**OVERVIEW**

The motivating context is the study of ecological communities comprised of species, where samples of organisms are collected over time (Christensen *et al.* 2018). We are interested in determining if the composition of communities changes over the course of the study, and if it does, we seek to quantify those dynamics. However, ecological communities are typically composed of many species relative to the number of samples collected (*i.e.,* the data are high-dimensional), which presents a challenge to time series modeling.

To address this problem, we reduce the dimensionality of the community data prior to time series analysis (Christensen *et al.* 2018). We accomplish this through a two-stage analysis referred to here as LDATS, wherein the first stage (LDA) uses Latent Dirichlet Allocation (Blei *et al*. 2003) to reduce (high-dimensional) species data to (low-dimensional) community data, which are then analyzed in the second stage (TS) using multinomial Time Series models, building upon the changepoint time series model of Western and Kleykamp (2004). Here, the TS models include discrete (changepoint) and continuous temporal changes as well as covariate impacts and are fit using parallel tempering Markov Chain Monte Carlo (ptMCMC) methods (Earl and Deem 2005).

This document describes the two-stage LDATS model in a unified mathematic setting and accompanies the codebase (Simonis *et al*. *In Development*). Because of the overlap in notation between classical LDA (Blei *et al.* 2013) and time series models (*e.g.*, both use but with different meanings, and further, the time series models use in describing both the statistical model and the fitting algorithm), we create a new notational set for use here and within the LDATS code and curate a translational list of terms (Table 1).

**STAGE ONE: DIMENSION REDUCTION**

The first stage of the analysis is focused on reducing the putatively high-dimensional species-level data to a lower dimensional representation of community types (Valle *et al*. 2014) using Latent Dirichlet Allocation (Blei *et al*. 2003). The following text describes the application of the LDA model derived and developed by Blei *et al*. (2003) that has been applied to ecological data by Valle *et al*. (2014) and Christensen *et al*. (2018), but with full derivation using notation that facilitates integration with time series models. Although the notation is non-standard for LDA, the application here has no novel developments (Blei *et al*. 2003; Table 1).

In relation to the linguistic models that motivated the original LDA description and notation (Blei *et al.* 2003), species are like terms (word options) in a vocabulary, communities are like linguistic topics, samples are like documents using the terms, and organisms within a sample are like words within documents (Valle *et al*. 2014). An ecological study consists of total samples comprising total organisms from total species. Each sample (in ) consists of organisms ( in ) assigned to one of in species. The total number of organisms in the study is the sum of the organisms within each sample:

LDA involves grouping the species into in total latent (unobserved) component communities (*sensu* Valle *et al*. 2014), such that “component community” means a group of species that tend to be found together in specific proportions (Christensen *et al*. 2018). The allocation process (Blei *et al.* 2003)allows that individual species can be assigned to multiple component communities (Valle *et al*. 2014). The total number of latent communities is also unknown, and for the present approach is fixed *a priori* within a given Stage 1 (LDA) model in .

Each organism within a sample has an observed species identity, notated as , and a latent community membership, notated as . Because there are varying numbers of organisms in each sample, we use a vector structure to hold organism-level data across the study. The species identities of all organisms within sample are (or just ) and the species identities of all organisms across all samples are (or simply ), an -length vector. Similarly, the community identities of all organisms in sample are represented as (or simply ) and the community identities of all organisms across all samples are (or just ), an -length vector. Across all organisms from all samples, is the latent state (community identity) and is what is observed (species identity).

For a specific model with a total number of communities , the organism-level community distribution within a sample (*i.e.*, the allocation of an organism among the possible communities) is a -dimension categorical random variable (equivalent to a multinomial variable with sample size 1) described by probabilities held in vector () and collated across samples into -dimension matrix . Thus, the community identity of an organism within a sample is

The vector of community probabilities within a sample () is described by a -dimensional Dirichlet distribution with concentration parameters ,which we assume do not change among samples (*i.e.*, ). Further, we assume that the concentration parameters are symmetric, thereby reducing the distribution to reliance on a single parameter (*i.e.,* ):

which is an unknown parameter to be estimated.

The organism-level species distribution within a sample (*i.e.*, the allocation of an organism among the possible species) is an -dimension categorical random variable contingent upon the community identity of the organism and defined by probabilities , where the sum across species within a community for a given sample is 1 (*i.e.*, ). The probabilities across all communities within a sample are held in an -dimension matrix (), which we assume is constant across samples, (), thus the species state of an organism within a sample is

which is equivalent to a categorical variable defined by the row of interest in given community identity

The components of are unknown parameters to be estimated.

The inferential problem of interest lies in determining the posterior distribution of the latent quantities (the community probabilities and states ), given the observations () and fitted parameters ( and )

an equation which obviously necessitates an estimation of the parameters and . From a parameter estimation standpoint, we are concerned with the probability of the suite of observations () given the parameters and , or the likelihood () of the parameters given the data:

However, for fitting purposes, we are interested in the log-likelihood () of the parameters, given the data

The log-likelihood of (the data across all samples) is the sum of the logged probabilities of each sample’s data given the parameters (under the assumption that all samples are derived from the same parameters):

The probability of a sample’s data given the parameters () can be decomposed into the product of [1] the organism-by-organism species-identity distributions () and [2] the sample-level community distribution (), integrated over the uncertainty in the latent community distribution, :

The organism-level species-identity distribution can be further decomposed into the product of [1] the species identity distribution given the community identity and the unknown parameter matrix () and [2] the community identity distribution given the latent community distribution (), integrated (summed due to discreteness) over the uncertainty in community type

The probability of a single sample’s observations can therefore be decomposed into the product of [A] the product of the organism-by-organism species distributions, each integrated over the uncertainty in community type (which are themselves the product of [1] the species identity given the community identity and the unknown parameter and [2] the community identity given the latent parameter ) and [B] the probability of the latent parameter given the unknown parameter , integrated over the uncertainty in

This is then scaled up to the probability of the entire set of samples under the unknown parameters and , which is the product of the sample-level probabilities given those parameters:

The probabilities for in cannot be tractably estimated due to the coupling of (and thus ) and in the summation over latent communities (Blei *et al*. 2003). To address this, we use a variational approximation to the equations that decouples the parameters, and which we fit using the expectation-maximization routine (aka VEM for Variational Expectation Maximization; Blei *et al*. 2003).

To accomplish this, we endow the model with free variational parameters ( and ) that characterize a family of distributions (notated by to distinguish from ) which provide a lower bound on the log likelihood. is an -dimension matrix similar to . However, the rows of (row corresponding to sample : or simply ) correspond to the concentration parameters of a -dimension Dirichlet distribution and therefore are not constrained to sum to 1. is an -dimensional matrix, where the rows correspond to the organisms across samples (indexed akin to and ) and the columns correspond to the communities. describes the probability that organism within sample is from community , and (or simply ) is a -length vector of probabilities that define the categorical distribution controlling that organism’s community identity, where . is composed of sample-specific matrices, where the matrix for a specific sample is -dimensional and notated .

In comparison to and , the variational parameters are sample-specific (able to vary among samples), and so are found separately for each sample. For a specific sample , the variational distribution is

As the Expectation Step (“E-Step”) in the VEM algorithm, the distribution can be used to find a tight lower bound on by optimizing the variational parameters and (*i.e.*,find and , where the asterisks notate optimal values) with respect to minimizing the Kullback-Leibler Divergence () between and :

Minimization of the distance is achieved through an iterative fixed-point method, where the derivative of the is set to zero, producing a pair of update equations. First, the parameters describing the community allocation of each organism () are updated based on the community distribution for the sample ():

where the expected value of the (log-scale) community probability is calculated using the digamma function (), which is the logarithmic derivative of the gamma function (), a quantity that is calculated through Taylor approximation:

And then, the parameters describing the community distribution for the sample () are updated based on the organism-level community distributions for the sample ():

The update equations are alternated until the bound converges (*i.e.*, the updates do not yield changes to the parameters), at which point the document-specific variation parameters have been optimized (and have been found) for the set of main parameters ().

Similarly to the overall log likelihood for the set of samples being equal to the sum of the log likelihoods for the individual samples, the variational lower bound () for all of the samples is equal to the sum of the variational lower bounds for the individual samples:

The complete E-Step maximizes this overall lower bound with respect to the full variational parameters and (*i.e.*, given the main model parameters). The Maximization Step (“M-Step”) maximizes the overall lower bound with respect to the main model parameters and (*i.e.*, given the optimal variational parameters). This corresponds to obtaining maximum likelihood values of the model parameters using expected sufficient statistics for each sample under the approximate posterior calculated in the E-Step.

The update for the community-level species distribution () can be written analytically:

where is an indicator variable based on the species identity of the organism:

The update for the concentration parameter underlying the sample-level community distribution () requires an iterative approach to find a station point estimate. The optimization is conducted using the Newton-Raphson method (Ronning 1989), which repeats

(where is the hessian matrix and is the gradient) until convergence. Having updated the main model parameters (the M-Step), a new iteration of the E-Step followed by the M-Step is conducted, and the E-Step and M-Step are alternated until the lower bound of the log-likelihood converges. Thus, the VEM approach can be considered coordinate ascent in (the space defined by the lower bound).

Once the log-likelihood has converged, the VEM algorithm has arrived at approximate maximum likelihood estimates for the model parameters ( and ) given the full set of observations () for the specific Stage 1 model (notated by the component in the superscripts), which forms the basis of Stage 1 model selection to determine the most parsimonious number of communities . Specifically, we use as our Stage 1 model selection criterion, defined for a specific LDA model :

where is the number of parameters in the model: corresponding to and corresponding to each entry in , a -dimensional matrix (Grün and Hornik 2011). If small sample size is a concern with respect to the degrees of freedom being used, we can use the correction based on the number of observations, here sample size (, Grün and Hornik 2011):

Because of the use of multiple iterative optimization routines (which require starting values to be drawn at random) to solve otherwise intractable likelihood functions, it is critical to account for the potential influence of starting values on analytical results. Here, we accomplish this by running multiple models with the same number of communities () using different starting values, assigned through the random number generator seed ().

Specifically, we use replicates ( in ) at each number of communities from 2 to , the total number of communities to be explored. The minimal number of communities is set to 2 by the current coding implementation of the LDA algorithm (Blei *et al.* 2003, Grün and Hornik 2011), although it is possible to generalize the code to include the limiting case of a single community. Thus, the total number of models in the Stage 1 () is

The optimal (according to ) LDA model () is determined by

and has a corresponding set of parameters:

Having found the optimal model, we can obtain the posterior point estimates for the sample-level community probabilities (held in a -length vector) by normalizing the vector of optimal values of the variational Dirichlet concentration parameters (taken from the final step of the VEM algorithm in model ) for the sample () so it sums to one and is thereby a proper proportion (notated by the dot accent ):

The posterior point estimates of the community proportions across the samples are then held in an -dimension matrix , corresponding to the optimal (from an “ based on VEM inference”-perspective) decomposition of the species-level data to community-level data. This matrix forms the multivariate response variable analyzed in the time series model, as outlined in the next section.

**MULTIVARIATE TIME SERIES MODEL**

The approach we take in analyzing the temporal dynamics of the data includes a combination of multinomial generalized regression (estimated via neural networks according to Ripley 1996, Venables and Ripley 2002), which allow us to focus on dynamics of community composition, and change point time series analyses (Western and Kleykamp 2004) that allow for continuous and discrete changes in dynamics.

The output from the decomposition detailed in the previous section () is a matrix of multinomial variables corresponding to proportions of each community in present in each sample in , where each sample has a weight , which is a function of the number of organisms collected relative to the maximum number collected across all samples:

thereby allowing us to account for variable efforts (numbers of organisms collected) among samples.

We are interested in quantifying changes in community composition over time, and so we define the time of sample to be and the vector of sample times to be . can be a discrete or continuous variable, but is always increasing from to , and defines the temporal relationship among samples. In addition, we collate total covariates, indexed as in and measured for each sample. The value of a particular covariate for a specific sample is and the set of covariates for the sample is a vector or simply . All of the covariates across all of the samples are held in , an -dimensional matrix.

Although the analysis of the community proportions is via “time series” models, the time of the sample () *per se* does not necessarily enter every Stage 2 (time series) model ( in ) directly. Rather, depending upon the model, the time of the samples may control the application of the predictor variables in the model (in the case of discrete change points), may directly influence quantitative values of predictors (in the case of continuous time impacts), and may impact the non-independence of the data (via autocorrelation structures).

Using Poisson log-linear model (Venables and Ripley 2002)

Multinomial generalized model (Ripley 1996, Venables and Ripley 2002)

Changepoint model (Western and Kleykamp 2004)

Parallel tempering MCMC (Earl and Deem 2005)

**LITERATURE CITED**

Blei, D. M., A. Y. Ng, and M. I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3**: 993-1022.

Christensen, E. M., D. J. Harris, and S. K. M. Ernest. 2018. Long-term community change through multiple rapid transitions in a desert rodent community. *Ecology* **99**:1523–1529.

Earl, D. J. and M. W. Deem. 2005. Parallel tempering: theory, applications, and new perspectives. *Physical Chemistry Chemical Physics* **7**: 3910-3916.

Grün, B. and K. Hornik. 2011. topicmodels: an R package for fitting topic models. *Journal of Statistical Software* **40**:13. http://dx.doi.org/10.18637/jss.v040.i13

Ripley, B. D. 1996. Pattern Recognition and Neural Networks. Cambridge.

Ronning, G. 1989. Maximum likelihood estimation of Dirichlet distributions. *Journal of Statistical Computation and Simulation* **34**:215-221.

Simonis, J. L., E. M. Christensen, D. J. Harris, H. Ye, R. Diaz, and S. K. Morgan Ernest. *In Development*. LDATS: Latent Dirichlet Allocation coupled with Time Series analyses. R package v 0.0.6. https://github.com/weecology/LDATS

Valle, D., B. Balser, C. W. Woodall, and R. Chazdon. 2014. Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical model. *Ecology Letters* **17**: 1591-1601.

Venables, W. N. and B. D. Ripley. 2002. Modern and Applies Statistics with S. Fourth Edition. Springer.

Western, B. and M. Kleykamp. 2004. A Bayesian change point model for historical time series analysis. *Political Analysis* **12**:354-374.