Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective

Roger Koenker

University of Illinois, Urbana-Champaign

University of Michigan: 9 October 2014

Joint work with Jiaying Gu (UIUC)
A Compound Decision 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 \(\mu_i\)'s under squared error loss. We might call this “incidental parameters with a vengeance.”
A Compound Decision Homework Problem

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

Not knowing any better, we assume that the $\mu_i$ 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:

$$\delta(y) = y + \frac{g'(y)}{g(y)}.$$
A Compound Decision 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 \( \mu_i \)'s under squared error loss. We might call this “incidental parameters with a vengence.”

- Not knowing any better, we assume that the \( \mu_i \) 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. This is the point of departure for Robbins’s empirical Bayes program.
Stein Rules I

Suppose that the $\mu_i$'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,

$$\delta(y) = \left(1 - \frac{1}{1 + \sigma_0^2}\right) y.$$
Suppose that the $\mu_i$’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 $\sigma_0^2$ is unknown, $S = \sum Y_i^2 \sim (1 + \sigma_0^2)\chi_n^2$, and recalling (!) that an inverse $\chi_n^2$ random variable has expectation, $(n - 2)^{-1}$, we obtain the Stein rule in its original form:

$$\hat{\delta}(y) = \left(1 - \frac{n - 2}{S}\right)y.$$
More generally, if $\mu_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$ we shrink instead toward the prior mean,

$$
\delta(y) = \mu_0 + \left(1 - \frac{1}{1 + \sigma_0^2}\right) (y - \mu_0),
$$
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 $S = \sum (Y_i - \bar{Y}_n)^2$. 
Needles and Haystacks

Johnstone and Silverman (2004) compare various thresholding rules with a parametric empirical Bayes procedure that estimates a prior mass at 0 and a scale parameter for a (non-null) Laplace density.

<table>
<thead>
<tr>
<th>Number nonzero</th>
<th>5</th>
<th>50</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value nonzero</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Exponential</td>
<td>36</td>
<td>32</td>
<td>17</td>
</tr>
<tr>
<td>Cauchy</td>
<td>37</td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>Postmean</td>
<td>34</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td>Exphard</td>
<td>51</td>
<td>43</td>
<td>22</td>
</tr>
<tr>
<td>$a = 1$</td>
<td>36</td>
<td>32</td>
<td>19</td>
</tr>
<tr>
<td>$a = 0.5$</td>
<td>37</td>
<td>34</td>
<td>17</td>
</tr>
<tr>
<td>$a = 0.2$</td>
<td>38</td>
<td>37</td>
<td>18</td>
</tr>
<tr>
<td>$a = 0.1$</td>
<td>38</td>
<td>37</td>
<td>18</td>
</tr>
<tr>
<td>SURE</td>
<td>38</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>Adapt</td>
<td>42</td>
<td>63</td>
<td>73</td>
</tr>
<tr>
<td>FDR $q = 0.01$</td>
<td>43</td>
<td>51</td>
<td>26</td>
</tr>
<tr>
<td>FDR $q = 0.1$</td>
<td>40</td>
<td>35</td>
<td>19</td>
</tr>
<tr>
<td>FDR $q = 0.4$</td>
<td>58</td>
<td>58</td>
<td>53</td>
</tr>
<tr>
<td>BlockThresh</td>
<td>46</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>NeighBlock</td>
<td>47</td>
<td>64</td>
<td>51</td>
</tr>
<tr>
<td>NeighCoeff</td>
<td>55</td>
<td>51</td>
<td>38</td>
</tr>
<tr>
<td>Universal soft</td>
<td>42</td>
<td>63</td>
<td>73</td>
</tr>
<tr>
<td>Universal hard</td>
<td>39</td>
<td>37</td>
<td>18</td>
</tr>
</tbody>
</table>
Nonparametric Empirical Bayes

Brown and Greenshtein (Annals, 2009) propose estimating $g$ by standard fixed bandwidth kernel methods and they compare performance to Johnstone and Silverman.

Jiang and Zhang (Annals, 2009) adopt the Kiefer and Wolfowitz (1956) non-parametric MLE for mixture models using Laird's (1978) EM implementation. 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:

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

and the implied Bayes rule becomes at convergence:

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

Brown and Greenshtein (Annals, 2009) propose estimating $g$ by standard fixed bandwidth kernel methods and they compare performance to Johnstone and Silverman.

Jiang and Zhang (Annals, 2009) adopt the Kiefer and Wolfowitz (1956) non-parametric MLE for mixture models using Laird’s (1978) EM implementation. 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:

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

Brown and Greenshtein (Annals, 2009) propose estimating $g$ by standard fixed bandwidth kernel methods and they compare performance to Johnstone and Silverman.

Jiang and Zhang (Annals, 2009) adopt the Kiefer and Wolfowitz (1956) non-parametric MLE for mixture models using Laird’s (1978) EM implementation. 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 via Convex Optimization

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,

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

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

- GMLEBIP
- GMLEBEM: $m = 10^2$
- GMLEBEM: $m = 10^4$
- GMLEBEM: $m = 10^5$

Roger Koenker (UIUC)
Unobserved Heterogeneity
UM: 9.10.2014 9 / 31
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.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>EM1</th>
<th>EM2</th>
<th>EM3</th>
<th>IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterations</td>
<td>100</td>
<td>10,000</td>
<td>100,000</td>
<td>15</td>
</tr>
<tr>
<td>Time</td>
<td>1</td>
<td>37</td>
<td>559</td>
<td>1</td>
</tr>
<tr>
<td>( L(g) - 422 )</td>
<td>0.9332</td>
<td>1.1120</td>
<td>1.1204</td>
<td>1.1213</td>
</tr>
</tbody>
</table>

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.
Performance of the MLE Bayes Rule

In the Johnstone and Silverman sweepstakes we have the following comparison of performance.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>k = 5</th>
<th></th>
<th></th>
<th></th>
<th>k = 50</th>
<th></th>
<th></th>
<th></th>
<th>k = 500</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td></td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$\hat{\delta}_{\text{MLE-IP}}$</td>
<td>33</td>
<td>30</td>
<td>16</td>
<td>8</td>
<td></td>
<td>153</td>
<td>107</td>
<td>51</td>
<td>11</td>
<td></td>
<td>454</td>
<td>276</td>
</tr>
<tr>
<td>$\hat{\delta}_{\text{MLE-EM}}$</td>
<td>37</td>
<td>33</td>
<td>21</td>
<td>11</td>
<td></td>
<td>162</td>
<td>111</td>
<td>56</td>
<td>14</td>
<td></td>
<td>458</td>
<td>285</td>
</tr>
<tr>
<td>$\tilde{\delta}$</td>
<td>37</td>
<td>34</td>
<td>21</td>
<td>11</td>
<td></td>
<td>173</td>
<td>121</td>
<td>63</td>
<td>16</td>
<td></td>
<td>488</td>
<td>310</td>
</tr>
<tr>
<td>$\tilde{\delta}_{1.15}$</td>
<td>53</td>
<td>49</td>
<td>42</td>
<td>27</td>
<td></td>
<td>179</td>
<td>136</td>
<td>81</td>
<td>40</td>
<td></td>
<td>484</td>
<td>302</td>
</tr>
<tr>
<td>J-S Min</td>
<td>34</td>
<td>32</td>
<td>17</td>
<td>7</td>
<td></td>
<td>201</td>
<td>156</td>
<td>95</td>
<td>52</td>
<td></td>
<td>829</td>
<td>730</td>
</tr>
</tbody>
</table>

Here MLE-EM is Jiang 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 \( \theta_i \) 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 \( \theta_a \) takes five possible values: The first 8 rows of the Table are taken directly from Table 1 of Castillo and van der Vaart (2012).

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

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 (2009)** For the normal location mixture
problem, with a (complicated) weak $p$th moment restriction on $\Theta$, the
approximate non-parametric MLE, $\hat{\theta} = \delta_{\hat{F}_n}(Y)$ 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 $p$th moment condition encompasses a broad class of both
deterministic and stochastic classes $\Theta$. Relatively little is still known about
the KWMLE beyond the original consistency result: no rates, no limiting
distributions.
Econometric Motivation: Duration Modeling

Heckman and Singer (1984) employed the Kiefer-Wolfowitz MLE to study durations $T_i$ of single spell unemployment data with (Weibull) density:

$$f(t | x_i, \alpha, \beta, \theta_i) = \alpha t^{\alpha-1} e^{x_i' \beta} \theta_i \exp(-t^\alpha e^{x_i' \beta} \theta_i), \quad \theta_i \sim H$$

Conclusions:

1. Neglecting heterogeneity in $\theta_i$ leads to misinterpretation of “duration dependence.”

2. Common parameters in the model $(\alpha, \beta)$ are sensitive to parametric assumptions imposed on $H(\theta)$.

3. EM is painful.
Econometric Motivation: Panel Data

Model:

\[ y_{it} = \alpha_i + \sqrt{\theta_i} u_{it}, \quad u_{it} \sim \mathcal{N}(0, 1) \]

Neyman and Scott (1948) showed that in the “fixed effect” model with \( \theta_i \equiv \theta_0 \), the MLE of \( \theta_0 \) is inconsistent.
Econometric Motivation: Panel Data

Model:

\[ y_{it} = \alpha_i + \sqrt{\theta_i} u_{it}, \quad u_{it} \sim \mathcal{N}(0, 1) \]

Neyman and Scott (1948) showed that in the “fixed effect” model with \( \theta_i \equiv \theta_0 \), the MLE of \( \theta_0 \) is inconsistent.

Kiefer and Wolfowitz (1956) then showed that consistency of \( \hat{\theta}_0 \) could be restored if we (simply!) replaced the fixed effect assumption by an iid \( \alpha_i \sim G_0 \) assumption, and proceeded with the MLE. Indeed, both \( \theta_0 \) and \( G_0 \) are consistently estimable.
Econometric Motivation: Panel Data

Model:

\[ y_{it} = \alpha_i + \sqrt{\theta_i} u_{it}, \quad u_{it} \sim \mathcal{N}(0, 1) \]

Neyman and Scott (1948) showed that in the “fixed effect” model with \( \theta_i \equiv \theta_0 \), the MLE of \( \theta_0 \) is inconsistent. Kiefer and Wolfowitz (1956) then showed that consistency of \( \hat{\theta}_0 \) could be restored if we (simply!) replaced the fixed effect assumption by an iid \( \alpha_i \sim G_0 \) assumption, and proceeded with the MLE. Indeed, both \( \theta_0 \) and \( G_0 \) are consistently estimable.

Using annual income data from the PSID, I’d like to now show how to extend these methods to incorporate:

- random scale \( \sqrt{\theta_i} \),
- additional covariates and dynamics,
- bivariate heterogeneity in \((\alpha, \theta)\),
- forecasting and prediction.

Roger Koenker (UIUC)
A Toy Example

Model

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

\[ \mu_i \sim \frac{1}{3}(\delta_{-0.5} + \delta_1 + \delta_3) \quad \perp \parallel \quad \theta_i \sim \frac{1}{3}(\delta_{0.5} + \delta_2 + \delta_4) \]
A Toy Example

Model

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

\[ \mu_i \sim \frac{1}{3}(\delta_{-0.5} + \delta_1 + \delta_3) \perp \perp \theta_i \sim \frac{1}{3}(\delta_{0.5} + \delta_2 + \delta_4) \]
In the Beginning was the Data

- PSID sample used by Meghir and Pistaferri (2004) Browning, Ejrnæs and Alvarez (2010), Hospido (2012), ...
- 2069 individuals between age 25-55 with at least 9 consecutive records,
- Further reduced to 938 individuals with records starting at age 25,
- Preliminary estimation of observable effects: quadratic age, race, education, region, marital status to obtain log earning residuals, $y_{it}$. 
QQ Plots of Partial Differences

Gaussian Quantiles

Empirical Quantiles

-2  0  2

-5  0  5

-5  0  5

-5  0  5
Scatter Plots of Partial Differences

![Scatter Plots of Partial Differences](image-url)
The Mixture Model

\[ y_{it} = \rho y_{it-1} + \alpha_i (1 - \rho) + \sqrt{\theta_i} \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(0, 1), \quad (\alpha_i, \theta_i) \sim \mathcal{H} \]

- We can re-write the model as

\[ y_{it} - \rho y_{it-1} := z_{it} \mid \alpha_i, \theta_i \sim \mathcal{N}((1 - \rho)\alpha_i, \theta_i) \]

- Fixing \( \rho \), we reduce the dimension via sufficient statistics

\[
\hat{\alpha}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} z_{it}, \quad \hat{\alpha}_i \mid \alpha_i, \theta_i \sim \mathcal{N}(\alpha_i, \theta_i/m_i) \\
\hat{s}_i = \frac{1}{T_i-1} \sum_{t=1}^{T_i} (z_{it} - \hat{\alpha}_i)^2, \quad \hat{s}_i \mid \theta_i \sim \gamma((T_i - 1)/2, 2\theta_i/(T_i - 1))
\]

- The likelihood factors:

\[
L(z_{i1}, \ldots z_{iT_i} \mid \rho) \propto \int \int f(\hat{\alpha}_i \mid \alpha, \theta) \gamma(s_i \mid \theta) \, dH_\rho(\alpha, \theta)
\]

\[ \gamma \]

\[ g_i \]
Estimation

For fixed $\rho$ the Kiefer-Wolfowitz MLE is

$$\hat{H}_\rho = \arg\max_{H \in \mathcal{H}} \sum_{i=1}^{n} \log \int \int f(\hat{\alpha}_i \mid \alpha, \theta) \gamma(s_i \mid \theta) \, dH(\alpha, \theta)$$

Given $\hat{H}_\rho$ we can estimate $\rho$ by profile likelihood,

$$\hat{\rho} = \arg\max_{\rho} \sum_{i=1}^{n} \log \int \int f(\hat{\alpha}_i \mid \alpha, \theta) \gamma(s_i \mid \theta) \, d\hat{H}_\rho(\alpha, \theta)$$

Note that $\hat{\alpha}_i$ and $s_i$ implicitly depend upon $\rho$ via the partial differencing.

- Identification for $H$ follows from a uniqueness of the characteristic function argument.
- Identification of $\rho$ follows from the quadratic approximation of profile likelihood.
The Heterogeneity Distribution $\hat{H}_\rho$ and $\hat{\rho}$

- Only mild persistence of $y_{it}$ once heterogeneity of scale is accounted for,
- Some negative dependence in $H(\alpha, \theta)$, but no apparent parametric approximation.
A financial advisor, who has witnessed many individual earning paths, wishes to forecast future income paths for a new client with earning history $y_0 = \{y_t : t = 1, \ldots, T_0\}$.

1. Draw one pair $(\alpha, \theta)$ from the posterior $p(\alpha, \theta | y_0)$,

2. Simulate $y_1 = \{y_t : t = T_0 + 1, \ldots, T\}$

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

$m$ times to obtain $m$ paths, $y_1$, then

3. Repeat steps 1 and 2 $M$ times.

Construct quantile prediction bands from the $mM$ trajectories.
Prediction Bands for Two Individuals

The advisor updates the (estimated) prior, $\hat{H}$, based on the first 9 years of income data, for ages 25-34, and then forecasts earnings to age 50.
Prediction Bands for Two (More) Individuals

Pointwise bands don’t always cover!

PSID ID Number 44

PSID ID Number 1
Uniform Prediction Bands for Two (More) Individuals

Uniform bands are safer!

PSID ID Number 44

PSID ID Number 1
Estimation of Random Effects

Estimation of \{((\alpha_i, \theta_i) : i = 1, \cdots, n}\} brings us back to the Tweedie (Eddington) formulae. Shrinkage rules of this type play an important role in insurance rating, e.g. B"uhlmann on “Credibility Theory,” see also Goldberger (1962) on Best Linear Unbiased Prediction aka BLUP.

- Recall
  \[ \hat{\alpha}_i \mid \alpha_i, \theta_i \sim N(\alpha_i, \theta_i/T_i) \]
  \[ s_i \mid \theta_i \sim \gamma((T_i - 1)/2, 2\theta_i/(T_i - 1)) \]

- Under $\mathcal{L}_2$ loss,
  \[ \min_{\delta} \mathbb{E}_{(\alpha, \theta)} \|\delta(y) - \alpha\|^2 \]

- The Bayes rule is
  \[ \delta_i = \mathbb{E}(\alpha \mid \hat{\alpha}_i, s_i) = \int_{\theta} \mathbb{E}(\alpha \mid \hat{\alpha}_i, \theta)f(\theta \mid \hat{\alpha}_i, s_i)d\theta \]
The Garlic Plot
Bayes Rule for $\alpha$ given various $s$
Conclusions

- More efficient computation of the Kiefer-Wolfowitz MLE opens the way to a variety of nonparametric mixture models of unobserved heterogeneity,
- Profile likelihood provides an attractive strategy for both estimation and testing in such models,
- Bivariate nonparametric heterogeneity in location and scale is a flexible framework for longitudinal data,
- Empirical Bayes provides natural forecasting and prediction apparatus.
Some References