Unobserved heterogeneity

Compound loss, risk and decision rule

Each individual has an observed α_i which is unobservable. We only observe Y_i as an estimate/sufficient statistics for α_i for n individuals. We know that

$$Y_i|\alpha_i \sim P_{\alpha_i}$$

We care about the **entire** vector

$$\alpha = (\alpha_1, \dots, \alpha_n)$$

and want to have a good estimate of the **entire** vector. This is why we call the problem **compound decision**.

We collect the **entire** observation as

$$Y = (Y_1, \dots, Y_n)$$

The **compound decision rule** is

$$\delta(Y) = (\delta_1(Y), \dots, \delta_n(Y_n))$$

where each $\delta_i(\cdot)$ is a decision rule for /an estimate of α_i with the **entire** vector Y as input.

Since we care about the performance of the **compound decision** (the estimate of the **entire** vector α), we need to define a loss function that combines the loss from each individual decision.

The individual loss is

$$L(\alpha_i, \delta_i(Y))$$

A simple combination is the sum

$$L_n(\alpha, \delta(Y)) = \sum_{i=1}^{n} L(\alpha_i, \delta_i(Y))$$

which we now call **compound loss**.

As usual, **risk is defined as the expected loss**. Thus, the compound risk is

$$\begin{aligned}
R_n(\alpha, \delta(Y)) &= E[L_n(\alpha, \delta(Y))] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[L(\alpha_i, \delta_i(Y))] \\
&= \frac{1}{n}\sum_{i=1}^{n} \int \dots \int L(\alpha_i, \delta_i(y_1, \dots, y_n)) P_{\alpha_1}(dy_1) \dots P_{\alpha_n}(dy_n)
\end{aligned}$$

Given the **compound risk**, our goal is to find a function/decision rule $\delta(\cdot)$ that minimizes the **compound risk**. This is the **optimal** compound decision rule for a given vector Y

$$\boldsymbol{\delta}^*(Y) = \arg\min_{\delta} R_n(\alpha, \delta(Y))$$

If $\delta^*(Y)$ is separable (the linear shrinkage class belongs to this class as well), which means that $\delta_i^*(Y) = \{t(Y_1), \dots, t(Y_n)\}$, the **compound risk** can be written as

$$R_n(\alpha, \delta(Y)) = \frac{1}{n}\sum \int \dots \int L(\alpha_i, \delta(y_1, \dots, y_n)) dP_{\alpha_1}(y_1) \dots dP_{\alpha_n}(y_n) = \int_{\alpha} \int L(\alpha_i, t(y_i)) dP_{\alpha_i}(y_i) dG_n(\alpha)$$

where $G_n(\alpha)$ is the empirical distribution of α.

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

ATTENTION: We don't know the true α_i's so there's no way we know the empirical distribution $G_n(\alpha) = 1/n \sum 1\{\alpha_i < u\}$. We want to non-parametrically estimate $G_n(\alpha)$.

Comparison

Name	Decision rule	Remarks
Naive	$\delta_i(Y) = Y_i$	ignore compound risk
James-Stein	$\delta_i(Y) = (1 - \frac{n-2}{S})Y_i$ with S = $\sum_{i=1}^n Y_i^2$	known as linear shrinkage

Historical view

Up til now, we didn't specify any views on the α. Conventionally in the literature, there are different "philosophical views" on the α_i's. 1. Fixed effect: $\alpha_i, \dots, \alpha_n$ are viewed as fixed unknown parameters. No assumption on distribution of α_i whatsoever. 2. Random effect: $\alpha_i, \dots, \alpha_n$ are viewed as i.i.d. draws (a realization of the random variable)from a common distribution G.

Appendix

James-Stein rule

FIXED EFFECT VIEW Consider all the linear shrinkage estimators of the form

$$\delta(Y) = ((1-b)Y_1, \dots, (1-b)Y_n)$$

In order to proceed, we assume that P_{α_i} is a normal distribution with mean α_i and variance $\sigma^2 = 1$. We specify the loss function as the squared error loss $L(\alpha_i, \delta_i(Y)) = (\alpha_i - \delta_i(Y))^2$.

The **compound risk** is

$$R_n(\alpha, \delta(Y)) = 1/n \sum \int \dots \int (\alpha_i - \delta_i(y_1, \dots, y_n))^2 dP_{\alpha_1}(dy_1) \dots dP_{\alpha_n}(dy_n) = 1/n \sum \int (\alpha_i - (1-b)y_i)^2 dP_{\alpha_i}(y_i) = 1/n \sum \alpha_i^2$$

Thus the **optimal** compound decision rule is

$$b^* = \arg\min_b R_n(\alpha, \delta_b(Y)) = n / \sum E(Y_i^2)$$

which depends on $\sum E[Y_i^2]$ only.

Since $Y_i | \alpha_i \sim N(\alpha_i, 1)$, we have $E[Y_i^2] = 1 + \alpha_i^2$.

An approximation of $\sum E[Y_i^2]$ is $\sum Y_i^2$, or to correct for the degree of freedom, $\frac{n}{n-2} \sum Y_i^2$.

This is the J-S rule where the optimal shrinkage term is approximated by data.

We call this the **frequentist view** which corresponds to the idea of **fixed effect**.

RANDOM EFFECT VIEW

> We can take the **Bayesian view** (idea of **random effect**) and assume that α_i are i.i.d. draws from a common distribution G.

We assume that $\alpha_i \sim G = N(0, A)$ The **Bayesian risk$ is

$$E_G(E_\alpha(L(\alpha, \delta(Y))))$$

Given a vector α the expected loss is$E_\alpha(L(\alpha, \delta(Y)))$.
Given a distribution of α the expected loss is $E_G(E_\alpha(L(\alpha, \delta(Y))))$. The optimal decision rule is

$$\delta_i^*(Y) = E(\alpha | Y_i)$$

Given the prior distribution G,

$$E(\alpha|Y_i) = (1 - \frac{1}{1+A})Y_i$$

To approximate/estimate $(A+1)$, we can use

$$S = \sum Y_i^2 \sim (A+1)\chi^2(n)$$

where $E(\frac{n-1}{S}) = A+1$. Therefore, the empirical optimal decision rule takes the form $\delta_i^*(Y) = (1 - \frac{n-2}{S})Y_i$.

All roads to the same estimator.

Conclusion

1. Fixed effect: JS mimics the optimal linear shrinkage estimator.
2. Random effect: Js mimics the optimal Bayesian estimator when G=N(0,A). > Pretty restrictive class & assumptions.