This document describes the two-stage LDATS model in a unified mathematic setting. Because of the overlap in notation between classical LDA and time series models (*e.g.*, both use but with different meanings), we create a new notational set for use within LDATS, and curate a translational list

The motivating context is the study of ecological communities comprised of species, where samples of organisms are collected. In relation to linguistic models (which motivate much of the LDA notation), communities are like topics, species are like terms, samples are like documents, and organisms are like words within documents.

An ecological study consists of total samples comprising total organisms from total species. Each sample (in ) consists of organisms ( in ) assigned to one of the species ( in ). The total number of organisms in the study is the sum of the organisms within each sample:

LDA involves the latent, grouping of species into total communities, such that communities, *per se*, are never observed and that individual species can be assigned to multiple communities. The total number of communities () is unknown, and for the present approach is fixed *a priori*. **[will expand to model level]**

Each organism within a sample has an observed state (species) noted as and a latent state (community), noted as . Because there are varying numbers of organisms in each sample, we use a vector structure to hold organism-level data (rather than a matrix structure). The species identities of all organisms within sample are (or just ) and the species identities of all organisms across all samples are (or simply, ), which is 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 ), which is an -length vector. is the latent state and is what is observed.

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 (*i.e.,* a multinomial variable with sample size 1) described by probabilities , which are held in vector () and collated across samples in an -dimension matrix (). Thus the community state 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 value 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, but is contingent upon the community identity of the organism and is 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 the community identity

The components of are unknown values 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 )

which necessitates an estimation of the parameters. From an 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

as then 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. 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).

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