(y_i|\theta_j^{(-i)}) \delta_{\theta_j^{(-i)}} \\ &= q_{i,0}\mathcal{B}(y_i + a, l_i - y_i + b) + \textstyle\sum_{j=1}^{k_{(-i)}} q_{i,j} \delta_{\theta_j^{(-i)}} \end{aligned}$$

with

$$q_{i,j} = \begin{cases} c \frac{\alpha}{\alpha+n-1} \frac{B(y_i+a, l_i-y_i+b)}{B(a,b)} & \text{if } j = 0 \\ c \frac{n_j^{(-i)}}{\alpha+n-1} (\theta_j^{(-i)})^{y_i} (1 - \theta_j^{(-i)})^{l_i - y_i} & \text{if } j > 0 \end{cases}$$

and normalizating constant,

$$c^{-1} = \frac{\alpha}{\alpha + n - 1} \frac{B(y_i + a, l_i - y_i + b)}{B(a, b)} + \frac{1}{\alpha + n - 1} \sum_{j=1}^{k_{(-i)}} n_i^{(-i)} (\theta_j^{(-i)})^{y_i} (1 - \theta_j^{(-i)})^{l_i - y_i}.$$

This second Gibbs sampling algorithm differs from the previous one in that instead of iteratively updating μ_i's, we update the configuration variable $\mathcal{S}$ according to its posterior distribution, that is,

$$P(\mathcal{S}_i = j \mid y, \mu_{(-i)}, \mathcal{S}_{(-i)}, k_{(-i)}) = q_{i,j}.$$

We initiate the chain by choosing $\mathcal{S}_i^{(0)} = i$ (i.e., each observation forms its own cluster) and $\theta^{(0)}$'s can be drawn from $\mathcal{B}(y_i + a, l_i - y_i + b)$. The Markov chain is simulated as,

(1) Given values for $\theta^{(0)}$ and $\mathcal{S}^{(0)}$, generate a new $\mathcal{S}^{(1)}$ according to posterior distribution specified above successively. For any index i that has $\mathcal{S}_i^{(1)} = 0$, draw a new μ_i from $\mathcal{B}(y_i + a, l_i - y_i + b)$. Count the number of clusters $k^{(1)}$.

(2) Given $k^{(1)}$ and $\mathcal{S}^{(1)}$, generate a new set of $\theta^{(1)}$ by sampling from

$$p(\theta_j | y, \mathcal{S}^{(1)}, k^{(1)}) \propto \prod_{r \in I_j^{(1)}} f(y_r|\theta_j) dG_0(\theta_j)$$

(3) Continue iterating ...

After discarding the first few steps of the Markov chain (burn-in), we can use the simulated sample for the following posterior analysis.

A.3. **Posterior Analysis.** Given the posterior distribution (7), the posterior mean for μ_i from the m^{th} step of the MCMC scan is

$$\begin{aligned} \mathcal{E}_i^{(m)} = \mathbb{E}(\mu_i^{(m)}|y, \mu_{(-i)}^{(m)}) &= \int \mu_i p(\mu_i|\mu_{(-i)}, y) d\mu_i \\ &= \frac{\alpha \frac{B(y_i+a+1, l_i-y_i+b)}{B(a,b)} + \sum_{j \neq i} (\mu_j^{(m)})^{y_i+1} (1 - \mu_j^{(m)})^{l_i - y_i}}{\alpha \frac{B(y_i+a, l_i-y_i+b)}{B(a,b)} + \sum_{j \neq i} (\mu_j^{(m)})^{y_i} (1 - \mu_j^{(m)})^{l_i - y_i}} \end{aligned} \tag{8}$$

The posterior mean of μ_i given data using the entire chain is thus estimated by,

$$\frac{1}{M} \sum_{m=1}^{M} \mathcal{E}_i^{(m)},$$

where M is the total number of MCMC scans after initial burn-in. As noted by (Gelfand and Smith 1990) this is essentially Rao-Blackwellization. The posterior variance can be found accordingly as,

$$\mathbb{V}(\mu_i|y) = \mathbb{E}(\mathbb{V}(\mu_i|\mu^{(-i)}, y)) + \mathbb{V}(\mathbb{E}(\mu_i|\mu^{(-i)}, y)).$$

A.4. **Predictive distribution.** Nonparametric Bayesian analysis often focuses on the predictive distribution of μ_{n+1} or y_{n+1} for a future experiment. It is also something routinely reported by software packages. As noted by Liu (1996), since F is an infinite-dimensional object, there is no easy way of exploring its full posterior distribution. However, we can look at its posterior mean, $\mathbb{E}(F|y)$, which also turns out to be the predictive distribution of a future μ_{n+1}. In particular, given the latent class model , the predictive density for a future μ in the m^{th} round of Markov Chain simulation, evaluated at grid points on the domain of $(0, 1)$, is,

$$\begin{aligned} p(\mu_{n+1}^{(m)} \mid y, s^{(m)}, k^{(m)}) &= \mathbb{E}(F \mid y_1, \ldots, y_n) \\ &= \tfrac{\alpha}{\alpha+n} G_0(\cdot) + \tfrac{1}{\alpha+n} \textstyle\sum_{j=1}^{k^{(m)}} n_j^{(m)} \mathcal{B}(\cdot; Z_j^{(m)} + a, L_j^{(m)} - Z_j^{(m)} + b) \end{aligned}$$

with $Z_j^{(m)} = \sum_{r \in I_j^{(m)}} y_r$ and $L_j^{(m)} = \sum_{r \in I_j^{(m)}} l_r$. The predictive distribution is then obtained as,

$$\frac{1}{M} \sum_{m=1}^{M} p(\mu_{n+1}^{(m)} \mid y, s^{(m)}, k^{(m)}).$$

Given this predictive density we can compute the Bayes rule for predicting μ given an observed y as in the binomial Kiefer Wolfowitz setting,

$$\mathbb{E}(\mu \mid y) = \frac{\int \mu f(y \mid \mu) \hat{f}(\mu) d\mu}{\int f(y \mid \mu) \hat{f}(\mu) d\mu}$$

The predictive density depends on the Dirichlet prior $\mathcal{D}(\alpha G_0)$. In the application, we take the distribution G_0 to be $\mathcal{B}(1, 1)$, i.e., the uniform distribution on the interval $[0, 1]$. The precision parameter α takes values 10 and 0.01. The bigger α is, the more confidence we have in G_0 The closer α is to zero, for a given n, the closer the predictive distribution of μ is to the mixing distribution estimated by the Kiefer-Wolfowitz MLE.