L'Hopital's (Selection) Rule:

An Empirical Bayes Application to French Hospital Efficiency

Fu Zixuan

Supervised by Thierry Magnac

July 4, 2024

Abstract

Something interesting

Contents

1 Introduction

It is almost of human nature to compare, rank and select. And competition, be it good or bad, emerges in the wake. As invidious as ranking and selection can be, in many cases it is one of the driving forces behind improvement in performances. The society itself is constantly constructing league table as well. It rewards the meritorious and question or even punishes the unsatisfactory. The measure based on which rank is constructed ranges from teacher's evaluation (Chetty et al., 2014), communities' mobility index (Chetty and Hendren, 2018) to firm discrimination (Kline et al., 2022).

The present article extends the practice to the health sectors. To be more specific, it studies the labor efficiency across all hospitals in France. By exploring a comprehensive database called *The Annual Statistics of Health Establishments (SAE)* of French hospitals, I first construct a measure of labor efficiency. Then based on the estimates, we compare the public and private hospitals by selecting the top-performing units. I borrow from the recent developments in Empirical Bayes method to achieve the comparison.

I found that out of the top 20% best performing hospitals, there are roughly 5 times more private units than the public, adjusted by the number of hospitals in each category. The difference is more pronounced when I also control for the expected number of wrongly selected. The takeaway is that public hospitals are in general less efficient than private ones. While the conclusion is in line with that of Croiset and Gary-Bobo (2024) that, now we have a granular perspective on the performance comparison.

The article bridges two fields of interests. The first one is on productivity analysis. The most popular methods in the field are Data Envelopment Analysis (Charnes et al., 1978) and Stochastic Frontier Analysis (Aigner et al., 1977; Meeusen and van Den Broeck, 1977). Yet I abstract from both of them and use the *conditional input demand function* specification stated in Croiset and Gary-Bobo (2024).[1] To put it simply, we estimate a linear function of how much labor input is needed to produce a give list of 8 hospital outputs. I only focus on the employment level of nurses because unlike medical doctors, this is a category that do not suffer from a shortage of labor supply.

The second area of interests is the Empirical Bayes Methods. I lean on a series of work by Jiaying Gu and Roger Koenker, chiefly the following two papers. Gu and Koenker (2017) discussed the usefulness of estimating a prior distribution in baseball batting average prediction. And Gu and Koenker (2023) has formally defined the selection problem as a compound decision on which the estimated prior can be of help as well. Kiefer and Wolfowitz (1956) has shown that non parametric maximun likelihood estimation of the prior is feasible and consistent. The computation of NPMLE is greatly improved by Koenker and Mizera (2014) by leveraging the recent development in convex optimization (Andersen and Andersen, 2010). I will be using the `REBayes` package (Koenker and Gu, 2017) in the estimation, which is based on software `MOSEK` developed by Andersen and Andersen (2010).

In Croiset and Gary-Bobo (2024), the authors argue that public hospital is less efficient than private counterpart in the sense that it would need a smaller size of personnel if it were

[1] I refer the reader to Croiset and Gary-Bobo (2024) for detailed reasons of adopting such an approach.

to use the input demand function of the private hospital, which is the main result of their counterfactuals.

Having roughly replicated the results after doubling the length of the panel, the paper differentiates itself by utilizing classical panel data methods in input demand function estimation, specifically the standard fixed-effect estimation and GMM. Though it is straightforward to include individual fixed effect in specification, estimation is not without challenge. For example, as Croiset and Gary-Bobo (2024) has correctly pointed out, within hospital variation is much smaller than between group variation. The former may be insufficient to obtain credible estimates. I extended the panel length in an attempt to mitigate the problem. Secondly, the strict exogeneity assumption required by the standard within group estimation is questionable. A natural way to relax it is to use the first difference GMM estimator proposed by Arellano and Bond (1991). The high persistency in the regressors poses another challenge of weak instruments. In response, Arellano and Bover (1995); Blundell and Bond (1998)'s system GMM is modified and implemented, and the estimation results are used for the rest of the section.

The benefit of the panel data estimator is that it gives us an estimate of the underlying heterogeneity, which opens door to individual comparisons. However, the fixed effect estimates are generally noisy, rendering the ensuing decision maker hand wavy in making choices. The EB methods are proposed in an attempt to rectify the situation by empirically estimating the prior distribution of the fixed effect.

For example, in Gu and Koenker (2023), we are given the task of selecting the top 20% fixed effect denoted by θ_i. If the θ_i follows a distribution G, this is to say we are selecting those $\theta_i > G^{-1}(0.8)$. The decision rule for individual i is an indicator function δ_i, determining whether i belongs to selection set. The task naturally falls under the compound decision framework pioneered by **?** if we define the loss function of the selection problem in such a way that takes into account the results of all the individual decisions δ_i.

$$\delta^* = \arg\min_{\delta} \mathbb{E}_G \mathbb{E}_{\theta|\hat{\theta}} \left(L_n \right).$$

Since we don't know the true value θ, we minimize the expected compound loss L_n over the distribution of θ given the observed $\hat{\theta}$.

In addition to the capacity constraint of the top 20%, Gu and Koenker (2023) further controls for the number of Type II mistakes made in the selection process. The false discovery rate (FDR) constraint is imposed to ensure that the expected number of wrongly selected units is below a certain level. The FDR constraint is a measure of the proportion of false positives among all the selected units defined as $\mathbb{P}(h_i = 0 | \delta_i = 1) \leq \gamma$.

Being interested in the top performing French hospitals, I define my selection problem as *Left tail selection* because the goal is to choose the bottom 20% of the hospital fixed effect θ_i. A smaller θ_i indicates that less labor input is needed to produce the same amount of output, as compared to hospitals with higher θ_i.

It is worth mentioning that classical empirical Bayes method assumes a parametric form of the prior distribution G which is computationally more attractive. Yet thanks to fast convex optimization algorithms, the non-parametric maximum likelihood estimation is now both

feasible and efficient. Nevertheless, we are completely free from imposing any parametric assumption. In fact, there are two *layers* of distribution. The lower hierarchy is the prior G with $\theta \sim G$ while the higher hierarchy is $\hat{\theta}|\theta \sim P_\theta$. It is when P_θ belongs to the exponential family that the Lindsay (1995) results hold. Usually in application, we need to impose assumptions or perform some transformation such that P_θ is normal. This kind of procedure is often questionable. Often times, researchers resort to asymptotics to justify the normality assumption, which may not be valid in small samples.

The rest of the paper is organized as follows. Section 2 briefly describes the data and lays out the reduced form estimation of the input demand function, treating the number of nurses as the dependent variable and a list of 9 output measures as the regressors. It then applies the classical panel data estimators to the same specification, distinguishing between whether strict exogeneity is assumed. In section 3, I introduce the compound decision framework and the method to non parametrically estimate G. In section 4, I specifically define the selection problem following the framework of Gu and Koenker (2023). Section 5 follows with a comparison of the different selection outcome. I try to draw preliminary conclusion on the comparative performance of public and private hospitals. Section 6 discusses potential issues and concludes.

2 Data and Estimation

2.1 Data

The data we used is called *The Annual Statistics of Health Establishments (SAE)*[2]. It is a comprehensive, mandatory administrative survey and the primary source of data on all health establishments in France. We primarily exploited the report of healthcare output (a list of 10 output measure) and labor input (registered and assistant nurses). The panel covers 9 years from 2013 to 2022, with 2020 missing due to the pandemic. The SAE data only distinguishes 3 types of units based on legal status.

1. *Public hospitals*
2. *Private for-profit hospitals*
3. *Private non-profit hospitals*

Following Croiset and Gary-Bobo (2024), I further single out/distinguish the *public teaching hospitals* from the public hospitals since it is intrinsically different from others in the French healthcare system.

As shown in Table 1, The number of hospitals in normal public, private for-profit, private non-profit are roughly equal and stable over the years. With respect to the teaching hospitals, it is worth mentioning that they not only provide treatments like other types of hospitals but spend a significant amount of resources on doctor training and research as well. Since

[2]La Statistique annuelle des établissements (SAE)

Year	Teaching	Normal Public	Private For Profit	Private Non Profit	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

Table 1: Number of hospitals in each category, 2013-2022

Output	Teaching	Normal Public	Private For Profit	Private Non Profit	Total
STAC inpatient	25.17%	43.09%	23.64%	8.1%	100%
STAC oupatient	18.4%	19.46%	52.95%	9.18%	100%
Sessions	14.49%	21.96%	34.4%	29.16%	100%
Outpatient Consultations	36.8%	52.45%	0.23%	10.52%	100%
Emergency	21.4%	60.06%	13.37%	5.17%	100%
Follow-up care and Long-term care	7.6%	19.47%	37.95%	34.98%	100%
Home hospitalization	13%	17.38%	12.4%	57.22%	100%
Psychiatry stays	6.53%	62.26%	12.93%	18.28%	100%

Table 2: Hospital share of output, 2013-2022

teaching hospitals have more missions on top of the regular healthcare provision, it is natural that they are in general larger in size. This latter point can be seen much more clearly we present the hospital's output share. Despite being relatively few in number, their share of output is quite substantial. The difference is more pronounced after being adjusted by the number of hospitals as shown in Table 3.

Moreover, we see that each type of hospital differs in terms of the mix of services they provide. For example, emergency care is mostly taken care of by public hospitals and private hospitals are strong in medical sessions.

2.2 Estimation

Regression without individual fixed effect Let $\log(x_{it})$ be the number of nurses in hospital i at time t, and $\log(y_{it})$ denote a vector of output levels. I estimate

$$\log(x_{it}) = \beta_0 + \beta_1 \log(y_{it}) + \varepsilon_{it} \tag{1}$$

First, having performed the regression separately for each type of hospital, it is without surprise that teaching hospitals have very different coefficients, as shown in Table 4. In addition to the differences in descriptive statistics from the last section, this intrinsic difference in

Output	Teaching	Normal Public	Private For Profit	Private Non Profit	Total
STAC inpatient	66.98%	19.29%	10.25%	3.48%	100%
STAC oupatient	57.91%	10.29%	27.13%	4.67%	100%
Sessions	50.12%	12.7%	20.18%	16.99%	100%
Outpatient Consultations	77.69%	18.64%	0.08%	3.59%	100%
Emergency	62.02%	29.26%	6.31%	2.41%	100%
Follow-up care and Long-term care	33.5%	14.37%	27.31%	24.82%	100%
Home hospitalization	47.83%	10.75%	7.46%	33.96%	100%
Psychiatry stays	29.65%	47.38%	9.6%	13.37%	100%

Table 3: Hospital share of output weighted by the number of hospitals, 2013-2022

For example, the value a_{ij} where i is STAC inpatient and j is teaching hospitals, is calculated by $a_{ij} = \frac{\text{Number of STAC inpatient in teaching hospitals}}{\text{Share of teaching hospitals} \times \text{Total number of STAC inpatient}}$.

input demand functions or equivalently in production function is another sign that teaching hospitals may not be directly comparable to other types of hospitals. For this reason, I will exclude teaching hospitals from the subsequent analysis.

By excluding the teaching hospitals from estimation, it becomes more reasonable to assume that all hospitals share the same set of coefficients, giving rise to the pooled regression results shown in Table 5.

Regression with individual fixed effect Let $\log(x_{it})$ and $\log(y_{it})$ the same as before. In addition, let θ_i be the fixed effect of hospital i. One interpretation of θ_i is the measure of labor *inefficiency*. The smaller the θ_i, the more efficient the hospital is in labor use. The estimate of θ_i will be used to rank and select the hospitals in the next section. The specification now is

$$\log(x_{it}) = \beta_0 + \beta_1 \log(y_{it}) + \theta_i + \varepsilon_{it} \tag{2}$$

I considered 5 types of estimator, within-group, first difference, fist difference GMM, system GMM and just identified system GMM. For the sake of exposition, the linear specification takes the general form

$$y_{it} = x_{it}\beta + \theta_i + \epsilon_{it} \quad \text{where} \quad E[\epsilon_{it}|x_{i1}, \ldots, x_{it-1}, \theta_i] = 0.$$

The system GMM makes use of two types of moment conditions. The first one is that from the first difference GMM estimator,

$$E[x_{i,t-2}(\Delta y_{it} - \beta \Delta x_{it})]$$

where lagged $x_{i,t-2}$ serves as instrument for Δx_{it}. If the persistency in x_{it} is high, that is to say $x_{it} = \alpha x_{i,t-1} + \eta_{it}$ with α close to 1. Then the reduced form relationship between Δx_{it} and $x_{i,t-2}$ is

$$\Delta x_{it} = (\alpha - 1)\alpha x_{i,t-2} + \alpha \eta_{i,t-1} + \eta_{i,t}$$

posing the problem of weak instrument.

Dependent Variable:	Nurses			
	Teaching	Public	Forprofit	Nonprofit
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	3.28***	1.40***	1.41***	1.00***
	(0.123)	(0.099)	(0.041)	(0.059)
STAC inpatient	0.108***	0.328***	0.261***	0.343***
	(0.016)	(0.018)	(0.007)	(0.012)
STAC outpatient	0.131***	0.078***	0.048***	0.046***
	(0.013)	(0.005)	(0.005)	(0.010)
Medical sessions	0.058***	0.049***	0.075***	0.093***
	(0.008)	(0.003)	(0.002)	(0.006)
External consultations	0.018**	0.025***	-0.003	0.002
	(0.009)	(0.004)	(0.006)	(0.005)
Emergency	0.049***	-0.008**	0.034***	0.024***
	(0.004)	(0.003)	(0.002)	(0.004)
Long-term & follow-up	0.058***	0.051***	0.057***	0.117***
	(0.005)	(0.003)	(0.003)	(0.007)
Home care	0.020**	0.029***	0.049***	-0.012
	(0.010)	(0.003)	(0.007)	(0.009)
Psychiatric care	0.029***	0.071***	0.076***	0.049***
	(0.005)	(0.004)	(0.007)	(0.017)
Fit statistics				
Observations	1,123	5,260	4,415	2,604
R^2	0.780	0.863	0.742	0.754

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 4: Separate estimation of input demand function, lagged value as IV, 2013-2022

Dependent Variable:	Nurses	
	Dummy	Dummy IV
Model:	(1)	(2)
Variables		
Constant	1.51***	1.50***
	(0.025)	(0.028)
STAC inpatient	0.291***	0.290***
	(0.004)	(0.005)
STAC outpatient	0.048***	0.048***
	(0.003)	(0.004)
Medical sessions	0.068***	0.068***
	(0.002)	(0.002)
External consultations	0.025***	0.028***
	(0.002)	(0.002)
Emergency	0.019***	0.018***
	(0.001)	(0.001)
Long-term & follow-up	0.066***	0.067***
	(0.002)	(0.002)
Home care	0.026***	0.025***
	(0.002)	(0.003)
Psychiatric care	0.072***	0.071***
	(0.003)	(0.004)
Private Forprofit	-0.258***	-0.245***
	(0.024)	(0.027)
Private Nonprofit	-0.178***	-0.160***
	(0.020)	(0.022)
Fit statistics		
Observations	14,067	12,279
R^2	0.820	0.821

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 5: Pooled regression with dummy variables, lagged value as IV, 2013-2022

The second moment condition makes another assumption, requiring that the correlation between x_{it} and θ_i is the same as that between $x_{i,t-1}$ and θ_i,

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

where the current level x_{it} is instrumented by lagged first difference $\Delta x_{i,t-1}$.

It is obvious that there's a large difference between the first two estimators and the GMM ones, a sign that the exogeneity assumption may not be valid. Second, the first difference GMM estimates gives mull results, possibly due to weak instruments. Though the estimate from system GMM looks more hopeful, the sargan-hansen test almost rejects over-identification null hypothesis for sure, indicating that some moment conditions are not in accordance with each other. The fifth just identified GMM only makes use of the second type of moment conditions from system GMM, abstracting from over-identification issue. Though the issues of weak instrument, rejection of over-identification are intriguing problems, I will set them aside for future investigation since the focus of the paper is more empirical bayes application. Believing in the validity of the assumption $\mathbb{E}\left[\Delta x_{i,t-1}(\theta_i + \varepsilon_{i,t})\right] = 0$, I will take as given the estimation results from the last column of Table **??** and proceed to the next section.

3 Empirical Bayes and Selection Problem

3.1 Compound decision framework

The idea of compound decision is pioneered by Robbins (1956), which takes into account the consequences of all individual decisions. Consider the case where each individual unit has an unobserved parameter θ_i. We are given a list of estimates $\hat{\theta}_i$ for each θ_i.

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n) \quad \text{where} \quad \hat{\theta}_i|\theta_i \sim P_{\theta_i}$$

For the moment, I will be agnostic to the specific decision to make and denote the decision rule by δ.

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

The next step is to define the loss function as the objective function to minimize. Since I care about the **collective performance** of my decision, I will define the loss function such that the attention to the compound decision is reflected. A natural choice would be to aggregate the individual losses. Therefore, the compound loss function is defined as

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

Correspondingly, the compound risk is defined as the expectation of compound loss

$$R_n(\theta, \delta(\hat{\boldsymbol{\theta}})) = \mathbb{E}_{\theta|\hat{\theta}}[L_n(\theta, \delta(\hat{\boldsymbol{\theta}}))]$$

Dependent Variable:	Nurses		
	Within Group	First Difference	System GMM
Model:	(1)	(2)	(3)
Variables			
STAC inpatient	0.10***	0.07***	0.51***
	(0.00)	(0.01)	(0.02)
STAC outpatient	0.02***	0.01***	0.06***
	(0.00)	(0.00)	(0.02)
Medical sessions	0.02***	0.02***	0.04***
	(0.00)	(0.00)	(0.01)
External consultations	0.00	0.00	0.07***
	(0.00)	(0.00)	(0.01)
Emergency	0.01***	0.01	−0.07**
	(0.00)	(0.00)	(0.03)
Long-term & follow-up	0.01***	0.01***	0.02
	(0.00)	(0.00)	(0.02)
Home care	0.01***	0.02**	0.01
	(0.00)	(0.01)	(0.02)
Psychiatric care	0.02***	0.01	0.04
	(0.00)	(0.01)	(0.03)
Fit statistics			
n	1690	1690	1690
T	9	9	9

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

Dependent Variable:	Nurses		
	Within Group	First Difference	System GMM
Model:	(1)	(2)	(3)
Variables			
STAC inpatient	0.10***	0.07***	0.74***
	(0.00)	(0.01)	(0.08)
STAC outpatient	0.02***	0.01***	−0.07
	(0.00)	(0.00)	(0.04)
Medical sessions	0.02***	0.02***	0.07***
	(0.00)	(0.00)	(0.02)
External consultations	0.00	0.00	0.03
	(0.00)	(0.00)	(0.02)
Emergency	0.01***	0.01	−0.11*
	(0.00)	(0.00)	(0.05)
Long-term & follow-up	0.01***	0.01***	−0.04
	(0.00)	(0.00)	(0.05)
Home care	0.01***	0.02**	0.04
	(0.00)	(0.01)	(0.06)
Psychiatric care	0.02***	0.01	−0.09
	(0.00)	(0.01)	(0.19)
Fit statistics			
n	1690	1690	1690
T	9	9	9

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

We further restrict our attention to the separable decision rule $\delta(\hat{\boldsymbol{\theta}}) = \{t(\hat{\theta}_1), \ldots, t(\hat{\theta}_n)\}$. In order to make the connection with the Bayesian view under which we assume that $\theta \sim G$, we can rewrite the compound risk as

$$R_n(\theta, \delta(\hat{\boldsymbol{\theta}})) = \int \int L(\theta_i, t(\hat{\theta}_i)) dP_{\theta_i}(\hat{\theta}_i) dG_n(\theta)$$

where $G_n(\theta)$ is the empirical distribution of θ. [3]

The Frequentist and Bayesian views differ slightly here in the definition of risk. The original compound decision formulation keeps the empirical distribution G_n in compound risk while the Bayesian risk replaces it with the prior distribution G. On a side note, the two views are somewhat related to the two assumptions in the fixed/random effect terminology, in the sense that the fixed effect view treats θ_i as fixed unknown parameters while the random effect view treats $\boldsymbol{\theta}$ as a random draw from a distribution G. However, in our context, it has nothing to do with whether θ_i is correlated with x_{it}.

The last step is to find the decision rule δ^* that minimizes the risk

$$\delta^* = \arg\min_{\delta} R_n(\theta, \delta(\hat{\boldsymbol{\theta}})) \tag{3}$$

subject to any constraints that we may have. Since G is unknown, whether we use G_n or G in risk does not matter much. In the rest of the section, I adopt the Bayesian risk as the objective of minimization and impose constraints relevant to the selection problem. Now I will turn to the non-parametric estimation of the prior distribution G.

3.2 Estimate G

Parametric G Most literature has imposed a parametric form of G. In the case of a Gaussian G, recall the hierarchical model

$$\hat{\theta}_i | \theta_i, \sigma_i \sim P_{\theta_i}$$
$$\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$$

There are two hyperparameters to be estimated μ_θ and σ_θ.

If we compare the performance of posterior mean estimator $\theta^* = \mathbb{E}\left[\theta | \hat{\theta}\right]$ with the original estimate $\hat{\theta}$, James and Stein (1992) has shown that there's always an improvement in the average performance if we assume G is Gaussian and replace it with an estimate $\hat{G}$. If we relax the normality assumption on G and adopt a NPMLE estimation as established by Kiefer and Wolfowitz (1956), there could be further improvements. For example, Jiang and Zhang (2009) has proven that a plugged in θ^* with a NPMLE $\hat{G}$ is asymptotically optimal among all separable estimators. A comparison between the parametric and non parametric $\hat{G}$ is demonstrated in Gilraine et al. (2020) on their teacher value added application.

[3] $E_{G_n}(f(x)) = 1/n \sum_i f(x_i)$

NPMLE G The initial NPMLE estimator defined in Kiefer and Wolfowitz (1956) takes the following form

$$\hat{G} = \arg\min_{G \in \mathcal{G}} \left\{ -\sum_{i=1}^{n} \log g(y_i) | g(y_i) = \int \mathbb{P}(y_i|\theta) dG(\theta) \right\}$$

where $\mathbb{P}(y_i|\theta)$ is the probability density function of y_i conditional on the true parameter θ $\longrightarrow$ $g(y_i)$ is the marginal pdf of y_i.

Though this is a convex optimization problem with strictly convex objective and a convex constraint set, it is of infinite dimension. In order to solve the primal problem, it is necessary to discretize. The algorithm proposed by Koenker and Mizera (2014) has taken advantage of the fixed point iteration method in convex optimization (Andersen and Andersen, 2010), thus greatly improved the computation efficiency over the fixed point EM iteration method by Jiang and Zhang (2009).

3.3 The selection problem

The definition of the selection problem is taken from the work of Gu and Koenker (2023). Instead of focusing on the right tail of the distribution, the top performers in my context corresponds to the left tail. The task at hand is to select the bottom 20% of the θ_i and compare the share of public and private in the meritorious group. This is another perspective/exercises on the public and private sectors different from that of Croiset and Gary-Bobo (2024).

On top of the constraint on the size of the selected group (20%), I further impose a constraint on the number of false positive mistakes made in the selection process. This leads to the false discovery constraint at level γ,

$$\frac{\mathbb{E}_G \left[h_i = 0, \delta_i = 1 \right]}{\mathbb{E}_G \left[\delta_i \right]} \leq \gamma$$

where $h_i = 1\left\{\theta_i < \theta_\alpha\right\}$ is the indicator function of whether the unit i is truly below the threshold θ_α. And $\delta_i = 1$ when unit i is selected.

All in all, we can formally define the loss function of selection problem as

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

and the optimal decision rule is given by

$$\begin{aligned}
\delta^* &= \arg\min_{\delta} \mathbb{E}_G \mathbb{E}_{\theta|\hat{\theta}} \left[L(\delta, \theta) \right] \\
&= \mathbb{E}_G \sum_i \mathbb{E}_{\theta|\hat{\theta}}(h_i)(1-\delta_i) + \tau_1 \left(\sum_i (1 - \mathbb{E}_{\theta|\hat{\theta}}(h_i))\delta_i - \gamma \delta_i \right) + \tau_2 \left(\sum_i \delta_i - \alpha n \right) \\
&= \mathbb{E}_G \sum_i v_\alpha(\hat{\theta})(1-\delta_i) + \tau_1 \left(\sum_i (1 - v_\alpha(\hat{\theta}))\delta_i - \gamma \delta_i \right) + \tau_2 \left(\sum_i \delta_i - \alpha n \right)
\end{aligned} \tag{4}$$

Here, the term $\mathbb{E}_{\theta|\hat{\theta}}(h_i)$ is called **posterior tail probability**. It is the probability of i being truly in the bottom $\alpha\%$ given the estimated $\hat{\theta}$. This is a posterior statistics different from the posterior mean $\mathbb{E}_{\theta|\hat{\theta}}(\theta_i)$ because the variable inside the expectation $h_i = 1\{\theta_i < G^{-1}(\alpha)\}$ is specific to the capacity constraint at α level. From the previous section, we have obtained an estimate of the prior distribution G so that we can derive the posterior tail probability $v_\alpha(\hat{\theta}_i)$

$$v_\alpha = P(\theta_i < \theta_\alpha|\hat{\theta})$$

If we know that $\hat{\theta}|\theta \sim P_\theta$ with density function p_θ, the posterior tail probability can be further written down

$$v_\alpha(y_i) = \frac{\int_{-\infty}^{\theta_\alpha} p_{\theta_i}(y_i)dG(\theta_i)}{\int_{-\infty}^{\infty} p_{\theta_i}(y_i)dG(\theta_i)}$$

From now on, the notation $\hat{\theta}_i$ is replaced by y_i. For example, if it follows a normal distribution $y_i|\theta_i \sim \mathcal{N}(\theta_i, \sigma_i)$ as is often the case in application, v_α takes the explicit form

$$= \frac{\int_{-\infty}^{\theta_\alpha} \varphi(y_i|\theta_i, \sigma_i^2)dG(\theta_i)}{\int_{-\infty}^{\infty} \varphi(y_i|\theta_i, \sigma_i^2)dG(\theta_i)}$$

where φ is the density function of y_i conditional mean θ_i and variance σ_i^2.

Here, we have assumed that σ_i is known meaning that P_θ only depends on θ_i. But sometimes σ_i is unknown meaning that P_θ depends on some other parameters. The two cases are distinguished when defining posterior tail probability and constraints.

The two constraints can be preliminarily written out as

- Capacity constraint:

$$\mathbb{P}(v_\alpha > \lambda_1^*) \leq \alpha \Rightarrow \lambda_1^* = H^{-1}(1-\alpha).$$

  Empirically, λ_1^* is found by the empirical cumulative distribution H of v_α.

- False Discovery constraint:

$$\mathbb{P}(\theta < \theta_\alpha|v_\alpha > \lambda_2^*) \leq \gamma \Rightarrow \frac{\sum_i \mathbb{E}[(1-v_{\alpha,i})\delta_i]}{\sum_i E[\delta_i]} \leq \gamma.$$

  We approximate $\mathbb{P}(\theta < \theta_\alpha|v_\alpha > \lambda_2^*)$ by $\frac{\sum_i E[(1-h_i)\delta_i]}{\sum_i E[\delta_i]}$. Then it needs to be shown that $\mathbb{E}[(1-h_i)\delta_i]$ is equivalent to $\mathbb{E}[(1-v_{\alpha,i})\delta_i]$. This is straightforward by the law of iterated expectation. Let $D_i = (Y_i, \sigma_i^2)$ when σ_i^2 is known and $D_i = (Y_i, S_i)$ when σ_i^2 is unknown.

$$\mathbb{E}[(1-h_i)\delta_i] = \mathbb{E}[\mathbb{E}[(1-h_i)\delta_i|D_i]] = \mathbb{E}[\delta_i(1-\mathbb{E}[h_i|D_i])] = \mathbb{E}[\delta_i(1-v_{\alpha,i})]$$

Known variance, $G(\theta)$ The true inefficiency value of hospital i is θ_i, We only observe a sequence of Y_i where

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

The tail probability v_α is a function of y_i only

$$v_\alpha(y_i) = \mathbb{P}(\theta_i < \theta_\alpha | y_i) = \frac{\int_{-\infty}^{\theta_\alpha} \varphi(y_i|\theta_i, \sigma_i^2) dG(\theta_i)}{\int \varphi(y_i|\theta_i, \sigma_i^2) dG(\theta_i)}$$

The cutoff λ^* is determined such that the constraints are satisfied

$$\begin{aligned} \mathbb{P}(v_\alpha > \lambda^*) &\leq \alpha \\ \mathbb{P}(\theta < \theta_\alpha | v_\alpha > \lambda^*) &\leq \gamma \end{aligned}$$

Unknown variance, $G(\theta, \sigma)$ We only a sequence of Y_{it} where

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

Neither θ_i nor σ_i^2 is known. But there exists two sufficient statistics for (θ_i, σ_i) such that

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

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

Now the posterior tail probability v_α is a function of both y_i and s_i

$$\begin{aligned} v_\alpha(y_i, s_i) &= \mathbb{P}(\theta_i < \theta_\alpha | y_i, s_i) \\ &= \frac{\int \int_{-\infty}^{\theta_\alpha} \Gamma(s_i | \frac{(T_i-1)}{2}, \frac{2\sigma_i^2}{T_i-1}) \varphi(y_i | \theta_i, \frac{\sigma_i^2}{T_i}) dG(\theta, \sigma^2)}{\int \int \Gamma(s_i | \frac{(T_i-1)}{2}, \frac{2\sigma_i^2}{T_i-1}) \varphi(y_i | \theta_i, \frac{\sigma_i^2}{T_i}) dG(\theta, \sigma^2)} \end{aligned}$$

The cutoff λ^* is found in the same way as before.

3.4 Selection results

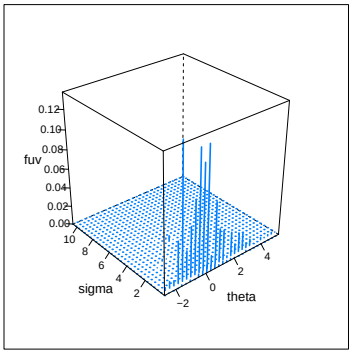
I have selected hospitals with than 6 years of observations. In the end, the sample contains 1661 hospitals, out of which 658 are public hospitals and 1003 are private ones. The Y_{it} in the last section is calculated as $\log(x_{it,\text{nurses}}) - \log(y_{it,\text{output}})\hat{\beta}$. Since the panel is too short to invoke central limit theorem, in order to apply the results above, I am obliged to impose normality assumption on the error term $\varepsilon \sim \mathcal{N}(0, \sigma_i^2)$. Thus Y_{it} follows a normal distribution $\mathcal{N}(\theta_i, \sigma_i^2)$ as the number of hospitals N tends to infinity.

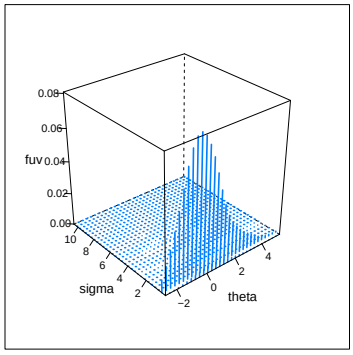
Instead of assuming that σ_i^2 (or the distribution of it) is known, in our setting it seems more reasonable to employ the estimates of it S_i in defining v_α. However, in baseball batting average (Gu and Koenker, 2017), teacher added value (Gilraine et al., 2020) and kidney dialysis center rating (Gu and Koenker, 2023), an estimate of the variance is taken to be the true value. A comparison selection results under the two different assumptions will be presented in this section.

Unknown variance, and TP rules Since we only observe the sample mean Y_i and sample variance S_i, the prior G is a two dimensional distribution on (θ, σ^2). Without assuming independence between the two parameters, we can assume a two dimension gridding /discretization in the defining the convex objective function.

$$\hat{G} = \underset{G \in \mathcal{G}}{\arg\min} \left\{ -\sum_{i=1}^{n} \log g(y_i, s_i) | g(y_i, s_i) = \int \int \mathbb{P}(y_i, s_i | \theta, \sigma^2) dG(\theta, \sigma^2) \right\}$$

This can be solved similarly by the interior point method as in the one dimensional case. The solution is an atomic distribution with fewer than n atoms. It is worth mentioning that the NPMLE method is self-regularizing because the mass points are determined by the solution without recourse to any tuning parameter. Further smoothing is justified by the fact that we have ignored the variability of G. The bandwidth of biweight kernel for smoothing was chosen as the mean absolute deviation from the median of the discrete $\hat{G}$.





With an estimated prior, the tail probability function as well as the constraints are well-defined. Though, given the discrete nature of selection, it is similar to discrete optimization as in knapsack problem. I follow the approach described in Basu et al. (2018) and thus consider only sequentially selecting the units until one constraint is violated.

In Figure 1, I present the results of selecting the top 20% hospitals with or without the FDR constraint at 20%. The selection rule is the posterior tail probability which is explained in the sections above, that is, the solution to the problem defined in 4. The prior G is taken to be the smoothed Kiefer-Wolfowitz estimate.

The left-hand side corresponds to the selection outcome without imposing the FDR constraints while the right-hand side controls the expected FDR at 20%. In the first case,

there are around 10 times more private hospitals in the top 20% while the total number of hospitals is less than twice of the public.

The FDR seems to have only impacted the private hospitals, leaving 18 out of the selection set. A stringent FDR constraint would lead to a smaller set as shown in 2.

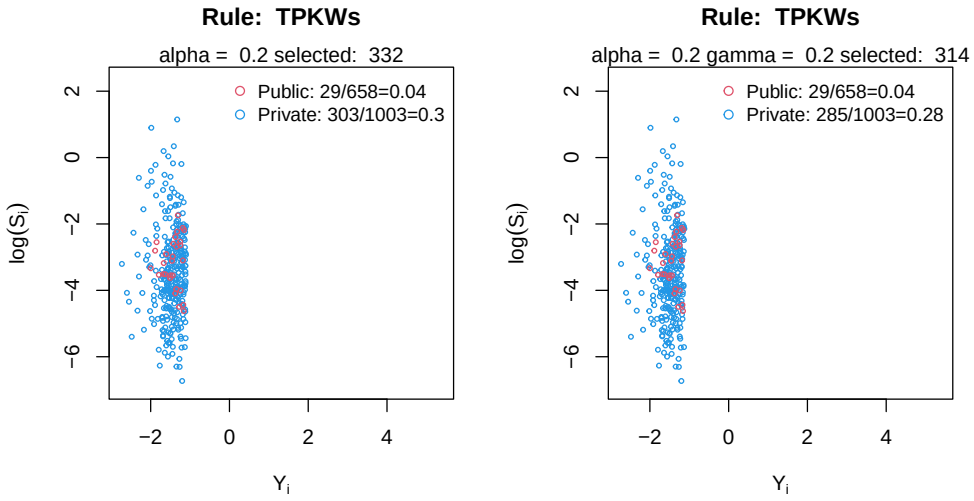


Figure 1: Tail probability rule, capacity 20%, FDR 20%, unknown variance

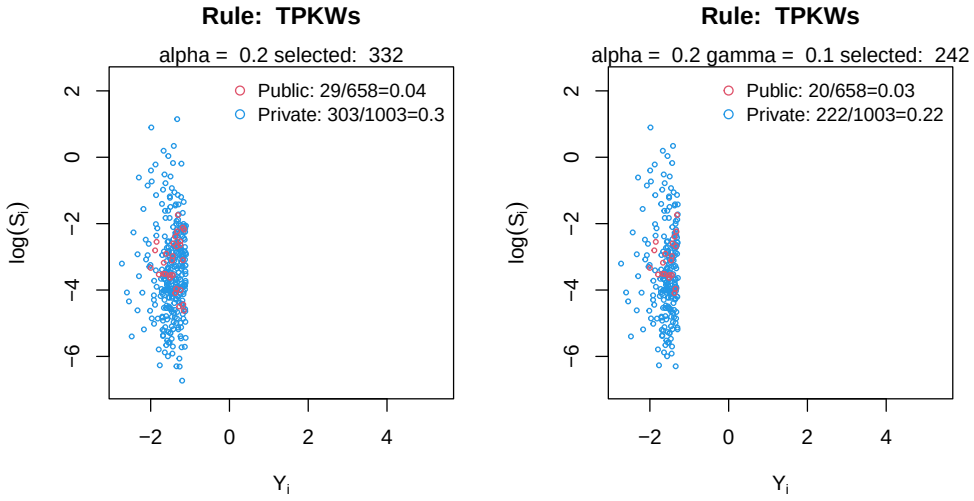
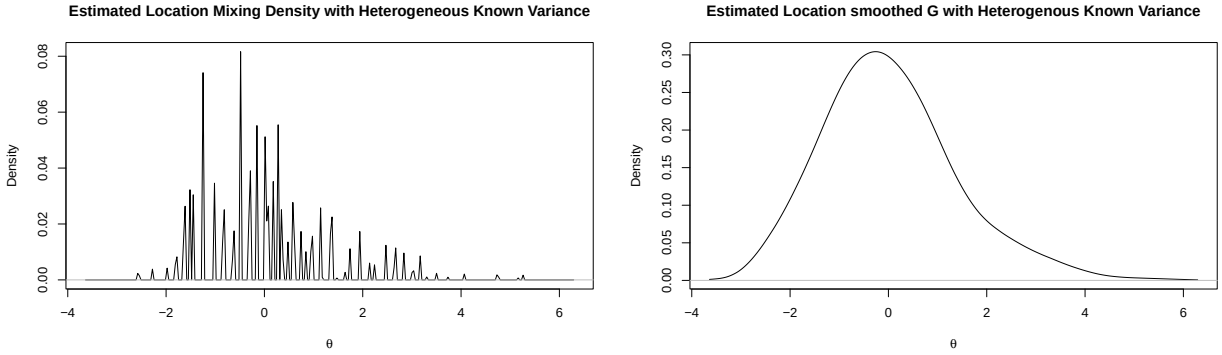


Figure 2: Tail probability rule, capacity 20%, FDR 10%, unknown variance

As of other selection rules using different ranking statistics (posterior mean, MLE face value, James-Stein linear shrinkage), see appendix A.4.1 for an overview.

Known variance, and TP rules In Gu and Koenker (2023), the authors apply the newly proposed selection method to the the selection of kidney dialysis centers studied by (Lin et al., 2006, 2009). However, the focus is on the quality of service, specifically on the mortality rate. Secondly, they assume the predictions of expected mortality is sufficiently accurate such that the variance is known and independent of $\theta \sim G$.



Though not desirable in the present setting, it would be interesting to see what the results would be if I take the S_i as the σ_i^2. For the moment, whether this assumption would lead to a more stringent selection outcome is unclear. Figure 3 presents the outcome under the posterior tail probability rule with smoothed estimated prior. It seems that the with only the capacity constraint, the outcome does not differ much. However, when FDR constraint is combined, the known variance assumption becomes too lenient to incorporate the newly imposed constraint. At the level $\alpha = 0.2$ and $\gamma = 0.2$, the FDR constraint is not binding. While a more stringent FDR constraint does bind as shwon in Figure 4.

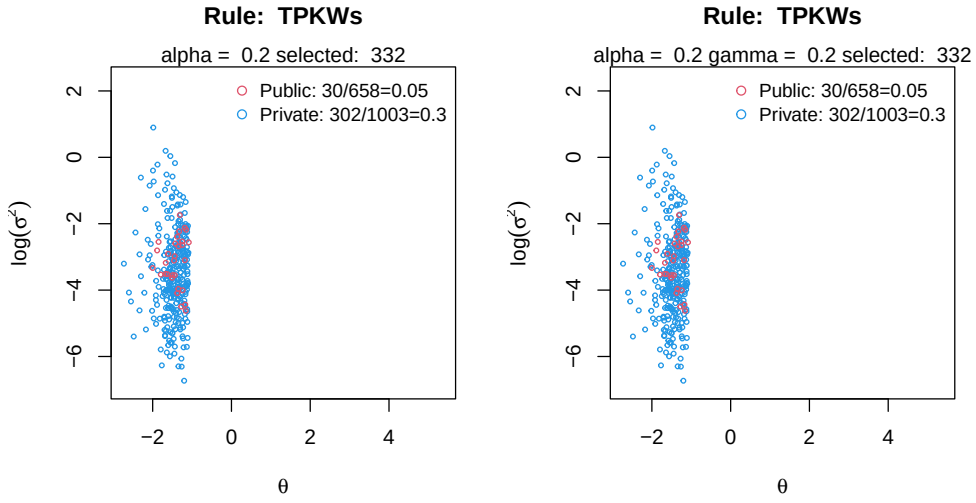


Figure 3: Tail probability rule, capacity 20%, FDR 20%, known variance

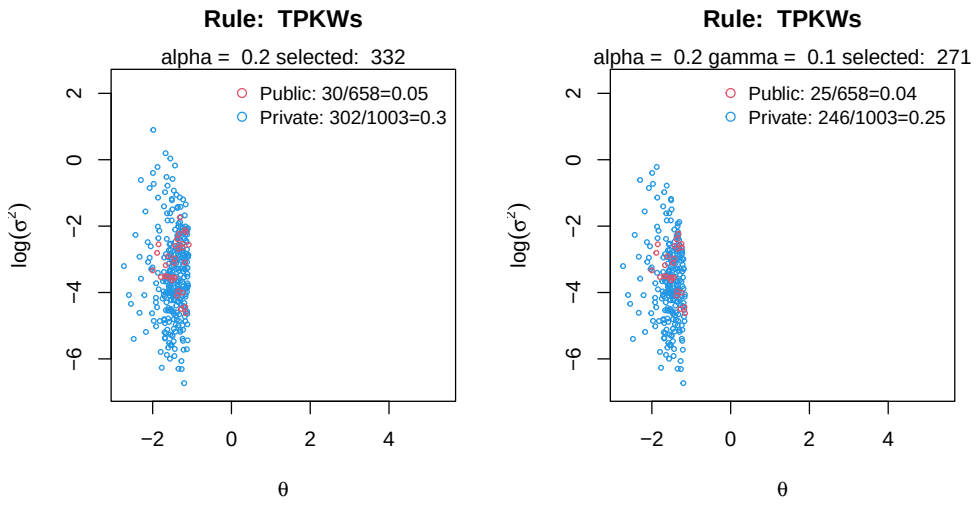


Figure 4: Tail probability rule, capacity 20%, FDR 10%, known variance

Appendix A.4.2 presents the results of other selection rule. In this case, a contour line can be drawn to highlight the differences between ranking statistics.

4 Conclusion

References

Aigner, D., Lovell, C. K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of econometrics*, 6(1):21–37.

Andersen, E. D. and Andersen, K. D. (2010). The mosek optimization tools manual, version 6.0.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of econometrics*, 68(1):29–51.

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*, 113(523):1172–1183.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics*, 87(1):115–143.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, 2(6):429–444.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9):2593–2632.

Chetty, R. and Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228.

Croiset, S. and Gary-Bobo, R. (2024). Are public hospitals inefficient? an empirical study on french data.

Gilraine, M., Gu, J., and McMillan, R. (2020). A new method for estimating teacher value-added. Technical report, National Bureau of Economic Research.

Gu, J. and Koenker, R. (2017). Empirical bayesball remixed: Empirical bayes methods for longitudinal data. *Journal of Applied Econometrics*, 32(3):575–599.

Gu, J. and Koenker, R. (2023). Invidious comparisons: Ranking and selection as compound decisions. *Econometrica*, 91(1):1–41.

James, W. and Stein, C. (1992). Estimation with quadratic loss. In *Breakthroughs in statistics: Foundations and basic theory*, pages 443–460. Springer.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647 – 1684.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906.

Kline, P., Rose, E. K., and Walters, C. R. (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, 137(4):1963–2036.

Koenker, R. and Gu, J. (2017). Rebayes: an r package for empirical bayes mixture methods. *Journal of Statistical Software*, 82:1–26.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685.

Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. (2006). Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis (Online)*, 1(4):915.

Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. (2009). Ranking usrds provider specific smrs from 1998–2001. *Health Services and Outcomes Research Methodology*, 9:22–38.

Lindsay, B. G. (1995). Mixture models: theory, geometry, and applications. Ims.

Meeusen, W. and van Den Broeck, J. (1977). Efficiency estimation from cobb-douglas production functions with composed error. *International economic review*, pages 435–444.

Robbins, H. E. (1956). An empirical bayes approach to statistics. In *Proceedings of the third berkeley symposium on mathematical statistics and probability*, volume 1, pages 157–163.

A Appendix

A.1 Data

The panel is first filtered by the following criteria

1. the number of nurses is positive,
2. at least one of STAC inpatient, STAC outpatient, Sessions is positive,
3. the number of observations is larger than 6

Second, I add one to every variable to avoid null value when taking log.

A.2 NPMLE G

Koenker and Mizera (2014) defined the primal problem as

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

where $T(f)=\int p(y_i|\theta)fd\theta$ and $K(f)=\int fd\theta$.
By discretizing the support,

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

where $A_{ij}=p(y_i|\theta_j)$ and $f=(f(\theta_1),f(\theta_2),\ldots,f(\theta_m))$.
It is straightforward to derive the dual problem

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

A.3 Assumption on $\hat{\theta}_i|\theta_i,\sigma_i$

If our specification and assumptions on exogeneity are correct, the consistency of $\hat{\beta}$ is guaranteed by N's asymptotic. However, our estimate of the fixed effect is

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

When T is relatively small (or even fixed), I am not in a good position to use central limit theorem to claim that $\hat{\theta}_i\xrightarrow{d}\mathcal{N}(\theta_i,\frac{\sigma_i^2}{T})$. A bold assumption that $\varepsilon_{it}\sim\mathcal{N}(0,\sigma_i^2)$ will save me from the T issue, which I will impose for the rest of the section (and abstract from whether that for each i is a testable/reasonable/feasible assumption).

A.4 Comparison of selection rules

A.4.1 Unknown variance

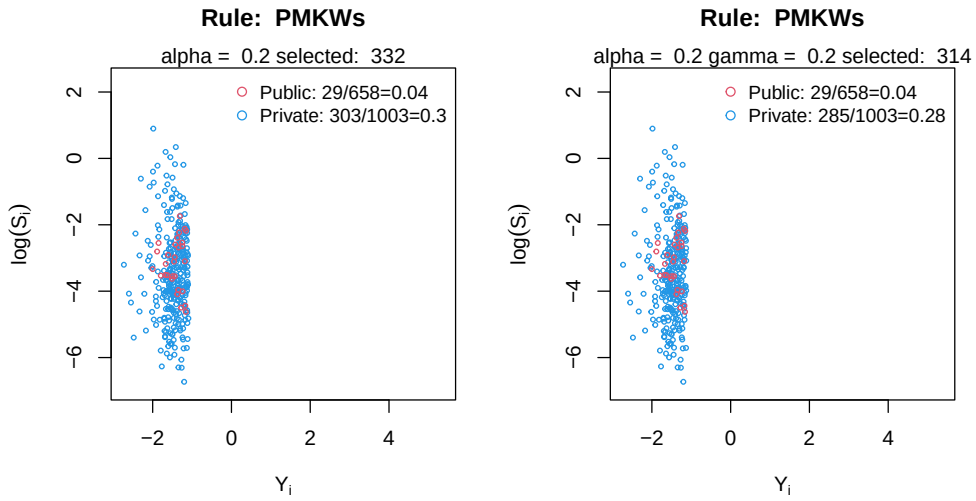


Figure 5: Posterior mean, capacity 20%, FDR 20%, unknown variance

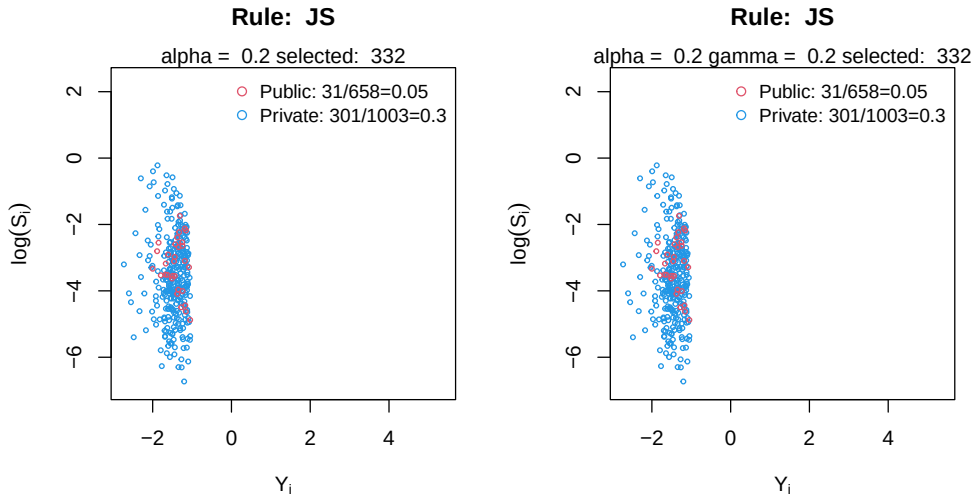


Figure 6: James-Stein Linear Shrinkage, capacity 20%, FDR 20%, unknown variance

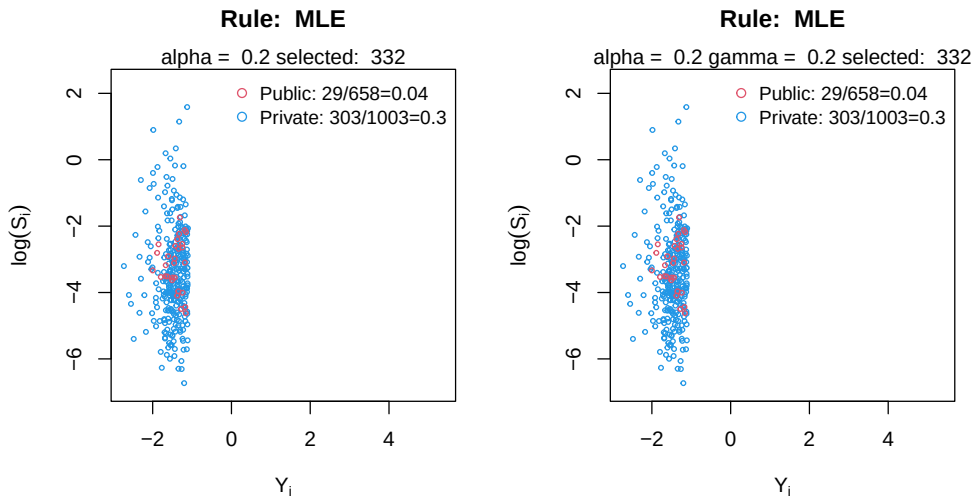


Figure 7: MLE, capacity 20%, FDR 20%, known variance

A.4.2 Known variance

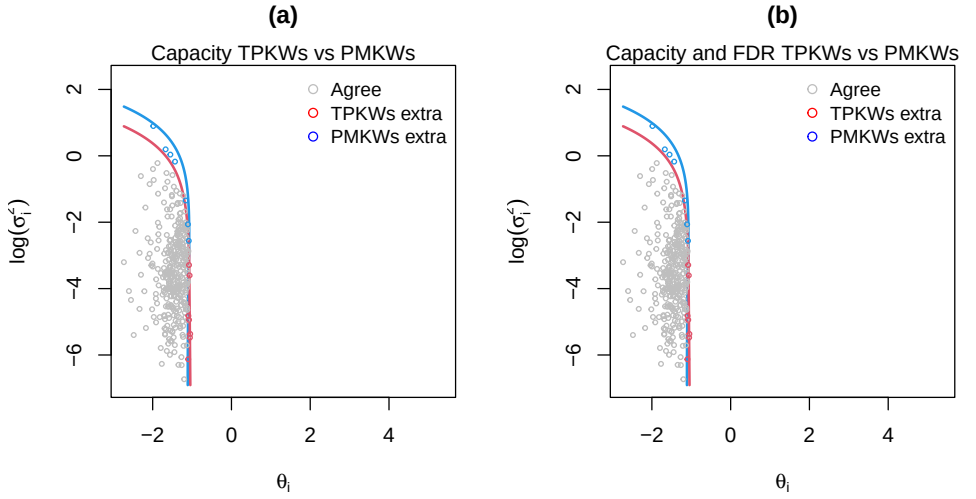


Figure 8: TP VS PM, capacity 20%, FDR 20%, known variance

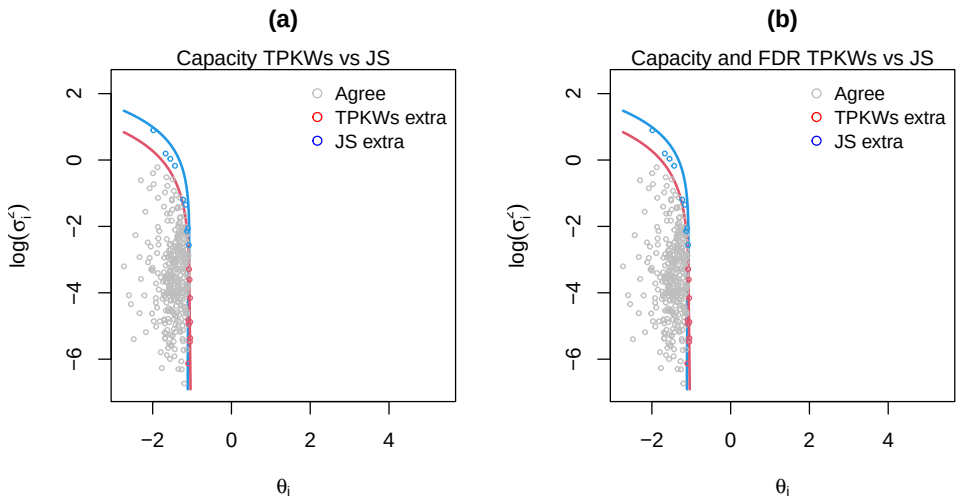


Figure 9: TP VS JS, capacity 20%, FDR 20%, known variance

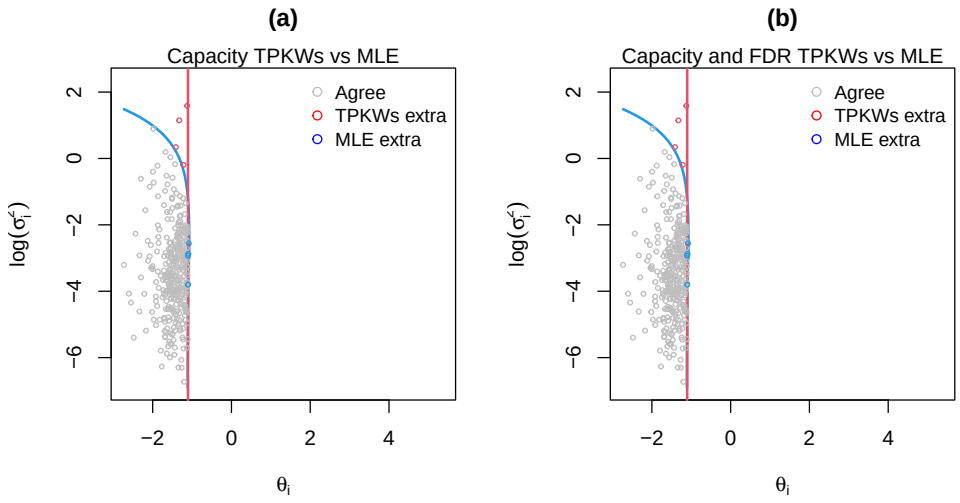


Figure 10: TP VS MLE, capacity 20%, FDR 20%, unknown variance