Bayesian and Frequentist Issues in Modern Inference

Bradley Efron

Stanford University
Small Data

- **Classical statistics**  Direct tests and estimates of individual parameters within well-defined models (MLE, Neyman–Pearson)

- **Not much:**
  - data-based model selection
  - Bayesian combination of related problems

- **Today**  Methodology (not Philosophy)
Bayesian Inference

- Parameter: $\mu \in \Omega$
- Observed data: $x$
- Prior: $g(\mu)$
- Probability distributions: $\{f_\mu(x), \mu \in \Omega\}$
- Parameter of interest: $\theta = t(\mu)$

$$E\{\theta|x\} = \frac{\int_{\Omega} t(\mu)f_\mu(x)g(\mu) \, d\mu}{\int_{\Omega} f_\mu(x)g(\mu) \, d\mu}$$
Jeffreysonian Bayes Inference

“Uninformative Priors”

- What if we don’t know prior $g$?
- Jeffreys: $g(\mu) = |I(\mu)|^{1/2}$ where $I(\mu) = \text{cov}\left\{ \nabla_\mu \log f_\mu(x) \right\}$ (the Fisher information matrix)
- Can still use Bayes theorem but how accurate are the estimates?
- Frequentist variability of $E \{ t(\mu) | x \}$
General Accuracy Formula

- $\mu$ and $x \in \mathcal{R}^p$
- $V_\mu = \text{cov}_\mu(x)$
- $\alpha_x(\mu) = \nabla_x \log f_\mu(x) = \left(\ldots, \frac{\partial \log f_\mu(x)}{\partial x_i}, \ldots\right)^T$

**Lemma**

$E = E \{t(\mu)|x\}$ has gradient $\nabla_x E = \text{cov} \{t(\mu), \alpha_x(\mu)|x\}$.

**Theorem**

The delta-method standard deviation of $E$ is

$$\text{sd}(E) = \left[\text{cov} \{t(\mu), \alpha_x(\mu)|x\}^T V_x \text{cov} \{t(\mu), \alpha_x(\mu)|x\}\right]^{1/2}.$$
Implementation

- Posterior sample from $\mu|x$
  $$\{\mu_1, \mu_2, \ldots, \mu_B\} \quad \text{(MCMC?)}$$
- Each $\mu_i$ gives $t_i = t(\mu_i)$ and $\alpha_i = \alpha_x(\mu_i)$
- $\hat{E} = \sum t_i / B \doteq E\{t(\mu)|x\}$
- $\hat{\text{cov}} = \sum_{i=1}^B (\alpha_i - \hat{\alpha})(t_i - \bar{t}) / B$
- $\hat{\text{sd}} = \left[\hat{\text{cov}}^T V_x \hat{\text{cov}}\right]^{1/2}$
- No additional sampling for $\hat{\text{cov}}$
Diabetes Data

Efron et al. (2004), “LARS”

- $n = 442$ subjects
- $p = 10$ predictors: age, sex, bmi, glu,…
- Response: $y =$ disease progression at one year
- Model: $\mathbf{y} = \mathbf{X} \beta + \mathbf{e}$ [$\mathbf{e} \sim \mathcal{N}_n(0, I)$]
Bayesian Lasso
Park and Casella (2008)

- Model: \( y \sim \mathcal{N}_n(X\beta, I) \)
- Prior: \( g(\beta) = e^{-\gamma L_1(\beta)} \) \( [\gamma = 0.37] \)
- Then posterior mode at Lasso \( \hat{\beta}_\gamma \)
- Subject 125: \( \theta_{125} = x_{125}^T \beta \)
- How accurate are Bayes posterior inferences for \( \theta_{125} \)?
Bayesian Analysis

- MCMC: posterior sample \( \{ \beta_i \text{ for } i = 1, 2, \ldots, 10,000 \} \)
- Gives \( \{ \theta_{125,i} = x_{125}^T \beta_i, \ i = 1, 2, \ldots, 10,000 \} \)

\[ \theta_{125,i} \sim 0.248 \pm 0.072 \]

- General accuracy formula frequentist sd 0.071 for \( E = 0.248 \)

\[ [\alpha_x(\mu) = X^T X \beta] \]
Posterior CDF for Subject 125

- \( \text{cdf}_y(c) = \Pr\{\theta_{125} \leq c | y\} \)
- \( s_i = \begin{cases} 
1 & \text{as } t_i \leq c \\
0 & \text{as } t_i > c 
\end{cases} \)
- \( \hat{\text{cdf}}_y(c) = \sum_{1}^{B} s_i / B \)
- For \( c = 0.3 \):
  \[
  \hat{\text{cdf}}_y(c) = 0.762 \pm 0.304
  \]
  Bayes frequentist estimate sd
Posterior cdf for mu125, Diabetes data, 10000 MCMC draws, Prior $\exp\{-0.37L_1(\beta)\}$; verts are $\pm$ One Frequentist Standard Dev

Upper 95% credible limit is 0.342 $\pm$ 0.069
c value

Prob{mu125 $< c$ | data}

Bradley Efron (Stanford University)
Exponential Families

- \( f_{\alpha}(\hat{\beta}) = e^{\alpha^T \hat{\beta} - \psi(\alpha)} f_0(\hat{\beta}) \), with \( \alpha, \hat{\beta} \) in \( \mathcal{R}^p \)

- Natural parameter \( \alpha \), sufficient statistic \( \hat{\beta} \), expectation \( \beta = E_\alpha \{ \hat{\beta} \} \)

- Poisson: \( f_\mu(x) = e^{-\mu} \mu^x / x! : x = \hat{\beta}, \mu = \beta, \alpha = \log(\mu) \)

- General accuracy formula: For \( E = E \{ t(\beta) | \hat{\beta} \} \),

\[
\hat{sd}(E) = \left\{ \text{cov}(t, \alpha | \hat{\beta})^T V_{\hat{\alpha}} \text{cov}(t, \alpha | \hat{\beta}) \right\}^{1/2}
\]

with \( V_{\hat{\alpha}} = \text{cov}_{\alpha=\hat{\alpha}}(\hat{\beta}) \).
Better Frequentist Inferences

- For $E = E\{t(\beta)\mid \hat{\beta}\}$ in exfam $f_\alpha (\hat{\beta}) = e^{\alpha^T \hat{\beta} - \psi(\alpha)} f_0 (\hat{\beta})$
- **Parametric bootstrap** $f_\hat{\alpha} (\cdot) \rightarrow [\hat{\beta}_1^*, \hat{\beta}_2^*, \ldots, \hat{\beta}_j^*, \ldots, \hat{\beta}_j^*]$
  $\longrightarrow \left[ \cdots E_j^* = E\{t(\beta)\mid \hat{\beta}_j^*\} \cdots \right] \longrightarrow$ bootstrap conf int for $E$
- **Trouble** Need new MCMC sample for each $\hat{\beta}_j^*$
- **Shortcut** Reweight original MCMC sample (importance sampling)
Digression: *Posterior Exponential Family*

- \( f_\alpha(\hat{\beta}) = e^{\alpha^T \hat{\beta} - \psi(\alpha)} f_0(\hat{\beta}) \): natural param \( \alpha \), suff stat \( \hat{\beta} \), “carrier” \( f_0(\hat{\beta}) \)
- Posterior exponential family

\[
\begin{align*}
  g(\alpha | \text{suff stat } b) &= e^{(b - \hat{\beta})^T \alpha - \phi(b)} g(\alpha | \hat{\beta}) \\
  \text{natural param } b, \text{ suff stat } \alpha, \text{ carrier } g(\alpha | \hat{\beta})
\end{align*}
\]

- Importance sampling  
  Reweight the original MCMC realizations:

\[
\begin{align*}
  \hat{E}\{t_i | b\} &= \frac{\sum_{i=1}^{B} t_i W_i(b)}{\sum_{i=1}^{B} W_i(b)} \\
  \left[ W_i(b) = e^{(b - \hat{\beta})^T \alpha_i} \right]
\end{align*}
\]
Bootstrap Intervals Without Bootstrapping

DiCiccio and Efron (1992)

“abc” Investigate $\hat{E}\{t|b\}$ for $b$ near $\hat{\beta}$

Requires $p + 2$ numerical 2nd derivatives of $\hat{E}$ function

Next: Applied to posterior cdf for mu125
Heavy curve is posterior cdf for mu125, diabetes data. Vertical bars frequentist central 68% abc confidence intervals. (Light lines show +− one frequentist standard error)
Estimation After Model Selection

- **Usually:**
  1. look at data
  2. choose model (linear, quad, cubic . . . ?)
  3. fit estimates using chosen model
  4. analyze as if pre-chosen

- **Today**  Include model selection process in the analysis

- **Question**  Effects on standard errors, confidence intervals, etc.?
**Cholesterol Data**

- $n = 164$ men took Cholestyramine for $\sim 7$ years
- $x =$ compliance measure (adjusted: $x \sim \mathcal{N}(0,1)$)
- $y =$ cholesterol decrease
- Wish to estimate regression values

$$\mu_j = E\{y|x = x_j\} \quad \text{for } j = 1, 2, \ldots, 164$$

$$\mu = (\mu_1, \mu_2, \ldots, \mu_{164})^T$$
Cholesterol data, n=164 subjects: cholesterol decrease plotted versus adjusted compliance; Green curve is OLS cubic regression; Red points indicate 5 featured subjects.
\( Cp \) Selection Criterion

- **Regression model**
  \[
  \mathbf{y}_{n \times 1} = \mathbf{X}_{n \times m} \mathbf{\beta}_{m \times 1} + \mathbf{e}_{n \times 1} \quad [e_i \sim (0, \sigma^2)]
  \]

- **\( Cp \) criterion**
  \[
  \| \mathbf{y} - \mathbf{X}\hat{\beta} \|^2 + 2m\sigma^2
  \]

  \( \hat{\beta} = \text{OLS estimate} \), \( m = \text{“degrees of freedom”} \)

- **Model selection**
  From possible models \( \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots \)
  choose the one minimizing \( Cp \).

- Then use OLS estimate from chosen model.
**C_p for Cholesterol Data**

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>( C_p - 80000 )</th>
<th>(Boot %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_1 ) (linear)</td>
<td>2</td>
<td>1132</td>
<td>(19%)</td>
</tr>
<tr>
<td>( M_2 ) (quad)</td>
<td>3</td>
<td>1412</td>
<td>(12%)</td>
</tr>
<tr>
<td>( M_3 ) (cubic)</td>
<td>4</td>
<td>667</td>
<td>(34%)</td>
</tr>
<tr>
<td>( M_4 ) (quartic)</td>
<td>5</td>
<td>1591</td>
<td>(8%)</td>
</tr>
<tr>
<td>( M_5 ) (quintic)</td>
<td>6</td>
<td>1811</td>
<td>(21%)</td>
</tr>
<tr>
<td>( M_6 ) (septic)</td>
<td>7</td>
<td>2758</td>
<td>(6%)</td>
</tr>
</tbody>
</table>

(\( \sigma = 22 \) from “full model” \( M_6 \))
Nonparametric Bootstrap Analysis

- data = {(x_i, y_i), \ i = 1, 2, \ldots, n = 164} gave original estimate

\[ \hat{\mu} = X_3\hat{\beta}_3 \]

- Bootstrap data set: data* = \{(x_j, y_j)^*, \ j = 1, 2, \ldots, n\} where
  (x_j, y_j)^* drawn randomly and with replacement from data:
  
  \[
  \text{data}^* \xrightarrow{C_p} m^* \xrightarrow{\text{OLS}} \hat{\beta}_{m^*}^* \xrightarrow{} \hat{\mu}^* = X_{m^*}\hat{\beta}_{m^*}^*
  \]

- I did this all B = 4000 times.
B=4000 nonparametric bootstrap replications for the model−selected regression estimate of Subject 1; boot (m, stdev)=(-2.63, 8.02); 76% of the replications less than original estimate 2.71

Red triangles are 2.5th and 97.5th boot percentiles bootstrap estimates for subject 1

Frequency

0 50 100 150 200 250

bootstrap estimates for subject 1
Red triangles are 2.5th and 97.5th boot percentiles
Smooth Estimation Model: 'y' is observed data; 
Ellipses indicate bootstrap distribution for 'y*';
Red curves level surfaces of equal estimation for \( \hat{\theta} = t(y) \)
Estimation with model selection: now the curves of equal estimation are discontinuous across the Model region boundaries.
Boxplot of Cp boot estimates for Subject 1; B=4000 bootreps; Red bars indicate selection proportions for Models 1–6.

Only 1/3 of the bootstrap replications chose Model 3.
**Bootstrap Smoothing**

- **Idea** Replace original estimator $t(y)$ with bootstrap average

\[
s(y) = \frac{1}{B} \sum_{i=1}^{B} t(y^*_i)
\]

- Model averaging
- Same as *bagging* ("bootstrap aggregation," Breiman)
- Removes discontinuities, reduces variance
- Approximate confidence interval: $s(y) \pm 1.96 \cdot \tilde{sd}$
Accuracy Theorem

- **Notation** \( s_0 = s(y), \quad t_i^* = t(y_i^*), \ i = 1, 2, \ldots B \)
- \( Y_{ij}^* = \# \) of times \( j \)th data point appears in \( i \)th boot sample
- \( \text{cov}_j = \sum_{i=1}^{B} Y_{ij}^* \cdot (t_i^* - s_0)/B \) (covariance \( Y_{ij}^* \) with \( t_i^* \))

**Theorem**

The delta method standard deviation estimate for \( s_0 \) is

\[
\tilde{sd} = \left[ \sum_{j=1}^{n} \text{cov}_j^2 \right]^{1/2},
\]

always \( \leq \left[ \sum_{i=1}^{B} (t_i^* - s_0)^2 / B \right]^{1/2} \), the boot stdev for \( t(y) \).
Projection Interpretation
Standard Deviation of smoothed estimate relative to original (Red) for five subjects; green line is stdev Naive Cubic Model.

Subject number

Relative stdev

- 7.9 - 3.9 - 4.1 - 4.7 - 6.8 -
Model Probability Estimates

- 34% of the 4000 bootreps chose the cubic model
- Poor man’s Bayes posterior prob for “cubic”
- How accurate is that 34%?
- Apply accuracy theorem to indicator function for choosing “cubic”
<table>
<thead>
<tr>
<th>Model</th>
<th>Boot %</th>
<th>± Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$ (linear)</td>
<td>19%</td>
<td>±24</td>
</tr>
<tr>
<td>$M_2$ (quad)</td>
<td>12%</td>
<td>±18</td>
</tr>
<tr>
<td>$M_3$ (cubic)</td>
<td>34%</td>
<td>±24</td>
</tr>
<tr>
<td>$M_4$ (quartic)</td>
<td>8%</td>
<td>±14</td>
</tr>
<tr>
<td>$M_5$ (quintic)</td>
<td>21%</td>
<td>±27</td>
</tr>
<tr>
<td>$M_6$ (sextic)</td>
<td>6%</td>
<td>±6</td>
</tr>
</tbody>
</table>