Project IV Proposal

Kenneth Wirjadisastra

May 24, 2025

1 Statistical Question

Time series analysis involves the use of previously recorded data to predict future outcomes. In most cases, such as stock market prediction and weather forecasting, this technique will require full data without gaps. In the former case, the data is generally well documented at various resolutions. However, the latter is prone to missing values. This can be due to a number of reasons, including faulty recording equipment or maintenance. This issue raises the need for the imputation of the data. There are many ways to accomplish this task. The goal of this project is to answer the following question: **How does the imputation method impact the performance of our model? Does this change as the number of missing values changes?**

2 Data

The goal is to test the effect of the number of missing values and the imputation methods used. The number of missing values k is based on the percentage of data we wish to alter p. We wish to test low, moderate, and high values of $p \in \{0.1, 0.25, 0.4\}$. The imputation methods are as follows: 1) Kalman Smoothing (KS), 2) unconditional mean imputation (UM), and 3) linear interpolation (LI).

We have a complete dataset of n points, indexed from $0, 1, \ldots, n-1$. We can specify the range of indices $[m, M]$, which can be altered. Generally, we set $m = 1$ and $M = n - 2$ to avoid edge cases in our imputation methods (the first and last values of the data cannot be missing values).

We can simulate data with missing values by sampling k indices from the uniform distribution over the integers $[m, M]$ without replacement. Then we insert missing values into the data set according to these indices. The data can then be imputed using three different imputation methods.

We can repeat this process multiple times to obtain different datasets for each $\langle p, \text{imputation} \rangle$ configuration. For now we will only be using data from a single station, and repeat the process 5 times (possibly more if runtime is not an issue).

3 Estimates

To evaluate the models, we will compare their RMSE, and bias. The RMSE will help us quantify the average error of our model. We pick the RMSE in particular because the model from Project 3 tended to underestimate high temperatures and overestimate low temperatures. The RMSE will penalize these errors more heavily. The RMSE was chosen over the MSE because it is in the original units, making it more interpretable. We also want to measure bias to determine whether the model is over or underestimating on average.

4 Methods

In this experiment, we will not be testing different models / model parameters against each other. Instead, we aim to evaluate how a single model performs based on the amount of missing data and the imputation methods used. The model we are testing is the same as the one from Project 3. It is single layer LSTM model that maps a sequence of previously observed temperatures to the following day's temperature. The model is trained over 50 epochs with an MSE loss function and the Adam optimizer. We expect the model's performance to decrease as the p increases. We also expect the model to perform best with KS imputation and worst with LI imputation.

5 Performance Criteria

To evaluate the methods, we will compare the mean and standard deviation of the RMSE and bias for each $\langle p, \text{imputation} \rangle$ configuration. We can show these results in a table, as well as graphically using box plots.

6 Simulation Plan

To perform this experiment, we generate 5 datasets for each $\langle p, \text{imputation} \rangle$ configuration. There are 3 imputation methods and values of p we wish to test. Therefore, we have a total of 9 $\langle p, \text{imputation} \rangle$ configurations, and 45 total datasets. The imputation methods and values of p are outlined in Section 2. To keep things consistent between all simulations, we will set the same seed before each training loop. We pick a seed of 42. We will record the loss for each model on their respective testing data. The loss is the MSE, so we can take its root to obtain our response variable in the RMSE. We can also obtain the bias by using the following formula

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y_i} - y_i^{\text{true}}).$$

Once we have the RMSE and bias for each model, we can compute the mean and standard deviation among the 5 models. Given that we want to find the bigger picture overall, we treat these 5 as a sample rather than the population of models.

7 Challenges & Limitations

The main issue with this experiment is the required runtime. We will need to train 45 different models. To speed things up, we can terminate if our loss has not improved after 5 epochs. We can also save the model parameters so that we won't have to train the models over and over again.