Detecting the Evidence of Red-Queen Hypothesis in Daphnia-parasite Interaction with Dynamic Regression

Jan Greve

Abstract

One of the fundamental questions in evolutionary biology is why majority of the species have evolved to reproduce sexually. Although the Fisher-Muller hypothesis which provides some justification to it has been around for a century, a more recent gene-level explanation to such phenomena has been proposed under the name of "Red Queen Hypothesis" by Van Valen (1973)[1].

In this project, I made a framework to assess this hypothesis through Dynamic regression and in the process made a missing value imputation algorithm due to the nature of the Daphnia-parasite interaction data I use. However, because of the sheer number of the missing data, no reasonable evidence in support of the hypothesis was found.

1. Introduction

The Red Queen Hypothesis, often referred to as "evolutionary arms race" hypothesizes a constant competition amongst opposing organisms for survival. This results in a dynamics called co-evolution characterized by constant adaptation to each others' survival strategy.

Although the hypothesis has a general scope, it is often used in the context of host-parasite interaction. Specifically, the co-evolution between host and parasite is expected to follow a dynamics in which a parasite starts to specialize in targeting a specific dominant genotype in a host, which results in a decrease in proportion of the host population with that gene, while allowing unaffected genotypes to grow in shares which eventually leads to a parasite switching the target to another dominant type and will go on and on in a similar manner.

This hypothesized dynamic change in gene shares within the host population is used to explain the benefit of sexual reproduction which experiences this dynamics in much tempered manner. This is due to high within and across generation diversity where the former contributes in robustness by ensuring that some genotypes remain unaffected while the latter in smoothness of transition by enabling unaffected genotypes to gradually increase in shares delaying the parasite's eventual switch in specialization.

Conversely, in order to clearly distinguish the hypothesized dynamics from random genetic drifts, we should use the data from asexually reproducing organisms. For this reason, I used the data from Turko et al (2018)[2] which deals with the Daphnia reproduction in relation to the parasite activities. In the next section, I will go over the details of this data.

2. Data Description

Variable Name	**Type**	**Description**
Daphnia Diversity	[0,1]	Diversity measure of Daphnia population
Parasite Infection Rate	[0,1]	Proportion of sampled Daphnia infected
Oxygen Density	$\mathbb{R}$	Oxygen density in the lake
Temperature	$\mathbb{R}$	Temperature in the lake

Table 1: Data set used in this project

The data set above is based on that of Turko et al (2018)[2]. The *Daphnia Diversity* variable was originally a categorical count time series data where each category represents a particular genotype. In order to utilize the normal framework of the DLM (we could have used Poisson framework too, but the dimension is quite high), I transformed this into univariate diversity measure using Simpson's diversity index, a typical metric for this purpose. Other data is taken directly from the original data set.

For analysis, I transformed the variables with 0-1 support with logit function so that all variables have real support.

The most serious challenge I encountered in this project is with the irregular spacing of the data. Detailed methodology to resolve (more like partially resolve) this issue is mentioned in later sections.

3. Model Specification

3.1. Overview

Fundamentally, we are interested in lerning the structure of the dynamic relationship between *Daphnia Diversity* (DD) and *Parasite Infection Rate* (PIR) which by domain knowledge is expected to be confounded by factors such as *Oxygen Density* and *Temperature*.

Therefore, the most straightforward approach is to use dynamic regression of DD on PIR and its lags while also including other factors to control for confounding. Since we are not aiming for exact causal inference, but rather on finding some potential evidence for co-evolution, this is a justifiable methodology. We are also not interested in forecasting, thus the importance of prior and discount specification is slightly less than that of more forecast oriented projects, provided that the innovation term can capture the heterogeneities that the model missed in some extent by making appropriate corrections in the filtering.

Instead, majority of my time on this project is spent on missing value imputation due to the irregular sampling interval of the data.

3.2. Missing Value imputation

Because of the sampling irregularity, our data when converted to equally spaced time-series has missing values on all four variables.

For this project, I resorted to impute missing values of the predictors separately, each with a model and then used the posterior samples of the predictors to impute the response series. All this is done within the gibbs sampling framework, and the detailed algorithm will be shown in next subsection.

3.3. Models and Sampling Algorithm

For PIR and other factor series, after initializing the missing value, I separately modeled each of them with TVAR(p) and run forward forecasting and smoothing. Once retrospectively smoothed parameter estimates were obtained, I used the multiple observation deletion (deleting the imputed/initialized missing values) and imputation procedure (under the discount weighted stochastic variance DLM) mentioned in Harrison and Veerapen (1993)[3]. The resulting algorithm for univariate predictor imputation is shown below.

Algorithm 1 Univariate Imputation (predictors)

```
procedure GIBBS FOR PREDICTOR $X \equiv \{X^{\text{OBS}}, X^{\text{MISS}}\}$
    Set prior,discount and model params $\phi$
    Initialize $X_0^{\text{miss}}$
    for $s = 1$ : number of Gibbs runs (=S) do          ▷ until good mixing
        $X^{\text{obs}} \sim X_{s-1}^{\text{miss}}, \phi$          ▷ by FFBS of TVAR(p)
        $X_s^{\text{miss}} \sim X^{\text{obs}}, \phi$          ▷ by deletion & imputation
    return $X_s \equiv \{X^{\text{obs}}, X_s^{\text{miss}}\} \forall s \in S$
```

Unfortunately, the potential confounding factor series (*Oxygen Density* and *Temperature*) had too many missing values so that regardless of any reasonable prior and discount specification, there where simply not enough degree of freedom. Thus, I dropped these from the analysis (which invalidates the inference on the following dynamic regression ...). For PIR series, I used TVAR(3) which seemed like an optimal balance between the speed of mixing and accuracy of the imputation after trying out various values of p. More detailed diagnostics on the missing value imputation procedure will be mentioned in later section.

Once we obtained posterior draws of the missing values of the predictors coupled with the observed values, we can use similar procedure to Algorithm 1 to obtain the missing values of the response, and at the same time generate samples of the parameters from the posterior (after some burn-in of course) with Dynamic regression as shown below.

Algorithm 2 Response imputation & Parameter estimation

```
procedure GIBBS FOR RESPONSE $Y \equiv \{Y^{\text{OBS}}, Y^{\text{MISS}}\}$ AND PARAMS
    Run Algorithm 1 and obatain $\mathrm{X}_s(\mathrm{T}\times S)$
    Set prior,discount and model params $\psi$
    Initialize $Y_0^{\text{miss}}$
    for $s = 1 : \mathrm{S}$ do
        $Y^{\text{obs}} \sim Y_{s-1}^{\text{miss}}, X_s, \psi$            ▷ by FFBS of Dynamic Regression
        $Y_s^{\text{miss}} \sim Y^{\text{obs}}, X_s, \psi$                    ▷ by deletion & imputation
    return $Y_s \equiv \{Y^{\text{obs}}, Y_s^{\text{miss}}\} \forall s \in S$ & Posterior of Coefficients etc.
```

Since we are not jointly imputing predictors and response, as shown above in the algorithm, we can first run the sampler for the predictor and once that's done plug in those values for each Gibbs run in response imputation. This is because imputed value of the response has no implication to the predictors in this particular setting. This specification of course comes with the cost of lower accuracy in imputation since we are not using the information between these two series to impute each other.

Before showing the result of these procedure in our main data, let's check how these algorithms perform in known toy data by randomly removing observations of two related series.

3.4. Test of Missing Value Imputation algorithms with Sales-Index Data

To check the performance of the imputation algorithms, I randomly deleted around 20% of the observations in both Index and Sales series from the homework (deletions are done separately, so it may or may not overlap).

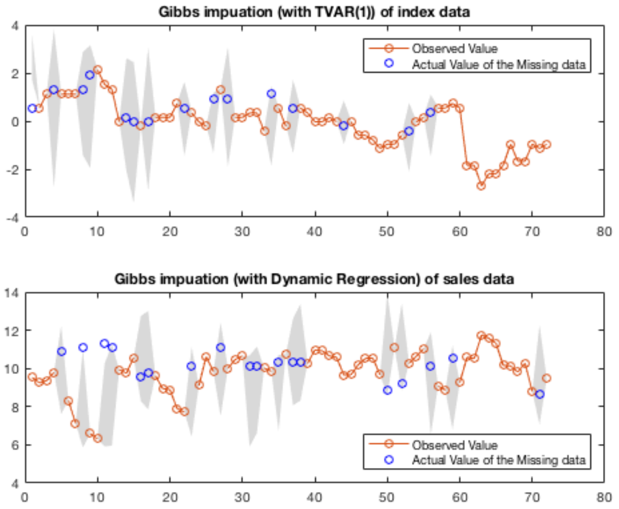

Figure 1: 95% intervals of missing values and original data (above:Index, below: Sales)

As we can see from the figure above, The imputation of univariate Index series with TVAR(1) gives reasonable prediction intervals for all missing values.

However, once we move on to the imputation of the Sales data with these posterior samples of predictors as well as levels and seasonalities (used the identical specifications as in the homework), our predictive intervals do seem to systematically miss the actual observations, especially those around time 10.

Although there may also be some model specification or sampling related issues behind it, my guess for the cause of this relatively poor result is because we are only conditioning the 2nd stage imputation (Sales data part) with the result in the 1st stage, but not using the other direction of the conditional dependence to infer the distribution of the 1st stage samples through samples of the 2nd stage.

Additionally, let's also check whether trajectory of the smoothed posterior distribution for parameters are consistent with the result without missing values.

The figure 3 is smoothed trajectory of bounds for the coefficient of the Index obtained from full data while that the figure 2 is the same thing but averaged out across each samples of Gibbs imputed FFBS. Both models uses the same prior and discount specification.

We can observe that these two figures have very close bounds most of the time, although the sudden drop in uncertainty around time 30 observed in the full data case (Fig3) is not present in the imputed case (Fig2). This might be because some influential points which lead to that drop in uncertainty was deleted from our missing data set. Therefore, the imputed estimate did not catch that regime switch and instead may have assumed

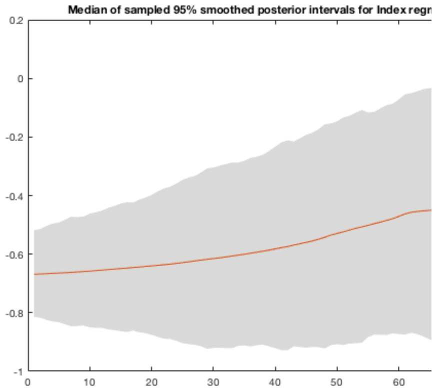

Figure 2: Trajectory with imputation

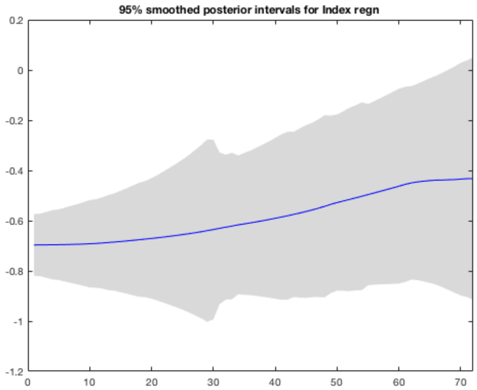

Figure 3: Trajectory without imputation

smooth transition by learning from imputed synthetic observations which replaced those deleted points. In fact, the imputed bounds are typically more smooth although it is consistent with the full data bounds in the general characteristics such as gradual increase in median and mostly graduate increase in the uncertainty etc.

Now that we checked with the toy data that our imputation algorithm is not too far off from the truth (although there are ways to improve it further), we can move on to the results in the real data of our interest.

4. Results from Daphnia Data

The first step is to impute the PIR series which will be used as a contemporary and lagged predictors to the DD series. The Figure next page shows the result of the imputation.

Although the data starts from 2002 to 2011, due to the sparse nature of the data, I decided to take a subset of it from July 2003 to December 2004. Also, since the winter to early spring period is where Daphnia reproduces sexually, the data around this time is generally lacking and thus I connected December 2003 to early June 2004. Generally, the sampling frequency increases from June due to the active asexual reproduction around this period. So focusing on the period starting from June to the end of the year is justifiable. Additionally, no special discount has been applied between the period of December 2003 cutoff to June 2004.

Unfortunately, the sparsity of the data coupled with the relative lack of clear trend (strong positive correlation) or counter-movements (strong

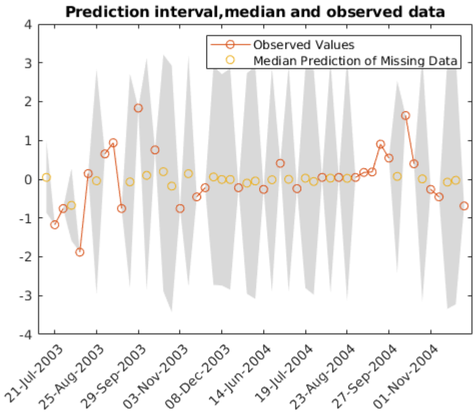

Figure 4: Imputation results (with TVAR(3)) of PIR series

negative correlation) made the prediction intervals of the data quite large and uninformative.

Although this result is already suggestive of the quality of the 2nd stage imputation and the subsequent parameter learning, I nonetheless tried dynamic regression with posterior of PIR series and its lags (up to lag-2). The resulting trajectory of the parameters are shown below.

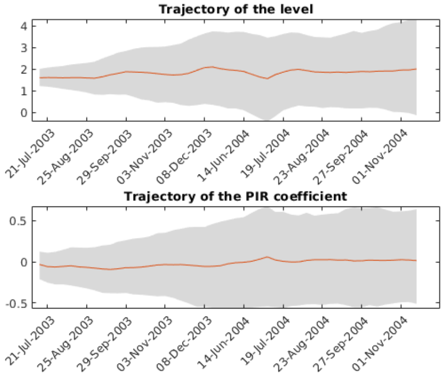

Figure 5: Trajectory of level and PIR coefficient

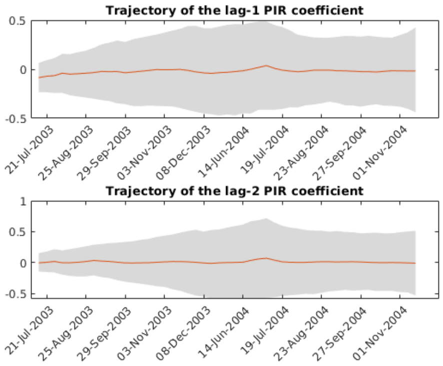

Figure 6: Trajectory of Lag 1,2 coefficients

As expected, none of the PIR series have a bound not containing 0. Apparently, the trajectory of the lag-1 coefficient throughout most of the time has negative median value as opposed to other lags which is centered around the zero.

If we were to make any inference on this result (ignoring the real possibility of confounding), the general tendency for the lag-1 PIR series to be more likely to be negative than positive is understandable. This is because initially, the increase in parasite infection rate with some time delay will reduce the diversity of the species by allowing only those that have some resistance to it to survive. Thus, by only allowing those to prosper, it reduces the genetic diversity. However, at some point, the hypothesized dynamics we expect is that parasites starts to target the most prevalent genotype which until that time thrived from the relative parasite resistance, but due to the co-evolution dynamics are now subject to parasite's specialization and is no longer immune. This effect will significantly reduce the number of dominant genotypes and allow the genetic diversity to increase temporarily, which leads to positive coefficient value for the PIR series. Although there is a slight sign of a surge in values for all three PIR coefficients around the period between 14-Jun 2004 to 19-Jul 2004, the uncertainty of the bounds as well as potential confounding effect from dropped factor series (*Oxygen Density*,*Temperature*) could very well invalidate this. Ideally, we should clearly observe the trajectory of the coefficients to temporarily be positive and back to negative again as explained above.

5. Conclusions & Comments

5.1. Summary

In this project, I attempted to find an evidence for co-evolution dynamics between Daphnia and Parasite population data with other relevant factors. However, due to the irregular sampling frequency of the original data, most of my time has been spent in missing value imputation and several compromises were made throughout this process.

Although, I was able to verify the effectiveness of the missing value imputation procedure in some extent (albeit inferior to the joint imputation) with the toy data, the main data contained too many missing values which prevented me from making any reasonable inference on the hypothesized dynamics. However, at least the framework to do this type of analysis is made. Hence, a better data or imputation methods (specifically the joint imputation) may shed light on some of the questions asked here.

5.2. Challenges

One of the challenges not fundamentally relevant to the result, but nonetheless had to be tackled were the efficiency issues with the Gibbs sampling imputation.

Since we are dealing with imputing time series observations, which could very well have strong correlations, once a sampler imputes one of the observations with a exceedingly high or low value, it is likely that other observations will also get values similar in magnitude. Essentially, this amounts to being conditioned in a region with very low posterior probability due to several dimensions having an extreme value. Thus, it will take some time for the Gibbs sampler to come back to the region with high posterior probability since full conditionals are almost like uniform distribution due to the low posterior probability of the space it is conditioned to.

In the end, thanks to the comment from the TA during the presentation, I managed to partially resolve this issue by trying out various prior specifications. The key was to adjust the prior degree of freedom and standard deviation in a theoretically consistent manner (low degree of freedom and relatively high sd will eventually lead to proposing big innovations etc.).

5.3. Furture Work

The most straightforward way to expand this work is by trying the joint imputation so that we may be able to impute even the factor variables utilizing the conditional normal structure.

In terms of model improvement, since the moment of co-evolution is temporal, and for other cases the parasite infection rate and genome diversity seems to have a negative relationship (if we assume that our smoothed estimate was not significantly confounded by other factors), we may need two processes to model this dynamics. Thus, a Markov switching processes may be a better choice to capture such dynamic systems.

In a broader perspective, the area of evolutionary biology until recently had too much emphasis on the deterministic force such as selection to play a major role in evolution. After the neutral theory of evolution which emphasizes the random and non-directional nature of the evolution was proposed and gradually accepted, the field has moved more towards the direction of stochastic modeling where the selection is considered as a highly situational and rare force which may depend on time (not necessarily the real time but could be generation) and environment. The reason behind choosing Red-Queen Hypothesis (although this hypothesis is quite deterministic rather than stochastic) was to try out the time series methodology in this new direction on the general area of evolutionary biology, and there may be more interesting theory that we can build a model on.

6. Bibliography

[1] L. Van Valen, A new evolutionary law., Evol Theory 1 (1973) 1–30.

[2] P. Turko, T. C., K. E., N. Tardent, Keller, P. B., Spaak, J. Wolinska, Parasites driving host diversity: Incidence of disease correlated with daphnia clonal turnover., Evolution 72.3 (2018) 619–629.

[3] P. Harrison, P. Veerapen, Incorporating and deleting information in dynamic models, in: Developments in Time Series Analysis: in Honour of Maurice B. Priestley, Chapman and Hall London, 1993, pp. 37–49.