Building Classical Procedures

Surya Tokdar

Heuristics for parametric models

- We have learned that for a given statistical model, one can build a statistical procedure by using heuristics and then turn it into a rigorous classical procedure by calibrating its frequentist properties.
- The ML is one such heuristic (which is pity!).
- Will see two other heuristics
  1. minimum contrast, and
  2. estimating equations

  commonly used in semi-parametric models.

Models to consider

- Consider $X_1, \cdots, X_n \overset{\text{IID}}{\sim} g(x_i|\theta)$, $\theta \in \Theta$
- Θ is possibly infinite dimensional
- Restrict interest to a finite dimensional quantity $\eta = h(\theta)$
- Example: $Y_i = Z_i^T \beta + \sigma\epsilon_i$
  - $\epsilon_i \overset{\text{IID}}{\sim} f$, $f \in \{\text{all densities with mean 0 and var 1}\}$
  - $\theta = (\beta, \sigma, f)$
  - Interested only in $\eta = A^T\beta$

Basic goal

- Find an "estimator" $\hat{\eta} = \hat{\eta}(x_1, \cdots, x_n)$ of η and
- a "standard error estimate" $\hat{se} = \hat{se}(x_1, \cdots, x_n)$ such that
- If $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$ and $\eta_0 = h(\theta_0)$ then
  - $\hat{\eta} \overset{P}{\to} \eta_0$ i.e., $\hat{\eta}$ is consistent
  - $\sqrt{n}(\hat{\eta} - \eta_0) \overset{d}{\to} N(0, se^2(\eta_0))$
  - $\hat{se} \overset{P}{\to} se(\eta_0)$
- $100(1-\alpha)\%$ confidence intervals for $\eta = \hat{\eta} \mp z(\alpha)\hat{se}$
- Size-α test for $H_0 : \eta = \eta_0$ rejects H_0 if $\frac{|\hat{\eta} - \eta_0|}{\hat{se}} > z(\alpha)$.

Minimum contrast estimators

- Let $\rho(X_i, \eta)$ measure "discrepancy" between η and X_i
- Average discrepancy when $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$:

$$D(\theta_0, \eta) = \mathbb{E}_{[X_1|\theta_0]}\rho(X_1, \eta),$$

- Suppose $D(\theta_0, \eta)$ is uniquely minimized at $\eta = \eta_0 = h(\theta_0)$
- Define $D_X(\eta) = \frac{1}{n}\sum_{i=1}^n \rho(X_i, \eta)$ and $\hat{\eta} = \arg\min_\eta D_X(\eta)$.
- By LLN, if $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$ then $D_X(\eta) \overset{P}{\to} D(\theta_0, \eta)$.
- Hope: minimizer $\hat{\eta}$ of $D_X(\eta) \approx$ minimizer η_0 of $D(\theta_0, \eta)$

Asymptotic normality

- Assume $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0$ and $\eta_0 = h(\theta_0)$
- Under regularity conditions on ρ and g,

$$\sqrt{n}(\hat{\eta} - \eta_0) \overset{d}{\to} N(0, \Sigma(\theta_0))$$

where

- $\Sigma(\theta) = \{B(\theta)\}^{-1}A(\theta)\{B(\theta)\}^{-1}$
- $A(\theta)$ is matrix with elements $\int \frac{\partial}{\partial\theta_i}\rho(x_1, \theta)\frac{\partial}{\partial\theta_j}\rho(x_1, \theta)g(x_i|\theta)dx_i$
- $B(\theta)$ is matrix with elements $\int \frac{\partial^2}{\partial\theta_i\partial\theta_j}\rho(x_i, \theta)g(x_i|\theta)dx_i$

Example: Least squares, linear and non-linear

- Data $X_i = (Z_i, Y_i)$, $i = 1, \cdots, n$ modeled as
  - $Y_i = m(Z_i, \beta) + \sigma\epsilon_i$,
  - $(Z_i, \epsilon_i) \overset{\text{IID}}{\sim} g(z_i, \epsilon_i)$.
  - $\theta = (\beta, \epsilon, g)$, care about β
  - g any pdf such that $(Z, \epsilon) \sim g \implies \mathbb{E}[\epsilon|Z] = 0$, $\mathbb{V}\text{ar}[\epsilon|Z] = 1$
  - m is a known function, potentially non-linear
  - Identifiable: $m(z, \beta_1) = m(z, \beta_2)\ \forall z \iff \beta_1 = \beta_2$
- Discrepancy measure:
  - $\rho(X_1, \beta) = (Y_1 - m(Z_1, \beta))^2$
  - $D(\theta_0, \beta) = \sigma^2 + \mathbb{E}\{h(Z_1, \beta_0) - h(Z_1, \beta)\}^2$
  - uniquely minimized at $\beta = \beta_0$.

Example: Least squares, linear and non-linear (contd)

- Minimum contrast estimator $\hat{\beta}$
  - minimizes $\sum_{i=1}^n [Y_i - h(Z_i, \beta)]^2$,
  - i.e., it is the least squares estimator
  - found by quasi-Newton optimization [e.g., `optim()`, `nlm()`]
- $\Sigma(\theta) = \sigma^2 \left[\mathbb{E}_{[X_1|\theta]} \left\{ \dot{m}(Z_1, \beta) \dot{m}(Z_1, \beta)^T \right\}\right]^{-1}$
  - $\dot{m}(z_1, \beta) = \frac{\partial}{\partial \beta} m(z_1, \beta)$
- Consistent estimator of $\Sigma(\theta_0)$ is $s^2_{y|z}(\tilde{Z}^T \tilde{Z})^{-1}$ where
  - $s^2_{y|z} = \frac{1}{n-p} \sum_{i=1}^n (y_i - m(z_i, \hat{\beta}))^2$
  - $\tilde{Z} = \begin{pmatrix} \dot{m}(z_1, \hat{\beta})^T \\ \vdots \\ \dot{m}(z_n, \hat{\beta})^T \end{pmatrix}$

Quantile regression

- Data $X_i = (Z_i, Y_i)$, $i = 1, \cdots, n$
- To model: τ-th conditional quantile of Y_i given Z_i is $Z_i^T \beta$
  - $Y_i = Z_i^T \beta + \epsilon_i$
  - $Z_i \overset{\text{IID}}{\sim} g_z(z_i)$,
  - $\epsilon_i | z_i \overset{\text{IND}}{\sim} g_\epsilon(\varepsilon_i | z_i)$, $\int_{-\infty}^0 g_\epsilon(\varepsilon | z_i) d\varepsilon = \tau$
  - i.e., $P(Y_i \le z_i^T \beta | Z_i = z_i) = \tau$ for all z_i.
- Discrepancy measure:
  - $\rho(X_i, \theta) = (Y_i - Z_i^T \beta)\{\tau - I(Y_i - Z_i^T \beta \le 0)\}$
  - $D_X(\theta_0, \beta)$ uniquely minimized at β_0
  - $\hat{\beta}$ found by linear programming
- $\Sigma(\theta) = \tau(1-\tau)[\mathbb{E}\ g_\epsilon(0|Z_i) Z_i Z_i^T]^{-1} \mathbb{E}[Z_i Z_i^T] [\mathbb{E}\ g_\epsilon(0|Z_i) Z_i Z_i^T]^{-1}$
- Not easy to estimate

MLE as minimum contrast estimator

- Discrepancy measure:
  - $\rho(x_1, \theta) = -\log g(x_1|\theta)$
  - negative of log-likelihood for single obs
- Assume $g(x|\theta_1) = g(x|\theta_2)\ \forall x \iff \theta_1 = \theta_2$.
- $D(\theta_0, \theta) = -\int g(x_1|\theta_0) \log g(x_1|\theta) dx_1$
- Uniquely minimized at $\theta = \theta_0$.
- The minimum contrast estimator is precisely $\hat{\theta}_{\text{MLE}}(X)$
- $\Sigma(\theta) = \{n I_1^F(\theta)\}^{-1}$.

Digression on Kullback-Leibler divergence

- For any two densities $f(x), g(x)$ the Kullback-Leibler (KL) divergence of g from f is defined as

$$K(f, g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

- $K(f, g) \ge 0$ and equals zero if and only if $f = g$
- A fundamental concept of "distance" between pdfs
- For $X_i \overset{\text{IID}}{\sim} f(x)$, "likelihood ratio" $\sum_{i=1}^n \log \frac{f(x_i)}{g(x_i)} \approx nK(f, g)$
- $D(\theta_0, \theta) = K(g(\cdot|\theta_0), g(\cdot|\theta)) + \text{const}$

First order condition

- Suppose for every x_1, $\theta \mapsto \rho(x_1, \eta)$ is differentiable in η.
- Then η_0 can be expected to uniquely solve $\frac{\partial}{\partial \eta} D(\theta_0, \eta) = 0$.
- Similarly we can expect $\hat{\eta}$ to solve $\frac{\partial}{\partial \eta} D_X(\eta) = 0$.
- This suggests another way of constructing estimates – known as *estimating equations estimate*.

Estimating equations

- $\eta \mapsto \psi(X_1, \eta)$: a $\mathbb{R}^p$ valued "score" function
- When $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$, with $\eta_0 = h(\theta_0)$
  - average score $V(\theta_0, \eta) = \mathbb{E}_{[X_1|\theta_0]}\psi(X_1, \eta)$
  - has unique soln at $\eta = \eta_0$.
- $V_X(\eta) = \frac{1}{n}\sum_{i=1}^n \psi(X_i, \eta) \xrightarrow{p} V(\theta_0, \eta)$
- $\hat{\eta}$ solves $V_X(\eta) = 0$
- Under regularity conditions $\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} N(0, \Sigma(\theta_0))$
  - $\Sigma(\theta) = \{B(\theta)\}^{-1}A(\theta)\{B(\theta)^T\}^{-1}$
  - $A(\theta) = \mathbb{E}_{[X_1|\theta]}\psi(X_1, \theta)\psi(X_1, \theta)^T$
  - $B(\theta) = \mathbb{E}_{[X_1|\theta]}\frac{\partial}{\partial\theta}\psi(X_1, \theta)$

Method of moments

- $X_1, \cdots, X_n \overset{\text{IID}}{\sim} g(x_i|\theta)$,
- X_i's are univariate, θ is a p-vector.
- Let $\mu_j(\theta) = \mathbb{E}_{[X_1|\theta]}X_1^j$, $j = 1, \cdots, p$.
- Suppose $\theta \mapsto (\mu_1(\theta), \mu_2(\theta), \cdots, \mu_d(\theta))$ is one-to-one.
- Then one can estimate θ by solving the equations:

$$\mu_j(\theta) = \frac{1}{n}\sum_{i=1}^n X_i^j, \quad j = 1, 2, \cdots, d$$

- This $\hat{\theta}$ is called the method of moments (MoM) estimator
- $\Sigma(\theta)$ may be difficult to estimate

Approximating standard error by bootstrapping

- $X_i \overset{\text{IID}}{\sim} g(x_i|\theta)$, $\eta = h(\theta)$
- Have estimator $\hat{\eta}$ such that $\sqrt{n}(\hat{\eta} - \eta_0) \xrightarrow{d} N(0, se(\eta_0))$ whenever $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$ with $\eta_0 = g(\theta_0)$.
- Finding a consistent estimator $\hat{se}$ of $se(\eta)$ might be difficult
- Bootstrapping provides an answer by resampling from the observed data $x_1, \cdots, x_n$.

Method and justification

- Randomly sample $x_1^*, \cdots, x_n^*$ from $x_1, \cdots, x_n$ with replacement
- Use your estimating procedure on $x_1^*, \cdots, x_n^*$ to get $\hat{\eta}^*$
- Repeat a B (large number) many times and record $\hat{\eta}_1^*, \cdots, \hat{\eta}_B^*$
- Use $\hat{se} = \frac{1}{B-1}\sum_{i=1}^B(\hat{\eta}_i^* - \bar{\eta}^*)^2$ where $\bar{\eta}^* = \frac{1}{B}\sum_{i=1}^B \hat{\eta}_i^*$.
- Under fairly mild conditions $\hat{se} \xrightarrow{p} se(\eta_0)$ whenever $X_i \overset{\text{IID}}{\sim} g(x_i|\theta_0)$ with $\eta_0 = h(\theta_0)$.

Second order correction

- Even when an $\hat{se}$ can be found directly with $\hat{se} \xrightarrow{p} se(\eta_0)$, bootstrap may be used to better approximate the distribution of $T = \frac{\hat{\eta} - \eta_0}{\hat{se}/\sqrt{n}}$ than the asymptotic limit
- This is due to a certain cancellation of second order terms in theoretical calculations of the approximation
- We now obtain $T^* = \frac{\hat{\eta}^* - \hat{\eta}}{\hat{se}^*/\sqrt{n}}$ from our bootstrap sample $x_1^*, \cdots, x_n^*$
- Repeat this B many times and record $T_1^*, \cdots, T_B^*$
- An approximately size-α test rejects $H_0 : \eta = \eta_0$ if $|T| > z^*_{1-\alpha,|T^*|}$: the $(1-\alpha)$-th quantile of $|T_1^*|, \cdots, |T_B^*|$.