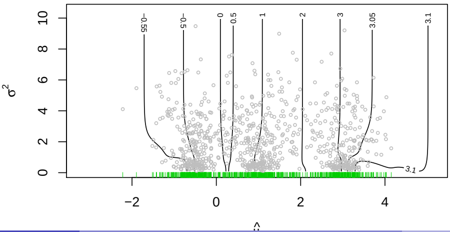
Unobserved Heterogeneity in Longitudinal Data
An Empirical Bayes Perspective

Roger Koenker

University of Illinois, Urbana-Champaign

CeMMaP: 18 June 2013

Joint work with Jiaying Gu (UIUC)



An Empirical Bayes Homework Problem

Suppose you observe a sample $\{Y_1, ..., Y_n\}$ and $Y_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, ..., n$, and would like to estimate all of the μ_i's under squared error loss. We might call this "incidental parameters with a vengence."

An Empirical Bayes Homework Problem

Suppose you observe a sample $\{Y_1, ..., Y_n\}$ and $Y_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, ..., n$, and would like to estimate all of the μ_i's under squared error loss. We might call this "incidental parameters with a vengence."

- Not knowing any better, we assume that the μ_i are drawn iid-ly from a distribution F so the Y_i have density,

$$g(y) = \int \phi(y - \mu) dF(\mu),$$

  the Bayes rule is then given by Tweedie's formula:

$$\delta(y) = y + \frac{g'(y)}{g(y)}$$

An Empirical Bayes Homework Problem

Suppose you observe a sample $\{Y_1, ..., Y_n\}$ and $Y_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, ..., n$, and would like to estimate all of the μ_i's under squared error loss. We might call this "incidental parameters with a vengence."

- Not knowing any better, we assume that the μ_i are drawn iid-ly from a distribution F so the Y_i have density,

  $$g(y) = \int \phi(y - \mu) dF(\mu),$$

  the Bayes rule is then given by Tweedie's formula:

  $$\delta(y) = y + \frac{g'(y)}{g(y)}$$

- When F is unknown, one can try to estimate g and plug it into the Bayes rule.

Stein Rules I

Suppose 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,

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

Stein Rules I

Suppose 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,

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

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 original form:

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

Stein Rules II

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

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

Stein Rules II

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

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

Estimating the prior mean parameter costs us one more degree of freedom, and we obtain the celebrated James-Stein (1960) estimator,

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

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

Nonparametric Empirical Bayes Rules

Brown and Greenshtein (Annals, 2009) propose estimating g by standard fixed bandwidth kernel methods and they compare performance of their *estimated* Bayes rule with various other methods including the various parametric empirical Bayes methods investigated by Johnstone and Silverman in their "Needles and Haystacks" (Annals) paper.

Nonparametric Empirical Bayes Rules

Brown and Greenshtein (Annals, 2009) propose estimating g by standard fixed bandwidth kernel methods and they compare performance of their *estimated* Bayes rule with various other methods including the various parametric empirical Bayes methods investigated by Johnstone and Silverman in their "Needles and Haystacks" (Annals) paper.

A drawback of the kernel approach is that it fails to impose a monotonicity constraint that should hold for the Gaussian problem, or indeed for any similar problem in which we have iid observations from a mixture density,

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

with φ an exponential family density with natural parameter $\theta \in \mathbb{R}$.

Back to the Homework

When φ is an exponential family density we may write,

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

Quadratic loss implies that the Bayes rule is a conditional mean:

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

Monotonicity of the Bayes Rule

When φ is of the exponential family form,

$$\delta_G'(y) = \frac{d}{dy}\left[\frac{\int \theta\varphi dF}{\int \varphi dF}\right] = \frac{\int \theta^2\varphi dF}{\int \varphi dF} - \left(\frac{\int \theta\varphi dF}{\int \varphi dF}\right)^2$$
$$= \mathbb{E}[\Theta^2|Y=y] - (\mathbb{E}[\Theta|Y=y])^2$$
$$= \mathbb{V}[\Theta|Y=y] \geqslant 0,$$

implying that δ_G must be monotone.

Monotonicity of the Bayes Rule

When φ is of the exponential family form,

$$\delta'_G(y) = \frac{d}{dy}\left[\frac{\int \theta\varphi dF}{\int \varphi dF}\right] = \frac{\int \theta^2 \varphi dF}{\int \varphi dF} - \left(\frac{\int \theta\varphi dF}{\int \varphi dF}\right)^2$$
$$= \mathbb{E}[\Theta^2 | Y = y] - (\mathbb{E}[\Theta | Y = y])^2$$
$$= \mathbb{V}[\Theta | Y = y] \geqslant 0,$$

implying that δ_G must be monotone. Or equivalently that,

$$K(y) = \log \hat{g}(y) - \log m(y)$$

is convex. Such problems are closely related to recent work on estimating log-concave densities, e.g. Cule, Samworth and Stewart (JRSSB, 2010), Koenker and Mizera (Annals, 2010), Seregin and Wellner (Annals, 2010), Dümbgen, Samworth and Schuhmacher (Annals, 2011).

Standard Gaussian Case

In our homework problem,

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

So $m(y) = e^{-y^2/2}$ and the logarithmic derivative yields our Bayes rule:

$$\delta_G(y) = \frac{d}{dy}\left[\frac{1}{2}y^2 + \log g(y)\right] = y + \frac{g'(y)}{g(y)}.$$

Estimating g by maximum likelihood subject to the constraint that

$$K(y) = \frac{1}{2}y^2 + \log \hat{g}(y)$$

is convex as discussed in Koenker and Mizera (2013).

Nonparametric MLE

Kiefer and Wolfowitz (1956) reconsidering the Neyman and Scott (1948) problem showed that non-parametric maximum likelihood could be used to establish consistent estimators even when the number of incidental parameters tended to infinity. Laird (1978) and Heckman and Singer (1984) suggested that the EM algorithm could be used to compute the MLE in such cases.

Nonparametric MLE

Kiefer and Wolfowitz (1956) reconsidering the Neyman and Scott (1948) problem showed that non-parametric maximum likelihood could be used to establish consistent estimators even when the number of incidental parameters tended to infinity. Laird (1978) and Heckman and Singer (1984) suggested that the EM algorithm could be used to compute the MLE in such cases.

Jiang and Zhang (Annals, 2009) adapt this approach for the empirical Bayes problem: Let $u_i : i = 1, \ldots, m$ denote a grid on the support of the sample Y_i's, then the prior (mixing) density f is estimated by the EM fixed point iteration:

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

Nonparametric MLE

Kiefer and Wolfowitz (1956) reconsidering the Neyman and Scott (1948) problem showed that non-parametric maximum likelihood could be used to establish consistent estimators even when the number of incidental parameters tended to infinity. Laird (1978) and Heckman and Singer (1984) suggested that the EM algorithm could be used to compute the MLE in such cases.

Jiang and Zhang (Annals, 2009) adapt this approach for the empirical Bayes problem: Let $u_i : i = 1, ..., m$ denote a grid on the support of the sample Y_i's, then the prior (mixing) density f is estimated by the EM fixed point iteration:

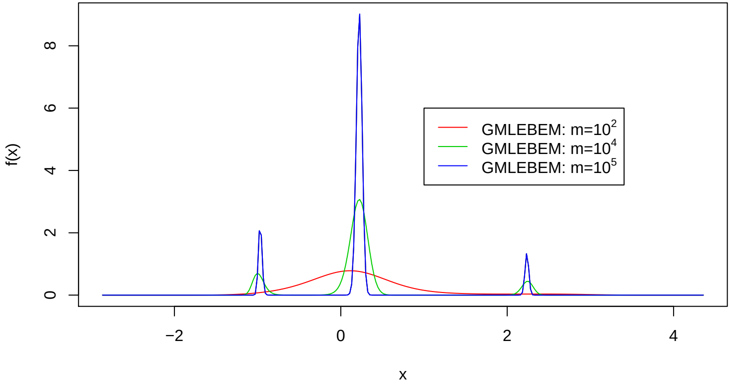
$$\hat{f}_j^{(k+1)} = n^{-1} \sum_{i=1}^{n} \frac{\hat{f}_j^{(k)} \phi(Y_i - u_j)}{\sum_{\ell=1}^{m} \hat{f}_\ell^{(k)} \phi(Y_i - u_\ell)},$$

and the implied Bayes rule becomes at convergence:

$$\hat{\delta}(Y_i) = \frac{\sum_{j=1}^{m} u_j \phi(Y_i - u_j) \hat{f}_j}{\sum_{j=1}^{m} \phi(Y_i - u_j) \hat{f}_j}.$$

The Incredible Lethargy of EM-ing

Unfortunately, EM fixed point iterations are notoriously slow and this is especially apparent in the Kiefer and Wolfowitz setting. Solutions approximate discrete (point mass) distributions, but EM goes ever so slowly. (Approximation is controlled by the grid spacing of the u_i's.)



Accelerating EM

There is a large literature on accelerating EM iterations, but none of the recent developments seem to help very much. However, the Kiefer-Wolfowitz problem can be reformulated as a convex maximum likelihood problem and solved by standard interior point methods:

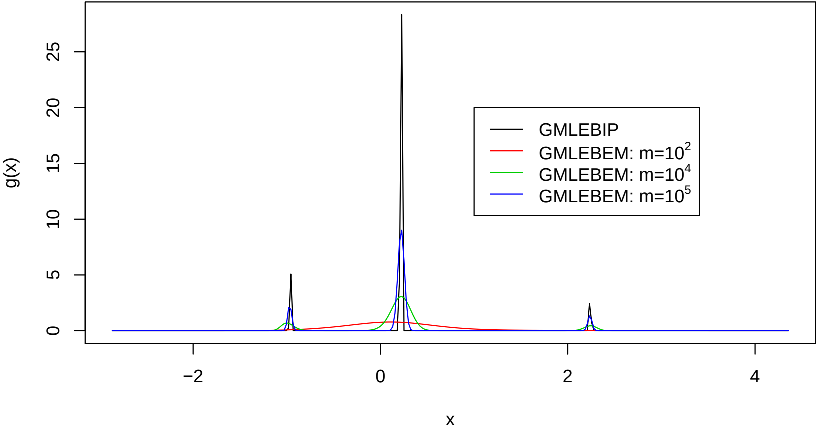
$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{m} \phi(y_i - u_j) f_j\Big),$$

can be rewritten as,

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

where $A = (\phi(y_i - u_j))$ and $\mathcal{S} = \{s \in \mathbb{R}^m | 1^\top s = 1,\ s \geqslant 0\}$. 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.

Interior Point vs. EM



Interior Point vs. EM

In the foregoing test problem we have $n = 200$ observations and $m = 300$ grid points. Timing and accuracy is summarized in this table.

Estimator	EM1	EM2	EM3	IP
Iterations	100	10,000	100,000	15
Time	1	37	559	1
L(g) - 422	0.9332	1.1120	1.1204	1.1213

Comparison of EM and Interior Point Solutions: Iteration counts, log likelihoods and CPU times (in seconds) for three EM variants and the interior point solver.

Scaling problem sizes up, the deficiency of EM is even more serious. Simulation performance of the Bayes Rule is improved over EM implementation.

But How Does It Work in Theory?

For the Gaussian location mixture problem empirical Bayes rules based on the Kiefer-Wolfowitz estimator are adaptively minimax.

Theorem: Jiang and Zhang For the normal location mixture problem, with a (complicated) weak pth moment restriction on Θ, the approximate non-parametric MLE, $\hat{\theta} = \hat{\delta}_{\hat{F}_n}(Y)$ is adaptively minimax, i.e.

$$\frac{\sup_{\theta} \mathbb{E}_{n,\theta} L_n(\hat{\theta}, \theta)}{\inf_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{n,\theta} L_n(\tilde{\theta}, \theta)} \to 1.$$

The weak pth moment condition encompasses a much broader class of both deterministic and stochastic classes Θ.

Gaussian Mixtures with Longitudinal Data

Model:

$$y_{it} = \mu_i + \sqrt{\theta_i} u_{it}, \; t = 1, \cdots, m_i, \; 1, \cdots, n, \; u_{it} \sim \mathcal{N}(0, 1)$$

Gaussian Mixtures with Longitudinal Data

Model:

$$y_{it} = \mu_i + \sqrt{\theta_i} u_{it}, \; t = 1, \cdots, m_i, \; 1, \cdots, n, \; u_{it} \sim \mathcal{N}(0, 1)$$

Sufficient Statistics:

$$\hat{\mu}_i = m_i^{-1} \sum_{t=1}^{m_i} y_{it} \sim \mathcal{N}(\mu_i, \theta_i / m_i)$$

$$\hat{\theta}_i = (m_i - 1)^{-1} \sum_{t=1}^{m_i} (y_{it} - \hat{\mu}_i)^2 \sim \Gamma(r_i, \theta_i / r_i), \; r_i = (m_i - 1)/2$$

Gaussian Mixtures with Longitudinal Data

Model:

$$y_{it} = \mu_i + \sqrt{\theta_i} u_{it}, \; t = 1, \cdots, m_i, \; 1, \cdots, n, \; u_{it} \sim \mathcal{N}(0, 1)$$

Sufficient Statistics:

$$\hat{\mu}_i = m_i^{-1} \sum_{t=1}^{m_i} y_{it} \sim \mathcal{N}(\mu_i, \theta_i / m_i)$$

$$\hat{\theta}_i = (m_i - 1)^{-1} \sum_{t=1}^{m_i} (y_{it} - \hat{\mu}_i)^2 \sim \Gamma(r_i, \theta_i / r_i), \; r_i = (m_i - 1)/2$$

Likelihood

$$L(F|y) = \prod_{i=1}^{n} \int \int \phi((\hat{\mu}_i - \mu_i)/\sqrt{\theta m_i}) / \sqrt{\theta_i m_i} \gamma(\hat{\theta}_i | r_i, \theta_i / r_i) dF_{\mu}(\mu) dF_{\theta}(\theta)$$

A Toy Example

Model

$$y_{it} = \mu_i + \sqrt{\theta_i} u_{it},\ t = 1, \cdots, m_i,\ 1, \cdots, n,\ u_{it} \sim \mathcal{N}(0, 1)$$

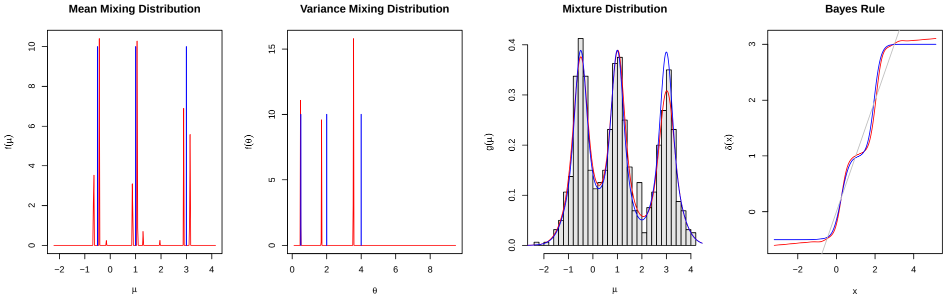
$$\mu_i \sim \tfrac{1}{3}(\delta_{-0.5} + \delta_1 + \delta_3) \perp\!\!\!\perp \theta_i \sim \tfrac{1}{3}(\delta_{0.5} + \delta_2 + \delta_4)$$

A Toy Example

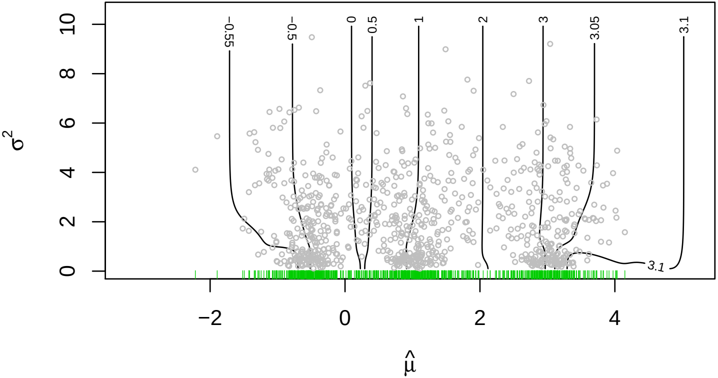
Model

$$y_{it} = \mu_i + \sqrt{\theta_i} u_{it}, \; t = 1, \cdots, m_i, \; 1, \cdots, n, \; u_{it} \sim \mathcal{N}(0,1)$$

$$\mu_i \sim \tfrac{1}{3}(\delta_{-0.5} + \delta_1 + \delta_3) \perp\!\!\!\perp \theta_i \sim \tfrac{1}{3}(\delta_{0.5} + \delta_2 + \delta_4)$$



Contour Plot for Joint Bayes Rule: $\delta(\hat{\mu}, \hat{\theta}) = \mathbb{E}(\mu \mid \hat{\mu}, \hat{\theta})$



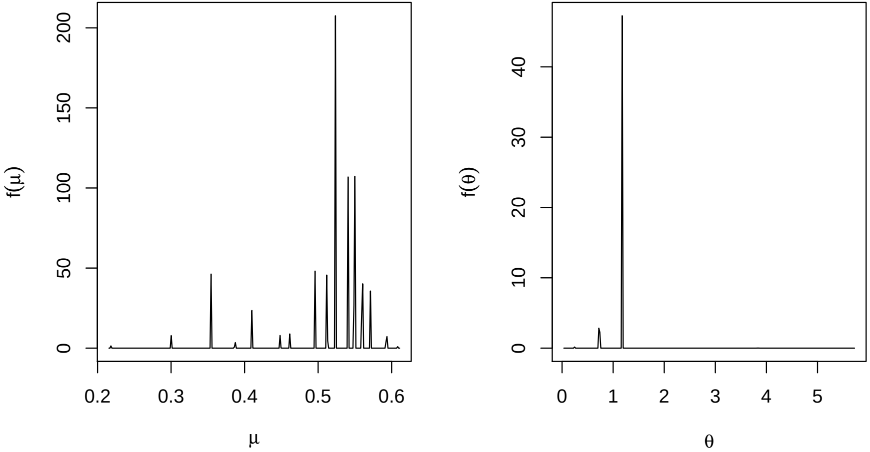
Empirical Bayesball

Using (ESPN) data we have constructed an unbalanced panel, 10,575 observations, on 1072 players from 2002-2011. Following standard practice, Brown (2009, AoAS) and Jiang and Zhang (2010, Brown Festschrift) we transform batting averages to (approximate) normality:

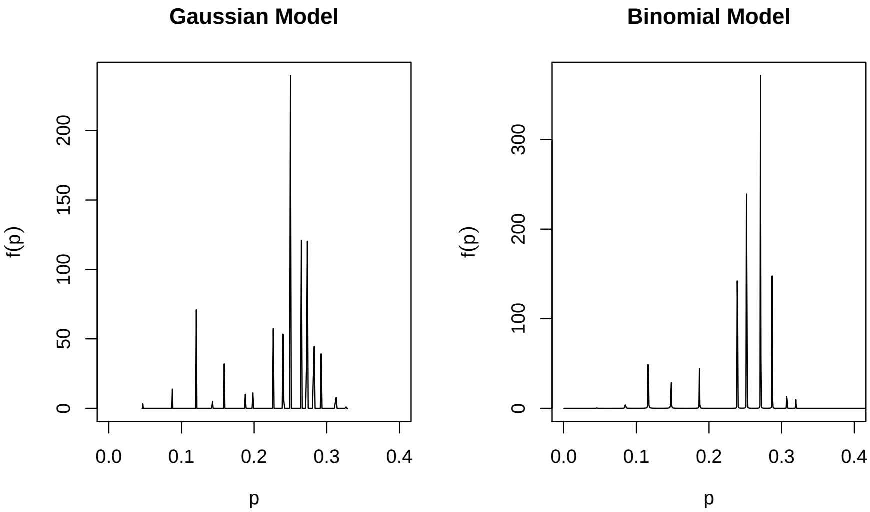
$$\hat{Y}_i = \text{asin}\left(\sqrt{\frac{H_{i1} + 1/4}{N_{i1} + 1/2}}\right) \sim \mathcal{N}(\text{asin}(\sqrt{\rho}), 1/(4N_{i1}))$$

Treating these observations as approximately Gaussian, we compute sample means and variances for each player through 2011, and estimate our independent prior model.

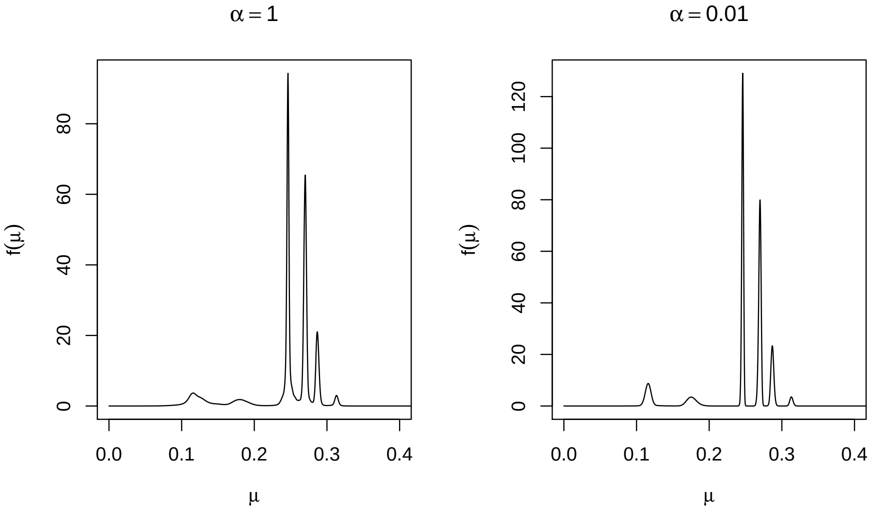
Prior Estimates on the Gaussian Scale



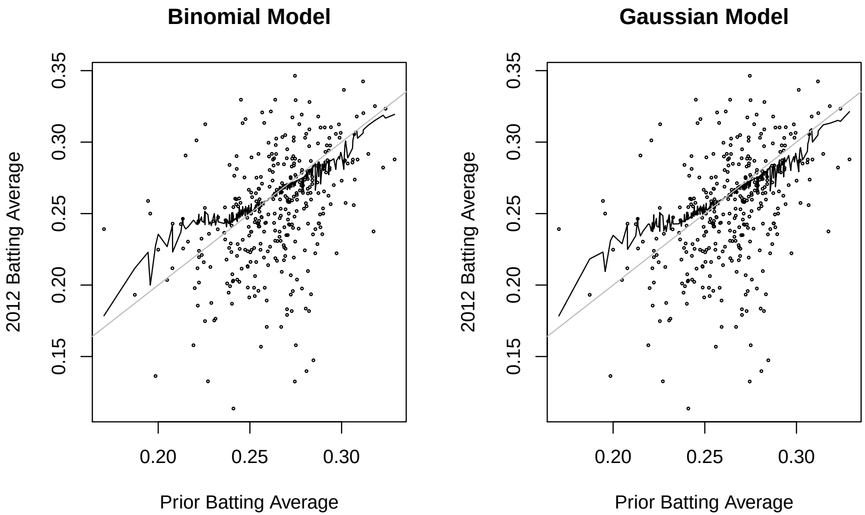
Prior Estimates on the Batting Average Scale



Dirichlet Prior Estimates on the Batting Average Scale



Bayes Rule Predictions



Covariate Effects

The location-scale mixture model is really just a starting point for more general panel data models with covariate effects and unobserved heterogeneity estimable by profile likelihood. Given the model,

$$y_{it} = x_{it}\beta + \alpha_i + \sigma_i u_{it},$$

and a fixed $\beta \in \mathbb{R}^p$, we have sufficient statistics $\bar{y}_i - \bar{x}_i\beta$, for α_i and

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

for σ_i^2. Clearly, $\bar{y}_i|\alpha_i, \beta, \sigma_i^2 \sim \mathcal{N}(\alpha_i + \bar{x}_i\beta, \sigma_i^2)$ and $S_i|\beta, \sigma_i^2 \sim \Gamma(r_i, \sigma_i^2/r_i)$, where, $r_i = (m_i - 1)/2$.

Profile Likelihood for Covariate Effects

Reducing the likelihood to sufficient statistics we have (almost) a decomposition in terms of “within” and “between” information:

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

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

Profile Likelihood for Covariate Effects

Reducing the likelihood to sufficient statistics we have (almost) a decomposition in terms of "within" and "between" information:

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

where $R_i = r_i S_i/\sigma_i^2$, $r_i = (m_i - 1)/2$, and $K = \prod_{i=1}^{n} \left(\frac{\Gamma(r_i)}{r_i^{r_i}} (1/\sqrt{2\pi})^{m_i - 1} \right)$. But note that the likelihood doesn't factor so the between and within information isn't independent.

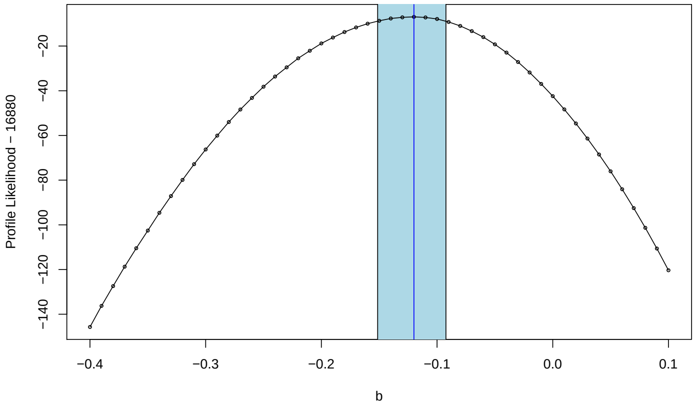
Age and Batting Ability

There is considerable controversy about the relationship between player's age and their batting ability. To explore this we collected (reported) birth years for each of the players and reestimated the model including both linear and quadratic age effects using the profile likelihood method.

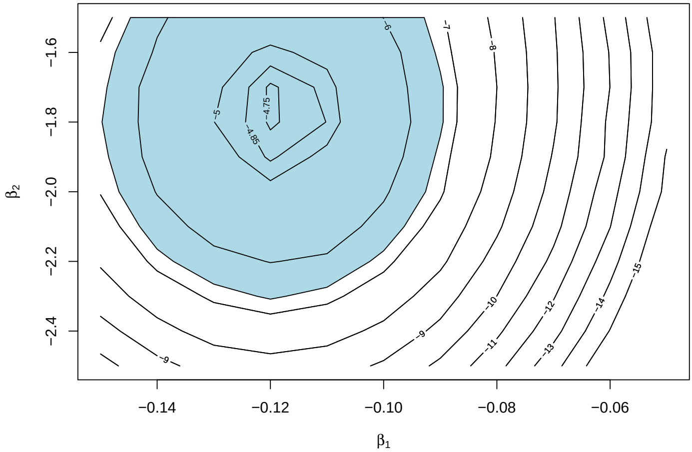
We evaluate the profile likelihood on a grid of parameter values, but as you will see the likelihood is quite smooth and well behaved so higher dimensional problems could be done with standard optimization software.

Evaluations of the profile likelihood are quick, a few seconds for our application, with grids of a few hundred points for the mixing distributions.

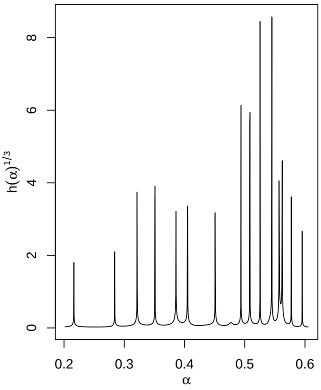
Profile Likelihood for the Linear Age Effect



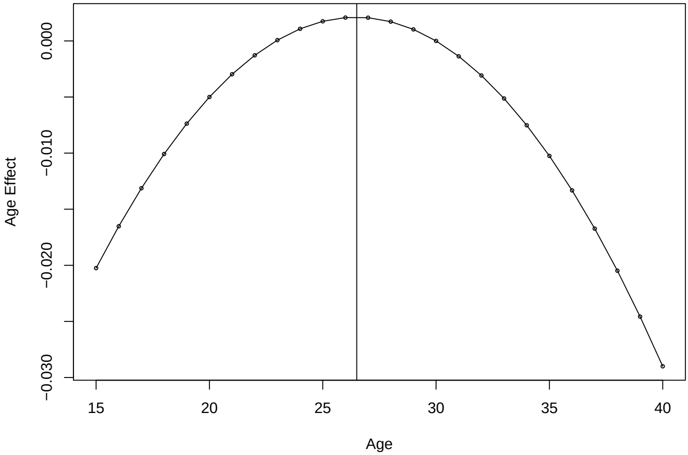
Contour Plot of the Quadratic Age Effect



The Mixing Densities at the Profile MLE



The Estimated Quadratic Age Effect



Conclusions and Extrapolations

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,

Conclusions and Extrapolations

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,
- Kernel based empirical Bayes rules can be improved with shape constrained MLEs and are computationally very efficient, but

Conclusions and Extrapolations

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,
- Kernel based empirical Bayes rules can be improved with shape constrained MLEs and are computationally very efficient, but
- Kiefer-Wolfowitz type non-parametric MLEs perform even better.

Conclusions and Extrapolations

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,
- Kernel based empirical Bayes rules can be improved with shape constrained MLEs and are computationally very efficient, but
- Kiefer-Wolfowitz type non-parametric MLEs perform even better.
- There are many opportunities for linking such methods to various semi-parametric estimation problems a la Heckman and Singer (1983) and van der Vaart (1996) as for the baseball problem,

Conclusions and Extrapolations

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,
- Kernel based empirical Bayes rules can be improved with shape constrained MLEs and are computationally very efficient, but
- Kiefer-Wolfowitz type non-parametric MLEs perform even better.
- There are many opportunities for linking such methods to various semi-parametric estimation problems a la Heckman and Singer (1983) and van der Vaart (1996) as for the baseball problem,
- It is all downhill after 27 in math and baseball,

Conclusions and Extrapolations

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,
- Kernel based empirical Bayes rules can be improved with shape constrained MLEs and are computationally very efficient, but
- Kiefer-Wolfowitz type non-parametric MLEs perform even better.
- There are many opportunities for linking such methods to various semi-parametric estimation problems a la Heckman and Singer (1983) and van der Vaart (1996) as for the baseball problem,
- It is all downhill after 27 in math and baseball,
- Be cautious about predicting baseball batting averages, or anything else about baseball.