**Coupling Latent Dirichlet Allocation to multivariate change point time series models to study macroecological time series patterns**

Juniper L. Simonis1,†, Erica M. Christensen1,2, David J. Harris1,3, Renata Diaz1, Hao Ye1, Ethan P. White1, and S. K. Morgan Ernest1

1Weecology Lab, University of Florida

2Current institution: the Jornada Rangeland Research Program, New Mexico State University

3Current institution: Wayfair, Inc.

†Corresponding author: juniper.simonis@weecology.org

**INTRODUCTION**

We endeavor to develop methods for analyzing time series of high-dimensional data, and are motivated context by the study of ecological communities comprised of species, where samples of organisms are collected over time (Christensen *et al.* 2018). Specifically, we are interested in determining if the composition of a community (relative composition of species) changes over the course of the study, and if it does, we seek to quantify those dynamics, which may occur abruptly (Williams *et al.* 2011) or smoothly (Tingley *et al.* 2009). However, ecological communities are typically composed of many species relative to the number of samples collected (*i.e.,* the data are high-dimensional; McCune and Grace 2002), 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. The first stage (LDA) uses Latent Dirichlet Allocation (Blei *et al.* 2003) to find the optimally simplified, latent representation of the data, which is then analyzed in the second stage (TS) using Bayesian change point Time Series models (Western and Kleykamp 2004) that we extend for multinomial data using softmax regression (Venables and Ripley 2002). This manuscript describes the two-stage LDATS model in a unified mathematic setting and accompanies an LDATS R package (Simonis *et al.* *In Development*)

Latent Dirichlet Allocation is a hierarchical Bayesian model that uses a generative classifier (*i.e.*, it uses the joint probability of the inputs and outputs with Bayes’ rules to calculate the posterior, as opposed to a discriminative classifier that calculates the posterior directly; see Ng and Jordan 2002) to decompose high-dimension data into a reduced number of latent groups (Blei *et al*. 2003) also known as topics (*i.e.*, LDA is a topic model). The LDA model originally derived and developed by Blei *et al.* (2003) for analysis of textual corpora has since been successfully applied to ecological data (Valle *et al.* 2014, Christensen *et al.* 2018, Valle *et al.* 2018), the domain of interest here. In relation to the original linguistic LDA description and notation (Blei *et al.* 2003), organisms within a sample are like words in a document, species are like terms (word options) in a vocabulary, component communities are like linguistic topics, samples are like documents using the terms, and the whole study is like the corpus of documents (Valle *et al.* 2014). Importantly, LDA is a mixed-membership model, such that terms (species) can be associated with multiple topics (component communities). For the sake of maintaining the relationship between our two-stage LDATS model and the topic model derivation of LDA, we retain the original naming (*e.g.*, observations of words within documents, latent grouping of terms into topics) of Blei *et al.* (2003)

The TS models used here to analyze the decomposed (via LDA) sample data build upon the Bayesian change point model of Western and Kleykamp (2004), which allows for discrete (change point) and continuous temporal changes as well as covariate impacts. This approach allows estimation of the continuous dynamics and covariate impacts to be estimated unconditionally with respect to discrete changes (*i.e.*, the model includes change point uncertainty when estimating regression coefficients; Western and Kleykamp 2004), but had a few components that needed expansion for our application. The original model included a single change point (Western and Kleykamp 2004), but we recognize that ecological communities can undergo multiple discrete shifts during a study (including temporary changes) or may not undergo any abrupt shifts (*i.e.,* change gradually or not at all; Ratajczak, *et al.* 2018). Therefore, we expanded the model to include potentially 0, as well as multiple (but currently restricted to a maximum of five for purely computational reasons), change points. Further, the original model assumed a univariate normal response variable (Western and Kleykamp 2004), whereas the output from the LDA in Stage 1 are the proportional parameters of a multinomial response, which are multivariate and non-normal. For use in LDATS, we therefore generalized (à la generalized linear modeling; McCullagh and Nelder 1989) the model of Western and Kleykamp (2004) using softmax regression (Venables and Ripley 2002) to predict multinomial probability variables. The TS models are fit with Bayesian techniques using parallel tempering Markov Chain Monte Carlo (ptMCMC) methods (Earl and Deem 2005) to locate the change points and neural networks (Ripley 1996) to estimate continuous time and covariate parameters.

By combining dimension reduction and time series analysis into a single mathematical framework and software pipeline, LDATS provides a robust and user-friendly methodology for evaluating complex dynamics of high-dimension timeseries (Fig. 1). Although we understand the pressing need for these models within ecology (Andersen *et al.* 2009, Karssenberg *et al.* 2017, Ratajczak, *et al.* 2018), we recognize their application may be broader, given the interest in regime shifts in financial, political, and engineering sectors, for example (Scheffer 2009, Gal and Anderson 2010). We therefore keep the model and its coding implementation as general as possible to facilitate application to other systems of interest. This manuscript details the mathematical model underlying the LDATS methodology and an example application from the motivating system (based on Christensen *et al.* 2018). The accompanying LDATS software package implements the model in R (R Core Team 2018). The code is currently available via the Weecology Lab GitHub repository (https://github.com/weecology/LDATS).

**MODEL DEVELOPMENT**

*General Terminology and Notation*

Because of the overlap in notation between LDA (Blei *et al.* 2013), the time series models used here (Western and Kleykamp 2004), and the ptMCMC method used to fit the time series models (Earl and Deem 2005) (*e.g.*, all use but with different meanings), we create a notational set for use here with an effort to minimize name reuse. Given that LDATS specifically uses LDA as the first stage and that our methods build upon topic models, in instances of notational overlap between the LDA and TS or ptMCMC components, we defer to the LDA usage. We do make one important deviation from the original LDA notation (Blei *et al.* 2003), however, to clarify the dimensionality of state variables and parameters. Specifically, we use the lowercase, regular type letter (*e.g.,* ) to indicate a singular value; the lowercase, boldface letter (*e.g.,* ) to indicate a vector of values; and the capital, boldface letter (*e.g.,* ) to indicate a matrix of values.

A corpus (set of documents) consists of total documents comprising total words from total terms. Each document (in ) consists of words ( in ) assigned to one of in terms. The total number of words in the corpus is the sum of the words within each document. The weight of document is the number of words it has relative to the maximum number of words in any document:

|  |  |  |
| --- | --- | --- |
|  |  | [1] |

allowing us to account for variable numbers of words among documents (with vector , or just ).

LDA involves grouping the terms into in total latent (unobserved) component topics, where “component topic” means a group of terms that tend to be found together in specific proportions. The allocation process (Blei *et al.* 2003)allows individual terms to be assigned to multiple component topics. The total number of latent topics is also unknown, and for the present approach is fixed *a priori* within a given Stage 1 (LDA) model in (note the difference between , the number of Stage 1 models and , the number of documents) as .

Each word within a document has an observed term identity and a latent topic membership that is fit in the Stage 1 model as . Because there are varying numbers of words in each document, we use a vector structure to hold word-level data across the corpus. The term identities of all words within document are (or ) and the term identities of all words across all documents are (or ), an -length vector. Similarly, the topic identities of all words in document are (or ) and the topic identities of all words across all documents are (or ), an -length vector. Thus  contains the topic identity (latent state) and the term identity (observed state) for all words in the corpus. Note that the term identity does not contain a rank-index subscript associated with the model (), indicating that the term identity stays the same for all models, whereas the topic identity does have the subscript, indicating that it varies among the Stage 1 models.

We are interested in temporal changes in topic composition, and so define the time of document to be and the vector of all document times to be or simply . defines the temporal relationship among documents, is used to locate change points, and presently must be a discrete (or discretizable) variable. For a given time series model (in ), we also collate total covariates (including an overall intercept), indexed as in , and measured for each document. The value of a particular covariate for a specific document is  and the set of covariates for the document is a vector or simply . All of the covariates (including the intercept) across all of the documents are held in , an matrix, that can vary among the time series models (hence the rank-index subscript).

*Stage One: Dimension Reduction*

Latent Dirichlet Allocation

The first stage of the LDATS analysis is focused on reducing the raw, high-dimensional data (counts of terms in documents over time) to a lower dimensional representation of the information contained in the data using Latent Dirichlet Allocation (Blei *et al.* 2003). Specifically, we use the Variational Expectation Maximization (Jordan *et al.* 1999) version of the LDA model derived and developed first by Blei *et al.* (2003).

For a Stage 1 model with a total number of topics , the word-level topic distribution within a document (*i.e.*, the allocation of words among the possible topics) is a -dimension categorical random variable described by probabilities held in vector () and collated across documents into the matrix . Thus, the realized topic identity of word within a document is

|  |  |  |
| --- | --- | --- |
|  |  | [2] |

The vector of topic probabilities within a document () is defined by a -dimensional Dirichlet distribution with concentration parameters ,which we assume do not change among documents (*i.e.*, ) and are symmetric (*i.e.,* ), reducing the set to a single unknown parameter for the model:

|  |  |  |
| --- | --- | --- |
|  |  | [3] |

The word-level term distribution (*i.e.*, the allocation of words to the possible terms) within a document is a -dimension categorical random variable contingent upon the topic identity of the word and defined by probabilities , where . The probabilities across all topics within a document are held in a matrix (), which we assume is constant across documents, (). The word-level term identity is defined based on either an unknown or known topic identity:

|  |  |  |
| --- | --- | --- |
|  |  | [4] |

is therefore defined by unknown parameters  (a scalar) and  (a matrix).

The inferential problem of interest therefore lies in determining the posterior distribution of the latent topic probabilities and states , given observations and estimated parameters and :

|  |  |  |
| --- | --- | --- |
|  |  | [5] |

which obviously necessitates values for and . From a parameter estimation standpoint then, we are concerned with the probability of observations given the parameters and , or the likelihood () of the parameters given the data (). And specifically for optimization, we need to determine the log-likelihood () of the parameters given the data (). We decompose the probability of the corpus given the parameters (steps in Appendix 1), resulting in

|  |  |  |
| --- | --- | --- |
|  |  | [6] |

which highlights the problematic coupling of (and thus ) and in the summation over latent topics (Blei *et al.* 2003), which prevents direct, tractable estimation of parameters.

To circumvent this issue, we use a variational approximation (Jordan *et al*. 1999) 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; Appendix 2). To accomplish this, we endow the model with free latent variational parameters  and (Appendix 2) that decouple the parameters and characterize a family of distributions ( to distinguish from ) providing a lower bound on the log likelihood (Jordan *et al.* 1999, Blei *et al.* 2003). 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 . This estimation procedure is executed using the LDA function in the topicmodels package (v0.2-7; Grun and Hornik 2011) in R (v 3.5.1; R Core Team 2018), which leverages C code written by Blei *et al.* (2003).

Multi-Model Inference

Given the fit of a specific Stage 1 model (), we can then consider multiple Stage 1 models to determine the most parsimonious number of topics . Specifically, we use as our Stage 1 model selection criterion (Christensen *et al*. 2018), defined for a specific LDA model :

|  |  |  |
| --- | --- | --- |
|  |  | [7] |

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

|  |  |  |
| --- | --- | --- |
|  |  | [8] |

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 topics () using different starting values, assigned through the random number generator seed (). Specifically, we use replicates ( in ) at each number of topics from 2 to , the total number of topics to be explored. The minimal number of topics is set to 2 by the current coding implementation of the LDA algorithm (Blei *et al.* 2003, Grün and Hornik 2011), although the underlying mathematics can include the limiting case of a single topic (*i.e.*, no dimension reduction). Thus, the total number of models in Stage 1 () is

|  |  |  |
| --- | --- | --- |
|  |  | [9] |

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

|  |  |  |
| --- | --- | --- |
|  |  | [10] |

and has the corresponding set of parameters .

Having found the optimal model, we can obtain the posterior point estimates for the document-level topic 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 ) within documents so that they sum to one and are proper proportions, notating the normalized parameter with the overbar accent as . Thus, for a general document :

|  |  |  |
| --- | --- | --- |
|  |  | [11] |

The posterior point estimates of the topic proportions across the documents are held in the matrix , which corresponding to the optimal (according to based on VEM inference) decomposition of the word-level data to topic-level data. This matrix forms the multivariate response variable analyzed in the time series model, as outlined in the next section.

*Stage Two: Multinomial Time Series*

Change Points: Segmenting the Time Series

The second stage of the LDATS model focuses on analyzing the time series of topic proportions estimated by the LDA (). Although the analysis of the topic proportions is via time series models, the times of the documents () *per se* do not necessarily enter every Stage 2 (time series) model ( in ) directly. Rather, depending upon the model, the time of the documents 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), or may not impact at all (in the case of a model with no change points and no continuous time impacts). Presently, temporal autocorrelation is not included in the time series models, but is planned for future work (see **FUTURE DEVELOPMENTS**).

A given Stage 2 model has a non-negative integer number of discrete change points () that divide the time series into distinct segments or “chunks” ( in ), where the number of chunks () is always one more than the number of change points (). If there are change points (*i.e.,* ), then their locations (for the in change points) are unknown parameters to be estimated (Western and Kleykamp 2004, Christensen *et al*. 2018). A specific change point ’s location is represented by and the set of change points are represented by the –length vector (or ). To define the full deconstruction of the time series into chunks, we augment the vector of change point locations with the time step before the minimum () and the maximum time step () in the series, generating the –length vector , where the overbrace accent references the addition of the fixed time range to the unknown change points. In the instance that there are no change points (*i.e.,* ), is still defined, but now is simply a length-2 vector including the minimum and maximum times, and therefore includes no unknown change point locations to be estimated.

We use the locations to assign the documents to specific chunks via a document segment mapping function (), which returns an indication () if document belongs to chunk (0 for no or 1 for yes). Crucially, the indication depends on the change point locations, , which are unknown and variable, even within the context of a fixed number of change points as defined by . Thus, whether a particular document belongs to a specific chunk of time can differ even if the number of change points stays consistent (*i.e.*, the assignment is conditional on the change point locations). The required strictly Bernoulli distribution of indicator function (returning only a 0 or 1), however, necessitates known change point locations, and so we must define the function in terms of fixed change points that constitute a realization drawn from the underlying multivariate distribution.

For a given assignment of the documents into chunks, we assume a fixed set of change points drawn from the , which we notate as , where the replacement of the subscript with  indicates a realization from the vector distribution of change point locations, given the model . Each realization can be considered a submodel of the model (which only defines the number of change points) that specifies the change point locations and their impacts on the segmentation of the documents. A realization in defines the segments ( in ) and allows for the operation of on the timestamp of the document () and the start () and end () times of a chunk :

|  |  |  |
| --- | --- | --- |
|  |  | [12] |

where the start time for a segment is the first time step after the previous change point and the end time is the timestep of the change point. For each chunk in , produces an -length vector of 0s and 1s ( or just ), which we collate into an matrix () that allows us to identify to which chunk each document belongs and deconstruct the full output matrix from the Stage 1 model () into matrices ( in ), corresponding to the chunks in .

Segment-Level Models: Multinomial Logistic Regression

At this point, we note a few important distinctions between the change point model of Western and Kleykamp (2004) and LDATS with respect to the application of the change points to the time series of data. In Western and Kleykamp’s (2004) model, the indicator function is only applied within the regressor terms but not the error term of a singular regression fit, such that all chunks are fit in one model and have the same error term but different covariate effects. In our model, however, the data are fully separated into multiple independent regression fits by the change points, as a result applying the indicator to both the covariate impacts and error components. This allows us to increase the number of discrete change points in the model beyond the singular point included in the original model (Western and Kleykamp 2004) without suffering rapidly increasing computational costs of large parameter covariance matrices.

Further, the original change point model (Western and Kleykamp 2004) assumed a univariate normal response variable, whereas our response data are multivariate and non-normal. Specifically, our response variable is a set of multinomial probabilities, each of which must be non-negative and which must sum to 1. To accommodate this structure, we take a generalized linear model approach (McCullagh and Nelder 1989) and model the data using a log-linear multinomial (aka multinomial logit or softmax) regression (Ripley 1996, Venables and Ripley 2002) within each given chunk of time.

We seek to predict , the matrix of topic proportions for all of the documents () within a given chunk of time , but we often work with the components (the length- vector corresponding to the topic distribution of a single document  in the chunk) or (the scalar probability that a word in document comes from topic ). Following linear modeling, we wish to define these proportions in terms of predictor covariates ( for a single document or for all documents in the chunk) and coefficients, which we generally notate as , where is a -length column vector of parameters for topic and is a -matrix of parameters for all topics. However, we must acknowledge two constraints on the multivariate response variable: [1] that each component value (topic proportion) be non-negative and [2] that they sum to 1 within a document. We address these issues by using the softmax function on a set of augmented parameters (, , and , where the acute accent indicates the augmentation) that define the first topic as a reference value (Appendix 3; Venables and Ripley 2002). We can use the softmax function to predict topic probabilities at any of the three levels (single topic within a document, all topics within a document, all topics within all documents) of a chunk of documents, given the realization of the model:

|  |  |  |
| --- | --- | --- |
|  |  | [13] |

where indicates the expected (predicted) value(s) of the proportion(s) and its subscript indicates the model being fit for the chunk of time , which is defined as the realization of the Stage 2 model .

This representation is aligned with the generalized linear model equation (McCullagh and Nelder 1989), wherein our link function is the multinomial logit (similar to the binomial logit of a logistic regression) and our inverse link is the softmax function (akin to the logistic function of a logistic regression). Recognizing the uncertainty in the relationship between the observations () and the predictions () allows us to formulate an estimation problem whereby we are interested in finding an optimal set of parameters, (the asterisk indicates the optimal value set), given the data. To do this, we leverage the conditional (in the sense of depending on the full input complex via the softmax function, see Appendix 3) probability for a document given the parameters and covariates for that chunk in that realization of that model (), which is also the conditional likelihood of the parameters, given the covariates and the observed proportions () and is the product of each topic’s predicted probability raised to power of the corresponding observed probability (Berger *et al*. 1996):

|  |  |  |
| --- | --- | --- |
|  |  | [14] |

Of particular importance here is that the “observed” values (which are in fact the estimated latent topic proportions) are probabilities, which means we retain the full version of this equation, as compared to when the observed data are categorized individuals and the likelihood equation can be simplified due to the one-hot coding of the observations (when all states are 0 except one that is 1 for each observation). The conditional likelihood for the set of parameters given topic proportions, covariates, and weights of all documents within the chunk is the weighted joint document-level probabilities (Venables and Ripley 2002):

|  |  |  |
| --- | --- | --- |
|  |  | [15] |

We use the log-likelihood () as our function for optimization:

|  |  |  |
| --- | --- | --- |
|  |  | [16] |

which is equivalent to the negative cross-entropy between the observed () and predicted () distributions (Berger *et al.* 1996, Malouf 2002).

Given that the log-likelihood is concave, there exists a set of specific parameter values where the multivariate derivative of the log-likelihood function is 0. However, the log-likelihood function does not have a closed form solution (Malouf 2002) and the log-likelihood surface may be quite complex, given the high number of potentially colinear parameters (Venables and Ripley 2002). Therefore, the parameters are jointly estimated as the maximum *a posteriori* (MAP; *i.e.*, posterior mode; Bassett and Deride 2018) values, where MAP estimation is an extension of maximum likelihood estimation with a regularization (*i.e.*, penalty) to smooth the likelihood surface, thereby addressing over-fitting and avoiding pathological solutions (local minima; Ripley 1993). We penalize the log-likelihood (creating ) using the classical “weight decay” (so called because the regressors in the neural networks employed to fit the model are confusingly named “weights”; Venables and Ripley 2002), which is also known as Tikhonov or L2 regularization (Tikhonov and Arsenin 1977, Ng 2004) or ridge regression (Hoerl 1962, Hoerl and Kennard 1970) in other applied settings, and is equivalent to taking a Bayesian approach by adding a Gaussian prior on the model parameters . A decay parameter (, the inverse variance of the Gaussian prior) scales the penalty, which is the sum of squared coefficients:

|  |  |  |
| --- | --- | --- |
|  |  | [17] |

The original formulation of LDATS (Christensen *et al.* 2018) did not explicitly include the decay penalization (that is, was assumed to be 0, resulting in functionally infinite variance for the prior), but including a small penalty (; Ripley 1993, Ripley 1996) can aid in finding the optimal solution in multinomial regressions (as long as all coefficients have been scaled to about ), and so we now include the option in the software (Simonis *et al.* *In Development*). Following Venables and Ripley (2002), is treated as a fixed value for each multinomial regression, and we also fix the value across the chunks and models (), resulting in an updated penalized log-likelihood function for the documents within segment , which is based on realization of model :

|  |  |  |
| --- | --- | --- |
|  |  | [18] |

which is strictly concave. We find the optimal parameters for the realization of the chunk by maximizing (although we actually minimize ):

|  |  |  |
| --- | --- | --- |
|  |  | [19] |

Generally then, the regression model for a given chunk of time based on a specific realization of a Stage 2 model is fit using the multinom and nnet functions within the nnet package (v7.3-12; Venables and Ripley 2002) in R (R Core Team 2018), which formulate the regression as a single-hidden-layer neural network with skip-layer connections (Venables and Ripley 2002). The solution is found with the gradient-based optimization algorithm known as the Broyden–Fletcher–Goldfarb–Shanno or BFGS Algorithm (Brayden 1970, Fletcher 1970, Goldfarb 1970, Shanno 1970), a quasi-Newtonian iterative searching method for non-linear optimization. The BFGS Algorithm works efficiently by not calculating the Hessian (matrix of partial second derivatives) at every step in the optimization, but rather approximating the Hessian by comparing successive iterations of the Jacobian matrix (matrix of partial first derivatives), whose components cover all combinations of input element (the coefficients associates with each covariate for each topic) and output element (loss). Thus, for each entry in the Jacobian, we must specify with respect which covariate-topic combination we want the partial derivative of the loss with respect to.

The full Jacobian of the loss equation has an extensive derivation (due to the multivariate application of the chain rule) that is based on the nuances of the data set being analyzed, but which collapses neatly because it is sparse. In Appendix 4, we fully derive the Jacobian for the loss function based on probability responses (rather than one-hot vector responses) with linear predictors (covariates) acting on the predicted probability through matrix multiplication with parameters (thereby allowing linear modeling) and with a flexible weight decay penalty. Here we highlight the derivation of the matrix.

To reduce clutter, we rename the cross entropy function ; the softmax function ; matrix multiplication ; and the penalty function , allowing us to write the negative penalized log-likelihood equation for all documents within a time chunk based on the realization of the Stage 2 model as

|  |  |  |
| --- | --- | --- |
|  |  | [20] |

Further, we collapse into , recognizing that the lower-order functions are still executed via the chain rule, arriving at

|  |  |  |
| --- | --- | --- |
|  |  | [21] |

The partial derivative of with respect to a specific parameter associated with covariate and topic is notated as , where the capital refers to the matrix structure being derived and the subscript indicates which covariate-topic combination the partial derivative is with respect to:

|  |  |  |
| --- | --- | --- |
|  | The derivation is distributed throughout the equation via the sum rule | [22] |

|  |  |  |
| --- | --- | --- |
|  |  | [23] |

giving us the components of the generalized partial derivative that need to be calculated for each instance: [1] the partial derivative of the cross-entropy function with respect to covariate and topic () and the partial derivative of the penalty function with respect to covariate and topic (). Combined across the documents within the chunk, we obtain the general entry in the chunk-specific model’s Jacobian for the loss calculation with penalty, a row-matrix with columns, each corresponding to a covariate-topic combination:

|  |  |  |
| --- | --- | --- |
|  |  | [24] |

In combination with our loss function (), this completed Jacobian () allows us to execute the BFGS Algorithm (Brayden 1970, Fletcher 1970, Goldfarb 1970, Shanno 1970) to estimate the optimal values for the parameters () and calculate the loss value for the documents in the chunk of time, , given the realization of the model, (Appendix 4).

Combining Segment-Level Models: Inference About Change Point Locations

The next set of steps is to scale up the chunk-level model of to each of the chunks,and then combine the models across the chunks. Our formulation of the chunk-level model and its expansion up across time chunks differs from Western and Kleykamp (2004) with respect in particular to parameter estimation. Most specifically, in our model, there is no covariance among regression parameters fit in different chunks for the same realization of (*e.g.*, and are uncorrelated), whereas the Western and Kleykamp (2004) model includes covariance among all of the regression parameters by fitting them together in a singular regression that includes the indicator function for each realization of the model. Estimating the covariance among many parameters is computationally very expensive (Genz 1992), and Western and Kleykamp’s (2004) model included a single change point, a few simple predictors, and a normal response variable—all three assumptions we have relaxed, causing the number of regression parameters to increase beyond present computational limitations. Further, given our hypothesis of the change points being discrete and abrupt, it is reasonable to represent that in our model formulation via the assumption of no covariance among parameters in different time chunks.

Building up the time series model across all of the documents, each of the chunks is fit with its own regression, arriving at a specific optimal parameter set () and minimized loss value (). Importantly, these parameter and loss values are estimated conditionally, given the locations of the change points in the realization of the model . As a result, while they inform us about the fit of the full time series, we will need to return to the chunk-level models to decondition the estimates once we have estimated the distributions for the change points. To accomplish that, we sum the minimized loss (negative log-likelihood) values across the chunks to calculate the total penalized log-likelihood of the documents across all segments based on realization of model ():

|  |  |  |
| --- | --- | --- |
|  |  | [25] |

Recognizing that [1] the negative log-likelihood of a chunk depends on the realized change point locations used to break apart the time series () and [2] the realized values are drawn from a multivariate distribution (), we use the negative log-likelihood as the loss function to estimate the multivariate distributions of change point locations unconditionally with respect to within-chunk parameters ( in ). Then, we estimate the within-chunk parameters unconditionally with respect to the change point locations (Western and Kleykamp 2004). However, because the negative log-likelihood function is discontinuous for discrete values of the change point locations, Bayesian methods are required for proper inference.

We use Markov Chain Monte Carlo (MCMC) methods to estimate the posterior distribution of the change points for each model in based on the latent topic probabilities from model . Because the change point negative log-likelihood surface has the high potential for multiple modes (if multiple change points are reasonably likely), we specifically employ parallel tempering MCMC (ptMCMC; also called Metropolis-coupled or replica-exchange MCMC), which endows the search with auxiliary chains that explore the negative log-likelihood surface more rapidly than the focal chain (Swendsen and Wang 1986, Geyer 1991, Earl and Deem 2005). Full ptMCMC methods are detailed in Appendix 5, but we briefly describe the application here.

The prior for the set of change points () is a multivariate discrete distribution, which could take any number of specific formulations (Western and Kleykamp 2004). The original LDATS model (Christensen *et al*. 2018) allowed only a uniform prior, and that requirement is presently maintained in the coding of the package (Simonis *et al*. *In Development*), although relaxing this assumption in the software is planned for future work (see **FUTURE DEVELOPMENTS**). The uniform prior allocates equal probability to each of the discrete time points from the time of the first document () to one time step before the last document (), and a realization then begins by sampling without replacement initial values of the change point locations (, where the in the superscript on indicates that these are initial values), which are then sorted chronologically. Although we are using a uniform prior for the change point locations (*i.e.*, no specific information is included in the prior), the Bayesian approach offers insight above a maximum likelihood approach, in particular with respect to estimating the regression coefficients, given the change point locations (Western and Kleykamp 2004). The Bayesian approach allows us to integrate over the uncertainty in the change point locations when estimating the regression coefficients, compared to a maximum likelihood approach that would condition the regression coefficients on the change point locations.

The search of the negative log-likelihood surface is explored using replica chains ( in ), each with its own “temperature” (), which describes the ability of the chain to search the surface: higher temperature chains have higher variances in their search step sizes (similarly, one can think of higher temperature chains having “flattened” surfaces to search) and are therefore more able to search the breadth of the negative log-likelihood surface. This comes at a cost, however, of instability of the resulting sample, as higher temperature chains are less likely to settle in to a stable distribution. We therefore use a range of temperatures for the chains grouped in a series () to balance search breadth and depth for the within-chain Monte Carlo searches, with the coolest chain () as the focal chain for inferential purposes (Earl and Deem 2005).

The specifics of the temperature regime are critical for fitting a model to data using ptMCMC (Kone and Kofke 2005, Rathone *et al.* 2005). Following the original LDATS code (Christensen *et* al. 2018), the current implementation of the algorithm allows for control over the parameters to facilitate fitting a wide range of potential time series (Simonis *et al*. *In Development*). In particular, the temperature series is defined as

|  |  |  |
| --- | --- | --- |
|  |  | [26] |

where is a sequence of values from to of length , is the exponent controlling the temperature sequence ( produces a geometric series, implies squaring before exponentiating), is the penultimate temperature, and is the ultimate temperature. The control inputs (, , , and ) are available to the user, although they are fixed values for a given fit of the Stage 2 models. A target for future mathematical and coding development is therefore expand the inputs and allow for an adaptive approach to ptMCMC, which will facilitate a more plug-and-play approach to model fitting (where the user does not need to set any control parameters; see **FUTURE DEVELOPMENTS**).

For each MCMC iteration in of realization of model (), there is a set of proposed change points that follow the Monte Carlo evolution based on the multivariate step occurrence and magnitude and which are evaluated based on the Metropolis rule, resulting in a set of accepted change points .

which occur via the Metropolis-Hastings Algorithm (Metropolis *et al*. 1953, Hastings 1970). The chains are allowed to swap configurations with their nearest neighbors in the temperature series between each within-chain Monte Carlo step (following the Metropolis criterion; Metropolis *et al*. 1953), allowing them to share information and search the surface in combination. The temperatures scale the update acceptance rates for each of the chains during chain-specific Monte Carlo steps as well as the swap acceptance rates for the sharing of information among chains in between Monte Carlo steps (hotter chains are more likely to accept update steps and chain swaps; Appendix 5).

Unwind it first!

We collate the parameter sets by row addition of the matrices across the chunks into a single matrix ():

|  |  |  |
| --- | --- | --- |
|  |  | [25] |

***Combine within an iteration and scale up***

***Unconditional estimate of eta***

*Not sure if needed exactly as is*

To do this, we first define an expanded version of the parameter matrix , which has columns and a number of rows equal to the number of coefficients across all of the time chunks:

We next expand the covariates to the block-diagonal matrix , which has a number of columns equal to the number of coefficients across all of the time chunks and rows:

Finally, we compute the full matrix of topic-by-document proportions:

which relates directly to the generalized linear modeling equation that is typically written as or , where is the so-called link function and is the inverse link function. The component from our model missing thus far is uncertainty, which we acknowledge now through the inclusion of the expected value notation ( is the expected value of ):

thereby producing a generalized linear model equation, wherein our link function is the multinomial logit (similar to the binomial logit being the link function of a logistic regression) and our inverse link is the softmax function (akin to the logistic function being the inverse link of a logistic regression).

Now can move to the rest of what W and K use to describe the model

Add the var-cov matrix and the distributions of the priors and the change points

**FUTURE DEVELOPMENTS**

**ACKNOWLEDGEMENTS**

The motivating study—known as the Portal Project—has been funded nearly continuously since 1977 by the National Science Foundation, most recently by DEB-1622425 to S. K. M. Ernest, which also supported (in part) E. Christensen’s time. Much of the computational work (including time of J. Simonis, D. Harris, and H. Ye) was supported by the Gordon and Betty Moore Foundation’s Data-Driven Discovery Initiative through Grant GBMF4563 to E. P. White. R. Diaz was supported in part by a National Science Foundation Graduate Research Fellowship (No. DGE-1315138 and DGE-1842473).

**AUTHOR CONTRIBUTIONS**

J. L. Simonis provided insight on LDA applications and feedback on technical writing during development of the first version of the LDATS model and application, led the coding and mathematical development of the model into an R package, and led the writing on this manuscript. E. M. Christensen led the project during development of the first version of the LDATS model and its application to the Portal data, specifically conceiving the project, coding the pipeline wrappers of the analysis, and writing and editing the first description of the model and its application (Christensen *et al*. 2018). D. J. Harris was involved in developing and applying the first version of the LDATS model, specifically suggesting the LDA and change point approaches, coding the first version of the change point model, and writing and editing the first description of the model (Christensen *et al*. 2018). R. Diaz contributed code to the LDATS package, provided insight into model development, and conducted end-user code application testing. H. Ye contributed code to the LDATS package and insight into data structures and LDA algorithms. E. P. White helped design, troubleshoot, and supervise initial methods development and provided big-picture feedback on development of the R package. S. K. Morgan Ernest provided managerial oversight and feedback on the project in both the initial and second stages of LDATS development, tested applications of the code to data sets, and assisted with writing and editing of the first description of the model and its application (Christensen *et al*. 2018).

**LITERATURE CITED**

Andersen, T., J. Carstensen, E. Hernández-Garcia, and C. M. Duarte. 2009. Ecological thresholds and regime shifts: approaches to identification. *Trends in Ecology and Evolution* **24**:49-57.

Bassett, R. and J. Deride. 2018. Maximum *a posteriori* estimators as a limit of Bayes estimators. *Mathematical Programming, Series B.* https://doi.org/10.1007/s10107-018-1241-0

Berger, A. L., V. J. Della Pietra, and S. A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* **22**:39-71.

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

Broyden, C. G. 1970. The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications* **6**:76–90.

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.

Fletcher, R. 1970. A new approach to variable metric algorithms. *Computer Journal* **13**:317-322.

Gal, G. and W. Anderson. 2010. A novel approach to detecting a regime shift in a lake ecosystem. *Methods in Ecology and Evolution* **1**:45-52.

Genz, A. 1992. Numerical computation of the multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**:141-150.

Geyer, C. J. 1991. Markov Chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. pp 156-163. American Statistical Association, New York, USA.

Goldfarb, D. 1970. A family of variable metric updates derived by variational means. *Mathematics of Computation* **24**:23-26.

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

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57**:97-109.

Hoerl, A. E. 1962. Application of ridge analysis to regression problems. Chemical Engineering Progress. **58**:54-59.

Hoerl, A. E. and R. W. Kennard. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**:55-67.

Jordan, M. Z. Ghahramani, T. Jaakkola, and L. Saul. 1999. Introduction to variational methods for graphical models. *Machine Learning* **37**:183-233.

Karssenberg, D., M. F. P. Bierkens, and M. Rietkerk. 2017. Catastrophic shifts in semiarid vegetation-soil systems may unfold rapidly or slowly. *The American Naturalist* **190**:e145-e155.

Kone, A. and D. A. Kofke. 2005. Selection of temperature intervals for parallel-tempering simulations. *The Journal of Chemical Physics* **122**:206101.

Malouf, R. 2002. A comparison of algorithms for maximum entropy estimation. *Proceedings of the 6th Conference on Natural Language Learning* **20**:1-7.

McCullagh, P. and J. A. Nelder. 1989. Generalized Linear Models. 2nd Edition. Chapman and Hall, New York, NY, USA.

McCune, B. and J.B. Grace. 2002. Analysis of Ecological Communities. MjM Software. Gleneden Beach, OR, USA.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087-1092

Ng, A. Y. and M. I. Jordan. 2002. On discriminative vs. generative classifiers: a comparison of logistic regression and naïve Bayes. In *Advances in Neural Information Processing Systems*, eds: T. G. Dietterich and S. Becker and Z. Ghahramani. pp 841-848. MIT Press.

Ng, A. 2004. Feature selection, L1 vs. L2 regularization, and rotation invariance. *Proceedings of the Twenty-first International Conference on Machine Learning*. http://doi.acm.org/10.1145/1015330.1015435

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ratajczak, Z., S. R. Carpenter, A R. Ives, C. J. Kucharik, T. Ramiadantsoa, M. A. Stegner, J. W. Williams, J. Zhang, and M. G. Turner. 2018. Abrupt change in ecological systems: inference and diagnosis. *Trends in Ecology & Evolution* **33**:513-526.

Rathone, N., M. Chopra., and J. J. de Pablo. 2005. Optimal allocation of replicas in parallel tempering simulations. *The Journal of Chemical Physics* **122**:024111.

Ripley, B. D. 1993. Statistical aspects of neural networks. In *Networks and Chaos: Statistical and Probabilistic Aspects*, eds: O. E. Barndorff-Nielsen, J. L. Jensen, and W. S. Kendall. pp 40-123. Chapman Hall, London.

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.

Scheffer, M. 2009. Critical transitions in nature and society. Princeton University Press, Princeton, NJ.

Shanno, D. F. July 1970. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation* **24**:647-656.

Simonis, J. L., E. M. Christensen, D. J. Harris, H. Ye, R. Diaz, E. P. White, 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

Swendsen, R. H. and J. S. Wang. 1986. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters* **57**:2607-2609.

Tikhonov, A. N.; V. Y. Arsenin. 1977. Solution of Ill-posed Problems. Washington: Winston & Sons

Tingley, M.W., W.B. Monahan, S.R. Beissinger, and C. Moritz. 2009. Birds track their Grinnellian niche through a century of climate change. *Proceedings of the National Academy of Sciences* **106**:19637–19643.

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.

Williams, J.W., J.L. Blois, and B.N. Shuman. 2011. Extrinsic and intrinsic forcing of abrupt ecological change: case studies from the late Quaternary. *Journal of Ecology* **99**:664–677.

**FIGURES**

**Fig 1.** Schematic representation of the data-model relation of the LDATS framework.

**Appendix 1: Decomposing the likelihood of an LDA model given a corpus**

The log-likelihood of the Stage 1 model (governed by parameters and ) given the term assignments of all words within all documents () is the sum of the log-likelihoods of the model given each document’s data :

|  |  |  |
| --- | --- | --- |
|  |  | [A1.1] |

which is the sum of the probabilities of each document’s data, given the parameters:

|  |  |  |
| --- | --- | --- |
|  |  | [A1.2] |

The probability of a document’s data given the parameters is decomposed into the product of [1] the word-level term-identity distributions, given the latent topic state and model parameters () and [2] the sample-level topic distribution given the model parameters (), integrated over the uncertainty in the latent topic distribution, :

|  |  |  |
| --- | --- | --- |
|  |  | [A1.3] |

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

|  |  |  |
| --- | --- | --- |
|  |  | [A1.4] |

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

|  |  |  |
| --- | --- | --- |
|  |  | [A1.5] |

We then scaled this up to the probability of the entire corpus under the unknown parameters and , which is the product of the sample-level probabilities given those parameters:

|  |  |  |
| --- | --- | --- |
|  |  | [A1.6] |

The probabilities for in cannot be tractably estimated due to the coupling of (and thus ) and in the summation over latent topics (Blei *et al.* 2003). To address this, we use the established variational approximation with an expectation-maximization routine (aka VEM for Variational Expectation Maximization; Blei *et al.* 2003), as outlined in Appendix 2.

**Appendix 2: Variational Expectation Maximization estimation of a Latent Dirichlet Allocation**

For a Stage 1 LDA model , is an -matrix akin to , where row corresponds to document : (or ), but contains the concentration parameters of a -dimension Dirichlet distribution and therefore are not constrained to sum to 1. is an -matrix, where the rows correspond to the words across documents (indexed akin to and ) and the columns correspond to the topics. is the probability that word within document is from topic , and (or ) is a -length vector of probabilities defining the categorical distribution of that word’s topic identity (). contains document-specific matrices (each is and notated ). In comparison to and the variational parameters ( and ) are document-specific and not problematically coupled.

For a specific document , the variational distribution () is

|  |  |  |
| --- | --- | --- |
|  |  | [A2.1] |

As the Expectation (“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 the true posterior :

|  |  |  |
| --- | --- | --- |
|  |  | [A2.2] |

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 topic allocation of each word () are updated based on the topic distribution for the document ():

|  |  |  |
| --- | --- | --- |
|  |  | [A2.3] |

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

|  |  |  |
| --- | --- | --- |
|  |  | [A2.4] |

And then, the parameters describing the topic distribution for the document () are updated based on the word-level topic distributions for the sample ():

|  |  |  |
| --- | --- | --- |
|  |  | [A2.5] |

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 documents being equal to the sum of the log likelihoods for the individual documents, the variational lower bound () for all of the documents is equal to the sum of the variational lower bounds for the individual documents:

|  |  |  |
| --- | --- | --- |
|  |  | [A2.6] |

The complete E-Step maximizes this overall lower bound with respect to the full variational parameters and given the main model parameters.

The Maximization Step (“M-Step”) maximizes the overall lower bound with respect to the main model parameters and 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 topic-level term distribution () can be written analytically:

|  |  |  |
| --- | --- | --- |
|  |  | [A2.7] |

where is an indicator variable based on the term identity () of the observed word ():

|  |  |  |
| --- | --- | --- |
|  |  | [A2.8] |

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

|  |  |  |
| --- | --- | --- |
|  |  | [A2.9] |

until convergence, where represents the full matrix element-wise partial derivatives of a multivariate function, such that is the Jacobian matrix and is the Hessian matrix. 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).

**Appendix 3: Softmax regression**

We accommodate the multivariate proportional responses in our time series model by using multinomial logistic regression, also known as Softmax regression. The model is “log-linear” in that it relates the log of the expected proportion () to the linear predictors (), although we formulate our model as the expected proportion () being a function of exponentiated predictors . And we handle the sum-to-1 constraint by normalizing with a document-specific partition function :

|  |  |  |
| --- | --- | --- |
|  |  | [A3.1] |

Given the sum-to-1 constraint (), we can simply define , where we replace the with to avoid confusion with the focal topic . This produces a generalized equation that is often referred to as the softmax function:

|  |  |  |
| --- | --- | --- |
|  |  | [A3.2] |

However, because of the sum-to-1 constraint, only of the proportions ( in ), and by extension only of the parameter vectors ( in ), are uniquely identifiable. Thus, we define an augmented parameter vectors (in ), where

|  |  |  |
| --- | --- | --- |
|  |  | [A3.3] |

setting the parameters associated with the first topic () to 0 (, ). This reduces the number of free parameter vectors (and number of proportions estimated) by 1 to the ( we are able to fit for this specific chunk of time , resulting in the modified probability equation

|  |  |  |
| --- | --- | --- |
|  |  | [A3.4] |

We combine all of the parameter vectors to (including the vector of 0s in ) into a matrix ( is capital ), which has columns and a number of rows equal to the number of coefficients in model () including the intercept (*i.e.*, the length of ).

This allows us to further condense the expected probability equation to

|  |  |  |
| --- | --- | --- |
|  |  | [A3.5] |

thereby facilitating use of the generalized linear modeling framework. We expand the model to predict the proportions across all of the topics within the document, which means we can drop the input and produce the full set of values from the softmax function, which is a length- row vector corresponding to the topic distribution of a single document:

|  |  |  |
| --- | --- | --- |
|  |  | [A3.6] |

We then expand the model across all documents within the chunk of time

|  |  |  |
| --- | --- | --- |
|  |  | [A3.7] |

where the covariates are held in a matrix () with the number of columns equal to the number of coefficients () and the number of rows equal to the number of documents in the chunk (). That is, is a series of row vectors. This equation relates directly to the generalized linear modeling equation that is typically written as or , where is the so-called link function and is the inverse link function (McCullagh and Nelder 1989).

**Appendix 4: Derivation of the Jacobian for the loss function.**

To construct the Jacobian of the loss function (negative penalized log-likelihood ), we first recognize the order of operations of the component multivariate functions, and to reduce clutter in our derivation, we name the cross entropy function , the softmax function , the matrix multiplication , and the penalty function :

|  |  |  |
| --- | --- | --- |
|  |  | [A4.1] |

which can be condensed via the summation across topics to

|  |  |  |
| --- | --- | --- |
|  |  | [A4.2] |

This highlights the chained aspect of the non-penalty functions (the cross entropy is calculated using the output of the softmax, which uses the output of the matrix multiplication), whose derivative can be expanded using the multivariate chain rule. For two general functions and chained as (where contains the multivariate input values), we can write the function composite using the ring operator as . We then take the multivariate derivative (denoted as function ) of the composite:

|  |  |  |
| --- | --- | --- |
|  |  | [A4.3] |

where is the dot product operator. Thus, the derivative of of of is the dot product of the derivative of evaluated at of and the derivative of evaluated at . Using the chain rule, we now expand the derivative of the loss function applied to a specific document within a specific chunk (). We start by expanding the outer layers ( and ):

|  |  |  |
| --- | --- | --- |
|  |  | [A4.4] |

Note that we drop and from the notation here because we are taking the derivative with respect to, and so while both are relevant to the functions ( for and for ), they are not the focus of the derivation and so are functionally constants. We next expand the inner layers ( and ) by working only with the right-hand-side of the dot product:

|  |  |  |
| --- | --- | --- |
|  |  | [A4.5] |

Combining these chained results gives the full Jacobian for the non-penalized component of the loss function applied to a single document:

|  |  |
| --- | --- |
|  | [A4.6] |

which is the dot product between [1] the dot product between [a] derivative of the cross-entropy function (evaluated at the softmax of the matrix multiplication of the coefficients) and [b] the derivative of the softmax (evaluated at the matrix multiplication of the coefficients), and [2] the derivative of the matrix multiplication evaluated at the coefficients.

We now define the derivative matrices (Jacobians) of each of the functions , , and . To aid in this, we consider that the Jacobian of a function contains the partial derivatives of each output (, a general component of ) with respect to each input ( a general component of , which can be written generally as or . The Jacobian for a given function then maps the input to the output, and so has dimensions equal to the number of output classes the number of input classes.

The function maps the matrix () to the dimensions of () by left-multiplying by the covariate row matrix (). Thus, its Jacobian has rows and columns:

|  |  |  |
| --- | --- | --- |
|  |  | [A4.7] |

For notekeeping purposes, entry in the coefficient matrix () corresponds to the column in the Jacobian. In effect, the coefficient matrix is linearized in column-major order (iterating through all covariates within a given topic before progressing to the next topic). Recall that the matrix multiplication used to generate an output element (row) (for in ) is just a linear combination of components

|  |  |  |
| --- | --- | --- |
|  |  | [A4.8] |

and therefore, the partial derivative of the output element with respect to an input element is simply the relevant covariate or 0 (when beyond the relevant part of the Jacobian):

|  |  |  |
| --- | --- | --- |
|  |  | [A4.9] |

Moving next to the softmax function, maps to , both of which are of dimension , because we are working within a single document. Thus, its Jacobian has rows and columns:

|  |  |  |
| --- | --- | --- |
|  |  | [A4.10] |

We can write a generalized equation for the entries by describing the partial derivative of output with respect to input , :

|  |  |  |
| --- | --- | --- |
|  |  | [A4.11] |

We decompose the generalized entry using the quotient rule, where for a function that is equal to the ratio of two other functions: , the derivative of the function is

|  |  |  |
| --- | --- | --- |
|  |  | [A4.12] |

Here, and and we differentiate each with respect to :

|  |  |  |
| --- | --- | --- |
|  |  | [A4.13] |

Regardless of the specific input that we are computing the partial derivative for with respect to, the value will always be :

|  |  |  |
| --- | --- | --- |
|  |  | [A4.14] |

For , however, the value of the partial derivative is 0 unless , in which case it is :

|  |  |  |
| --- | --- | --- |
|  |  | [A4.15] |

Thus, when ,

|  |  |  |
| --- | --- | --- |
|  |  | [A4.16] |

Similarly, when ,

|  |  |  |
| --- | --- | --- |
|  |  | [A4.17] |

Combining these conditions, we have

|  |  |  |
| --- | --- | --- |
|  |  | [A4.18] |

We can use the Kronecker delta function to condense the conditional equation to

|  |  |  |
| --- | --- | --- |
|  |  | [A4.19] |

where

|  |  |  |
| --- | --- | --- |
|  |  | [A4.20] |

The function set , which we simplify to for notational clarity, maps the matrix () to the dimensions of () and so, like , its Jacobian has rows and columns. We combine the Jacobians of and to define the Jacobian of :

|  |  |  |
| --- | --- | --- |
|  |  | [A4.21] |

For a general entry, the partial derivative is

|  |  |  |
| --- | --- | --- |
|  |  | [A4.22] |

Recalling that is 0 except for when (in which case it is ), we can simplify this equation to

|  |  |  |
| --- | --- | --- |
|  |  | [A4.23] |

And recalling that , we can write this equation as

|  |  |  |
| --- | --- | --- |
|  |  | [A4.24] |

Finally, The cross-entropy function, , maps (dimension ) to the cross-entropy (loss) for the document, which is a scalar value. Thus, the Jacobian of is of dimension :

|  |  |  |
| --- | --- | --- |
|  |  | [A4.25] |

A general entry in the Jacobian (*i.e.*, the partial derivative of the cross entropy loss with respect to topic ’s probability) is

|  |  |  |
| --- | --- | --- |
|  |  | [A4.26] |

Notably, the only instance where appears in the function being derived is when , in which case, the derivative is

|  |  |  |
| --- | --- | --- |
|  |  | [A4.27] |

Otherwise (*i.e.*, when ), the function being derived is a constant and therefore has a derivative = 0.

Combining these conditions, we have

|  |  |  |
| --- | --- | --- |
|  |  | [A4.28] |

For notation, we identify the element as .

We can verify that the dimensionalities of the Jacobians are proper for combination via dot products: . is , is , and is . Thus, each of the two dot products has proper component matrices. In addition, the resulting matrix is , which heuristically matches the fact that the composite of the three functions maps the set of parameters ( in total) to a single scalar value of the cross-entropy loss. Having verified the dimensionsNext, we combine the elements across the three Jacobians to determine the derivative of (of of) with respect to : . To reduce clutter, we simplify the notation of to and the corresponding Jacobian of interest is then . At this condensed level, maps the parameters (each entry in ) to a scalar output (cross-entropy loss), so the resulting Jacobian is of dimensions :

|  |  |  |
| --- | --- | --- |
|  |  | [A4.29] |

Similar to the rows in , the single row of partial derivatives in corresponds to the column-major order linearized . Following the indexing of , we will index with and (as ), where refers to column (element) in . For a general entry , then

|  |  |  |
| --- | --- | --- |
|  |  | [A4.30] |

Since only the element of is non-0, in which case it is , we can simplify this equation to

|  |  |  |
| --- | --- | --- |
|  |  | [A4.31] |

Substituting in the derivative of the softmax (of the matrix multiplication) of the parameters,

|  |  |  |
| --- | --- | --- |
|  |  | [A4.32] |

Noting that, by our definition, , we can further simplify this equation:

|  |  |  |
| --- | --- | --- |
|  |  | [A4.33] |

This row matrix holds the partial derivatives associated with the parameters and the topics. With these functions, we can calculate the full set of partial derivatives required to evaluate the Jacobian for the loss equation without the penalty, the remaining component to be added.

Following the sum rule, given that we need to calculate the partial derivative of the loss with respect to each of the model parameters , we start with

|  |  |  |
| --- | --- | --- |
|  |  | [A4.34] |

and take the derivative with respect to a particular parameter in the set,

|  |  |  |
| --- | --- | --- |
|  |  | [A4.35] |

Therefore, we need to calculate the partial derivative of the penalty function with respect to a particular parameter . Remembering that the penalty function is,

|  |  |  |
| --- | --- | --- |
|  |  | [A4.36] |

the partial derivative with respect to is

|  |  |  |
| --- | --- | --- |
|  |  | [A4.37] |

which is

|  |  |  |
| --- | --- | --- |
|  |  | [A4.38] |

Given that , and that , the partial derivative of loss with respect to input () is now fully defined:

|  |  |  |
| --- | --- | --- |
|  |  | [A4.39] |

This equation determines the gradient of the loss across with respect to each input (parameter-topic combination ) within a single document. Acknowledging multiple documents fall under the same parameter combination, we simply sum across all :

|  |  |  |
| --- | --- | --- |
|  |  | [A4.40] |

and distribute the summation

|  |  |  |
| --- | --- | --- |
|  |  | [A4.41] |

to achieve the completely defined general entry to the Jacobian for the penalized loss equation used to fit a multinomial model to a set of documents’ topic proportions.

**Appendix 5: Parallel Tempering Markov Chain Monte Carlo**