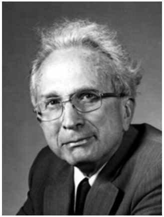
Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective

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 μ_i's under squared error loss. We might call this "incidental parameters with a vengence."

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 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:

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

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 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. This is the point of departure for Robbins's empirical Bayes program.

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

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.

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	293	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

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.

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:

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

Nonparametric 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:

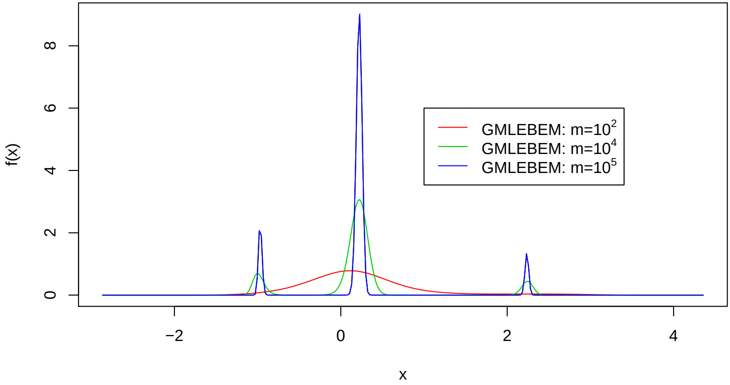
$$\hat{f}_j^{(k+1)} = n^{-1} \sum_{i=1}^{n} \frac{\hat{f}_j^{(k)} \phi(Y_i - u_j)}{\sum_{\ell=1}^{m} \hat{f}_\ell^{(k)} \phi(Y_i - u_\ell)},$$

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:

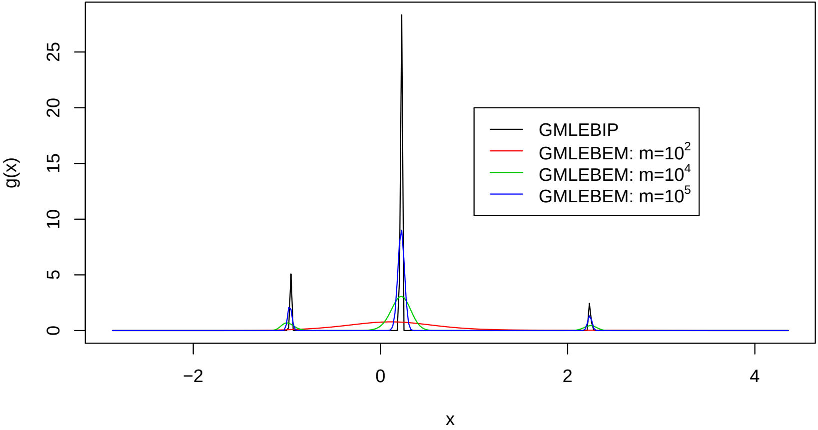
$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{m} \phi(y_i - u_j) f_j\Big),$$

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.

Performance of the MLE Bayes Rule

In the 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
$\hat{\delta}_{\mathrm{MLE-IP}}$	33	30	16	8	153	107	51	11	454	276	127	18
$\hat{\delta}_{\mathrm{MLE-EM}}$	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 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 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 (2009) 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 broad class of both deterministic and stochastic classes Θ. 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:

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

Conclusions:

1. Neglecting heterogeneity in θ_i leads to misinterpretation of "duration dependence."
2. Common parameters in the model (α, β) 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 θ_0 is inconsistennt.

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 θ_0 is inconsistennt.
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 θ_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 θ_0 is inconsistennt.

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 θ_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 (α, θ),
- forecasting and prediction.

A Toy Example
Model

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

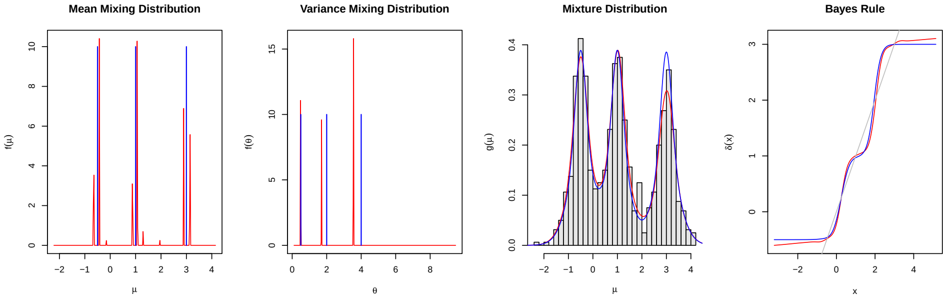
$$\mu_i \sim \tfrac{1}{3}(\delta_{-0.5} + \delta_1 + \delta_3) \perp\!\!\!\perp \theta_i \sim \tfrac{1}{3}(\delta_{0.5} + \delta_2 + \delta_4)$$

A Toy Example

Model

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

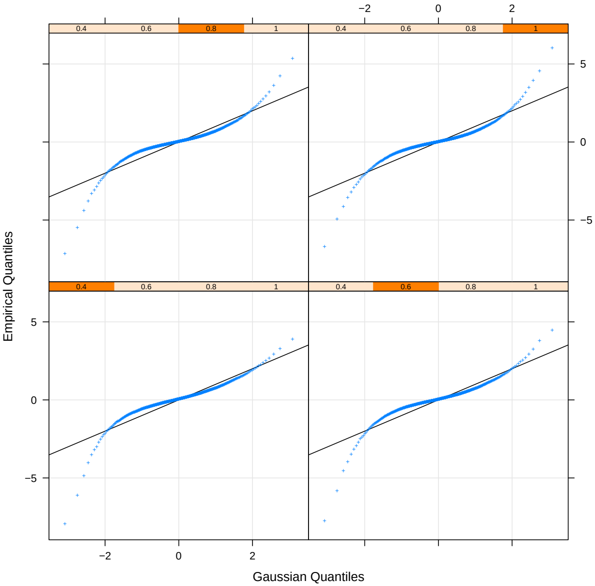
$$\mu_i \sim \tfrac{1}{3}(\delta_{-0.5} + \delta_1 + \delta_3) \perp\!\!\!\perp \theta_i \sim \tfrac{1}{3}(\delta_{0.5} + \delta_2 + \delta_4)$$



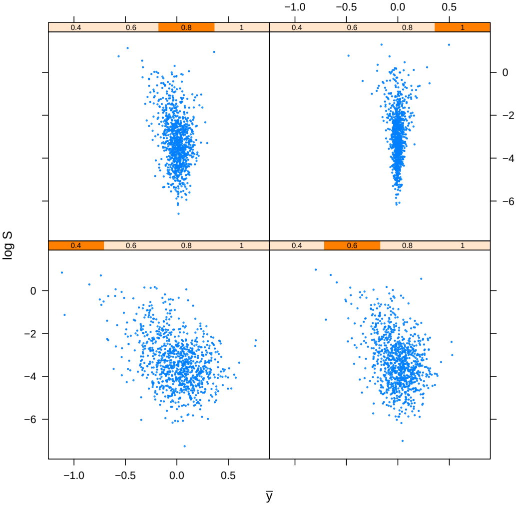
In the Beginning was the Data

- 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



Scatter Plots of Partial Differences



The Mixture Model

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

- 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 ρ, we reduce the dimension via sufficient statistics

$$\begin{aligned} \hat{\alpha}_i &= \tfrac{1}{T_i}\textstyle\sum_{t=1}^{T_i} z_{it}, & \hat{\alpha}_i \mid \alpha_i, \theta_i \sim \mathcal{N}(\alpha_i, \theta_i/m_i) \\ s_i &= \tfrac{1}{T_i-1}\textstyle\sum_{t=1}^{T_i}(z_{it} - \hat{\alpha}_i)^2, & s_i \mid \theta_i \sim \gamma((T_i-1)/2, 2\theta_i/(T_i-1)) \end{aligned}$$

- The likelihood factors:

$$L(z_{i1}, \ldots z_{iT_i} \mid \rho) \propto \underbrace{\int\int \underbrace{f(\hat{\alpha}_i \mid \alpha, \theta)}_{\mathcal{N}} \underbrace{\gamma(s_i \mid \theta)}_{\gamma} dH_\rho(\alpha, \theta)}_{g_i}$$

Estimation

For fixed ρ the Kiefer-Wolfowitz MLE is

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

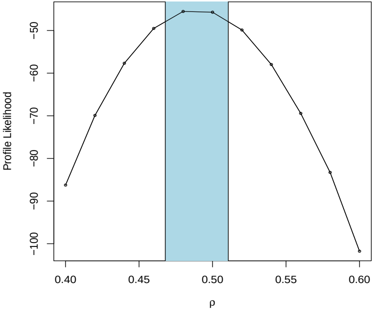
Given $\hat{H}_\rho$ we can estimate ρ by profile likelihood,

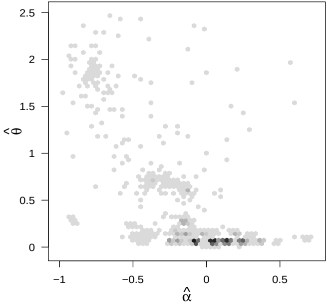
$$\hat{\rho} = \operatorname*{argmax}_{\rho} \sum_{i=1}^{n} \log \int\int f(\hat{\alpha}_i \mid \alpha, \theta)\gamma(s_i \mid \theta) d\hat{H}_\rho(\alpha, \theta)$$

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

- Identification for H follows from a uniqueness of the characteristic function argument.
- Identification of ρ follows from the quadratic approximation of profile likelihood.

The Heterogeneity Distribution $\hat{H}_{\hat{\rho}}$ and $\hat{\rho}$





- Only mild persistence of y_{it} once heterogeneity of scale is accounted for,
- Nice quadratic approximation of profile likelihood, e.g. Murphy and van der Vaart (1995), van der Vaart (1996), gives a narrow Wilks confidence interval.
- Some negative dependence in $H(\alpha, \theta)$, but no apparent parametric approximation.

Forecasting Income Trajectories

A financial advisor, who has witnessed many individual earning paths, wishes to forecast future income paths for a new client with earning history $\mathcal{Y}_0 = \{y_t : t = 1, \dots, T_0\}$.

1. Draw one pair (α, θ) from the posterior $p(\alpha, \theta \mid \mathcal{Y}_0)$,
2. Simulate $\mathcal{Y}_1 = \{y_t : t = T_0 + 1, \dots, T\}$

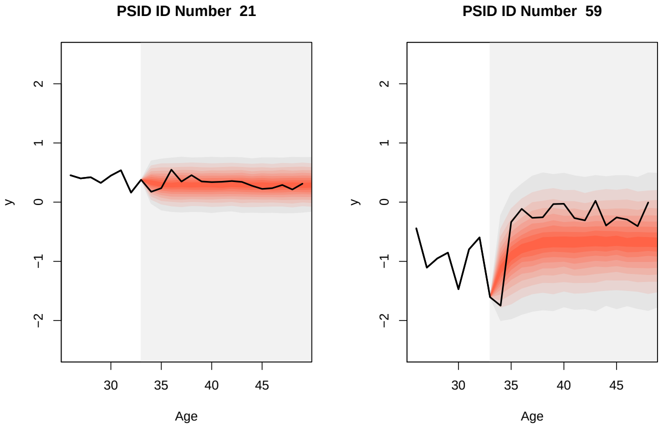
   $$y_{T_0+s} = \alpha + \hat{\rho} y_{T_0+s-1} + \sqrt{\theta} u_s, \; s = 1, \cdots, T - T_0, \text{ and } u_s \sim \mathcal{N}(0, 1),$$

   m times to obtain m paths, $\mathcal{Y}_1$, then
3. Repeat steps 1 and 2 M times.

Construct quantile prediction bands from the mM trajectories.

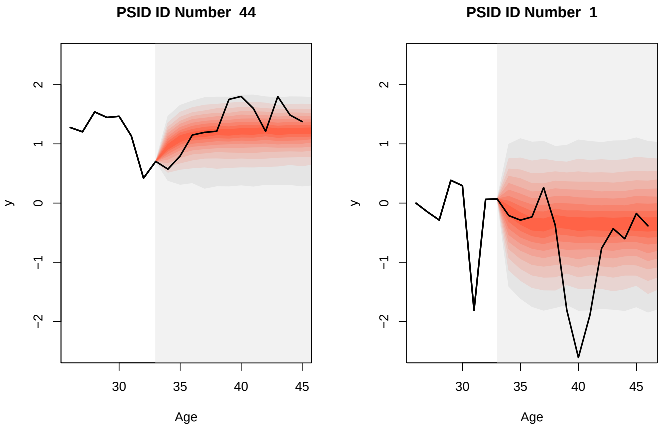
Prediction Bands for Two Individuals

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.



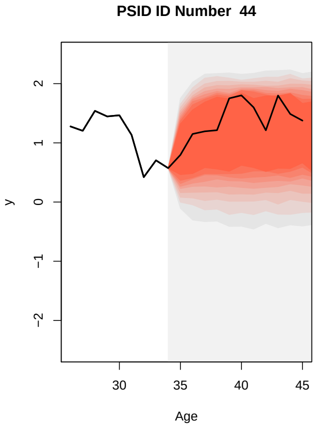
Prediction Bands for Two (More) Individuals

Pointwise bands don't always cover!



Uniform Prediction Bands for Two (More) Individuals

Uniform bands are safer!



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. Bühlmann on "Credibility Theory," see also Goldberger (1962) on Best Linear Unbiased Prediction aka BLUP.

- Recall

$$\begin{aligned} \hat{\alpha}_i \mid \alpha_i, \theta_i & \sim \mathcal{N}(\alpha_i, \theta_i / T_i) \\ s_i \mid \theta_i & \sim \gamma((T_i - 1)/2, 2\theta_i/(T_i - 1)) \end{aligned}$$

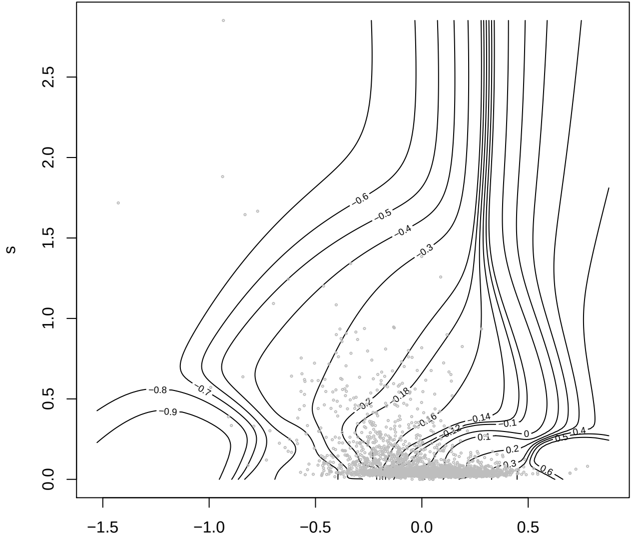
- Under $\mathcal{L}_2$ loss,

$$\min_{\delta} \; \mathbb{E}_{(\alpha,\theta)} \|\delta(y) - \alpha\|^2$$

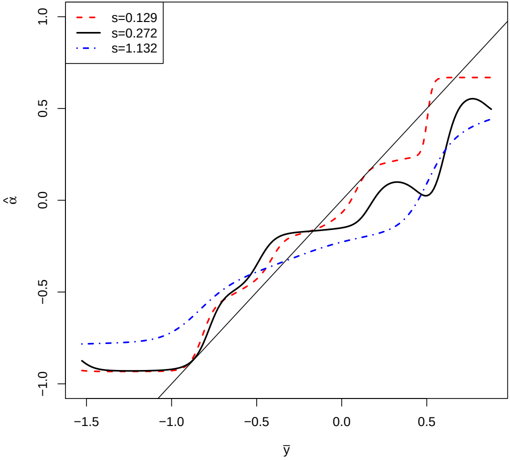
- The Bayes rule is

$$\delta_i = \mathbb{E}(\alpha \mid \hat{\alpha}_i, s_i) = \int_{\theta} \mathbb{E}(\alpha \mid \hat{\alpha}_i, \theta) f(\theta \mid \hat{\alpha}_i, s_i) d\theta$$

The Garlic Plot



Bayes Rule for α 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

GU, J., AND R. KOENKER (2014): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," preprint.

KOENKER, R. (2014a): "A Gaussian Compound Decision Bakeoff," *Stat*, 3, 12–16.

——— (2014b): "REBayes: An R package for empirical Bayes methods," Available from CRAN `http://www.r-project.org`.

KOENKER, R., AND J. GU (2014): "Frailty, Profile Likelihood and Medfly Mortality," in *Contemporary Developments in Statistical Theory: A Festschrift for Hira Lal Koul*, ed. by S. Lahiri, A. Schick, A. SenGupta, and T. Sriram, pp. 227–238. Springer.

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