Hierarchical Bayesian Markov Switching Models with Application to Predicting Spawning Success of Shovelnose Sturgeon

Scott Holan
*University of Missouri - Columbia*

Joint work with

**Ginger Davis** - University of Virginia

and


*Washington University - December 4, 2008*
Outline

- Background and Motivation
- Data
- Model and Methods
- Results
- Summary and Future Directions
Recruitment of pallid sturgeon to the adult population is limited in the Missouri River.

Species is rare in the Missouri River and was listed as an endangered species in 1990.

Possible reasons for decline in pallid sturgeon population:

1. Commercial Harvest
2. Habitat Alteration
3. Pollution
4. Impoundment (Dam Construction)

Similar to the pallid sturgeon, the shovelnose sturgeon is declining and is at risk of extirpation.
Determine the ecological requirements for reproduction and survival of pallid and shovelnose sturgeon in the Missouri River.

Shovelnose sturgeon closely related to the pallid sturgeon, and spawning requirements and behavior are similar in many respects.

Use shovelnose as a surrogate species to develop new research tools, or to examine the impacts of management actions, or environmental variables on sturgeon biology and habitat use.

Understanding the difference in successful and unsuccessful spawners within the shovelnose sturgeon should provide us with some knowledge concerning the spawning success of the closely related pallid sturgeon.
USGS Columbia Environmental Research Center - Study Objectives

1. **Determine the direction, magnitude, and habitat used during spawning migrations** for shovelnose sturgeon at two geologically and hydrologically distinct reaches of the lower Missouri River

2. **Describe the reproductive physiology of shovelnose sturgeon** prior to and after successful and unsuccessful spawning

3. **Identify and rank proximate cues necessary for successful spawning** by Missouri River sturgeon
Study Area - Missouri River Basin
Study Area - Two geologically and hydrologically distinct segments of the Lower Missouri River
Data Collection Process

- Study Subjects
  - 2004: 9 female shovelnose sturgeon
  - 2005: 15 female shovelnose sturgeon
  - 2006: 20 female shovelnose sturgeon

- Biologists track sturgeon using two types of implanted telemetry devices

- Ultrasonic transmitters provide the location of fish, which are tracked through the suspected spawning period

- Archival data storage tags (DST’s) record the temperature and depths of the fish every 15 minutes

- Goal: Use collected data to compare behavioral and environmental factors for spawning and non-spawning sturgeon
Ultrasound

USGS Fisheries Biologist is checking for female and readiness to spawn (note spawning only occurs every 2-3 years).
Reproductive Stage V
Telemetry and DST Devices
Telemetry Device Implantation
Fish Recapture
Biological Variables of Interest

- **Sl** = standard length of fish in mm
- **Fl** = fork length of fish in mm
- **Wt** = weight of fish in kg
- **PI** = polarization index
  - percent distance the germinal vesicle is to the edge of the egg
  - The lower the number the farther the nucleus has migrated and the closer the fish is to spawning.
- **Cape2** = capture estradiol level in pg/mL
- **Cap11kt** = capture 11-ketotestosterone level in pg/mL
- **Capc** = capture cortisol level in ng/mL
Blood Sample

Removal of blood from a shovelnose sturgeon
Environmental and Behavioral Variables of Interest

- **Transcode** = unique fish number (transmitter code number)

- **Capture segment** = denotes whether the fish was caught in the south or north section of the river

- **Year** = year fish was caught

- **Depth** = depth of fish

- **Temperature** = temperature of fish

- **Location** = river location
Example Time Series Plot of Depth, Temperature, and Location
Response Variables

- **Recapoocyteratio** = ratio of mature oocytes (eggs) to early stage oocytes
  - Lower ratios are indicative of more complete the spawning.
  - Also known as **spawning index**
  - Logit transformation: \( \text{logitratio} = \log \frac{\text{recapoocyteratio}}{1-\text{recapoocyteratio}} \)

- **Recapspawn** = Categorical variable of the continuous recapocyteratio variable
  - 0 – 35% = complete spawn
  - 35 – 75% = incomplete spawn
  - > 75% = no spawn

- Note that this choice of threshold was determined through extensive empirical investigation by expert fisheries biologists
Partial and Complete Spawn

Partial-Spawner

Complete-Spawner
Model Motivation

After examining depth profiles for successful and non-successful spawners, it is hypothesized that the variability of their depth profiles could be useful in predicting spawning success.
Model Motivation - Exploratory Analysis

- Fit Markov Switching Stochastic Volatility Models - Smith 2002
  - 2-regime (high and low) model for volatility
  - Switching between regimes is governed by probability transition matrix
    \[
    \begin{pmatrix}
    1 - e_i & e_i \\
    e_i & 1 - e_i
    \end{pmatrix}
    \]
  - It was observed that the values $e_i$ were higher for fish who had no spawn or only a partial spawn
  - univariate measurement = 2nd eigenvalue = $1 - e_i - e_i$
  - What does this mean?
    - Fish who do not spawn or spawn only partially transition between the high and low variability states more frequently than the fish who spawn completely
Let $\lambda_i$ and $z_i$ denote the eigenvalue and logit ratio for the $i^{th}$ fish respectively.

Linear Regression Model

$$ z_i = \gamma_0 + \gamma_1 wt_i + \gamma_2 sl_i + \gamma_3 Pl_i + \gamma_4 cape2_i + \gamma_5 cap11kt_i + \gamma_6 capci + \gamma_7 \lambda_i + \epsilon_i $$

What is the problem here?

- $\lambda_i$ is not given, but estimated
- differing levels of uncertainty / sample sizes (profile lengths)

One solution: Take a Bayesian approach
Hierarchical Bayesian Model Definition

Heuristically the model can be thought of as follows

1. Higher level of hierarchy
   - Generate draws from distribution of the second eigenvalue of the transition probability matrix in a Markov switching volatility model

2. Lower level of hierarchy
   - Given these eigenvalues, we estimate the regression parameters of interest in a linear model
Two-state Markov Switching Volatility Model Definition

- $d_{it}$ = depth of fish $i$ at time $t$ and $D_i$ = collection of all depth values for fish $i$

- Two-state Markov switching model with different GARCH - dynamics:

$$d_{it} = \begin{cases} \beta_{i1} \sqrt{h_{it}} + \sqrt{h_{it}} \epsilon_{it}, & h_{it} = \alpha_{i0} + \alpha_{i1} h_{i,t-1} + \alpha_{i2} a_{i,t-1}^2, \text{ if } s_{it} = 1; \\ \beta_{i2} \sqrt{h_{it}} + \sqrt{h_{it}} \epsilon_{it}, & h_{it} = \alpha_{i20} + \alpha_{i21} h_{i,t-1} + \alpha_{i22} a_{i,t-1}^2, \text{ if } s_{it} = 2, \end{cases}$$

where $a_{it} = \sqrt{h_{it}} \epsilon_{it}$, $\{\epsilon_{it}\}$ is a sequence of standard normal white noise random variables and the parameters $\alpha_{ijk}$ satisfy some regularity conditions so that the unconditional variance of $a_{it}$ exists.
The probability that a fish transitions from one state to another, is governed by the following transition probabilities

\[
P(s_{it} = 2 | s_{i,t-1} = 1) = e_{i1},
\]
\[
P(s_{it} = 1 | s_{i,t-1} = 2) = e_{i2}
\]

where \(0 < e_{ij} < 1\) for \(j = 1, 2\)

Small values of \(e_{ij}\) indicate that fish \(i\) has a tendency to stay in the \(j^{th}\) state with expected duration \(\frac{1}{e_{ij}}\)
Estimating the Model

- Bayesian method using Gibbs sampling approach

- Assume $h_{i1}$ and equal to the sample variance of $d_{it}$. The effect of this assumption is negligible when the sample size is large.

- Parameters to estimate:
  - $\beta_{i1}, \beta_{i2}$
  - $\alpha_{i0}, \alpha_{i11}, \alpha_{i12}, \alpha_{i20}, \alpha_{i21}, \alpha_{i22}$
  - $e_{i1}, e_{i2}$
  - state vector $S_i = (s_{i1}, s_{i2}, \ldots, s_{in_i})$
  - volatility vector $H_i = (h_{i2}, \ldots, h_{in_i})$
Estimating the Model - Prior Distributions

- **Gibbs sampling approach** - only the following conditional posterior distributions are needed:
  - \( f(\beta_i|D_i, S_i, H_i, \alpha_{i1}, \alpha_{i2}) \)
  - \( f(\alpha_{i1}|D_i, S_i, H_i, \alpha_{i2}) \)
  - \( f(\alpha_{i2}|D_i, S_i, H_i, \alpha_{i1}) \)
  - \( P(S_i|D_i, h_{i1}, \alpha_{i1}, \alpha_{i2}) \)
  - \( f(e_{i1}, e_{i2}|S_i) \)

- For simplicity, we impose **conjugate priors** for \( \beta_{ij} \) and \( e_{ij} \) \((j = 1, 2)\):
  - \( \beta_{ij} \sim N(\beta_{j0}, \sigma_{j0}^2) \), for \( j = 1, 2 \)
  - \( e_{ij} \sim \text{Beta}(\delta_{j1}, \delta_{j2}) \), for \( j = 1, 2 \)

- The prior distribution of \( \alpha_{ijk} \) is uniform over a properly specified interval
The posterior distribution of $\beta_{ij}$ ($j = 1, 2$) only depends on the data in state $j$.

For ($j = 1, 2$), let

$$
d_{it}^{(j)} = \frac{d_{it}}{\sqrt{h_{it}}} \text{ if } s_{it} = j \text{ and } 0 \text{ otherwise,}
$$

$$
\bar{d}_{it}^{(j)} = \sum_{s_{it}=1} d_{it}^{(j)}
$$

$$
n_{ij} = \text{number of data points in state } j \text{ for fish } i
$$

Then the conditional posterior distribution of $\beta_{ij}$ is

$$
\beta_{ij} \sim N\left(\sigma_{ij}^2 \left(n_{ij} \bar{d}_{it}^{(j)} + \beta_{j0}/\sigma_{j0}^2\right), \sigma_{ij}^2\right),
$$

where $\frac{1}{\sigma_{ij}^2} = n_{ij} + \frac{1}{\sigma_{j0}^2}$.
Posterior Distribution of $\alpha_{ijk}$

- In order to draw realizations of $\alpha_{ijk}$ we use the Griddy Gibbs Method - Ritter and Tanner (1992)
- Given $h_{i1}, S_i$, all other elements in $\alpha$, we have that

$$f(\alpha_{ijk}|\cdot) \propto -\frac{1}{2} \left( \log h_{it} + \frac{(d_{it} - \beta_{ij} \sqrt{h_{it}})^2}{h_{it}} \right), \quad \text{if } s_{it} = j$$

- Evaluate this function at a grid of points for $\alpha_{ijk}$ over a properly specified interval.
- Define the following
  - $m_{i1} =$ the number of switches from state 1 to state 2
  - $m_{i2} =$ the number of switches from state 2 to state 1
- posterior distribution of $e_{ij} \sim \text{Beta}(\delta_{j1} + m_{ij}, \delta_{j2} + n_{ij} - m_{i1})$
Posterior Distribution of $S_i$

- Elements of $S_i$ drawn one by one
- Let $S_i^{(-l)}$ be the vector obtained by removing $s_{il}$ from $S_i$
- Given $S_i^{(-l)}$ and other information, the conditional posterior distribution of $s_{il}$ is

$$P(s_{il} \mid \cdot) \propto \prod_{t=l}^{n_i} (a_{it} \mid H_i) P(s_{il} \mid S_i^{(-l)}).$$

- $L(s_{il} = j) \equiv \prod_{t=l}^{n_i} f(a_{it} \mid H_i) \propto \exp(f_{ilj})$, where

$$f_{ilj} = \sum_{t=l}^{n_i} - \frac{1}{2} \left( \ln h_{it} + \frac{a_{it}^2}{h_{it}} \right).$$

and $a_{it} = d_{it} - \beta_{ij} \sqrt{h_{it}}$ if $s_{it} = j$ for $j = 1, 2$

- Finally, the conditional posterior probability of $s_{il} = j$ is

$$P(s_{il} = j \mid \cdot) = \frac{P(s_{il} = j \mid s_{i,l-1}, s_{i,l+1}) L(s_{il} = j)}{P(s_{il} = 1 \mid s_{i,l-1}, s_{i,l+1}) L(s_{il} = 1) + P(s_{il} = 2 \mid s_{i,l-1}, s_{i,l+1}) L(s_{il} = 2)}$$

- Therefore state $s_{il}$ can be drawn from a Uniform(0,1) distribution
Estimating the Regression Model - Priors

- Let \( z = \text{logitratio} = X\gamma + \epsilon \) denote our regression in matrix notation, with \( \epsilon \sim N(0, \sigma_\epsilon^2 I) \)

- Additionally, let \( \gamma = \mu + \nu \) with \( \nu \sim N(0, V_\gamma) \)

- Following Rossi, Allenby, and McCulloch (2005) we impose the following conjugate priors:
  - \( \sigma_\epsilon^2 \sim \nu_\epsilon s_0^2/\chi^2_{\nu_\epsilon} \)
  - \( V_\gamma \sim IW(\nu, V) \)
  - \( \mu|V_\gamma \sim N(\mu, V_\gamma \otimes A^{-1}) \) where \( \mu = 0 \) is a matrix of prior means and \( A = .01I \) is a matrix for prior precision
  - \( \nu_\epsilon \) is a degree of freedom (df) parameter for \( \sigma_\epsilon^2 \) defined equal to 3
  - \( \nu \) is the df parameter for \( V_\gamma \) defined equal to the number of variables plus 3
We use a **Gibbs sampling** technique to first draw \((\gamma, \sigma_\epsilon^2)\) given the parameters of the first stage prior, \(\mu, V_\gamma\), and then draw the prior parameters conditional on \((\gamma, \sigma_\epsilon^2)\).

The **posterior distribution** for the regression parameters of interest is

\[
\gamma | z, X, \mu, V_\gamma, \sigma_\epsilon^2 \sim N \left\{ \gamma^*, (X^* X^* + V_\gamma^{-1})^{-1} \right\}
\]

where

\[
\gamma^* = (X^* X^* + V_\gamma^{-1})^{-1} (X^* z^* + V_\gamma^{-1} \mu),
\]

with \(z^* = z / \sigma_\epsilon\) and \(X^* = X / \sigma_\epsilon\).
Parallel Computing

- Large number of Markov switching stochastic volatility models required in estimation (1 per fish)

- These models are computationally expensive due to the grid estimation technique (high dimensional grid)

- Good news: Each model is independent!

- Parallel computation using Rmpi in R
Parallel Computation Algorithm

Steps in algorithm

1. Master processor generates information for joint model (which is conditional on the previous iteration).
2. Master processor reports this information to each slave.
4. Master collect estimates of the fish specific eigenvalues and uses them in estimation of lower level parameters.

Computational cost benefit

- 1 iteration in serial: 57.7 minutes
- 1 iteration in parallel: 1.7 minutes
- 97% reduction in computing time!
Alternative models

▶ OLS Linear Regression Model without eigenvalue predictor

\[ \text{logitratio}_i = \gamma_0 + \gamma_1 wt_i + \gamma_2 sl_i + \gamma_3 Pl_i + \gamma_4 cape2_i + \gamma_5 cap11kt_i + \gamma_6 capc_i + \epsilon_i \]

▶ Hierarchical Bayesian Regression Model 1

\[ \text{logitratio}_i = \gamma_0 + \gamma_1 wt_i + \gamma_2 sl_i + \gamma_3 Pl_i + \gamma_4 cape2_i + \gamma_5 cap11kt_i + \gamma_6 capc_i + \gamma_7 \lambda_i + \epsilon_i \]

▶ Hierarchical Bayesian Regression Model 2

\[ \text{logitratio}_i = \gamma_0 + \gamma_1 wt_i + \gamma_2 Pl_i + \gamma_3 cape2_i + \gamma_4 cap11kt_i + \gamma_5 capc_i + \gamma_6 \lambda_i + \epsilon_i \]

▶ Hierarchical Bayesian Regression Model 3

\[ \text{logitratio}_i = \gamma_0 + \gamma_1 wt_i + \gamma_2 \lambda_i + \epsilon_i \]
OLS Linear Regression Model without eigenvalue parameter estimates predictor

| Parameter | Estimate | Std. Error | Pr(>|t|) |
|-----------|----------|------------|---------|
| Intercept | -8.2099  | 21.4056    | 0.706   |
| wt        | 0.0083   | 0.0062     | 0.198   |
| sl        | -0.0081  | 0.0471     | 0.866   |
| Pl        | 8.1889   | 15.6906    | 0.608   |
| cape2     | -0.0004  | 0.0003     | 0.224   |
| cap11kt   | 0.0003   | 0.0005     | 0.572   |
| capc      | -0.0557  | 0.0498     | 0.278   |
Preferred Hierarchical Bayesian Regression Model

Model 3 was preferred based on:

1. exploratory analysis of 95% credible intervals of model parameters in a fully saturated model,
2. underlying biological considerations supplied by expert fisheries biologists,
3. DIC (Note: DIC(Model 1)=533.35, DIC(Model 2)=532.46),
4. Model 3 has lowest mean squared error as well as best in-sample classification

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>Posterior Std. Dev.</th>
<th>95% Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.180</td>
<td>0.989</td>
<td>(-1.44,1.80)</td>
</tr>
<tr>
<td>wt*</td>
<td>0.00268</td>
<td>0.00130</td>
<td>(0.000582,0.00485)</td>
</tr>
<tr>
<td>eigenvalue*</td>
<td>-0.851</td>
<td>0.186</td>
<td>(-1.16,-0.546)</td>
</tr>
</tbody>
</table>

Note: In all Bayesian models eigenvalue was significant and in Models 2 and 3 wt was significant as well.
Comparison of Confusion Matrix for OLS regression model and the hierarchical Bayesian Model 3

<table>
<thead>
<tr>
<th>Actual (Model)</th>
<th>Predict (Model) Successful Spawn</th>
<th>Predict (Model) Unsuccessful Spawn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Successful Spawner (OLS)</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>Unsuccessful Spawner (OLS)</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Successful Spawner (Bayesian M3)</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>Unsuccessful Spawner (Bayesian M3)</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>
Probability estimates of being in the low variability regime

**Spawner**

**Non-spawner**
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Non-spawner Posterior Mean (95% CI)</th>
<th>Spawner Posterior Mean (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low Variability Regime</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{10}$</td>
<td>1.33 (0.714, 2.09)</td>
<td>1.30 (0.0781, 1.69)</td>
</tr>
<tr>
<td>$\alpha_{11}$</td>
<td>0.292 (0.174, 0.396)</td>
<td>0.485 (0.340, 0.631)</td>
</tr>
<tr>
<td>$\alpha_{12}$</td>
<td>0.432 (0.242, 0.644)</td>
<td>0.140 (0.00795, 0.366)</td>
</tr>
<tr>
<td></td>
<td>High Variability Regime</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{20}$</td>
<td>0.025 (0.0144, 0.0395)</td>
<td>0.122 (0.0979, 0.157)</td>
</tr>
<tr>
<td>$\alpha_{21}$</td>
<td>0.569 (0.425, 0.722)</td>
<td>0.857 (0.679, 0.998)</td>
</tr>
<tr>
<td>$\alpha_{22}$</td>
<td>0.431 (0.355, 0.521)</td>
<td>0.105 (0.0343, 0.191)</td>
</tr>
</tbody>
</table>
Note that these posterior densities are disjoint and have endpoints as follows: (0.717, 0.884) for the non-spawners and (0.887, 0.998) for the spawners.
Histogram - Proportion of Time Each Fish Spent in the High Variability Regime

Non-spawners

Spawners

Note that we work with proportion of time rather than length of time because of unequal observation lengths due to different recapture times.
Summary

- Developed a Bayesian hierarchical model for predicting spawning success capable of utilizing Data Storage Tag data
- Model incorporates an eigenvalue predictor from the transition probability matrix in a two-state Markov switching model with GARCH dynamics as a generated regressor in a linear regression model
- Outperforms model without DST information
- OLS model is insufficient: does not find relevance of weight to spawning success
- Our results support the hypothesis that spawners exhibit lower levels of depth variability in their swimming pattern during the spawning season
- Clear distinction between depth variability (95% CIs of eigenvalue estimates) in spawners and non-spawners
- Computationally expensive, but cost minimized using parallel computing
Future Work

- Incorporate temperature profile into model
  - The fact that there is a “preferred” temperature might help predict spawning occurrence
- Evaluate importance of upstream or downstream movement for spawning prediction
- Further data collection for developing model for not only the occurrence, but the timing of spawning - currently in progress