Classical Statistics beyond ML

Surya Tokdar

When ML fails

- Is LSAT correlated with GPA?
- Data (X_i, Y_i), $i = 1, \cdots, n$
- Model: $(X_i, Y_i) \overset{\text{IID}}{\sim} f$, $f \in \mathcal{F}_2 = \{\text{all densities on } \mathbb{R}^2\}$
- $\rho = \rho(f) = \mathbb{C}\text{or}_{[(X_1,Y_1)|f]}(X_1, Y_1)$.
- $H_0 : \rho = 0$.
- $\hat{f}_{\text{MLE}}$ does not exist! $\sup_{f \in \mathcal{F}_2} \ell_{x,y}(f) = \infty$.

A quick check

- Obs: (x_i, y_i), $i = 1, \cdots, n$,
- Denote $z_i = (x_i, y_i)^T$.
- For every $k \geq 1$ define $f_k = \frac{1}{n} \sum_{i=1}^{n} N_2(z_i, \frac{1}{k} I_2)$
- For large k, $\ell_{x,y}(f_k) \approx n \log k$

But all is not lost

- Let $r_{x,y}$ denote the sample correlation coefficient
- If H_0 were true, we would expect $r_{x,y}$ to be close to zero and if it were untrue, $|r_{x,y}|$ should be not close to zero
- A test: Reject H_0 if $|r_{x,y}| > c$ with

$$\text{size} = \sup_{f \in \mathcal{F}_2, \rho(f)=0} P_{[X,Y|f]}(|r_{X,Y}| > c)$$

- How to calculate or approximate these probabilities?
- Asymptotic helps

Fisher's result

- Let (X_i, Y_i), $i = 1, \cdots, n$ are IID with correlation ρ
- For large n, $r_{X,Y} \sim AN(\rho, \frac{\{1-\rho^2\}^2}{n})$
- A better approximation: $\tanh^{-1}(r_{X,Y}) \sim AN(\tanh^{-1}(\rho), \frac{1}{n-3})$ [where $\tanh^{-1}(r) = \frac{1}{2} \log \frac{1+r}{1-r}$, an increasing function]
- See here for a proof based on the Delta theorem. Try verifying by simulations!

Tests and intervals for correlation

- An approximately size-α test rejects $H_0 : \rho = 0$ if

$$|r_{x,y}| > z(\alpha)/\sqrt{n}$$

- Also, an approximately size-α test rejects $H_0 : \rho = \rho_0$ if

$$\sqrt{n-3} \times |\tanh^{-1}(r_{X,Y}) - \tanh^{-1}(\rho_0)| > z(\alpha)$$

- A $100(1-\alpha)\%$ confidence interval for $\tanh^{-1}(\rho)$ is

$$\tanh^{-1}(r_{x,y}) \mp \frac{z(\alpha)}{\sqrt{n-3}}$$

- What's an interval for ρ?

The classical kingdom

- The above example illustrates the reach of classical statistics.
- For a complex model, to test a certain hypothesis H_0, we could use heuristics to come of with a statistic $T(x)$ that is expected to be large when the hypothesis is false and small when the hypothesis is true.
- Heuristics turn into a rigorous testing procedure the moment we are able to calculate its power function, and in particular, its size.
- These calculations can be challenging, but asymptotic theory and/or simulation techniques (e.g., the bootstrap) help out.

Going ahead

- The classical kingdom is really vast as the basic principles are easy to implement and have been widely implemented in many different contexts.
- We can't enumerate all of these, but will see some selected examples.
- In most of these examples, there would be a "non-parametric" flavor, i.e., the models we will consider are collections of pdfs/pmfs that cannot be indexed by a finite dimensional parameter.

Least squares in linear models

- Consider data (Y_i, z_i), $i = 1, \cdots, n$ modeled as

$$Y_i = z_i^T \beta + \sigma \epsilon_i, \epsilon_i \overset{\text{IID}}{\sim} f$$

  with $f \in \mathcal{F} = \{\text{all densities on } \mathbb{R} \text{ with mean 0 and var 1}\}$
- Under mild regularity conditions,
  - $\hat{\beta}_{\text{LS}} \sim AN(\beta, \sigma^2 (Z^T Z)^{-1})$, [CLT + Delta + ...]
  - $s^2_{y|z} \xrightarrow{P} \sigma^2$
  - (but they need not be independent)
- And so the confidence intervals and testing procedures that we developed for the linear Gaussian model under ML considerations can be used for this more general model with same (asymptotic) guarantees.

Rank-sum test for location shift

- Data: $X_1, \cdots, X_n, Y_1, \cdots, Y_m$
- Model: $X_i \overset{\text{IID}}{\sim} f_1$, $Y_i \overset{\text{IID}}{\sim} f_2$,
  - X_i's and Y_j's independent
  - $f_1, f_2 \in \{\text{all densities on } \mathbb{R}\}$
  - restrict to: either $f_1 = f_2$ or $P_{[(X_1, Y_1) | f_1 \times f_2]}(X_1 < Y_1) \neq 0.5$.
- $H_0 : f_1 = f_2$
- $U = \sum_{i=1}^{n} \text{rank}(x_i)$
  - $\text{rank}(x_i)$ gives the rank of x_i in the pooled sample $\{x_1, \cdots, x_n, y_1, \cdots, y_m\}$ arranged from smallest to largest.
  - Under H_0, U is expected to be around $nm/2$.
- Under H_0, $T = \frac{U - nm/2}{\sqrt{\frac{nm(n+m+1)}{2}}} \sim AN(0, 1)$
- Size-α test: Reject H_0 if $|T| > z(\alpha)$.

Testing complete spatial randomness (CSR)

- Data $X_1, \cdots, X_n$: spatial locations of n events in $S \subset \mathbb{R}^2$
- Is the pattern
  - aggregate (exhibits clustering)?
  - regular (grid-like, exhibits repulsion)?
  - CSR (placed at random)?
- The K-statistic at distance r

$$\hat{K}(r) = \frac{A}{n^2} \sum_{i \neq j} w_{ij} I(\|x_i - x_j\| < r)$$

  where $A = \text{area}(S)$, $w_{ij} = 1/\text{area}(S \cap B(x_i, r))$
- CSR $\implies$ $K(r)$ is close to πr^2. Take $\hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}} - r$

Size calculation via simulation

- The exact distribution of $\hat{L}(r)$ is difficult to assess (depends on the shape of S in an intractable way)
- But, can simulate CSR patterns of n points in S
- Gives a simulation approximation to the distribution of $\hat{L}(r)$
- Get quantiles $L_{\alpha/2,r}$ and $L_{1-\alpha/2,r}$ by simulation
- Approximately size-α test rejects H_0 : CSR at (range r) if $\hat{L}(r) \notin [L_{\alpha/2,r}, L_{1-\alpha/2,r}]$
- See here for more details and here for an application to shrub patterns.