Lecture 12

IDS575: Statistical Models and Methods
Theja Tulabandhula

1 Course Evaluation

- You should have received an email with subject line: "UIC Student Evaluation of Teaching [Subject Code and Number] [Instructor Name] [Semester, Year]."
- Some reasons to submit feedback are:
  - Positive feedback helps keep industry relevant content.
  - Helps improve the course offering to meet the needs of the students in the future and strengthen the degree programs at UIC.
- Pleas do submit before the deadline!

Notes derived from the course material titled "Time Series" (2010) and "Applied Time Series Analysis" (2010)

We will look at a brief overview of time series modeling.

2 Time Series and Supervised Learning

The idea of time series is to capture dependence across observations. So far, we have not modeled any dependency across examples (say (x_i, y_i) and (x_j, y_j)) in a dataset. We will now explicitly model this.

Example 1. One way to measure dependency is via correlation.

Note 1. Because we want to model dependency across observations, the ordering of observations becomes important.

Time series models are applicable in many realistic settings, including:

- Demographic projections
- Financial analysis
- Dynamics of viral media (e.g., view-count of a youtube video)
- Speech and video
- ...

While supervised learning is a broad area, the emphasis is typically on predictive ability and not on modeling dependence across observations. One can certainly model dependencies in a supervised learning setting (say by casting an appropriate maximum likelihood problem). Although, the collection of techniques that explicitly do this have been historically put under the time series modeling umbrella.

Note 2. Also, while the concept of input and output variables is important in supervised learning, it is less important, or even ignored, in time series modeling.

Example 2. Time series of lap times of a Nascar driver is shown in Figure 1.

The various phenomena that we care about with time series modeling include:

- finding trends and seasonality,
- dependence across time, and
- properties of random fluctuations around trend and seasonality patterns.

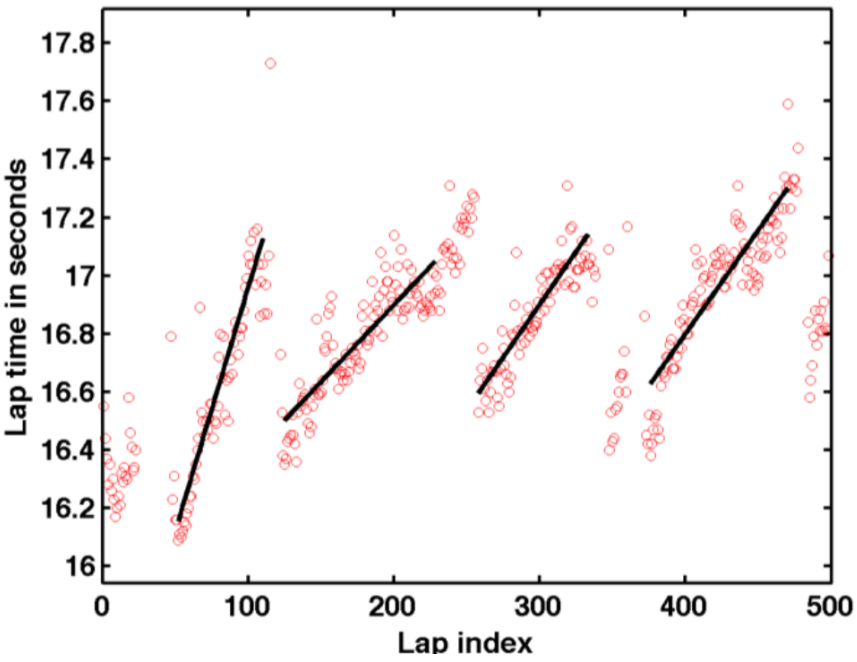

Figure 1: An example of a time series realization.

2.1 Moments vs Distributions

We want to model dependencies across observations. Lets denote random variable W_1 for the first observation, W_2 for the second observation and so on. Say we have N observations.

Note 3. A sequence of random variables is also called a *stochastic process*.

We could define a parametric joint distribution of these random variables as:

$$P_\theta(W_1, .., W_N).$$

The issue with this is that it may be very difficult to specify such a parametric model. A slightly simpler modeling approach is to only focus on *moments*, which are functions such as $E[W_i]$ and $cov(W_{t+\tau}, W_t)$ etc.

A sequence of random variables is called weakly stationary if the following conditions hold:

- $E[W_t] = \mu$ is a constant.
- $cov(W_{t+\tau}, W_t) = \gamma_\tau < \infty$ is a function that only depends on the relative *lag* and not on the absolute index t.

Note 4. These are okay assumptions to make once the predictability using features has already been accounted for. In other words, roughly speaking, first do supervised learning and then apply time series modeling on the residuals.

Note 5. Check: (a) $var(W_t) = \gamma_0$, and (b) $\gamma_{-\tau} = \gamma_\tau$.

The function γ_τ is called the autocovariance function. The function $\rho_\tau = \gamma_\tau/\gamma_0$ is called the autocorrelation function.

Note 6. A sequence $W_1, ..., W_N$ is called white noise if $E[W_t] = 0$ and $E[W_t^2] = \sigma^2$. If the W_t are independent and identically distributed, then they are sometimes called *i.i.d noise*.

Example 3. An example of a non-stationary time series is as follows: Let W_t be a white noise sequence. Let $S_0 = 0$ and $S_t = W_1 + W_2 + \ldots + W_t$. Then the sequence $S_0, S_1, ...$ is called a random walk. We can check that $cov(S_t, S_{t+\tau}) = t\sigma^2$.

Why stationarity? Well, it allows us to average. We are only observing one realization of each random variable. Still, we are allowed to average the values of these realizations across (time).

Example 4. Sample autocorrelation and sample autocovariance functions can be estimated similar to estimating sample mean, which we have seen before.

How do we estimate seasonality and trend effects? Well, there are approaches that filter out "noise" such that trends or seasonality patterns remain. We will skip the details of these techniques for now.

3 The Autoregressive Moving Average (ARMA) Model

First we describe two linear time series models: the autoregressive model and the moving average model.

By linearity, we mean that the random variable at time t is linearly related to other random variables.

3.1 Autoregressive (AR) Model

An autoregressive (AR) model assumes that the random variable at time t is related to the random variables before time t. For instance, an AR(1) model is written as follows:

$$W_t = \phi W_{t-1} + \epsilon_t,$$

where ϵ_t is a white noise random variable with $var(\epsilon_t) = \sigma^2$.

If ϵ_t is independent of $W_{t-1}, W_{t-2}, ...$, then the sequence is Markovian. That is, $P(W_t|W_{t-1}, W_{t-2}, ...) = P(W_t|W_{t-1})$.

If we repeatedly *back-substitute*, we get the following expression:

$$W_t = \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \phi^{t-1}\epsilon_1 + \phi^t W_0.$$

For this sum to be finite for any realization (so that stationarity holds), we need $|\phi| < 1$.

Example 5. The mean of the t^{th} random variable in an AR(1) model is:

$$E[W_t] = \sum_{j=0}^{t} \phi^j E[\epsilon_{t-j}] = 0.$$

The autocovariance function is:

$$\begin{aligned}
\gamma_\tau &= cov(W_{t+\tau}, W_t) \\
&= E[\sum_{j=0}^{t+\tau} \phi^k \epsilon_{t+\tau-j} \sum_{k=0}^{t} \epsilon_{t-k}] \\
&\approx \sigma^2 \sum_{k=0}^{\infty} \phi^{k+\tau}\phi^k \\
&= \sigma^2\phi^\tau \sum_{k=0}^{\infty} \phi^{2k} = \frac{\sigma^2\phi^h}{1-\phi^2}.
\end{aligned}$$

3.2 Moving Average (MA) Model

Say $W_t = c_t\epsilon_t + c_{t-1}\epsilon_{t-1} + ...$, where ϵ_t is white noise. This is called a moving average (MA) model. Then γ_τ for the W_t time series is given by:

$$\gamma_\tau = \sigma^2 \sum_t c_t c_{t+\tau}.$$

More specifically, a MA(q) model would be:

$$W_t = \theta_0\epsilon_t + \theta_1\epsilon_{t-1} + ... + \theta_q\epsilon_{t-q}.$$

The mean of MA(q) would be $E[W_t] = 0$.
The covariance would be:

$$\gamma_\tau = \sigma^2 \sum_{t=0}^{q-\tau} \theta_t\theta_{t+\tau}; \;\; 0 \le \tau \le q.$$

3.3 Combining AR and MA

We can combine to two models above to get the following model:

$$W_t - \phi_1 W_{t-1} - ... - \phi_p W_{t-p} = \epsilon_t + \theta_1\epsilon_{t-1} + ... + \theta_q\epsilon_{t-q},$$

where ϵ_t represents white noise.

Why do we combine? Because, MA and AR(1) capture certain covariances, but not all. For example, for AR(1), the covariances are all determined using a single parameter. For MA(q), the number of parameters is q.

In terms of notation, we can compactly write this as:

$$\phi(B)W_t = \theta(B)\epsilon_t,$$

where $\phi(B)W_t = W_t - \phi_1 W_{t-1} - ... - \phi_p W_{t-p}$, similarly for $\theta(B)\epsilon_t$.

Note 7. B is a shorthand symbol for the backshift operation. For example, $BW_t = W_{t-1}$ and $B^k W_t = W_{t-k}$.

Not all coefficients can go into the above model equation. In particular, we will need the following conditions:

- The roots of the polynomial $\theta(B)$ are different from the roots of the polynomial $\phi(B)$ (if this was not true, terms would cancel out on both sides).
- $\phi(B) \neq 0$ for all $|B| \leq 1$ (allowing B to be a complex number). This ensures a property called causality (W_t only depends on variables before time t)
- $\theta(B) \neq 0$ for all $|B| \leq 1$ (allowing B to be a complex number). This ensures a property called identifiability (otherwise there are multiple θ coefficients that correspond to the same autocovariance values.)
- Ultimately, we are writing $W_t = \psi(B)\epsilon_t$, where $\psi(B) = \theta(B)/\phi(B)$. We need to ensure that this polynomial is nice (it will have infinite terms, but hopefully the coefficients are small).
- We will also assume for simplicity that $\phi(0) = \theta(0) = 1$.

3.4 Estimation

How does one estimate all these parameters? How to determine p and q? The estimation of both these is based on covariance or correlation estimates.

Example 6. For example for MA(q), if you plot the covariance values for different τ values, you will observe that for $\tau > q$, the covariance will be nearly zero.

Example 7. For AR(p), there is another quantity called the partial autocorrelation function, that can be plotted to determine p. This is defined as:

$$\begin{aligned}\phi_{11} &= corr(W_1, W_0)\\ \phi_{\tau\tau} &= corr(W_\tau - W_\tau^{\tau-1}, W_0 - W_0^{\tau-1}), \quad \tau \geq 2,\end{aligned}$$

where $W_\tau^{\tau-1}$ is the regression of W_τ on $(W_{\tau-1}, ..., W_1)$ and $W_0^{\tau-1}$ is the regression of W_0 on $(W_1, ..., W_{\tau-1})$. Seems complicated, but it just ensures that beyond p, the plot of partial autocorrelation dies down.

Example 8. Estimation of the parameters of, say, an AR(p) model can be done using a specific *method of moments* technique called the system of *Yule-Walker equations*. In general, such estimation can be done using the method of maximum likelihood as well.

Note 8. In the literature, you may come across a methodology called the *Box-Jenkins* approach, which describes the steps relating to estimating p and q and then estimating the model parameters and then testing their validity.

Note 9. There is another model called ARIMA which just means that you may have to *difference* the time series a few times before fitting an ARMA model. Differencing means subtracting successive values.

4 Summary

We learned the following things:

- Relation between time series models and supervised learning.
- The Autoregressive Moving Average (ARMA) model.
- There are many time series specific techniques beyond what we saw here. For instance, time domain and frequency domain methods, methods based on distributional assumptions etc are all very useful to know when dependency between observations is a key aspect of your dataset.

A Sample Exam Questions

1. In time series models, why do we focus on moments of the distribution rather than the distribution itself?
2. What are the key differences between an autoregressive (AR) model and a moving average (MA) model?

B List of Topics

1. Expectation Maximization
2. Sampling from the posterior
3. Generalized additive models
4. Tree based methods
5. Classification and regression trees, their issues
6. Missing data
7. Adaboost

8. Gradient boosting models
9. Interpretation using relative importance and partial dependence plots
10. Random Forests and bagging
11. MARS
12. SVMs and the dual formulation
13. The kernel trick
14. Association rules
15. Clustering: dissimilarity and algorithms for clustering
16. Principal components
17. Spectral clustering
18. Basics of time series