Stat 624: Computational Topics in Statistics - Project Guidelines

This project serves as an opportunity for students to engage in advanced computational topics pertinent to their graduate program and beyond. You are encouraged to select a topic from the list provided or propose your own, contingent on instructor approval. The project has two key deliverables: an in-class presentation and a written report. The written report should be compiled using LaTeX and encompass the following sections:

1. **Introduction**: Introduce the research question and articulate its significance. Outline the methodology and data sets involved. Also, identify additional fields where the methodology could be applicable.

2. **Methodology**: Elaborate on the methodology. What statistical techniques (e.g., optimization, regression, hypothesis testing) will be utilized to address the research questions?

3. **Simulation Study**: Validate the efficacy of your chosen methodology in meeting the research objectives. Provide evidence.

4. **Results**: Discuss the data analysis and findings. Make your results easily comprehensible through visualizations or tables.

5. **Conclusion**: Summarize key insights and possible extensions for future research.

The presentation should be created using Beamer and last between 5 to 8 minutes. It should encompass:

1. **Problem Definition**: Highlight the research question and describe the dataset used.

2. **Simulation Study**: Summarize the study design, its relevance to your research question, and key findings.

3. **Results**: Discuss the outcomes when your chosen methodology is applied to the dataset.

1 One Variable Data Set - Uncorrelated

1.1 Estimation Procedures

Research Goal: Investigate various estimation techniques and evaluate their performance metrics using real-world data.

Select a distribution that matches the domain of the data. The distribution needs to have two parameters. It cannot be the normal, gamma, or beta distribution. Evaluate the effectiveness of three different estimators. These estimators may include the maximum likelihood estimate, the Bayesian estimate under squared error loss, method of moments estimators, percentile matching estimators, or others discovered through personal research. For each of the estimators determine uncertainty intervals. This can be done via bootstrap or MCMC sampling.

Evaluate the effectiveness of these estimators using:

- Bias of the point estimate
- Mean squared error of the point estimates
- Coverage of the uncertainty interval

In this simulation study, use at least three different settings for the true underlying parameters and at least three different sample sizes for each set of parameters (at least 9 total settings). Be sure to use effective figures or tables to help the reader visualize the problem and the results.

Explore the data and apply the three estimators and uncertainty intervals using the chosen distribution to the data. **Additional Research Questions**:

- How do the estimators behave when data is sparse or abundant?
- Can the estimators be modified or improved to offer better performance?

1.2 Empirical Model Fit

Research Goal: Utilize computational techniques to assess which probabilistic model best aligns with a given data set.

Select three different distributions that match the domain of the data. Two distributions will have no restrictions, but at least one of the distributions must be a two parameter distribution that is not the normal, gamma, or beta. The goal of this project is to use empirical model selection techniques to select which model fits the data best.

Conduct a simulation study where the data is drawn randomly from one of the three distributions and then all three models are fit. Determine which model fits the data better using two different methods:

- Kolmogorov-Smirnov (KS)
- Akaike's Information Criterion (AIC)
- Continuous Rank Probability Score (CRPS)

Note that the KS and AIC scores use the MLE parameter estimates and the CRPS uses samples from a Bayesian posterior distribution. Use this simulation study to determine how effective these metric are in selecting the correct model. Either use different parameter settings for the distribution you are drawing data from, or try simulating data from one of the other distributions being compared/

Explore the selected data and apply this approach to determine which model fits the data better. Illustrate how the fitted model compares to the data for all three models including using an empirical CDF plotted against the fitted CDF.

Additional Research Questions:

- How do different sample sizes affect the model selection process?
- Are there other model selection metrics that could be more effective in certain cases?

2 One Variable Data Set - Time Dependent

Research Goal: Explore time series model fitting techniques and apply to real data.

If a data set is time dependent, then instead of just applying a probability distribution, a time series model can be built. An AR(p) model looks like

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \epsilon_t, \quad \epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

One important decision is how to determine the lag. There are some techniques that exist in the literature and we can compare how well each of those work. To do this, perform a simulation study with a fixed number of lags and then see how often the techniques return the correct number of lags.

These techniques include

- Testing each lag and comparing AIC
- Testing each lag and comparing BIC
- Using the partial auto-correlation function

The first two require fitting several AR models, so if the third method is at least comparable in performance, it may be preferable. Use three different lags and three different data set sizes to examine this. Apply these methods to a real data set and fit the model with the determined lags.

3 Multivariate Data Set with a Dependent Variable

Research Goal: Explore regression modeling beyond OLS, including developing methods for parameter estimation and hypothesis testing.

This is a regression setting but linear regression using a normal model is not allowed. However, there does need to be a clear dependent variable and at least one independent variable. The data must be fit using either a non-Gaussian likelihood function or a nonlinear model. Some possible examples of this are:

- Minimize least absolute deviations. A linear regression model is the equivalent of minimizing squared residuals $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i})^2$. To minimize least absolute deviations, minimize the sum of the absolute value of the residuals.

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_{1,i} - \beta_1 2 x_{2,i} - ... - \beta_p x_{p,i}|$$

- When the response data in a regression model lies in the domain of positive integers, it can fit in the framework of poisson regression. In a poisson regression model

$$\begin{aligned} f(y_i|\lambda_i) &= \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} && (1) \\ \log(\lambda_i) &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_p x_{p,i} && (2) \end{aligned}$$

Create a hypothesis test for the coefficients to see if they are equal to 0. Do this via a bootstrap confidence interval for the coefficient or Bayesian confidence interval depending on you estimation method. Test this method on two simulated data sets, one where the actual value for the coefficients is 0 and one where it is not 0. Apply this method to your data set to see which variables are significant.

Also compare analysis of the real data set using OLS. Compare model fits using MSPE.

4 Poisson Process

A poisson process is a process that defines how often an event occurs in a specified time. It is controlled by an intensity function $\lambda(t)$. The number of events that occur between time a and time b is distributed as a Poisson random variable with intensity function $\int_a^b \lambda(t)$, i.e.

$$Pr(N(a,b) = \text{Pois}\left(\int_a^b \lambda(t)\right).$$

This is often simplified by assuming the intensity function is constant, $\lambda(t) = \lambda$. Then the number of events between a and b is

$$Pr(N(a,b) = Pois\left(\lambda(b-a)\right).$$

Suppose that the intensity function is instead piecewise, where there are certain change points. For example,

$$\lambda(t) = \begin{cases} \lambda_1 & t < k_1 \\ \lambda_2 & k_1 < t < k_2 \\ \lambda_3 & t > k_2 \end{cases}$$

Alternatively, λ could be a function of t, such as $a + bt$ or $a + bt + ct^2$. Another approach could be estimating parameters in an intensity function such as this.

Data: 75 years of British accidents. Determine when the change points are to determine when the intensity of accidents changed. Perhaps this can gain some insight into what policies or practices of the different periods caused more frequent accidents.

Research Questions: Determine a way to estimate the intensity function in both the case when it is constant and when it is piecewise with 3 pieces, as above. Determine a way to estimate where the change points are.

5 Pairs Trading

The key idea underlying pairs trading is that the movement of the ratio away from its historical average represents an opportunity to make money. For example, if stock 1 is doing better than it typically does, relative to stock 2, then we should sell stock 1 and buy stock 2. This is called "opening a position." Then, when the ratio returns to its historical average, we should buy stock 1 and sell stock 2. This is called "closing the position." The reasoning is quite simple: when stock 1 is priced sufficiently higher than usual, it is likely to go down in value and the price of stock 2 is likely to go up, at least relative to the price of stock 1, since they are positively correlated. Of course, both could increase, but we are interested in relative change as we are looking at the ratio.

Let m be the historical ratio between two stocks. If the ratio moves above or below the the long term average by k standard deviations, then the strategy would be to buy one and sell the other to make a profit.

Data: Dow Jones (DJIA) and S&P 500 (SP500) stock indices.

Research Question: Determine a way to find the optimal value for k to maximize profits. Is that value for k at all related to the correlation between the two stocks? How does the correlation between the stocks affect the final profit?

6 Markov Chain: SIR

An SIR (Susceptible - Infected - Recovered) model is a way that epidemiologists model diseases and their progression over time. The idea is that the proportion in each group evolves over time according to

$$\begin{pmatrix} p_{S,t} \\ p_{I,t} \\ p_{R,t} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ 0 & a_{32} & 1 \end{pmatrix} \begin{pmatrix} p_{S,t-1} \\ p_{I,t-1} \\ p_{R,t-1} \end{pmatrix} \tag{3}$$

which can also be written as $\boldsymbol{p}_t = \boldsymbol{A}\boldsymbol{p}_{t-1}$. The rows in $\boldsymbol{A}$ must sum to 1 to ensure that $\boldsymbol{p}_t$ sums to 1 as probabilities should. The elements in $\boldsymbol{A}$ are transition probabilities of staying or moving to different groups. The initial condition for these are typically a $p_{S,0}$ being close to 1, $p_{I,0}$ being close to 0 an $p_{R,0}$ being equal to 0.

Data: Influenza data at an English boarding school. The data reports the number total of 743 individuals at the school were infected with the disease. This data can tell us about how this disease spreads so we can know what to expect if introduced in a different population.

Research Questions: From being given only the total number and the number affected at each time point, how would you construct the SIR model and learn the parameters? For a given matrix $\boldsymbol{A}$ and total size of a population, how could you determine the expected number affected at any given time. What is a reasonable range for the maximum infected. How long does it take until the maximum number of infected is observed? How are these values affected by the parameters in $\boldsymbol{A}$.

7 Expectation Maximization - Missing Data

The multivariate normal distribution with dimension d has a mean function, μ, and a covariance matrix Σ. Estimating these parameters are not terribly difficult. It gets more complicated when some data is missing, though. Suppose some of the d entries in the observations are missing. Three possible methods are

- Throw out all the records with any missing observations in it
- Use the Expectation-Maximization algorithm and iteratively determine the estimate by calculating the conditional expectations and variances of the missing data

- Instead of using the expected values as above, draw samples for the missing values based on conditional normal theory and then estimate the mean. This method will not converge like the E-M algorithm will, but you can collect samples and evaluate the set of samples afterwards. This is a somewhat Bayesian approach (if you included priors it would be fully Bayesian).

Uses multivariate normal data with some missing value.

Data 1: Characteristics of individuals with Hepatitis. Many of the variables are factor variables. Use the continuous or integer variables, which include age, bibirubin, alkphosphate, sgot, albumin, and protime.

Data 2: Automobile sales. Many of the variables are factor variables. Leave out the factor variables such as make, fuel type, aspiration, doors, style, wheels, engine location and type, cylinders and fuel system.

Research Questions: The above algorithms may work differently when there is a reason why data is missing. For example, perhaps the ones that are missing are missing particularly because they were difficult to measure, too high, or too extreme. Determine a method of guessing if the data is missing at random or if there is a pattern to the missingness. Find estimates of the means and covariances between the variables. Determine which variables are most and least correlated.

8 Stochastic Differential Equations

Some data, such as financial or experiment data, follows something called a stochastic differential equations. A stochastic differential equation in it's most general form can be written as

$$dX_t = \alpha(X_t, t; \theta)dt + \sigma(X_t, t; \theta)dW_t$$

This allows for more specific structures such as the Vasicek model that is commonly used for interest rate data:

$$dX_t = \theta_1(\theta_2 - X_t)dt + \sigma dW_t,$$

the Cox Ingersoll Ross model also used for interest rates:

$$dX_t = \theta_1(\theta_2 - X_t)dt + \sigma\sqrt{X_t}dW_t,$$

or geometric Brownian motion commonly used for stock prices:

$$dX_t = \theta X_t dt + \sigma X_t dW_t.$$

For this project, you can find interest rate or stock data and use the appropriate formula, or some other SDE that you come across. There are a number of approximation techniques used to fit these SDEs. The task is to compare the approximations in how well they produce model fits to return the parameters. To do this, you will simulate data from the SDEs for a given parameter set and then compare the bias and MSE of the different model fits for the different approximation schemes. The approximation schemes to compare are:

- Euler
- Millstein
- Runge-Kutta

Then fit the real data using those schemes.

9 Hamiltonian Monte Carlo or other advanced MCMC method

Hamiltonian Monte Carlo is built on the idea that the proposals can be made in a much smarter approach. It introduces another variable, called an auxiliary variable, p. The joint distribution of the parameter of interest θ and auxiliary variable p is $\pi(p, \theta) = \pi(p|\theta)\pi(\theta)$. Then we let $V(\theta)$ be a negative likelihood, $V(\theta) = -\log(\pi(\theta|X))$. The prior for p is $p \sim N(0, M)$, where M is a covariance matrix, typically the identity, with the dimension equal to the number of unknown parameters.

The way you propose a new value for θ and p is using a leapfrog method:

$$\begin{aligned} p_{1/2} &= p_0 - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}(\theta_0) \\ \theta_1 &= \theta_0 + \epsilon M^{-1} p_{1/2} \\ p_1 &= p_{1/2} - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}(\theta_1) \end{aligned}$$

Then θ_1 and p_1 are accepted or rejected together using the probability

$$min(1, \exp H(\theta_1, p_1|X) - H(\theta_0, p_0|X)$$

where $H(\theta, p|X) = -V(\theta|X) + N(p|0, M)$.

The idea is that by using these advanced proposal schemes, Hamiltonian Monte Carlo is better suited for likelihoods where it is harder to sample the parameters. From the data you chose, estimate the parameters using Bayesian estimation. Try both Metropolis Hastings as well as Hamiltonian Monte Carlo. Explore the differences in model diagnostics, such as effective sample size.

10 Exploring Machine Learning Methods: Neural Networks and Random Forests

Research Goal: Investigate the applicability, effectiveness, and interpretability of machine learning algorithms such as neural networks and random forests in comparison to traditional regression techniques.

Machine learning algorithms like neural networks and random forests have found applications in various fields including natural language processing, computer vision, and bioinformatics. However, their 'black-box' nature often raises questions about their interpretability compared to traditional statistical models.

- Neural Networks: A feedforward neural network can be represented as a composite function

  $$f(x) = f^{(L)} \circ f^{(L-1)} \circ \cdots \circ f^{(1)}(x)$$

  where $f^{(l)}$ represents the function at layer l.

- Random Forests: A random forest aggregates the predictions of multiple decision trees and selects the mean outcome in continuous targets or the most common outcome for categorical. The prediction for a given input x is

  $$y = \text{mode}\{T_i(x)\}$$

where $T_i(x)$ is the prediction from the i^{th} tree for categorical or

$$y = \frac{1}{n}\sum_{i=1}^{n} T_i(x)$$

for continuous targets.

Specific Research Questions:

- Compare the effect of outliers in various models including these machine learning models.
- When we compare feature importances to significance levels in regression models, when will they agree/disagree.
- How accurate are techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to interpret complex models like neural networks? i.e. I can simulate data knowing certain features are significant, can these measures recover that information?
- Are there specific scenarios or types of data where machine learning models outperform traditional models?