A Short Course on Computationally Intensive Probability and Statistics

Murali Haran

Department of Statistics
Penn State University

University of Split
September 2017

Organization of the Course

- Probability
  - Computational challenge: simulation from probability distributions, approximating integrals
  - Focus in this course:
    1. Monte Carlo
    2. Markov chain Monte Carlo
- Statistics (Statistical Inference)
  - Computational challenge: optimization, approximating integrals, approximating sampling distributions, standard errors
  - Focus in this course:
    1. basic optimization via Newton-Raphson
    2. bootstrap for standard errors
    3. E-M algorithm (time permitting)

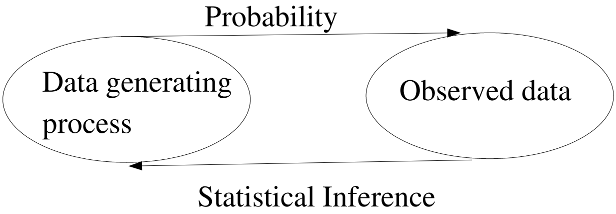
References

- Computational Statistics by Givens and Hoeting
- Numerical Analysis for Statistics by Kenneth Lange
- Monte Carlo Statistical Methods by Robert and Casella

Probability and Statistical Inference

Can think of them as “forward” and “inverse” problem

- Probability: Given a model for data, what can we say about the probability of various outcomes (data) from that model?
- Statistics: Given data (observations), what can we infer about the process that generated the data?



Probability

Probability: Given a data generating process, what are the properties of the outcomes (observations)? Simple examples:

- Example 1: If you know the probability p of “Heads” on a coin toss is 0.7, and X is the number of heads on 100 independent coin tosses, what is the distribution of X?
- Example 2: (X, Y) is bivariate normal with mean 0, $(\sigma_1^2, \sigma_2^2) = (1, 4)$, covariance=1. What is P$(X > Y^2)$?

These are not very complicated calculations. But they are easy to approximate via simulation/Monte Carlo.

Why is Computing Important in Probability?

- For a complicated statistical model, a probability distribution f, understanding the properties of the distribution can be mathematically challenging
- However, simulating from f can often be much easier. Can use Monte Carlo approximations
- Goal: Approximate expectation w.r.t. f
- Principle: If we can simulate exactly from f or construct (Harris-ergodic) Markov chain with stationary distribution f, we can approximate this expectation
- Monte Carlo avoids analytically intractable calculations, numerical integration. Convergence rate of estimator is not a function of dimension of f (less affected by dimension of the distribution)

Computing Probabilities via Monte Carlo

Simple idea:

1. Simulate ("pseudo") random variables from the probability model $f(x)$ using computer code, $X_1, X_2, \cdots \sim f(x)$
2. Approximate expected value, $\mu = E_f(g(X))$, for real-valued function g, compute corresponding sample mean,

$$\hat{\mu}_m = \frac{\sum_{i=1}^m g(X_i)}{m}$$

Basic probability theory:

- **Law of large numbers**: sample mean converges to its expectation, $\hat{\mu}_m \to \mu$ as $m \to \infty$, as long as $E_f(g(X)) < \infty$
- **Central Limit Theorem**: to obtain approximate confidence interval for μ based on $\hat{\mu_m}$

Writing Computer Code for Monte Carlo

- Example 1: need command for Binomials, `rbinom`
- Example 2: simulate bivariate normals, choices:
  - Use `R` package `MASS` and `mvrnorm`
  - (To understand ideas better) Write our own code: (i) enter covariance matrix Σ, (ii) generate iid Norm(0,1) random variates, $\mathbf{x}$ (iii) compute choleski factor, C s.t. $CC^T = \Sigma$, and (iv) obtain via multiplication $\mathbf{z} = C\mathbf{x} \sim N(0, \Sigma)$
- In order to use Monte Carlo, repeat above (procedure for generating Binomials or multivariate Normal) many times, then calculate sample means

Probability: More Examples

- Example 3: Draw values at 30 locations from Gaussian process with mean 0 and stationary covariance, cov(X_i, X_j)=exp($-d_{ij}/\phi$), where d_{ij} is Euclidean distance. Call 30-dim r.v., **X**. What is $P(X_3 > X_4^2 + X_5)$?
- Example 4: Stochastic model for the spread of an infectious disease: Q1. study how disease will spread over space and time for different initial conditions/parameter values. Q2. How likely is an epidemic?
- Example 5: Ising model: dependence model for binary data on a lattice/undirected graph. Study behavior of system by simulation via Markov chain Monte Carlo (MCMC)

Statistical Inference

Inference: Given the outcomes (our observations), what can we say about the process that generated the data?

- Example 1 (simple): If you know X (# of heads on 100 independent coin tosses) what can you say about p?
  - Point estimate for p, interval estimate for p
- Example 2: Given space-time data on the spread of an infectious disease, what are the (unknown) parameters of the stochastic model?
  - Point and interval estimates for parameters
  - Does the model “fit” the observations? Idea: simulate from fitted model, compare to data. (Back to Monte Carlo)
- Example 3: Observe pairs $(X_1, Y_1), \ldots (X_n, Y_n)$. Relationship between $X's$ and $Y's$, allowing for flexibility (non-linearity + errors)? Predict Y based on X?

Estimation and Prediction

Estimation: What we say about our fitted model should reflect our uncertainty (variability) in estimates of the parameter.

Prediction: We will often use fitted model to make predictions. Since there is typically randomness in how the data generating process produces the data, our inferences and predictions should account for this variability

- Sampling distributions, confidence intervals
- standard errors, bias correction
- prediction intervals: variability in predictions from fitted model combined with variability from parameter estimates often easy to do via simulation-based approach

Why is Computing Important in Statistics?

- Parameter estimation in a model often involves minimizing/maximizing an objective function
    - Maximum likelihood/penalized maximum likelihood
    - Nonparametric regression: minimizing an objective function to fit a curve to a data set
- Computing standard errors for parameter estimates can be difficult – asymptotic theory may be hard to derive or inapplicable
    - Re-sampling methods (bootstrap) for approximating sampling distributions, approximate confidence intervals
- Bayesian inference (alternative to maximum likelihood)
    - Approximating posterior distribution (back to probability) by Markov chain Monte Carlo methods
    - Maximizing posterior distribution (back to optimization)

Advanced Topics/Challenges

- Likelihood functions are often expensive to evaluate, may involve high-dimensional integration
  - Surrogate function methods: minorization-maximization (M-M) algorithms, expectation-maximization (E-M) algorithms (special case of MM), composite likelihood,...
  - Monte Carlo maximum likelihood, Monte Carlo E-M
- Computational complexity and storage requirements for large data, complicated models
  - Divide-and-conquer algorithms for parallelizing computing/storage
  - Monte Carlo/MCMC with approximate likelihoods
- New constrained optimization algorithms
- + Many research problems in modern probability/statistics