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.

Sample in total samples (akin to documents)

Species in total species (akin to terms)

Community  in total communities (akin to topics)

is unknown

Organism in total organisms (akin to words)

The number of organisms within sample is and

The prior weight of community in a sample:

Typically a number less than 1 (e.g. 0.1)

Collected into a vector

We follow the general assumption that

The probability of sample containing community :

Probability between 0 and 1

Vector of the distribution of probabilities of all communities in sample : (or simply )

The vector of community probabilities within a sample is a -dimensional Dirichlet distribution:

The probabilities of samples containing the suite of communities is housed in an matrix () of values, such that each row is described by the above equation.

The prior weight of species in a community:

Collected into a vector

We follow the general assumption that

The probability of community containing species :

Probability between 0 and 1

Vector of the distribution of probabilities of all species in community : (or just )

The vector of species probabilities within a community () is described by an -dimensional Dirichlet distribution:

The probabilities of communities containing the suite of species is housed in an matrix () of values, such that each row is described by the above equation.

The community identity of organism in sample is represented by and is defined by a -dimensional categorical distribution:

Because there are varying numbers of organisms observed during each sample, we avoid using a matrix structure, and instead retain the vector structure for , notating the community identities of all organisms within sample (or simply ) and the community identities of all organisms across all samples (or just ), which is an -length vector.

The species identity of organism in sample () is defined by an -dimensional categorical distribution, conditioned on the community identity:

We use a parallel structure for as was used for , notating the species identities of all organisms within sample (or just ) and the community identities of all organisms across all samples (or simply, ), which is an -length vector. is what is observed, and what we will use to make inference about the underlying parameters and latent states based.

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

which requires an estimation of the parameters and . The likelihood () of the fitted parameters given the data is equal to the probability of the suite of samples given the parameters:

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

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

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

The general organism-level species-identity distribution is composed of the community-by-community product of the species identity given the community identity and the unknown parameter () and the community identity given the latent parameter (, ), 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

The probability of the entire set of samples under the unknown parameters and is then 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).

This is weird tho, since it looks like gamma really just is a single parameter, but then you get a matrix out that’s called gamma, but really is it? Or is it a posterior estimation of z?

To accomplish this, we endow the model with free variational parameters:

is a Dirichlet parameter that describes the community distribution and are categorical parameters that define the distribution of community states

and , generating a family of distributions (notated by to distinguish from ) for the latent variables.

We treat as a random variable, held in a -dimension matrix () of values with a posterior distribution (derived below) that is dependent upon the variational parameter , which is held in a -dimension matrix () of values, which are in turn dependent upon the hyperparameter .

In particular, we consider a variational approximation that places a distribution on , , and