Introduction

The traditional statistical testing with a null hypothesis H_0 and an alternative H_1 where Neyman-Pearson lemma is used is not adequate for many other tasks. For example, the task of selecting the "best" Gaussian population ("best" refers to the highest mean) and the task of multiple testing (test which ones of all α_i do not belong to the set A).

For the first task, the problem is formulated as choosing the weights w_i so as minimize

$$L(\theta, w) = \theta * - \sum w_i \theta_k / \sum w_i$$

where $\theta*$ is the true maximum of the θ_k and θ_k is the true mean of the k-th Gaussian population. The approach where a preliminary test of equality followed by let $w_i > 0$ if equality is accepted in not **admisible**. While the approach where the picking that with the sample mean is preferred (assuming homogenous variance). Later, there's extension of heterogenous variance.

The **Hierarchical Bayes** approach means that there are two hierichies. The lower hierarchy is the G_α which is a distribution of the unobserved heterogeneity α_i across the population. The upper hierarchy is the H_{α_i} which is the distribution of the estimate/statistics of the unobserved heterogeneity $\hat{\alpha}_i$. Previously, the lower layer is usually assummed a parametric form (Gaussian :) But the recent development in Nonparametric estimation makes the NP estimation of G_α possible.

We will study a specific use case of the Hierarchical Bayes approach–compound decision. We want to improve the performance **compound decision**/minimize the loss of our **compound decision** which means considering the **overall loss** of one's decision. Here, the decision is selecting the most meritous α_i. Applying the Hierarchical Bayes approach (assuming the existence of G and H), we are able to achieve smaller **overall loss**.

Selection (ranking) is ubiquitous in reality. Notable work by Chetty et al on ranking teachers/community by unobserved heterogeneity α_i has pinoeered the field. Later work proposed innovatitve ways to construct confidence interals for the α_i. In a nutshell, they focus on the interval for each α_i while Gu&Koenker focus on the **compound** decision rule of selecting the most meritous α_i. More specifically the **compound** decision rule tries to take into account the lower hierarchy G, rather than only focusing on the H_{α_i}.

Previous related work in **compound rank/selection** constructs posterior mean (the updated mean after having an estimation of G). In Gu&Koenker, they construct **posterior tail probability** which is the updated probability of the α_i being the most $x\%$ meritous given an estimation of G. It seems that from simulation using tail probability is preferred to using means when selecting the top $x\%$ percent.

A remark is that when G and H are both Gaussian, the classical linear shrinkage (embodied in the James-Stein formula) can improve the compound loss of selection decision by a lot. Yet, improvement is said with reference to the naive maximun likelihood estimator. The false discovery rate is still alarmingly high when the **signal to noise ratio** is low. More specifically, when the the variance in H is comparable to the variance in G. When the variance in G is much larger, the ***signal to noise ratio*** is high. It is easier to select the most meritous.
A second (ensuing) remark is that the construction of ranking and the selection decision should be done with caution. While mostly because the data availability issue and the uncontrolable nature of low signal to noise ratio, there is some room for methodological improvement (such as what the Gu&Koenker paper suggested). But still, it's improvement over something really unsatisfactory.

Outlines

1. How we can make use of G in the compound decision and how to nonparametric estimate G_α.
2. If we assume homogeneous variance of H.
3. Assume heterogenous known variance of H.
4. Assume heterogenous unknown variance of H (variance and mean independent or not).

Section 1: Hierachical Bayes in Compound Decision

Define a compound decision and make use of hierarchical bayes in decision rule

Consider the simplest compound decision, we observe a set of $\hat{\theta}_i$, each following a Gaussian distribution $H := N(\theta_i, 1)$ (**higher hierarchy**).

The distribution of θ_i is $G := p(\theta_i = 1) = 1 - p(\theta_i = -1)$ (**lower hierarchy**).

The **loss function** is

$$L(\hat{\theta}, \theta) = 1/n \sum |\hat{\theta}_i - \theta_i|$$

Then $p(\theta = 1|y) = \frac{p\varphi(y-1)}{p\varphi(y-1)+(1-p)\varphi(y+1)}$ where φ is the standard normal density.

The loss function takes into account overall loss (**Compound decision**). The new decision rule based on $p(\theta = 1|y) > 1/2$ takes into account $P(\theta = 1)$ (the lower hierarchy) (**Hierarchical Bayes**).

When G only takes two value, $\hat{G}$ is essentially just $\hat{p}$, which we can find a way to estiamte.
When G takes value in the real line, we are facing a more complex problem. We can non parametrically estimate $\hat{G}$ by utlizing the recent deveopment in convex optimization.
The problem is a **infinite dimensional convex optimization problem** with a strictly convex objective subject to linear constraints.

Estimates of G

Several estiamtes of G + KW with Koenker Mizera: G is atomic, a discrete distribution with fewer than n atoms. + Efron: log-spline sieve approach taht yields smooth estimates of G. + Other smoothed estimates of G.

Loss functions and the corresponding decision rules

If the loss function is

$$L(\hat{\theta}, \theta) = 1/n \sum |\hat{\theta}_i - \theta_i|^2$$

Then the decision rule is given by **Posterior mean**. To be more specific,

$$\delta(y) = E(\theta|y) = y + \frac{f'(y)}{f(y)}$$

where

$$f(y) = \int \varphi(y - \theta) dG(\theta)$$

If the loss function is

$$L(\delta_i, \theta_i) = \lambda 1\{h_i = 0, \delta_i = 1\} + 1\{h_i = 1, \delta_i = 0\}$$

The compound decision rule is given by **Posterior tail probability**. We will go into details in the next section.

Section 2

2.1 Introducing a new problem (Selection), the loss function (Lagrangian multiplier) and the decision rule (Posterior tail probability)

The ***new problem*** is to select the best α% populations. Define θ_α as the α-th quantile of G

$$\theta_\alpha = G^{-1}(1 - \alpha)$$

Thus, we can formulate the problem as a multiple testing problem where $H_0 = \{\theta_i \leq \theta_\alpha\}$ and $H_1 = \{\theta_i \geq \theta_\alpha\}$. Let $h_i = 1\{\theta_i \geq \theta_\alpha\}$, then the ***loss function*** of obersvation i is

$$L(\delta_i, \theta_i) = \lambda 1\{h_i = 0, \delta_i = 1\} + 1\{h_i = 1, \delta_i = 0\}$$

where δ_i is the decision rule of observation i. Note that there are two types of error. The first term is the error of false discovery. The second term is non discovery. The compound decision rule is to minimize the **overall expected loss**

$$E[\sum L(\delta_i, \theta_i)] = \sum E[L(\delta_i, \theta_i)] = \sum (\int_{h_i=0} \int \lambda \delta(y) p(y|\theta) dy dG_\theta + \int_{h_i=1} \int (1 - \delta(y)) p(y|\theta) dy dG_\theta)$$

Recall that $h_i = 0$ is equivalent to $\theta < \theta_\alpha$. Then the expected loss for obersvation i is

$$\int_{-\infty}^{\theta_\alpha} \int \lambda \delta(y) p(y|\theta) dy dG_\theta - \int_{\theta_\alpha}^{\infty} \int \delta(y) p(y|\theta) dy dG_\theta + \int_{\theta_\alpha}^{\infty} \int 1 p(y|\theta) dy dG_\theta = A - B + \alpha$$

For $A - B$, we exchange the order of intergration and get

$$A - B = \int \lambda \delta(y) \int_{-\infty}^{\theta_\alpha} p(y|\theta) dG_\theta dy - \int \delta(y) \int_{\theta_\alpha}^{\infty} p(y|\theta) dG_\theta dy$$

$$= \int \delta(y) (\lambda \int_{-\infty}^{\theta_\alpha} p(y|\theta) dG_\theta - \int_{\theta_\alpha}^{\infty} p(y|\theta) dG_\theta) dy$$

A remark: when we assume that y is normally distributed, we have

$$p(y|\theta, \sigma) = \phi(y|\theta, \sigma) = \varphi((y - \theta)/\sigma)/\sigma$$

To minimize the expected loss, we want to minimize $A - B$, which gives essentially ***decision rule***

$$\delta(y) = 1\{\lambda \int_{-\infty}^{\theta_\alpha} p(y|\theta) dG_\theta < \int_{\theta_\alpha}^{\infty} p(y|\theta) dG_\theta\}.$$

We can define the posterior tail probability as

$$v_\alpha(y) = P(\theta > \theta_\alpha | y) = \frac{\int_{\theta_\alpha}^{\infty} p(y|\theta) dG_\theta}{\int_{-\infty}^{\theta_\alpha} p(y|\theta) dG_\theta + \int_{\theta_\alpha}^{\infty} p(y|\theta) dG_\theta} = \int_{\theta_\alpha}^{\infty} p(y|\theta) dG_\theta$$

Then the ***decision rule*** is

$$\delta(y) = 1\{v_\alpha(y) > \frac{\lambda}{1 + \lambda}\}$$

Provdied that $v_\alpha(y)$ is monotonic in y, a λ^* can be found such that $P(v_\alpha(Y) > \frac{\lambda^*}{\lambda^* + 1}) = \alpha$

We can see that it is imperative to have an estimate of G which gives us first θ_α then $v_\alpha(y)$.

This λ^* is chosen so that probability of being chosen for each is exactly the capacity constraint α. In reality, the probability of being chosen is approximated by the number of i being chosen divided by the total number of i (from my understanding).

A remark: nestedness is guaranteed here. Nestedness means that if a population i is chosen under α_1 then it is also chosen under $\alpha_2 > \alpha_1$.

A second remark: since $v_\alpha(y)$ is monotonic in y, the ranking is the same as ranking by y and the selection of top α percent is equivalent to selecting the top α percent of y. The result is not interesting but paved the way for the next section where we introduce a penalty on false discovery.

Guarding against the false discovery rate

We define the **marginal false discovery rate** as

$$mFDR = P(\theta < \theta_\alpha | \delta(Y) = 1)$$

Gu&Koenker has shown that if $H := N(\theta_i, 1)$ and $G := N(0, 1)$ The mFDR is alarmingly high especially for small α. A natural question would be to guard against the false discovery rate by incorporating the false discovery into the loss function.

The **new loss function** is defined as followed:

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

If we set $\tau_1 = 0$ then the $E(L(\delta, \theta))$ is essentially the same as what is discussed in the previous section and the decision rule is the same except that $\frac{\lambda}{1+\lambda}$ is replaced by τ_2.

τ_2^* is chosen such that

$$\tau_2^* = \min\{\tau_2 : P(v_\alpha(Y) > \tau_2) - \alpha \leq 0\}$$

Similarly, we can define

$$\begin{aligned} t_2^* &= \min\{t_2 : P(v_\alpha(Y) > v_\alpha(t_2)) - \alpha \leq 0\} \\ &\Leftrightarrow \min\{t_2 : P(Y > t_2) - \alpha \leq 0\} \\ &\Leftrightarrow \min\{t_2 : \int P(Y > t_2 | \theta, \sigma) dG(\theta, \sigma) - \alpha \leq 0\} \\ &\Leftrightarrow \min\{t_2 : \int (1 - \Phi((t_2 - \theta)/\sigma)) dG - \alpha < 0\} \end{aligned}$$

If we set $\tau_2 = 0$ then the problem is minimizing the expected numbero of non discoverie subject to the constraint that the marginal FDR rate is controlled at γ.

$$E[\sum(1 - h_i)\delta_i]/E[\sum \delta_i] \leq \gamma$$

which is equivalent to

$$\frac{E((1 - h_i)\delta_i)}{E(\delta_i)} \leq \gamma$$

$$\Leftrightarrow \frac{\int\int P(Y > t_1, h_i = 0 | \theta, \sigma) dG = \int\int_{-\infty}^{\theta_\alpha} P(Y > t_1 | \theta, \sigma) dG}{\int P(Y > t_1 | \theta, \sigma) dG} \leq \gamma$$

Question: does the left hand side represent false discovery rate? The decision rule will be

$$\delta(y) = 1\{v_\alpha(y) > \tau_1(1 - v_\alpha(y) - \gamma)\}$$

If none of τ is zero, then we have the expected loss function as

$$\min_{\delta(y)} E[\sum h_i(1 - \delta_i)] + \tau_1 E(\sum\{(1 - h_i)\delta_i - \gamma\delta_i\}) + \tau_2 E(\sum h_i - \alpha n)$$

A remark: given the discrete nature of the problem, it looks like knapsack problem. We will consider a relaxed version, where units are selected sequentially until one or the other constraint would be violated. (This is not the same as the original problem because the minimum can be achieved while none of the constraint is violated. . . ?)

2.3 Consider heterogenous known variance of H

Similar to the previous section, yet now posterior tail probability becomes $v_\alpha(y,\sigma) = \int_{\theta_\alpha}^{\infty} p(y|\theta,\sigma)dG$ where σ is known. T

τ_2^* is chosen such that

$$\begin{aligned}\tau_2^* &= \min\{\tau_2 : P(v_\alpha(Y,\sigma) > \tau_2) - \alpha \leq 0\} \\ &= \min\{\tau_2 : P(v_\alpha(Y,\sigma) > v_\alpha(t_2(\tau_2,\sigma),\sigma)) - \alpha \leq 0\} \\ &\Leftrightarrow \min\{\tau_2 : \int_\sigma P(Y > t_2(\tau_2,\sigma)|\sigma)dG_\sigma - \alpha \leq 0\} \\ &\Leftrightarrow \min\{\tau_2 : \int (1 - \Phi(t_2(\tau_2,\sigma) - \theta/\sigma))dG(\theta,\sigma)\}\end{aligned}$$

The other τ_1 is chosen in the similar way. Now the threshold value is indirectly affected by σ because τ is chosen such that $P(v_\alpha(Y,\sigma) > \tau_2) \leq \alpha$ and $v_\alpha(y,\sigma)$ is affected by σ.

Section 3: Examples with different G

3.1 Gaussian G

Since $y|\theta,\sigma^2 \sim N(\theta,\sigma^2)$ (higher hierachy) and $\theta|\sigma_\theta^2 \sim N(0,\sigma_\theta^2)$ (lower hierarchy), we have the marginal distribution of $y|\sigma^2,\sigma_\theta^2 \sim N(0,\sigma^2+\sigma_\theta^2)$. And $\sigma \sim H$ with density $h(\sigma)$. The joint density of (y,θ) takes the form

$$f(y,\sigma) = f(y|\sigma)h(\sigma) = \phi(y|0,\sigma^2+\sigma_\theta^2)h(\sigma)$$

The posterior probability

$$v_\alpha(y,\sigma) = P(\theta > \theta_\alpha|y,\sigma) = \int\int_{\theta_\alpha}^{\infty} p(y|\theta,\sigma^2)dG(\theta)dH(\sigma^2)$$

Because we are assuming that G is also Gaussian, then $\theta|y,\sigma^2 \sim N(\rho y,\rho\sigma^2)$ where $\rho = \frac{\sigma^2}{\sigma^2+\sigma_\theta^2}$. Then the last step of calculation can be directly obtained by the CDF of the normal distribution

$$1 - F(\theta_\alpha) = 1 - \Phi((\theta_\alpha - \rho y)/\sqrt{\rho\sigma^2}) = \Phi((\theta_\alpha - \rho y)/\sqrt{\rho\sigma^2})$$

In order to get the joint density of $(v_\alpha(y,\sigma),\sigma)$, let's recall a lemma if the density of a random variable X is $f(x)$ and $X = g(Y)$ then the density of Y is $f(g(y))|g'(y)|$.

Previously, we have $f(y|\sigma)h(\sigma)$, and $y = \psi^{-1}(v,\sigma)$. Conditioned on σ, we have $y = \psi^{-1}(v)$. Then, the density of v is found by $f(\psi^{-1}(v)|\sigma)h(\sigma)|\nabla_v\psi^{-1}(v)|$. If we integrate out σ, we get the **marginal density** of v.

The **capacity constraint** is

$$P(v \geq \tau_2^*) \leq \alpha$$

The **marginal false discovery rate constraint** is

$$\frac{\int_{\tau_1}^1 (1-v)f_v(v)dv}{P(v \geq \tau_1)} \leq \gamma$$

which is derived from the definition of mFDR

$$mFDR = P(\theta \leq \theta_\alpha|\delta_\alpha(Y) = 1)$$

Thus, the final $\tau^* = \max\{\tau_1,\tau_2\}$ and the final t is the solution to the equation $v_\alpha(t,\sigma) = \tau^*$.

$$\Phi((\theta_\alpha - \rho t)/\sqrt{\rho\sigma^2}) = \tau^*\theta_\alpha - \rho t = \Phi^{-1}(\tau^*)\sqrt{\rho\sigma^2}$$

The corresponding selection region is $\{(y,\sigma) : v_\alpha(y,\sigma) > \tau^*\} = \{(y,\sigma) : y > t(\tau^*,\sigma)\}$