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

Juniper L. Simonis[1,†], Erica M. Christensen[1,2], David J. Harris[1,3], Renata Diaz[1], Hao Ye[1], Ethan P. White[1], and S. K. Morgan Ernest[1]

[1]Weecology Lab, University of Florida
[2]Current institution: the Jornada Rangeland Research Program, New Mexico State University
[3]Current 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 (Ruggieri 2013). 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 β but with different meanings), we create a notational set for use here with an effort to minimize name reuse (Table 1). 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.,* β) to indicate a singular value; the lowercase, boldface letter (*e.g.,* $\boldsymbol{\beta}$) to indicate a vector of values; and the capital, boldface letter (*e.g.,* $\boldsymbol{B}$) to indicate a matrix of values.

A corpus (set of documents) D consists of M total documents comprising N total words from V total terms. Each document d (in $1 \ldots M$) consists of N_d words (n in $1 \ldots N_d$) assigned to one of v in $1 \ldots V$ terms. The total number of words in the corpus is the sum of the words within each document. The weight of document u_d is the number of words it has relative to the maximum number of words in any document:

$$u_d = \frac{N_d}{\arg\max_{1 \le j \le M}(N_j)} \quad [1]$$

allowing us to account for variable numbers of words among documents (with vector $\boldsymbol{u}_{1 \ldots M}$, or just $\boldsymbol{u}$).

LDA involves grouping the terms into i in $1 \ldots k$ 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 m_1 in $1 \ldots \mathcal{M}_1$ (note the difference between $\mathcal{M}_1$, the number of Stage 1 models and M, the number of documents) as k_{m_1}.

Each word within a document has an observed term identity w_{d,n_d} and a latent topic membership that is fit in the Stage 1 model m_1 as z_{m_1,d,n_d}. 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 d are $\boldsymbol{w}_{d,1\ldots N_d}$ (or $\boldsymbol{w}_d$) and the term identities of all words across all documents M are $\boldsymbol{w}_{1\ldots M,1\ldots N_d}$ (or $\boldsymbol{w}$), an N-length vector. Similarly, the topic identities of all words in document d are $\boldsymbol{z}_{m_1,d,1\ldots N_d}$ (or $\boldsymbol{z}_{m_1,d}$) and the topic identities of all words across all documents M are $\boldsymbol{z}_{m_1,1\ldots M,1\ldots N_d}$ (or $\boldsymbol{z}_{m_1,}$), an N-length vector. Thus $\boldsymbol{z}_{m_1}$ contains the topic identity (latent state) and $\boldsymbol{w}$ 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 (m_1), indicating that the term identity stays the same for all models, whereas the topic identity does have the m_1 subscript, indicating that it varies among the $\mathcal{M}_1$ Stage 1 models.

We are interested in temporal changes in topic composition, and so define the time of document d to be t_d and the vector of all document times to be $\boldsymbol{t}_{1\ldots M}$ or simply $\boldsymbol{t}$. $\boldsymbol{t}$ 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 m_2 (in $1 \ldots \mathcal{M}_2$), we also collate C_{m_2} total covariates (including an overall intercept), indexed as c in $1 \ldots C_{m_2}$, and measured for each document. The value of a particular covariate c for a specific document d is $x_{m_2,d,c}$ and the set of C_{m_2} covariates for the document is a vector $\boldsymbol{x}_{m_2,d,1\ldots C_{m_2}}$ or simply $\boldsymbol{x}_{m_2,d}$. All of the covariates (including the intercept) across all of the documents are held in $\boldsymbol{X}_{m_2}$, an $M \times (C_{m_2} + 1)$ matrix, that can vary among the $\mathcal{M}_2$ time series models (hence the rank-index subscript).

Stage One: Dimension Reduction

Latent Dirichlet Allocation: Single Model

The first stage of the LDATS analysis reduces 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 (LDA; 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 m_1 with a total number of topics k_{m_1}, the distribution of topics within a document d is a k_{m_1}-dimension categorical random variable described by probabilities $\theta_{m_1,d,1} \ldots \theta_{m_1,d,k_{m_1}}$ held in the vector $\boldsymbol{\theta}_{m_1,d}$ ($\sum \boldsymbol{\theta}_{m_1,d} = 1$) and collated across documents into the $M \times k_{m_1}$ matrix $\boldsymbol{\theta}_{m_1}$. Thus, the realized topic identity (z) of word n within document d under model m_1 is

$$z_{m_1,d,n} \sim \mathrm{Cat}_{k_{m_1}}\left(\boldsymbol{\theta}_{m_1,d}\right) \quad [2]$$

The vector of topic probabilities within a document ($\boldsymbol{\theta}_{m_1,d}$) is defined by a k_{m_1}-dimensional Dirichlet distribution with concentration parameters $\boldsymbol{\alpha}_{m_1,d}$, which (following Blei *et al.* 2003) we assume do not change among documents (*i.e.*, $\boldsymbol{\alpha}_{m_1,1} = \cdots = \boldsymbol{\alpha}_{m_1,M} = \boldsymbol{\alpha}_{m_1}$) and are symmetric (*i.e.,* $\alpha_{m_1,d,1} = \cdots = \alpha_{m_1,d,k_{m_1}} = \alpha_{m_1}$), reducing the set to a single unknown parameter α_{m_1} for the model. Thus, the topic probabilities are

$$\boldsymbol{\theta}_{m_1,d} \sim \mathrm{Dir}_k\left(\alpha_{m_1}\right) \quad [3]$$

which is drawn separately for each document (Blei *et al.* 2003). The word-level term distribution within a document is a V-dimension categorical random variable contingent upon the topic identity of the word and

defined by probabilities $\beta_{m_1,d,1,1} \ldots \beta_{m_1,d,k_{m_1},V}$, where $\sum_v \boldsymbol{\beta}_{m_1,d,i} = 1$. The probabilities across all topics within a document are held in a $k_{m_1} \times V$ matrix ($\boldsymbol{B}_{m_1,d}$), which we assume is constant across documents, ($\boldsymbol{B}_{m_1,d} = \cdots = \boldsymbol{B}_{m_1,M} = \boldsymbol{B}_{m_1}$; Blei *et al.* 2003). The word-level term identity (w) is then generally defined:

$$w_{d,n} \sim \text{Cat}_V\left(z_{m_1,d,n}, \boldsymbol{B}_{m_1}\right) \quad [4]$$

$\boldsymbol{w}$ is therefore a function of unknown parameters α_{m_1} (a scalar) and $\boldsymbol{B}_{m_1}$ (a $k_{m_1} \times V$ matrix).

Despite the fact that the observations ($\boldsymbol{w}$; word-level term identities) are a statistical function of the unknown parameters α_{m_1} and $\boldsymbol{B}_{m_1}$, we are actually interested in the latent components of the model, not the base parameters. Specifically, we would like to estimate the posterior probability distribution for the latent topic probabilities $\boldsymbol{\theta}_{m_1}$ (and thus also states $\boldsymbol{z}_{m_1}$) given the observations $\boldsymbol{w}$, but that distribution depends on the parameters fit by the model, and so our inferential problem becomes determining the posterior probability distribution of the latent components, given the observations and the estimated parameters:

$$\mathcal{P}\left(\boldsymbol{\theta}_{m_1}, \boldsymbol{z}_{m_1} \middle| \boldsymbol{w}, \alpha_{m_1}, \boldsymbol{B}_{m_1}\right) = \frac{\mathcal{P}\left(\boldsymbol{\theta}_{m_1}, \boldsymbol{z}_{m_1}, \boldsymbol{w} \middle| \alpha_{m_1}, \boldsymbol{B}_{m_1}\right)}{\mathcal{P}\left(\boldsymbol{w} \middle| \alpha_{m_1}, \boldsymbol{B}_{m_1}\right)} \quad [5]$$

where $\mathcal{P}$ is used generally as a probability distribution function. Clearly, estimating the probability distribution of the latent states necessitates calculation of α_{m_1} and $\boldsymbol{B}_{m_1}$, which facilitates expansion of the probability of observations $\boldsymbol{w}$ given the parameters α_{m_1} and $\boldsymbol{B}_{m_1}$ ($\mathcal{P}\left(\boldsymbol{w} \middle| \alpha_{m_1}, \boldsymbol{B}_{m_1}\right)$; Appendix 1):

$$\mathcal{P}\left(\boldsymbol{w} \middle| \alpha_{m_1}, \boldsymbol{B}_{m_1}\right) = \prod_{d=1}^{M} \left[\int \mathcal{P}\left(\boldsymbol{\theta}_{m_1,d} \middle| \alpha_{m_1}\right) \left(\prod_{n=1}^{N_d} \sum_{z_{m_1,d,n}} \mathcal{P}\left(w_{d,n} \middle| z_{m_1,d,n}, \boldsymbol{B}_{m_1}\right) \mathcal{P}\left(z_{m_1,d,n} \middle| \boldsymbol{\theta}_{m_1,d}\right) \right) d\boldsymbol{\theta}_{m_1,d} \right] \quad [6]$$

This equation highlights the problematic coupling of $\boldsymbol{\theta}_{m_1,d}$ (and thus α_{m_1}) and $\boldsymbol{B}_{m_1}$ in the summation over latent topics (Blei *et al.* 2003), which prevents direct, tractable estimation of parameters (and latent states).

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 $\boldsymbol{\Gamma}_{m_1}$ and $\boldsymbol{\Phi}_{m_1}$ (Appendix 2) that decouple the terms and characterize a family of distributions ($\mathcal{Q}$ to distinguish from $\mathcal{P}$) providing a lower bound on the probabilities (Jordan *et al.* 1999, Blei *et al.* 2003). Once the VEM algorithm has converged, we achieve approximate maximum likelihood estimates for the model parameters ($\alpha^*_{m_1}$ and $\boldsymbol{B}^*_{m_1}$) given the full set of observations ($\boldsymbol{w}$) for model m_1. This estimation procedure is executed using the LDA function in the topicmodels package (v0.2-7; Grün and Hornik 2011) in R (v 3.5.1; R Core Team 2018), which leverages C code written by Blei *et al.* (2003).

Latent Dirichlet Allocation: Multi-Model Inference

Given the fit of a specific Stage 1 model (m_1), we can then consider multiple Stage 1 models to determine the model with the most parsimonious number of topics $k_{m_1^*}$. Specifically, we use AIC as our Stage 1 model selection criterion (Christensen *et al.* 2018), defined for a specific LDA model m_1:

$$\text{AIC}_{m_1} = -2\ell\left(\alpha^*_{m_1}, \boldsymbol{B}^*_{m_1} \middle| \boldsymbol{w}\right) + 2k_{m_1} \quad [7]$$

where ℓ is the log likelihood ($\ell\left(\alpha^*_{m_1}, \boldsymbol{B}^*_{m_1} \middle| \boldsymbol{w}\right) = \log \mathcal{P}\left(\boldsymbol{w} \middle| \alpha^*_{m_1}, \boldsymbol{B}^*_{m_1}\right)$) and $k_{m_1} = \left(1 + k_{m_1} V\right)$ is the

number of parameters in the model: 1 for $\alpha^*_{m_1}$ and $k_{m_1}V$ corresponding to each entry in $\boldsymbol{B}^*_{m_1}$, a $k_{m_1} \times V$ matrix (Grün and Hornik 2011). If small sample size is a concern with respect to the degrees of freedom being consumed by the model, one can use the AICc correction based on the number of observations, here corpus size (M; Grün and Hornik 2011):

$$\mathrm{AICc}_{m_1} = -\mathrm{AIC}_{m_1} + \frac{2k_{m_1}{}^2 + 2k_{m_1}}{M - k_{m_1} - 1} \quad [8]$$

Because of the use of multiple iterative optimization routines (which require starting values to be drawn at random) to estimate otherwise intractable probability distributions, 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 (k_{m_1}) using different starting values, assigned through the random number generator seed ($\int$). Specifically, we use $\mathcal{N}$ replicates (n in $1 \ldots \mathcal{N}$) at each number of topics from 2 to $k_{\max}$, 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 ($\mathcal{M}_1$) is

$$\mathcal{M}_1 = \mathcal{N}(k_{\max} - 1) \quad [9]$$

The optimal (according to AIC) LDA model (m^*_1) is determined by

$$m^*_1 = \underset{m_1 \mathrm{in} 1 \ldots \mathcal{M}_1}{\arg\min}\ \mathrm{AIC}(m_1) \quad [10]$$

and has the corresponding set of parameters $\{\alpha^*_{m^*_1}, \boldsymbol{B}^*_{m^*_1}, k_{m^*_1}, \int_{m^*_1}\}$.

Reduced-Dimension Data

Having found the optimal Stage 1 model, we obtain the posterior estimates for the document-level topic probabilities (held in a $k_{m^*_1}$-length vector), which we will use as the response in our Stage 2 models. Because we are using a variational approach, we actually obtain posterior point estimates for the variational parameters $\boldsymbol{\Gamma}_{m^*_1}$ (rather than the base model parameters $\boldsymbol{\theta}_{m^*_1}$; Grün and Hornik 2011), taken from the final step of the VEM algorithm in model m^*_1. Recognizing that $\boldsymbol{\Gamma}_{m^*_1}$ contains Dirichlet concentration parameters within documents ($\boldsymbol{\gamma}^*_{m^*_1,d}$), we must normalize the values so they are proper proportions (sum to one) and can be modeled using a multinomial distribution. We notate the normalized parameters with the overbar accent as $\overline{\boldsymbol{\gamma}}^*_{m^*_1,d}$:

$$\overline{\boldsymbol{\gamma}}^*_{m^*_1,d} = \frac{\boldsymbol{\gamma}^*_{m^*_1,d}}{\sum \boldsymbol{\gamma}^*_{m^*_1,d}} \quad [11]$$

The normalized posterior point estimates of the topic proportions across all of the M documents in the corpus are held in the $M \times k_{m^*_1}$ matrix $\overline{\boldsymbol{\Gamma}}_{m^*_1}$, which corresponds to the optimal (according to AIC 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

The second stage of the LDATS model analyzes the time series of topic proportions estimated by the LDA ($\overline{\boldsymbol{\Gamma}}_{m_1^*}$) to quantify temporal dynamics and in particular to identify abrupt change point. The times of the documents ($\boldsymbol{t} = t_1 \ldots t_M$) have the potential to influence topic proportions in multiple manners: the time 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 (if continuous time impacts are included in the regression model), or may not impact $\overline{\boldsymbol{\Gamma}}_{m_1^*}$ 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**). We base our model on that indicator regression approach of Western and Kleykamp (2004), but make notable alterations, namely allowing for multiple change points, fitting segment-level regressions individually (following Ruggeri 2013), and generalizing the segment-level regression to model multinomial response data.

A Stage 2 model m_2 has a non-negative integer number of discrete change points (P_{m_2}) that divide the corpus into distinct temporal segments (s in $1 \ldots S_{m_2}$, aka "chunks") such that the number of chunks is always one more than the number of change points ($S_{m_2} = P_{m_2} + 1$; see Change Points: Segmenting the Time Series). If there are change points (*i.e.,* $P_{m_2} > 0$), then their locations (for the p in $1 \ldots P_{m_2}$ change points) are unknown parameters to be estimated (Western and Kleykamp 2004, Christensen *et al.* 2018). A specific change point p's location is represented by $\rho_{m_2,p}$ and the set of change point locations is the P_{m_2}–length vector $\boldsymbol{\rho}_{m_2,1\ldots P}$ (or $\boldsymbol{\rho}_{m_2}$). In addition to the number of change points, m_2 has a within-chunk regression model defined by a set of covariates (including the intercept indicator, temporal, and non-temporal covariates) $\boldsymbol{X}_{m_2}$ that impact the topic proportions through a set of parameters $\acute{\boldsymbol{H}}_{m_2}$ à la generalized linear models (McCullagh and Nelder 1989; see Segment-Level Models: Multinomial Logistic Regression). In combination then, we seek to estimate the posterior probability distribution of the change point locations and the regression coefficients given the fitted topic probabilities from m_1^* ($\mathcal{P}\left(\boldsymbol{\rho}_{m_2}, \acute{\boldsymbol{H}}_{m_2} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$).

We estimate $\mathcal{P}\left(\boldsymbol{\rho}_{m_2}, \acute{\boldsymbol{H}}_{m_2} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$ by leveraging the dual nature of the change points as both parameters of the model and parameters that structure the model ($\acute{\boldsymbol{H}}_{m_2}$) and calculating the marginal posterior probabilities of $\boldsymbol{\rho}_{m_2}$ and $\acute{\boldsymbol{H}}_{m_2}$ in sequence (Western and Kleykamp 2004). First, we estimate the marginal posterior probability distribution of the change point locations, given the topic proportions

$$\mathcal{P}\left(\boldsymbol{\rho}_{m_2} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right) = \int_{\acute{\boldsymbol{H}}_{m_2}} \mathcal{P}\left(\boldsymbol{\rho}_{m_2}, \acute{\boldsymbol{H}}_{m_2} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right) d\acute{\boldsymbol{H}}_{m_2} \qquad [12]$$

using Markov Chain Monte Carlo (MCMC; see Combining Segment-Level Models: Inference About Change Point Locations). Then, we define model realizations (r in $1 \ldots R_{m_2}$, as in m_2^r in $m_2^1 \ldots m_2^{R_{m_2}}$) to specify the actual change point locations ($\boldsymbol{\rho}_{m_2^r}$; see Change Points: Segmenting the Time Series). In a corpus of M documents with a range of time stamps from t_1 to t_M, the P_{m_2} change points have $\binom{(t_M-1)-t_1}{P_{m_2}}$ ("$(t_M - 1) - t_1$ choose P_{m_2}") possible values for $\boldsymbol{\rho}_{m_2}$, constituting R_{m_2} unique realizations of m_2. Each realization has a posterior probability value itself ($\mathcal{P}\left(\boldsymbol{\rho}_{m_2} = \boldsymbol{\rho}_{m_2^r} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$ and produces a conditional posterior probability distribution of the regressors given the change point locations it specifies ($\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2} \middle| \boldsymbol{\rho}_{m_2^r}, \overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$). We then combine these conditional distributions via Bayesian Model Averaging (BMA; Bartels 1997, Western and Kleykamp 2004, Hobbs and Hooten 2015) considering the realizations as submodels and using their posterior probabilities as weights to estimate the marginal posterior probability of the regression coefficients, given just the topic proportions:

$$\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2}\middle|\overline{\boldsymbol{\Gamma}}_{m_1^*}\right) = \sum_{r=1}^{R_{m_2}} \mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2}\middle|\boldsymbol{\rho}_{m_2^r}, \overline{\boldsymbol{\Gamma}}_{m_1^*}\right)\mathcal{P}\left(\boldsymbol{\rho}_{m_2} = \boldsymbol{\rho}_{m_2^r}\middle|\overline{\boldsymbol{\Gamma}}_{m_1^*}\right) \quad [13]$$

(see Combining Segment-Level Models: Inference About Within-Segment Parameters).

Change Points: Segmenting the Time Series

To deconstruct the time series into chunks, we augment the vector of change point locations $\boldsymbol{\rho}_{m_2}$ with the time step before the minimum ($\min \boldsymbol{t} - 1$) and the maximum time step ($\max \boldsymbol{t}$), generating the $\left(P_{m_2} + 2\right)$–length vector $\overbrace{\boldsymbol{\rho}}_{m_2}$ where the overbrace indicates the addition of the fixed range. In the instance that there are no change points (*i.e.,* $P_{m_2} = 0$), $\overbrace{\boldsymbol{\rho}}_{m_2}$ 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 assign the documents into segments via a mapping function ($f_{d\to s}$), which returns an indication ($\xi_{m_2^r,d,s}$) of whether or not document d belongs to chunk s (0 for no or 1 for yes) based on its timestamp t_d and the start ($\overbrace{\rho}_{m_2^r,s,1}$, the first time step after the previous change point) and end ($\overbrace{\rho}_{m_2^r,s,2}$, the timestep of the change point) times of the chunk:

$$\xi_{m_2^r,d,s} = f_{d\to s}\left(\overbrace{\boldsymbol{\rho}}_{m_2^r}, t_d, s\right) = \begin{cases} 0, & t_d < \overbrace{\rho}_{m_2^r,s,1} \\ 1, & \overbrace{\rho}_{m_2^r,s,1} \le t_d \le \overbrace{\rho}_{m_2^r,s,2} \\ 0, & t_d > \overbrace{\rho}_{m_2^r,s,2} \end{cases} \quad [14]$$

For each chunk of time, $f_{d\to s}$ produces a length-M vector of 0s and 1s ($\boldsymbol{\xi}_{m_2^r,1\ldots M,s}$ or $\boldsymbol{\xi}_{m_2^r,s}$), which are collated across chunks into an $M \times S_{m_2}$ matrix ($\boldsymbol{\Xi}_{m_2^r}$) that identifies to which chunk each document belongs (a document only belongs to one segment, such that the columns of $\boldsymbol{\Xi}_{m_2^r}$ each sum to 1). $\boldsymbol{\Xi}_{m_2^r}$ deconstructs the Stage 1 output ($\overline{\boldsymbol{\Gamma}}_{m_1^*}$) into S_{m_2} submatrices ($\overline{\boldsymbol{\Gamma}}_{m_1^*,s}$ in $\overline{\boldsymbol{\Gamma}}_{m_1^*,1} \ldots \overline{\boldsymbol{\Gamma}}_{m_1^*,S_{m_2}}$), corresponding to the chunks.

Segment-Level Models: Multinomial Logistic Regression

In LDATS, the indicator function is used to segment the data prior to analyses, such that under a given model m_2, each of the chunks of documents is fit with a separate version of the same regression (*i.e.*, the regression models all include the same predictor variables). Our approach corresponds to the assumption of no covariance among parameters across segments, which is reasonable given our conceptualization of the change points as discrete and abrupt. In this, LDATS follows Ruggieri (2013) but deviates from the approach of Western and Kleykamp (2004), who fit a single regression across the (two) chunks that applied the indicator internally, thereby allowing non-zero covariance among parameters between the segments. However, Western and Kleykamp (2004) had a singular change point, modeled a normal response variable, and used a relatively simple linear regression model with few covariates, all assumptions that we relax in LDATS (we allow for multiple change points, have a multinomial response variable, and permit complex predictive models) leading to a substantial increase in the number of parameters fit by the model, which would be computationally prohibitive to fit under a single model allowing for full covariance due to the size of the variance-covariance matrix (Genz 1992). (Although see **FUTURE DEVELOPMENTS**.)

The within-chunk component of a Stage 2 model m_2 predicts the matrix of topic proportions for the $M_{m_2^r,s}$ documents belonging to segment s ($\overline{\boldsymbol{\Gamma}}_{m_1^*,s}$) in terms of C_{m_2} predictors ($\boldsymbol{X}_{m_2^r,s}$) and $C_{m_2}k_{m_1^*}$ coefficients ($\boldsymbol{H}_{m_2^r,s}$), recognizing that the segmentation (*i.e.*, what s indexes over) depends on the identification matrix $\boldsymbol{\Xi}_{m_2^r}$, which varies among realizations of model m_2. Although the original change point

model (Western and Kleykamp 2004) assumed a univariate normal response variable, our response data are multivariate and non-normal. Specifically, our response variable is a set of $k_{m_1^*}$ multinomial probabilities, each of which must be non-negative and which must sum to 1 within a document. We address these constraints on the response variable by taking a generalized linear model approach (McCullagh and Nelder 1989) and modeling the data using a log-linear multinomial (aka multinomial logit or softmax) regression (Ripley 1996, Venables and Ripley 2002) based on a set of augmented parameters $\acute{\boldsymbol{H}}_{m_2^r,s}$ (the acute accent indicates augmentation) that define the first topic as a reference value (Appendix 3; Venables and Ripley 2002):

$$\mathrm{E}\left[\overline{\boldsymbol{\Gamma}}_{m_1^*,s}\right]_{m_2^r} = \mathrm{softmax}\left(\boldsymbol{X}_{m_2^r,s}\acute{\boldsymbol{H}}_{m_2^r,s}\right) \quad [15]$$

where E indicates the expected (predicted) value(s) of the proportion(s) and $\mathrm{softmax}$ is the normalized exponential function that generalizes the logistic function to multiple dimensions (Bishop 2006). This representation is aligned with the generalized linear model equation (McCullagh and Nelder 1989), wherein our link and inverse link functions are the multinomial logit and softmax (akin to the binomial logit and logistic functions for a logistic regression). Recognizing the uncertainty in the relationship between the "observations" ($\overline{\boldsymbol{\Gamma}}_{m_1^*,s}$) and predictions ($\mathrm{E}\left[\overline{\boldsymbol{\Gamma}}_{m_1^*,s}\right]_{m_2^r}$) allows us to estimate the probability distribution of the regression parameters for the segment given the change point locations and the segment's topic proportions ($\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\middle|\boldsymbol{\rho}_{m_2^r,s}, \overline{\boldsymbol{\Gamma}}_{m_1^*,s}\right)$) and measure the segment-level contribution to the probability of the realization of change point locations (used to calculate $\mathcal{P}\left(\boldsymbol{\rho}_{m_2} = \boldsymbol{\rho}_{m_2^r}\middle|\overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$; see <u>Combining Segment-Level Models: Inference About Within-Segment Parameters</u>). This flexible formulation of the regression model is a substantial generalization of the original LDATS application, which imposed a seasonal (within-year) dynamic modeled as a Fourier series that made biological sense for the system of focus (Christensen *et al.* 2018). However, many systems experience dynamics that are not seasonal (*e.g.*, decadal cycles) or are not stationary (*i.e.*, directional changes), and so generalizing the regression model is a key component to generalizing LDATS more broadly.

Following Bayes' rule, the posterior probability distribution of the regression parameters for a segment is proportional to the probability distribution of the topic proportions given the parameters (*i.e.*, the model likelihood) times the prior probability distribution for the parameters:

$$\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\middle|\boldsymbol{\rho}_{m_2^r,s}, \overline{\boldsymbol{\Gamma}}_{m_1^*,s}\right) \propto \mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*,s}\middle|\boldsymbol{\rho}_{m_2^r,s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\middle|\boldsymbol{\rho}_{m_2^r,s}\right) \quad [16]$$

where all of the probabilities are still conditional on the realized change point locations (*i.e.*, the submodel). The probability of a single document $\boldsymbol{d}_s$ given the parameters is the product of each topic's predicted probability raised to power of the corresponding observed probability, which is equivalent to the negative cross entropy between the observed and predicted distributions (Berger *et al.* 1996, Malouf 2002). Of particular importance here is that the "observed" values (which are actually estimated latent topic proportions) are probabilities, and so we retain the full version of the probability equation (compared to when the observed data are categorized individuals and the equation can be simplified due to the fact that all states are 0 except one that is 1 for each observation). The probability of the $\boldsymbol{M}_{m_2^r,s}$ documents in the segment is then the weighted (by $\boldsymbol{u}$) product of the document probabilities:

$$\mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*,s}\middle|\boldsymbol{\rho}_{m_2^r,s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) = \prod_{d_s=1}^{M_{m_2^r,s}} e^{u_{d_s}} \prod_{i=1}^{k_{m_1^*}} \left(\mathrm{E}\left[\bar{y}^*_{m_1^*,d_s,i}\right]_{m_2^r}\right)^{\bar{y}^*_{m_1^*,d_s,i}} \quad [17]$$

The weights are exponentiated on the probability scale and thus linear on the additive log-probability (*i.e.*, log-likelihood, loss) scale.

Generally speaking, we use a multivariate Gaussian prior for the regression parameters, following standard Bayesian approaches (Gelman and Hill 2007). The current iteration of LDATS assumes a mean-0 distribution with common precision (inverse variance) across segments λ and no covariance (but see **FUTURE DEVELOPMENTS**). Recalling that the model parameters $\acute{\boldsymbol{H}}_{m_2^r,s}$ are defined with the first topic as a reference (intercept), however, there are some entries in $\acute{\boldsymbol{H}}_{m_2^r,s}$ that are fixed at 0 and not free for estimation. Specifically, for any covariate c, the parameter associated with the first topic $i = 1$ is 0: $\acute{\eta}_{m_2,s,i=1,c} = 0$. Thus, we define the variance-covariance matrix using the augmented identity matrix $\acute{\boldsymbol{I}}$ where the accent matches $\acute{\boldsymbol{H}}$ and signifies that all of the entries associated with the first topic are set to 0. Including the set-to-0 regressors, there are $C_{m_2}k_{m_1^*}$ entries in $\acute{\boldsymbol{H}}_{m_2^r,s}$, and thus the multivariate normal distribution is of dimension $C_{m_2}k_{m_1^*}$:

$$\acute{\boldsymbol{H}}_{m_2^r,s} \sim \mathrm{MVN}_{C_{m_2}k_{m_1^*}}\left(0, \lambda\acute{\boldsymbol{I}}\right) \quad [18]$$

where MVN is defined by the mean vector (0) and precision matrix ($\lambda\acute{\boldsymbol{I}}$). The original formulation of LDATS (Christensen *et al.* 2018) assumed $\lambda = 0$ (*i.e.*, a fully vague prior) but including a small increase in precision ($\lambda \approx 10^{-4} - 10^{-2}$) can aid in finding the optimal solution as long as all coefficients have been scaled to about $[0,1]$ (*e.g.*, normalized; Ripley 1993, Ripley 1996). This approach to estimating $\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\middle|\boldsymbol{\rho}_{m_2^r,s}, \overline{\boldsymbol{\Gamma}}_{m_1^*,s}\right)$ is equivalent to the method called ridge regression (Hoerl 1962, Hoerl and Kennard 1970), Tikhonov regularization (Tikhonov and Arsenin 1977), L2 regularization (Ng 2004) or joint maximum *a posteriori* (MAP; Bassett and Deride 2018) estimation of a neural network using a weight decay (Venables and Ripley 2002), depending on the applied setting.

Computationally, the regression for a chunk of time based on a 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 posterior distribution for the parameters is found with the gradient-based optimization routine 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 models. Here, we use the negative log of the prior-penalized likelihood as the loss value ($\mathcal{L}$):

$$\mathcal{L}_{m_2^r,s} = -\log\left(\mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*,s}\middle|\boldsymbol{\rho}_{m_2^r,s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\middle|\boldsymbol{\rho}_{m_2^r,s}\right)\right) \quad [19]$$

that is minimized to find the optimal parameter value set $\acute{\boldsymbol{H}}^*_{m_2^r,s}$

$$\acute{\boldsymbol{H}}^*_{m_2^r,s} = \underset{\acute{\boldsymbol{H}}_{m_2^r,s}}{\arg\min}\, \mathcal{L}_{m_2^r,s} \quad [20]$$

which has the loss value $\mathcal{L}^*_{m_2^r,s}$ and corresponds to the mode of the posterior (Venables and Ripley 2002). The BFGS Algorithm works efficiently by not calculating the Hessian (matrix of partial second derivatives) of the prior-penalized loss function at every step in the optimization, but rather approximating it by comparing successive iterations of the Jacobian matrix (matrix of partial first derivatives), whose components correspond to the partial derivatives the loss with respect to the coefficients associated with each combination of covariate and topic. 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 (Appendix 4).

Combining Segment-Level Models: Inference About Change Point Locations

We then define the full time series model for the realization by collating the chunk-level models. Specifically, we row-wise stack the chunk-specific parameter matrices $\acute{\boldsymbol{H}}_{m_2^r,1}$ to $\acute{\boldsymbol{H}}_{m_2^r,S_{m_2}}$ creating $\acute{\boldsymbol{H}}_{m_2^r}$ (dimension: $\left(S_{m_2}C_{m_2}\right) \times k_{m_1^*}$) and we place the chunk-specific covariate matrices $\boldsymbol{X}_{m_2^r,1}$ to $\boldsymbol{X}_{m_2^r,S_{m_2}}$ into a diagonal square ($S_{m_2} \times S_{m_2}$) matrix to produce a covariate matrix for the full time series $\boldsymbol{X}_{m_2^r}$ (dimension: $M \times \left(S_{m_2}C_{m_2}\right)$). Recalling that the splitting of the corpus into segments and thus the fitting of chunk-level parameters is governed by the realized change point locations $\boldsymbol{\rho}_{m_2^r}$ via the indication matrix $\boldsymbol{\Xi}_{m_2^r}$, we can then define the generalized linear equation for the full time series as

$$\mathrm{E}\left[\overline{\boldsymbol{\Gamma}}_{m_1^*}\right]_{m_2^r} = f_{f(s)}\left(\mathrm{softmax}, \boldsymbol{X}_{m_2^r}\acute{\boldsymbol{H}}_{m_2^r}, \boldsymbol{\Xi}_{m_2^r}\right) \quad [21]$$

where the segment function mapper ($f_{f(s)}$) applies the $\mathrm{softmax}$ function to the segments of $\boldsymbol{X}_{m_2^r}\acute{\boldsymbol{H}}_{m_2^r}$ as defined by $\boldsymbol{\Xi}_{m_2^r}$ (Appendix 5). The use of the mapping function is needed here, even for a simple prediction, as it defines the values included in the $\mathrm{softmax}$ function, whose output depends on the full set of inputs (Appendix 3). Further, because the $\mathrm{softmax}$ is log-linear, entries in $\boldsymbol{X}_{m_2^r}\acute{\boldsymbol{H}}_{m_2^r}$ that are 0 are actually meaningful (they are equal to the reference topic) when included in the regression, and so we must only include true 0s, and not the filler-entry 0s, from $\boldsymbol{X}_{m_2^r}\acute{\boldsymbol{H}}_{m_2^r}$ in the calculation. Equation 21 can be used to calculate the optimal model-estimated topic proportions for all documents in the corpus ($\mathrm{E}\left[\overline{\boldsymbol{\Gamma}}_{m_1^*}\right]^*_{m_2^r}$) by simply substituting the MAP estimated parameter matrix into the segment function mapper, as in $f_{f(s)}\left(\boldsymbol{X}_{m_2^r}\acute{\boldsymbol{H}}^*_{m_2^r}, \mathrm{softmax}, \boldsymbol{\Xi}_{m_2^r}\right)$.

Inferentially, we are interested in determining the (marginal posterior) probability distribution of the change point locations, given topic proportions $\mathcal{P}\left(\boldsymbol{\rho}_{m_2}|\overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$, which we relate to the probability of the topic proportions, given the change point locations (*i.e.*, the model likelihood, $\mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*}|\boldsymbol{\rho}_{m_2}\right)$) and the prior distribution of the change points ($\mathcal{P}\left(\boldsymbol{\rho}_{m_2}\right)$) via Bayes' theory:

$$\mathcal{P}\left(\boldsymbol{\rho}_{m_2}|\overline{\boldsymbol{\Gamma}}_{m_1^*}\right) \propto \mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*}|\boldsymbol{\rho}_{m_2}\right)\mathcal{P}\left(\boldsymbol{\rho}_{m_2}\right) \quad [22]$$

We use the segment-level models' loss values to construct the likelihood of the topic proportions under the change point locations. Specifically, we sum the minimized loss (negative log prior-penalized likelihood) values across the chunks to calculate the total minimized loss for a realization of the model ($\mathcal{L}^*_{m_2^r}$), which is equal to the negative log likelihood for the topic proportions given the change point locations

$$\mathcal{L}^*_{m_2^r} = \sum_{s=1}^{S_{m_2}} \mathcal{L}^*_{m_2^r,s} = -\ell\left(\overline{\boldsymbol{\Gamma}}_{m_1^*}|\boldsymbol{\rho}_{m_2^r}\right) \quad [23]$$

Remembering that a realization of a model simply specifies the change point locations, we can define the distribution of minimized loss $\mathcal{L}^*_{m_2}$ for the model as a function of change point locations based on the $R_{m_2} = \binom{(t_M-1)-t_1}{P_{m_2}}$ unique realizations of, each of which has its own $\mathcal{L}^*_{m_2^r}$ (Western and Kleykamp 2004):

$$\mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*}|\boldsymbol{\rho}_{m_2} = \boldsymbol{\rho}_{m_2^r}\right) = \frac{e^{\mathcal{L}^*_{m_2^r}}}{\sum_{r=1}^{R_{m_2}} e^{\mathcal{L}^*_{m_2^r}}} \quad [24]$$

Although because of the sheer number of possible realizations for even a modest time series with multiple change points (Table 2), we use Markov Chain Monte Carlo (MCMC) methods to sample the probability distributions efficiently (rather than sample it systematically).

The prior probability distribution for the change point locations ($\mathcal{P}(\boldsymbol{\rho}_{m_2})$) is a multivariate discrete distribution, which could take any number of specific formulations (Western and Kleykamp 2004, Ruggieri 2013). 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 (see **FUTURE DEVELOPMENTS**). The uniform prior allocates equal probability to each of the discrete time points from the time of the first document ($\min \boldsymbol{t}$) to one time step before the last document ($\max \boldsymbol{t} - 1$), and the selected times are then sorted chronologically

Because the probability distribution has a high potential for multiple modes (if multiple change points are reasonably likely), standard MCMC approaches may have difficulty fitting the model (Sambridge 2014). Thus, we employ parallel tempering MCMC (ptMCMC; also called Metropolis-coupled or replica-exchange MCMC), which endows a standard MCMC search with auxiliary chains that explore the distribution surface more rapidly than the focal chain (Swendsen and Wang 1986, Geyer 1991, Earl and Deem 2005). ptMCMC is a robust methodology that works well for generalized models (Guo *et al.* 2016) and is an efficient sampler of rough probability landscapes (Machata and Ellis 2011). Here, the posterior probability distribution's surface is explored using H chains (h in $1 \ldots H$), each with its own temperature (a_h) defining its search ability: higher temperature chains have higher variances in their step sizes (they have flattened surfaces to search; Gupta *et al.* 2018) and are therefore more easily able to navigate the surface. This comes at a cost of instability, however, as higher temperature chains are less likely to settle in to stable distributions. We therefore use a range of temperatures grouped in a series ($\boldsymbol{a} = a_1 < a_2 < \cdots < a_H$) to balance search breadth and depth, with the coolest (a_1) sampling the true surface (Earl and Deem 2005).

The specifics of the temperature regime are obviously then critical for fitting a model to data using ptMCMC (Kone and Kofke 2005, Rathore *et al.* 2005, Nagata and Watanabe 2008). Following the original LDATS model (Christensen *et* al. 2018), the current implementation allows for control over the temperature sequence control parameters to facilitate fitting a wide range of potential corpus time series (Simonis *et al. In Development*) and defines the temperatures as

$$\begin{gathered} \boldsymbol{a}_{1 \ldots H-1} = 2^{\frac{(\mathrm{seq}(0, \log_2 a_{H-1}, H-1))^{1+q}}{(\log_2 a_{H-1})^q}} \\ a_H = a_H \end{gathered} \qquad [25]$$

where $\mathrm{seq}(0, \log_2 a_{H-1}, H-1)$ is a sequence of values from 0 to $\log_2 a_{H-1}$ of length $H-1$, q is the exponent controlling the temperature series ($q = 0$ produces a geometric sequence, $q = 1$ implies squaring before exponentiating), a_{H-1} is the penultimate temperature, and a_H is the ultimate temperature. Currently, the control inputs (a_H, a_{H-1}, H, and q) are available to the user, but are fixed for a given fit of the $\mathcal{M}_2$ Stage 2 models. A target for future mathematical and coding development is to 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**).

The ptMCMC algorithm works by coupling the chains (which are taking their own walks on the distribution surface) through "swaps", where neighboring chains exchange configurations in between steps (Geyer 1991, Falcioni and Deem 1999). Over the course of the search, each of the chains proceeds through G iterations (steps; g in $1 \ldots G$) of the Metropolis-Hastings (MH) algorithm (Metropolis *et al.* 1953, Hastings 1970), a classical MCMC method. Step g on chain h is some realization of the model, such that the values of the change points for $r = h, g$ are $\boldsymbol{\rho}_{m_2^{h,g}}$. The chains are each initialized ($g = 0$) with a draw from $\mathcal{P}(\boldsymbol{\rho}_{m_2})$

and the best fit draw (determined by likelihood) is put in the focal chain with the next best fit draw is put in the next hottest chain until the worst fit draw is put in the hottest chain. Then, from each step until $g = G$, for each chain, a new set of change points is proposed ($\boldsymbol{\rho}_{m_2^{h,\breve{g}}}$; the breve accent indicates a proposal (pre)-step), evaluated (loss is calculated), and then either accepted ($\boldsymbol{\rho}_{m_2^{h,g}} = \boldsymbol{\rho}_{m_2^{h,\breve{g}}}$) or rejected ($\boldsymbol{\rho}_{m_2^{h,g}} = \boldsymbol{\rho}_{m_2^{h,g-1}}$). The proposed set of change point locations for step g is generated from the proposal distribution $\mathcal{R}$ conditional on the previous change point locations ($\mathcal{R}\left(\boldsymbol{\rho}_{m_2^{h,\breve{g}}} \middle| \boldsymbol{\rho}_{m_2^{h,g-1}}\right)$). The proposal distribution describes the movement of a single change point via a symmetric geometric distribution with a user-controlled average step size. Functionally, $\mathcal{R}$ is a joint distribution representing three steps: [1] selecting one of the P_{m_2} change points to move via a multinomial distribution with equal probabilities ($\frac{1}{P_{m_2}}$), [2] determining the directionality of movement (earlier or later in the time series) using a binomial distribution with equal probability for the outcomes -1 and 1, and [3] calculating the magnitude of the movement (number of discrete time steps) with a geometric distribution (mean step size is κ). The multiplication of the three distributions results in the P_{m_2}-length vector representing the proposal step from the current change point locations:

$$\boldsymbol{\rho}_{m_2^{h,\breve{g}}} \sim \boldsymbol{\rho}_{m_2^{h,g-1}} + \mathrm{Geom}\left(1, \frac{1}{\kappa}\right) \mathrm{Binom}_{\{-1,1\}}(1, 0.5) \mathrm{Multinom}_{P_{m_2}}\left(1, \frac{1}{P_{m_2}}\right) \quad [26]$$

where Geom is the version of the geometric distribution that has 1 as its minimum returned value. The only parameter that is available to the user to set is the average step size of the geometric distribution, κ. Ostensibly, this parameter could be adaptively set based on the data set, but that functionality is not presently included (although see **FUTURE DEVELOPMENTS**).

Because the proposal distribution is symmetric ($\mathcal{R}\left(\boldsymbol{\rho}_{m_2^{h,\breve{g}}} \middle| \boldsymbol{\rho}_{m_2^{h,g-1}}\right) = \mathcal{R}\left(\boldsymbol{\rho}_{m_2^{h,g-1}} \middle| \boldsymbol{\rho}_{m_2^{h,\breve{g}}}\right)$), we are able to use the simplified Metropolis acceptance rule for the proposal (Metropolis *et al.* 1953, Hastings 1960). Acceptance of the proposal is probabilistic and based on the difference in energy ($\Delta\mathcal{E}$) between the current ($m_2^{h,g}$) and proposed ($m_2^{h,\breve{g}}$) realizations of the model:

$$\Delta\mathcal{E}_{m_2^{h,\breve{g}}} = \mathcal{E}_{m_2^{h,\breve{g}}} - \mathcal{E}_{m_2^{h,g}} \quad [27]$$

where the energy of m_2^r is the log of the inverse posterior evaluated at $\boldsymbol{\rho}_{m_2} = \boldsymbol{\rho}_{m_2^r}$,

$$\mathcal{E}_{m_2^r} = \log \frac{1}{\mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*} \middle| \boldsymbol{\rho}_{m_2} = \boldsymbol{\rho}_{m_2^r}\right) \mathcal{P}\left(\boldsymbol{\rho}_{m_2^r}\right)} \quad [28]$$

which is equal to the difference between the total minimized loss ($\mathcal{L}^*_{m_2^r}$; negative log likelihood), and the log of the prior:

$$\mathcal{E}_{m_2^r} = \mathcal{L}^*_{m_2^r} - \log \mathcal{P}\left(\boldsymbol{\rho}_{m_2^r}\right) \quad [29]$$

(Metropolis *et al.* 1953, Gupta *et al.* 2018). The change in energy associated with a proposal is converted to an acceptance probability for the step:

$$u_{m_2^{h,\breve{g}}} = \min\left(1, e^{-b_h \Delta\mathcal{E}_{m_2^{h,\breve{g}}}}\right) \quad [30]$$

where b_h is the inverse temperature of chain h (*i.e.*, $b_h = \frac{1}{a_h}$), such that higher temperature chains have higher acceptance probabilities for the same energy difference. A random number from the standard uniform distribution is then drawn ($\mathrm{U}(0,1)$), with the proposal being accepted if the random number is less than or equal to $\mathcal{u}_{m_2^{h,g}}$, and rejected if the random number is larger than $\mathcal{u}_{m_2^{h,g}}$.

After each Metropolis MCMC iteration, the chains are able to swap configurations with their nearest neighbors in the temperature series (following the Metropolis criterion; Metropolis *et al.* 1953), allowing them to share information and search the surface in combination (Earl and Deem 2005). Starting with the hottest pair of chains (H and $H-1$) and descending in temperature, the chains (generally, h and $h-1$) swap information, and similar to the within-chain steps, the temperatures scale the swap acceptance rates for the sharing of information between chains, such that hotter chains are more likely to accept swaps (Earl and Deem 2005, Gupta *et al.* 2018). The acceptance probability for the swap between the neighboring chains is the exponentiated product of the difference in the chains' inverse temperatures and energies:

$$\mathcal{u}_{m_2^{h:h-1,g}} = \min\left(1, e^{\Delta b_{m_2^{h:h-1,g}} \Delta \mathcal{E}_{m_2^{h:h-1,g}}}\right) \quad [31]$$

where

$$\Delta b_{m_2^{h:h-1,g}} = \frac{1}{a_h} - \frac{1}{a_{h-1}} \quad [32]$$

and

$$\Delta \mathcal{E}_{m_2^{h:h-1,g}} = \mathcal{E}_{m_2^{h,g}} - \mathcal{E}_{m_2^{h-1,g}} \quad [33]$$

Currently, the schedule of Metropolis iterations and chain swaps and the inclusion of the chains in each swap are set (as "swaps between each iteration" and "all neighbor pairs included in each swap"), although they could become adaptive or user-controlled to allow for more efficient sampling (this is the most computationally intensive step in the LDATS modeling process; see **FUTURE DEVELOPMENTS**).

We remove the first steps $G_{\text{burn-in}}$ (set by the user) steps as burn-in and thin the resulting sample to a fraction of τ ($\tau = 1$ equates to no thinning and $\tau \in (0,1]$) steps (Link and Eaton 2012, Hobbs and Hooten 2015), with the resulting change point location vectors we store as rows in the matrix $\boldsymbol{P}_{m_2^{h=1}}$ ($\boldsymbol{P}$ as in capital $\boldsymbol{\rho}$), which is dimension $\mathrm{floor}\left(\tau(G - G_{\text{burn-in}} + 1)\right) \times P_{m_2}$, where floor is the round-down function. $\mathcal{P}\left(\boldsymbol{\rho}_{m_2} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$ is then defined by the proportional representation of each of the R_{m_2} realizations of $\boldsymbol{\rho}_{m_2^r}$ in the rows of $\boldsymbol{P}_{m_2^{h=1}}$. Throughout the ptMCMC algorithm, we track step and swap acceptances and count the trips made by particles (the bits of information that can move among replica chains during swaps) from the hottest to the coolest chain (Katzgraber *et al.* 2006), and calculate a final rate for each based on the full set of G iterations (Earl and Deem 2005). These rates, along with trace plots (Kruschke 2015), constitute the ptMCMC diagnostics presently provided in LDATS.

Combining Segment-Level Models: Inference About Within-Segment Parameters

Having defined $\acute{\boldsymbol{H}}_{m_2^r}$ as the collated matrix of segment-level parameter matrices ($\acute{\boldsymbol{H}}_{m_2^r,s}$) fit based on the realized change point locations and having estimated the probability of that specific set of change point locations using ptMCMC, we can now calculate the marginal posterior distribution of the regression parameters given the topic proportions ($\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$), acknowledging the uncertainty in the change point

locations (Western and Kleykamp 2004). In order to accomplish this, however, we must account for the variance-covariance structure among the parameters within $\acute{\boldsymbol{H}}_{m_2^r}$, which is facilitated by first unwinding the matrix into a vector $\widetilde{\acute{\boldsymbol{\eta}}_{m_2^r}}$, as indicated by the tilde accent, which constitutes the mean vector for the parameter distributions ($\widetilde{\acute{\boldsymbol{\eta}}_{m_2^r,s}}$ is the unwound segment-specific matrix $\acute{\boldsymbol{H}}_{m_2^r,s}$). Then, we place the precision (inverse variance-covariance) matrices for the segment-level model fits (generally, $\boldsymbol{\Lambda}_{\widetilde{\acute{\eta}_{m_2^r,s}}} = \boldsymbol{\Sigma}^{-1}_{\widetilde{\acute{\eta}_{m_2^r,s}}}$) along the diagonal of a square ($S_{m_2} \times S_{m_2}$) matrix, producing the block-diagonal precision matrix for the full time series model given the realized change points, $\boldsymbol{\Lambda}_{\widetilde{\acute{\eta}_{m_2^r}}}$, whose off-diagonal entries are ∞ (*i.e.*, no covariance). The probability distribution for the full set of regression parameters, given a realized set of change point locations and the topic proportions is then described by a multivariate normal distribution (MVN; Tiao and Zellner 1964) with mean vector $\widetilde{\acute{\boldsymbol{\eta}}_{m_2^r}}$ and precision matrix $\boldsymbol{\Lambda}_{\widetilde{\acute{\eta}_{m_2^r}}}$:

$$\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2} \middle| \boldsymbol{\rho}_{m_2^r}, \overline{\boldsymbol{\Gamma}}_{m_1^*}\right) = \mathcal{P}_{\text{MVN}_{S_{m_2} C_{m_2} k_{m_1^*}}}\left(\widetilde{\acute{\boldsymbol{\eta}}_{m_2^r}}, \boldsymbol{\Lambda}_{\widetilde{\acute{\eta}_{m_2^r}}}\right) \quad [34]$$

where $\mathcal{P}_{\text{MVN}}$ is the probability density function of an MVN defined for a set of specific parameter values $\boldsymbol{x}$

$$\mathcal{P}_{\text{MVN}_{S_{m_2} C_{m_2} k_{m_1^*}}}\left(\widetilde{\acute{\boldsymbol{\eta}}_{m_2^r}}, \boldsymbol{\Lambda}_{\widetilde{\acute{\eta}_{m_2^r}}}\right) = \frac{1}{\sqrt{\left|2\pi \boldsymbol{\Lambda}^{-1}_{\widetilde{\acute{\eta}_{m_2^r,s}}}\right|}} e^{-\frac{1}{2}\left(\boldsymbol{x} - \widetilde{\acute{\boldsymbol{\eta}}_{m_2^r}}\right)' \boldsymbol{\Lambda}_{\widetilde{\acute{\eta}_{m_2^r}}} \left(\boldsymbol{x} - \widetilde{\acute{\boldsymbol{\eta}}_{m_2^r}}\right)} \quad [35]$$

and $S_{m_2} C_{m_2} k_{m_1^*}$ is the dimensionality of the specific MVN distribution here (one parameter per covariate per topic per segment). The value of $\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2} \middle| \boldsymbol{\rho}_{m_2^r}, \overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$ is determined for each of the R_{m_2} realizations, which are then combined via Eq. 13 to estimate the marginal posterior probability of the regression coefficients, given just the topic proportions ($\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right)$). The original LDATS application did not calculate or return the posterior estimates of the regression parameters as the research was focused on the change point locations (Christensen *et al.* 2018), and so this constitutes a key update to the application.

Multi-Model Inference

Given the fit of a specific Stage 2 model (m_2) defined by the number of change points (P_{m_2}) and the within-segment covariate model (the specific C_{m_2} predictors comprising $\boldsymbol{X}_{m_2}$), we can then consider multiple Stage 2 models to determine the optimal configuration of the time series component.

As with Stage 1, we use AIC as our Stage 2 model selection criterion (Christensen *et al.* 2018, based on Gelman *et al.* 2014), defined for a specific time series model m_2:

$$\text{AIC}_{m_2} = -2\ell\left(\boldsymbol{\rho}_{m_2}, \acute{\boldsymbol{H}}_{m_2} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right) + 2k_{m_2} \quad [36]$$

where ℓ is the log likelihood ($\ell\left(\boldsymbol{\rho}_{m_2}, \acute{\boldsymbol{H}}_{m_2} \middle| \overline{\boldsymbol{\Gamma}}_{m_1^*}\right) = \log \mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*} \middle| \boldsymbol{\rho}_{m_2}, \acute{\boldsymbol{H}}_{m_2}\right)$) estimated from the ptMCMC samples retained after burn-in and thinning, and $k_{m_2} = \left(S_{m_2} C_{m_2}\left(k_{m_1} - 1\right) + P_{m_2}\right)$ is the number of parameters in the model: for each segment (in S_{m_2} total), there are $C_{m_2}\left(k_{m_1} - 1\right)$ parameters fit for the multinomial regression (one for every covariate for every topic, save the first to account for the sum-to-1 constraint, Appendix 3), plus the P_{m_2} change point locations (Western and Kleykamp 2004, Christensen *et al.* 2018). The total number of models in Stage 2 ($\mathcal{M}_2$) is

$$\mathcal{M}_2 = (P_{\max} + 1)\mathcal{U}_{\boldsymbol{X}} \quad [37]$$

where $P_{\max}$ is the maximum number of change points included (and the $+1$ to account for $P_{m_2} = 0$) and $\mathcal{U}_{\boldsymbol{X}}$ is the number of unique configurations of the covariate matrix ($\boldsymbol{X}$). The optimal (according to AIC) time series model (m_2^*) is determined by

$$m_2^* = \underset{m_2 \text{in} 1 \ldots \mathcal{M}_2}{\arg\min}\ \text{AIC}(m_2) \quad [38]$$

and has the corresponding set of parameters $\{\boldsymbol{\rho}_{m_2^*}, \acute{\boldsymbol{H}}_{m_2^*}\}$.

MODEL EVALUATION

{This section will focus on evaluating the model under known parameters}

MODEL APPLICATION

{This section will include one or two brief example applications of LDATS to real data sets}

DISCUSSION

{This section will do the normal discussing of the present work in the context of the literature}

FUTURE DEVELOPMENTS

The LDATS modeling framework (as presented here) and code package (Simonis *et al. In Development*) are stable and robust, but as with any methodology, improvements can be made through future developments. We have noted multiple components slated for developments throughout this manuscript; we add some additional components and briefly describe our current plans in this section.

Although the VEM approach to LDA model fitting is well developed in its application (Blei *et al.* 2003, Grün and Hornik 2011), there is one notable area of improvement that could be included: adaptive selection of the optimal number of topics $k_{m_1^*}$. The other common method to parameter estimation under LDA models, Gibbs sampling, provides a flexible approach to estimating the number of topics by considering it a free parameter to be fit (Griffiths and Steyvers 2004, Valle *et al.* 2018). However, there is also a method for adaptive selection of the number of topics under a VEM approach to LDA based on topic densities (Cao *et al.* 2009), which could be integrated into LDATS without needing to alter the underlying inference machinery.

The time series model options (based on Western and Kleykamp 2004) in LDATS do not presently include temporal autocorrelation, which is a feature of many time series (Cressie and Wikle 2011), and could be mistaken for (statistically speaking) change points, especially if the magnitude of the change is small (Jarušková 1997, Wang 2008, Beaulieu *et al.* 2012). Although there are methods for including temporal autocorrelation into change point models (Lund *et al.* 2007, Jandhyala *et al.* 2010, Beaulieu *et al.* 2012, Pandya *et al.* 2012), they all operate under the assumption of a univariate normal response variable, which is a much simpler segment-level regression model than the present softmax and offers the simple entrée of the error term for autocorrelation. Methods exist for including autocorrelation structure in multinomial regressions (Linderman *et al.* 2015), but have yet to be integrated with a change point model.

By splitting the corpus prior to fitting the segment-level models, LDATS makes the important assumption of no covariance among regressors across segments, although it is possible that there are substantially-non-0 covariances. It could therefore improve the predictive capacity of the model to fit the model across-segments in a fashion to allow covariances. Such a change to LDATS would require a shift in the underlying inferential machinery and is likely to considerably slow computation (Genz 1992), however.

LDATS also assumes mean-0, uncorrelated, vague priors for the regression parameters and a vague prior for the change point locations. Although the vagueness of the regression prior can be toggled by the user, no other control on the priors is available, and therefore the full strength of the Bayesian approach is not yet being leveraged in the time series component of LDATS (Chin Choy and Broemeling 1980, Western and Kleykamp 2004). Including more nuanced control of the priors necessitates substantial changes to the inferential code underlying the model, and in particular will require a shift in the fitting of the multinomial models away from an established code base (Venables and Ripley 2003).

Although the parameters controlling the ptMCMC algorithm (a_H, a_{H-1}, H, q, and κ) are available to the user to input, they are limited to fixed values, and so require the user to *a priori* know appropriate values or spend considerable time testing them. In addition, the swap schedule (in terms of frequency of swaps and the inclusion of chains within a swap) used for the ptMCMC is presently fixed and not available to control at all by the user. All of these controls have the potential to influence the convergence of the ptMCMC algorithm and the speed at which convergence is achieved. Therefore, the model's application could be improved by allowing flexibility in and control over the swap schedule and by taking an adaptive approach to ptMCMC that allows the controls to optimize themselves via the fitting process (Katzgraber *et al.* 2006, Trebst *et al.* 2006, Hasenbusch and Schaefer 2010, Miasojedow *et al.* 2013).

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.

Bartels, L. M. 1997. Specification uncertainty and model averaging. *American Journal of Political Science* **41**:641-674.

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

Beauleiu, C., J. Chen, and J. L. Sarmiento. 2012. Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A* **370**:1228-1249.

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.

Bishop, C. M. 2006. Pattern Recognition and Machine Learning. Springer.

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.

Cao, J., T. Xia, J. Li, Y. Zhang, and S. Tang. 2009. A density-based method for adaptive LDA model selection. *Neurocomputing* **72**:1775-1781.

Chin Choy, J. H. and L. D. Broemeling. 1980. Some Bayesian inferences for a changing linear model. *Technometrics* **22**:71–78.

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.

Cressie, N. and C. K. Wikle. 2011. Statistics for Spatio-Temporal Data. Wiley, Hoboken, NJ, USA.

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

Falcioni, M. and M. W. Deem. 1999. A biased Monte Carlo scheme for zeolite structure solution. *Journal of Chemical Physics* **110**:1754–1766.

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.

Gelman, A. and J. Gill. 2007. Data Analyis aUsing Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge, UK.

Gelman, A., J. Hwang, and A. Vehtari. 2014. Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24**:997–1016.

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.

Griffiths, T. L. and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101:5228-5235.

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

Guo, G., W. Shao, L. Lin, and X. Zhu. 2016. Parallel tempering for dynamic generalized linear models. *Communications in Statistics—Theory and Methods* **45**:6299-6310

Gupta, S., L. Hainsworth, J. S. Hogg, R. E. C. Lee, and J. R. Faeder. 2018. Evaluation of parallel tempering to accelerate Bayesian parameter estimation in systems biology. arXiv:1801.09831 [q-bio.QM]

Hasenbusch, M. and S. Schaefer. 2010. Speeding up parallel tempering simulations. *Physical Review* **82**:046707.

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

Hobbs, N. T. and M. B. Hooten. 2015. Bayesian Models: A Statistical Primer for Ecologists. Princeton University Press, Princeton, NJ, USA.

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.

Jandhyala, V. K., S. B. Fotopoulos, and J. You. 2010. Change-point analysis of mean annual rainfall data from Tucumán, Argentina. *Environmetrics* **21**:687-697.

Jarušková, D. 1997. Some problems with application of change-point detection methods to environmental data. *Environmetrics* **8**:469–483.

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.

Katzgraber, H. G., S. Trebst, D. A. Huse. And M. Troyer. 2006. Feedback-optimized parallel tempering Monte Carlo. *Journal of Statistical Mechanics: Theory and Experiment* **3**:P03018.

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

Kruschke, J. K. 2015. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition. Elsevier Academic Press, New York, NY, USA.

Linderman, S., M. Johnson, and R. P. Adams. 2015. Dependent multinomial models made easy: stick-breaking with the Pólya-gamma augmentation. *Advances in Neural Information Processing Systems* **28**: 3456-3464.

Link, W. A. and M. J. Eaton. 2012. On thinning of chains in MCMC. *Methods in Ecology and Evolution* **3**:112-115.

Lund, R., X. L. Wang, Q. Lu, J. Reeves, C. Gallagher, and Y. Feng. 2007. Changepoint detection in periodic and autocorrelated time series. *Journal of Climate* **20**:5178-5190.

Machata, J. and R. S. Ellis. 2011. Monte Carlo methods for rough free energy landscapes: population annealing and parallel tempering. *Journal of Statistical Physics* **144**:541-553.

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

Miasojedow, B., E. Moulines, and M. Vihola. 2013. An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics* **22**:649-664.

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

Nagata, K. and S. Watanabe. 2008. Asymptotic behavior of exchange ratio in exchange Monte Carlo method. *Neural Networks* **21**:980-988.

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, Boston, MA, USA.

Ng, A. 2004. Feature selection, L1 vs. L2 regularization, and rotation invariance. *Proceedings of the Twenty-first International Conference on Machine Learning*: 78.

Pandya, M., K. Bhatt, and H. C. Thakar. 2012. Bayesian estimation of change point in autoregressive process. *International Journal of Research and Reviews in Applied Sciences* **13**:41-52.

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.

Rathore, 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, UK.

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

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

Ruggieri, E. 2013. A Bayesian approach to detecting change points in climactic records. *International Journal of Climatology* **33**:520-528.

Sambridge, M. 2014. A Parallel Tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International* **196**:357-374.

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

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.

Tiao, G. C. and A. Zellner. 1964. On the Bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society: Series B (Methodological)* **26**:277-285 .

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

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.

Trebst, S., M. Troyer, and U. H. E. Hansmann. 2006. Optimized parallel tempering simulations of proteins. *The Journal of Chemical Physics* **124**:174903.

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.

Valle, D., P. Albuquerque, Q. Zhao, A. Barberan, and R. J. Fletcher Jr. 2018. Extending the Latent Dirichlet Allocation model to presence/absence data: a case study on North American breeding birds and biogeographic shifts expect from climate change. *Global Change Biology* **24**:5560-5572.

Venables, W. N. and B. D. Ripley. 2002. Modern and Applied Statistics with S. Fourth Edition. Springer, New York, NY, USA.

Wang, X. L. L. 2008. Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal t or F test. *Journal of Applied Meteorology and Climatology* **47**:2423–2444.

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.

TABLES

Table 1. Definitions of the notation used in LDATS.

Definition	Parameter	Expansions
Corpus (set of documents)	D	
Total documents	M	
Specific document	d	
Total words	N	
Total words in a document	N_d	
Specific word	n	
Total terms	V	
Specific term	v	
Weight of a document	u_d	$\boldsymbol{u}$
Time of a document	t_d	$\boldsymbol{t}$
Probability	$\mathcal{P}$	
Loss	$\mathcal{L}$	
Log-likelihood	ℓ	
Expected value	E	
Number of fitted parameters	k	
Optimal value of a fit (parameter or model)	$*$	
Replicates of a model with different starting values	$\mathcal{N}$	
Derivative	d	$\mathcal{D}$
General index variables	i, j	
Total Stage 1 models	$\mathcal{M}_1$	
Specific Stage 1 model	m_1	
Total latent topics in a model	k_{m_1}	
Specific latent topic	i	
Observed word identity	w_{d,n_d}	$\boldsymbol{w}_d, \boldsymbol{w}$
Latent topic membership	z_{m_1,d,n_d}	$\boldsymbol{z}_{m_1,d}, \boldsymbol{z}_{m_1}$
Within-document topic probability	$\theta_{m_1,d,i}$	$\boldsymbol{\theta}_{m_1,d}, \boldsymbol{\theta}_{m_1}$
Document-level topic concentration	$\alpha_{m_1,i}$	$\alpha_{m_1}, \boldsymbol{\alpha}_{m_1,d}$
Word-level term probability	$\beta_{m_1,d,i,v}$	$\boldsymbol{B}_{m_1,d}, \boldsymbol{B}_{m_1}$
Random number generator seed	$\int$	
Variational distribution	Q	
Variational lower bound	L	
Kullback-Leibler Divergence	$\mathrm{D_{KL}}$	
Within-document topic concentration (variational)	$\gamma_{m_1,d,i}$	$\boldsymbol{\gamma}_{m_1,d}, \boldsymbol{\Gamma}_{m_1}$
Within-document topic probability (variational)	$\bar{\gamma}_{m_1,d,i_{m_1}}$	$\overline{\boldsymbol{\gamma}}_{m_1,d}, \overline{\boldsymbol{\Gamma}}_{m_1}$
Word-level topic probability (variational)	$\phi_{m_1,d,n,i}$	$\boldsymbol{\phi}_{m_1,d,n}, \boldsymbol{\Phi}_{m_1}$
Gamma function	Γ	
Digamma function	Ψ	
Indication of term identity for an observed word	ω_{d,n_d}^{v}	
Total Stage 2 models	$\mathcal{M}_2$	

Specific Stage 2 model	m_2	
Total realizations of a specific Stage 2 model	R_{m_2}	
Specific realization of a specific Stage 2 model	r	
Total change points	P_{m_2}	
Specific change point	p	
Total segments of the time series	S_{m_2}	
Specific segment	s	
Change point location	$\rho_{m_2,p}$	$\boldsymbol{\rho}_{m_2}$
Change point locations augmented with time range		$\widetilde{\boldsymbol{\rho}}_{m_2}$
Indication if a document is from a segment	$\xi_{m_2^r,d,s}$	$\boldsymbol{\xi}_{m_2^r,s}$, $\Xi_{m_2^r}$
Total covariates in a model	C_{m_2}	
Specific covariate	c	
Value of a covariate	$x_{m_2,d,c_{m_2}}$	$\boldsymbol{x}_{m_2,d}$, $\boldsymbol{X}_{m_2,s}$, $\boldsymbol{X}_{m_2}$
Within-segment regressors	$\eta_{m_2^r,s,i,c}$	$\boldsymbol{\eta}_{m_2^r,s,i}$, $\boldsymbol{H}_{m_2^r,s}$
Augmented within-segment regressors	$\acute{\eta}_{m_2^r,s,i,c}$	$\acute{\boldsymbol{\eta}}_{m_2^r,s,i}$, $\acute{\boldsymbol{H}}_{m_2^r,s}$, $\acute{\boldsymbol{H}}_{m_2^r}$, $\acute{\boldsymbol{H}}_{m_2}$
Precision on the prior for the regressors	λ	
General augmented within-segment regressor values	$\mathfrak{x}$	$\boldsymbol{\mathfrak{x}}$
Total MCMC chains	H	
Specific MCMC chain	h	
Temperature of a chain	a_h	$\boldsymbol{a}$
Inverse temperature of a chain	b_h	$\boldsymbol{b}$
Exponent controlling the temperature sequence	q	
Final temperature	a_H	
Penultimate temperature	a_{H-1}	
Total MCMC iterations	G	
Specific MCMC iteration	g	
Proposal iteration for the step to g	$\breve{g}$	
Statistical energy	$\mathcal{E}$	
Acceptance probability	$\mathcal{u}$	
Proposal distribution	$\mathcal{R}$	
Average step size in the proposal distribution	κ	
Number of burn-in iterations	$G_{\text{burn-in}}$	
Fraction of samples remaining after thinning	τ	
Document to segment mapping function	$f_{d\to s}$	
Cross entropy function	f_{CE}	
Softmax function	softmax, f_{S}	
Matrix multiplication function	f_{MM}	
Softmax of matrix multiplication function	$f_{\text{S(MM)}}$	
Cross entropy of softmax of matrix multiplication	$f_{\text{CE(S(MM))}}$	
Penalty function	f_{P}	
Partition function	f_{part}	
Segment function mapper	$f_{f(s)}$	
Kronecker Delta	δ_{ij}	

Table 2. Number of unique model realizations for a combination of time series length and number of change points.

Time steps → / Change points ↓	50	100	200	500
1	49	99	199	499
2	1176	4851	1970	124251
3	18424	156849	1293699	10507399
4	211876	3764376	63391251	2552446876
5	1906884	71523144	2472258789	252692240724

FIGURES

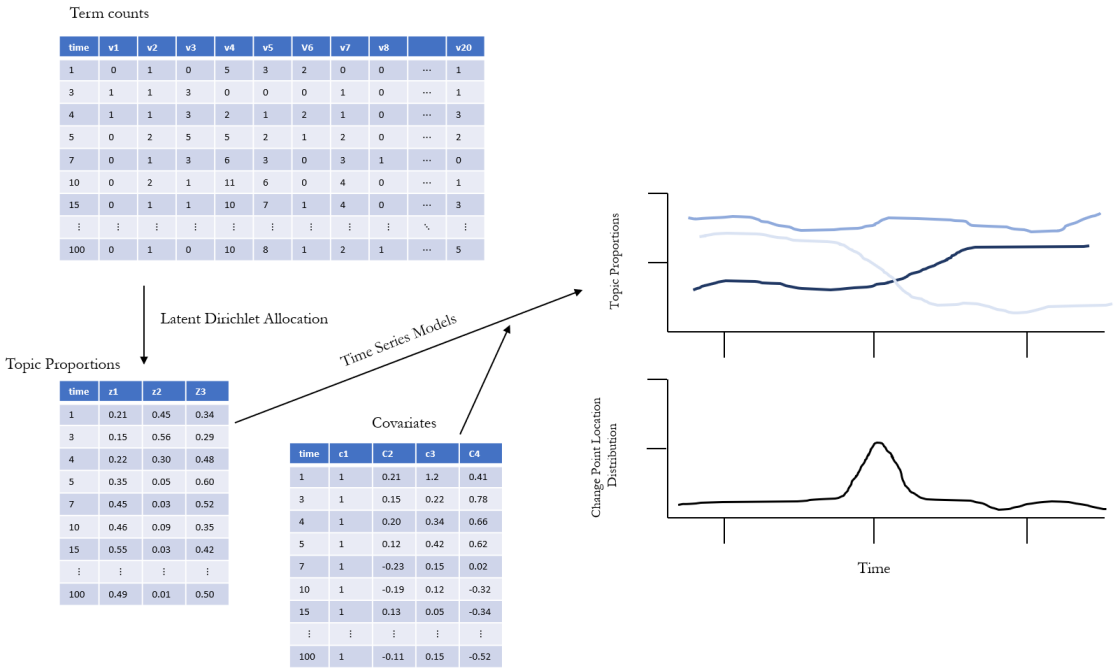

Fig 1. Schematic representation of the data-model relation of the LDATS framework. Documents are described by high-dimensional term counts, which are first reduced to low-dimension topic proportions. Then, the topic proportions are analyzed with time series models that include both covariate and change point dynamics.

Appendix 1: Expanding the probability of a corpus given an LDA model

For a given Latent Dirichlet Allocation (LDA; Blei *et al.* 2003) m_1, the probability of the entire corpus, given the parameters, is the product of the probabilities of each document, given the parameters:

$$\mathcal{P}\left(\boldsymbol{w}\middle|\alpha_{m_1}, \boldsymbol{B}_{m_1}\right) = \prod_{d=1}^{M} \mathcal{P}\left(\boldsymbol{w}_d\middle|\alpha_{m_1}, \boldsymbol{B}_{m_1}\right) \quad \text{[A1.1]}$$

The probability of a document's data given the parameters is the product of [1] the word-level term-identity distributions, given the model parameters governing the within-document topic distribution and word-level term distributions ($\mathcal{P}\left(w_{d,n}\middle|\boldsymbol{\theta}_{m_1,d}, \boldsymbol{B}_{m_1}\right)$) and [2] the within-document topic distribution parameters given the document-level concentration parameter ($\mathcal{P}\left(\boldsymbol{\theta}_{m_1,d}\middle|\alpha_{m_1}\right)$), integrated over the uncertainty in the within-document topic distribution ($\boldsymbol{\theta}_{m_1,d}$):

$$\mathcal{P}\left(\boldsymbol{w}_d\middle|\alpha_{m_1}, \boldsymbol{B}_{m_1}\right) = \int \mathcal{P}\left(\boldsymbol{\theta}_{m_1,d}\middle|\alpha_{m_1}\right)\left(\prod_{n=1}^{N_d} \mathcal{P}\left(w_{d,n}\middle|\boldsymbol{\theta}_{m_1,d}, \boldsymbol{B}_{m_1}\right)\right) d\boldsymbol{\theta}_{m_1,d} \quad \text{[A1.2]}$$

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 word-level term distribution parameters ($\mathcal{P}\left(w_{d,n}\middle|z_{m_1,d,n}, \boldsymbol{B}_{m_1}\right)$) and [2] the topic identity distribution given the parameters governing the within-document topic distribution ($\mathcal{P}\left(z_{m_1,d,n}\middle|\boldsymbol{\theta}_{m_1,d}\right)$), integrated (summed due to discreteness) over the uncertainty in topic type $z_{m_1,d,n}$

$$\mathcal{P}\left(w_{d,n}\middle|\boldsymbol{\theta}_{m_1,d}, \boldsymbol{B}_{m_1}\right) = \sum_{z_{m_1,d,n}} \mathcal{P}\left(w_{d,n}\middle|z_{m_1,d,n}, \boldsymbol{B}_{m_1}\right)\mathcal{P}\left(z_{m_1,d,n}\middle|\boldsymbol{\theta}_{m_1,d}\right) \quad \text{[A1.3]}$$

Substituting Eq. A1.3 into Eq. A1.2,

$$\mathcal{P}\left(\boldsymbol{w}_d\middle|\alpha_{m_1}, \boldsymbol{B}_{m_1}\right) = \int \mathcal{P}\left(\boldsymbol{\theta}_{m_1,d}\middle|\alpha_{m_1}\right)\left(\prod_{n=1}^{N_d} \sum_{z_{m_1,d,n}} \mathcal{P}\left(w_{d,n}\middle|z_{m_1,d,n}, \boldsymbol{B}_{m_1}\right)\mathcal{P}\left(z_{m_1,d,n}\middle|\boldsymbol{\theta}_{m_1,d}\right)\right) d\boldsymbol{\theta}_{m_1,d} \quad \text{[A1.4]}$$

and then Eq. A1.4 into Eq 1.1,

$$\mathcal{P}\left(\boldsymbol{w}\middle|\alpha_{m_1}, \boldsymbol{B}_{m_1}\right) = \prod_{d=1}^{M}\left[\int \mathcal{P}\left(\boldsymbol{\theta}_{m_1,d}\middle|\alpha_{m_1}\right)\left(\prod_{n=1}^{N_d} \sum_{z_{m_1,d,n}} \mathcal{P}\left(w_{d,n}\middle|z_{m_1,d,n}, \boldsymbol{B}_{m_1}\right)\mathcal{P}\left(z_{m_1,d,n}\middle|\boldsymbol{\theta}_{m_1,d}\right)\right) d\boldsymbol{\theta}_{m_1,d}\right] \quad \text{[A1.5]}$$

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

For the variational expansion of a Stage 1 LDA model m_1, $\boldsymbol{\Gamma}_{m_1}$ is an $M \times k_{m_1}$-matrix akin to $\boldsymbol{\theta}_{m_1}$, where row d corresponds to document d: $\gamma_{m_1,d,1} \ldots \gamma_{m_1,d,k_{m_1}}$ (or $\boldsymbol{\gamma}_{m_1,d}$), but contains the concentration parameters of a k_{m_1}-dimension Dirichlet distribution and therefore are not constrained to sum to 1, and $\boldsymbol{\Phi}_{m_1}$ is an $N \times k_{m_1}$-matrix whose rows correspond to the words across documents (indexed akin to $\boldsymbol{w}_d$ and $\boldsymbol{z}_d$) and whose columns correspond to the k_{m_1} topics. $\phi_{m_1,d,n,i}$ is the probability that word n within document d is from topic i, and $\boldsymbol{\phi}_{m_1,d,n;1\ldots k_{m_1}}$ (or $\boldsymbol{\phi}_{m_1,d,n}$) is a k_{m_1}-length vector of probabilities defining the categorical distribution of that word's topic identity ($\sum \boldsymbol{\phi}_{m_1 d,n} = 1$). $\boldsymbol{\Phi}_{m_1}$ contains M document-specific matrices (each is $N_d \times k_{m_1}$ and notated $\boldsymbol{\Phi}_{m_1,d}$). In comparison to α_{m_1} and $\boldsymbol{B}_{m_1}$, the variational parameters $\boldsymbol{\Gamma}_{m_1}$ and $\boldsymbol{\Phi}_{m_1}$ are document-specific and not coupled, thereby allowing estimation.

For a specific document d, the variational distribution ($\mathcal{Q}$) is

$$\mathcal{Q}\left(\boldsymbol{\theta}_{m_1,d}, \boldsymbol{z}_{m_1,d} \middle| \boldsymbol{\gamma}_{m_1,d}, \boldsymbol{\Phi}_{m_1,d}\right) = \mathcal{Q}\left(\boldsymbol{\theta}_{m_1,d} \middle| \boldsymbol{\gamma}_{m_1,d}\right) \prod_{n=1}^{N_d} \mathcal{Q}\left(z_{m_1,d,n} \middle| \boldsymbol{\phi}_{d,n}\right) \quad \text{[A2.1]}$$

In the Expectation ("E") Step in the VEM algorithm, the distribution $\mathcal{Q}\left(\boldsymbol{\theta}_{m_1,d}, \boldsymbol{z}_{m_1,d} \middle| \boldsymbol{\gamma}_{m_1,d}, \boldsymbol{\Phi}_{m_1,d}\right)$ is used to find a tight lower bound on $\mathcal{P}\left(\boldsymbol{w}_d \middle| \alpha_{m_1}, \boldsymbol{B}_{m_1}\right)$ by optimizing the variational parameters $\boldsymbol{\gamma}_{m_1,d}$ and $\boldsymbol{\Phi}_{m_1,d}$ (*i.e.*, finding $\boldsymbol{\gamma}^*_{m_1,d}$ and $\boldsymbol{\Phi}^*_{m_1,d}$, where the asterisks notate optimal values) with respect to minimizing the Kullback-Leibler Divergence ($\mathrm{D_{KL}}$) between $\mathcal{Q}\left(\boldsymbol{\theta}_{m_1,d}, \boldsymbol{z}_{m_1,d} \middle| \boldsymbol{\gamma}_{m_1,d}, \boldsymbol{\Phi}_{m_1,d}\right)$ and $\mathcal{P}\left(\boldsymbol{\theta}_{m_1}, \boldsymbol{z}_{m_1} \middle| \boldsymbol{w}, \alpha_{m_1}, \boldsymbol{B}_{m_1}\right)$:

$$\left(\boldsymbol{\gamma}^*_{m_1,d}, \boldsymbol{\Phi}^*_{m_1,d}\right) = \underset{\boldsymbol{\gamma}_{m_1,d}, \boldsymbol{\Phi}_{m_1,d}}{\arg\min}\ \mathrm{D_{KL}}\left(\mathcal{Q}\left(\boldsymbol{\theta}_{m_1,d}, \boldsymbol{z}_{m_1,d} \middle| \boldsymbol{\gamma}_{m_1,d}, \boldsymbol{\Phi}_{m_1,d}\right) \middle\| \mathcal{P}\left(\boldsymbol{\theta}_{m_1}, \boldsymbol{z}_{m_1} \middle| \boldsymbol{w}, \alpha_{m_1}, \boldsymbol{B}_{m_1}\right)\right) \quad \text{[A2.2]}$$

Minimization of the distance is achieved through an iterative fixed-point method, where the derivative of $\mathrm{D_{KL}}$ is set to zero, producing a pair of update equations (Blei *et al.* 2003). First, the parameters describing the topic allocation of each word ($\phi_{m_1,d,n,i}$) are updated based on the topic distribution for the document ($\boldsymbol{\gamma}_{m_1,d}$):

$$\phi_{m_1,d,n,i} \propto \beta_{m_1,d,w_{d,n},i} e^{\mathrm{E}_{\mathcal{Q}}\left[\log\left(\theta_{m_1,d,i}\right) \middle| \boldsymbol{\gamma}_{m_1,d}\right]} \quad \text{[A2.3]}$$

where $\mathrm{E}_{\mathcal{Q}}$ is the expected value of the (log-scale) topic probability and is calculated using the digamma function (Ψ), which is the logarithmic derivative of the gamma function ($\Psi(a) = \frac{d}{da}\log\left(\Gamma(a)\right)$), a quantity that is calculated through Taylor approximation:

$$\mathrm{E}_{\mathcal{Q}}\left[\log\left(\theta_{m_1,d,i}\right) \middle| \boldsymbol{\gamma}_{m_1,d}\right] = \Psi\left(\gamma_{m_1,d,i}\right) - \Psi\left(\sum \boldsymbol{\gamma}_{m_1,d}\right) \quad \text{[A2.4]}$$

Then, the parameters describing the topic distribution for the document ($\gamma_{m_1,d,i}$) are updated based on the word-level topic distributions for the sample ($\boldsymbol{\Phi}_{m_1,d}$):

$$\gamma_{m_1,d,i} = \alpha_{m_1} + \sum_{n=1}^{N_d} \phi_{m_1,d,n,i} \quad \text{[A2.5]}$$

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 ($\boldsymbol{\gamma}^*_{m_1,d}$ and $\boldsymbol{\Phi}^*_{m_1,d}$ have been found) for the set of main parameters ($\alpha_{m_1}, \boldsymbol{B}_{m_1}$)..

The Maximization ("M") Step in the VEM algorithm maximizes the overall lower bound with respect to the main model parameters α_{m_1} and $\boldsymbol{B}_{m_1}$ given the optimal variational parameters (Blei *et al.* 2003). 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 (Blei *et al.* 2003). The update for the topic-level term distribution ($\boldsymbol{B}_{m_1}$) is, analytically:

$$\beta_{m_1,i_{m_1}v} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d} \phi^*_{m_1,d,n,}\, \omega^v_{d,n} \qquad \text{[A2.7]}$$

where $\omega^v_{d,n}$ is an indicator variable based on the term identity (v) of the observed word ($w_{d,n}$):

$$\omega^v_{d,n} = \begin{cases} 1, & w_{d,n} = v \\ 0, & w_{d,n} \neq v \end{cases} \qquad \text{[A2.8]}$$

The update for the concentration parameter underlying the document-level topic distribution (α_{m_1}) requires an iterative approach to find a stationary point estimate. The optimization is conducted using the Newton-Raphson method (Ronning 1989), which repeats

$$\alpha^{\text{new}}_{m_1} = \alpha^{\text{old}}_{m_1} - \mathrm{DD}\left(\alpha^{\text{old}}_{m_1}\right)^{-1}\mathrm{D}\left(\alpha^{\text{old}}_{m_1}\right) \qquad \text{[A2.9]}$$

until convergence, where D represents the Jacobian matrix of first derivatives of a multivariate function and DD is the Hessian matrix of second derivatives of the function. 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 $\mathcal{Q}$ converges.

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 ($\log \mathrm{E}\left[\bar{\gamma}^*_{m_1^*,d_s,i}\right]_{m_2^r}$) to the linear predictors ($\boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,i}$), although we formulate our model as the expected proportion ($\mathrm{E}\left[\bar{\gamma}^*_{m_1^*,d_s,i}\right]_{m_2^r}$) being a function of exponentiated predictors $e^{\boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,i}}$. And we handle the sum-to-1 constraint by normalizing $e^{\boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,i}}$ with a document-specific partition function $f_{\mathrm{part}_{m_2^r,d_s}}$:

$$\mathrm{E}\left[\bar{\gamma}^*_{m_1^*,d_s,i}\right]_{m_2^r} = \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,i}}}{f_{\mathrm{part}_{m_2^r,d_s}}} \quad [A3.1]$$

Given the sum-to-1 constraint ($\sum_{i=1}^{k_{m_1^*}} \bar{\gamma}^*_{m_1^*,d_s,i} = \sum_{i=1}^{k_{m_1^*}} \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,i}}}{f_{\mathrm{part}_{m_2^r,d_s}}} = 1$), we can simply define $f_{\mathrm{part}_{m_2^r,d_s}} = \sum_{j=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,j}}$, where we replace the i with j to avoid confusion with the focal topic i. This produces a generalized equation that is often referred to as the softmax function:

$$\mathrm{E}\left[\bar{\gamma}^*_{m_1^*,d_s,i}\right]_{m_2^r} = \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,i}}}{\sum_{j=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,j}}} = \mathrm{softmax}\left(i, \boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,1}, \ldots, \boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,k_{m_1^*}}\right) \quad [A3.2]$$

However, because of the sum-to-1 constraint, only $k_{m_1^*} - 1$ of the proportions ($\bar{\gamma}^*_{m_1^*,d_s,i}$ in $\bar{\gamma}^*_{m_1^*,d_s,1} \ldots \bar{\gamma}^*_{m_1^*,d_s,k_{m_1^*}}$), and by extension only $k_{m_1^*} - 1$ of the parameter vectors ($\boldsymbol{\eta}_{m_2^r,s,i}$ in $\boldsymbol{\eta}_{m_2^r,s,1} \ldots \boldsymbol{\eta}_{m_2^r,s,k_{m_1^*}}$), are uniquely identifiable. Thus, we define an augmented parameter vectors $\acute{\boldsymbol{\eta}}_{m_2^r,s,i}$ (in $\acute{\boldsymbol{\eta}}_{m_2^r,s,1} \ldots \acute{\boldsymbol{\eta}}_{m_2^r,s,k_{m_1^*}}$), where

$$\acute{\boldsymbol{\eta}}_{m_2^r,s,i} = \boldsymbol{\eta}_{m_2^r,s,i} - \boldsymbol{\eta}_{m_2^r,s,1} \quad [A3.3]$$

setting the parameters associated with the first topic ($i = 1$) to 0 ($\acute{\boldsymbol{\eta}}_{m_2^r,s,1} = 0, e^{\boldsymbol{x}_{m_2^r,d_s}\boldsymbol{\eta}_{m_2^r,s,1}} = 1$). This reduces the number of free parameter vectors (and number of proportions estimated) by 1 to the $k_{m_1^*} - 1$ ($\acute{\boldsymbol{\eta}}_{m_2^r,s,2} \ldots \acute{\boldsymbol{\eta}}_{m_2^r,s,k_{m_1^*}}$) we are able to fit for this specific chunk of time $s_{m_2^r}$, resulting in the modified probability equation

$$\mathrm{E}\left[\bar{\gamma}^*_{m_1^*,d_s,i}\right]_{m_2^r} = \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}}}{\sum_{j=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}} = \mathrm{softmax}\left(i, \boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,1}, \ldots, \boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,k_{m_1^*}}\right) \quad [A3.4]$$

We combine all of the $k_{m_1^*}$ parameter vectors $\acute{\boldsymbol{\eta}}_{m_2^r,s,1}$ to $\acute{\boldsymbol{\eta}}_{m_2^r,s,k_{m_1^*}}$ (including the vector of 0s in $\acute{\boldsymbol{\eta}}_{m_2^r,s,1}$) into a matrix $\acute{\boldsymbol{H}}_{m_2^r,s}$, which has $k_{m_1^*}$ columns and a number of rows equal to the number of coefficients in model m_2 (C_{m_2}) including the intercept (*i.e.*, the length of $\boldsymbol{x}_{m_2^r,d_s}$).

This allows us to further condense the expected probability equation to

$$\mathrm{E}\left[\bar{\gamma}^{*}_{m_1^*,d_s,i}\right]_{m_2^r} = \mathrm{softmax}\left(i, \boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right) \quad [A3.5]$$

thereby facilitating use of the generalized linear modeling framework. We expand the model to predict the proportions across all of the $k_{m_1^*}$ topics within the document, which means we can drop the i input and produce the full set of values from the softmax function, which is a length- $k_{m_1^*}$ row vector corresponding to the topic distribution of a single document:

$$\mathrm{E}\left[\overline{\boldsymbol{\gamma}}^{*}_{m_1^*,d_s}\right]_{m_2^r} = \mathrm{softmax}\left(\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right) \quad [A3.6]$$

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

$$\mathrm{E}\left[\overline{\boldsymbol{\Gamma}}_{m_1^*,s}\right]_{m_2^r} = \mathrm{softmax}\left(\boldsymbol{X}_{m_2^r,s}\acute{\boldsymbol{H}}_{m_2^r,s}\right) \quad [A3.7]$$

where the covariates are held in a matrix ($\boldsymbol{X}_{m_2^r,s}$) with the number of columns equal to the number of coefficients (C_{m_2}) and the number of rows equal to the number of documents in the chunk ($M_{m_2,s}$). That is, $\boldsymbol{X}_{m_2^r,s}$ is a series of $\boldsymbol{x}_{m_2^r,d_s}$ row vectors. This equation relates directly to the generalized linear modeling equation that is typically written as $\mathrm{g}(\boldsymbol{Y}) = \boldsymbol{X\beta}$ or $\mathrm{E}[\boldsymbol{Y}] = \mathrm{g}^{-1}(\boldsymbol{X\beta})$, where g is the so-called link function and g^{-1} 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 (which generates the negative prior-penalized log-likelihood ($\mathcal{L}_{m_2^r,s} = -\log\left(\mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*,s}\middle|\boldsymbol{\rho}_{m_2^r,s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\middle|\boldsymbol{\rho}_{m_2^r,s}\right)\right)$), we first recognize that the loss for the segment is (weighted) additive across the documents within the chunk:

$$\mathcal{L}_{m_2^r,s} = -\sum_{d_s=1}^{M_{m_2^r,s}} u_{d_s}\left[-\log\left(\mathcal{P}\left(\overline{\boldsymbol{\Gamma}}_{m_1^*,s,d_s}\middle|\boldsymbol{\rho}_{m_2^r,s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\mathcal{P}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\middle|\boldsymbol{\rho}_{m_2^r,s}\right)\right)\right] \quad [A4.1]$$

We next acknowledge the order of operations of the component multivariate functions that comprise $\mathcal{L}_{m_2^r,s}$. To reduce notational clutter, we rename the cross entropy, softmax, matrix multiplication, and penalty functions f_{CE}, f_{S}, f_{MM}, and f_{P}, as well as rename the nested functions: the softmax of the matrix multiplication and the cross entropy of the softmax of the matrix multiplication become $f_{\mathrm{S(MM)}}$ and $f_{\mathrm{CE(S(MM))}}$. This allows us to write the loss equation for all documents within a time chunk as

$$\mathcal{L}_{m_2^r,s} = -\sum_{d_s=1}^{M_{m_2^r,s}} u_{d_s}\left[\sum_{i=1}^{k_{m_1^*}} f_{\mathrm{CE(S(MM))}}\left(\bar{\gamma}^*_{m_1^*,d_s,i}, i, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) + f_{\mathrm{P}}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\right)\right] \quad [A4.2]$$

which can be condensed via the summation across topics to

$$\mathcal{L}_{m_2^r,s} = -\sum_{d_s=1}^{M_{m_2^r,s}} u_{d_s}\left(f_{\mathrm{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) + f_{\mathrm{P}}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\right)\right) \quad [A4.3]$$

This highlights the chained (nested) 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 f and g chained as $\mathrm{f}(\mathrm{g}(\boldsymbol{a}))$ (where $\boldsymbol{a}$ contains the multivariate input values), we can write the function composite using the ring operator as $(\mathrm{f}\circ\mathrm{g})(\boldsymbol{a})$. We then take the multivariate derivative (denoted as function $\mathcal{D}$) of the composite:

$$\begin{aligned}\mathcal{D}\left(\mathrm{f}(\mathrm{g}(\boldsymbol{a}))\right) &= \mathcal{D}\left((\mathrm{f}\circ\mathrm{g})(\boldsymbol{a})\right)\\ &= (\mathcal{D}(\mathrm{f})\circ\mathrm{g})(\boldsymbol{a})\cdot\mathcal{D}\left(\mathrm{g}(\boldsymbol{a})\right)\\ &= \mathcal{D}(\mathrm{f})\left(\mathrm{g}(\boldsymbol{a})\right)\cdot\mathcal{D}\left(\mathrm{g}(\boldsymbol{a})\right)\\ &= \mathcal{D}(\mathrm{f})\left(\mathrm{g}(\boldsymbol{a})\right)\cdot\mathcal{D}(\mathrm{g})(\boldsymbol{a})\end{aligned} \quad [A4.4]$$

where $\cdot$ is the dot product operator. Thus, the derivative of f of g of $\boldsymbol{a}$ is the dot product of the derivative of f evaluated at g of $\boldsymbol{a}$ and the derivative of g evaluated at $\boldsymbol{a}$. Using the chain rule, we now expand the derivative of the loss function applied to a specific document within a specific chunk (d_s). We start by expanding the outer layers (f_{CE} and f_{S}):

$$\mathcal{D}\left(f_{\mathrm{CE(S(MM))}}\left(\bar{\gamma}^*_{m_1^*,d_s,i}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right) = \mathcal{D}(f_{\mathrm{CE}})\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)\cdot\mathcal{D}\left(f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right) \quad [A4.5]$$

We next expand the inner layers (f_{S} and f_{MM}) by working with the right-hand-side of the dot product:

$$\mathcal{D}\left(f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s},\acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)=\mathcal{D}(f_{\mathrm{S}})\left(f_{\mathrm{MM}}\left(\boldsymbol{x}_{m_2^r,d_s},\acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)\cdot\mathcal{D}(f_{\mathrm{MM}})\left(\boldsymbol{x}_{m_2^r,d_s},\acute{\boldsymbol{H}}_{m_2^r,s}\right) \quad [A4.6]$$

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

$$\begin{aligned}&\mathcal{D}\left(f_{\mathrm{CE(S(MM))}}\left(\bar{\gamma}^*_{m_1^*,d_s,i},\boldsymbol{x}_{m_2^r,d_s},\acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)\\&\quad=\mathcal{D}(f_{\mathrm{CE}})\left(\bar{\boldsymbol{\gamma}}^*_{m_1^*,d_s},f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s},\acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)\cdot\mathcal{D}(f_{\mathrm{S}})\left(f_{\mathrm{MM}}\left(\boldsymbol{x}_{m_2^r,d_s},\acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)\cdot\mathcal{D}(f_{\mathrm{MM}})\left(\boldsymbol{x}_{m_2^r,d_s},\acute{\boldsymbol{H}}_{m_2^r,s}\right)\end{aligned} \quad [A4.7]$$

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 singular functions f_{MM}, f_{S}, and f_{CE} and the composite functions $f_{\mathrm{S(MM)}}$ and $f_{\mathrm{CE(S(MM))}}$. To aid in this, we consider that the Jacobian of a function f contains the partial derivatives of each output (f_i, a general component of f) with respect to each input (a_j a general component of $\boldsymbol{a}$), which can be written generally as $\frac{\partial \mathrm{f}_i}{\partial a_j}$ or $\mathcal{D}_j\mathrm{f}_i$. 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 f_{MM} maps the $\acute{\boldsymbol{H}}_{m_2^r,s}$ matrix ($C_{m_2}\times k_{m_1^*}$) to the dimensions of $\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}$ ($1\times k_{m_1^*}$) by left-multiplying $\acute{\boldsymbol{H}}_{m_2^r,s}$by the covariate row matrix $\boldsymbol{x}_{m_2^r,d_s}$ ($1\times C_{m_2^r}$). Thus, its Jacobian has $k_{m_1^*}$ rows and $C_{m_2}k_{m_1^*}$ columns:

$$\mathcal{D}(f_{\mathrm{MM}})=\begin{bmatrix}\mathcal{D}_1 f_{\mathrm{MM}_1} & \cdots & \mathcal{D}_{C_{m_2}k_{m_1^*}} f_{\mathrm{MM}_1}\\ \vdots & \ddots & \vdots\\ \mathcal{D}_1 f_{\mathrm{MM}_{k_{m_1^*}}} & \cdots & \mathcal{D}_{C_{m_2}k_{m_1^*}} f_{\mathrm{MM}_{k_{m_1^*}}}\end{bmatrix} \quad [A4.8]$$

For notekeeping purposes, entry c,i in the coefficient matrix $\acute{\boldsymbol{H}}_{m_2^r,s}$ ($\acute{\eta}_{m_2^r,s,c,i}$) corresponds to the column $(i-1)C_{m_2}+c$ in the Jacobian. In effect, the coefficient matrix $\acute{\boldsymbol{H}}_{m_2^r,s}$ 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) f_{MM_l} (for l in $1\ldots k_{m_1^*}$) is just a linear combination of components

$$f_{\mathrm{MM}_l}=x_{m_2^r,d_s1}\acute{\eta}_{m_2^r,s,1,l}+x_{m_2^r,d_s,2}\acute{\eta}_{m_2^r,s,2,l}+\cdots+x_{m_2^r,d_s,C_{m_2}}\acute{\eta}_{m_2^r,s,C_{m_2},l} \quad [A4.9]$$

and therefore, the partial derivative of the output element $l_{m_1^*}$ with respect to an input element $c_{m_2^r},i_{m_1^*}$ is simply the relevant covariate or 0 (when beyond the relevant part of the Jacobian):

$$\mathcal{D}_{c,i}f_{\mathrm{MM}_l}=\begin{cases}x_{m_2^r,d_s,c}, & i=l\\ 0, & i\neq l\end{cases} \quad [A4.10]$$

Moving to the softmax function, f_{S} maps $\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}$ to $\mathrm{E}\left[\bar{\boldsymbol{\gamma}}^*_{m_1^*,d_s}\right]_{m_2^r}$, both of which are of dimension $1\times k_{m_1^*}$, because we are working within a single document. Thus, its Jacobian has $k_{m_1^*}$ rows and $k_{m_1^*}$ columns:

$$\mathcal{D}(f_{\mathrm{S}}) = \begin{bmatrix} \mathcal{D}_1 f_{\mathrm{S}_1} & \cdots & \mathcal{D}_{k_{m_1^*}} f_{\mathrm{S}_1} \\ \vdots & \ddots & \vdots \\ \mathcal{D}_1 f_{\mathrm{S}_{k_{m_1^*}}} & \cdots & \mathcal{D}_{k_{m_1^*}} f_{\mathrm{S}_{k_{m_1^*}}} \end{bmatrix} \quad [A4.11]$$

We can write a generalized equation for the entries by describing the partial derivative of output j with respect to input i, $\mathcal{D}_i f_{\mathrm{S}_j}$:

$$\begin{aligned} \mathcal{D}_i f_{\mathrm{S}_j} &= \frac{\partial f_{\mathrm{S}_j}}{\partial\left(\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right)_i} \\ &= \frac{\partial \dfrac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}}}{\partial\left(\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right)_i} \\ &= \frac{\partial}{\partial\left(\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right)_i}\frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}} \end{aligned} \quad [A4.12]$$

We decompose the generalized entry using the quotient rule, where for a function $\mathrm{f}(\boldsymbol{a})$ that is equal to the ratio of two other functions: $\mathrm{f}(\boldsymbol{a}) = \frac{\mathrm{g}(\boldsymbol{a})}{\mathrm{h}(\boldsymbol{a})}$, the derivative of the function is

$$\mathcal{D}\big(\mathrm{f}(\boldsymbol{a})\big) = \frac{\mathcal{D}\big(\mathrm{g}(\boldsymbol{a})\big)\mathrm{h}(\boldsymbol{a}) - \mathcal{D}\big(\mathrm{h}(\boldsymbol{a})\big)\mathrm{g}(\boldsymbol{a})}{[\mathrm{h}(\boldsymbol{a})]^2} \quad [A4.13]$$

Here, $\mathrm{g} = e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}$ and $\mathrm{h} = \sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{d_{s_{m_2^r}}}\acute{\boldsymbol{\eta}}_{s_{m_2^r},l}}$ and we differentiate each with respect to $\left(\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right)_i$:

$$\mathcal{D}_i f_{\mathrm{S}_j} = \frac{\dfrac{\partial e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\partial\left(\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right)_i}\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}} - \dfrac{\partial \sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}}{\partial\left(\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right)_i} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\left[\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}\right]^2} \quad [A4.14]$$

Regardless of the specific input i that we are computing the partial derivative for h with respect to, the value will always be $e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}}$:

$$\frac{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}}{\partial\left(\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right)_i} = e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}} \quad [A4.15]$$

For g, however, the value of the partial derivative is 0 unless $i = j$, in which case it is $e^{\boldsymbol{x}_{m_2^r,d_s j}}$:

$$\frac{\partial e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\partial\left(\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{H}}_{m_2^r,s}\right)_i} = \begin{cases} e^{\boldsymbol{x}_{m_2^r,d_s j}} & i = j \\ 0 & i \neq j \end{cases} \quad [A4.16]$$

Thus, when $i = j$,

$$\begin{aligned}
\mathcal{D}_i f_{\mathrm{S}_j} &= \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}} \sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}} - e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\left[\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}\right]^2} \\
&= \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}} \left(\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}} - e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}}\right)}{\left[\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}\right]^2} \\
&= \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}} \frac{\left(\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}} - e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}}\right)}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}} \\
&= \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}} \left(\frac{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}} - \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}}\right) \\
&= \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}} \left(1 - \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}}\right) \\
&= f_{\mathrm{S}_j}\left(1 - f_{\mathrm{S}_i}\right)
\end{aligned} \qquad \text{[A4.17]}$$

Similarly, when $i \neq j$,

$$\begin{aligned}
\mathcal{D}_i f_{\mathrm{S}_j} &= \frac{0 - e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\left[\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}\right]^2} \\
&= -\frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,i}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}} \frac{e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,j}}}{\sum_{l=1}^{k_{m_1^*}} e^{\boldsymbol{x}_{m_2^r,d_s}\acute{\boldsymbol{\eta}}_{m_2^r,s,l}}} \\
&= -f_{\mathrm{S}_i} f_{\mathrm{S}_j} \\
&= -f_{\mathrm{S}_j} f_{\mathrm{S}_i}
\end{aligned} \qquad \text{[A4.18]}$$

Combining these conditions, we have

$$\mathcal{D}_i f_{\mathrm{S}_j} = \begin{cases} f_{\mathrm{S}_j}\left(1 - f_{\mathrm{S}_i}\right) & i = j \\ -f_{\mathrm{S}_j} f_{\mathrm{S}_i} & i \neq j \end{cases} \qquad \text{[A4.19]}$$

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

$$\mathcal{D}_i f_{\mathrm{S}_j} = f_{\mathrm{S}_j}\left(\delta_{ij} - f_{\mathrm{S}_i}\right) \qquad \text{[A4.20]}$$

where

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad \text{[A4.21]}$$

The function set $f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)$ maps the $\acute{\boldsymbol{H}}_{m_2^r,s}$ matrix ($C_{m_2} \times k_{m_1^*}$) to the dimensions of

$\mathrm{E}\left[\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}\right]_{m_2^r}$ $(1 \times\ k_{m_1^*})$ and so, like f_{MM}, its Jacobian has $k_{m_1^*}$ rows and $C_{m_2}\ k_{m_1^*}$ columns. We combine the Jacobians of f_{S} and f_{MM} to define the Jacobian of $f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)$:

$$\mathcal{D}\left(f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right) = \begin{bmatrix} \mathcal{D}_1 f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)_1 & \cdots & \mathcal{D}_{C_{m_2}\ k_{m_1^*}} f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)_1 \\ \vdots & \ddots & \vdots \\ \mathcal{D}_1 f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)_{k_{m_1^*}} & \cdots & \mathcal{D}_{C_{m_2}\ k_{m_1^*}} f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)_{k_{m_1^*}} \end{bmatrix} \quad [A4.22]$$

For a general entry, the partial derivative is

$$\mathcal{D}_{c,i} f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)_j = \sum_{l=1}^{k_{m_1^*}} \mathcal{D}_l f_{\mathrm{S}_j} \cdot \mathcal{D}_{c,i} f_{\mathrm{MM}_l} \quad [A4.23]$$

Recalling that $\mathcal{D}_l f_{\mathrm{S}_j} \cdot \mathcal{D}_{c,i} f_{\mathrm{MM}_l}$ is 0 except for when $i = l$ (when it is $\boldsymbol{x}_{d_s,c,i}$), we can simplify this equation to

$$\mathcal{D}_{c,i} f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)_j = D_i f_{\mathrm{S}_j} x_{m_2^r,d_s,c,i} \quad [A4.24]$$

And recalling that $\mathcal{D}_i f_{\mathrm{S}_j} = f_{\mathrm{S}_j}\left(\delta_{ij} - f_{\mathrm{S}_i}\right)$, we can write this equation as

$$\mathcal{D}_{c,i} f_{\mathrm{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)_j = f_{\mathrm{S}_j}\left(\delta_{ij} - f_{\mathrm{S}_i}\right) x_{m_2^r,d_s,c,i} \quad [A4.25]$$

The cross-entropy function, f_{CE}, maps $\mathrm{E}\left[\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}\right]_{m_2^r}$ (dimension $1 \times k_{m_1^*}$) to the cross-entropy (loss) for the document, which is a scalar value. Thus, the Jacobian of f_{CE} is of dimension $1 \times k_{m_1^*}$:

$$\mathcal{D}(f_{\mathrm{CE}}) = \begin{bmatrix} \mathcal{D}_1 f_{\mathrm{CE}} & \cdots & \mathcal{D}_{k_{m_1^*}} f_{\mathrm{CE}} \end{bmatrix} \quad [A4.26]$$

A general entry in the Jacobian $\mathcal{D}_l f_{\mathrm{CE}}$ (*i.e.*, the partial derivative of the cross entropy loss with respect to topic l's probability) is

$$\mathcal{D}_l f_{\mathrm{CE}} = -\sum_{i=1}^{k_{m_1^*}} \frac{\partial \bar{\gamma}^*_{m_1^*,d_s,i} \log\left(\mathrm{E}\left[\bar{\gamma}^*_{m_1^*,d_s,i}\right]_{m_2^r}\right)}{\partial \mathrm{E}\left[\bar{\gamma}^*_{m_1^*,d_s,l}\right]_{m_2^r}} \quad [A4.27]$$

Notably, the only instance where $\mathrm{E}\left[\bar{\gamma}^*_{m_1^*,d_s,l}\right]_{m_2^r}$ appears in the function being derived is when $l = i$, in which case, the derivative is

$$\mathcal{D}_{l=i} f_{\mathrm{CE}} = -\frac{\bar{\gamma}^*_{m_1^*,d_s,i}}{\mathrm{E}\left[\bar{\gamma}^*_{m_1^*,d_s,l}\right]_{m_2^r}} \quad [A4.28]$$

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

Combining these conditions, we have

$$\mathcal{D}_l f_{\text{CE}} = \begin{cases} -\dfrac{\bar{\gamma}^*_{m_1^*,d_s,i}}{\text{E}\left[\bar{\gamma}^*_{m_1^*,d_s,l}\right]_{m_2^r}} & l = i \\ 0 & l \neq i \end{cases} \quad \text{[A4.29]}$$

For notation, we identify the element $l = i$ as $\tilde{\imath}$.

We can verify that the dimensionalities of the Jacobians are proper for combination via dot products: $\mathcal{D}(f_{\text{CE}})\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s} f_{\text{S(MM)}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right) \cdot \mathcal{D}(f_{\text{S}})\left(f_{\text{MM}}\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right) \cdot \mathcal{D}(f_{\text{MM}})\left(\boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)$. $\mathcal{D}(f_{\text{CE}})$ is $1 \times k_{m_1^*}$, $\mathcal{D}(f_{\text{S}})$ is $k_{m_1^*} \times k_{m_1^*}$, and $D(f_{\text{MM}})$ is $k_{m_1^*} \times C_{m_2} k_{m_1^*}$. Thus, each of the two dot products has proper component matrices. In addition, the resulting matrix is $1 \times C_{m_2} k_{m_1^*}$, which heuristically matches the fact that the composite of the three functions maps the set of parameters ($C_{m_2} k_{m_1^*}$ in total) to a single scalar value of the cross-entropy loss. Having verified the dimensions, we combine the elements across the three Jacobians to determine the derivative of f_{CE} of f_{S} of f_{MM}: $\mathcal{D}\left(f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)$. $f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)$ maps the $C_{m_2} k_{m_1^*}$ parameters (each $\acute{\eta}_{s_{m_2^r},c_{m_2^r},i_{m_1^*}}$ entry in $\acute{\boldsymbol{H}}_{m_2^r,s}$) to a scalar output (cross-entropy loss), so the resulting Jacobian is of dimensions $1 \times C_{m_2} k_{m_1^*}$:

$$\begin{aligned} &\mathcal{D}\left(f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right) \\ &= \left[\mathcal{D}_1 f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) \quad \cdots \quad \mathcal{D}_{C_{m_2} k_{m_1^*}} f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right] \end{aligned} \quad \text{[A4.30]}$$

Similar to the rows in $\mathcal{D}(f_{\text{MM}})$, the single row of partial derivatives in $\mathcal{D}\left(f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)$ corresponds to the column-major order linearized $\acute{\boldsymbol{H}}_{m_2^r,s}$. Following the indexing of $\mathcal{D}(f_{\text{MM}})$, we will index $\mathcal{D}\left(f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)$ with c and i (as c, i), where c, i refers to column (element) $cC_{m_2} + i$ in $\mathcal{D}\left(f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right)\right)$. For a general entry c, i, then

$$\mathcal{D}_{c,i} f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) = \sum_{j=1}^{k_{m_1^*}} \mathcal{D}_j(f_{\text{CE}})\left(\text{E}\left[\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}\right]_{m_2^r}\right) \cdot \mathcal{D}_{ci} f_{\text{S(MM)}}\left(j, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) \quad \text{[A4.31]}$$

Since only the $\tilde{\imath}$ element of $\mathcal{D}_j f_{\text{CE}}\left(\text{E}\left[\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}\right]_{m_2^r}\right)$ is non-0, in which case it is $-\dfrac{\bar{\gamma}^*_{m_1^*,d_s,\tilde{\imath}}}{\text{E}\left[\bar{\gamma}^*_{m_1^*,d_s,\tilde{\imath}}\right]_{m_2^r}}$, we can simplify this equation to

$$\mathcal{D}_{c,i} f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) = -\frac{\bar{\gamma}^*_{m_1^*,d_s,\tilde{\imath}}}{\text{E}\left[\bar{\gamma}^*_{m_1^*,d_s,\tilde{\imath}}\right]_{m_2^r}} \cdot \mathcal{D}_{ci} f_{\text{S(MM)}}\left(\tilde{\imath}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) \quad \text{[A4.32]}$$

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

$$\mathcal{D}_{c,i} f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) = -\frac{\bar{\gamma}^*_{m_1^*,d_s,\tilde{\imath}}}{\text{E}\left[\bar{\gamma}^*_{m_1^*,d_s,\tilde{\imath}}\right]_{m_2^r}} \cdot f_{\text{S}_{\tilde{\imath}}}\left(\delta_{i\tilde{\imath}} - f_{\text{S}_i}\right) x_{m_2^r,d_s,c,i} \quad \text{[A4.33]}$$

Noting that, by our definition, $f_{\mathrm{S}\tilde{\imath}} = \mathrm{E}\left[\bar{\gamma}^{*}_{m^{*}_{1},d_s,\tilde{\imath}}\right]_{m^{r}_{2}}$, we can further simplify this equation:

$$\begin{aligned}\mathcal{D}_{c,i}\, f_{\mathrm{CE(S(MM))}}\left(\bar{\boldsymbol{\gamma}}^{*}_{m^{*}_{1},d_s}, \boldsymbol{x}_{m^{r}_{2},d_s}, \acute{\boldsymbol{H}}_{m^{r}_{2},s}\right) &= -\frac{\bar{\gamma}^{*}_{m^{*}_{1},d_s,i}}{f_{\mathrm{S}\tilde{\imath}}} \cdot f_{\mathrm{S}\tilde{\imath}}\left(\delta_{i\tilde{\imath}} - f_{\mathrm{S}i}\right) x_{m^{r}_{2},d_s,c,i} \\ &= -\bar{\gamma}^{*}_{m^{*}_{1},d_s,i} \cdot \left(\delta_{i\tilde{\imath}} - f_{\mathrm{S}i}\right) x_{m^{r}_{2},d_s,c,i} \\ &= \bar{\gamma}^{*}_{m^{*}_{1},d_s,i} \cdot \left(f_{\mathrm{S}i} - \delta_{i\tilde{\imath}}\right) x_{m^{r}_{2},d_s,c,i}\end{aligned} \quad [\mathrm{A4.34}]$$

This row matrix $\mathcal{D}\left(f_{\mathrm{CE(S(MM))}}\left(\bar{\boldsymbol{\gamma}}^{*}_{m^{*}_{1},d_s}, \boldsymbol{x}_{m^{r}_{2},d_s}, \acute{\boldsymbol{H}}_{m^{r}_{2},s}\right)\right)$ holds the $C_{m_2} k_{m^{*}_{1}}$ partial derivatives associated with the C_{m_2} parameters and the $k_{m^{*}_{1}}$ 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 $C_{m_2} k_{m^{*}_{1}}$ model parameters $\acute{\boldsymbol{H}}_{m^{r}_{2},s}$, we start with

$$\mathcal{L}^{*}_{m^{r}_{2},s} = -\sum_{d_s=1}^{M_{m^{r}_{2},s}} u_{d_s}\left(f_{\mathrm{CE(S(MM))}}\left(\bar{\boldsymbol{\gamma}}^{*}_{m^{*}_{1},d_s}, \boldsymbol{x}_{m^{r}_{2},d_s}, \acute{\boldsymbol{H}}_{m^{r}_{2},s}\right) + f_{\mathrm{P}}\left(\acute{\boldsymbol{H}}_{m^{r}_{2},s}\right)\right) \quad [\mathrm{A4.35}]$$

focus in on a specific document d_s,

$$\mathcal{L}^{*}_{m^{r}_{2},d_s} = -\left(f_{\mathrm{CE(S(MM))}}\left(\bar{\boldsymbol{\gamma}}^{*}_{m^{*}_{1},d_s}, \boldsymbol{x}_{m^{r}_{2},d_s}, \acute{\boldsymbol{H}}_{m^{r}_{2},s}\right) + f_{\mathrm{P}}\left(\acute{\boldsymbol{H}}_{m^{r}_{2},s}\right)\right) \quad [\mathrm{A4.36}]$$

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

$$\mathcal{D}_{c,i}\left(\mathcal{L}^{*}_{m^{r}_{2},d_s}\right) = \mathcal{D}_{c,i}\, f_{\mathrm{CE(S(MM))}}\left(\bar{\boldsymbol{\gamma}}^{*}_{m^{*}_{1},d_s}, \boldsymbol{x}_{m^{r}_{2},d_s}, \acute{\boldsymbol{H}}_{m^{r}_{2},s}\right) + \mathcal{D}_{c,i} f_{\mathrm{P}}\left(\acute{\boldsymbol{H}}_{m^{r}_{2},s}\right) \quad [\mathrm{A4.37}]$$

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

$$f_{\mathrm{P}}\left(\acute{\boldsymbol{H}}_{m^{r}_{2},s}\right) = \sum_{i=i}^{k_{m^{*}_{1}}} \sum_{c=1}^{C_{m_2}} \lambda_{m_2}\left(\acute{\eta}_{m^{r}_{2},s,i,c}\right)^2 \quad [\mathrm{A4.38}]$$

the partial derivative of the penalty with respect to c, i is

$$\mathcal{D}_{c,i} f_{\mathrm{P}}\left(\acute{\boldsymbol{H}}_{m^{r}_{2},s}\right) = \sum_{i=i}^{k_{m^{*}_{1}}} \sum_{c=1}^{C_{m_2}} \frac{\partial \lambda_{m_2}\left(\acute{\eta}_{m^{r}_{2},s,i,c}\right)^2}{\partial\left(\acute{\eta}_{m^{r}_{2},s,i,c}\right)} \quad [\mathrm{A4.39}]$$

which evaluates to

$$\mathcal{D}_{c,i} f_{\mathrm{P}}\left(\acute{\boldsymbol{H}}_{m^{r}_{2},s}\right) = \sum_{i=i}^{k_{m^{*}_{1}}} \sum_{c=1}^{C_{m_2}} 2\lambda_{m_2} \acute{\eta}_{m^{r}_{2},s,i,c} \quad [\mathrm{A4.40}]$$

Given that we also know that $\mathcal{D}_{c,i}\left(\mathcal{L}^*_{m_2^r,d_s}\right) = \mathcal{D}_{c,i} f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) + \mathcal{D}_{c,i} f_{\text{P}}\left(\acute{\boldsymbol{H}}_{m_2^r,s}\right)$ and $\mathcal{D}_{c,i} f_{\text{CE(S(MM))}}\left(\overline{\boldsymbol{\gamma}}^*_{m_1^*,d_s}, \boldsymbol{x}_{m_2^r,d_s}, \acute{\boldsymbol{H}}_{m_2^r,s}\right) = \bar{\gamma}^*_{m_1^*,d_s,i} \cdot \left(f_{\text{S}_i} - \delta_{i\bar{\imath}}\right) x_{m_2^r,d_s,c,i}$, the partial derivative of penalized loss with respect to input ci ($\mathcal{D}_{c,i}\left(\mathcal{L}^*_{m_2^r,d_s}\right)$) is now fully defined:

$$\mathcal{D}_{c,i}\left(\mathcal{L}^*_{m_2^r,d_s}\right) = \bar{\gamma}^*_{m_1^*,d_s,i} \cdot \left(f_{\text{S}_i} - \delta_{i\bar{\imath}}\right) x_{m_2^r,d_s,c,i} + \sum_{i=i}^{k_{m_1^*}} \sum_{c=1}^{C_{m_2}} 2\lambda \acute{\eta}_{m_2^r,s,i,c} \qquad \text{[A4.41]}$$

This equation determines the gradient of the loss across with respect to each input (parameter-topic combination c, i) within a single document. Acknowledging multiple documents fall under the same parameter combination, we simply sum across all $d_{s_{m_2^r}}$:

$$\mathcal{D}_{c,i}\left(\mathcal{L}^*_{m_2^r,s}\right) = \sum_{d_s}^{M_{m_2^r,s}} \left[\bar{\gamma}^*_{m_1^*,d_s,i} \cdot \left(f_{\text{S}_i} - \delta_{i\bar{\imath}}\right) x_{m_2^r,d_s,c,i} + \sum_{i=i}^{k_{m_1^*}} \sum_{c=1}^{C_{m_2}} 2\lambda \acute{\eta}_{m_2^r,s,i,c} \right] \qquad \text{[A4.42]}$$

and distribute the summation

$$\mathcal{D}_{c,i}\left(\mathcal{L}^*_{m_2^r,s}\right) = \sum_{d_s}^{M_{m_2^r,s}} \bar{\gamma}^*_{m_1^*,d_s,i} \cdot \left(f_{\text{S}_i} - \delta_{i\bar{\imath}}\right) x_{m_2^r,d_s,c,i} + \sum_{d_s=1}^{M_{s_{m_2^r}}} \sum_{i=i}^{k_{m_1^*}} \sum_{c=1}^{C_{m_2}} 2\lambda \acute{\eta}_{m_2^r,s,i,c} \qquad \text{[A4.43]}$$

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

Appendix 5: Segment Function Mapper

The function $f_{f(s)}$ is used to map a function to segments of data and has three inputs: [1] the function to map, [2] the matrix multiplication to map it onto, and [3] the indication matrix that defines the mapping. Specifically here, we use $\boldsymbol{\Xi}_{m_2^r}$ to map the $\mathsf{softmax}$ function to the matrix multiplication $\boldsymbol{X}_{m_2^r}\acute{\boldsymbol{H}}_{m_2^r}$:

$$\mathrm{E}\left[\overline{\boldsymbol{\Gamma}}_{m_1^*}\right]_{m_2^r} = f_{f(s)}\left(\mathsf{softmax}, \boldsymbol{X}_{m_2^r}\acute{\boldsymbol{H}}_{m_2^r}, \boldsymbol{\Xi}_{m_2^r}\right) \quad \text{[A5.1]}$$

The operation of $f_{f(s)}$ can be considered algorithmically:

For each column in $\boldsymbol{\Xi}_{m_2^r}$, which corresponds to a segment s (in $1 \ldots S_{m_2}$)
  Select the rows from $\boldsymbol{X}_{m_2^r}$ where $\boldsymbol{\xi}_{m_2^r,s} = 1$ (the document is in the segment): $\boldsymbol{X}_{m_2^r,s}$
  Select columns $C_{m_2}(s-1)+1$ to $C_{m_2}s$ from $\acute{\boldsymbol{H}}_{m_2^r}$ (the regressors are for the segment): $\acute{\boldsymbol{H}}_{m_2^r,s}$
  Apply the $\mathsf{softmax}$ function to the selected entries: $\boldsymbol{X}_{m_2^r,s}\acute{\boldsymbol{H}}_{m_2^r,s}$