On the Ugliness of Ecological Monitoring

Computational Constraints Arising from Ecological Data and Inference Methods

James D. Nichols
Patuxent Wildlife Research Center

Barroom Messages

- All data are not created equal.
- Computational methods for one discipline cannot necessarily be transferred to another.

Presentation Focus

Consequences of ecological data and inference methods for selection of computational approaches.

Outline

- How to monitor?
  - Spatial variation
  - Detection probability
  - Example: occupancy modeling
- Why monitor?
  - Science: stochastic dynamic optimization for learning
  - Management: stochastic dynamic optimization for making smart decisions
- What to monitor?
  - Selection of system components to provide information about entire system
    - Attractor-based methods
    - Information-theoretic methods
- Summary

How to Monitor?
Basic Sampling Issues

- Geographic variation
  - Frequently counts/observations cannot be conducted over entire area of interest
  - Proper inference requires a spatial sampling design that:
    - Permits inference about entire area, based on a sample, and/or
    - Provides good opportunity for discriminating among competing hypotheses

How to Monitor?
Basic Sampling Issues

- Detectability
  - Counts represent some unknown fraction of animals in sampled area
  - Proper inference requires information on detection probability

Detectability: Monitoring Based on Some Sort of Count

- Ungulates seen while walking a line transect
- Tigers detected with camera-traps
- Birds heard at point count
- Small mammals captured on trapping grid
- Bobwhite quail harvested during hunting season
- Kangaroos observed while flying aerial transect
- Number of locations at which a species is detected

Detectability: Conceptual Basis

- N = abundance
- C = count statistic
- p = detection probability; P(member of N appears in C)

$$E(C) = pN$$

Detectability: Inference

- Inferences about *N* (and relative *N*) require inferences about *p*

$$\hat{N} = \frac{C}{\hat{p}}$$

Inference from Ecological Data

WYSIWYG

(What You See Is What You Get)

Doesn’t Work in Ecology

Inference Example: Species Distribution and Habitat Relationships

- Basic field situation: single season
  - From a population of S sampling units, s are selected and surveyed for the species.
  - Units are closed to changes in occupancy during a common 'season'.
  - Units must be repeatedly surveyed within a season.

Single Season: Data

- Obtain detection history data for each site visited
- Possible detection histories, 3-visits:

  101 000

- Key issue for inference: ambiguity of 000

  (1) absence or

  (2) presence with nondetection

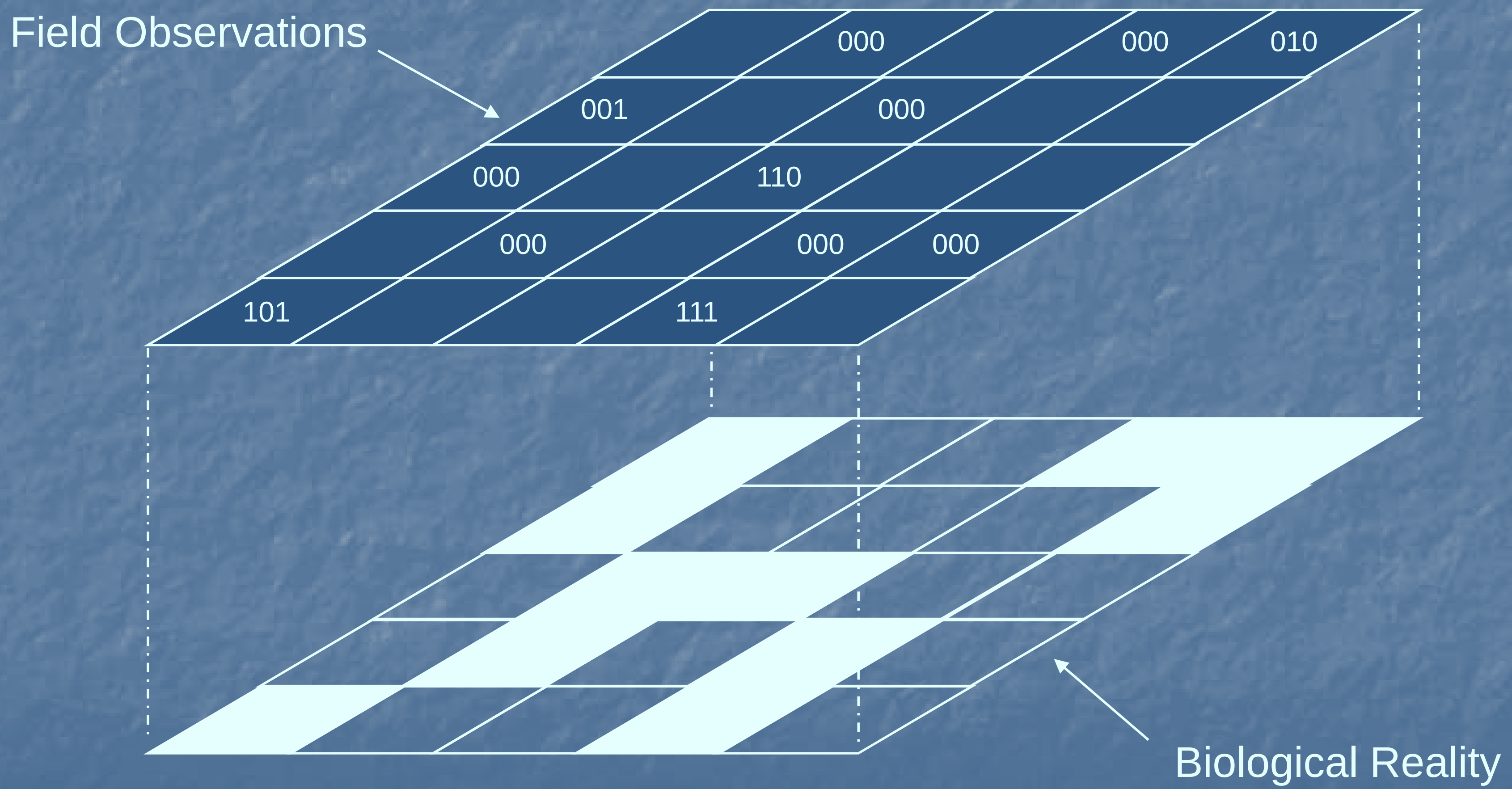
Single Season Model

- Consider the data as consisting of 2 'layers'
  1. True presence/absence of the species.
  2. Observed data, conditional upon species distribution.

- Knowledge about the first layer is imperfect.

- Must account for the observation process to make reliable inferences about occurrence.

Model Development



Single Season Model Parameters

- ψ = probability a unit is occupied.
- p_j = probability species is detected at a unit in survey j (given presence).

Single Season Modeling

- Basic idea: develop probabilistic model for process that generated the data

$$\Pr\left(\mathbf{h}_1 = 101\right) = \psi p_1 \left(1 - p_2\right) p_3$$

$$\Pr\left(\mathbf{h}_2 = 000\right) = \psi \prod_{j=1}^{3} \left(1 - p_j\right) + \left(1 - \psi\right)$$

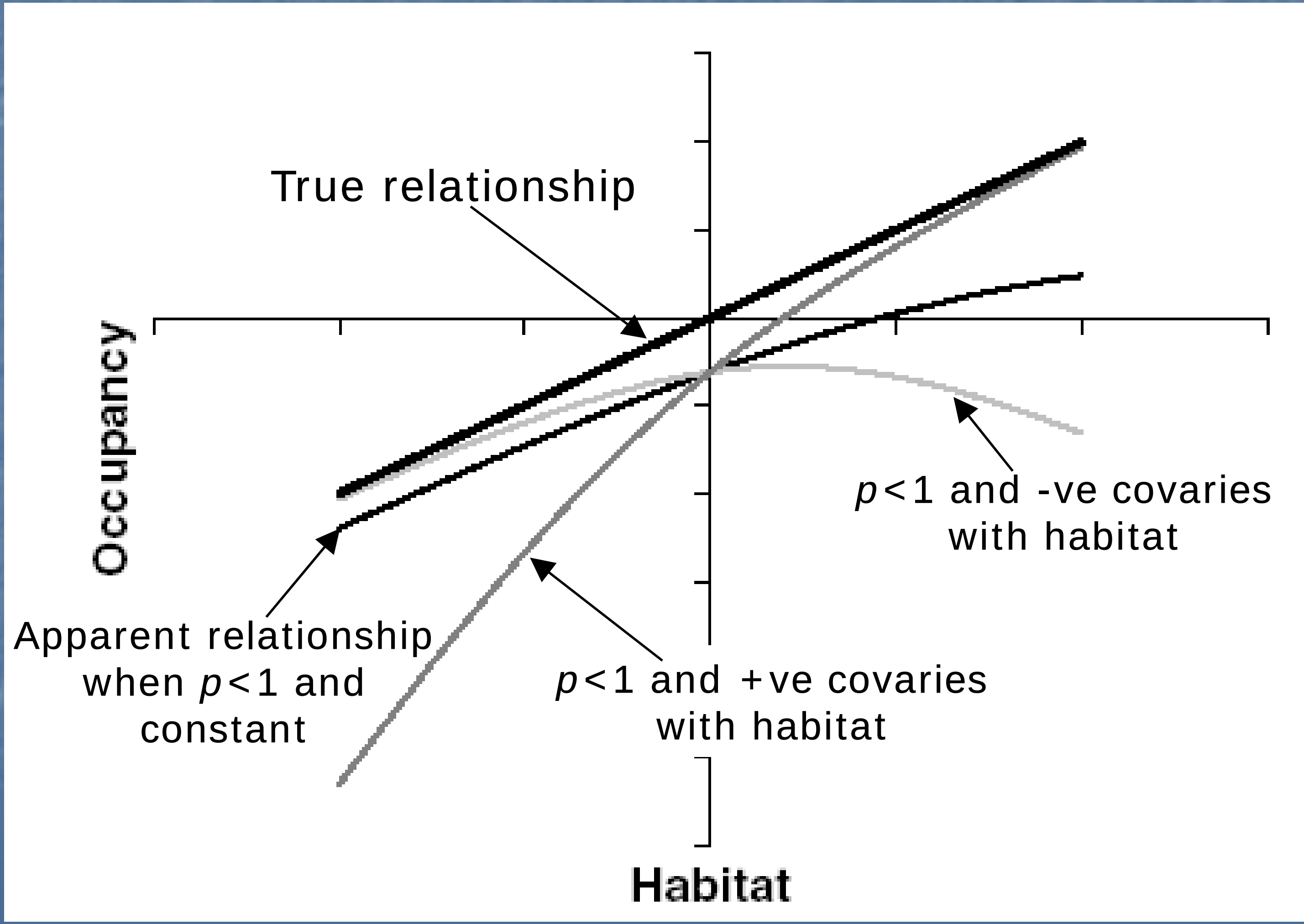
Single Season Model: Inference

- Given:
  - (1) detection history data for each site,
  - (2) probabilistic model for each detection history
- Inference:
  - Maximum likelihood
  - State space approach (e.g., hierarchical Bayes implemented using MCMC)
- Relevance to computations using estimates:
  - Estimates (e.g., of occupancy) have non-negligible variances and covariances
  - Typically, cov $(\hat{\psi}_t, \hat{\psi}_{t+1}) \neq 0$

Detection Probability and Occupancy: Why Bother?

- Methods that ignore *p*<1 produce:
  - Negative bias in occupancy estimates
  - Positive bias in estimates of local extinction
  - Biased estimates of local colonization
  - Biased estimates of incidence functions and derived parameters
  - Misleading inferences about covariate relationships

Habitat Relationships and Resource Selection



Inference Example: Species Distribution & Habitat Relationships

- Geographic variation and detection probability are not statistical fine points
- They must be dealt with for proper inference
- Proper inference methods yield estimates (e.g., of occupancy) that have non-negligible variances and covariances
- Computational algorithms (e.g., for dynamic optimization) that use such estimates must deal with this variance-covariance structure resulting from ecological sampling

Why Monitor?

- Monitoring is not a stand-alone activity but is most useful as a component of a larger program
- (1) Science
  - Understand ecological systems
  - Learn stuff
- (2) Management/Conservation
  - Apply decision-theoretic approaches
  - Make smart decisions

Key Step of Science: Confront Predictions with Data

- Deduce predictions from hypotheses
- Observe system dynamics via monitoring
- Confrontation: Predictions vs. Observations
  - Ask whether observations correspond to predictions (single-hypothesis)
  - Use correspondence between observations and predictions to help discriminate among hypotheses (multiple-hypothesis)

Single-Hypothesis Approach to Science

- Develop hypothesis
- Use model to deduce testable prediction(s), typically relative to a null hypothesis
- Carry out suitable test
- Compare test results with predictions (confront model with data)
- Reject or retain hypothesis

Multiple-Hypothesis Approach to Science

- Develop set of competing hypotheses
- Develop/derive prior probabilities associated with these hypotheses
- Use associated models to deduce predictions
- Carry out suitable test
- Compare test results with predictions
- Based on comparison, compute new probabilities for the hypotheses

Single Hypothesis Science & Statistics: Historical Note

- Much of modern experimental statistics seems to have been heavily influenced by single-hypothesis view of science
- Fisherian experimental design
  - Emphasis on expectations under H_0 (replication, randomization, control)
  - Objective function for design: maximize test power within hypothesis-testing framework
- Result: statistical inference and design methods
  - Well-developed for:
    - single-hypothesis approaches
    - single experiments
  - Not well-developed for:
    - multiple hypothesis approaches
    - accumulation of knowledge for sequence of experiments

Science and the Accumulation of Knowledge

- Science has long been viewed as a progressive enterprise
  - “I hoped that each one would publish whatever he had learned, so that later investigations could begin where the earlier had left off.” (Descartes 1637)
- How does knowledge accumulate in single- and multiple-hypothesis science?

Accumulation of Knowledge

- No formal mechanism under single hypothesis science
- *Ad hoc* approach: develop increased faith in hypotheses that withstand repeated efforts to falsify
- Popper's (1959, 1972) "Natural Selection of Hypotheses" analogy
  - Subject hypotheses to repeated efforts at falsification: some survive and some don't

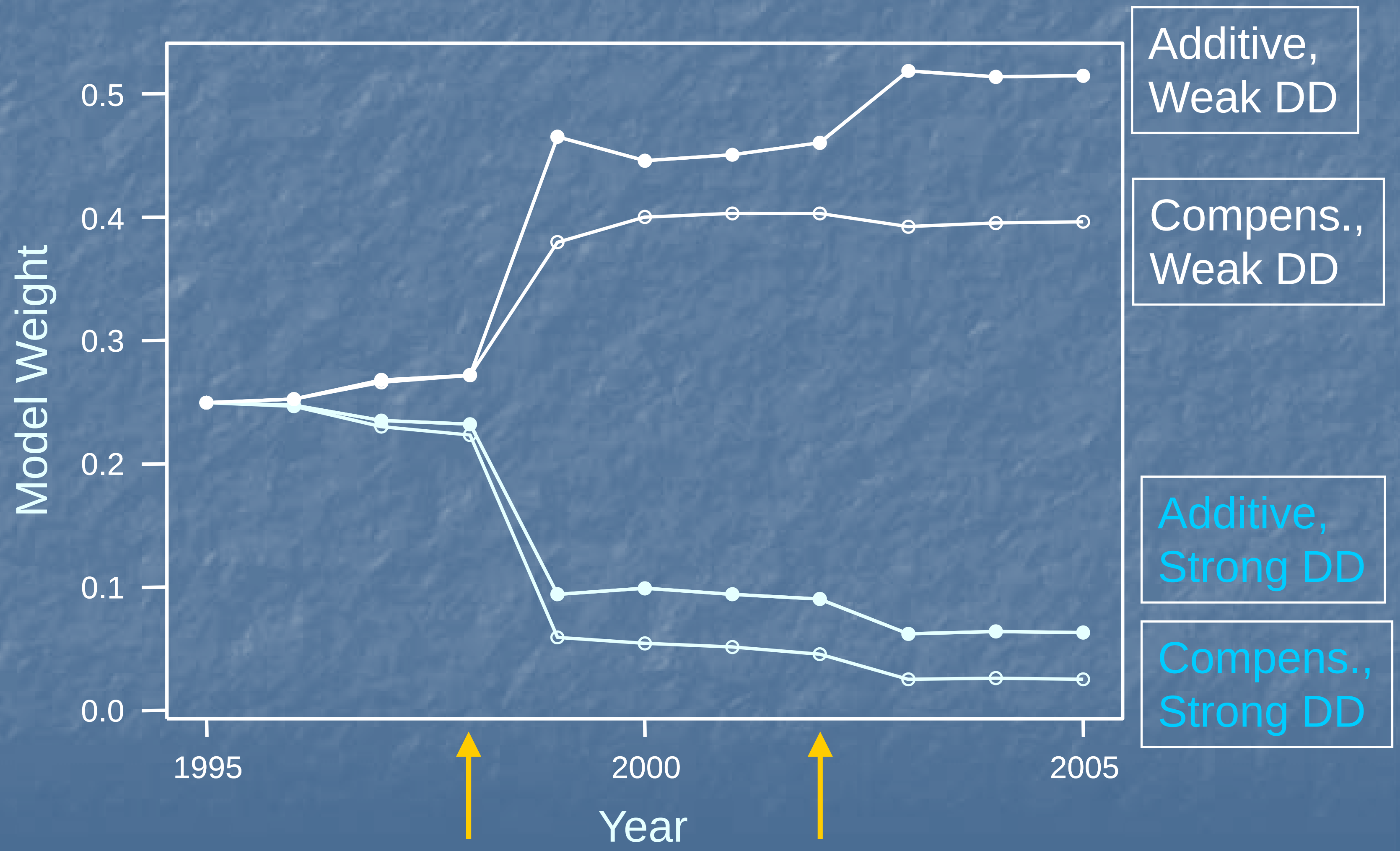
Accumulation of Knowledge

- Mechanism built directly into multiple hypothesis approach
- Model probabilities updated following each study, reflecting changes in relative degrees of faith in different models
  - "Natural Selection of Hypotheses": view changes in model probabilities as analogous to changes in gene frequencies
- Formal approach under multiple hypothesis science based on Bayes' Theorem

Updating Model Probabilities: Bayes' Formula

$$p_{t+1}(\text{model } i \mid \text{data}_{t+1}) = \frac{p_t(\text{model } i)\, P(\text{data}_{t+1} \mid \text{model } i)}{\sum_j p_t(\text{model } j)\, P(\text{data}_{t+1} \mid \text{model } j)}$$

Adaptive Harvest Management



Study Design Considerations: Multiple Hypothesis Science

- Envisage a sequence of studies or manipulations
- Make design decisions at each time *t*, depending on the information state (model probabilities) at time *t*
- When studies are on natural populations, design decisions will likely also depend on system state (e.g., population size)

Study Design Considerations: Multiple Hypothesis Science

- Proposal (Kendall): use methods for optimal stochastic control (dynamic optimization) to aid in aspects of study design (e.g., selection of treatments) at each step in the program of inquiry
- Objective function focuses on information state, the vector of probabilities associated with the different models

Objective Functions for Learning Over T Experiments

- Maximize sum of squares of posterior model probabilities (likelihood)
  - Same as minimizing Simpson's index

Choose decision that max $\sum_{t=1}^{T} \sum_{i=1}^{\#\text{mod}} p_{i,t+1}^{2}$

- Minimize Shannon-Wiener index

Choose decision that

min $-\sum_{t=1}^{T} \sum_{i=1}^{\#\text{mod}} p_{i,t+1} \log_2(p_{i,t+1})$

Dynamic Optimization: Computational Issues

- Partial observability
  - sampling variances and covariances
- Problem dimension
  - limited number of state variables,
  - limited categories for discretizing state variables, etc.

Management/Conservation: Key Step in Process

- Monitoring provides estimates of system state for state-dependent decisions

- Dynamic optimization uses these estimates, together with objectives, available actions and models to yield optimal decisions

Dynamic Optimization: Computational Issues

- Partial observability
  - sampling variances and covariances
- Problem dimension
  - limited number of state variables,
  - limited categories for discretizing state variables, etc.
- Order of Markov process
  - Higher order processes characterize some ecological systems (e.g., 10-year maturation time for horseshoe crabs)
- Nonstationarity of Markov process
  - Climate change
  - Human activities and associated land-use changes

What to Monitor?

- Answer is inherited from answer to "Why?" question
- Straightforward for small (1-3) number of species
- What about focus on an ecological system with many components (e.g., species x location subpopulations)?

What to Monitor in Ecological Systems?

- We can’t monitor all populations of all species everywhere in a large system
- How do we select species x location components that provide more information about system dynamics
- Relevant to ideas about “indicator” species *and* locations.

Dynamical Interdependence

- Data: time series of 2 (or more) different state variables
- Question: what can we learn about 1 (or more) state variable by following another?
- Ecological applications:
  - Monitoring program design (indicator species, indicator locations, etc.)
  - Population synchrony and its cause(s)
  - Food web connectance
  - Competitive interactions

Dynamical Interdependence: Nonlinear Systems

- Attractor-based methods
  - If 2 state variables are dependent and belong to same system, then by Takens (1981) embedding theorem, their attractors should exhibit similar geometries
  - Continuity: focus on function relating 2 attractors
- Information-based methods
  - Mutual information
  - *Transfer entropy*

Example Method: Transfer Entropy

- Consider a Markov process in which value of random variable, *Y*, at any time depends on past values (*k* time units into the past)
- Consider another possible system variable, *Z*, and ask whether it contributes information about *Y*
- Absence of information flow from *Z* to *Y*:

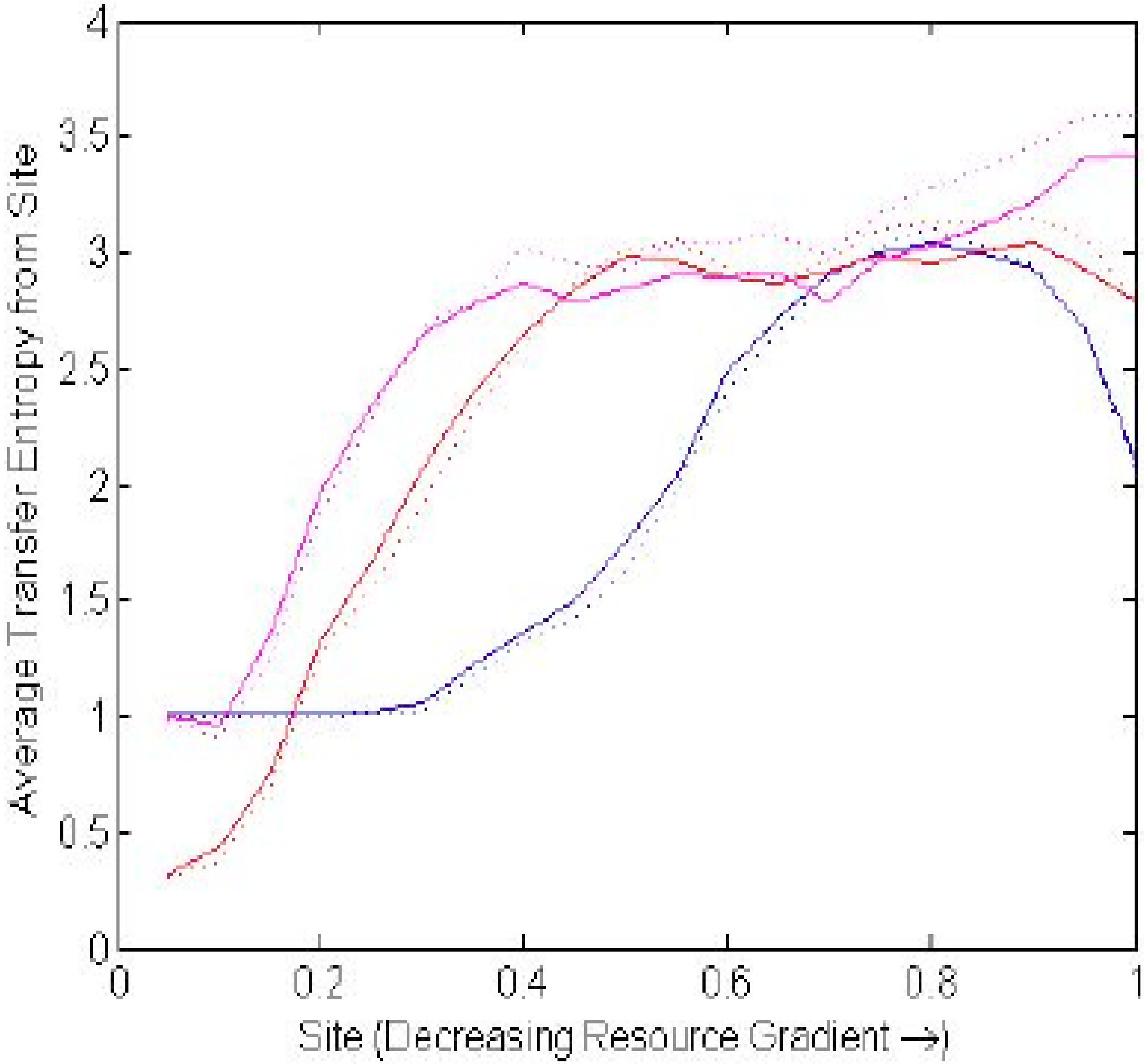
$$p(y_{t+1} \mid y_t^{(k)}) = p(y_{t+1} \mid y_t^{(k)}, z_t^{(l)})$$

Transfer Entropy (Schreiber 2000)

- Transfer Entropy, $T_{Z \to Y}$, measures the extra information about transitions of *Y* obtained by knowing *Z*
- Transfer Entropy is not symmetric
- Transfer Entropy is a Kullback entropy that focuses on the deviation of the process from the generalized Markov property

$$T_{Z \to Y} = \sum_{yz} p(y_{t+1}, y_t^{(k)}, z_t^{(l)}) \log_2 \frac{p(y_{t+1} \mid y_t^{(k)}, z_t^{(l)})}{p(y_{t+1} \mid y_t^{(k)})}$$

L.J. Moniz, J.D. Nichols, J.M. Nichols, E.G.Cooch, and L.M.Pecora



Computational Methods for Inference About Dynamical Interdependence

- Both attractor-based and information-based (e.g., transfer entropy) approaches are usually computed assuming stationarity and using:
  - Long time series
  - Direct observations with no sampling variances-covariances
- Example, the probability distributions for transfer entropy are developed using binning approach

Computational Methods for Inference About Dynamical Interdependence

- Many of these methods not yet ready for ecological prime-time
- Approaches to nonlinear analysis of time series that are noisy, nonstationary and short include:
  - surrogate data sets for bootstrap-type approach to inference
  - kernel density estimation approaches instead of “bin counting”
  - use of symbolic dynamics
  - information-based approaches for deterministic signal extraction in the presence of noise

On the Ugliness of Ecological Monitoring: Summary

- Inference from ecological monitoring data requires methods that deal with geog. variation & detection probability
  - WYSIWYG won't work!
- These inference methods have been well-developed, *but* resulting estimates are typically few and characterized by sampling variance-covariance structures

On the Ugliness of Ecological Monitoring: Summary

- Many ecological processes are also characterized by relatively high dimension and dynamics are governed by higher order Markov processes
- Some algorithms that would be especially useful to ecologists (dynamic optimization, attractor- and information-based approaches to assessing coupling) were not designed with such data and processes in mind