Time series and data cloning slides

Subhash Lele

2024-08-27

Time series data: Missing observations, measurement error and prediction

- We will demonstrate how one can use MCMC and data cloning to analyze time series data sets.
- These examples can be extended to longitudinal data sets, spatial data sets quite easily.
- These models also tend to have missing observations. The code can be modified easily to account for the missing observations for estimating the parameters.
- We may also want to *predict* the values of the missing observations. We will demonstrate how to predict missing data or forecast future observations.

Auto-regression of order 1 (AR(1) process)

This is one of the most basic time series models. The AR(1) model can be written as:

$$Y_i = \rho Y_{i-1} + \epsilon_i$$

where $i = 1, 2, ..., n$ and $\epsilon_i \sim N(0, \sigma^2)$ are indepedent random variables.

This model says that the next year's value is related to the past year's value. That is, the value in the past year is a good predictor for the next year's value. Hence the term 'auto-regression'.

More applications

- This model can be used to model many different phenomena that proceed in time. For example, tomorrow's temperature can be predicted using today's temperature (except in Alberta!). Next day's stock price is likely to be related to today's price and so on.
- This model can be modified to include covariates. Hence, we can write:

$$Y_i = X_i\beta + \rho(Y_{i-1} - X_{i-1}\beta) + \epsilon_i$$

This allows for correlated environmental noise in regression.

- This model has been used to model population changes in wildlife populations. It is useful in epidemiology and so on. Many econometric models are derivatives (derived from) of this basic model.

Measurement error

Of course, reality is most of the times more complicated than this model. For example, one may not observe the response without error. This is called an observation error. We may not (most of the times, we will not) observe the true population size but only an estimate of the true population size. Thus, the observed value is not the true response. Such cases are modelled using a hierarchical structure:

Hierarchy 1 (True state model):

$X_i = \rho X_{i-1} + \epsilon_i$

Hierarchy 2 (Observation model):

$Y_i = X_i + \eta_i$ where η_i is observation error and $\eta_i \sim N(0, \tau^2)$ are independent random variables.

Kalman filter

This is what is called a 'Kalman filter' after a famous Hungarian (Yeah, Peter!) electrical engineer Professor Rudolf Kalman. This is a particular case of the model class 'State space models'. They consist of at least two hierarchies: one models the true underlying phenomenon and the other the observation process that models the error due to observation process.

- Under the Normal distribution assumption, the mathematics can be worked out for the simple linear model to conduct the likelihood inference.
- But once we enter the non-linear time series modelling or non-Gaussian observation processes, mathematics become nearly impossible.

Can we use MCMC and data cloning?

For the time being, let us avoid all the mathematics and see if we can use JAGS and dclone to conduct the statistical analysis. NO MATH PLEASE!! is our motto.

Identifiability

Try running the above code when true rho=0. Are the parameters estimable? Does the Bayesian approach tell you that? Without the estimability diagnostics, you could be easily mislead by the Bayesian approach. You will merrily go around with the scientific inference when the parameters are not estimable.

Clipped or Censored time series

Suppose the underlying process is AR(1) but the observed process is a clipped process such that it is 1 if X is positive and 0 if X is negative. This is called a clipped time series. Similarly you may observe Y to belong to an interval. This is an interval censored data. Above model can be modified to accommodate such a process. An easier way to model binary, count or proportion time series is as follows.

Modelling binary and count data time series:

For modelling binary time series, we can consider the observation process as:

$Y_t \sim Bernoulli(p_t)$ where $log(\frac{p_t}{1-p_t}) = \gamma * X_t$

For modelling count data time series, we can consider

$Y_t \sim Poisson(\lambda_t)$ where $log \lambda_t = \gamma * X_t$.

This is a time series generalization of the GLMM that we considered before.

Caveat: It is extremely important that you check for the estimability of the parameters for these models. Because of clipping and other observation processes, you are more likely to run into estimability issues. As far as we know, data cloning is the only method that allows estimability diagnostics as part of the estimation process. See Lele (2010, Ecology).

Non-linear time series analysis

Now we will consider a non-linear time series model, Beverton-Holt growth model, that is commonly used in ecology.

In ecology and population biology, one wants to understand how abundance changes over time. Following Malthus' thinking, it is also evident that, in a finite environment, abundance cannot increase without limit. Thus, the growth usually is exponential in the beginning (low population, ignore the Allee effect for now) and then it slows down as we approach the carrying capacity.

Beverton-Holt model

One commonly used model is the Beverton-Holt model (discrete analog to the continuous Logistic model).

Let $Log(N_t) = X_t$ where N_t is the abundance at time t. A general form for the population growth models is:

$$X_t = m_t + \sigma Z_t$$

where Z_t is $Normal(0, 1)$ random variable. This is similar to the AR(1) process but with a non-linear mean structure.

If $m_t = log(\lambda) + x_t - log(1 + \beta N_t$, the population growth model is called the Beverton-Holt model. It has an upper limit $K = \frac{\lambda - 1}{\beta}$ called the Carrying capacity, the maximum population size that can be attained (with some perturbation).

Estimated population abundance, not the true abundance

In practice, we usually have to conduct some sampling to *estimate* the abundance. Hence there is measurement error. We can represent this process by using hierarchical model:

Hierarchy 1: Process model

$X_t|X_{t-1} = x_{t-1} \sim Normal(m(x_{t-1}), \sigma^2)$

Hierarchy 2: Observation model

$Y_t|N_t \sim Poisson(N_t)$

Likelihood function and high dimensional integration

One can use other observation models as well. We can write the likelihood function for this using a T_max (length of the time series) dimensional integral. If the time series is of length 30, this will be a 30 dimensional integral. In order to compute the MLE, we will need to evaluate this integral repeatedly until the numerical optimization routine converges. This is a nearly impossible task.

The code to analyze this non-linear time series with non-Gaussian observation error can be written as follows.

Prediction for future states (Extinction?)

Parameters for this model are estimable. It will be interesting to see if one can use Negative Binomial distribution (one additional parameter) instead of the Poisson distribution. Are the parameters still estimable? (TBD!!!)

We are also interested in predicting the true population abundances as well as forecasting the future trajectory of the abundances. This is quite easy under the Bayesian paradigm. The frequentist paradigm involves an additional step.

Let us see how to use the Bayesian paradigm and MCMC to do this. In the Bayesian paradigm, there is no difference between parameters and the unobserved states. They both are considered random variables.

On the other hand, in the frequentist paradigm parameters are fixed but unknown (not random) whereas the unobserved states are true random variables. We (the instructors) consider these to be different.

Key features of prediction versus estimation

1. Information about the parameters converges to infinity as the sample size increases. Thus, we can *estimate* them with high degree of confidence. The *confidence* intervals shrink as we increase the sample size.
2. Information about the states (random variables) does not converge to infinity as the sample size increases. The *prediction* intervals do not shrink.

This should be familiar to most of you from your regression class.

Back to R coding

Summary

- We have illustrated how one can use R, JAGS (implicitly MCMC) to obtain statistical inference for any complex model without doing the hard mathematics.
- If you can do Bayesian inference, you can also do the frequentist inference
- Which one should we use?