UNIVERSITY OF CALIFORNIA
SANTA CRUZ

A METAMODELING APPROACH FOR BIAS ESTIMATION OF BIOLOGICAL REFERENCE POINTS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Nicholas Grunloh

June 2024

The Dissertation of Nicholas Grunloh
is approved:

Professor Herbert Lee, Chair

Edward J. Dick Ph.D.

Professor Paul Parker

Professor Bruno Sanso

Dean Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

A Metamodeling Approach for Bias estimation of Biological Reference Points

by

Nicholas Grunloh

Stock assessments often assume a two-parameter functional form (e.g., Beverton-Holt or Ricker) for the expected recruitment produced by a given level of spawning output. Mangel et al. [14] and others have shown that biological reference points such as $\frac{F^*}{M}$ and $\frac{B^*}{B(0)}$ are largely determined by a single parameter (steepness) when using two-parameter relationships. These functions introduce strong correlations between reference points (RP) that are pre-determined by the functional form, rather than a biological characteristic of the stock. Mangel et al. note that use of a three-parameter stock-recruitment relationship allows for independent estimation of these reference points. This research seeks to understand the nature of biases in reference points resulting from fitting a two-parameter logistic functional form when the true relationship follows a three-parameter stock-recruitment relationship (SRR). This work demonstrates the useful limits of the misspecified Schaefer model, and the mechanisms of model failure which arise from mapping a three-dimensional parameter space into two dimensions.

To myself,

Perry H. Disdainful,

the only person worthy of my company.

Acknowledgments

I want to “thank” my committee, without whose ridiculous demands, I would have graduated so, so, very much faster.

Chapter 1

Introduction

Generalize beyond production model

The most fundamental model in modern fisheries management is the surplus-production model. These models focus on modeling population growth via nonlinear parametric ordinary differential equations (ODE). Key management quantities called reference points (RPs) are commonly derived from the ODE equilibrium equations and depend upon the parameterization of biomass production. Two-parameter forms of the production function have been shown to limit the theoretical domain of RPs [14]. The limited RP-space of two-parameter models makes these models vulnerable to model misspecification with respect to RPs, and thus the limiting stucture of two-parameter models may in and of itself induce bias in RP estimation using these models. The behavior of RP estimation is not well understood and as a result patterns of bias in RP estimation may easily go unnoticed. A metamodeling approach is developed here to describe RP biases and explore mechanisms of model failure under the most common two-parameter models.

Data for a typical surplus-production model comes in the form of an index of abundance through time which is assumed to be proportional to the reproducing biomass for the modelled population that is vulnerable to fishing. The index is often observed alongside a variety of other known quantities, but at a minimum, each index will be observed in the presence of some known catch for the period. Figure (1.1) shows the classic Namibian Hake dataset [15, 11, 13] exemplifying the form.

Indices are assumed to have multiplicative log-normal errors, and thus the following observation model arises naturally,

$$I_t = qB_t e^{\varepsilon} \quad \varepsilon \sim N(0, \sigma^2). \tag{1.1}$$

Above q is often referred to as the "catchability parameter"; it serves as the proportionality constant mapping between the observed index of abundance and biomass. σ^2 models residual variation. Biologically speaking q and σ^2 are often treated as nuisance parameters with the "biological parameters" entering the model through a process model on biomass.

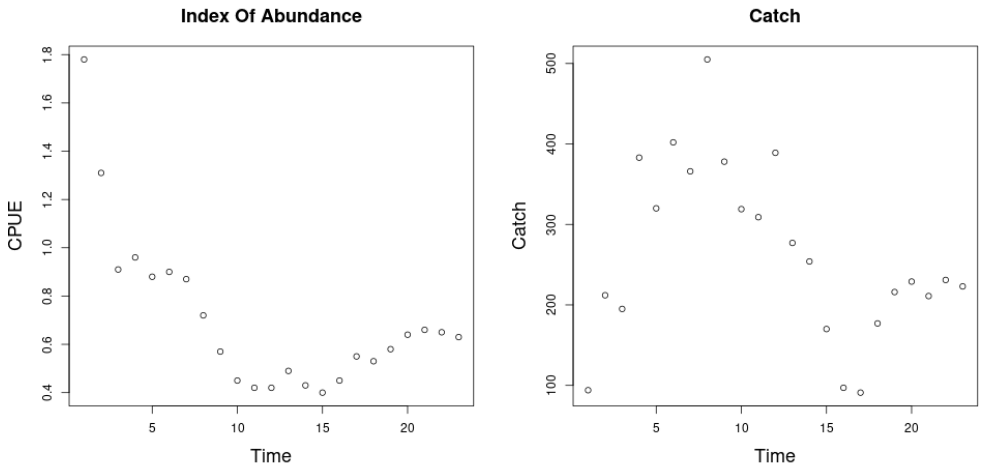

Figure 1.1: *left*: An index of abundance data, catch per unit effort (CPUE), for Namibian Hake from 1965 to 1987. *right*: The associated catch data for Namibian Hake over the same time period.

Biomass is assumed to evolve as an ODE; in this case I focus on the following form

$$\frac{dB}{dt} = P(B(t);\boldsymbol{\theta}) - Z(t)B(t). \tag{1.2}$$

Here biomass is assumed to change in time by two processes, net production of biomass into the population, $P(B)$, and various sources of biomass removal, Z, from the population.

Firstly, the population grows through a production function, $P(B)$. Production in this setting is defined as the net biomass increase due to all reproduction and maturation processes. The production function is assumed to be a parametric (generally non-linear) function relating the current biomass of the population to an aggregate production of biomass.

Secondly, the population decreases as biomass is removed by various sources that are assumed to remove biomass linearly with biomass. Above, $Z(t)$, is an aggregate rate of removal. When the fishing rate, $F(t)$, is the only source of removal $Z(t) = F(t)$, however some models will also included other linear terms in $Z(t)$. Commonly the rate of "natural mortality", M, is also included as an additional term so that $Z(t) = M + F(t)$.

From a management perspective a major goal of modeling is to accurately infer a

quantity known as *maximum sustainable yield* (MSY). One could maximize simple yield at a particular moment in time (and only for that moment) by fishing all available biomass in that moment. This strategy is penny-wise but pound-foolish (not to mention ecologically devastating) since it doesn't leave biomass in the population to reproduce in the future. We seek to fish in a way that allows (or even encourages) future productivity in the population. This is accomplished by maximizing the equilibrium level of catch over time. Equilibrium yield is considered by replacing the steady state biomass ($\bar{B}$) in the assumed form for catch, so that $\bar{Y} = F\bar{B}(F)$, where $\bar{}$ indicates a value at steady state. MSY is found by maximizing $\bar{Y}(F)$ with respect to F, and F^* is the fishing rate at MSY. Going forward let * decorate any value derived under the condition of MSY.

Fisheries are very often managed based upon reference points which serve as simplified heuristic measures of population behavior. The mathematical form of RPs depends upon the model assumptions through the production function. While a number of different RPs exist which describe the population in different (but related) ways, the most common RPs revolve around the concept of MSY (or robust ways of measuring MSY [10, 16]). Here the focus is primarily on the RPs $\frac{B^*}{\bar{B}(0)}$ and F^* ($\frac{F^*}{M}$ when appropriate) for their pervasive use in modern fisheries [17].

F^* is the afore mentioned fishing rate which results in MSY. $\frac{B^*}{\bar{B}(0)}$ is the depletion of the stock at MSY. That is to say $\frac{B^*}{\bar{B}(0)}$ describes the fraction of the unfished population biomass that will remain in the equilibrium at MSY. In general $F^* \in \mathbb{R}^+$ and $\frac{B^*}{\bar{B}(0)} \in (0,1)$, however under the assumption of two-parameter production, models will be structurally unable to capture the full theoretical range of RPs.

Many of the most commonly used production functions depend only on two-parameters. For example, the Schaefer model depends only on the biological parameters r and K, and limits RP inference so that under the Schaefer model $\left(F^*, \frac{B^*}{\bar{B}(0)}\right) \in \left(\mathbb{R}^+, \frac{1}{2}\right)$. The two-parameter Fox model [6] limits $\left(F^*, \frac{B^*}{\bar{B}(0)}\right) \in \left(\mathbb{R}^+, \frac{1}{e}\right)$. Similarly the two-parameter Cushing [4], Beverton-Holt [1, BH] and Ricker [20] production functions do not model the full theoretical space of RPs [14, 26].

The bias-variance trade-off [18] makes it clear that the addition of a third parameter in

the production function will necessarily reduce estimation bias. However the utility of this bias reduction is still under debate because the particular mechanisms and behavior (direction and magnitude) of these biases for key management quantities are not fully understood or described. Lee et al. [12] provides some evidence that estimation of productivity parameters are dependent on changes in biomass trend through time (i.e. contrast) as well as model specification. Conn et al. [2] comes to similar conclusions via calibration modeling techniques. These studies indicate important factors that contribute to inferential failure. However they do not offer mechanisms of model failure, nor do their experimental designs allow for the control of different types of model misspecification.

In this study I consider the behavior of inference when index data are simulated from three-parameter PT and Schnute production models, but the simulated data are fit using intentionally misspecified two-parameter logistic or BH production models. The work begins with a derivation of RPs under the three-parameter models. A method is then presented for generating simulation designs based on the parametric form of RPs which serves as a control on the nature of simulated model misspecification. Finally a Gaussian Process (GP) metamodel [7] is constructed for exploration and analysis of RP biases.

A key insight of this approach is that bias is considered broadly across RP-space to uncover patterns and correlations between RPs. The GP metamodel is explicit about trade-offs between RPs so as to inform the full utility of reducing bias, as well as to suggest mechanisms for understanding what causes bias. Further, the effect of contrast on estimation is considered together with model misspecification.

Chapter 2

Pella-Tomlinson Model

2.1 Introduction

2.2 Methods

2.2.1 Model

The three-parameter Pella-Tomlinson (PT) family has a convenient form that includes, among others [6, 19], the logistic production function as a special case. PT production function is parameterized so that $\boldsymbol{\theta} = [r, K, \gamma]$ and the family takes the following form,

$$P_p(B; [r, K, \gamma]) = \frac{rB}{\gamma - 1}\left(1 - \left(\frac{B}{K}\right)^{(\gamma-1)}\right). \tag{2.1}$$

γ is a parameter which breaks PT out of the restrictive symmetry of the logistic curve. In general $\gamma \in (1, \infty)$, with the logistic model appearing in the special case of $\gamma = 2$, and the Fox model appearing as a limiting case as $\gamma \to 1$. The parameter r controls the maximum per-capita growth rate of the population in the absence of competition for resources (i.e. the slope of production function at the origin). K is the so called "carrying capacity" of the population. In this context the carrying capacity can be formally stated as steady state biomass in the absence of fishing (i.e. $\bar{B}(0) = K$). In Figure (2.1) PT production is shown for a range of parameter values so as to demonstrate the various productivity shapes that can be achieved under PT.

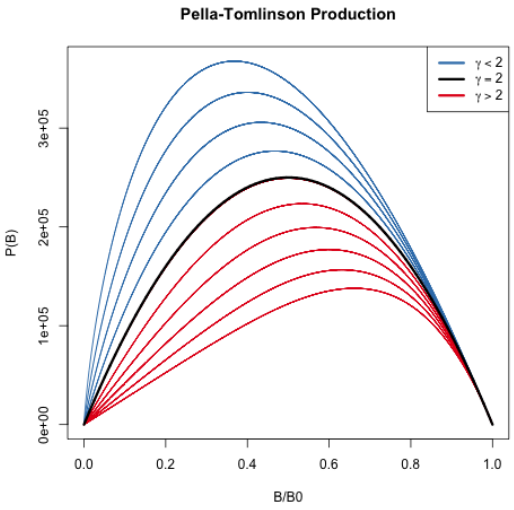

Figure 2.1: The Pella-Tomlinson production function plotted across a variety of parameter values. The special cases of Logistic production is shown in black, and the left-leaning and right-leaning regimes are shown in blue and red respectively.

While the form of the PT curve produces some limitations [5], importantly the introduction of a third parameter allows enough flexibility to fully describe the space of reference

points used in management. To see this, the reference points are analytically derived for the PT model below.

2.2.2 Reference Points

With $B(t)$ representing biomass at time t, under PT production, the dynamics of biomass are defined by the following ODE,

$$\frac{dB}{dt} = \frac{rB}{\gamma - 1}\left(1 - \left(\frac{B}{K}\right)^{\gamma - 1}\right) - FB. \tag{2.2}$$

An expression for the equilibrium biomass is attained by setting Eq (2.2) equal to zero, and rearranging the resulting equation to solve for B. Thinking of the result as a function of F gives,

$$\bar{B}(F) = K\left(1 - \frac{F(\gamma - 1)}{r}\right)^{\frac{1}{(\gamma - 1)}}. \tag{2.3}$$

At this point it is convenient to notice that $\bar{B}(0) = K$. The expression for B^* is given by evaluating Eq (2.3) at F^*. To get an expression for F^*, the equilibrium yield is maximized with respect to F,

$$F^* = \underset{F}{\operatorname{argmax}}\, F\bar{B}(F). \tag{2.4}$$

In the case of PT production this maximization can be done analytically, by differentiating the equilibrium yield with respect to F as follows,

$$\frac{d\bar{Y}}{dF} = \bar{B}(F) + F\frac{d\bar{B}}{dF} \tag{2.5}$$

$$\frac{d\bar{B}}{dF} = -\frac{K}{r}\left(1 - \frac{F(\gamma - 1)}{r}\right)^{\frac{1}{\gamma - 1} - 1}. \tag{2.6}$$

Setting Eq (2.5) equal to 0, substituting $\bar{B}(F)$ and $\frac{d\bar{B}}{dF}$ by Equations (2.3) and (2.6) respectively, and solving for F produces the following expression for the fishing rate required

to produce MSY,

$$F^* = \frac{r}{\gamma} \tag{2.7}$$

Plugging the above expression for F^* back into Eq (2.3) gives the following expression for biomass at MSY,

$$B^* = K\left(\frac{1}{\gamma}\right)^{\frac{1}{\gamma-1}}. \tag{2.8}$$

The above derived expressions for $\bar{B}(0)$, B^*, and F^* can then be used to build a specific analytical form for the biological reference points in terms of only productivity parameters.

$$F^* = \frac{r}{\gamma} \qquad\qquad \frac{B^*}{\bar{B}(0)} = \left(\frac{1}{\gamma}\right)^{\frac{1}{\gamma-1}} \tag{2.9}$$

2.2.3 Simulation

Generating simulated indices of abundance from the PT model requires inverting the relationship between $\left(F^*, \frac{B^*}{\bar{B}(0)}\right)$, and (r,γ). It is not generally possible to analytically invert this relationship for many three-parameter production functions [17, 21]. Most three-parameter production functions lead to RPs that require expensive numerical methods to invert; more over the numerical inversion procedure can often be unstable. That said, for the case of PT this relationship is analytically invertible, and leads to the following relationship

$$r = \gamma F^* \qquad\qquad \gamma = \frac{W\left(\frac{B^*}{\bar{B}(0)}\log\left(\frac{B^*}{\bar{B}(0)}\right)\right)}{\log\left(\frac{B^*}{\bar{B}(0)}\right)}. \tag{2.10}$$

Above W is the Lambert product logarithm function. More details about this derivation, and the Lambert product logarithm, are given in Appendix (A).

Using Eq. (2.10) to obtain production parameters, a PT production model can be fully defined for any combination of the RPs F^* and $\frac{B^*}{\bar{B}(0)}$. Since K does not enter the RP calculation its value is fixed arbitrarily at 10000.

Indices of abundance are simulated from the three-parameter PT production model broadly over the space of F^* and $\frac{B^*}{\bar{B}(0)}$ via a space filling design as described in Section (3.2.3). A small amount of residual variation, $\sigma = 0.01$, is added to the simulated index, and these data are then fit with a Schaefer model, at various degrees of misspecification, so as to observe the effect of productivity model misspecification upon RP inference.

2.2.4 Design

Letting $\mathcal{F}$ and $\mathcal{B}$ be regular grids, of size $n = 100$, on $F^* \in (0.1, 0.7)$ and $\frac{B^*}{B_0} \in (0.2, 0.6)$ respectively, a LHS design of size 100 is collected among the cells produced by $\mathcal{F} \times \mathcal{B}$.

Each of the sampled LHS design locations represent a unique PT model with the sampled RP values. Since the relationship mapping RPs analytically to productivity parameters can be found for the PT model, LHS designs the the PT model are computed directly in RP space and Eq. (2.10) is used to map the sampled RP design locations to PT productivity parameters.

2.2.5 Gaussian Process Metamodel

At its core, a metamodel is simply a model of some mapping of inputs to outputs (the mapping itself is typically defined by a computer model). By modeling the mapping with a statistical model (that explicitly defines the relevant features of the mapping) a metamodel defines a specific ontology for the mapping. By simulating examples of the mapping, the inferential infrastructure of the statistical model is used to empirically learn an effective emulation of the mapping within the ontology defined by the statistical model. The predictive infrastructure of the statistical model is then useful as an approximate abstraction of the system itself to better understand the system through further data collection, cheap approximation of the mapping, and/or study of the mapping itself.

In this setting, the aim of metamodeling is to study how well RPs are inferred when typical two-parameter models of productivity (Logistic and BH) are misspecified for populations that are actually driven by more complicated dynamics. The simulation design, $\boldsymbol{X}$, provides a sample of different population dynamics that are driven by three-parameter production

functions broadly in RP space. By simulating index of abundance data from the three parameter model, and fitting those data with the two-parameter production model, we observe particular instances of how well RPs are inferred at the given misspecification of the two-parameter model relative to the true three-parameter production model. By gathering all of the simulated instances of how RPs are inferred (under the two-parameter model), we form a set of example mappings to train a metamodel which represents the mapping of true RPs (under the three-parameter model) to estimates of RPs under the misspecified two-parameter production model. The metamodel is essentially a surrogate for inference under the misspecified two-parameter production model that controls for the specific degree of model misspecification.

A flexible GP model is assumed for the structure of the metamodel to describe the mapping of RPs under misspecified two-parameter models of productivity. A GP is a stochastic process generalizing the multivariate normal distribution to an infinite dimensional analog. GP models are often specified primarily through the choice of a covariance (or correlation) function which defines the relationship between locations in the input space. Typically correlation functions are specified so that points closely related in space result in correlated effects in the model. In this setting the inputs to the GP metamodel are the space of reference points which define the simulated three-parameter production models.

While index of abundance data are generated from three-parameter models, at each design location of the simulation, fitting the restricted two-parameter model results in a maximum likelihood estimate (MLE; and associated estimation uncertainty) of each of the productivity parameters (i.e. Schaefer:[$log(r)$, $log(K)$], BH:[$log(\alpha)$, $log(\beta)$]). To simplify the specification of the metamodel, let $\mathbf{y}$ be a vector collecting the fitted MLEs for one of the productivity parameters, and let $\boldsymbol{\omega}$ be a vector of estimates of the estimator variances (via the inverted Fisher information) at each $\mathbf{y}$. Each of the fitted productivity parameter estimates are then modeled

using independent instances of the following GP metamodel.

$$\begin{aligned} \mathbf{y} &= \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v} + \boldsymbol{\epsilon} \\ \boldsymbol{v} &\sim N_n(\mathbf{0}, \tau^2 \boldsymbol{R}_\ell) \\ \boldsymbol{\epsilon} &\sim N_n(\mathbf{0}, \boldsymbol{\omega}' \boldsymbol{I}) \end{aligned} \tag{2.11}$$

$\boldsymbol{X}$ is the n x 2 LHS design matrix of RPs for each simulated three-parameter data generating model as described in Section (3.2.4.1). ε models independent normally distributed error, which provides an ideal mechanism for propagating uncertainty from inference in the simulation step into the metamodel. By matching each y_i with an observed ω_i variance term, ε serves to down weight the influence of each y_i in proportion to the inferred production model sampling distribution uncertainty. This has the effect of smoothing the GP model in a way similar to the nugget effect [8], although the application here models this effect heterogeneously.

The term, $\boldsymbol{v}$, contains spatially correlated GP effects. The correlation matrix, $\boldsymbol{R}_\ell$ describes how RPs close together in the simulation design are more correlated than those that are far away. This spatial effect is modeled with a squared exponential correlation function,

$$R(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \exp\left(\sum_{i=1}^{2} \frac{-(x_i - \tilde{x}_i)^2}{2\ell_j^2} \right). \tag{2.12}$$

R has an anisotropic separable form which allows for differing length scales, ℓ_1 and ℓ_2, in the different RP axes. The flexibility to model correlations separately in the different RP axes is key due to the differences in the extent of the RP domains marginally. The metamodel parameters β_0, $\boldsymbol{\beta}$, τ^2, ℓ_1 and ℓ_2 are fit via MLE against the observations $\mathbf{y}$, $\boldsymbol{X}$, and $\boldsymbol{\omega}$ from simulation fits.

Fitting the metamodel allows for a full predictive description of inference under the misspecified restricted models. Predictive estimates are obtained via kriging [3]

$$\hat{y}(\mathbf{x}) = \beta_0 + \mathbf{x}\boldsymbol{\beta} + \mathbf{r}(\mathbf{x})' \boldsymbol{R}_{\boldsymbol{\ell}}^{-1} \left(\mathbf{y} - \left(\beta_0 + \boldsymbol{X}\boldsymbol{\beta} \right) \right) \tag{2.13}$$

$\hat{y}(\mathbf{x})$ is the predicted value of the modeled productivity parameter MLE under the two-parameter production model, when the index of abundance is generated from the three-parameter production model at RP location $\mathbf{x}$. $\mathbf{r(x)}$ is a vector-valued function of correlation function evaluations for the predictive location $\mathbf{x}$ against all observations in $\boldsymbol{X}$ (i.e. $\mathbf{r(x)} = \boldsymbol{R}(\mathbf{x}, \boldsymbol{x}_i) \; \forall \; \boldsymbol{x}_i \in \boldsymbol{X}$).

While metamodeling occurs on the inferred productivity parameters of the restricted production model, the metamodel can also be used to build estimates of major biological RPs. For the BH model the relevant transformations for relating productivity parameters with RPs are given in Eqs. (3.5, 3.8) with γ fixed to -1; for the Schaefer model $\hat{B}^* = \frac{\hat{K}}{2}$ and $\hat{F}^* = \frac{\hat{r}}{2}$. Applying the metamodel predictive surfaces on the scale of RP estimates allows for the quantification of estimation bias that is induced by fitting a misspecified two-parameter production model to indices of abundance generated under three-parameter productivity.

2.2.6 Catch

It is known that contrast in the observed index and catch time series can effect inference on the productivity parameters [9]. In this setting contrast refers to changes in the long term trends of index data. Figure (2.2, *right*) demonstrates an example of biomass that includes contrast induced by catch. It is not well understood how contrast may factor into inferential failure induced by model misspecification. Thus catch is parameterized so as to allow for a spectrum of possible contrast simulation settings.

Catch is parameterized so that $F(t)$ can be controlled with respect to F^*. Recall that catch is assumed to be proportional to biomass, so that $C(t) = F(t)B(t)$. To control $F(t)$ with respect to F^*, $C(t)$ is specified by defining the quantity $\frac{F(t)}{F^*}$ as the relative fishing rate. $B(t)$ is defined by the solution of the ODE, and F^* is defined by the biological parameters of the model. By defining $\frac{F(t)}{F^*}$, catch can then be written as $C(t) = F^* \left(\frac{F(t)}{F^*}\right) B(t)$.

Intuitively $\frac{F(t)}{F^*}$ describes the fraction of F^* that $F(t)$ is specified to for the current $B(t)$. When $\frac{F(t)}{F^*} = 1$, $F(t)$ will be held at F^*, and the solution of the ODE brings $B(t)$ into equilibrium at B^*. When $\frac{F(t)}{F^*}$ is held constant in time biomass comes to equilibrium as an exponential decay from K approaching B^*. When $\frac{F(t)}{F^*} < 1$, $F(t)$ is lower than F^* and $B(t)$ is

pushed toward $\bar{B} > B^*$. Contrarily, when $\frac{F(t)}{F^*} > 1$, $F(t)$ is higher than F^* and $B(t)$ is pushed toward $\bar{B} < B^*$; the precise values of $\bar{B}$ can be calculated from the steady state biomass equations provided above and depend upon the specific form of the production function.

For the simulations presented here, a family of fishing behaviors are considered where the fishing rate accelerates as technology and fishing techniques improve rapidly until management practices are applied, which ultimately brings fishing into equilibrium at F^*. This is parameterized as three distinct phases, over a total of 45 units of time, with each phase lasting 15 time units. The specific form is given below.

$$\frac{F(t)}{F^*} = ae^{bt}\mathbf{1}_{0\leq t<15} + (d-ct)\mathbf{1}_{15\leq t<30} + \mathbf{1}_{30\leq t\leq 45} \tag{2.14}$$

The first term of Eq (2.14) is an exponential increase in fishing, the second term is a linear decline in relative fishing as initial management practices are applied, and the third term, $\mathbf{1}_{30\leq t\leq 45}$, simply holds the fishing rate at F^* there after. These three phases are controlled by the four parameters a, b, c, and d. By enforcing that the interface of the phases meet at χ_{max} and 1 respectively the relative fishing series is reduced to a two-parameter family.

$$a = e^{\log(\chi_{max})-15b} \qquad b = \frac{1}{t-15}\log\left(\frac{\chi_{min}}{\chi_{max}}\right) \tag{2.15}$$

$$c = \frac{\chi_{max}-1}{15-1} \qquad d = 15c+\chi_{max} \tag{2.16}$$

By further specifying $\chi_{max} = 1.6^{\chi}$ and $\chi_{min} = 0.4^{\chi}$ the two-parameters χ_{max}, and χ_{min} can be reduced to the single parameter χ. The tuning parameter χ then singularly controls contrast that appears in time series data.

When $\chi = 0$, the relative fishing rate is a constant at 1 to create a low contrast simulation environment. As χ increases Eq (2.14) induces more and more contrast in the observed index and catch time series until $\chi = 1$ which produces a high contrast simulation environment. Figure (2.2) demonstrates a spectrum of contrast simulation environments as well as the time series data they induce in the solution of the production model ODE.

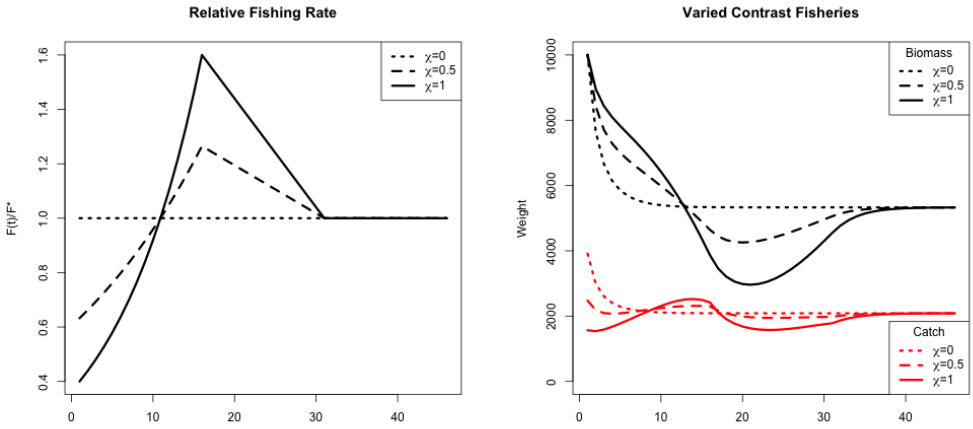

Figure 2.2: (*left*) Relative fishing with low, medium, and high contrast. (*right*) Population biomass and catch at each associated level of contrast.

2.2.7 Two-Parameter Production Model Inference

The simulated mapping results from fitting an intentionally misspecified two parameter production model to index of abundance data that are generated from a more complex three-parameter model of productivity. Thus, let I_t be an index of abundance simulated from the three-parameter PT or Schnute production models at time $t \in \{1,2,3,...,T\}$. However the fitted model is specified to be intentionally misspecified so that the fitted model is driven by a two-parameter Schaefer, or BH production model respectively.

The observation model for the fitted model is log-normal such that,

$$I_t|q,\sigma^2,\boldsymbol{\theta} \sim LN(qB_t(\boldsymbol{\theta}),\sigma^2). \tag{2.17}$$

$B_t(\boldsymbol{\theta})$ is defined by the solution of the ODEs defined by the Schaefer, or BH models. For the Schaefer model $\boldsymbol{\theta} = [r,K]$, and for the BH model $\boldsymbol{\theta} = [\alpha,\beta]$. From the perspective of the fitted model, the observed I_t are assumed independent conditional on q, σ^2, r, K and the two-parameter ODE model for biomass. Thus the log likelihood can be written as

$$\log \mathcal{L}(q,\sigma^2,\boldsymbol{\theta};I) = -\frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_t \log\left(\frac{I_t}{qB_t(\boldsymbol{\theta})}\right)^2. \tag{2.18}$$

In this setting, q is fixed at 0.0005 and M is fixed at 0.2, to focus on the inferential

effects of model misspecification on biological parameters. σ^2 and θ are reparameterized to the log scale and fit via MLE. Reparameterizing the parameters to the log scale improves the reliability of optimization, in addition to facilitating the use of Hessian information for estimating MLE standard errors.

Given that the biological parameters enter the likelihood via a nonlinear ODE, and further the parameters themselves are related to each other nonlinearly, the likelihood function can often be difficult to optimize. A hybrid optimization scheme is used to maximize the log likelihood to ensure that a global MLE solution is found. The R package GA [22, 23] is used to run a genetic algorithm to explore parameter space globally. Optimization periodically jumps into the L-BFGS-B local optimizer to refine optima within a local mode. The scheme functions by searching globally, with the genetic algorithm, across many initial values for starting the local gradient-based optimizer. The genetic algorithm serves to iteratively improve hot starts for the local gradient-based optimizer. Additionally, optimization is only considered to be converged when the optimum results in an invertible Hessian at the found MLE.

2.2.8 Continuous model formulation

An important (and often overlooked) implementation detail is the solution to the ODE which defines the progression of biomass through time. As a statistical model it is of paramount importance that this ODE not only have a solution, but also that the solution be unique. Of primary concern, uniqueness of the ODE solution is necessary for well conditioned inference.

If the form of $\frac{dB}{dt}$ is at least Lipschitz continuous, then the Cauchy-Lipschitz-Picard theorem provides local existence and uniqueness of $B(t)$. Recall from Eq(1.2) that $\frac{dB}{dt}$ is separated into a term for biomass production, $P(B)$, and a term for removals, $Z(t)B(t)$. For determining Lipschitz continuity of $\frac{dB}{dt}$, the smallest Lipschitz constant of $\frac{dB}{dt}$ will be the sum of the constants for each of the terms $P(B)$ and $Z(t)B(t)$ separately. Typically any choice of $P(B)$ will be continuously differentiable, which implies Lipschitz continuity. At a minimum $Z(t)$ typically contains fishing mortality as a function of time $F(t)$ to model catch in time as $C(t) = F(t)B(t)$. $Z(t)$ may or may not contain M, but typically M is modeled as stationary in time and does not pose a continuity issue, unlike some potential assumptions for $C(t)$.

In practice $C(t)$ is determined by a series of observed, assumed known, catches. Catch observations are typically observed on a quarterly basis, but in practice may not be complete for every quarter (or year) of the modeled period. It is overwhelmingly common to discretize the ODE in time via Euler's method with integration step sizes to match the observation frequency of the modeled data. This is often computationally convenient when the underlying species dynamics are resonably well behaved, however when the dynamics model is used as a statistical model, with the goal of inferring the behavior of the underlying species dynamics, the regularity of the dynamics are not guaranteed. An implicit assumption of continuity of catch in time provides the neccessary regularity for the statistical model. Furthermore a continuous handling of the dynamics provides improved accruacy in evaluating the ODE, particually when inferring productivity parameters which largely control the regularity of the dynamics.

While there are many ways to handle catch continuity, here I assume that catches accrue linearly between observed catches. This assumption defines the catch function as a piecewise linear function of time, with the smallest Lipschitz constant for the catch term defined by the steepest time segment of the catch function. This assumption represents one of the simplest ways of handling catch, while retaining Lipschitz continuity overall. Furthermore linearly interpolated catch is adequately parsimonious for the typical handling of catches.

2.2.8.1 Integration and Stiffness

As previously mentioned, the overwhelming majority of implementations of stock assessment models discretized the ODE using Euler's method with the integration step sized fixed so as to match the observation frequency. In this setting we explore model parameterizations that explore the full extent of biologically relevant reference points. This exercise produces some combinations of parameters that result in numerically stiff ODEs.

The concept of stiffness in ODEs is hard to precisely characterize. Hairer and Wanner [25, p.2] describe stiffness in the following pragmatic sense, "Stiff equations are problems for which explicit methods don't work". It is hard to make this definition more mathematically precise, but this a consistent issue for models of very productive species in the low contrast simulation. Euler's method, as often implemented, is particularly poorly suited for these stiff

regions of parameter space. In these stiff regions it is necessary to integrate the ODE with an implicit integration method.

Several of the most common implicit methods were tried including the Livermore Solver for ODEs (lsode), and the Variable Coefficient ODE Solver (vode) as implemented in the deSolve package of R [24]. The difference between implicit solvers is negligible, while explicit methods result in wildly varying solutions to the ODE in stiff regions of parameter space. Results shown here are computed using the lsode integration since it runs relatively quickly and has a relatively smaller footprint in system memory.

2.3 Results

2.3.1 An *MSY*-Optimal Catch History

When $F(t)$ is held constant at F^*, as it is in the "low contrast" simulation setting, $B(t)$ comes to equilibrium as an exponential decay from K to B^*. Understanding model misspecification bias is simplified in this setting due to the relative simplicity that this induces in $B(t)$. However this simplicity is known to poorly inform estimates of r, and thus F^*, due to the limited range of the production function that is observed [9].

Figure (2.3) shows four of the most misspecified example production function fits as compared to the true data generating PT production functions. The rug plots below each set of curves show how the observed biomasses decay exponentially from K to B^* in each case. In particular, notice how observations only exist where the PT biomass is greater than B^*. Due to the leaning of the true PT curves, and the symmetry of the logistic parabola, the logistic curve only observes information about its slope at the origin from data observed on the right portion of the PT curves. The top two panels of Figure (2.3) shows PT data generated such that $\frac{B^*}{\bar{B}(0)} > 0.5$; in these cases PT is steeper to the right of B^* than it is on the left, and so the the logistic curve over-estimates r, and consequently also over-estimates F^*. The bottom two panels of Figure (2.3) show PT data generated with $\frac{B^*}{\bar{B}(0)} < 0.5$ and where the vice versa phenomena occurs. PT is shallower to the right of B^* than it is on the left and so the logistic parabola estimate tends to under estimate F^*.

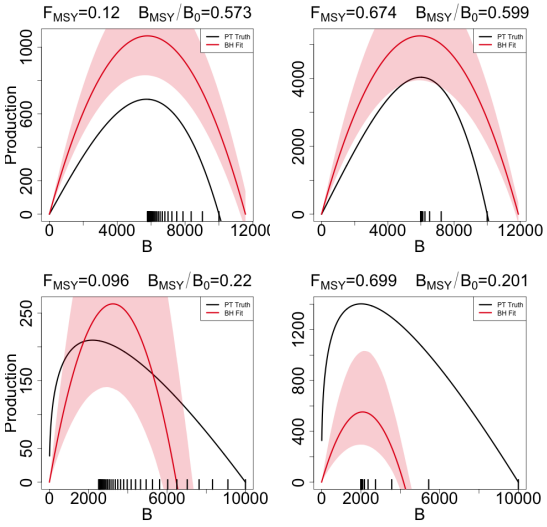

Figure 2.3: A comparison of the true PT production function (in black) and the estimated logistic curve (in red) with 95% CI shown. The examples shown represent the four corners of maximum model misspecification in the simulated RP-space. Observed biomasses are plotted in the rug plots below the curves.

2.3.1.1 Metamodeled Trends

Each point in the space of the RPs F^* and $\frac{B^*}{B(0)}$ uniquely identifies a complete PT model with different combinations of parameters values. Recall that when $\gamma = 2$ for the PT model, the PT curve becomes a parabola and is equivalent to the logistic curve of the Schaefer model. Since the logistic curve is symmetric about B^*, the Schaefer model must fix the value of $\frac{B^*}{B(0)}$ at the constant 0.5 for any value of F^*. So the line through RP space defined by $\frac{B^*}{B(0)} = 0.5 \ \forall \ F^*$, defines the subset of RP space where $\gamma = 2$ and where the PT model is equivalent to the Schaefer model. For brevity this subset of RP where $\frac{B^*}{B(0)} = 0.5$ will be referred to as the "Schaefer set". Thus simulated data that are generated along the Schaefer set will be the only data that are not misspecified relative to the Schaefer model; as PT data are simulated farther and farther away from this line at $\frac{B^*}{B(0)} = 0.5$ model misspecification of the Schaefer model becomes worse and worse.

While Figure (2.3) demonstrates a real trend in simulation results, individual simulation runs will at best show jittery trends due to the stochastic nature of statistical inference. The GP process metamodel accounts for this stochasticity to focus analysis on the signal in the simulation results. Recall that metamodeling occurs on the scale of the inferred productivity parameters of the restricted production model, by transforming metamodel predictions via Eq. (2.9), metamodeled predictions are obtained for Schaefer RPs. By further subtracting the true data generating PT RPs from the predicted Schaefer RPs at each point in RP space a pattern of inferential RP bias, induced by model misspecification of the Schaefer model, can be seen.

Figure (2.4) shows the pattern of biases the Schaefer model creates when fit to PT data generated at each point of RP space. An equivalent way to think of Figure (2.4) is that since the Schaefer model must estimate RPs in the Schaefer set, the metamodel arrows indicate the mapping that is created by inferring RPs under a misspecified Schaefer model fit to PT data generated at each point over the pictured region.

Since $\frac{B^*}{B_0}$ must be 0.5 under the Schaefer model, biases in the $\frac{B^*}{B_0}$ direction must simply map vertically onto the Schaefer set. Due to this simplified RP geometry under the Schaefer model, the degree of bias in $\frac{B^*}{B_0}$ estimation is defined solely by the degree of model misspeci-

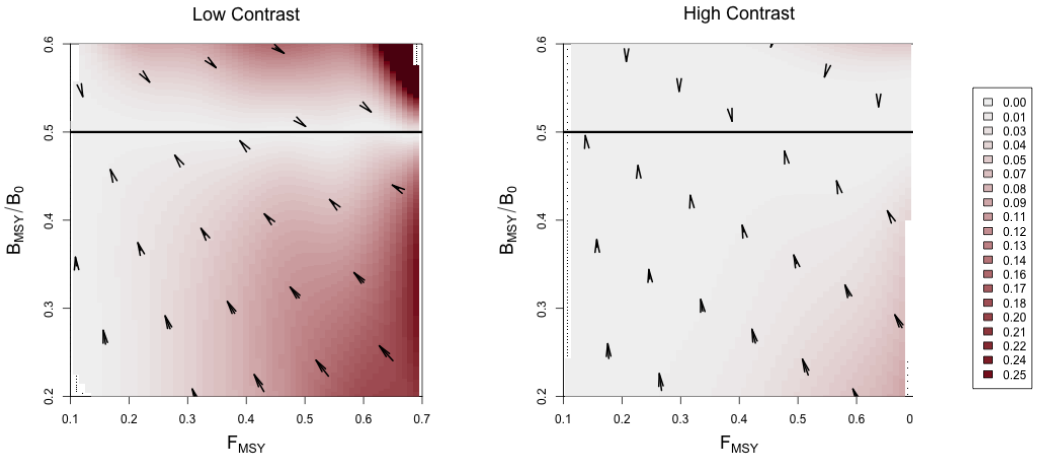

Figure 2.4: Joint bias direction for $(F^*, \frac{B^*}{B_0})$ estimates under the misspecified Schaefer Model. The intensity of color represents the excess bias relative to the shortest possible mapping. Results in the low contrast setting are shown $left$, and the high contrast setting is shown $right$.

fication irrespective of F^*. Furthermore, the closest possible point along the Schaefer set that Schaefer model inference could map RPs would be the perfectly vertical mapping. This pattern only contains the strictly necessary bias present in $\frac{B^*}{B_0}$, and zero bias in F^*. Any deviation from this minimal bias pattern is necessarily due to added bias in F^*.

The two simulation settings shown in Figure (2.4) are identical except for the amount of contrast present in the simulated index. The left panel of Figure (2.4) shows RP biases in the low contrast setting, while the right panel shows the high contrast setting. Notice that in the low contrast setting the RP bias pattern is far from the minimum distance mapping, however when contrast is added the mapping becomes much closer to a minimal vertical bias mapping. In the low contrast setting the observed bias is consistent with the pattern and mechanism described in Figure (2.3), where F^* is underestimated for data generated below the Schaefer line and overestimated above the Schaefer set. In the high contrast simulation the mapping is nearly minimal distance with the exception of PT data generated with simultaneously low $\frac{B^*}{B_0}$ and high F^*.

Figure (2.5) demonstrates how bias in F^* estimation decreases as contrast is added to PT data as generated in the low $\frac{B^*}{B_0}$ and high F^* regime. By including additional contrast F^*

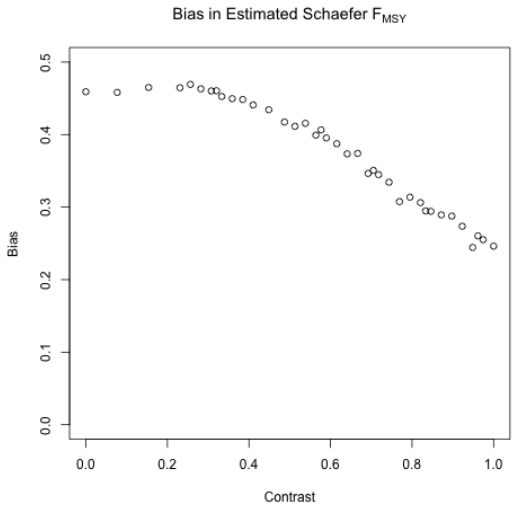

Figure 2.5: Bias in F^* under the Schaefer model when PT data are generated with increasing contrast so that F^* and $\frac{B^*}{B_0}$ are fixed at 0.699 and 0.201 respectively.

bias is decreased, however parameterizing contrast so as to fully extinguish F^* bias may require a more complex model of fishing.

2.4 Discussion

Tease Out BH

Results presented here generally agree with what is known about estimating population growth rate parameters [12, 2, **?**]. These studyies appreciate the role of contrast for estimating growth rates, however they struggle to make generally extensible conclusions since they focus only on a handful of stocks that fall short of forming a random sample of the greater population of possible stock behaviors. The LHS design methods presented here are designed specifically to simulate a representative sample of stocks broadly across the space of possible RPs. Furthermore, the simulation design, taken together with the GP metamodel of productivity parmater estimates, allows this study to control the degree of model misspecification and generalize conclusions about the behavior of productivity estimation within the production model setting presented.

In the presence of contrast, F^* estimation can enjoy very low bias even for a wide

range of poorly specified models; conversely in the absence of contrast F^* estimation can suffer very large bias even for slightly misspecified models. This pattern is particularly true for low-contrast inference under the Schaefer model where the geometry of the restricted RP set isolates estimation failure of F^* from $\frac{B^*}{B(0)}$. While contrast has a similar impact on F^* estimation under the BH model, the geometry of the BH RP set correlates estimation bias of F^* and $\frac{B^*}{B(0)}$. The GP metamodeling approach reveals a more general pattern that highly informative data sets (high contrast) produces a nearly minimal distance mapping of RPs onto the constrained RP set.

In all cases when model misspecification is removed, even with weakly informative data, RP estimation is unbiased and well estimated. Thus contrast alone is not the only factor leading to inferential failure. Model misspecification is a necessary but not sufficient condition for inducing RP estimation bias. The particular RP bias present depends on the RP geometry of the fitted model and how that geometry is misspecified relative to the data. The RP mapping is then oriented to the RP geometry of the fitted model.

While the relative fishing rate parameterized in Section (2.2.6) captures a usefully broad spectrum of relevant fishing behaviors, it is still limiting in the amount of information that it can induce. Improved methods for quantifying contrast in fisheries data, and/or methods of discovering more informative fishing behavior, could improve this analysis. In the absence of a maximally informative dataset simulation methods will not fully describe how inference fails, but the methods presented here tell the most complete picture yet, with explicit control of the degree model misspecification, contrast, and a simulation design that allows for uniform representative data generation across biologically meaningful stocks. The results presented here suggest the conjecture that under a maximally informative dataset, RP inference with a two parameter production function will be biased in the direction a shortest distance map from the true RPs onto restricted set of RPs under the two-parameter model.

Given the potential for model misspecification of RPs, a minimal distance mapping of RPs represents a best-case scenario where the total bias of RPs, when measured jointly, is minimized. That said, without recognizing the geometry of how two-parameter models of productivity limit RP space this may lead to unintuitive implications in RP estimation. For example, due to the shape of the BH RP set a minimal distance mapping ensures that if there

is bias in one of $\frac{B^*}{B_0}$ or F^*, there will necessarily be bias in the other RP. However under the Schaefer model, since the RP set is a constant in $\frac{B^*}{B_0}$, bias in F^* is not adulterated in the same way by bias in $\frac{B^*}{B_0}$ estimation. While models with constant RPs, such as the logistic model $\frac{B^*}{B_0} = \frac{1}{2}$ or the Fox model $\frac{B^*}{B_0} = \frac{1}{e}$, are extremely limited, they can be valuable tools for developing intuition precisely because they isolate RP estimation in their free RPs from the correlated RP biases present in models like the BH or Ricker model.

When one considers the implications of RP bias, overestimation of RPs carries the severe implication of management recommendations potentially leading to overfishing, while underestimation of RP leads to overly conservative management. In this sense, when the true model is not known, the geometry of the BH set together with the metamodeled bias trends makes the BH model a naturally conservative estimator of RPs for most stocks. For most non-BH populations the BH model is likely to make conservative errors in its estimates of F^* and $\frac{B^*}{B_0}$. The one notable exception to the conservatism of the BH model stands for data generated in the Cushing-like regime of Schnute RPs. In this regime the BH model tends to be fairly unbiased overall, however the bias that is present for these populations tends to be overestimation in both RPs, leading to much more severe management consequences for those populations.

The RP bias trends of the Schaefer model demonstrate much less conservatism than the BH overall. For any population with $\frac{B^*}{B_0} < 0.5$, $\frac{B^*}{B_0}$ will be overestimated. When the population comes from the regime where $\frac{B^*}{B_0} > 0.5$, $\frac{B^*}{B_0}$ will be under estimated, but F^* is likely to be overestimated depending on the degree of contrast present in the data. So while the Schaefer model is an intuitive model, it tends to lead to much less conservative RP estimation.

While it is important to recognize these limitations of two-parameter models of productivity, we should not solely accept conservativism as a rational of choosing a BH model of productivity. Increasing the flexibility of the production function by moving toward three-parameter models would release the underlying structural limitations [14] that cause these RP biases in the first place. Punt & Cope [17] considers a suite of possible three-parameter curves which could be used instead of current two-parameter curves. For all of their benefits, three parameter production functions have their own complicating factors, and the structure present in the Schnute model explored here makes it an intuitive bridge model for developing three-

parameter models going forward.

Chapter 3

Schnute Model

3.1 Introduction

3.2 Methods

3.2.1 Model

The Schnute production function is a three-parameter generalization of many of the most common two-parameter production functions [**?**, **?**]. It can be written in the following form, with parameters α, β, and γ,

$$P_s(B;[\alpha,\beta,\gamma]) = \alpha B(1-\beta\gamma B)^{\frac{1}{\gamma}}. \tag{3.1}$$

The BH and Logistic production functions arise when γ is fixed to -1 or 1 respectively. The Ricker model is a limiting case as $\gamma \to 0$. For $\gamma < -1$ a family of strictly increasing Cushing-like curves arise, culminating in linear production as $\gamma \to -\infty$. These special cases form natural regimes of similarly behaving production functions as seen in Figure (3.1).

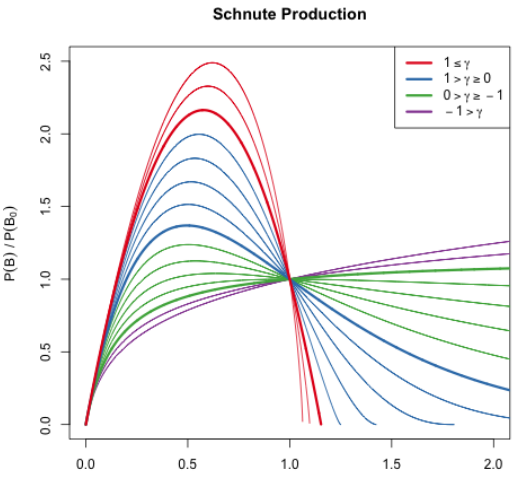

Figure 3.1: The Schnute production function plotted across a variety of parameter values. Regimes of similarly behaving curves are grouped by color.

The behavior of RP inference under the BH model is of particular interest due to the overwhelming popularity of the BH assumption in fisheries models. Since Schnute production models can represent a quantifiably wide variety of possible productivity behaviors, they present an ideal simulation environment for inquiry of the reliability of inference under the BH assumption.

Under Schnute production, biomass dynamics evolve according to the following ODE,

$$\frac{dB}{dt} = P_s(B;\theta) - (M+F)B. \tag{3.2}$$

This equation largely takes the same form as previously described, except that P_s is the Schnute production function and natural mortality, M, is modeled explicitly here. Natural mortality models the instantaneous rate of mortality from all causes outside of fishing. While Eq. (3.2) models M explicitly, natural mortality is implicit to the structure of the previously decribed Schaefer, Fox, and PT production models. Explicitly modeling natural mortality allows for the production function not to approach (or intersect) 0 for large biomasses (e.g. BH production). In turn, the Schunte model requires the addition of the term $-MB$ to form an interpretable yeild curve and make RPs well defined over the relevant domain of γ.

The derivation of RPs under Eq. (3.2) follows a similar logic as under the PT model. An expression for equilibrium biomass is attained by setting $\frac{dB}{dt} = 0$ and rearranging the resulting expression to solve for B

$$\bar{B}(F) = \frac{1}{\gamma\beta}\left(1 - \left(\frac{M+F}{\alpha}\right)^{\gamma}\right). \tag{3.3}$$

The above expression quickly yields B_0, B^* by evaluation at $F = 0$ and F^* respectively,

$$B_0 = \frac{1}{\gamma\beta}\left(1 - \left(\frac{M}{\alpha}\right)^{\gamma}\right) \tag{3.4}$$

$$\frac{B^*}{B_0} = \frac{1 - \left(\frac{M+F^*}{\alpha}\right)^{\gamma}}{1 - \left(\frac{M}{\alpha}\right)^{\gamma}}. \tag{3.5}$$

Attaining an expression for F^* requires maximization of equilibrium yield, $\bar{Y} = F\bar{B}(F)$,

with respect to F. Analytically maximizing proceeds by differentiating $\bar{Y}$ to produce

$$\frac{d\bar{Y}}{dF} = \bar{B}(F) + F\frac{d\bar{B}}{dF} \tag{3.6}$$

$$\frac{d\bar{B}}{dF} = -\frac{1}{\beta}\left(\frac{\left(\frac{M+F}{\alpha}\right)^{\gamma}}{F+M}\right). \tag{3.7}$$

Setting $\frac{d\bar{Y}}{dF} = 0$, filling in the expressions for $\bar{B}(F)$ and $\frac{d\bar{B}}{dF}$, then rearranging to solve for F^* is less yielding here than it was in the case of the PT model. This procedure falls short of providing an analytical solution for F^* directly in terms of θ, but rather shows that F^* must respect the following expression,

$$0 = \frac{1}{\gamma} - \left(\frac{1}{\gamma} + \frac{F^*}{F^*+M}\right)\left(\frac{F^*+M}{\alpha}\right)^{\gamma}. \tag{3.8}$$

The lack of an analytical solution here is understood. Schnute & Richards [21, pg. 519] specifically point out that F^* cannot be expressed analytically in terms of productivity parameters, but rather gives a partial analytical expression for the inverse relationship. Although parameterized slightly differently, Schnute & Richards derive expressions for α and β as a function of RPs and γ.

Since RPs are left without a closed form expression, computing RPs from productivity parameters amounts to numerically solving the system formed by collecting the expressions (3.8), (3.4), and (3.5).

3.2.2 Simulation

For the purpose of simulation, it is not necessary to completely know the precise relationships mapping RPs $\mapsto \theta$ or $\theta \mapsto$ RPs. Simulation only requires enough knowledge of these mappings to gather a list of (α, β, γ) tuples, for data generation under the Schnute model, and the corresponding RPs in some reasonable space-filling design over RP space.

Similarly to Schnute & Richards [21], expressions (3.8) and (3.4) are solved for α and β respectively. This leads to the partial mapping $\left(F^*, B_0\right) \mapsto \left(\alpha(\cdot,\gamma),\ \beta(\cdot,\cdot,\gamma)\right)$ in terms of

RPs and γ. By further working with Eq. (3.5), to identify γ, the following system is obtained,

$$\begin{aligned}
\alpha &= (M+F^*)\left(1+\frac{\gamma F^*}{M+F^*}\right)^{1/\gamma} \\
\beta &= \frac{1}{\gamma B_0}\left(1-\left(\frac{M}{\alpha}\right)^{\gamma}\right) \\
\frac{B^*}{B_0} &= \frac{1-\left(\frac{M+F^*}{\alpha}\right)^{\gamma}}{1-\left(\frac{M}{\alpha}\right)^{\gamma}}.
\end{aligned} \tag{3.9}$$

For a population experiencing natural mortality M, by fixing F^*, B_0, and $\frac{B^*}{B_0}$ the above system can fully specify α and β for a given γ. Notice for a given γ a cascade of closed form solutions for α and β can be obtained. First $\alpha(\gamma)$ can be computed, and then $\beta(\alpha(\gamma),\gamma)$ can be computed. If $\alpha(\gamma)$ is filled back into the expression for $\frac{B^*}{B_0}$, the system collapses into a single onerous expression for $\frac{B^*}{B_0}(\alpha(\gamma),\gamma)$. For brevity, define the function $\zeta(\gamma) = \frac{B^*}{B_0}\big(\alpha(\gamma),\gamma,F^*,M\big)$ based on Eq. (3.5).

Inverting $\zeta(\gamma)$ for γ, and computing the cascade of $\alpha(\gamma)$, and then $\beta(\alpha(\gamma),\gamma)$, fully defines the Schnute model for a given $(\frac{F^*}{M},\frac{B^*}{B_0})$. However inverting ζ accurately is extremely difficult. Inverting ζ analytically is not feasible, and numerical methods for inverting ζ are unstable and can be computationally expensive. Rather than numerically invert precise values of $\zeta(\gamma)$, γ is sampled so that the overall simulation design is space filling as described in Section (3.2.4).

Each design location defines a complete Schnute production model with the given RP values. Indices of abundance are simulated from the Schnute model at each design location, a small amount of residual variation, $\sigma = 0.01$, is added to the simulated index, and the data are then fit with a misspecified BH production model. The design at large captures various degrees of model misspecification relative to the BH model, so as to observe the effect of productivity model misspecification upon RP inference.

3.2.3 Latin Hypercube Sampling

The goal of space filling design in this setting is to extend the notion of the random sample (and its desirable parameter estimation properties) across the simulated RP domain so as to represent the simulated space as well as possible [7]. The simple random sample is the classical approach to unbiased parameter estimation, however simple randomness is patchy, often sampling some regions of design space quite densely, while leaving other regions of design space empty. Space filling designs aim to preserve (or enhance) parameter estimation properties across the simulated domain [**?**, **?**], while constraining samples to be spaced in some notion of spread over the entire space. Latin hypercube sampling [**?**, LHS] is among the most foundational of space filling designs used in computer experiments.

A LHS of size n, in the 2 dimensional space defined by RPs, distributes samples so as to spread points across a design region in a broadly representative way. A LHS design extends the notion of a univariate random uniform sample across multiple dimensions so that each margin of the design space enjoys a uniform distribution.

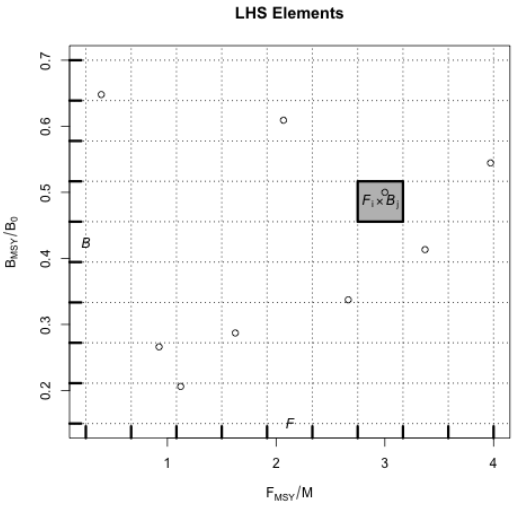

Figure 3.2: LHS grids. Intersecting $\mathcal{F}$ and $\mathcal{B}$ produces n^2 cells; a particular cell $\mathcal{F}_i \times \mathcal{B}_j$ is shown in grey. One point is in each of the marginal $\mathcal{F}_i$ and $\mathcal{B}_j$ grid elements.

LHS designs achieve this notion of uniformity by first partitioning each dimension of the design space into regular grids of size n. By intersecting the grids of each dimension, cells are produced that evenly partition the design space. In two dimensions n^2 cells are produced, from which a total of n samples are taken. Crucially only one point is randomly sampled from a given element of each grid in each dimension so as to reduce clumping of the n samples across the design space.

3.2.4 Design

Due to the lack of an analytical relationship mapping RPs $\mapsto \theta$, analogous to the PT model's Eq. (2.10), producing a LHS design over Schnute RPs requires a more tactful approach. The structured relationship between the RPs and productivity parameters, described in Section (3.2.2), allows an approximate LHS to be obtained by a careful navigation of the system of equations seen in Eq. (3.9).

Under the Schnute model, let $\mathcal{F}$ and $\mathcal{B}$ represent regular grids on $\frac{F^*}{M} \in (0.25, 4)$ and $\frac{B^*}{B_0} \in (0.15, 0.7)$ respectively which can serve as the scaffolding for computing an approximate LHS.

Since it is not practical to invert $\zeta(\gamma)$, a uniform sample in $\frac{B^*}{B_0}$ can be obtained by modeling γ as a random variable, with realization γ^*, and thinking of $\zeta(\gamma)$ as its cumulative distribution function (CDF). The aim is to model γ as an easily sampled random variable with a CDF that closely approximates ζ, so that $\zeta(\gamma^*) \tilde{\sim} U(\zeta_{min}, 1)$ as closely as possible. There may be many good models for the distribution of γ, but in this setting the following distribution is very effective,

Given B_0, M, and F^*:

1) Draw $\gamma^* \sim \gamma | F^*, M$.
2) Compute $\frac{B^*}{B_0} = \zeta(\gamma^*)$
3) Compute $\alpha^* = \alpha(\gamma^*, F^*, M)$
4) Compute $\beta^* = \beta(\alpha^*, \gamma^*, M, B_0)$

Figure 3.3: An outline of the sampling procedure for γ given B_0, M, and F^*.

$$\gamma \sim \zeta_{min}\delta(\gamma_{min}) + t(\mu, \sigma, \nu)\mathbf{1}_{\gamma > \gamma_{min}}. \tag{3.10}$$

Above, t is the density of the three-parameter location-scale family Student's t distribution with location μ, scale σ, and degrees of freedom ν. $\mathbf{1}_{\gamma > \gamma_{min}}$ is an indicator function that serves to truncate the Student's t distribution at the lower bound γ_{min}. $\delta(\gamma_{min})$ is the Dirac delta function evaluated at γ_{min},

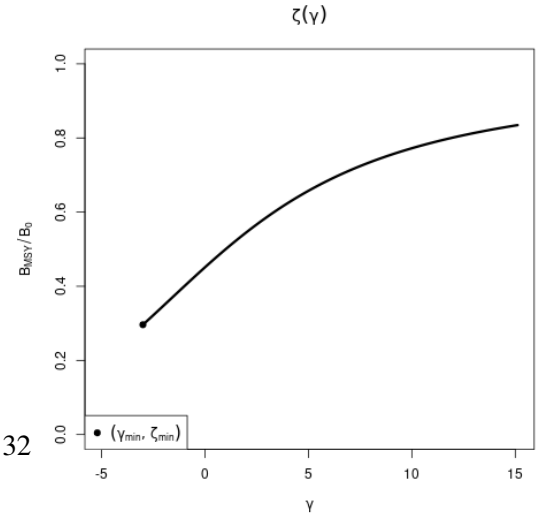

Figure 3.4: $\zeta(\gamma)$ Plotted for $F^* = 0.1$ and $M = 0.2$. The point $(\gamma_{min}, \zeta_{min})$ shows the lowest biologically meaningful value of γ; below which productivity is negative.

which is scaled by the known value ζ_{min}; this places probability mass ζ_{min} at the point γ_{min}. Since sampling from a Student's t distribution is readily doable, sampling from a truncated Student's t mixture only requires slight modification.

Let T be the CDF of the modeled distribution of γ. Since the point $(\gamma_{min}, \zeta_{min})$ is known from the dynamics of the Schnute model at a given RP, full specification of Eq. (3.10) only requires determining the values for μ, σ, and ν which make T best approximate $\zeta(\gamma)$. Thus, the values of μ, σ, and ν are chosen by minimizing the L^2 distance between $T(\gamma)$ and $\zeta(\gamma)$.

$$[\hat{\mu}, \hat{\sigma}, \hat{\nu}] = \underset{[\mu,\sigma,\nu]}{\arg\min} \int_{\Gamma} \big(T(\gamma;\mu,\sigma,\nu) - \zeta(\gamma)\big)^2 d\gamma \tag{3.11}$$

The distribution $T(\gamma|\hat{\mu},\hat{\sigma},\hat{v})$ is fit for use in generating γ^* random variates at a specific F^* and M. This approximation releases the need to invert ζ w.r.t. γ by using samples of γ^* values to generate approximatly uniform samples of $\zeta(\gamma^*)$. By sampling approximatly uniform $\zeta(\gamma^*)$ random variates in this way, and making use of the structure in Eq. (3.9), an approximate LHS sample can be collected via Algorithm (1).

Algorithm 1 LHS of size n on rectangle R.

procedure $LHS_n(R)$
  Define n-grids $\mathcal{F}, \mathcal{B} \in R$
  for each grid element i **do**
    Draw $\frac{F^*}{M} \sim Unif(\mathcal{F}_i)$
    Compute $[\hat{\mu},\hat{\sigma},\hat{v}]$ given F^* & M
    while $\mathcal{B}_j$ not sampled **do**
      Draw $\gamma^* \sim T(\gamma|\hat{\mu},\hat{\sigma},\hat{v})$
      Compute $\zeta^* = \zeta(\gamma^*)$
      Compute j such that $\zeta^* \in \mathcal{B}_j$
    end while
    Compute $\alpha^* = \alpha(\gamma^*, F^*, M)$
    Compute $\beta^* = \beta(\alpha^*, \gamma^*, M, B_0)$
    Save $(\frac{F^*}{M}, \zeta^*) \Leftrightarrow (\alpha^*, \beta^*, \gamma^*)$ in $\mathcal{F}_i \times \mathcal{B}_j$
  end for
end procedure

For a given i, $\frac{F^*}{M}$ is drawn uniformly from within $\mathcal{F}_i$. Conditioning on the sample of F^*, and M, $T(\gamma|\hat{\mu},\hat{\sigma},\hat{v})$ is fit and γ^* is sampled. ζ^* is then computed and placed into the appropriate grid element $\mathcal{B}_j$. Given γ^*, the cascade $\alpha(\gamma^*)$, and $\beta(\alpha(\gamma^*),\gamma^*)$, can be computed. The algorithm continues until all of the design elements, $(\frac{F^*}{M}, \zeta^*) \Leftrightarrow (\alpha^*, \beta^*, \gamma^*)$, have been computed for all $i \in [1, ..., n]$.

3.2.4.1 Design Refinement

Since the behavior of RP inference, under misspecified models, will vary in yet-unknown ways, the exact sampling design density may be hard to know a priori. Several factors, including the particular level of observation uncertainty, high variance (i.e. hard to resolve) features of the response surface, or simply "gappy" instantiations of the initial LHS design may necessitate adaptive design refinement, to accurately describe RP biases. Given the temperamental relationship between RPs and productivity parameters in the Schnute model, a recursive refinement algorithm that makes use of the previously described LHS routine, is developed.

While LHS ensures uniformity in the design margins, and a certain degree of spread,

it is widely recognized that particular LHS instantiations may leave substantive gaps in the simulation design. To correct this, LHS is often paired with design elements of maximin design [**?**, **?**]. Maximin designs sample the design space by maximizing the minimum distance between sampled points. This has the advantage of definitionally filling holes in the design, however because no points are ever drawn outside of the design domain, samples tend to clump around edges (particularly corners) of the design domain. Since LHS ensures uniformity in the margins and maximin designs enjoys a certain sense of optimality in how they define and fill gaps [**?**], the methods are quite complimentary when combined.

Making use of this complimentary relationship, holes in the existing LHS design of RPs are identified based on maximin design principles. New design points are collected based on areas of the RP design space which maximizes the minimum distance between all pairs of points in the current design, based on the following distance function

$$d(\boldsymbol{x}, \boldsymbol{x'}) = \sqrt{(\boldsymbol{x} - \boldsymbol{x'})^T \boldsymbol{D}^{-1} (\boldsymbol{x} - \boldsymbol{x'})} \tag{3.12}$$
$$\boldsymbol{D} = \textbf{diag}\left[\left(max(\mathcal{F}) - min(\mathcal{F})\right)^2,\ \left(max(\mathcal{B}) - min(\mathcal{B})\right)^2 \right].$$

Above, d is a scaled distance function that defines the distance between points in the differing scales of $\frac{B^*}{B_0}$ and $\frac{F^*}{M}$. $\boldsymbol{D}$ is a diagonal matrix that measures the squared size of the domain in each axis of so as to normalize distances to a common scale.

If $\boldsymbol{X_n}$ is the initial design, computed on R_{full}, let $\boldsymbol{x_a}$ be the augmenting point which maximizes the minimum distance between all of the existing design points,

$$\boldsymbol{x_a} = \underset{\boldsymbol{x'}}{\text{argmax}}\ \ \min\{d(\boldsymbol{x_i}, \boldsymbol{x'}) : i = 1, ..., n\}. \tag{3.13}$$

The point $\boldsymbol{x_a}$ is used as an anchor for augmenting $\boldsymbol{X_n}$. An additional $LHS_{n'}$ (via Algorithm (1)) is collected, adding n' design points, centered around $\boldsymbol{x_a}$, to the overall design. The augmenting region, $R_{(x_a, d_a)}$, for collecting $LHS_{n'}$ is defined based on the square centered at $\boldsymbol{x_a}$ with side length $2d_a$, where $d_a = \min\{d(\boldsymbol{x_i}, \boldsymbol{x_a}) : i = 1, ..., n\}$, in the space defined by the metric d.