WEEK 7: DEC. 15, 2022

Note Writer: Yu-Chieh Kuo[†]

[†]Department of Information Management, National Taiwan University

Recap

Bayesian Methods

Given the *i.i.d.* data $y_1, \cdots, y_n$, the density/likelihood $f(y_1, \cdots, y_n \mid \theta) = \prod_{i=1}^n f(y_i \mid \theta)$ (Prob (data | θ)), the prior density $g(\theta)$, and the marginal density $\int_\theta f(y_1, \cdots, y_n \mid \theta)g(\theta)d\theta = f(y_1, \cdots, y_n)$ (Prob (data)), we can conduct the posterior density

$$\overbrace{f(\theta \mid y_1 \cdots y_n)}^{\text{Prob}(\theta|\text{data})} = \frac{f(y_1, \cdots, y_n \mid \theta)g(\theta)}{f(y_1, \cdots, y_n)}$$

by Baye's rule.

Bayesian, Empirical Bayes, and James-Stein Shrinkege

(This part refers to the Ch1, Large-Scale Inference, Bradley Efron. Note that the usage of $g()$ and $f()$ below is different with above.)

Settings

The setting is described as following. The parameters follow the distribution of density $\mu \sim g(\ )$ (prior), and the data is given by $z \sim f(z \mid \mu)$ (likelihood), the marginal density is, therefore,

$$f(z) = \int_\mu f(z \mid \mu)g(\mu)d\mu.$$

The posterior density is computed by

$$g(\mu \mid z) = \frac{g(\mu)f(z \mid \mu)}{f(z)}.$$

The statistical problem is that we have independent observations from the distribution $z_1, \cdots, z_N$, and we want to estimate $\mu_1 \cdots, \mu_N$.

Assumptions

The prior of parameters follows $\mu \sim \mathcal{N}(0, A)$, and the likelihood of observations is $z \mid \mu \sim \mathcal{N}(\mu, 1)$. Under such assumptions, we can find the marginal distribution follows the normal distribution with the different variance $z \sim \mathcal{N}(0, 1 + A)$. Moreover, the posterior is

$$\mu \mid z \sim \mathcal{N}(Bz, B),$$

where $B \equiv \frac{A}{A+1}$. Note that the posterior mean is Bz.

We then estimate the parameters by maximum likelihood estimation (least information) of μ_i:

$$\begin{aligned}
\hat{\mu}_i^{ML} &= z_i \\
\mathbb{E}\left[\hat{\mu}_i^{ML} \mid \mu_i\right] &= \mathbb{E}[z_i \mid \mu_i] = \mu_i
\end{aligned}$$

Another approach is the Bayesian estimation (having information of prior) of μ_i, which calculates the posterior mean:

$$\hat{\mu}_i^{Bayes} = Bz_i = \frac{A}{A+1} z_i.$$

Note that the prior here is known.

Additionally, the Empirical Bayes estimation (partial information) of μ_i is also approachable. Note that the prior and A here are unknown. That is,

$$\hat{\mu}_i^{EB} = \hat{B} z_i,$$

when $\hat{B}$ is an estimation of B.

Loss function

We define the loss function to evaluate the performance of estimations. Define

$$\mu \equiv \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_N \end{pmatrix} \quad \text{and} \quad \hat{\mu} \equiv \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_N \end{pmatrix},$$

we yeild the loss function

$$L(\mu, \hat{\mu}) = ||\hat{\mu} - \mu||^2 = \sum_{i=1}^{N} (\hat{\mu}_i - \mu_i)^2.$$

In addition, we define the risk function as the conditional expectation of loss function:

$$R(\mu) = \mathbb{E}[L(\mu, \hat{\mu}) \mid \mu].$$

ML approach

The estimated μ under the ML approach is

$$\hat{\mu}^{ML} = z \equiv \begin{pmatrix} z_1 \\ \vdots \\ z_N \end{pmatrix},$$

and the risk function is therefore

$$\begin{aligned}
\mathbb{E}[L(\mu, \hat{\mu}) \mid \mu] &= \mathbb{E}\left[\sum_{i=1}^{N} (\hat{\mu}_i - \mu_i)^2 \mid \mu\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N} (z_i - \mu_i)^2 \mid \mu\right] \\
&= n.
\end{aligned}$$

Therefore, the overall Bayes risk is

$$\mathbb{E}[\mathbb{E}[L(\hat{\mu}, \mu) \mid \mu] \mid A] = N.$$

Bayesian approach

The estimated μ under the Bayesian approach is

$$\hat{\mu}^{Bayes} = Bz = \frac{A}{A+1}z,$$

and the risk function is

$$\begin{aligned}
\mathbb{E}[L(\mu,\hat{\mu}) \mid \mu] &= \mathbb{E}\left[\sum_{i=1}^{N}(Bz_i - \mu_i)^2 \mid \mu\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\left(B^2z^2 + \mu_i^2 - 2Bz_i\mu_i\right) \mid \mu\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\left(B^2z_i^2 + B^2\mu_i^2 - 2B^2z_i\mu_i + \mu_i^2 - B^2\mu_i^2 - 2Bz_i\mu_i + 2B^2z_i\mu_i\right) \mid \mu\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\left(B^2(z_i-\mu_i)^2 + \mu_i^2 - B^2\mu_i^2 - 2Bz_i\mu_i + 2B^2z_i\mu_i\right) \mid \mu\right] \\
&= B^2N + \left(1 - B^2 - 2B + 2B^2\right)\sum_{i=1}^{N}\mu_i^2 \\
&= B^2N + (1-B)^2\sum_{i=1}^{N}\mu_i^2.
\end{aligned}$$

The overall Bayes risk is therefore

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[L(\mu,\hat{\mu}) \mid \mu] \mid A] &= \mathbb{E}\left[B^2N + (1-B)^2\sum_{i=1}^{N}\mu_i^2 \mid A\right] \\
&= \left(\frac{A}{1+A}\right)^2 N + \left(\frac{1}{1+A}\right)^2 NA \\
&= BN \leq N.
\end{aligned}$$

Comparing the overall Bayes risk between ML approach and Bayesian,

$$N - NB = \frac{1}{A+1}N.$$

That is, if $A = 1$, the difference (or the improvement) is $\frac{1}{2}N$.

Empirical Bayes

Under this setting, B is unknown, and we need to derive an unbiased estimation of B:

$$z \mid \mu \sim \mathcal{N}(\mu, I_N) \quad \text{and} \quad \mu \sim \mathcal{N}(0, AI_N).$$

Note that z, μ, and I_N here are $N\times 1, N\times 1$, and $N\times N$, respectively. Then, the posterior is

$$z \sim \mathcal{N}(0, (A+1)I_N).$$

We define the auxiliary and variance-like

$$S = \sum_{i=1}^{N} z_i^2 \quad \text{and} \quad S \sim (A+1)\chi_N^2,$$

where χ^2_N is the Chi-square with the degree of freedom N. Now, we have

$$\mathbb{E}\left[\frac{N-2}{S}\right] = \frac{1}{A+1} = 1 - B.$$

James-Stein Estimator

We consider a particular empirical bayes estimator called James-Stein estimator, which is defined as

$$\hat{\mu}^{JS} = \left(1 - \frac{N-2}{S}\right)z \quad \text{and} \quad \hat{\mu}_i^{JS} = \left(1 - \frac{N-2}{S}\right)z_i.$$

We can also evaluate James-Stein estimator by calculating the overall Bayes risk:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[L(\mu,\hat{\mu}) \mid \mu] \mid A] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N}\left(\left(1 - \frac{N-2}{S}\right)z_i - \mu\right)^2 \mid \mu\right] \mid A\right] \\ &= N\frac{A}{A+1} + \frac{2}{A+1}.\end{aligned}$$

That is, the order of overall Bayes risk is ML>James-Stein>Bayesian, due to their corresponding size of information.

Theorem. If $N \geq 3$, the James-Stein estimmator $\hat{\mu}^{JS}$ everywhere (for all μ) dominates the ML estimator $\hat{\mu}^{ML}$ in terms of expected total squared error:

$$\mathbb{E}\left[||\hat{\mu}^{JS} - \mu||^2 \mid \mu\right] < \mathbb{E}\left[||\hat{\mu}^{ML} - \mu||^2 \mid \mu\right] \quad \text{i.e.} \quad \mathbb{E}\left[\sum_{i=1}^{N}\left(\hat{\mu}_i^{JS} - \mu_i\right)^2 \mid \mu\right] < \mathbb{E}\left[\sum_{i=1}^{N}\left(\hat{\mu}_i^{ML} - \mu_i\right)^2 \mid \mu\right].$$

The details of proof is not concluded here, but we mention that one common trick is to substract auxiliary term

$$\begin{aligned}\mathbb{E}\left[||\hat{\mu}^{JS} - \mu||^2 \mid \mu\right] &= \mathbb{E}\left[||\hat{\mu}^{JS} - \hat{\mu} + \hat{\mu} - \mu||^2 \mid \mu\right] \\ &= N - \mathbb{E}\left[\frac{(N-2)^2}{S} \mid \mu\right],\end{aligned}$$

which yeilds the result of the theorem. □

Remark. For the *shrinkege estimator*, we have

$$\begin{aligned}\hat{\mu}_i^S &= (1-\xi_i)z_i + \xi_i\left(\frac{\sum_{i=1}^N z_i}{N}\right) \\ &= z_i - \xi_i\left(z_i - \frac{\sum_{i=1}^N z_i}{N}\right).\end{aligned}$$

□

Regularized Estimation

In the least square, we minimize the objective function

$$\min \quad \sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2.$$

A similar one is the Ridge regression, or called L2 regulization:

$$\min \quad \overbrace{\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2}^{\text{normal}} + \lambda \overbrace{\sum_{j=1}^{k}(\beta_j - 0)^2}^{\text{normal prior}}.$$

(We try to give it the Bayesian interpretation) The last term can be regarded as the shrinkege to 0.

Moreover, the LASSO, or called L1 regulization is to

$$\min \quad \overbrace{\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2}^{\text{normal}} + \quad \overbrace{\lambda \sum_{j=1}^{k}|\beta_j - 0|}^{\text{double expected Laplace}} \quad .$$

Review of the Final

Least Square

For the Least Square, we taught the linear and nonlinear models, and the objective function is to minimize

$$\min \quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

For a linear model, almost everything has a closed-form solution. For example,

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) \xrightarrow{p} \mathbb{E}\left[x_i x_i'\right]^{-1}\mathbb{E}[x_i y_i] = \beta_\infty,$$

which yeild the unbiasedness and consistency. Additionally,

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\,\mathbb{E}\left[x_i x_i'\right]^{-1}\right)$$

for $\mathbb{E}\left[e_i^2\right] = \sigma^2$ and $\mathbb{E}\left[e_i e_j\right] = 0$.

For a nonlinear model, we may not have a closed-form soultion. Hence, $\hat{\beta}$ is defined by the FOC:

$$\frac{\partial Q(\hat{\theta})}{\partial \theta} = \frac{-2}{n}\frac{\partial f(x_i;\hat{\theta})}{\partial \theta'}\left(y - f(x_i;\hat{\theta})\right) = 0.$$

If $\hat{\theta} \xrightarrow{p} \theta$ i.e., consistent, then we can use the Mean Value Theorem

$$\frac{\partial Q_n\left(\hat{\theta}\right)}{\partial \theta} = \frac{\partial Q_n(\theta_\infty)}{\partial \theta} + \frac{\partial^2 Q_n(\theta_m)}{\partial \theta \partial \theta'}\left(\hat{\theta} - \theta_\infty\right).$$

Since $\hat{\theta} \xrightarrow{p} \theta_0$, it yeilds $\theta_\infty \xrightarrow{p} \theta_0$, and the distribution is

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, Cov\right).$$

Note that the correct covariance matrix is required.

Maximum likelihood

Given the density $f(y_i \mid x_i, \theta)$, the objective function is

$$Q_\infty(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i \mid x_i, \theta),$$

and $\hat{\theta}_\infty$ is defined by $\frac{\partial Q_\infty(\hat{\theta})}{\partial \theta} = 0$. Suppose $\hat{\theta}^{ML} \xrightarrow{p} \theta_0$, we also use the MVT to derive

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, Cov\right).$$

GMM and Minimum Distance estimators

There are ℓ moment conditions/equations, and we have

$$\mathbb{E}[g(y_i, x_i, z_i, \theta)] = 0 \quad \text{and} \quad \overline{g_n} \equiv \frac{1}{n}\sum_{i=1}^{n} g(y_i, x_i, z_i, \theta) \approx 0,$$

note that ℓ is larger than the number of parameters. The objective here is

$$Q_\infty(\theta) = \overline{g_n}'\hat{W}\overline{g_n},$$

where W is $\ell \times \ell$. Read Hayashi Ch7.3 to see the asymptotic normality, most efficient weight matrix, linear GMM (Ch3), and nonlinear GMM (Ch7.3).

Model selection

We have talked about the in-sample and out-sample prediction errors. The exam will only cover the linear model.

Hypothesis testing

We have discussed the Wald statistics, the Lagrangian multiplier, and the likelihood ratio. All of them above converges to $\chi^2(k)$.

Shrinkege and Bayesian