Quantifying Uncertainty in the Estimation of Probability Distributions

Jimena Lamanda Davis

Department of Energy Computational Science Graduate Fellow

In Collaboration with H.T. Banks

June 18, 2008

References

- H.T. Banks, V.A. Bokil, S. Hu, A.K. Dhar, R.A. Bullis, C.L. Browdy, and F.C.T. Allnutt, "Modeling shrimp biomass and viral infection for production of biological countermeasures," CRSC-TR05-45; NC State University, December, 2005; *Mathematical Biosciences and Engineering,* **3**, pp. 635-660, 2006.

- H.T. Banks, L.W. Botsford, F. Kappel, and C. Wang, "Modeling and estimation in size structured population models," LCDS-CC Rpt. 87-13, Brown University; *Proceedings 2nd Course on Mathematical Ecology,* (Trieste, December 8-12, 1986) World Press, Singapore, pp. 521-541, 1988.

- H.T. Banks and J.L. Davis, "A comparison of approximation methods for the estimation of probability distributions on parameters," CRSC-TR05-38; NC State University, October, 2005; *Applied Numerical Mathematics,* **57**, pp. 753-777, 2007.

References

- H.T. Banks and J.L. Davis, “Quantifying uncertainty in the estimation of probability distributions with confidence bands,” CRSC-TR07-21; NC State University, December, 2007; *Mathematical Biosciences and Engineering* (accepted).
- H.T. Banks, J.L. Davis, S.L. Ernstberger, S. Hu, A.K. Dhar, and C.L. Browdy, “Comparison of probabilistic and stochastic approaches in modeling growth uncertainty and variability,” CRSC-TR08-03; NC State University, February, 2008; *Journal of Biological Dynamics* (accepted).
- G.A.F. Seber and C.J. Wild, *Nonlinear Regression,* John Wiley & Sons, New York, 1989.
- J.W. Sinko and W. Streifer, “A new model for age-size structure for a population,” *Ecology,* **48**, pp. 910-918, 1967.

Motivation

- Development of an inverse problem computational methodology for the estimation of functional parameters in the presence of model and data uncertainty
- Applications involve the estimation of growth rate distributions in size-structured marine populations (Type II problem - aggregate or population level longitudinal data)
- Extension of the asymptotic standard error theory for finite-dimensional ordinary least squares (OLS) estimators to “functional” confidence bands that will aid in quantifying the uncertainty in estimated probability distributions

Application: Size-Structured Shrimp Population

- Use of shrimp as a scaffold organism to produce large amounts of a vaccine rapidly in response to a toxic attack on populations [Banks et. al. 2006]

- Joint project with ABN (Advanced Bionutrition Corporation) involving the development of a hybrid model of the shrimp biomass/countermeasure production system
- Being able to accurately model the dynamics of the size-structured shrimp population is important since the output of the biomass model will serve as input to the vaccine production model

Sinko-Streifer (SS) Model for Size-Structured Populations (1967)

- Widely used to describe various age and size-structured populations (cells, plants, and marine species)

$$v_t(t,x;g) + (g(t,x)v(t,x;g))_x = -\mu(t,x)v(t,x;g), \quad \underline{x} < x < \bar{x} \quad (1)$$

- Initial Condition

$$v(0,x) = v_0(x;g)$$

- Boundary Condition

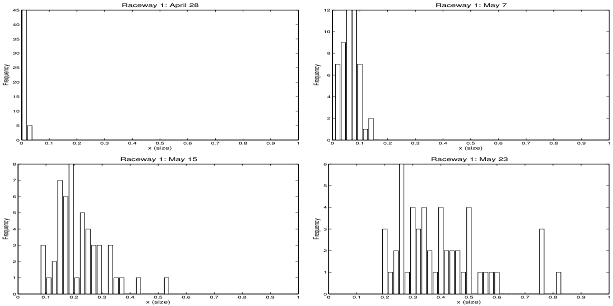
$$g(t,\underline{x})v(t,\underline{x};g) = \int_{\underline{x}}^{\bar{x}} K(t,\xi)v(t,\xi;g)d\xi$$

- Deterministic Growth Rate Model

$$\frac{dx}{dt} = g(t,x)$$

Example of Aggregate Type Longitudinal Shrimp Data

- Previous size-structured population data from a group in Texas indicates variability in size that could be a result of variability in growth rates and might suggest the use of GRD model (2)



Growth Rate Distribution (GRD) Model (1988)

- Deterministic growth model of (1) is not biologically reasonable when modeling populations that exhibit a great deal of variability in aggregate type longitudinal data as time progresses
- GRD model, a modification of the SS model, was developed by Banks et. al. [Banks et. al. 1988] to account for the variability observed in populations, such as size-structured mosquitofish populations that exhibit both dispersion and bifurcation in time
- Assumption of GRD: Individual growth rates vary across the population

$$u(t,x;P) = \int_{\mathcal{G}} v(t,x;g)\,dP(g) \tag{2}$$

Early Growth Dynamics of Shrimp

- We assume that mortality rate and reproduction rate in (1) are both zero
- We also assume that the size-dependent growth rate function of the shrimp has the form

$$g(x; b, c) = b(x + c),$$

which was shown to provide reasonable fits to average size data for 50 randomly sampled shrimp in [Banks et. al. 2008]

- Intrinsic growth rate b is a random variable taking values in a compact set $\mathcal{B}$
- Analysis of previous data also suggested that the assumption of a normal distribution on the intrinsic growth rates leads to a lognormal distribution in size
- We choose a truncated normal distribution with mean μ_b and standard deviation σ_b

Standard Parametric Approach - PAR(M, N)

- In the parametric approach, we assume that we know the distribution of the growth rates

- Assuming P is (absolutely) continuous ($\frac{dP}{db} = p$), the population density from the GRD model (2) is given by

  $$u(t, x; \theta) = \int_{\mathcal{B}} v(t, x; g(x; b))p(b; \theta)db,$$

  where $\theta \in \mathbb{R}_+^M$ represents the parameters (μ_b, σ_b) that are associated with the a priori probability density and distribution

- M represents the number of parameters in θ and N represents the number of quadrature nodes used to approximate the integral above

Parameter Estimation with PAR(M,N)

- Ordinary Least Squares Formulation (assuming constant variance noise model)
- We wish to solve for $\hat{\theta}$

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}_+^M} J(\theta) = \arg \min_{\theta \in \mathbb{R}_+^M} \sum_{i,j} |u(t_i, x_j; \theta) - \hat{u}_{ij}|^2$$

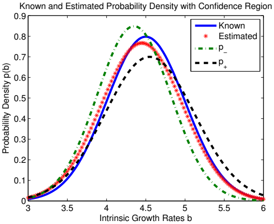
- We use MATLAB **fmincon** to determine the optimal values of $\theta = (\mu_b, \sigma_b)$ used to generate the estimated probability density and distribution

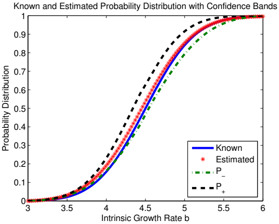
Confidence Intervals... Confidence Bands

- Since we reduced infinite dimensional estimation problem to finite dimensional problem for θ, we are able to compute standard errors based on the established asymptotic standard error theory for OLS estimators [Seber and Wild 1989]

- Standard errors are used to compute confidence intervals to quantify the uncertainty in the estimated finite dimensional parameter θ

- How does one use the confidence intervals computed in the finite dimensional setting to construct confidence bands in the infinite dimensional setting?





Monte Carlo Sampling Study to Aid in the Design of Experiments

- Goal: Determine the sampling *size* N_s and sampling *frequency* N_t needed to obtain reliable estimates of the probabilistic growth rate parameters in the GRD model (2) - experiments to be carried out at ABN and SCDNR (South Carolina Department of Natural Resources) [Banks et. al. 2008]

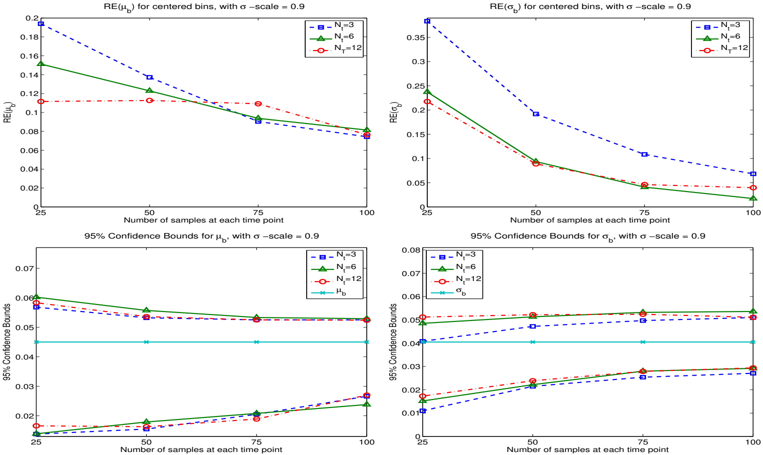
- Population data (total number of shrimp in each size class) used in inverse problem calculations

$$N_{GRD}(t, x; \theta) \approx u(t, x; \theta)\Delta x,$$

  where Δx is the length of the size class interval

- We simulated population data, where N_s varied from $25, 50, 75$ to 100 and N_t varied from twice a week, once a week to once every two weeks

Monte Carlo Sampling Study Results



- Conclusions: Most desirable experiment involved using $N_s = 100$ once a week; however there appears to be little loss in accuracy if one uses $N_s = 50$

Parameter Estimation Results with ABN Data

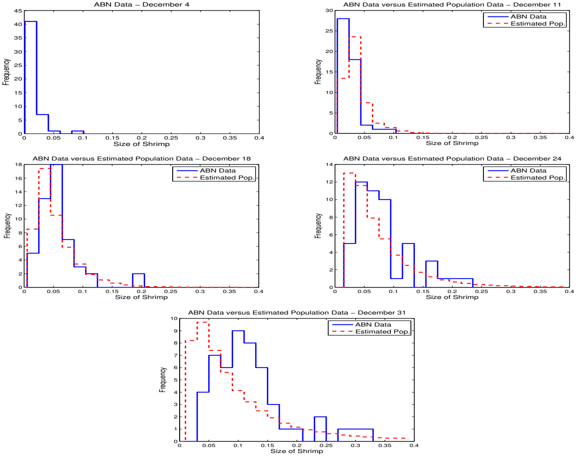
- Inverse problem with data (subsequently) collected from shrimp cultured in tanks at ABN

- Fifty shrimp were randomly sampled and measured once a week under relatively constant tank conditions

- Using our methodology, we determined estimates of the growth rate distribution and quantified the uncertainty associated with these estimates with confidence bands

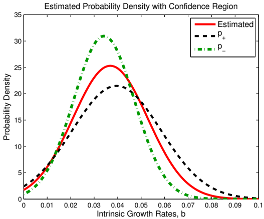
PAR(2,128) Results with Complete December Data

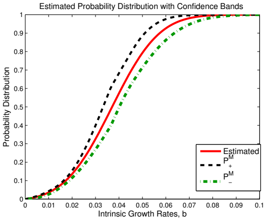
- $\hat{\mu}_b \pm 1.96 SE(\hat{\mu}_b) : 0.0010 \pm 0.0535$
  $\hat{\sigma}_b \pm 1.96 SE(\hat{\sigma}_b) : 0.0324 \pm 0.0313$
- $J^* = 574.8315, \hat{\sigma}^2 = 17.4191$



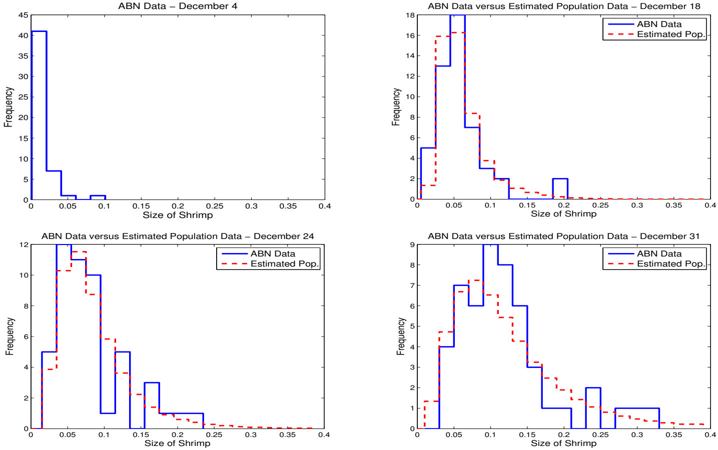
PAR(2,128) Results Excluding December 11 Data

- $\hat{\mu}_b \pm 1.96 SE(\hat{\mu}_b) : 0.0369 \pm 0.0027$
  $\hat{\sigma}_b \pm 1.96 SE(\hat{\sigma}_b) : 0.0159 \pm 0.0030$
- $J^* = 87.4619, \hat{\sigma}^2 = 3.1236$


Estimated Probability Density with Confidence Region


Estimated Probability Distribution with Confidence Bands

PAR(2,128) Results Excluding December 11 Data



Summary and Ongoing Work

- We have demonstrated how mathematical and statistical tools can be used to gain insight into the early growth dynamics of shrimp.
- We are working on improving the model predictions to the shrimp population data by considering different parametric and non-parametric approaches [Banks and Davis 2005] in the GRD model.
- Following the work of Seber and Wild, we are also working on fully developing the mathematical and asymptotic statistical theory ("functional" confidence bands) for OLS inverse problems where the parameter of interest is a probability distribution.
- We would also like to determine if the confidence bands constructed in the non-parametric approximation methods (not discussed here today) are converging to some "true" smooth confidence bands.

Acknowledgements

- Department of Energy Computational Science Graduate Fellowship and the Krell Institute Staff
- Collaborators:
  - NCSU
    - Dr. H.T. Banks
    - Dr. Shuhua Hu
    - Stacey Ernstberger
  - Advanced Bionutrition Corporation
    - Dr. Elena Artimovich
    - Dr. Arun K. Dhar
  - Marine Resources Research Institute
    - Dr. Craig L. Browdy