DS 102 Data, Inference, and Decisions **Spring 2020**

Lecture 22: Time Series Modeling

Lecturer: Jacob Steinhardt

22.1 Recap and Motivation

In the last two lectures, we introduced the idea of sequential decision-making by way of the multi-armed bandits problem. We saw two different bandits algorithms for multi-armed bandits, Upper Confidence Bound (UCB) and Thompson Sampling. While these algorithms work with sequential data, multi-armed bandits is an example of a "stateless" process because we assume that the same thing happens at each time step regardless of what has happened in the past (ie. rewards are independent and identically distributed).

Today, we will focus on modeling "stateful" processes that change over time. In these processes, what happens at the current time depends on what happened on previous time points. In future lectures, we will combine both stateless and stateful processes: reinforcement learning studies processes that change over time, but we also want to make decisions that control these processes.

	Stateless	Stateful
Predictions	Regression	Time series modeling
Decisions	Multi-armed bandits (UCB, Thompson sampling)	Reinforcement learning, control theory

The above table summarizes different types of algorithms and modeling techniques, and whether they apply to prediction vs. decision-making problems and stateless vs. stateful processes. In this lecture, we will focus on the top-right corner: we will introduce time series modeling, building up from simple exponential and logistic growth models. We will also discuss common issues with time series modeling of count data, and explore them by way of a case study of COVID-19 growth data.

22.2 Time series modeling

In the basic setting for time series modeling, we observe variables $x_1, x_2, \ldots, x_T$ up to some time T. Depending on our application, we may also have latent variables $z_1, \ldots, z_T$ in our model. In general, we think of the latent variables z_t as governing the evolution of some process that we might not fully observe.

Example 22.1. We saw an example of this setting when we looked at Hidden Markov Models (HMMs) in Lecture 9. Recall that, in that example, z_t was the total number of fish in a pond at time t, and x_t was the observed number of fish in a random sample from the pond at time t. Note that, in general, we do not have to assume that our time series model has a Markovian structure. However, the examples we look at today will indeed be Markovian.

Tasks that we might want to perform with a time series model include:

- Prediction
  - What will happen next?
  - Given $x_1, \ldots, x_T$, what will x_{T+1} be?
- Filtering
  - Suppose you are an autonomous vehicle and you have LIDAR readings giving noisy measurements of where the cars around you were up to a few seconds ago. Where are you now relative to the other cars?
  - What is z_T given $x_1, \ldots, X_T$?
- Smoothing
  - An astronomer might be interested in asking where the planets were at some previous point in time based on observed planetary measurements up to the current point in time.
  - What is z_t given $x_1, \ldots, x_T$ for some $t \in [1, T-1]$?

Potential real-world applications of time series modeling include population growth, financial forecasting, supply chain optimization, climate modeling, and epidemiology.

22.3 Population growth

We will begin by examining simple models of population growth. For now, we will assume that there are no latent variable—everything is observed.

Let x_t represents the population size at time t. One common model for population growth is an **exponential** model wherein each existing individual creates r offspring at each time step:

$$x_{t+1} = (1+r)x_t,$$
$$x_t = x_0(1+r)^t.$$

Although this exponential model is nice and simple, it is not such an attractive model as stated. For instance, it is a deterministic model — it is highly unlikely that our observed data will *exactly* follow exponential growth. Moreover, this model ignores saturation. Realistically, there is likely some maximum population size (e.g. animal populations will eventually exhaust available resources and

be unable to continue growing). Finally, the model also ignores death in the population. Today, we will discuss ways to address the first two of these issues.

A slight improvement over the simplistic exponential model is a logistic growth model,

$$x_{t+1}x_t + r \cdot x_t \cdot \left(1 - \frac{x_t}{N}\right),$$

where N is the maximum population size. Relative to the exponential model, the $\left(1 - \frac{x_t}{N}\right)$ factor provides a mediating effect on the growth rate. In particular, the effective growth rate will decrease as x_t increases, and will be 0 when x_t reaches the maximum population size N. When r is small, we have the following approximate solution:

$$x_t \approx \frac{N}{1 + \frac{N - x_0}{N} e^{-rt}}$$

We can also consider a stochastic version of this logistic growth model. Instead of the number of individuals being born being a deterministic function, we assume that it is a random variable with the correct expectation. A common choice is to use a Poisson random variable:

$$\begin{aligned} x_{t+1} &= x_t + \Delta_t, \\ \Delta_t &\sim \text{Poisson}\left(r \cdot x_t \cdot \left(1 - \frac{x_t}{N}\right)\right). \end{aligned}$$

The Poisson distribution is an attractive choice because it is the distribution that naturally arises when we have a rare event that happens independently at random for each individual, and we attempt to model the total number of events that actually occur. Notice that this implicitly makes an independence assumption, and when that independence is violated we might observe higher variance than is accounted for by a Poisson. Later today, we will see an example where replacing this Poisson assumption with a different distributional assumption leads to a better model.

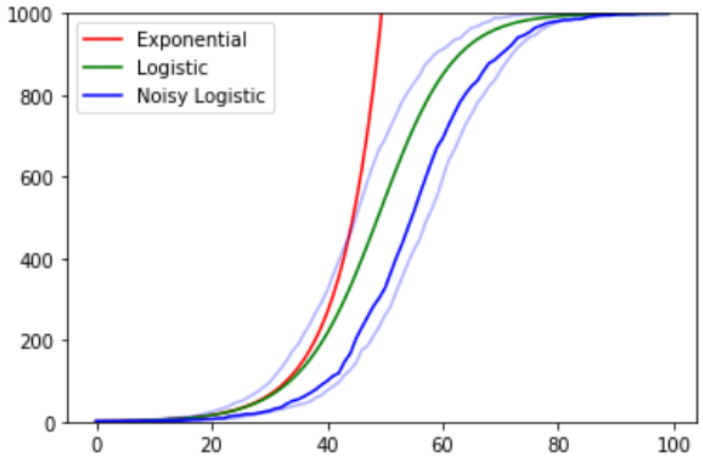

Figure 22.1: Comparison of exponential (red), logistic (green), and stochastic logistic (blue) models of population growth in simulation.

Figure 22.1 compares these different models of population growth. In general, exponential and logistic models look similar up to some point in time and then begin to diverge. In addition, since the Poisson random variable's standard deviation is the square root of its mean, the random fluctuations in the stochastic logistic models tend to decrease in magnitude as the population size saturates.

22.4 Case study: COVID-19

We can apply these basic models of population growth to modeling hospitalizations due to COVID-19. We would like to use a time series model of population growth to do smoothing: given that we observe noisy count data representing COVID-19-related hospitalizations, is there some way to infer whether there is a smooth, underlying growth process?

Let x_t be the observed number of hospitalizations due to COVID-19 on day t. x_t is a noisy, random subset of cases that lead to hospitalizations, so we also include in our model latent variables z_t representing the expected number of hospitalizations on day t. One reasonable model is

$$x_t \sim \text{Poisson}(z_t), \qquad\qquad z_{t+1} = (1+r)z_t,$$

which assumes that z_t grows exponentially.

Is an exponential growth model approximation for z_t reasonable? Some key considerations in answering this question include:

- A logistic growth model would say $z_{t+1} = z_t + r \cdot z_t \cdot (1 - z_t/N)$, which is very similar to exponential growth if $z_t/N << 1$. This suggests it may be reasonable to approximate logistic growth as exponential growth as long as the fraction of infected individuals is quite small. In New York City, current estimates suggest fewer than 10% of individuals are infected with SARS-CoV-2.
- An exponential growth model implicitly assumes that there is unlimited hospital capacity, since it does not account for saturation in the number of hospitalizations.
- It might be the case that there are different subpopulations that grow at different rates. For instance, a denser region might have faster growth than a less dense region. A subpopulation with faster growth may quickly hit saturation in that subpopulation, and then the remaining growth in the overall number of hospitalizations would be due to the slower growth of the other subpopulation(s). This would be better captured by a slight tweak on the logistic model than it is by the exponential model.

Notice that our model also assumes that the latent variable z_t is deterministic. This approximation is reasonable if the population size is large enough that random effects from any individual will wash out. However, this assumption still may not be ideal; for instance, the growth rate may not actually be the same on every day of the week (individuals my interact more on weekends), and this fluctuation in rate over time would cause z_t to deviate from our deterministic model.

When this model is used to analyze real COVID-19 hospitalization data from New York, the resulting confidence intervals for the inferred growth rate are extremely over-confident, and do not accurately capture the magnitude of the fluctuations actually observed in the count data. It turns out that these phenomena arise largely due to the variance of the Poisson distribution being too narrow. This phenomenon of count data having more variability than captured by a vanilla Poisson distribution is called **overdispersion**.

To fix these issues, we can replace the Poisson distribution with a Negative Binomial distribution which also models count data, but has a second parameter called the dispersion parameter which allows it to model greater spread/variance relative to a Poisson distribution with the same mean. In fact, there is a nice interpretation of Negative Binomial distributions in terms of Poisson distributions: Negative Binomials are the distribution obtained by sampling the mean of a Poisson from a Gamma distribution and then sampling from the Poisson. This is analogous to the Beta-Binomial model, which we saw in Discussion 4 is used to model data where a Binomial model would seem appropriate, but the data has higher variance than a vanilla Binomial random variable.

Even after using the Negative Binomial distribution to improve our model, we need to be careful about how we interpret our results. Can we actually use this model to understand whether the growth rate of COVID-19 is leveling off? Some factors to be wary of in interpreting our results include:

- The number of confirmed COVID-19 cases is confounded by underreporting (most infected people do not get tested) and rapid increase in testing capabilities (the total number of people tested increases).
- Death counts for COVID-19 cases have a low sample size and so are very noisy. Moreover, early deaths are often associated with atypical cases.
- All types of COVID-19 data (confirmed cases, deaths, and hospitalizations) come on a time lag. There is a lag of approximately 11 days for hospitalizations and 20 days for deaths, but this lag is itself stochastic.

Combining all of these factors, even with good modeling techniques it is nearly impossible to determine whether policies like shelter-in-place mandates are having an effect on the underlying growth rate until about two weeks later.

22.5 Summary

Time series models give us a formal way to aggregate noisy data from different points in time into a more stable estimate. We saw that it can be very important to model variance/variability in the data (overdispersion in count data) in order to avoid over-confident estimates and misinterpretations of our data. In thinking about interpreting the results of our time series models, we must also keep in mind that data is not reality and may be confounded by underreporting, time lags, and other factors. In the next lecture, we will consider decision-making with time series.