Metamodeling for Bias Estimation of Biological Reference Points.

Nick Grunloh

Introduction

- Hello. My name is Nick Grunloh.
- Thanks to everyone in attendance, and thanks to my committee for coming together today on π day 2022 to listen to my advancement to candidacy talk.
- I'll be talking to you today about a Metamodeling approach for assessing estimation bias in population dynamics models, but also some extensions to this work as I finish up my PhD.
- This is work in collaboration with NOAA NMFS, and largely funded by NOAA Sea Grant, and I'd like to thank my collaborators with NOAA.

Data and Basic Modeling Structure

- The modeling context here are single species population dynamics models as they might be used for managing fisheries.
- The simplest of these models (which really captures the essence of the managing objectives) is the surplus-production model.
- **Left Panel:** Data for a typical surplus-production model comes in the form of an index of abundance through time.
- **Right Panel:** The index is often observed alongside a variety of other known quantities, but at a minimum, each observed index will be observed in the presence of some known catch for the period.
- We can't observe all of the fish in the sea, but we can measure indicators of population biomass up to a proportionality constant, q.
- q is the proportionality nuisance parameter (often called catchability) which relates our index of abundance to actual biomass in the population.
- And naturally the nonnegative index of abundance is observed with some uncertainty, which is typically assumed to have lognormal errors.
- Most of the action in these models comes in through a process model on Biomass.
- Biomass is modeled as a nonlinear ODE.
  - the population grows through a (typically non-linear) production function, P(B), and decreases as biomass is removed due to catch, C(t).

  - Production in this setting is defined as the net change in biomass due to basically all reproduction, maturation, and mortality processes other than the recorded fishing from humans.
  - Map the current biomass to some change (growth) in biomass

Schaefer Model

- One classic choice of the production function is the logistic growth curve.
- Choosing P to be Logistic production function in the fisheries setting creates the Schaefer model.
- In ecology logistic growth is very commonly parameterized in terms of the parameters (r, K)
- r controls the maximum reproductive rate of the population in the absence of competition for resources (i.e. the slope at the origin).
- K is the so called "carrying capacity" of the population.
- The quadratic shape of the logistic growth curve encodes density dependence
  - when the population biomass is low there is little competition for resources and growth is maximized (thus r).
  - when the population biomass is high there is competition for resources and growth declines until growth completely stops when the population reaches K.
  - In the absense of fishing K is a stable equilibrium point, above K production is negative to bring the population down, and below K production is positive to bring the populaiton up to K.

Biological Reference points

- Reference points are simplified heuristic measures of population behavior, that are used to make decisions about how to manage the fishery.
- We want to mangage fisheries to allow (and promote) future productivity.
- The key idea is that we want to fish in a way to move the stable equilibrium of the population to a place along this curve that maximizes productivity in the steady state over time.
- Ex) Maximize simple yield at a particular moment V. Maximize sustainable yeild.
- The most common RPs are different ways of noticing that the population is at MSY.
- Any quantity decorated with a star reresent that quantity at MSY.
- Here I focus on the reference points Fmsy (fishing rate to result in MSY) and Bmsy (biomass of the populaiton at MSY) (Or rather Bmsy as a fraction of K aka. Depletion at MSY)

RP Constraints

- Conceptually $\frac{B^*}{B_0}$ and $\frac{F^*}{M}$ is an entire 2D space
- (Mangel et.al., 2013) Canadian Journal of Fisheries and Aquatic Science
- The primary insight of Mangel et.al. that I'd like to focus on is that, if we are not careful, we quickly limit ourselves as we model.
- With a 2 parameter production function, such as BH, if M is fixed this RP space is limited to a 1D curve.
- **Right Panel:** On the right I show an emperical way of noticing this
  - The black are posterior samples of the RPs for a 3 parameter Shepherd-like model fit to cowcod rockfish data
  - The red are posterior samples of the RPs for a BH model with fixed M
  - Notice how the posterior has been squashed into the red curve $\left(\frac{1}{\frac{Fmsy}{M}+2}\right)$
  - Mangel et. al. suggests looking into 3-parameter curves to avoid unintentionally overspecifying our models in the prior.
- **Next:** The Schaefer Model is a two parameter curve that suffers similarly from a constrained RP-Space.
  - model misspecification in this context limits the space of RPs.
  - This is the "Shaefer Line".
  - this over constrains variance of estimates
  - and induces severe bias in estimated RPs.

Pella-Tomlinson Production Model

- Here I formulate a setting for investigating how models in the 2 parameter limited setting might be biased when fit to data from a 3 parameter production function.
- In particular I have a PT Production Model
- The model starts with a nuisance observation layer
- Biomass is driven by the given ODE.
- Production is given by P; That's the PT 3 parameter production function.
- **Right Panel:** On the right you can see the PT production function and how it uses its third γ parameter to lean the production function to the left or right.
  - When $\gamma = 2$; PT=Logistic Prodcution function and the curve is symmetric about $Bmsy$.
- Simulate data under PT and fit those data under the Schaefer Model.
- Every point in RP-space corresponds to a set of parameters of the PT model.

Catch

- To complete the model specification, I assume a synthetic catch series.
- The information content of a given data series in this setting is known to be heavily dependent on catch and "contrast" in the time seeries.

- To control contrast I have catch parameterized in terms of a series of fishing rates relative to $Fmsy$.
- by varying $F/Fmsy$ you can get more or less contrast in the data.
- **Next:**
  - **Left:** By specifying $F/Fmsy = 1$ it represents a low contrast, relatively low information, setting.
  - Below I show the catch in red and the biomass this induces in black.
  - We might call this a "one way trip"
  - **Right:** On the right $F/Fmsy$ is varied to induce a high contrast, relatively high information, setting.
  - This of the right panel as hypothetical stock where fishing rate accelerates as technology and fishing techniques improve rapidly until management practices are applied to bring the stock into equilibrium at MSY.
  - Below again you can see that this population is exposed to a varied catch history, which induce contrast in the generated indices and allows the fitting model to observe a decrease in the population followed by a rebuild of the stock
  - a so called "two-way trip"
- I've looked other catch histories, but we'll focus on these tow extremes

Simulation

- The goal is to investigate bias induced my fitting PT data with the restricted shaefer model.
- For the PT model the relationship between the parameters and the RPs can be inverted so that I can directly write θ in terms of the RPs
- I simulate data off of this Shaefer line and subsequently fit those data with a limited 2 parameter model
- The initial population size is assumed to be at carrying capacity, and then dropped into a catch history.
- Here I show a grid of locations in RP space where I will simulate data and once the shaefer model is fit, that estimate will neccessarily have to fit on that $\frac{1}{2}$ line.
- **Next**
- In particular I'll point out these 4 red X's in the 4 corners.
- These are examples of large model misspecification relative to the scheafer model.
- I show an example biomass series of in the high contrast setting in each corner.
- $Bmsy/B0$: describes where the biomass comes to equilibrium
- $Fmsy$: describes how quickly the stock responds to fishing and how fast it rebuilds.

Metamodel

- Particular model fits are only as helpful as their standard errors allow, but when you observe trends on repeated sampling you can start to gain confidence.
- a squared exponential GP model is used as a flexible, stochastic interpolator of Schaefer model parameter estimates over RP space
- Since estimates are random variables the GP residual variation provides an ideal mechanism for propagating uncertainty from inference in the simulation step into the metamodel.
- While the constrained RP space limits the extend of RP standard errors, accounting for estimate uncertainty has a smoothing effect that focus on the metamodel on the mean behavior.
- While previous studies have considered the factors neccisary to estimate RPs, the limiting constraints of model misspecification have not been explicatly considered.
- This metamodeling approach explicatly highlights the inferencial trade-offs imposed by productivity model misspecification in terms of the bottom line metrics of managing fisheries.

Directionally

- Start with analysis of the low contrast, low information, catch setting.
- Here I'm showing the entire space of PT data and when those data are fit with a Scheafer model how do RPs map onto the Schaeffer line?
- Give example, of data generated at location and mapped onto shaefer line.
- Arrows indicate direction of bias, and color indicates the magnitude of bias.
- Below the Shaefer line we underestimate $Fmsy$ and above the line we overestimate $Fmsy$.

Components

- For all other plots Red indicates over estimation of the modeled quanity and blue indicates underestimates of the modeled quantity.
- **Left** In the bottom right you can see the generalized version of the our F^* story
  - Below the line we underestimate F^*
  - Above the line we overestimate F^*
- **Right** I show the bias in $\frac{B^*}{B_0}$
  - this picture is a law of nature for this simulation setting (estimate must land on the line)

$Fmsy$ Curves

- This slide visualizes the posterior fit of those data in the 4 large model misspecification corners of RP space
- In all cases the red lines represent the posterior fit of the Scheafer model, and the black is the truth of each of quantity.
- I show the production curves of those 4 corner max misspecification fits with each plot in the relative position of each corner.
- When the data are generated with $\frac{B^*}{B_0} > 1/2$ (above the scheaffer line) F^* is over estimated
  - We see that in the slope at the origin being too steep
- When the data are generated with $\frac{B^*}{B_0} > 1/2$ (above the scheaffer line) F^* tends to be under estimated
- Looking at these pictures you can start to understand why
  - When fishing is held at F^* the population simply declines exponentially from K to B^*.
  - The model only observes the right half of the true SRR
  - Due to the leaning of the true PT curves, and the symmetry of the logistic parabola, the logistic curve is learning about its slope at the origin entirely from data where depletion$> \frac{1}{2}$, and above the schaefer line PT is steeper that on the right half than it is on the left, and so we over estimate F^* for data generated above the line.

  - The vice versa phenomena occurs below the schaffer line.
  - Data is only obsevered on the right half of the production function $\Rightarrow$ PT is shallower on the right than on the left $\Rightarrow$ and so the logistic parabola estimate tends to under estimate F^*.

Ratio

- What's more interesting is what is the behavior of bias in the numerator (B^*) and the demoniator (K) not divided but independently
- Whatever the individual patterns are they need to divide back up to give this top middle picture.
- **Left:** On the left I show those quantities, the bias in the quantity B^* is shown *top left*, and the bias in K is shown on the *bottom left*.
- Interestingly, B^* shows large swaths of relatively little bias, and most of the pattern in the top middle panel comes from bias in K.
- The world did not have to be this way, but this says that $Bmsy$ is often a robustly estimated quantity, but due to restrictive model misspecification, its a zero sum game and accuray in B^* often come at the cost of estimates of K.

Contrast

Summary

- The statistician George Box famously said "All model are wrong, but some models are useful"

- My question is how useful are our models?, and if our models are useful. How useful are they? and when are they useful?
- This is a simulation based method for starting to understand that those questions.
- We want to expand these methods to more commonly used production functions
  - BH and Ricker
- and we want to try to get a similar analysis of these assumptiomtions when embeded inside of an age structured or delay difference model.
- but in this simple PT/Shaeffer setting we can see
  - as model misspecification increases biases in some quantities can become very large
  - estimates in B^* are tending to be less sensative to model misspecification than K
  - and F^* bias is going to tend to be strongly catch dependent

Productivity Extension

Growth Extension

Catch Interpolation