Inference for finite-sample trajectories in dynamic multi-state site-occupancy models using hidden Markov model smoothing

Ian J. Fiske · J. Andrew Royle · Kevin Gross

Abstract Ecologists and wildlife biologists increasingly use latent variable models to study patterns of species occurrence when detection is imperfect. These models have recently been generalized to accommodate both a more expansive description of state than simple presence or absence, and Markovian dynamics in the latent state over successive sampling seasons. In this paper, we write these multi-season, multi-state models as hidden Markov models to find both maximum likelihood estimates of model parameters and finite-sample estimators of the trajectory of the latent state over time. These estimators are especially useful for characterizing population trends in species of conservation concern. We also develop parametric bootstrap procedures that allow formal inference about latent trend. We examine model behavior through simulation, and we apply the model to data from the North American Amphibian Monitoring Program.

Handling Editor: Pierre Dutilleul.

I. J. Fiske · K. Gross (✉)
Department of Statistics, North Carolina State University, Raleigh, NC, USA
e-mail: kevin_gross@ncsu.edu

J. A. Royle
USGS Patuxent Wildlife Research Center, Beech Forest Road, Laurel, MD 12100, USA

1 Introduction

Modelling species occurrence is a fundamental objective of many ecological studies. At its most basic, species occurrence is the binary state variable that describes whether a study species is present or absent at each of a collection of sites. Statistical inference can focus either on the expected occurrence rate over all sites or on assessing which site-level variables may be associated with occurrence. The predominant challenge in the development of occurrence models is that detection is usually imperfect. A record of "presence" can either result from an accurate detection of a present species or a false detection of a species that is actually absent (perhaps by mistaking a closely related species). Similarly, a record of "absence" can either result from an accurate detection of absence or from failure to detect a species that is present. A standard approach to dealing with imperfect detection is to revisit a site multiple times within a single "season" and record presence or absence at each sampling occasion. One then assumes that (a) the true population status does not change within a single season, (b) visits are independent both within and among sites, (c) probabilities of mis-detection are constant across sites, and (d) the probability of falsely detecting a species that is truly absent is negligible compared to the probability of failing to detect a species that is present. Although this latter assumption may seem strong from a statistical perspective, it is broadly accepted in wildlife biology because of the elusive nature of many species. With these assumptions, one can then develop statistical models to simultaneously estimate rates of both occurrence and detection (MacKenzie et al. 2002, 2006).

General interest in species occurrence and the simplicity of associated sampling designs have led to several extensions of this basic idea (e.g., Royle and Link 2005; Royle and Kéry 2007; MacKenzie et al. 2006). One important extension allows the latent state to belong to more than two categories (e.g., Royle and Link 2005; Nichols et al. 2007). For example, in avian studies, sites in which the species is present might be more finely classified by whether or not the site contains a breeding pair. A second extension considers multiple consecutive seasons and allows the latent state at each site to change from one season to the next (MacKenzie et al. 2003; Royle and Kéry 2007). In these models, the latent state at a site is assumed to change from season to season according to a Markov model, i.e., the future fate of a population at a given site depends only on the population's current status and is independent of the population's past history conditional on the current status. Recently, MacKenzie et al. (2009) used a Bayesian formulation to publish the first model that synthesizes both the multi-state and multi-season extensions into a general dynamic multi-state model.

In this article, we develop multi-state, multi-season occupancy models motivated by data from the North American Amphibian Monitoring Program (NAAMP; Weir et al. (2005)). NAAMP is a joint effort among state agencies and the United States Geological Survey begun in 2001 to assess the status of anuran (frog and toad) populations in over 20 states. Frogs and other amphibians are thought to be particularly vulnerable to land-use change, and thus biologists have keen interest in detecting changes in their abundance. In NAAMP, surveyors visit multiple sites several times each year and measure the presence and abundance of specific anuran species by listening for their calls. At each visit to each site, the intensity of calls for each species is recorded using an

ordinal variable with the following four levels: no calling detected (intensity = 0), single individuals can be counted, with space between calls (intensity = 1), multiple distinct individuals can be distinguished, with some overlapping of calls (intensity = 2), or a full chorus of constant, continuous and overlapping calls (intensity = 3). This measure of calling intensity is used because it is thought to correspond to the abundance of the particular species, and because more precise measures (such as counting the number of distinct calling individuals) are impossible when calls overlap. Previous work has analyzed a single year of NAAMP data (Royle and Link 2005). Multiple consecutive years of regular monitoring data are now available, and we are interested in using these data to characterize temporal patterns of occurrence and abundance. Thus, we require statistical models to conduct inference about latent transition rates and the trajectory of an ordinal occupancy variable over time. These models must handle a multinomial response with possible misclassification errors and a longitudinal sampling structure in which the same sites are revisited yearly.

Although our occupancy models share structural similarities with those of MacKenzie et al. (2009), our analytical approach differs in usefully complementary ways. Here, we make use of the fact that dynamic occurrence models with imperfect detection belong to a class of models termed hidden Markov models (HMMs). Formulating these models as HMMs allows us to make use of a wealth of literature on estimation, filtering, smoothing, and prediction already developed for the general hidden Markov model (Rabiner 1989; Cappé et al. 2005; Bartolucci et al. 2012). In this paper, we show how HMM methods can be used for frequentist inference on the actual fluctuations in the latent population state. Following the terminology of Link and Sauer (1997) and Royle and Kéry (2007), we call this finite-sample trajectory inference. Finite-sample trajectory inference may be of particular importance in conservation or management, where legal classification of species' conservation status and consequent management actions can hinge on fluctuations in species abundance. One statistical concern with multi-state, multi-season occupancy models is that they can become highly parameterized as the number of latent states increases. Our models make the important assumption that misclassification only takes the form of under-detection, and that over-detection is negligible (Royle and Link 2005). This assumption is the logical multi-state extension of the conventional presence/absence assumption that false detections are rare, and is commonly supported by biologists.

Importantly, the adoption of HMM methods opens the door to a full suite of inferential procedures for the finite-sample trajectory. Towards this end, in addition to using HMM smoothing for estimation, this article also presents a parametric bootstrap procedure for estimating the sample variance of the finite-sample trajectory estimator. The bootstrap uses the Viterbi algorithm (Rabiner 1989) to identify the most probable latent trajectory at each site. The Viterbi-based bootstrap is also used to formally test for a temporal trend in univariate summaries of population abundance.

The rest of this article is structured as follows. In Sects. 2.1 and 2.2, we describe the model and maximum-likelihood procedures for estimating its parameters. Section 2.3 develops the finite-sample trajectory estimator. A parametric bootstrap procedure for the estimator's sampling variance is introduced in Sect. 2.4, and Sect. 2.5 develops tests for trends in univariate summaries of population abundance. A simulation study of the finite-sample trajectory estimator is presented in Sect. 3. Simulation studies

detailing the behavior of the bootstrap error estimation and trend tests can be found in the supplementary material. Finally, Sect. 4 estimates trajectory and tests for trends in the abundance of *Pseudacris crucifer* (Spring Peeper).

2 Dynamic multi-state hidden Markov models

2.1 Preliminaries

The underlying framework of the multi-year, multi-state occurrence model is the finite space hidden Markov model (Rabiner 1989; Cappé et al. 2005; Zucchini and MacDonald 2009; Bartolucci et al. 2012), which we briefly summarize here in the context of species occurrence data. We suppose that sampling occurs at a sample of M spatial units ("sites") over the course of T consecutive seasons. (A note about terminology: In general, a sampling *season* is the period of time during which the latent population status at a site is assumed to remain unchanged. For anurans, there is one sampling season per calendar year.) At site i, during season t, a surveyor records the categorical observation $Y_{itj} \in \{1, 2, \ldots, K\}$ for replicate visits $j = 1, 2, \ldots, J_{it}$. The unobservable or partially observed true state of site i during year t is $N_{it} \in \{1, 2, \ldots, K\}$. Because the latent state N_{it} does not change across the J_{it} visits during a given season, no j index is required for the latent state. Although observations Y_{itj} and latent states N_{it} often have the same number of possible values, the HMM framework can accommodate scenarios where this is not the case. When both Y_{itj} and N_{it} do have the same number of values, and when both are ordinal, then $Y_{itj} < N_{it}$ represents under-detection, and $Y_{itj} > N_{it}$ represents over-detection. Here, we impose the common constraint that over-detection is impossible, i.e., $Y_{itj} \leq N_{it}$ (Royle and Link 2005).

The full data-generation model for Y_{itj} consists of process and observation components. The process model describes how the distribution of the latent state N_{it} evolves over time, and the observation model describes the distribution of the observed data Y_{itj} conditional on the latent state. With a Markovian assumption, the process model is fully described by the distribution of the initial latent state N_{i1}, and the conditional distribution of N_{it} given $N_{i,t-1}$. We take both of these distributions to be multinomials with $\Pr(N_{i1} = k) = \psi_{1k}$ for $k = 1, 2, \ldots, K$ and $\Pr(N_{it} = k | N_{i,t-1} = j) = \Phi_{jk}$. Thus, the evolution of N_{it} over time is a first-order Markov chain with the row-stochastic transition matrix $\boldsymbol{\Phi} = (\Phi_{jk})$ and initial state distribution vector $\boldsymbol{\psi}_1 = (\psi_{11}, \psi_{12}, \ldots, \psi_{1K})'$.

In the observation model, the observations Y_{itj} are modeled as conditionally independent realizations from a multinomial distribution that depends only on the corresponding latent state. In the simplest case, when detection probabilities are constant across seasons, sites, and visits, we set $q_{ky} = \Pr(Y_{itj} = y | N_{it} = k)$, and form the row-stochastic detection matrix $\boldsymbol{Q} = (q_{ky})$. Perfect detection would render $\boldsymbol{Q}$ equal to the identity matrix. To allow detection probabilities to vary across seasons, sites, or visits, we define $q_{ky}^{(itj)} = \Pr(Y_{itj} = y | N_{it} = k)$, and $\boldsymbol{Q}_{itj} = \left(q_{ky}^{(itj)}\right)$. Often, detection probabilities may be modeled as functions of known covariates (e.g., time of year, weather conditions, observer identity). To make this full generality clear, write these observation covariates as $\boldsymbol{x}_{itj}$, and let $\boldsymbol{Q}_{itj} = \boldsymbol{Q}(\boldsymbol{x}_{itj}; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a vector of

Table 1 Model parameters and notation

Notation or Parameter	Definition
$i = 1, \ldots, M$	Sites
$t = 1, \ldots, T$	Seasons (years)
$j = 1, \ldots, J_{it}$	Visits at site i in season t
$Y_{itj} \in \{1, 2, \ldots, K\}$	Observed data
$N_{it} \in \{1, 2, \ldots, K\}$	Latent state
$N_{it} = (I(N_{it} = 0), I(N_{it} = 1), \ldots, I(N_{it} = K))'$	Indicator vector for N_{it}
$\psi_{1k} = Pr(N_{i1} = k)$	Initial probability for latent state
$\boldsymbol{\psi}_1 = (\psi_{11}, \psi_{12}, \ldots, \psi_{1K})'$	Vector of ψ_{1k}
$\Phi_{jk} = \Pr(N_{it} = k \mid N_{i,t-1} = j)$	Transition probability among latent states
$\boldsymbol{\Phi} = (\Phi_{jk})$	Transition probability matrix
$q_{ky}^{(itj)} = \Pr(Y_{itj} = y \mid N_{it} = k)$	Detection probability
$\boldsymbol{Q}_{itj} = \left(q_{ky}^{(itj)} \right)$	Detection probability matrix
$\boldsymbol{x}_{itj}$	Detection covariates
$\boldsymbol{\beta}$	Regression parameters in detection model
$\boldsymbol{\psi}_t = E(N_{it})$	Population-level trajectory at season t
$\underline{\psi}_t^{(fs)} = M^{-1} \sum_{i=1}^{M} N_{it}$	Finite-sample trajectory at season t

parameters in the model for $\boldsymbol{Q}$ as a function of $\boldsymbol{x}_{itj}$. In the simulations and application below, we use multinomial logistic regression (Agresti 2002) to connect covariates to detection probabilities.

Table 1 provides a summary of model parameters and notation.

2.2 Parameter estimation

Define a state path for site i as a vector defining a sequence of possible latent states, $(N_{i1}, N_{i2}, \ldots, N_{iT})'$. Computation of the HMM likelihood for site i requires integrating over all possible latent state paths. Denote the vector of observations taken at site i, year t as $\boldsymbol{y}_{it} = (y_{it1}, y_{it2}, \ldots, y_{itJ_{it}})'$, and denote all observations for site i as $\boldsymbol{y}_i$. The naïve approach to calculating the likelihood for site i would be to explicitly sum over all paths using the law of total probability.

$$
\begin{aligned}
L_i(\boldsymbol{\Phi}, \boldsymbol{\psi}_1, \boldsymbol{Q} \mid \boldsymbol{y}_i) = \sum_{k_1} \sum_{k_2} \ldots \sum_{k_T} \Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid N_{i1} = k_1, \\
N_{i2} = k_2, \ldots, N_{iT} = k_T; \boldsymbol{Q}) \\
\times \Pr(N_{i1} = k_1, N_{i2} = k_2, \ldots, N_{iT} = k_T; \boldsymbol{\Phi}, \boldsymbol{\psi}_1). \quad (1)
\end{aligned}
$$

In Eq. 1 and throughout, $\boldsymbol{Q}$ is taken to be either $\boldsymbol{Q}$ itself or observation-model parameters $\boldsymbol{\beta}$ depending on whether or not $\boldsymbol{Q}$ depends on covariates. The number of terms in Eq. 1 grows rapidly and thus this calculation quickly becomes computationally

infeasible as either K or T become large. To avoid this problem, we use a computationally efficient method for computing the likelihood for a finite space HMM known as forward recursion (Baum and Egon 1967). The mathematics underpinning the forward recursion algorithm are described in depth elsewhere (e.g. Rabiner 1989; Bartolucci et al. 2012). Using forward recursion, the likelihood for site i can be written compactly as

$$L_i(\boldsymbol{\Phi}, \boldsymbol{\psi}_1, \boldsymbol{Q} \mid \boldsymbol{y}_i) = \boldsymbol{\psi}_1' \left[\prod_{t=1}^{T-1} \mathrm{diag}(\boldsymbol{p}_{it})\boldsymbol{\Phi} \right] \boldsymbol{p}_{iT} \qquad (2)$$

where $\boldsymbol{p}_{it}$ is a K-dimensional vector whose kth element is $\Pr(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid N_{it} = k)$, and $\mathrm{diag}(\boldsymbol{v})$ is the diagonal matrix formed by the vector $\boldsymbol{v}$. Assuming independence among replicate visits to site i during year t, the elements of $\boldsymbol{p}_{it}$ are given by

$$\Pr(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid N_{it} = k) = \prod_{j=1}^{J_{it}} \Pr(Y_{itj} = y_{itj} | N_{it} = k) = \prod_{j=1}^{J_{it}} q_{k, y_{itj}}^{(itj)}. \qquad (3)$$

MacKenzie et al. (2009) provide a formulation of the likelihood that is equivalent to Eq. (2).

Maximum likelihood estimates $(\hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{Q}})$ are then found by numerically minimizing $-\sum_i (\log L_i(\boldsymbol{\Phi}, \boldsymbol{\psi}_1, \boldsymbol{Q} \mid \boldsymbol{y}_i))$. In the HMM literature, the Baum-Welch algorithm (an expectation-maximization procedure) is commonly used to obtain maximum likelihood estimates (Baum et al. 1970; Welch 2003). Although the Baum-Welch algorithm can accommodate observation covariates (Bartolucci et al. 2012), our particular example used multinomial logistic regression to incorporate observation covariates, thus necessitating a numerical solution to the m-step. Consequently, we found maximum likelihood estimates using the flexible Broyden-Fletcher-Goldfarb-Shanno algorithm.

2.3 Estimating the latent trajectory

Users of multi-state, multi-season occurrence models are likely to be interested in estimating the fluctuations in the latent states through time. We follow Link and Sauer (1997) and call these fluctuations the "'trajectory" of the study species. Two fundamentally different estimands provide information about the latent trajectory. The distinction between these estimands is subtle, yet critically important. The first estimand concerns the average trajectory of a theoretically infinite population of sites from which the site-specific trajectories are randomly selected realizations. This population-level estimand can be written by first letting N_{it} be the multinomial indicator vector $(I(N_{it} = 0), I(N_{it} = 1), \ldots, I(N_{it} = K))'$ that has one element equal to one and zeros elsewhere. Then, the population-level trajectory is the sequence of expected latent states $\boldsymbol{\psi}_t = E(N_{it})$ over $t = 1, 2, \ldots, T$. A population-level trajectory may be a suitable target of inference when the sites included in a sample can be regarded as a randomly selected subset of sites, and inference is desired that extends to the population from which the sites were selected.

In contrast, the "finite-sample" estimand describes the sequence of latent states for the specific sites included in the sample. This estimand can be written as $\psi_t^{(fs)} = M^{-1} \sum_{i=1}^{M} N_{it}$ for $t = 1, 2, \ldots, T$, where the *fs* superscript refers to "finite sample". Finite-sample estimands are preferable when inference is to be focused on the specific sites monitored, perhaps because the selected sites are of particular conservation or management concern.

The population and finite-sample estimands of latent trajectory will converge as the number of sites grows large (Fiske 2012). For finite sample sizes, however, they motivate different estimators with different uncertainty. In particular, finite-sample estimators are generally more precise than their population-level counterparts (Royle and Kéry 2007), although the scope of inference for finite-sample estimators is correspondingly reduced. The broad geographic coverage of NAAMP suggests that the finite-sample trajectory is appropriate for its data.

Although published methods exist for Bayesian estimation of $\psi_t^{(fs)}$ (Royle and Kéry 2007; MacKenzie et al. 2009), we are not aware of any existing frequentist method for estimating $\psi_t^{(fs)}$ for dynamic multi-state occupancy models. Here, we use maximum-likelihood estimation and HMM smoothing for frequentist estimation of $\psi_t^{(fs)}$. First, we obtain the MLEs $\hat{\boldsymbol{\Phi}}$, $\hat{\boldsymbol{\psi}}_1$, $\hat{\boldsymbol{Q}}$ using the methods described in Sect. 2.2 above. Then, we use a two-pass smoothing algorithm from the HMM literature (available in e.g., Rabiner 1989; Ephraim and Merhav 2002; Cappé et al. 2005) to calculate $\hat{N}_{it} = E(N_{it} | Y_i; \hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{Q}})$, the empirical best predictor (EBP) of the site-specific sequence of multinomial cell probabilities given the estimated model parameters and observed data for site i. The smoothed quantity $\hat{N}_{it}$ is an EBP because $E(N_{it} | Y_i; \boldsymbol{\Phi}, \boldsymbol{\psi}_1, \boldsymbol{Q})$ provides the best predictor of N_{it} in terms of minimizing mean-squared error (e.g., Williams 1991), and thus plugging in MLEs for unknown parameter values yields the conventional EBP. Finally, we estimate $\psi_t^{(fs)}$ by simply averaging $\hat{N}_{it}$ over sites:

$$\hat{\boldsymbol{\psi}}_t^{(fs)} = M^{-1} \sum_{i=1}^{M} \hat{N}_{it} \quad \text{for } t = 1, 2, \ldots, T. \tag{4}$$

In Sect. 3, we conduct simulations to evaluate the performance of $\hat{\boldsymbol{\psi}}_t^{(fs)}$.

2.4 Variance estimation for $\hat{\boldsymbol{\psi}}_t^{(fs)}$

Define $\hat{\boldsymbol{\psi}}^{(fs)}$ as the vector obtained by stacking $\hat{\boldsymbol{\psi}}_1^{(fs)}, \hat{\boldsymbol{\psi}}_2^{(fs)}, \ldots, \hat{\boldsymbol{\psi}}_T^{(fs)}$. There are at least two candidate variances that might be of interest for an EBP such as $\hat{\boldsymbol{\psi}}^{(fs)}$: the unconditional variance $\text{Var}(\hat{\boldsymbol{\psi}}^{(fs)})$ or the conditional variance $\text{Var}(\hat{\boldsymbol{\psi}}^{(fs)} | N)$, where N is the collection of N_{it} over all i and t. The quantity $\text{Var}(\hat{\boldsymbol{\psi}}^{(fs)} | N)$ captures the uncertainty in $\hat{\boldsymbol{\psi}}^{(fs)}$ attributable strictly to the error-prone observation process, while $\text{Var}(\hat{\boldsymbol{\psi}}^{(fs)})$ marginalizes over all possible realizations of the latent HMM. For the purpose of wildlife monitoring studies where the finite-sample trajectory is of interest,

the conditional variance is more relevant because we wish to condition on the sequence of latent states that actually occurred.

A parametric bootstrap estimator of $\text{Var}(\hat{\boldsymbol{\psi}}^{(fs)}|\boldsymbol{N})$ may be constructed by predicting $\boldsymbol{N}$ and then simulating observations $\boldsymbol{Y}_{it}|N_{it}$ for all i and t. One candidate for predicting N_{it} is the smoothed predictor $\hat{N}_{it}$. However, as $\hat{N}_{it}$ is a distribution and not an integer, additional levels of simulation would be required to use $\hat{N}_{it}$ in a parametric bootstrap. Instead, we predict each N_{it} by the exact sequence of latent states that maximizes the probability of the observed data for each site. Using the notation $\hat{N}_i^{boot}$ to denote this particular sequence for site i, we have

$$\hat{N}_i^{boot} = \underset{n_1, n_2, \ldots, n_T}{\arg\max} \, \Pr(\boldsymbol{Y}_i | N_{i1} = n_1, N_{i2} = n_2, \ldots, N_{iT} = n_T; \hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{Q}}). \quad (5)$$

We compute $\hat{N}_i^{boot}$ using a dynamic programming method known as the Viterbi algorithm (Rabiner 1989).

Let $b = 1, 2, \ldots, B$ index the B bootstrap iterations. Using $\hat{N}_i^{boot}$, we construct B bootstrap data sets $\boldsymbol{Y}_i^{(b)}$ for i in $1, 2, \ldots, M$. For each b, we then calculate the MLEs $\hat{\boldsymbol{\Phi}}^{(b)}, \hat{\boldsymbol{\psi}}_1^{(b)}, \hat{\boldsymbol{Q}}^{(b)}$ and a smoothed finite-sample trajectory estimate $\hat{\boldsymbol{\psi}}^{(fs),(b)}$. The desired conditional variance $\text{Var}(\hat{\boldsymbol{\psi}}^{(fs)}|\boldsymbol{N})$ is then estimated simply by the sample covariance matrix of the $\hat{\boldsymbol{\psi}}^{(fs),(b)}$.

Simulations demonstrate that the parametric bootstrap variance estimator performs suitably in a variety of contexts, as long as one of two conditions is true: either observation is reasonably accurate or sites are visited several times per season. In the worst-case scenario, the combination of badly inaccurate sampling combined with few (1–2) visits per year can lead to underestimates of $\text{Var}(\hat{\boldsymbol{\psi}}^{(fs)}|\boldsymbol{N})$ by as much as 50 %. These simulations appear in the supplementary material.

2.5 Univariate trend assessment

To perform formal inference about the significance of trend, it is helpful to summarize $\hat{\boldsymbol{\psi}}_t^{(fs)}$ in terms of a scalar summary statistic. Many such univariate statistics may be constructed. For example, the change in the proportion of sites in state k over the course of the study is quantified by $\psi_T^{(fs)}(k) - \psi_1^{(fs)}(k)$. Alternatively, a linear trend in the proportion of sites in state k can be assessed by using the linear polynomial contrast in $\psi_t^{(fs)}(k)$. A test of whether or not this linear contrast is equal to zero is equivalent to fitting a linear regression of $\psi_t^{(fs)}(k)$ versus t and testing whether the slope equals zero.

A point estimate for a univariate trend statistic may be obtained simply by plugging in appropriate components of $\hat{\boldsymbol{\psi}}_t^{(fs)}$. Inference, both in terms of confidence intervals and hypothesis tests, can be obtained using the parametric bootstrap procedure developed in Sect. 2.4. In Sect. 4, we illustrate use of univariate trend statistics for *P. crucifer*. Simulations show that confidence intervals and hypothesis tests for univariate

trends exhibit satisfactory performance under a variety of generative models, including latent trajectories that are unlikely to have been generated by a Markov process. The supplementary material details these simulations.

3 Simulation study

3.1 Study design

We examined the performance of $\hat{\boldsymbol{\psi}}_t^{(fs)}$ using simulated data structured comparably to NAAMP data. The simulation envisioned a sampling regime with three observations per year ($J = 3$), 8 years ($T = 8$), and the number of sites taking values $M \in \{100, 200, 400\}$. We used a four-state HMM ($K = 4$) with latent states corresponding to levels of abundance. We specified a transition structure in which gradual transitions among latent states were more likely than sudden transitions:

$$\boldsymbol{\Phi} = \begin{pmatrix} 0.64 & 0.24 & 0.09 & 0.03 \\ 0.20 & 0.53 & 0.20 & 0.07 \\ 0.02 & 0.11 & 0.83 & 0.04 \\ 0.01 & 0.09 & 0.24 & 0.66 \end{pmatrix}. \tag{6}$$

The initial state vector was defined as $\boldsymbol{\psi}_1 = (0.338, 0.250, 0.226, 0.185)'$, so that unoccupied sites were most prevalent and highly used sites were least prevalent.

We varied the detection process over four scenarios: (1) accurate detection with no covariates, (2) poor detection with no covariates, (3) accurate detection with three 'real' covariates that influenced the response, (4) accurate detection with one real covariates and two sham covariates that were unrelated to the response. 'Accurate' and 'poor' refer to the probability of under-detection when all covariates are equal to their mean. Poor detection has an increased chance of under-detecting all latent states compared to accurate detection (Table 2). For the two scenarios with detection covariates, all covariates were generated as independent standard normal random variables. We modeled the relationship between covariates and detection probabilities using multinomial logistic regression (Agresti 2002), as detailed in the supplementary material. Each detection covariate required six regression parameters in the observation component. Thus, the largest of these models was then highly parametrized, using

Table 2 Q for two simulated detection scenarios	Detection scenario	Detection matrix Q
	Accurate detection	$\begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 \\ 0.20 & 0.80 & 0.00 & 0.00 \\ 0.08 & 0.12 & 0.80 & 0.00 \\ 0.04 & 0.08 & 0.08 & 0.80 \end{pmatrix}$
	Inaccurate detection	$\begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 \\ 0.50 & 0.50 & 0.00 & 0.00 \\ 0.12 & 0.38 & 0.50 & 0.00 \\ 0.08 & 0.08 & 0.33 & 0.50 \end{pmatrix}$

12 parameters for $\boldsymbol{\Phi}$, 18 for $\boldsymbol{Q}$, and three for $\boldsymbol{\psi}_1$. We sampled 100 data sets from each scenario-sample size combination.

Typical Monte Carlo studies of estimators use mean-squared error to evaluate the performance of the estimator. However, because the rows $\boldsymbol{\Phi}$ and $\boldsymbol{Q}$ each describe a probability distribution, we constructed a loss function based on the Kullback-Leibler (KL) divergence measure. Specifically, for both $\hat{\boldsymbol{\Phi}}$ and $\hat{\boldsymbol{Q}}$, we measured the distance from the estimate to the true value by the average KL distance over the rows of the each matrix. For example, let $\boldsymbol{q}_k$ be the kth row of $\boldsymbol{Q}$. Then, we used

$$\overline{\mathrm{KL}}(\boldsymbol{Q}, \hat{\boldsymbol{Q}}) = K^{-1} \sum_{k=1}^{K} \mathrm{KL}(\boldsymbol{q}_k, \hat{\boldsymbol{q}}_k)$$

to measure the deviation between an $\boldsymbol{Q}$ and $\hat{\boldsymbol{Q}}$ where

$$\mathrm{KL}(\boldsymbol{q}_k, \hat{\boldsymbol{q}}_k) = \sum_{i=1}^{4} q_{ki} \log \frac{q_{ki}}{\hat{q}_{ki}}. \tag{7}$$

We computed $\overline{\mathrm{KL}}(\boldsymbol{Q}, \hat{\boldsymbol{Q}})$ with detection covariates set to zero. We used an equivalent divergence measure for $\boldsymbol{\Phi}$, $\overline{\mathrm{KL}}(\boldsymbol{\Phi}, \hat{\boldsymbol{\Phi}})$.

3.2 Simulation results

Figure 1 illustrates the performance of $\hat{\boldsymbol{\Phi}}$ and $\hat{\boldsymbol{Q}}$ under each of the simulated scenarios. Both $\overline{\mathrm{KL}}(\boldsymbol{\Phi}, \hat{\boldsymbol{\Phi}})$ and $\overline{\mathrm{KL}}(\boldsymbol{Q}, \hat{\boldsymbol{Q}})$ decrease as the number of sites increases, suggesting that the estimators are well-behaved in the range of scenarios we tested. To analyze our results, we fit an ordinary least squares regression model with $\log(\overline{\mathrm{KL}}(\boldsymbol{\Phi}, \hat{\boldsymbol{\Phi}}))$ as the response and sample size and detection scenario as a continuous and categorical predictors, respectively. The poor detection scenario was significantly worse than each of the others at estimating $\boldsymbol{\Phi}$. Comparing just the two no-covariate scenarios, both $\boldsymbol{Q}$ and $\boldsymbol{\Phi}$ were estimated more accurately with $M = 200$ under the poor detection than with $M = 100$ and accurate detection. This result suggests that increasing the sample size can offset the cost of poor detection. Also, the addition of sham covariates did not appear to affect the estimates of $\boldsymbol{\Phi}$. We fit a similar regression model with $\overline{\mathrm{KL}}(\boldsymbol{Q}, \hat{\boldsymbol{Q}})$ as the response. Estimation of $\boldsymbol{Q}$ was most successful when the true structure of $\boldsymbol{Q}$ contained fewer covariates. Adding noise covariates did not affect the estimation, and poor versus good detection did not compromise the estimate, either. Unfortunately, as the true detection structure is highly dependent on context and largely beyond the investigator's control, this result should serve simply to warn of the difficulty in estimating a complicated detection structure.

Figure 2 displays average performance of $\hat{\boldsymbol{\psi}}_t^{(fs)}$. Smoothing is the better method in most scenarios. Interestingly, $\hat{\boldsymbol{\psi}}_t^{(fs)}$ is a more accurate predictor of the actual finite-sample trajectory in the middle of the time series than at either its beginning or end. We speculate that this phenomenon may result from a tail effect analogous to the

Fig. 1 Mean simulated average Kullback-Leibler divergence between true and estimated $\boldsymbol{\Phi}$ and $\boldsymbol{Q}$. Asymptotic 95 % CIs for the mean are too small to be visible

Detection Scenario
- • accurate with no covariates
- ▲ inaccurate with no covariates
- ■ all real covariates
- + real and noise covariates
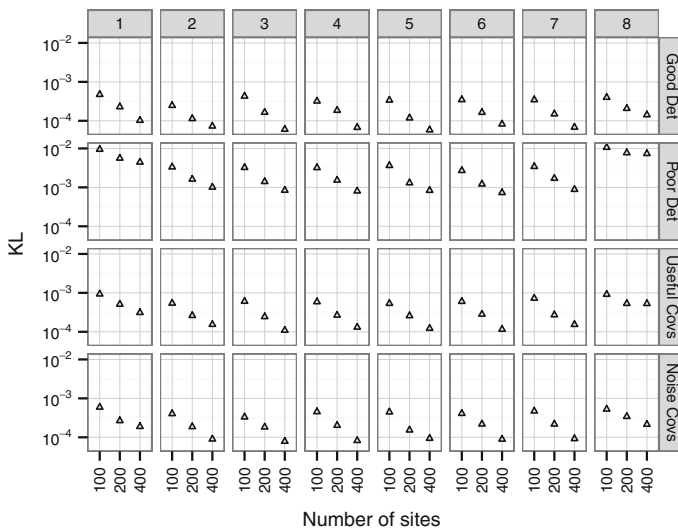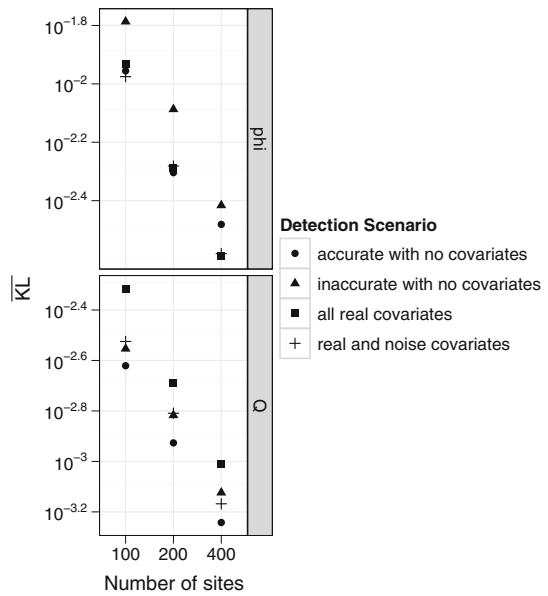
Fig. 2 Mean simulated Kullback-Leibler divergence between true and estimated finite-sample trajectory. Years are *plotted* by *grid column* and simulation scenarios are *plotted* by *grid row*. Asymptotic 95 % CIs for the mean are too small to be visible

poor performance of a local regression fit at the tails of a data sequence. That is, smoothing uses all data before and after a year to estimate the latent state for that year. Thus, the tails are effectively estimated using fewer data than the middle of the sequence. Not surprisingly, $\hat{\boldsymbol{\psi}}_t^{(fs)}$ was most accurate as the number of sites increased,

and when detection was more accurate. The use of observation covariates did not have a noticeable impact on the performance of the estimator.

4 Example: NAAMP data for *Pseudacris crucifer*

To illustrate an HMM analysis for the NAAMP data, we fit an HMM to *Pseudacris crucifer* (Spring Peeper) data from 2001 to 2009. *P. crucifer* is a common, widely distributed early breeder that congregates in large numbers and makes a loud breeding chorus. This nine-year data set has 2,051 sites with most sites visited 3–4 times per year. Geographic coverage includes 10 U.S. states (ME, NH, VT, MA, PA, NJ, MD, DE, VA, and WV) that extend from New England to the mid-Atlantic. As in the simulation, we use a latent state with $K = 4$ increasing levels of abundance. We define the levels of the latent state as the maximum calling potential of a population at a site over the course of the breeding season. Thus, $N_{it} = 1$ corresponds to sites that are vacant, and $N_{it} = 4$ corresponds to sites capable of producing a chorus. Exploratory analysis indicated that calling intensity consistently peaked during the breeding season in mid-to-late spring (Fiske 2012). To account for this, we modeled each of the six detection detection parameters in Q as a function of two predictor variables corresponding to day-of-year and day-of-year squared. These two covariates plus an intercept yield a detection model with 18 parameters.

Table 3 presents the estimated transition matrix $\hat{\boldsymbol{\Phi}}$ along with standard errors estimated from the observed information matrix. The matrix $\hat{\boldsymbol{\Phi}}$ shows that transitions among latent states were relatively rare, as most populations at most sites tended to remain in the same latent state from one year to the next. This is especially true for sites in the full-chorus state, which were estimated to have a 0.942 (s.e. = 0.010) probability of remaining in that state between successive years. Several transitions appear to be exceedingly unlikely in a single time step, including transitions from an unoccupied state to either a full or partial chorus. Table 3 also presents the estimate of the detection matrix Q for April 15, a day during the peak of the breeding season. This estimate shows that under-detection was not unusual, especially at sites that were occupied by less than a full chorus-producing population. The supplementary material shows how the diagonal elements of $\hat{Q}$ varied over the course of the year.

Table 3 List of parameter estimates for *Pseudacris crucifer* with standard errors in parentheses

Parameter	Estimate (Standard Errors)
$\boldsymbol{\psi}_1$	$(0.016(0.019), 0.06(0.022), 0.138(0.013), 0.786(0.013))'$
$\boldsymbol{\Phi}$	$\begin{pmatrix} 0.765(0.176) & 0.235(0.176) & 0.000(0.006) & 0.000(0.007) \\ 0.058(0.075) & 0.777(0.070) & 0.000(0.004) & 0.165(0.034) \\ 0.038(0.008) & 0.000(0.007) & 0.779(0.018) & 0.182(0.017) \\ 0.002(0.008) & 0.011(0.009) & 0.045(0.009) & 0.942(0.010) \end{pmatrix}$
$\boldsymbol{Q}$ (day = Apr 15)	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.855(0.029) & 0.145(0.029) & 0 & 0 \\ 0.311(0.015) & 0.287(0.012) & 0.401(0.016) & 0 \\ 0.133(0.003) & 0.078(0.003) & 0.161(0.004) & 0.628(0.006) \end{pmatrix}$

Fig. 3 Smoothed finite-sample estimates of latent trajectory for *P. crucifer* from 2001 to 2009. *Bars* are parametric-bootstrap standard errors
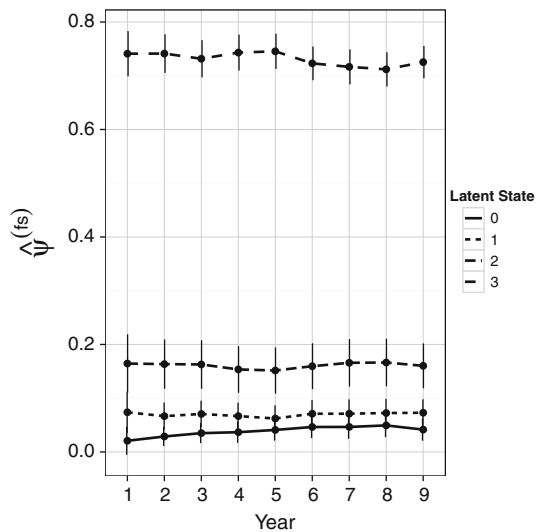
Figure 3 shows the estimated latent trajectory $\hat{\boldsymbol{\psi}}_t^{(fs)}$ for $t = 1, 2, \ldots, 9$ along with the 95 % confidence interval estimated using the parametric bootstrap. The latent trend suggests that nearly three-quarters of the surveyed sites contained populations in the chorus state, while only a small percentage of sites were unoccupied. This pattern is maintained over the course of the study, suggesting that this species is fairly stable. To rigorously test for temporal changes in *P. crucifer* abundance, we estimated and conducted inference for two univariate trend statistics. First, we estimated $\psi_T^{(fs)}(1) - \psi_1^{(fs)}(1)$, the change in the proportion of unoccupied sites between the first and last years. We estimated this change to be $+0.0208$ with 95 % CI $(-0.0520, 0.0594)$, suggesting no significant evidence for a change in the number of unoccupied sites. We also fit a linear polynomial contrast in $\psi_t^{(fs)}(K)$ to investigate whether there was evidence of a linear trend in the proportion of maximally occupied sites. This contrast was estimated as -0.201 with 95 % confidence interval $(-0.8350, -0.0439)$, suggesting that the proportion of sites capable of producing a full chorus declined slightly over time. This small reduction in the number of maximally occupied sites may or may not be enough of a decline to warrant conservation action, but the test does show that the decline is statistically significant.

5 Discussion

Modeling species occurrence is a common objective of ecological studies. Occurrence data are relatively inexpensive to collect, and, depending on species biology, may be a more feasible alternative to expensive tracking of individual animals. Here, we have shown how longitudinal, multi-state occurrence data may be analyzed across a collection of spatial sites using a hidden Markov model (HMM). HMMs have been used to analyze longitudinal data in a variety of applications (Altman 2007; Scott et al. 2005;

Bartolucci et al. 2012), and within ecology the methods in this paper are closely related to models for multi-state mark-recapture data when state classification is either imperfect or uncertain (Fujiwara and Caswell 2002; Pradel 2005; Runge et al. 2007; Conn and Cooch 2009; Kendall et al. 2012). However, existing applications have focused mainly on estimating the transition rates that control the evolution the hidden process. While estimates of transition rates are certainly useful for learning about a species' underlying biology, we focus here on using HMM smoothing methods to estimate a finite-sample population trajectory (Cappé et al. 2005). Finite-sample trajectory estimation is an important task in conservation settings where accurate assessment of recent trends in population abundance is imperative for successful management. The usefulness of finite-sample trajectory estimates is apparent in our illustrative example, where the finite-sample trajectory estimate for the Spring Peeper shows evidence of a slight decline in the proportion of maximally occupied states.

Data from large-scale monitoring programs such as NAAMP have been routinely analyzed using either standard regression-type models (which ignore the discrete nature of the dependent variable), or using logistic regression models with reduced-information summaries of the data. In particular, a generalized version of logistic regression which allows for false non-detection—so called "occupancy models" (MacKenzie et al. 2002, 2006) have been widely adopted in animal surveys, including for inference in NAAMP and related surveys (MacKenzie et al. 2002; Weir et al. 2005, 2009). The application of such models requires converting all of the observations that correspond to various levels of abundance into an observation of simply "detected" ($y = 1$). As has been noted elsewhere (Royle 2004; Royle and Link 2005) this conversion entails a loss of efficiency and can bias trend estimates or covariate effects, especially for common species with broad geographic ranges or that exist at high abundance levels, such as Spring Peeper. In such cases, most of the dynamics occurs in the higher range of the state-space (e.g., Fig. 3). Therefore, it is important to use observations that are more closely linked to abundance rather than simple occurrence as such models outperform estimators based on occupancy and extinction or colonization fitted to binary data (Dorazio 2007).

MacKenzie et al. (2009) recently developed a framework for inference in multi-state models such as those considered here, and used Bayesian methods to estimate the finite sample trajectory by calculating summary statistics of the latent site states within each iteration of a Gibbs sampler. Although Bayesian methods are very flexible and potentially useful, there are several aspects of their implementation that some researchers remain wary of. For example, Bayesian inference relying on Markov chain Monte Carlo (MCMC) integration can be quite slow to converge, taking hours or days. While this may be acceptable for a single model fit, researchers often require fits to many models, either to analyze many datasets or to perform comparisons among predictions of many competing models. In contrast, the frequentist estimate presented here can be found relatively quickly. Additionally, Bayesian and MCMC-based inference must always address questions of prior sensitivity and convergence assessment. Thus, although we do not promote any single inferential paradigm as best for all purposes, we believe that the HMM-based frequentist methods presented here will be a useful and practical alternative to Bayesian finite-sample trend estimation.

Several caveats to these methods deserve mention. First, and perhaps foremost, parameter identifiability requires some form of an assumption about the relationship between the data observed and the unobservable latent state. Here, we have invoked the assumption that over-detection is sufficiently rare that it can be ignored. This assumption is the natural multi-state generalization of the common assumption that rates of false detections are negligible in occupancy studies. While this assumption may seem severe from a statistical standpoint, it is an assumption that biologists are often willing to support. One obvious way in which this assumption could be violated is if surveyors are unable to distinguish multiple species calling simultaneously. Thus, whether over-detection is truly negligible must be determined on a case-by-case basis. Recent work discussing the ramifications of false positive detections can be found in McClintock et al. (2010) and Miller et al. (2012).

A second caveat is that, even with the assumptions mentioned above, parameter identifiability is not guaranteed, and identifiability will need to be explored on case-by-case basis (Pradel 2005; Giménez et al. 2004) until broad principles are uncovered. The identifiability of a particular model can be roughly ascertained by examining the condition of the Hessian of the negative-log-likelihood at the MLE. Finally, we note that the decision about whether finite-sample or population inference is desired should be determined at the time of data collection. Finite-sample trajectory estimators should not be used merely as a means of reducing the reported standard error, as such a post hoc decision would yield biased estimates of precision.

References

Agresti A (2002) Categorical data analysis. Wiley series in probability and statistics. Wiley Interscience, Hoboken, NJ

Altman RM (2007) Mixed hidden Markov models. J Am Stat Assoc 102(477):201–210

Bartolucci F, Farcomeni A, Pennoni F (2012) Latent Markov models for longitudinal data. CRC Press, Boca Raton FL

Baum L, Egon J (1967) An inequality with applications to statistical estimation for probabilistic functions of a markov process and to a model for ecology. Bull Am Meteorol Soc 73:360–363

Baum L, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 41:164–171

Cappé O, Moulines E, Rydén T (2005) Inference in hidden Markov models. Springer, Berlin

Conn PB, Cooch EG (2009) Multistate capture-recapture analysis under imperfect state observation: an application to disease models. J Appl Ecol 46:486–492

Dorazio R (2007) On the choice of statistical models for estimating occurrence and extinction from animal surveys. Ecology 88(11):2773–2782

Ephraim Y, Merhav N (2002) Hidden Markov processes. IEEE Trans Inf Theory 48(6):1518–1569

Fiske IJ (2012) Characterizing spatiotemporal trends in amphibian abundance using latent variable models. PhD thesis, North Carolina State University.

Fujiwara M, Caswell H (2002) Estimating population projection matrices from multi-stage mark-recapture data. Ecology 83(12):3257–3265

Giménez O, Viallefont A, Catchpole EA, Choquet R, Morgan BJT (2004) Methods for investigating parameter redundancy. Anim Biodivers Conserv 27(1):561–572

Kendall W, White G, Hines J, Langtimm C, Yoshizaki J (2012) Estimating parameters of hidden markov models based on marked individuals: use of robust design data. Ecology 93:913–920

Link W, Sauer J (1997) New approaches to the analysis of population trends in land birds: comment. Ecology 78(8):2632–2634

MacKenzie D, Nichols J, Seamans M, Gutiérrez R (2009) Modeling species occurrence dynamics with multiple states and imperfect detection. Ecology 90(3):823–835

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. Ecology 83(8):2248–2255

MacKenzie DI, Nichols JD, Hines JE, Knutson MG, Franklin AB (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84(8):2200–2207

MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE (2006) Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Academic Press, USA

McClintock B, Bailey L, Pollock K, Simons T (2010) Experimental investigation of observation error in anuran call surveys. J Wildl Manag 74:1882–1893

Miller D, Weir L, McClintock B, Grant E, Bailey L, Simons T (2012) Experimental investigation of false positive errors in auditory species occurrence surveys. Ecol Appl 22:1665–1674

Nichols JD, Hines JE, MacKenzie DI, Seamans ME, Gutiérrez R (2007) Occupancy estimation and modeling with multiple states and state uncertainty. Ecology 88(6):1395–1400

Pradel R (2005) Multievent: an extension of multistate capture-recapture models to uncertain states. Biometrics 61:442–447

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286

Royle J (2004) Modeling abundance index data from anuran calling surveys. Conserv Biol 18(5):1378–1385

Royle JA, Kéry M (2007) A bayesian state-space formulation of dynamic occupancy models. Ecology 88(7):1813–1823

Royle JA, Link WA (2005) A general class of multinomial mixture models for anuran calling survey data. Ecology 86(9):2505–2512

Runge JP, Hines JE, Nichols JD (2007) Estimating species-specific survival and movement when species identification is uncertain. Ecology 88(2):282–288

Scott SL, James GM, Sugar CA (2005) Hidden Markov models for longitudinal comparisons. J Am Stat Assoc 100(470):359–370

Weir L, Fiske IJ, Royle JA (2009) Trends in anuran occupancy from northeastern states of the north American Amphibian monitoring program. Herpetol Conserv Biol 4(3):389–402

Weir LA, Royle JA, Nanjappa P, Jung RE (2005) Modeling anuran detection and site occupancy on north American Amphibian monitoring program (NAAMP) routes in Maryland. J Herpetol 39(4):627–639

Welch L (2003) Hidden Markov models and the baum-welch algorithm. IEEE Inf Theory Soc Newslett 53:1–13

Williams D (1991) Probability with martingales. Cambridge University Press, Cambridge

Zucchini W, MacDonald IL (2009) Hidden Markov models for time series: an introduction using R. CRC Press, Boca Raton, FL

Author Biographies

Ian J. Fiske Ph.D., is a statistician who currently works in the private sector.

J. Andrew Royle Ph.D., is a Research Statistician and the USGS Patuxent Wildlife Research Center in Patuxent, MD, USA.

Kevin Gross Ph.D., is an Associate Professor in the Department of Statistics at North Carolina State University in Raleigh, NC, USA.