Section 1.
Bayesian Inference

Vincent Dorie

New York University

Overview

What is Bayesian Statistics?

Method of **statistical inference** that incorporates **prior knowledge** to make probabilistic statements about quantities of interest

- Write a probabilistic **model**/data generating process
- Add prior/penalize **parameters** in model
- Fit model to **data** (experiment)
- Use fitted model to produce **estimate**; quantify **uncertainty** in estimate
- Model checking, diagnostics

Contrast with: machine learning, classical statistics, econometrics

Why Bayes?

- Computation
- Prior information
- Regularization
- Coherence
- Explicit uncertainty, theory

Why Not Bayes?

- Computation
- Prior subjectivity
- DGP subjectivity
- Ideology over accuracy
- Parsimony over accuracy

Models

Probabilistic Models

- Joint probability distribution over data
- Sequence of distributional and independence assumptions

$$\begin{aligned} z \mid x, u &\sim \text{Bernoulli}(\operatorname{logit}^{-1}(x\beta^z + \zeta^z u)), \\ y \mid x, u, z &\sim \text{Normal}(x\beta^y + \zeta^y u + \tau z, \sigma_y^2), \\ u &\sim \text{Bernoulli}(\theta). \end{aligned}$$

- Plausible story for data generation and measurement noise
- Model (roughly) determines subsequent steps

Running Example: Trump

- Sample random New Yorkers and poll support for Trump
- Collect other data points, income, age, height, sex, ...

Build a model for height:

$$\text{height} \sim \mathsf{Normal}\left(\text{baseline} + \beta_1 \cdot \log(\text{income}) + \beta_2 \cdot \text{vote\_trump} + \cdots, \sigma^2\right)$$

Notation

- Random variables: y, u
- Observed data: $y = \{y_1, \ldots, y_N\}$
- Covariates $x : N \times P$ matrix, x_i column vector

y is **modeled data**, x is **unmodeled data**

- Inference is *conditional* on x
- Random variables express what *could have happened* (repetition)

y is “overloaded”, meaning context-specific

Notation Continued

- α, β **parameters**; θ vector
- $p(\cdot;\theta)$ probability density *indexed* by θ
- $p(\cdot \mid \theta)$ conditional probability density

Overload p: $p(y \mid \alpha, \beta, \sigma)$, $p(\alpha, \beta \mid \sigma)$, $p(\sigma)$

Common parameters:

- μ: mean, expected value
- σ: scale, standard deviation
- α, β: regression intercept, slope

Notation Continued

Predictive quantities:

- $\tilde{x}$: point for which we would like to make a prediction
- $\tilde{y}$: prediction at $\tilde{x}$

Simulated quantities:

- $\theta^{(m)}$: draw of θ from a distribution, $p(\theta \mid y)$

Estimated quantities:

- $\hat{\theta}$: point estimate of θ (MLE, MAP, MM)

Common assumptions:

- y independent after controlling for covariates: $y_i \perp\!\!\!\perp y_j \mid x$ for $i \neq j$

Running Example: Trump

- $y_1, \ldots, y_N$: height of individuals in sample, 1 through N
- $x_1, \ldots, x_N$: column vectors of predictors for individuals
- x: $N \times P$ matrix of predictors, 1st column 1s
- g: $N \times Q$ matrix, Q num of zip-codes, rows of g "select" zip-code for individuals
- β: individual coefficients
- α: zip-code offsets

$$y \sim \text{Normal}(x\beta + g\alpha, \sigma^2)$$

Comparison with Comp-sci & Econ

Objective function:

$$f(\alpha, \beta) = \|y - x\beta - g\alpha\|^2$$

- Find $\operatorname{argmin}_{\alpha,\beta} f$
- Possibly regularize adding penalty term to f

Estimating equation:

$$y = x\beta + g\alpha + \epsilon$$

- Estimate α, β by BLUP
- Correct correlations in error by using cluster robust standard errors

Where Do Models Come From?

- Sometimes model comes first, based on substantive considerations
  - toxicology, economics, ecology, . . .
- Sometimes model chosen based on data collection
  - traditional statistics of surveys and experiments
- Other times the data comes first
  - observational studies, meta-analysis, . . .
- Usually its a mix

Bayesian Modeling

Bayesian Overview

"Classically":

- Write down model/data generating process
- Find parameter values that maximize likelihood of observations
- Use interval procedure to quantify uncertainty

Bayesian:

- Model/DGP
- Incorporate priors over parameters
- Compute posterior distribution of parameters (or QOIs)
- Summarize posterior distribution by mean, quantiles

Likelihood

- Probability density/mass of data *as viewed as function of its parameters*
- $p(y; \theta) \equiv L(\theta)$
- Not a distribution
- Central role in classical statistics
- Maximize for estimate, curvature for approx uncertainty

Bayesian: prior + likelihood → posterior

Bayesian Differences

- Estimands are point and intervals
- Make probabilistic statements about procedures used

vs

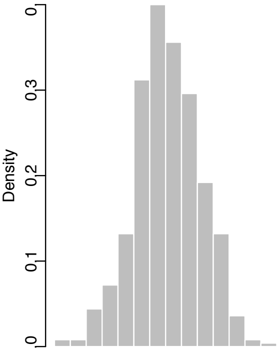
- Estimand is a distribution
- Make probabilistic statements about parameters

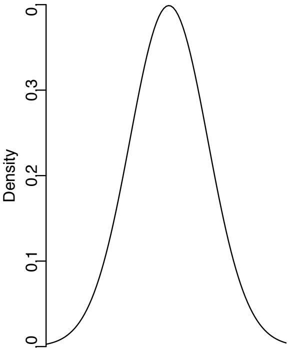
Distributions As Estimands

- Use Bayes' rule to compute posterior:

$$\begin{aligned} p(\theta \mid y) &= p(y, \theta) / p(y), \\ &= \frac{p(y \mid \theta) p(\theta)}{p(y)}. \end{aligned}$$

- Occasionally, can compute $p(\theta \mid y)$ directly
- Most often, use **samples** from $p(\theta \mid y)$ as summary





$$\Pr[\theta \le t \mid y] \approx \frac{1}{M} \sum_{m=1}^{M} \mathrm{I}[\theta^{(m)} \le t],$$

$$\tilde{\theta} \sim p(\theta \mid y).$$

Linear Model Example

$$y \mid \beta \sim \text{Normal}(x\beta, \sigma^2),$$
$$\beta \sim \text{Normal}(0, 5^2).$$

σ^2 fixed

$$\begin{aligned}
p(\beta \mid y) &= p(y \mid \beta)p(\beta)/p(y), \\
&= \prod_{i=1}^{N} \left[\sqrt{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(y_i - x_i^\top \beta)^2} \right] \times \\
&\quad \prod_{j=1}^{P} \left[\sqrt{2\pi 5^2} e^{-\frac{1}{2\cdot 5^2}\beta_j^2} \right] \Big/ p(y).
\end{aligned}$$

Linear Model Example Cont

$$p(\beta \mid y) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^\top\beta)^2\right\} \times$$
$$(2\pi 5^2)^{-P/2} \exp\left\{-\frac{1}{2\cdot 5^2}\sum_{j=1}^{P}\beta_j^2\right\} \Big/ p(y)$$

Bundle constants into C

$$p(\beta \mid y) = C \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - x_i^\top\beta)^2 + \frac{1}{5^2}\sum_{j=1}^{P}\beta_j^2\right]\right\}.$$

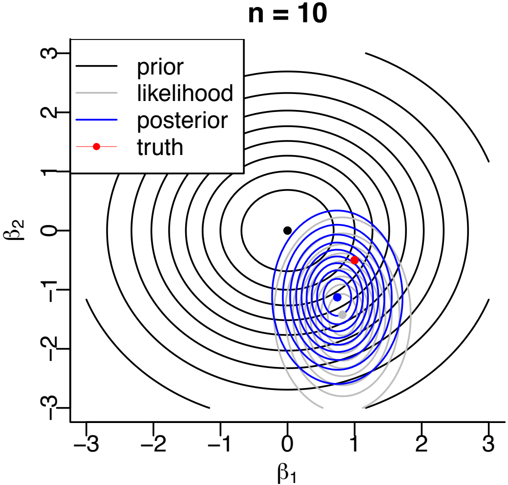
Write $p(\beta \mid y) \propto$ and drop C

Linear Model Example Cont

$$
\begin{aligned}
p(\beta \mid y) &\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}(y - x\beta)^\top (y - x\beta) + \frac{1}{5^2}\beta^\top\beta\right]\right\}, \\
&\propto \exp\left\{-\frac{1}{2}\left[\beta^\top x^\top x\beta/\sigma^2 + \beta^\top\beta/5^2 - 2\beta^\top x^\top y/\sigma^2\right]\right\}, \\
&\propto \exp\left\{-\frac{1}{2}(\beta - \Sigma_{\beta|y} x^\top y/\sigma^2)^\top \Sigma_{\beta|y}^{-1}(\beta - \Sigma_{\beta|y} x^\top y/\sigma^2)\right\}.
\end{aligned}
$$

Where $\Sigma_{\beta|y} = (x^\top x/\sigma^2 + \mathrm{I}_p 1/5^2)^{-1}$.

$$
\beta \mid y \sim \text{Normal}\left(\Sigma_{\beta|y} x^\top y/\sigma^2, \Sigma_{\beta|y}\right).
$$



n = 10
prior
likelihood
posterior
truth
β₂
β₁
−3
−2
−1
0
1
2
3

Logistic Example

Instead predict voting

$$
\begin{aligned}
y \mid \beta &\sim \text{Bernoulli}\left(\text{logit}^{-1}(x\beta)\right), \\
\beta &\sim \text{Normal}(0, 5^2).
\end{aligned}
$$

$$
\begin{aligned}
p(\beta \mid y) &\propto p(y \mid \beta) p(\beta), \\
&\propto \prod_{i=1}^{N} \left[y_i \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}} + (1 - y_i) \frac{1}{1 + e^{x_i^\top \beta}} \right] \times \\
&\quad \exp\left\{ -\frac{1}{2 \cdot 5^2} \beta^\top \beta \right\}.
\end{aligned}
$$

Prior Choice

- Prior information
- Conjugate
- Uninformative
- Jeffreys'
- Weakly informative

Posterior Predictive Distribution

- Predict new data $\tilde{y}$ based on observed data y
- Marginalize out parameters from posterior

$$\begin{aligned} p(\tilde{y} \mid y) &= \int p(\tilde{y}, \theta \mid y)\, \mathrm{d}\theta, \\ &= \int p(\tilde{y} \mid \theta) p(\theta \mid y)\, \mathrm{d}\theta. \end{aligned}$$

- Averages predictions $p(\tilde{y} \mid \theta)$ weighting by posterior $p(\theta \mid y)$
- Allows continuous, discrete, or mixed parameters
  - integral notation shorthand for sums and/or integrals

Diagnostics

Model Checking

- Do the inferences make sense?
  - are parameter values consistent with model's prior?
  - does simulating from parameter values produce reasonable fake data?
  - are marginal predictions consistent with the data?
- Do predictions and event probabilities for new data make sense?
- ***Not***: Is the model true?
- ***Not***: What is $\Pr[\text{model is true}]$?
- ***Not***: Can we "reject" the model?

Model Improvement

- Expanding the model
  - hierarchical and multilevel structure . . .
  - more flexible distributions (overdispersion, covariance)
  - more structure (geospatial, time series)
  - more modeling of measurement methods and errors
  - . . .
- Including more data
  - breadth (more predictors or kinds of observations)
  - depth (more observations)