Analysis of Unevenly Spaced Time Series

Suman Chakraborty

University of North Carolina, Chapel hill

sumanc@live.unc.edu

November 16, 2016

Plan for the talk

- Spectroscopically Confirmed Quasar data.
- Observed as an unevenly Spaced time series.
- Traditional Methods of modeling these data.
- Another method and analysis of the data.

Quasar Data

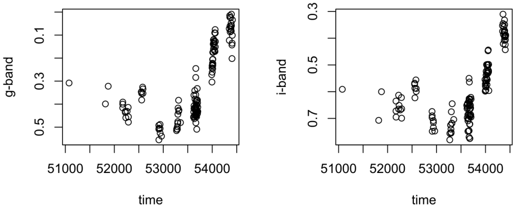


Figure: Plot of the Quasar data under two bands.

- Notice that the time series are unevenly spaced.
- Also the data is "bursty" in nature, i.e. the observations are made in short bursts of time followed by a relatively large gap due to instrumental constraints.
- Our aim is to propose a method to model these data.

Existing Methods(that I know of)

- Using a damped Random walk Model (Proposed by by Kelly et.al, 2009), it is also known as OrnsteinUhlenbeck process given by,

  $$dx_t = \theta(\mu - x_t)dt + \sigma dw_t,$$

  with $\theta, \mu, \sigma > 0$, and w_t is a SBM.The covariance tructure of this Gaussian process is given by $\frac{\sigma^2}{2\theta}\left[e^{-\theta(t-s)} - e^{-\theta(t+s)}\right]$.
- Another one put forward by Butler and Bloom, 2011 using Gaussian process with a slightly different Kernel .
- Both of the works assume that the observations are coming from a Gaussian Process and the problem reduces to the parameter estimation problem related to the process.

Pros and Cons

- These models are characterized by parameters.
- Limiting behaviors are well understood.
- Extensive literature on estimation of Gaussian processes.

On the other hand..

- These models are stationary.
- It does not account for the underlying data generating process.

The TVAR model

- The time varying AR model was proposed in Azrak and Melard, 1999 particularly for non-stationary and unevenly spaced model.
- The model assumes that there are finitely many states and each time point is associated with one of the states(These type of models were studied in Probability as Markov switching processes).
- First suppose that there are d states denoted by $\{1, \dots, d\}$ and each time point is associated with one of the state. Let τ_k be the state at k-th time point and x_k be the observation made at k-th time point,

$$x_k - \mu(\tau_k) = \sum_{i=1}^{p} \phi_i(\tau_k)\{x_{t-i} - \mu(\tau_{k-i})\} + \epsilon_k$$

- One may assume the error also depend on the state .

When there is no notion of states available

We try to formulate a model that will capture the scenario when there is no natural notion of states.

- Assume that we observe the data at time points $\{s_t\}_{t=1,2,\dots}$ and $s_t \leq s_{t+1}$, also, for defining purpose we assume that the states are known(More on it later).
- Suppose x_t be the observation made at time point t. Then the order-p model is given by ,

$$x_{t+1} - \mu(s_{t+1}) = \sum_{i=0}^{p-1} \phi_i(|s_{t+1} - s_{t-i}|)\{x_{t-i} - \mu(s_{t+1})\} + \epsilon_t.$$

- We will assume the errors are iid with mean zero and variance σ^2.

A simple version of the Model

For simplicity we will work with the model with constant mean $= \mu$ given by,

$$x_{t+1} - \mu = \sum_{i=0}^{p} \phi_i(|s_{t+1} - s_{t-i}|)\{x_{t-i} - \mu\} + \epsilon_t.$$

Also, motivated from evenly spaced AR model we consider the coefficients to be parametric function,

$$\phi_i(|s_{t+1} - s_{t-i}|) = c_i^{|s_{t+1} - s_{t-i}|},$$

for $i = 1, \dots, p$.

Coming back to the data

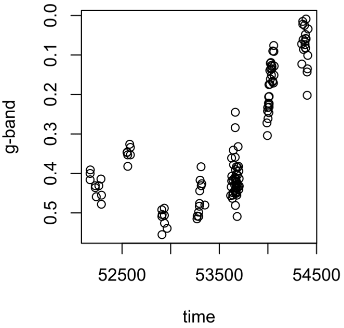


Figure: Plot of the Quasar band(except first three points.)

We will fit the simple version of the model of order 2.

Fitted Model

We fit the model using simple least square minimization. Also we estimate the variance by standard mean square error.

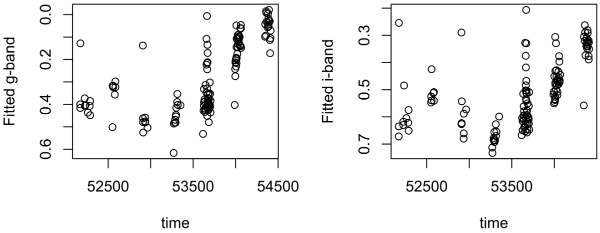


Figure: Plot of the fitted Model

For the g-band, the parameter estimates are $\mu = -0.0298589971, c_1 = 1.0005915992, c_2 = 0.0006319854$. For i-band $\mu = -0.04744262,\ c_1 = 1.00005692, c_2 = 0.20619338$.

Original vs Fitted

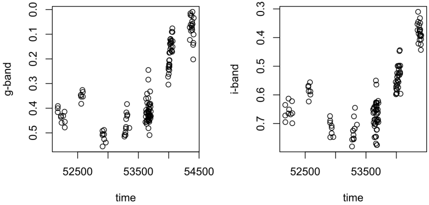
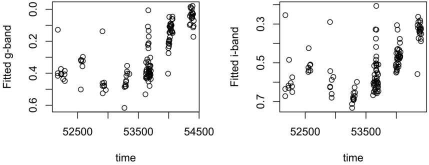


Figure: Plot of the Quasar band(except first three points.)



Prediction

To test the prediction-ability of this model we fit the model on first 104 observation and predict the rest 30 values using the fitted values, the parameter estimates are: $\mu = -0.007571102$, $c_1 = 0.999603893$, $c_2 = 0.001061915$.

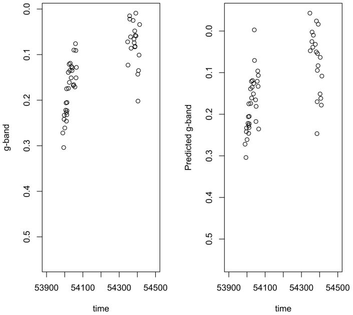


Figure: Predicted Vs Original

Predicted Vs Original

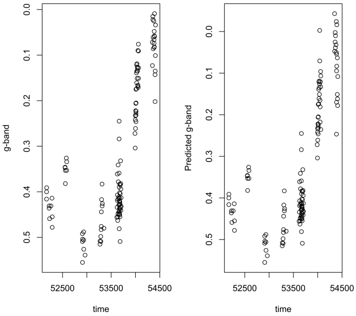


Figure: Predicted Vs Original

Remarks and Questions(without answers)

The empirical performance of the model seems satisfactory but there are many unanswered questions and scope of improvements.

- We have considered the parameters to be a function of time lag. Is there any better choice for this?
- We assumed a particular parametric(exponential) form of the coefficient function. What are other “natural” choices? What about non-parametric methods?
- How does the model compare with other available models?
- The computation was expensive even for the small data set. Better methods? Other risk functions?
- What about theoretical properties of the estimator?

Currently we do not have definite answer to the above questions and some of them are part of the presenter’s ongoing investigation.

Data Generating Process: “Bursty Data”

Throughout the above analysis we have assumed that the data generating process is deterministic. It has at least two drawbacks for general purpose:

- It does not capture if there is an inherent pattern in the data generating process.
- For prediction purpose there is no other way to predict the next observation without knowing(or modeling) the data generating process.

The process looks like “bursty traffic data” in case of the Quasar data,although for different reason altogether.

“Bursty traffic data”

It is an extremely useful model in telecommunication , it particularly captures the scenario when data come in short bursts . For example a facebook group of a particular team will have huge number of posts on game days. The model is used for the purpose:

- The gap between each bursts of data follows independent exponential distribution with mean λ.
- The number of observations in a particular burst follows a geometric distribution with parameter p.
- The gap between observations within a burst i follows exponential distribution with mean λ_i.

Conditional Time Series

Note that with sufficient number of bursts and sufficient number of data within each burst one can consistently estimate all the parameters mentioned above.
Now we can proceed with the analysis of the time series conditioned on this process exactly as above and achieve extrapolation capability with the data generating process.

Happy Thanksgiving

THANK YOU.