L'hôpital

Fu Zixuan

Supervised by Prof.Thierry Magnac

June 20, 2024

Outline

Introduction

Data and Estimation

Empirical Bayes Compound decision: The Selection Problem

Conclusion

Motivation: Invidious decision

- *League table mentality*: Ranking & Selection.
- *Noisy estimates*: Unobserved heterogeneity.
- *Bayesian view*: Prior distribution.

Motivation: French Hospitals

- *Productivity/Efficiency*: Factories, Schools, Hospitals etc.
- *Methodology*:
  1. (Parametric) Stochastic Frontier → How far it is to the frontier.
  2. (Non-Parametric): Data Envelopment → Compare with other units.
  3. Input demand function: [**?**].
- *Ownership*: Public (Teaching, Ordinary) vs. Private (For profit, Non-profit).

Questions

- Out of the top 20% hospitals in France[1], how many of them are public hospitals/private hospitals?
- What would be the selection outcome if I also control for the **False Discovery Rate**?
- Does different ranking/selection rule produce contradicting results? And to what degree?

[1] in terms of labor (nurses) employment efficiency

Roadmap

1. Data: The Annual Statistics of Health Establishments (SAE)
2. Estimation of efficiency
   - Y: Labor input (number of full time equivalent nurses).
   - X: Hospital output (e.g., inpatient/outpatient stays, medical sessions).
3. Selection problem: Compound decision framework and optimal decision rule
4. Comparison of outcomes under different decision rules
5. Conclusion

Outline

Introduction

Data and Estimation

Empirical Bayes Compound decision: The Selection Problem

Conclusion

Data

Year	Teaching	Normal Public	Private FP	Private NP	Total
2013	198	1312	1305	1382	4197
2014	201	1274	1293	1349	4117
2015	211	1275	1297	1349	4132
2016	212	1266	1297	1313	4088
2017	211	1249	1297	1306	4063
2018	214	1247	1296	1288	4045
2019	214	1236	1287	1281	4018
2021	219	1222	1293	1264	3998
2022	220	1220	1296	1259	3995

First glance

Dependent Variable:	Nurses	
Model:	(1)	(2)
Variables		
Constant	1.59***	1.58***
	(0.067)	(0.069)
STAC inpatient	0.278***	0.277***
	(0.012)	(0.013)
...	...	...
Private Forprofit	-0.303***	-0.280***
	(0.061)	(0.065)
Private Nonprofit	-0.215***	-0.188***
	(0.056)	(0.055)
Teaching	0.717***	0.709***
	(0.056)	(0.056)
Fit statistics		
Observations	15,335	13,402
R^2	0.835	0.837

Clustered (FI) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Naive counterfactual

Panel data Estimator

- Strict exogeneity: Within Group/First Diffrence

$$E[\epsilon_{it}|x_{i1}, \ldots, x_{iT}, \theta_i] = 0$$

- Relaxed: First Difference GMM [**?**], System GMM [**?**, **?**].

$$E[\epsilon_{it}|x_{i1}, \ldots, x_{it-p}, z_{i1}, \ldots, z_{it-p}, \theta_i] = 0$$

I choose **System GMM**, using lagged difference as instruments for level.

$$\mathbb{E}[\Delta x_{i,t-1}(y_{it} - \alpha y_{i,t-1} - \beta x_{it})] \quad \text{if} \quad \mathbb{E}\Delta x_{i,t-1}\varepsilon_{i,t} = 0$$

Panel Data Estimation

Dependent Variable:	log(ETP_INF)		
	Within Group	First Difference	**System GMM**
Model:	(1)	(2)	(3)
Variables			
log(SEJHC_MCO)	0.10^{***}	0.07^{***}	0.54^{***}
	(0.00)	(0.01)	(0.02)
log(SEJHP_MCO)	0.02^{***}	0.01^{***}	0.02
	(0.00)	(0.00)	(0.02)
...	...	...	...
log(SEANCES_MED)	0.02^{***}	0.02^{***}	0.06^{***}
	(0.00)	(0.00)	(0.01)
log(SEJ_PSY)	0.00	0.00	0.07^{***}
	(0.00)	(0.00)	(0.01)
Fit statistics			
Observations	15335	13502	15335
n	1833	1833	1833
T	9	9	9

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Outline

Introduction

Data and Estimation

Empirical Bayes Compound decision: The Selection Problem

Conclusion

Compound Decision Framework

Observe:

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$$

$$\text{where} \quad \hat{\theta}_i | \theta_i \sim P_{\theta_i}$$

Decision:

$$\delta(Y) = (\delta_1(\hat{\boldsymbol{\theta}}), \ldots, \delta_n(\hat{\boldsymbol{\theta}}))$$

Compound Loss and Risk

Loss:

$$L_n(\theta, \delta(\hat{\boldsymbol{\theta}})) = \sum_{i=1}^{n} L(\theta_i, \delta_i(\hat{\theta})).$$

Risk (Expectation of risk):

$$\begin{aligned} R_n(\theta, \delta(\hat{\boldsymbol{\theta}})) &= \mathbb{E}[L_n(\theta, \delta(\hat{\boldsymbol{\theta}}))] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[L(\theta_i, \delta_i(\hat{\boldsymbol{\theta}}))] \\ &= \frac{1}{n}\sum_{i=1}^{n} \int \dots \int L(\theta_i, \delta_i(\hat{\theta}_1, \dots, \hat{\theta}_n)) dP_{\theta_1}(\hat{\theta}_1) \dots dP_{\theta_n}(\hat{\theta}_n). \end{aligned}$$

The Selection Task

- Select the bottom 20% [2] out of the population of θ_i, those i whose $\alpha_i > G^{-1}(0.8)$

- Control the overall false discovery rate at 20%,

$$\frac{\mathbb{E}_G\left[1\left\{\theta_i > \theta_\alpha, \delta_i = 1\right\}\right]}{\mathbb{E}_G\left[\delta_i\right]} \leq \gamma$$

[2] The most efficient 20%.

Problem Formulation

The **loss** function is

$$L(\delta, \theta) = \sum h_i(1 - \delta_i) + \tau_1 \left(\sum (1 - h_i)\delta_i - \gamma \delta_i \right) + \tau_2 \left(\sum \delta_i - \alpha n \right)$$

Therefore, the problem is to find δ such that

$$\begin{aligned}
\min_{\delta} \quad & \mathbb{E}_G \mathbb{E}_{\theta|\hat{\theta}} \left[L(\delta, \theta) \right] \\
=& \mathbb{E}_G \ \sum \mathbb{E}(h_i)(1 - \delta_i) + \tau_1 \left(\sum (1 - \mathbb{E}(h_i))\delta_i - \gamma \delta_i \right) \\
& + \tau_2 \left(\sum \delta_i - \alpha n \right) \\
=& \mathbb{E}_G \sum v_\alpha(\hat{\theta})(1 - \delta_i) + \tau_1 \left(\sum (1 - v_\alpha(\hat{\theta}))\delta_i - \gamma \delta_i \right) + \tau_2 \left(\sum \delta_i - \alpha n \right)
\end{aligned}$$

where $v_\alpha(\hat{\theta}) = \mathbb{P}(\theta < \theta_\alpha | \hat{\theta})$ is the **posterior tail probability**.

Empirical Bayes G

Observe

$$Y_{it} = \theta_i + \varepsilon_{it} \quad \varepsilon_{it} \sim \mathcal{N}(0, \sigma_i^2) \quad (\theta_i, \sigma_i^2) \sim G$$

Neither θ_i nor σ_i^2 is known. But the sufficient statistics are

$$Y_i = \frac{1}{T_i} \sum_{t=1}^{T_i} Y_{it} \quad \text{where} \quad Y_i | \theta_i, \sigma_i^2 \sim \mathcal{N}(\theta_i, \sigma_i^2 / T_i)$$

$$S_i = \frac{1}{T_i - 1} \sum_{t=1}^{T_i} (Y_{it} - Y_i)^2 \quad \text{where} \quad S_i | \sigma_i^2 \sim \Gamma(r_i = (T_i - 1)/2, 2\sigma_i^2/(T_i - 1))$$

Tail probability

Given the two sufficient statistics, the posterior tail probability is

$$\begin{aligned} v_\alpha(Y_i, S_i) &= P(\theta_i > \theta_\alpha | Y_i, S_i) \\ &= \frac{\int_{-\infty}^{\theta_\alpha} \Gamma(s_i | r_i, \sigma_i^2) f(y_i | \theta_i, \sigma_i^2) dG(\theta_i, \sigma_i^2)}{\int_{-\infty}^{\infty} \Gamma(s_i | r_i, \sigma_i^2) f(y_i | \theta_i, \sigma_i^2) dG(\theta_i, \sigma_i^2)} \end{aligned}$$

We want to find a cutoff λ such that both constraints are satisfied:

- Capacity: $\int \int P(v_\alpha(Y_i, S_i) > \lambda) dG(\theta_i, \sigma_i^2) \leq \alpha$
- FDR: $\int \int \frac{E[1\{v_\alpha(Y_i,S_i) > \lambda\}(1 - v_\alpha(Y_i,S_i))]}{E[1\{v_\alpha(Y_i,S_i) > \lambda\}]} dG(\theta_i, \sigma_i^2) \leq \gamma$

Estimate G

The primal problem:

$$\min_{f=dG} \left\{ -\sum_i \log g(y_i) \middle| g_i = T(f_i),\ K(f_i) = 1,\ \forall i \right\}$$

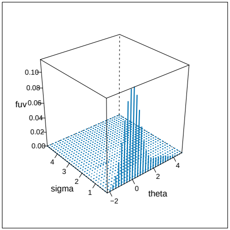
where $T(f_i) = \int p(y|\alpha) f_i d\alpha$ and $K(f_i) = \int f_i d\alpha$.
Discretize the support:

$$\min_{f=dG} \left\{ -\sum_i \log g(y_i) \middle| g = Af,\ 1^T f = 1 \right\}$$

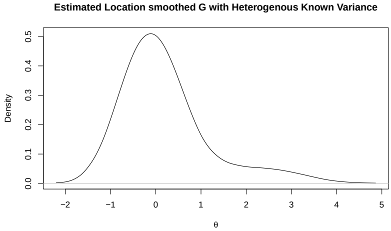
where $A_{ij} = p(y_i|\alpha_j)$ and $f = (f(\alpha_1), f(\alpha_2), \ldots, f(\alpha_m))$.
The dual problem:

$$\max_{\lambda,\mu} \left\{ \sum_i \log \lambda_1(i) \middle| A^T \lambda_1 < \lambda_2 1,\ (\lambda_1 > 0) \right\}$$

$\hat{G}$

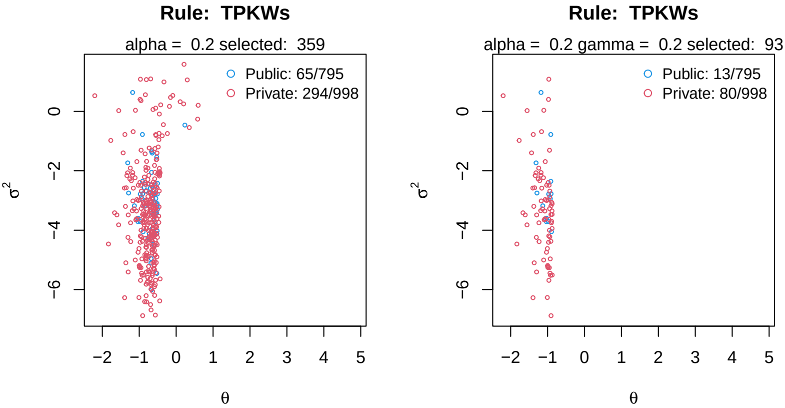


fuv
0.10
0.08
0.06
0.04
0.02
0.00
sigma
1
2
3
4
theta
−2
0
2
4

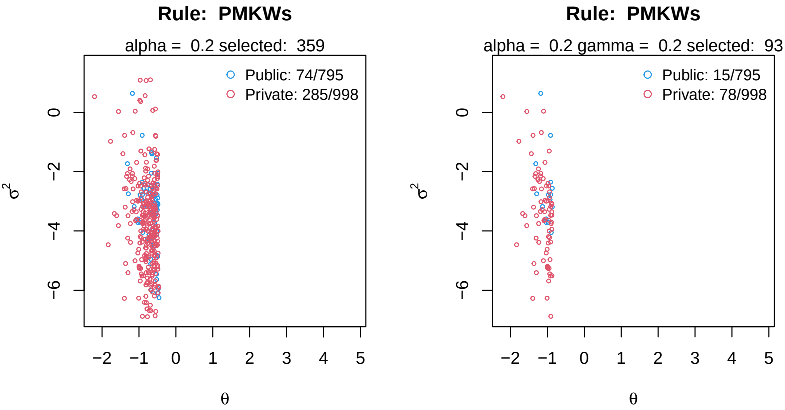


Estimated Location smoothed G with Heterogenous Known Variance
Density
0.0
0.1
0.2
0.3
0.4
0.5
−2
−1
0
1
2
3
4
5
θ

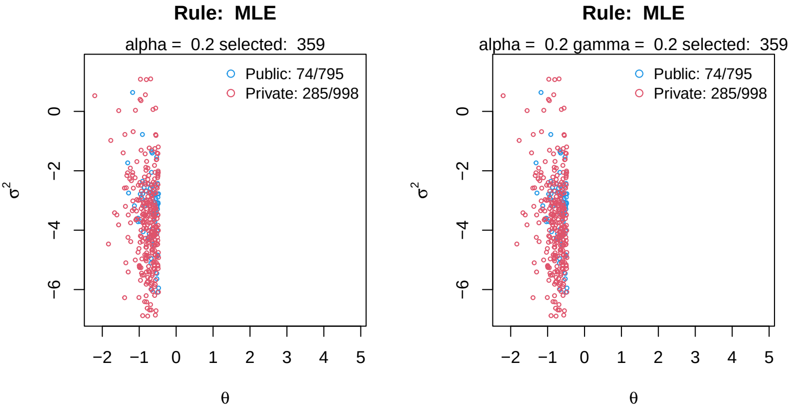
Posterior Tail probability



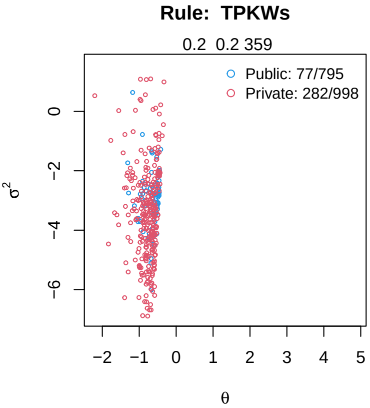
Posterior Mean

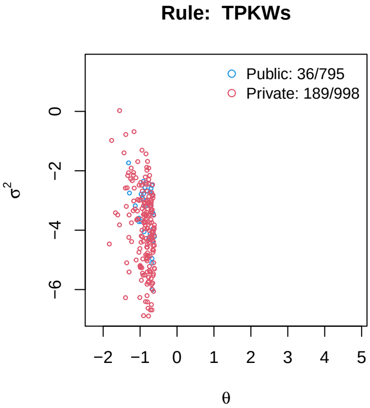


Linear shrinkage

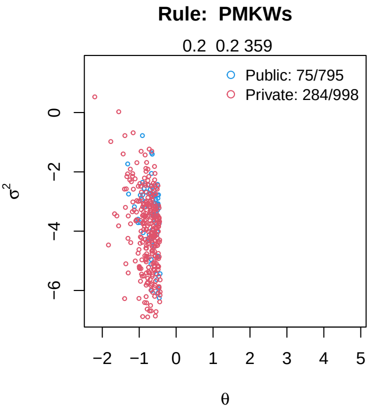


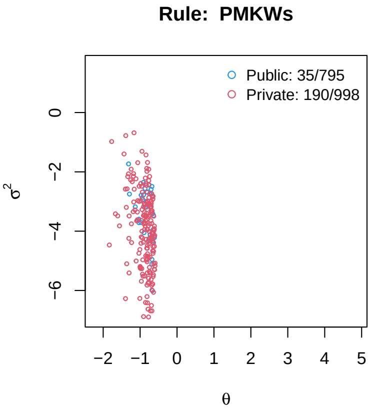
Posterior Tail probability



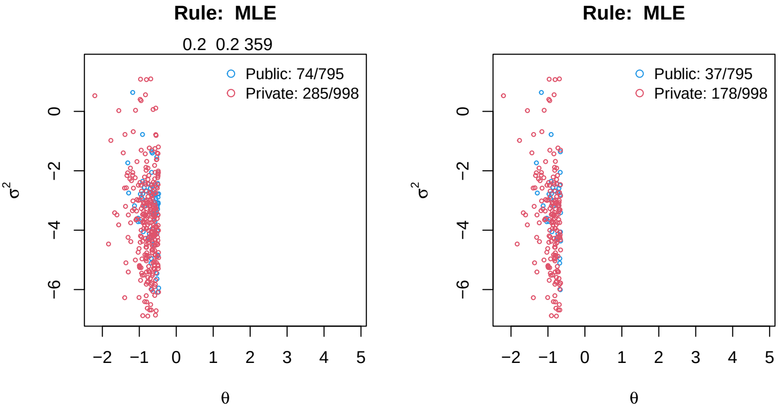


Posterior Mean





Linear shrinkage



Outline

Introduction

Data and Estimation

Empirical Bayes Compound decision: The Selection Problem

Conclusion

Limitation

- Specification
- Endogeneity issue
- Normality assumption on ε_{it}
- Interpretation of the θ_i.

References I

Normality assumption on ε_{it}

Estimate the fixed effect θ_i by

$$\begin{aligned}\hat{\theta}_i =& \frac{1}{T}\sum(\theta_i + \varepsilon_{it} + x_{it}(\beta - \hat{\beta})) \\ \xrightarrow{N\to\infty}& \theta_i + \frac{1}{T}\sum_t \varepsilon_{it}\end{aligned}$$

When T is relatively small (or even fixed), can't use central limit theorem to claim that $\hat{\theta}_i \xrightarrow{d} \mathcal{N}(\theta_i, \frac{\sigma_i^2}{T})$. $\longleftarrow$ Assume that $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_i^2)$.