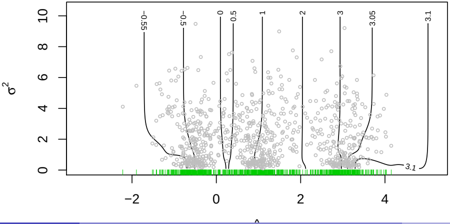
Unobserved Heterogeneity in Longitudinal Data
An Empirical Bayes Perspective

Roger Koenker

University of Illinois, Urbana-Champaign

University of Tokyo: 25 November 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,

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

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,

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

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 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 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,

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

and φ is 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, 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).

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 is 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 $u_i : i = 1, ..., m$ 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:

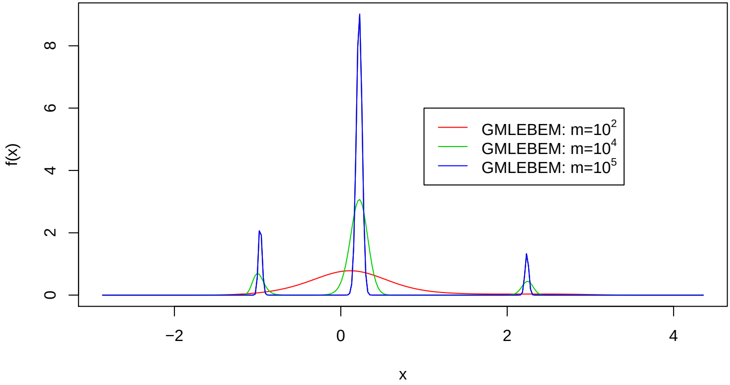
$$\hat{f}_j^{(k+1)} = n^{-1} \sum_{i=1}^{n} \frac{\hat{f}_j^{(k)} \phi(Y_i - u_j)}{\sum_{\ell=1}^{m} \hat{f}_\ell^{(k)} \phi(Y_i - u_\ell)},$$

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:

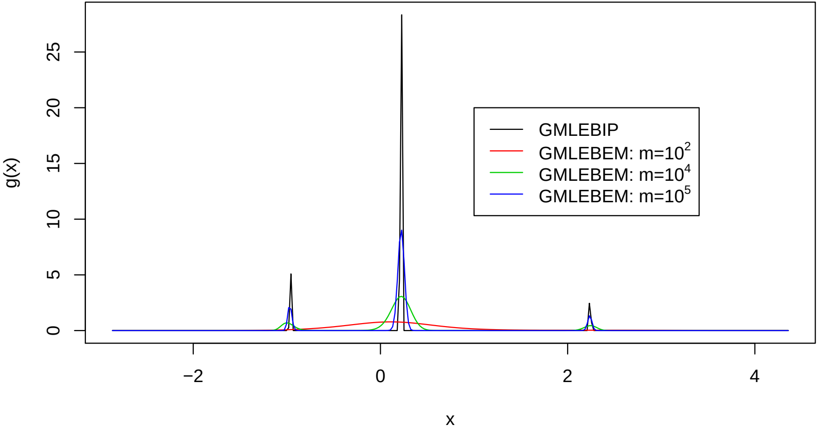
$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} \phi(y_i - u_j) f_j\right),$$

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 the EM approach is even more serious.

Johnstone and Silverman Simulation Design

Data is generated from 12 distinct models, all of the form:

$$Y_i = \mu_i + u_i, \quad u_i \sim \mathcal{N}(0, 1), \quad i = 1, ..., 1000.$$

Of the $n = 1000$ observations $n - k$ of the $\mu_i = 0$, and the remaining k take one of the four values $\{3, 4, 5, 7\}$. There are three choices of k: $\{5, 50, 500\}$. There are 50 replications for each of the 12 experimental settings and 18 different competing estimators.

Johnstone and Silverman Simulation Design

Data is generated from 12 distinct models, all of the form:

$$Y_i = \mu_i + u_i, \quad u_i \sim \mathcal{N}(0, 1), \quad i = 1, ..., 1000.$$

Of the $n = 1000$ observations $n - k$ of the $\mu_i = 0$, and the remaining k take one of the four values $\{3, 4, 5, 7\}$. There are three choices of k: $\{5, 50, 500\}$. There are 50 replications for each of the 12 experimental settings and 18 different competing estimators.

Performance is measured by the mean (over replications) of the sum (over the $n = 1000$ observations) of squared errors, so a score of 500 means that the mean squared prediction error is 0.5, or half of what the naïve prediction $\hat{\mu}_i = Y_i$ would yield if the μ_i were all zero.

Johnstone and Silverman Simulation Results

Number nonzero	5				50				500			
Value nonzero	3	4	5	7	3	4	5	7	3	4	5	7
Exponential	36	32	17	8	214	156	101	73	857	873	783	658
Cauchy	37	36	18	8	271	176	103	77	922	898	829	743
Postmean	34	32	21	11	201	169	122	85	860	888	826	708
Exphard	51	43	22	11	273	189	130	91	998	998	983	817
$a = 1$	36	32	19	15	213	166	142	135	994	1099	1126	1130
$a = 0.5$	37	34	17	10	244	158	105	92	845	878	884	884
$a = 0.2$	38	37	18	7	299	188	95	69	1061	730	665	656
$a = 0.1$	38	37	18	6	339	227	102	60	1496	798	609	579
SURE	38	42	42	43	202	209	210	210	829	835	835	835
Adapt	42	63	73	76	417	620	210	210	829	835	835	835
FDR $q = 0.01$	43	51	26	5	392	299	125	55	2568	1332	656	524
FDR $q = 0.1$	40	35	19	13	280	175	113	102	1149	744	651	644
FDR $q = 0.4$	58	58	53	52	298	265	256	254	919	866	860	860
BlockThresh	46	72	72	31	444	635	600	299	1918	1276	1065	983
NeighBlock	47	64	51	26	427	543	439	227	1870	1384	1148	972
NeighCoeff	55	51	38	32	375	343	219	156	1890	1410	1032	870
Universal soft	42	63	73	76	417	620	720	746	4156	6168	7157	7413
Universal hard	39	37	18	7	370	340	163	52	3672	3355	1578	505

Performance of the NP-MLE Bayes Rule

In the (now familiar) Johnstone and Silverman sweepstakes we have the following comparison of performance.

Estimator	k = 5				k = 50				k = 500			
	3	4	5	7	3	4	5	7	3	4	5	7
$\hat{\delta}_{MLE-IP}$	33	30	16	8	153	107	51	11	454	276	127	18
$\hat{\delta}_{MLE-EM}$	37	33	21	11	162	111	56	14	458	285	130	18
$\hat{\delta}$	37	34	21	11	173	121	63	16	488	310	145	22
$\tilde{\delta}_{1.15}$	53	49	42	27	179	136	81	40	484	302	158	48
J-S Min	34	32	17	7	201	156	95	52	829	730	609	505

Here MLE-EM is Jaing and Zhang's (2009) Bayes rule with their suggested 100 EM iterations. It does somewhat better than the shape constrained estimator, but the interior point version MLE-IP does even better.

The Castillo and van der Vaart Experiment

The setup is quite similar to the first earlier ones,

$$Y_i = \theta_i + u_i, i = 1, \cdots n$$

the θ_i are most zero, but s of them take one of the values from the set $\{1, 2, \cdots, 5\}$. The sample size is $n = 500$, and $s \in \{25, 50, 100\}$ and θ_a takes five possible values: The first 8 rows of the Table are taken directly from Table 1 of Castillo and van der Vaart (2012).

	s = 25					**s = 50**					**s = 100**				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
PM1			111	96	94			176	165	154			267	302	307
PM2			106	92	82			169	165	152			269	280	274
EBM			103	96	93			166	177	174			271	312	319
PMed1			129	83	73			205	149	130			255	279	283
PMed2			125	86	68			187	148	129			273	254	245
EBMed			110	81	72			162	148	142			255	294	300
HT			175	142	70			339	284	135			676	564	252
HTO			136	92	84			206	159	139			306	261	245
EBMR	30	77	89	65	35	50	123	136	92	48	79	185	193	127	62
EBKM	27	71	80	57	30	46	113	122	81	40	74	171	174	112	53

MSE based on 1000 replications

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

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

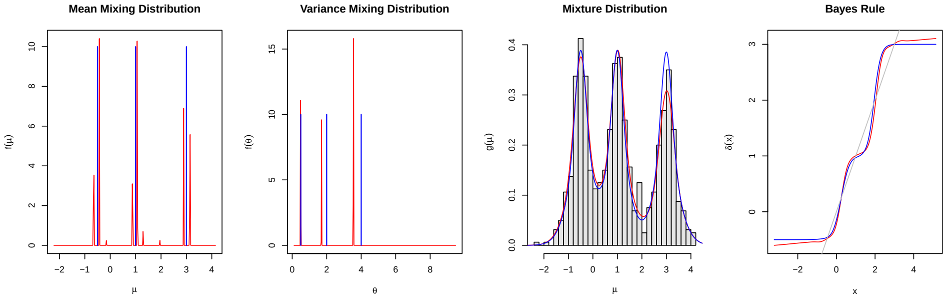
$$\mu_i \sim \tfrac{1}{3}(\delta_{-0.5} + \delta_1 + \delta_3) \perp\!\!\!\perp \theta_i \sim \tfrac{1}{3}(\delta_{0.5} + \delta_2 + \delta_4)$$

A Toy Example

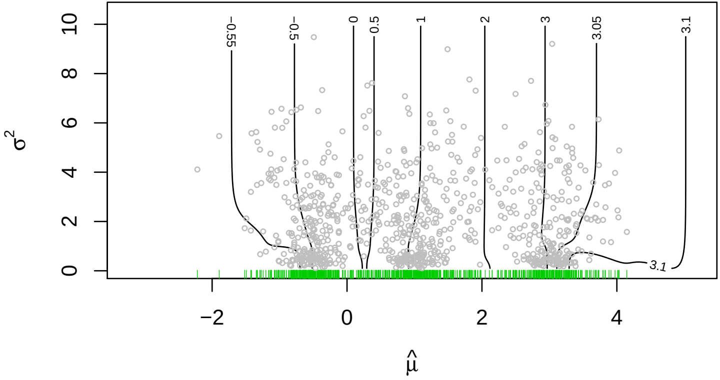
Model

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

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



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



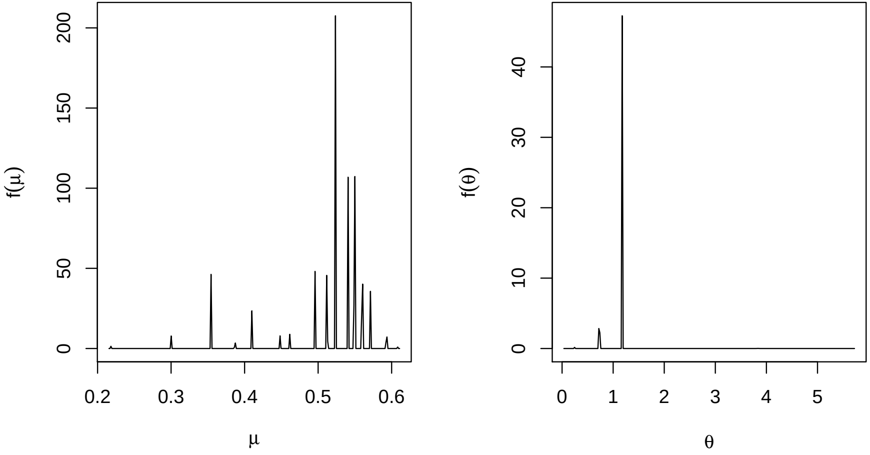
Empirical Bayesball

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:

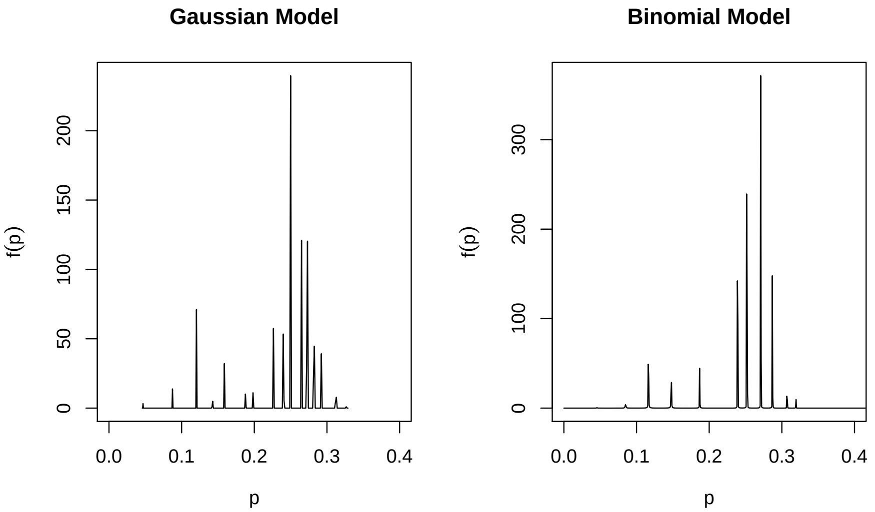
$$\hat{Y}_i = \operatorname{asin}\left(\sqrt{\frac{H_{i1} + 1/4}{N_{i1} + 1/2}}\right) \sim \mathcal{N}(\operatorname{asin}(\sqrt{\rho}), 1/(4N_{i1}))$$

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

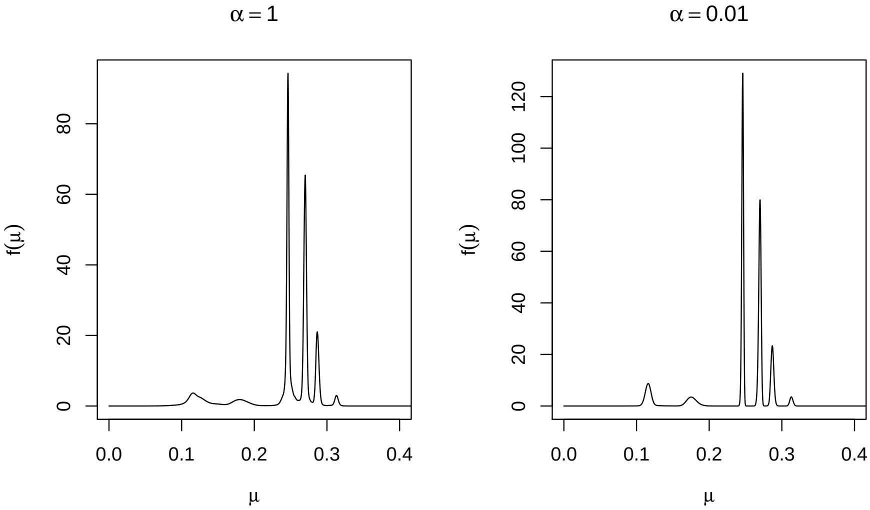
Prior Estimates on the Gaussian Scale



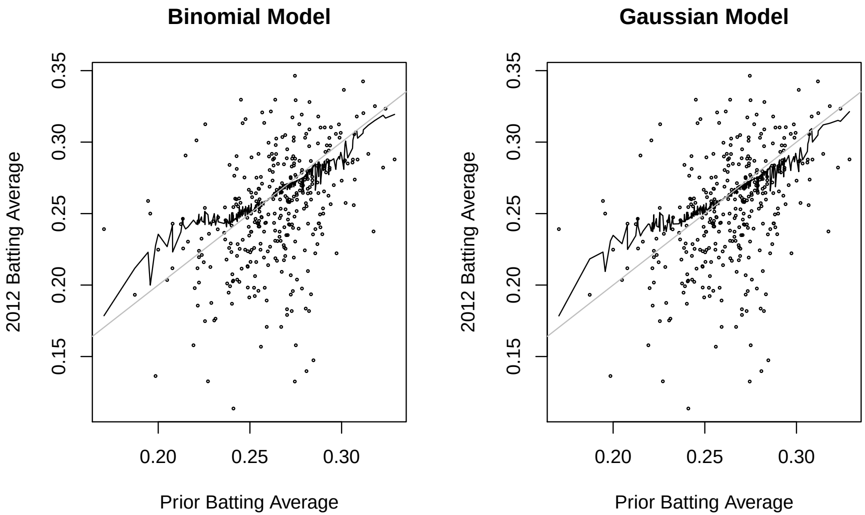
Prior Estimates on the Batting Average Scale



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:

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

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

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

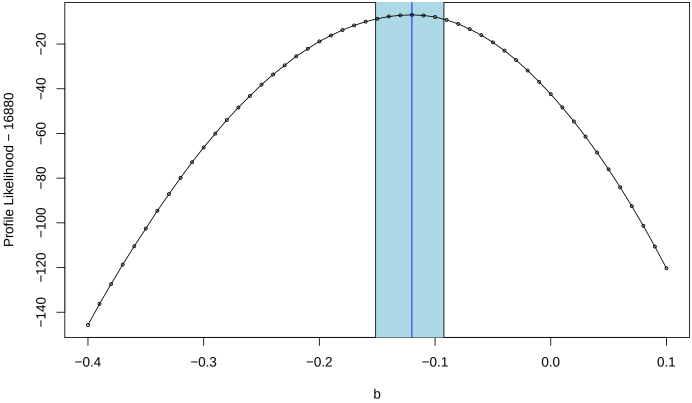
Age and Batting Ability

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.

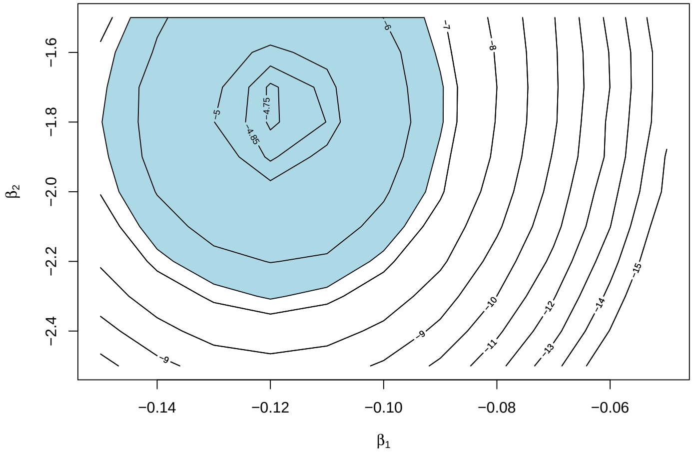
We evaluate the profile likelihood on a grid of parameter values, but as you will see the likelihood is quite smooth and well behaved so higher dimensional problems could be done with standard optimization software.

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

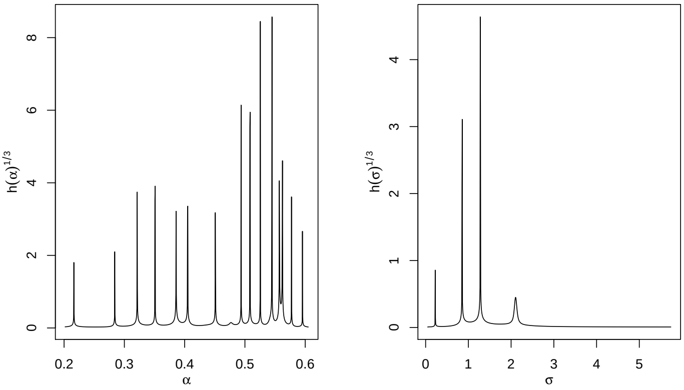
Profile Likelihood for the Linear Age Effect



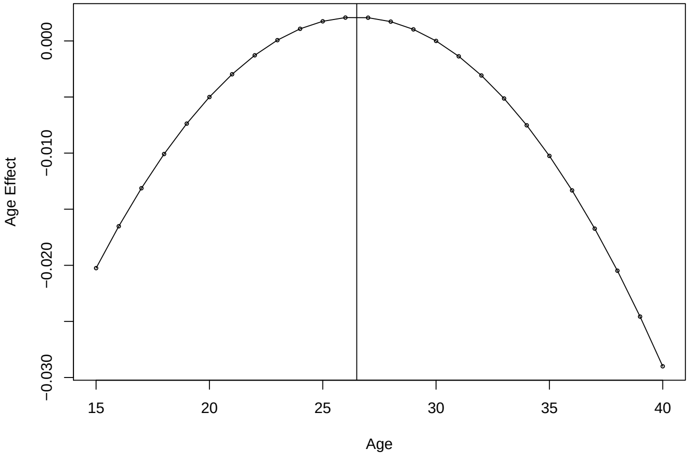
Contour Plot of the Quadratic Age Effect



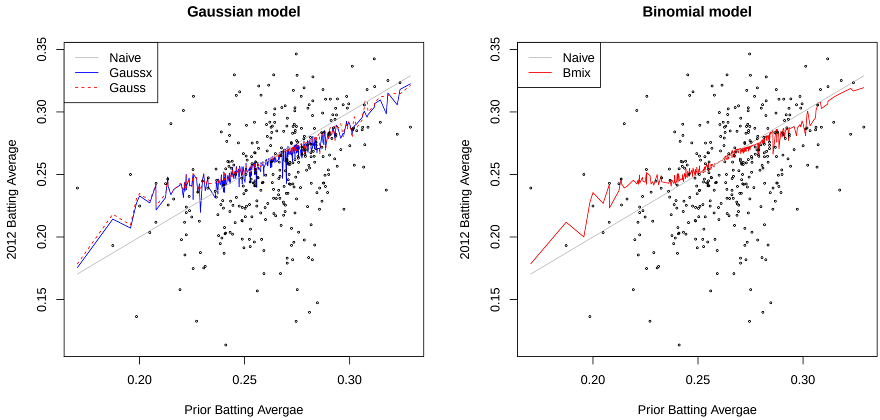
The Mixing Densities at the Profile MLE



The Estimated Quadratic Age Effect



Predictive Performance



Predictive Performance

Root Mean Squared Prediction Error

Gaussian Age Effects	Gaussian	Binomial	Dirichlet ($\alpha = 1$)	Dirichlet ($\alpha = 0.01$)	naive
0.0378	0.0393	0.0395	0.0395	0.0395	0.0394

- Dismal performance due to multimodality of the estimated mixing distribution not much shrinkage compared to naive estimator.
- Prior performance is not very useful for predicting future performance
- Model with age covariates performs slightly better.

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 mathematics 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 mathematics and baseball,
- Be cautious about predicting baseball batting averages, or anything else about baseball.