ALLOCATIVE POISSON FACTORIZATION FOR COMPUTATIONAL SOCIAL SCIENCE

A Dissertation Presented

by

AARON JOSEPH STERIADE SCHEIN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2019

College of Information and Computer Sciences

ALLOCATIVE POISSON FACTORIZATION FOR COMPUTATIONAL SOCIAL SCIENCE

A Dissertation Presented

by

AARON JOSEPH STERIADE SCHEIN

Approved as to style and content by:

Hanna Wallach, Chair

Dan Sheldon, Member

Ben Marlin, Member

Pat Flaherty, Member

Mingyuan Zhou, Member

David Blei, Member

James Allan, Department Chair
College of Information and Computer Sciences

DEDICATION

To Hanna Wallach, who believed in me before she probably should have.

ABSTRACT

ALLOCATIVE POISSON FACTORIZATION FOR COMPUTATIONAL SOCIAL SCIENCE

MAY 2019

AARON JOSEPH STERIADE SCHEIN

B.A., POLITICAL SCIENCE, UNIVERSITY OF MASSACHUSETTS AMHERST

B.A., LINGUISTICS, UNIVERSITY OF MASSACHUSETTS AMHERST

M.A., LINGUISTICS, UNIVERSITY OF MASSACHUSETTS AM HERST

M.S., COMPUTER SCIENCE, UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Hanna Wallach

Social science data often comes in the form of high-dimensional discrete data such as categorical survey responses, social interaction records, or text. These data sets exhibit high degrees of sparsity, missingness, overdispersion, and burstiness, all of which present challenges to traditional statistical modeling techniques. The framework of Poisson factorization (PF) has emerged in recent years as a natural way to model high-dimensional discrete data sets. This framework assumes that each observed count in a data set is a Poisson random variable $y_{\boldsymbol{\delta}} \sim \text{Pois}(\mu_{\boldsymbol{\delta}})$ whose rate parameter $\mu_{\boldsymbol{\delta}}$ is a function of shared model parameters. This thesis examines a specific subset of Poisson factorization models that constrain $\mu_{\boldsymbol{\delta}}$ to be a multilinear function of shared model parameters. This subset of models—hereby referred to as *allocative*

Poisson factorization (APF)—enjoys a significant computational advantage: posterior inference scales linearly with only the number of non-zero counts in the data set. A challenge to constructing and performing inference in APF models is that the multilinear constraint on $\mu_{\boldsymbol{\delta}}$—which must be non-negative, by the definition of the Poisson distribution—means that the shared model parameters must themselves be non-negative. Constructing models that capture the complex dependency structures inherent to social processes—e.g., networks with overlapping communities of actors or bursty temporal dynamics—without relying on the analytic convenience and tractability of the Gaussian distribution requires novel constructions of non-negative distributions—e.g., gamma and Dirichlet—and innovative posterior inference techniques. This thesis presents the APF analogue to several widely-used models—i.e., CP decomposition (Chapter 4), Tucker decomposition (Chapter 5), and linear dynamical systems (Chapters 6 and 7) and shows how to perform Bayesian inference in APF models under local differential privacy (Chapter 8). Most of these chapters introduce novel auxiliary-variable augmentation schemes to facilitate posterior inference using both Markov chain Monte Carlo and variational inference algorithms. While the task of modeling international relations event data is a recurrent theme, the models presented are applicable to a wide range of tasks in many fields.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Motivation

Much of the work in this thesis is motivated by modeling political dyadic event data—i.e., micro-records of the form "country i took action a to country j at time t"—that political scientists have collected and analyzed for several decades. Dyadic event data concretize the phenomena of international relations into a series of pairwise country–country interactions. These interactions are recorded in a standardized scheme wherein country actors and possible actions types are assigned categorical codes. Such data permit empirical and reproducible approaches to answering political science questions—e.g., what causes war?—via statistical modeling. However, dyadic events present challenges to the standard statistics toolkit. Some of these challenges stem from the data consisting of *rare events*, as observed by King and Zeng [2001], who explain political scientists' interest in such data:

> "Many of the most significant events in international relations—wars, coups, revolutions, massive economic depressions, economic shocks—are **rare events**. They occur infrequently but are considered of great importance."

The types of events political scientists choose to track are those that are significant or noteworthy and these tend to be rare. Dyadic event data sets are thus naturally *sparse*—i.e., most event types are never observed—since most countries do not take most actions towards most other countries at most times (see Fig. 1.1a). Further challenges are introduced by the source from which dyadic event data sets are

collected—i.e., the news—which strains the interpretation that they directly measure international relations. Schrodt [1995] instead offers a more nuanced interpretation:

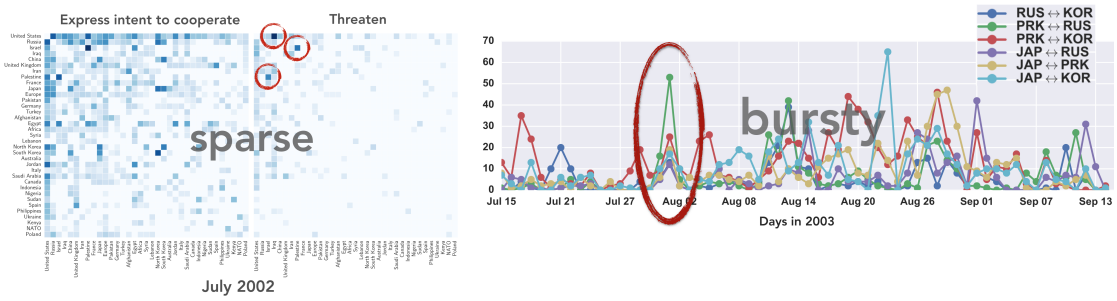
> "Event data are a formal method of **measuring the phenomena that contribute to foreign policy perceptions**. Event data are generated by examining thousands of newspaper reports on the day to day interactions of nation-states and assigning each reported interaction a numerical score or a categorical code."

Dyadic events do not directly provide a complete or "objective" account of international relations. Rather, they provide an account of what shaped collective *perception of* international relations—i.e., what the media paid attention to, when, and to what degree. What the media deems newsworthy and when is thus a critical factor influencing the empirical patterns in dyadic event data sets. Boydstun [2013] describes the dynamics of media attention:

> "[The news] does not ebb and flow; rather it fixates and explodes. In turn, the explosive nature of media dynamics exacerbates the degree of skew in news coverage across policy issues, such that a few issues receive the lion's share of coverage while most issues go unnoticed. These patterns—**explosiveness and skew**—are endemic to the media as an institution, they have far-reaching implications for politics and society."

Dyadic event data sets are *bursty*—i.e., many similar events are observed suddenly around the same time (see Fig. 1.1b). Burstiness is itself a form of *overdispersion*—i.e., unobserved heterogeneity—across time. Other modes of the data also exhibit overdispersion—e.g., some countries are observed sending many more actions than others. Moreover, these data sets conflate the inherent burstiness of international relations phenomena with the "explosive and skewed" dynamics of media attention. Finally, while many political events are truly rare, events that do occur may go unreported in the news—thus, these data sets conflate *missingness* with non-occurrence.

In addition to the challenging statistical properties of dyadic event data, the presence of what King [2001] refers to as "complex dependence structures in international relations" tie observations in a way that violates the independence assumptions of the standard regression toolkit. Stewart [2014] explains:



(a) Two sender-by-receiver slices of the data tensor for a single time step and action type. The circled points correspond to the counts for the pairs USA→Iraq, Israel→Palestine, and Palestine→Israel for the action "Threaten" during July 2002. Countries are ordered by activity; only the top 50×50 are displayed. The tensor becomes sparse quickly moving from out from the top left.

(b) Count time-series for action type "Consult" taken between all six undirected pairs of Russia, South Korea, North Korea, and Japan. These four countries, along with the USA and China, were involved in the Six Party Talks in 2003, a series of multilateral talks about disarming North Korea. The event counts for these pairs are correlated due to an underlying multilateral relation. The dynamics they exhibit are bursty—i.e., they remain near zero for extended periods, abruptly burst, and then return.

Figure 1.1: Two ways to represent dyadic event counts: slices of a count tensor (*left*) or count time-series (*right*). These two representations highlight two different characteristic properties of high-dimensional discrete data: sparsity and burstiness.

> "Regression, in particular the generalized linear model (GLM), plays a central role in the social sciences as the default statistical method for the analysis of relationships in quantitative data. GLMs leverage the assumption that observations are conditionally independent given the covariates in order to allow for tractable inference. Methodologists have periodically warned of the inaccuracy of these standard regression tools in the **presence of unmodeled dependence** between units."

Stewart [2014] continues by advocating for models of latent dependence structure:

> "One way to think about dependence is as arising due to **unobserved heterogeneity** between repeated units within the data. Thus if we had **the right set of control variables**, we could treat the remaining stochastic error as independent across observations. Subject matter experts often have an implicit understanding of unmodeled dependence and are able to specify the important groups within the data."

Over the past twenty years, political scientists have debated whether international relations and the inherently *multilateral* phenomena therein (e.g., treaties, alliances)

can be validly studied using only dyadic (i.e., pairwise or *bilateral*) events. This debate began with Green et al. [2001]'s demonstration that many regression analyses based on dyadic events were biased due to implausible independence assumptions. Researchers continued to expose such biases, e.g., [Erikson et al., 2014], and some, like Poast [2010], have even advocated eschewing dyadic data on principle, calling instead for the development of multilateral event data sets. Taking the opposite viewpoint—i.e., that dyadic events can be used conduct meaningful analyses of multilateral phenomena—other researchers have been developing latent factor regression models which explicitly infer unobserved dependence structure and then condition on it in subsequent regression analyses. Hoff and Ward [2004] pioneered this approach for analyzing dyadic events with a latent factor model for inferring (and thus controlling for) latent network efforts. This approach has seen an increase in interest and activity over the past few years e.g., [Stewart, 2014, Hoff, 2015, Hoff et al., 2016].

The challenging statistical properties of dyadic event data—sparsity, missingness, overdispersion, burstiness—are generally characteristic of *high-dimensional discrete data* which arise throughout the social sciences. Quantitative social science was traditionally dominated by the analysis of social surveys which solicit categorical responses—e.g., What is your party affiliation? Circle one: {Dem, Rep, Green, ...}—on a series of questions and demographic variables. These responses are analyzed collectively as contingency tables of sparse counts [Yvonne et al., 1975, Clogg and Goodman, 1984, Wickens, 2014] wherein zeros conflate missingness with true non-occurrence as well as with "structural zeros" [Yvonne et al., 1975, Chapter 5]—i.e., impossible event types. Analysis of US Senate voting behavior is often performed on the basis of a categorical Senators-by-bills voting matrix [Bafumi et al., 2005] and may additionally involve the actual text of bills [Gerrish and Blei, 2011]. More generally, the analysis of text—perhaps the must abundant source of high-dimensional discrete data in any field—is ubiquitous throughout social science [Grimmer and Stewart,

2013]. High-dimensional discrete data arises in many other fields beyond the social sciences as well. Sparse and overdispersed count matrices arise in both population genetics [Pritchard et al., 2000] and analysis of RNA sequencing [Lee et al., 2013]. Sparse and bursty neural spike trains are the study of intense study in neuroscience motivating models for multivariate time-series [Macke et al., 2011] and latent network structure [Linderman and Adams, 2014, Dunson and Xing, 2009]. Burstiness is a property observed in discrete events sequences across many other social and physical processes as well [Kleinberg, 2003, Goh and Barabási, 2008]. In ecology, discrete data sets that conflate missingness with true non-occurrence are commonplace and referred to as "presence-only data" [Pearce and Boyce, 2006].

In summary, due to the inherent rareness of noteworthy political events and the explosive dynamics of media attention, dyadic event data sets exhibit certain statistical properties that challenge the distributional assumptions of the traditional regression toolkit while the presence of "complex dependence structures" violates standard independence assumptions. Empirical and reproducible approaches to meaningfully test hypotheses about international relations phenomena require the development of "good" models for dyadic event data—i.e., models that are robust to the challenging statistical properties and ones that can faithfully infer hypothesized dependencies like temporal and network structures. Such models have the potential to address similar problems throughout the social and physical sciences that feature high-dimensional discrete data with complex dependence structure.

1.2 Contributions

This thesis defines and develops *allocative Poisson factorization (APF)*, a class of statistical models that are tailored to but general within the class of problems involving *high-dimensional discrete data*. Many models in statistics, machine learning, and signal processing are unified under APF by the basic probabilistic assumption (some-

times made implicitly) that observations are conditionally independent Poisson random variables—$y_{\boldsymbol{\delta}} \sim \text{Pois}\,(\mu_{\boldsymbol{\delta}})$—whose latent rates are multilinear functions of shared model parameters $\mu_{\boldsymbol{\delta}} = \sum_{\boldsymbol{\kappa} \in \mathcal{K}} \mu_{\boldsymbol{\delta}\boldsymbol{\kappa}}$. This assumption is natural for discrete data of rare events and yields models that are robust to sparsity, missingness, overdispersion, and burstiness. Moreover this basic assumption naïvely yields inference algorithms whose time and memory complexity scale linearly with only the set of observed events as opposed to the number of possible event types. The non-negative constraint on model parameters permits their interpretation in multiple overlapping ways. This facilitates the way in which social scientists use models to encode substantive hypotheses about complex dependence structure underlying their data.

The main challenge to constructing and performing inference in APF models stems from the multilinear structure of the latent rate $\mu_{\boldsymbol{\delta}}$ which must be non-negative, by definition of the Poisson distribution. This multilinear structure is necessary in obtaining the attractive computational property of APF models but excludes the use of non-linear functions to link real-valued parameters to the non-negative rate. Constructing models that capture the complex dependence structures inherent to social processes—e.g., networks that link overlapping communities of actors or excitatory temporal patterns—without relying on the analytic convenience and tractability of the Gaussian distribution requires novel constructions of non-negative distributions—e.g., gamma and Dirichlet—and innovative posterior inference techniques. Recent advances in auxiliary variable augmentation schemes and newfound connections different probability distributions have rapidly expanded the class of APF models that admit tractable posterior inference. The main technical goal of this thesis is to facilitate that expansion by providing concrete examples of novel APF models capable of capturing complex dependence structure and exploit new auxiliary variable techniques for tractable posterior inference. Towards that goal, this thesis also introduces

several new augmentation schemes and introduces novel properties of probability distributions that have applicability beyond APF.

- Chapter 2 provides background on *Bayesian latent variable modeling* (BLVM) and *tensor decomposition*, the two frameworks within which APF is defined. The search for a "good" model is an iterative cycle for which these frameworks are well-suited. BLVM cleanly distinguishes three phases of modeling [Blei, 2014]—i.e., model building, model fitting, and model critiquing—while tensor decomposition further organizes the model building phase by distinguishing a model's relational assumptions from its distributional ones. These complementary frameworks promote modularity and highlight the connections between models thus facilitating the iterative cycle of modeling. In this chapter, I also provide background on that iterative cycle—i.e., the *hypothetico-deductive approach*—and how concept of *measurement* in social science motivates it.

- In Chapter 3, I define APF and describe its core properties. I discuss the trade-offs between the two common choices of non-negative prior distributions—gamma and Dirichlet—notably, the difference in the way missing data can be handled under each. I also sketch different motifs of posterior inference in APF models—i.e., thinning, latent source aggregation, masking, imputation, and conjugacy. Finally, I provide historical notes that trace the lineage of APF—which unifies convergent threads in statistics, machine learning, and signal processing—and highlight connections to and applications in many fields.

- Chapter 4 introduces an APF version of *CP decomposition*—the most widely used form of tensor decomposition—along with a scalable variational inference algorithm. I demonstrate that this model infers interpretable dependence structure in dyadic event data—i.e., *multilateral relations*—and that its out-of-sample predictive performance is more robust to the inherent sparsity of dyadic

event data than widely-used non-APF equivalents. I also detail some of its connections to other popular methods and models.

- Chapter 5 introduces an APF version of *Tucker decomposition*—the second most widely used form of tensor decomposition. When applied to dyadic event data, this model infers dependence structure that yields classic network structures—i.e., *communities* of countries—separate from the structure it learns along other modes (e.g., action types). It further obtains superior out-of-sample predictive performance as compared to comparable models for network data. For posterior inference, I derive a Gibbs sampling algorithm—i.e., *compositional allocation*—that exploits the structure of the Tucker decomposition to significantly speed up the main bottleneck in APF inference (allocation). I also introduce a novel hierarchical shrinkage prior over the core tensor and introduce a novel auxiliary variable augmentation schemes for inference.

- Chapter 6 introduces an APF version of *linear dynamical systems* (LDS)—a widely used model for multivariate time series. I apply this model to several discrete multivariate time-series, including ones based on dyadic events, and demonstrate that this model outperforms the standard Gaussian LDS at both smoothing and forecasting, particularly when the time series are bursty. This model relies on a non-conjugate construction that chains gamma-distributed latent states together via their shape parameters. I introduce a novel augmentation scheme that yields an efficient *backward filtering forward sampling* for block updates of these latent states. This scheme relies on the iterative application of an augmentation scheme [Zhou and Carin, 2012] that I provide exposition on and offer insight into why its special structure facilitates tractable inference in a range of different models.

- Chapter 7 introduces another APF version of linear dynamical systems. This model introduces intermediate Poisson-distributed latent states that sit between the gamma-distributed latent states linked to the data. This change results in substantial improvements over the previous model, including superior smoothing and forecasting performance on tensors of dyadic event data and MCMC inference that can be easily extended to scalable stochastic alternatives. This model also introduces the general motif of *gamma–Poisson–gamma chains* that has wide applicability. I derive multiple algorithms for efficient closed-form inference in such chains. Under a certain setting of the hyperparameters, this chain permits true sparsity—i.e., values of exactly zero—in the gamma latent states. This setting motivates the derivation of a Rao-Blackwellized algorithm—i.e., one which integrates out the gamma states to re-sample the Poisson states. This algorithm relies on sampling from a novel probability distribution—hereby called the *size-biased confluent hypergeometric (SCH) distribution*—which I derive and describe.

- Chapter 8 imports the framework of differential privacy [Dwork et al., 2014] and introduces a method for locally private Bayesian inference in any APF model. In this setting, the true count data are sensitive and the model only conditions on a noisy version of them. This chapter introduces a novel relationship between the Bessel and Skellam distributions to derive a scheme for sampling the underlying sensitive counts as latent variables. This approach outperforms the naïve approach—which treats the noised data as if it were the true data—in two case studies on topic modeling and community detection. In some settings, the private method even outperforms the non-private method.

- Finally, in Chapter 9, I sketch promising future directions for expanding the framework of APF and suggest other applications.

CHAPTER 2

BACKGROUND

2.1 Computational social science

Computational social science (CSS) is a growing interdisciplinary research area dedicated to the development of data science methods for the study of *complex social processes* [Lazer et al., 2009, Alvarez, 2016]. The role of data science within CSS research can be broken down into three categories [Shmueli, 2010, Wallach, 2012]:

1. **Exploratory**—Surfacing patterns in data sets that help social scientists generate hypotheses and build theories about unobserved structures.

2. **Explanatory**—Testing existing hypotheses and theories.

3. **Predictive**—De-noising corrupted observations, imputing missing observables, or forecasting future ones.

In many cases, the scope of social science questions is so broad—e.g., *What are the precursors to civil war?*—that a directly predictive approach is untenable due to the scarcity of data on the object of interest. A central concept in social science research is thus *measurement*—i.e., the "[most] neglected topic in statistics" according to Gelman [2015] who provides the following definition:

> "[Measurement] is the idea of considering the connection between the data you gather and the underlying object of your study."

The concept of measurement applies both to case where the data is corrupted—i.e., what Gelman et al. [2013, Chapter 8] call "measurement error"—as well as to

the process of constructing a model whose parameters and latent variables have a close correspondence to unobserved but hypothesized objects in the real world. This latter process conforms to the theory of *representational measurement* which Hand [1996] describes in terms of a prescribed mapping between hypothesized objects and "numbers" (i.e., parameters or latent variables in a model); once this mapping is set:

> "Statistical operations can then be carried out on the numbers and the aim is that conclusions reached about relationships between the numbers will reflect corresponding relationships between the objects."

Competing theories of measurement exist, in particular *operational measurement* which is also described by Hand [1996]:

> "Operationalism defines scientific concepts in terms of the operations used to identify or measure them. It avoids assuming an underlying reality and so is fundamentally different from representationalism, which is based on a mapping from an assumed underlying reality. In operationalism, things start with the measurement procedure."

In terms of dyadic event data, the representationalist may hypothesize the presence of unobserved country community structure (i.e., the "object") that is confounding their regression analysis. They may then design a model whose latent variables (i.e., the "numbers") are supposed to measure those communities. Before using this measurement model to correct their regression however, they may perform *construct validation* "which assesses the extent to which the measure conforms with the theoretical predictions of relationships with other variables" [Hand, 1996]. The key point here is that the representationalist uses the model only if its inferred representation is consistent with their theory. The operationalist, on the other hand, does not constrain the model's inferred representation to conform to any theory or hypothesis of international relations. Instead, a model is "good" so long as performs well at some measurable operation—e.g., forecasting future dyadic events. The representationalism versus operationalism distinction is analogous to a rift in early artificial intelligence research—i.e., reductionism versus connectionism [Minsky, 1991].

Representational measurement is central to social science where direct data of the object of interest is scarce but theory about their connection is abundant. Computational social scientists carefully design exploratory (or "descriptive" [Shmueli, 2010]) methods so that the patterns they surface measure the latent structures whose presence is predicted by social scientific theory. In so doing, they encode their theory into a data analysis method and may then compare it to competing theories (encoded as methods) on predictive tasks [Hopkins and King, 2010, Schrodt, 2014, Wallach, 2012]. This results in an iterative process by which theories are encoded as methods, critiqued, and then updated. It is thus critical for computational social scientists to have an array of data analysis methods that are:

1. **Interpretable**—amenable to measurement tasks.

2. **Modular**—built from reusable or repurposable sub-units.

3. **Translatable**—described in a consistent and unified language that facilitates comparisons and surfaces connections between methods.

2.2 Bayesian latent variable modeling

Bayesian latent variable modeling (BLVM) is a framework for formalizing the relationships among observable data and hypothesized latent variables. It cleanly separates three phases of modeling—model building, model fitting, and model critiquing—which facilitates the iterative cycle of generating, testing, and refining substantive hypotheses about the data [Blei, 2014]. The approach to modeling wherein we restrict attention to a subset of models whose structure encode substantive hypotheses—i.e., representational measurement—and then choose among them by comparing their predictive performance is known as the *hypothetico-deductive approach* for which Gelman et al. [2013] motivate the BLVM framework.

2.2.1 Model building

A model formalizes the relationships among observables $\mathcal{D}$ and latent variables $\mathcal{Z}$ via a joint probability distribution $P(\mathcal{D}, \mathcal{Z} \mid \mathcal{H})$ that conditions on the set of prior assumptions and fixed hyperparameters $\mathcal{H}$. This joint distribution is often constructed via its factorization,

$$P(\mathcal{D}, \mathcal{Z} \mid \mathcal{H}) = \underbrace{P(\mathcal{D} \mid \mathcal{Z}, \mathcal{H})}_{\text{likelihood}} \underbrace{P(\mathcal{Z} \mid \mathcal{H})}_{\text{prior}}, \tag{2.1}$$

where the *likelihood* formalizes our hypothesis about how the latent variables influence the observables and the *prior* formalizes our hypothesis about the structure of the latent variables.

2.2.2 Model fitting

Given a model (i.e., joint distribution), the goal of Bayesian inference is to compute the posterior distribution which describes our uncertainty about the values of the latent variables conditioned on observables and assumptions. By Bayes' rule, it equals

$$\underbrace{P(\mathcal{Z} \mid \mathcal{D}, \mathcal{H})}_{\text{posterior}} = \frac{P(\mathcal{D}, \mathcal{Z} \mid \mathcal{H})}{\underbrace{P(\mathcal{D} \mid \mathcal{H})}_{\text{evidence}}}. \tag{2.2}$$

In most cases, the posterior is *intractable*—i.e., cannot be computed analytically—due to the evidence term in the denominator which is itself an intractable integral—i.e., $P(\mathcal{D} \mid \mathcal{H}) = \int \mathbf{d}\mathcal{Z}\, P(\mathcal{D}, \mathcal{Z} \mid \mathcal{H})$. In these cases, the posterior must be approximated; the field of Bayesian inference is concerned with the task of computing surrogate distributions that approximate the posterior in some way—i.e., $Q(\mathcal{Z}) \approx P(\mathcal{Z} \mid \mathcal{D}, \mathcal{H})$.

2.2.2.1 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods (e.g., Robert and Casella [2013], Gelman et al. [2013]) approximate the posterior using a set of samples $Q(\mathcal{Z}) \triangleq$

$\{\mathcal{Z}^{(s)}\}_{s=1}^{S}$ drawn from the exact posterior $\mathcal{Z}^{(s)} \sim P\left(\mathcal{Z} \,|\, \mathcal{D}, \mathcal{H}\right)$. One can interpret this set of samples as a histogram whose contours converge to the probability density (or mass) function of the exact posterior as the number of samples grows large $S \to \infty$. While the exact posterior may be intractable, we can generate samples from it by constructing a Markov chain that iteratively explores the state space of the latent variables. This chain is defined by its transition operator $T(\mathcal{Z}^{(s)} \to \mathcal{Z}^{(s+1)})$ which specifies the probability of transitioning from one state to the next—if it satisfies mild conditions, the chain is asymptotically guaranteed to converge to its stationary distribution, which is the exact posterior.

The most common example of a valid transition operator is one which re-samples each latent variable from its *complete conditional*—i.e.,

$$(z_n \,|\, -) \sim P(z_n \,|\, \mathcal{Z}_{\setminus n}, \mathcal{D}, \mathcal{H}), \tag{2.3}$$

where $\mathcal{Z}_{\setminus n}$ is the set of all latent variables except the n^{th} one (which we are re-sampling). MCMC based on this transition operator is called *Gibbs sampling*.

2.2.2.2 Mean-field variational inference

Variational methods [Jordan et al., 1999] turn posterior inference into an optimization problem. In this case, we approximate the full posterior with a surrogate distribution from a simpler parametric family $P\left(\mathcal{Z} \,|\, \mathcal{D}, \mathcal{H}\right) \approx Q\left(\mathcal{Z}; \boldsymbol{\gamma}\right)$ whose parameters $\boldsymbol{\gamma}$ we optimize to minimize some divergence between it and the exact posterior. The most common divergence to minimize is the KL divergence:

$$\boldsymbol{\gamma}^* = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\ \mathrm{KL}\Big(Q\left(\mathcal{Z}; \boldsymbol{\gamma}\right) \,||\, P\left(\mathcal{Z} \,|\, \mathcal{D}, \mathcal{H}\right)\Big). \tag{2.4}$$

We select the parametric family of surrogate distributions to facilitate optimization. In mean-field variational inference, this family factorizes over all latent variables:

$$Q\left(\mathcal{Z};\boldsymbol{\gamma}\right)=\prod_{n=1}^{N}Q\left(z_n;\boldsymbol{\gamma}_n\right). \tag{2.5}$$

A common algorithm for mean-field is *coordinate ascent variational inference* (CAVI) wherein the parameters of each factor $Q\left(z_n;\boldsymbol{\gamma}_n\right)$ are iteratively optimized to minimize the conditional objective. As shown by Bishop [2006, Chapter 10], the optimal factor $Q^*\left(z_n;\boldsymbol{\gamma}_n^*\right)$—i.e., the one that minimizes the KL divergence in Eq. (2.4) when the parameters to all other factors are fixed—is proportional to the *geometric expected value* of the complete conditional for z_n,

$$Q^*(z_n;\boldsymbol{\gamma}_n^*)\propto\mathbb{G}_{Q_{\setminus z_n}}\left[P(z_n\,|\,\mathcal{Z}_{\setminus n},\,\mathcal{D},\mathcal{H})\right], \tag{2.6}$$

where the geometric expected value—i.e., $\mathbb{G}[\cdot]=\exp\left(\mathbb{E}[\log\cdot]\right)$—is taken with respect to all factors of the surrogate distribution except $Q(z_n;\boldsymbol{\gamma}_n)$. Blei et al. [2017] provide an alternative derivation of this fact along with a statistical perspective on CAVI.

2.2.2.3 Auxiliary variables and augmentation schemes

Sometimes we *augment* our model (i.e., joint distribution) with a set of *auxiliary (latent) variables* $\mathcal{A}$ such that the original joint is equal to the augmented joint with the auxiliary variables marginalized out—i.e., $P\left(\mathcal{D},\mathcal{Z}\,|\,\mathcal{H}\right)=\int d\mathcal{A}\left(\mathcal{D},\mathcal{Z},\mathcal{A}\,|\,\mathcal{H}\right)$. Auxiliary variables can be incorporated into either Gibbs sampling or CAVI by treating them as any other latent variable.

In some cases, the complete conditional for a latent variable z_n in the original model—i.e., $P\left(z_n\,|\,\mathcal{Z}_{\setminus n},\mathcal{D},\mathcal{H}\right)$—is not available in closed form but its complete conditional in the augmented model—i.e., $P\left(z_n\,|\,\mathcal{Z}_{\setminus n},\mathcal{D},\mathcal{A},\mathcal{H}\right)$—is. This is typically the scenario that motivates augmentation schemes.

2.2.3 Model critiquing: Prediction as validation

Given some set of models that each encode competing substantive hypotheses, the hypothetico-deductivist may then choose among them by comparing their *posterior predictive* distributions on heldout data $\mathcal{D}'$—i.e.,

$$\underbrace{P(\mathcal{D}' \mid \mathcal{D}, \mathcal{H})}_{\text{posterior predictive}} = \int \mathbf{d}\mathcal{Z} \; \underbrace{P(\mathcal{D}' \mid \mathcal{Z}, \mathcal{H})}_{\text{heldout likelihood}} \underbrace{P(\mathcal{Z} \mid \mathcal{D}, \mathcal{H})}_{\text{posterior}}, \tag{2.7}$$

which is equal, in expectation, to the heldout likelihood averaged over samples of the latent variables drawn from the exact posterior (e.g., those returned by MCMC):

$$\approx \frac{1}{S} \sum_{s=1}^{S} P(\mathcal{D}' \mid \mathcal{Z}^{(s)}, \mathcal{H}). \tag{2.8}$$

It is also common to use samples from the surrogate distribution $\mathcal{Z}^{(s)} \sim Q(\mathcal{Z}; \boldsymbol{\gamma})$ when using variational inference—though this estimator of the posterior predictive likelihood may be biased if the variational family does not contain the exact posterior—or a point estimate—e.g., $\mathcal{Z}^* = \mathbb{E}_Q[\mathcal{Z}]$.

Gelman and Shalizi [2013] advocate for *posterior predictive checks* (PPCs) [Rubin, 1984, Gelman et al., 1996] as a means of model comparison. This framework provides a flexible way to compare "relevant" functions of models' posterior predictive distributions; it also permits (though does not require) this comparison without the use of a heldout data set. In this thesis, I will instead use prediction of heldout data as the primary means of model comparison, which can also be used to compare to non-Bayesian or non-probabilistic approaches.

2.3 Tensor decomposition

Tensor decomposition is a framework for representing an observed tensor—i.e., a multidimensional array—as a multilinear function of latent factors (that are fewer

than the number of data entries). Tensor decompositions represent compressions of the data that yield patterns and signatures of underlying dependence structure. This framework originated in psychometrics but was developed largely in signal-processing. Several thorough surveys and textbooks explain and catalog these methods—e.g., Kolda and Bader [2009], Cichocki et al. [2009], Sidiropoulos et al. [2017]. In this section, I introduce some terms and concepts that I refer to throughout this thesis.

2.3.1 Matrix decomposition

A special case of tensor decomposition is matrix decomposition (or matrix factorization) which assumes that an observed matrix $Y \in \mathbb{R}^{D \times V}$ is a noisy version of the product of two factor matrices $\Theta \in \mathbb{R}^{D \times K}$ and $\Phi \in \mathbb{R}^{V \times K}$ that each have a latent mode of *cardinality* K:

$$Y \approx \Theta \Phi^T. \tag{2.9}$$

This assumption can be equivalently made in terms of the individual entries:

$$y_{dv} \approx \sum_{k=1}^{K} \theta_{dk} \phi_{kv}. \tag{2.10}$$

Given an observed matrix, the goal is then to find "good" values of the factor matrices which are traditionally defined as those that minimize some divergence $D(Y, \Theta\Phi^T)$ plus some regularization constraint on the parameters $R(\Theta, \Phi^T)$.

2.3.2 Probabilistic matrix decomposition

Probabilistic matrix decomposition assumes some likelihood $P(Y \mid \Theta, \Phi)$ under which a sufficient statistic is decomposed—e.g., the expected value $\mathbb{E}[Y] = \Theta\Phi^T$. The goal is then to find parameter values that maximize this likelihood. These two

approaches coincide when minimizing the loss function yields equivalent solutions to maximizing the likelihood—i.e.,

$$\underset{\Theta,\Phi}{\operatorname{argmin}}\ D(Y,\, \Theta\Phi^T) = \underset{\Theta,\Phi}{\operatorname{argmax}}\ P(Y \mid \Theta, \Phi). \tag{2.11}$$

An example is the equivalence between minimizing Euclidean distance and maximizing the Gaussian likelihood—i.e.,

$$\underset{\Theta,\Phi}{\operatorname{argmin}} \sum_{d,v} \left(y_{dv} - \sum_{k=1}^{K} \theta_{dk}\phi_{kv} \right)^2 = \underset{\Theta,\Phi}{\operatorname{argmax}} \prod_{d,v} \mathcal{N}\left(y_{dv}; \sum_{k=1}^{K} \theta_{dk}\phi_{kv},\, \sigma \right). \tag{2.12}$$

Probabilistic matrix decomposition may further assume prior distributions over the factors and then find their values that maximize the joint distribution $P(Y, \Theta, \Phi)$—i.e., maximum a posteriori (MAP) estimation. Analogous equivalences exist between regularizers and priors. For instance, minimizing any divergence plus an ℓ_2-regularizer is equivalent to performing MAP estimation with Gaussian priors.

2.3.3 Canonical polyadic decomposition

Canonical polyadic (CP) decomposition, also known as CANDECOMP or PARAFAC [Harshman, 1970], is a generalization of matrix factorization to tensors. In this case, an observed M-mode tensor $\boldsymbol{Y} \in \mathbb{R}^{L_1 \times \cdots \times L_M}$ is assumed to be the noisy version of the tensor product of M factor matrices $\Theta^{(1)} \in \mathbb{R}^{K \times L_1}, \ldots, \Theta^{(M)} \in \mathbb{R}^{K \times L_M}$ that all share a K-dimensional latent mode. This is equivalent to the following assumption about a single element in the observed tensor:

$$y_{\delta_1 \ldots \delta_M} \approx \sum_{k=1}^{K} \prod_{m=1}^{M} \theta^{(m)}_{k\delta_m}. \tag{2.13}$$

The indices $\delta_1 \in \{1, \ldots, L_1\}, \ldots, \delta_M \in \{1, \ldots, L_M\}$ collectively form a *multi-index* that can be written as $\boldsymbol{\delta}$ and so that the data is written as $y_{\boldsymbol{\delta}} \equiv y_{\delta_1 \ldots \delta_M}$.

We may also introduce per-component weights λ_k, so that the decomposition is:

$$y_{\boldsymbol{\delta}} \approx \sum_{k=1}^{K} \lambda_k \prod_{m=1}^{M} \theta^{(m)}_{k\delta_m}. \tag{2.14}$$

Singular value decomposition is an example of a 2-mode (matrix) version of this.

2.3.4 Tucker decomposition

Tucker decomposition [Tucker, 1964] is a different generalization of matrix factorization to tensors. In this case, the M factor matrices are not constrained to share the same cardinality K—i.e., $\Theta^{(1)} \in \mathbb{R}^{K_1 \times L_1}, \ldots, \Theta^{(M)} \in \mathbb{R}^{K_M \times L_M}$. To combine the (potentially differently sized) factor matrices, we then introduce an M-mode core tensor $\boldsymbol{\Lambda} \in \mathbb{R}^{K_1 \times \cdots \times K_M}$. This makes the following assumption about a single element in the observed tensor,

$$y_{\boldsymbol{\delta}} \approx \sum_{\kappa_1=1}^{K_1} \cdots \sum_{\kappa_M=1}^{K_M} \lambda_{\boldsymbol{\kappa}} \prod_{m=1}^{M} \theta^{(m)}_{\kappa_m \delta_m}, \tag{2.15}$$

where here multi-index notation is additionally employed to describe the core tensor elements $\lambda_{\boldsymbol{\kappa}} \equiv \lambda_{\kappa_1 \ldots \kappa_M}$ for $\kappa_m \in \{1, \ldots, K_m\}$. The CP and Tucker decompositions coincide when $K_1 = \cdots = K_M$ and the off-diagonal entries of the core tensor are all zero. When the number of modes equals 2, tensor decomposition is matrix decomposition and Tucker decomposition corresponds to *bilinear* matrix decomposition [Nickel et al., 2015], also known as matrix trifactorization [Yoo and Choi, 2009].

2.3.5 Non-negative tensor decomposition

When the observed tensor is non-negative $\boldsymbol{Y} \in \mathbb{R}_{+}^{L_1 \times \cdots \times L_M}$ we may encode that prior knowledge into the model by constraining the model parameters to also be non-negative. Along with encoding a natural inductive bias, this constraint also allows us to interpret the model parameters in a number of different ways. See Cichocki et al.

[2007]'s textbook for a survey on these methods. Here, I highlight the multiple overlapping interpretations of model parameters that the non-negative constraint permits.

2.3.5.1 Embeddings and mixed-membership.

We may view the columns of the factor matrices—under both CP and Tucker decomposition—as representing *embeddings* of the various entities present in the data. The i^{th} column in the m^{th} factor matrix $\boldsymbol{\theta}^{(m)}_{:i} = (\theta^{(m)}_{1i}, \ldots, \theta^{(m)}_{K_m i})$ is vector-based representation of the i^{th} entity in the m^{th} mode. We should expect entities that exhibit similar behavior to be described by similar vectors. The embeddings interpretation is available under all forms of tensor decomposition. Non-negative tensor decomposition allows another interpretation of $\boldsymbol{\theta}^{(m)}_{i}$ as describing the mixed membership of entity i in each of K_m different components or clusters. In some applications, we may believe that entities truly do have mixed membership in different components. In other applications, we may believe that entities are members of only one, in which case the non-negative vector $\boldsymbol{\theta}^{(m)}_{:i}$ may be interpreted as our posterior belief about i's single membership—this is sometimes referred to as *soft clustering*.

2.3.5.2 Topics, components, and latent classes

We may also interpret the rows of the factor matrices in which case the interpretations under the CP and Tucker decompositions diverge. Under Tucker decomposition, the m^{th} factor matrix has its own latent cardinality K_m. The k^{th} row $\boldsymbol{\theta}^{(m)}_{k:} = (\theta^{(m)}_{k1}, \ldots, \theta^{(m)}_{kL_m})$ can be interpreted as describing how relevant each entity in the m^{th} mode is to the k^{th} component or cluster of that mode. The elements of the core tensor then describe the rate of interactions between the different components of the M modes. In CP decomposition, each factor matrix is constrained to have the same cardinality K. In this case, we interpret the k^{th} rows of all M factor matrices as collectively providing a signature for a *latent class* of the data. The latent class interpreta-

tion is also available under Tucker decomposition; in this case though, a latent class is composed of a unique combination of M components or clusters, i.e., $\boldsymbol{\kappa} = (\kappa_1, \dots \kappa_M)$.

2.4 Representing discrete data: tokens, types, tables, tensors

A discrete dataset consists of N *tokens* $\mathcal{D} \equiv \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_N\}$. Each token $\boldsymbol{e}_n$ records a single observed co-occurrence of M discrete (categorical or ordinal) variables $\boldsymbol{e}_n = (e_{n1}, \dots, e_{nM})$ where the m^{th} variable takes values in a discrete set $e_{nm} \in [L_m]$. The notation $[L_m] \equiv \{1, \dots, L_m\}$ refers to a discrete set with L_m *levels*.

There are $\prod_{m=1}^{M} L_m$ possible values—i.e., *types*—that a token can take. Each type $\boldsymbol{\delta} \equiv (\delta_1, \dots \delta_M)$ is a unique value in the Cartesian product of the M discrete sets $\boldsymbol{\delta} \in [L_1] \times \cdots \times [L_M]$. This thesis will employ *multi-index notation*—i.e., $\boldsymbol{\delta}$—when discussing tensor decomposition at a general level.

The *type count* $y_{\boldsymbol{\delta}}$ represents the total number of tokens taking type $\boldsymbol{\delta}$:

$$y_{\boldsymbol{\delta}} = \sum_{n=1}^{N} \mathbb{1}\left[\boldsymbol{e}_n = \boldsymbol{\delta}\right], \tag{2.16}$$

where the indicator function $\mathbb{1}[x]$ equals 1 if the predicate x is true and 0 otherwise. The counts for all types $\boldsymbol{\delta}$ can be represented in a *contingency table* (also known as a *pivot table*) which simply enumerates all possible types in an unstructured manner. Alternatively, the type counts can be organized into an M-mode count *tensor* $\boldsymbol{Y} \in \mathbb{N}_0^{L_1 \times \cdots \times L_M}$. The notation $\mathbb{N}_0$ refers to the set of possible counts—i.e., the natural numbers and zero, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ where the natural numbers are $\mathbb{N} = \{1, 2, 3, \dots\}$.

Example 1: Text corpora. The "bag-of-words" representation of a text corpus represents it as a collection of 2-dimensional tokens. A token $\boldsymbol{e}_n = (d, v)$ taking type (d, v) records a single instance of word type v occurring in document d. If there are D documents and V unique vocabulary items, the type counts can be organized into a 2-mode count tensor (or, matrix) $Y \in \mathbb{N}_0^{D \times V}$.

Example 2: Multinetworks. A multinetwork is a network with multiple edge types. A network with V actors and A possible edge types can be represented as a collection of 3-dimensional tokens where a token $\boldsymbol{e}_n = (i, j, a)$ taking type (i, j, a) records a single instance of actor i being connected to actor j with edge type a. The type counts can be organized into a 3-mode count tensor $\boldsymbol{Y} \in \mathbb{N}_0^{V \times V \times A}$.

CHAPTER 3

ALLOCATIVE POISSON FACTORIZATION

"God made the natural numbers; all else is man's work." –Leopold Kronecker, 1893

"Ah, ah, ah!" –Count Von Count, Sesame Street

Allocative Poisson factorization defines a class of models wherein the factorization of a count tensor *coincides with* an allocation of event tokens. Inference in APF models thus scales with the number of event tokens while tensor decomposition, in general, scales linearly (or worse) with the size of the tensor—i.e., the number of possible event *types*. APF models yield two equivalent mathematical representations: a multilinear decomposition of the Poisson rate parameter (factorization) and the multinomial thinning of the total observed count (allocation).

3.1 Basic definition of APF: factorization

APF is a subset of *Poisson factorization* [Canny, 2004, Titsias, 2008, Cemgil, 2009, Zhou and Carin, 2012, Gopalan and Blei, 2013, Paisley et al., 2014], a broad class of models for learning latent structure from discrete data. Poisson factorization assumes that each type count is a Poisson random variable,

$$y_{\boldsymbol{\delta}} \sim \text{Pois}\left(\mu_{\boldsymbol{\delta}}\right), \tag{3.1}$$

where $\mu_{\boldsymbol{\delta}} \geq 0$ is the rate parameter where if $\mu_{\boldsymbol{\delta}} = 0$ then $y_{\boldsymbol{\delta}} = 0$ almost surely. The rate $\mu_{\boldsymbol{\delta}}$ is then defined to be a function of shared model parameters—i.e., $\mu_{\boldsymbol{\delta}} = f(\mathcal{Z}_{\boldsymbol{\delta}})$—where $\mathcal{Z}_{\boldsymbol{\delta}}$ are the set of model parameters indexed by some subset of the multi-index $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_M)$ and $f(\cdot)$ is an arbitrary link function whose range is non-negative.

Definition 3.1: Poisson distribution

A Poisson random variable $y \sim \text{Pois}(\mu)$ is a count $y \in \mathbb{N}_0$ whose distribution is defined by non-negative rate parameter $\mu \geq 0$ and PMF:

$$\text{Pois}(y;\, \mu) = \frac{\mu^y}{y!} e^{-\mu}. \tag{3.2}$$

The expected value and variance are equal:

$$\mathbb{E}[y; \mu] = \mathbb{V}[y; \mu] = \mu. \tag{3.3}$$

The standard definition of the Poisson assumes a positive rate parameter $\mu > 0$. However, if we adopt the common convention that $0^0 = 1$, the distribution is defined for $\mu = 0$; in this case, the count is zero $y = 0$ almost surely:

$$\text{Pois}(y = 0;\, \mu = 0) = 1. \tag{3.4}$$

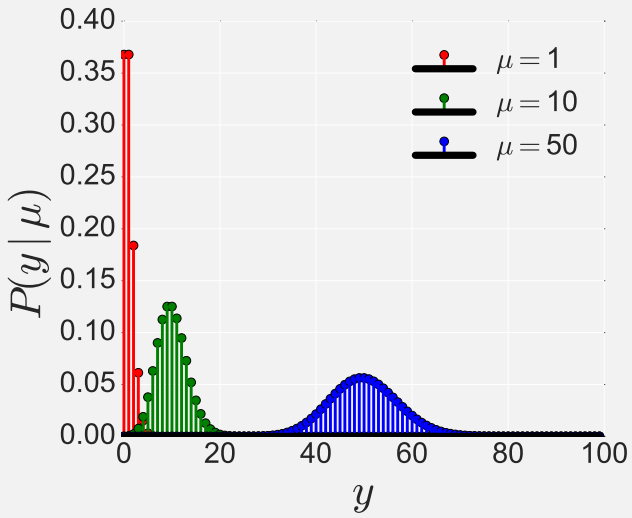


Figure 3.1: Probability mass function of the Poisson distribution for three different values of the rate μ which defines both the expected value and variance.

APF is the subset of Poisson factorization models for which the rate parameter is defined to be a *multilinear function* of shared model parameters. In this case, the rate may be defined as a sum over *latent classes*—$\mu_{\boldsymbol{\delta}} = \sum_{\boldsymbol{\kappa} \in \mathcal{K}} \mu_{\boldsymbol{\delta}\boldsymbol{\kappa}}$—where $\boldsymbol{\kappa}$ is the multi-index for a latent class, $\mathcal{K}$ is the set of all latent classes, and $\mu_{\boldsymbol{\delta}\boldsymbol{\kappa}}$ is some function of shared model parameters that are indexed by both the *type* $\boldsymbol{\delta}$ and *latent class* $\boldsymbol{\kappa}$. The canonical form for allocative Poisson factorization is thus

$$y_{\boldsymbol{\delta}} \sim \text{Pois}\left(\sum_{\boldsymbol{\kappa} \in \mathcal{K}} \mu_{\boldsymbol{\delta}\boldsymbol{\kappa}}\right). \tag{3.5}$$

Example 1: Poisson matrix factorization is typically written as

$$y_{ij} \sim \text{Pois}\left(\sum_{k=1}^{K} \theta_{ik}\phi_{kj}\right). \tag{3.6}$$

We may re-express this model as in Eq. (3.5) by defining the type as $\boldsymbol{\delta} = (i, j)$, the latent class as $\boldsymbol{\kappa} = k$ and then the rates $\mu_{\boldsymbol{\delta}\boldsymbol{\kappa}} = \theta_{ik}\phi_{kj}$.

Example 2: Poisson bilinear matrix factorization for community detection in networks (e.g., [Karrer and Newman, 2011, Zhou, 2015]) is typically described as

$$y_{ij} \sim \text{Pois}\left(\sum_{k_1=1}^{K} \theta_{ik_1} \sum_{k_2=1}^{K} \pi_{k_1k_2}\theta_{jk_2}\right). \tag{3.7}$$

We may re-express this model as in Eq. (3.5) by defining the type as $\boldsymbol{\delta} = (i, j)$, each latent class as a unique pair of communities $\boldsymbol{\kappa} = (k_1, k_2)$ and then the rates $\mu_{\boldsymbol{\delta}\boldsymbol{\kappa}} = \theta_{ik_1}\pi_{k_1k_2}\theta_{jk_2}$.

3.2 Latent source representation: allocation

The defining assumption of APF, given in Eq. (3.5), can be equivalently made by assuming that the observed type count is the sum of latent sub-counts—or *latent sources* [Cemgil, 2009]—each of which is an independent Poisson random variable:

$$y_{\boldsymbol{\delta}} \triangleq \sum_{\boldsymbol{\kappa} \in \mathcal{K}} y_{\boldsymbol{\delta}\boldsymbol{\kappa}}, \tag{3.8}$$
$$y_{\boldsymbol{\delta}\boldsymbol{\kappa}} \sim \text{Pois}\left(\mu_{\boldsymbol{\delta}\boldsymbol{\kappa}}\right) \text{ for } \kappa \in \mathcal{K}.$$

This alternative representation is available due to *Poisson additivity*.

Definition 3.2: Poisson additivity

The Poisson distribution is closed under convolution. Define the sum of K count random variables $y_{\cdot} \triangleq \sum_{k=1}^{K} y_k$. If each y_k is an independent Poisson random variable $y_k \sim \text{Pois}\left(\mu_k\right)$ then their sum is marginally Poisson distributed $y_{\cdot} \sim \text{Pois}\left(\mu_{\cdot}\right)$ with rate parameter equal to the sum of their rates $\mu_{\cdot} \triangleq \sum_{k=1}^{K} \mu_k$.

A recurrent motif in APF is *latent source aggregation* wherein arbitrary sums of latent sources are marginally Poisson-distributed (again due to Poisson additivity). A common example of this is to consider the sum of all latent sources whose observed index at the m^{th} mode takes a particular value—e.g., $\delta_m = d$—and whose latent index at the m^{th} mode takes a particular value—e.g., $\kappa_m = k$:

$$y_{dk}^{(m)} \triangleq \sum_{\boldsymbol{\delta} \in \Delta} \sum_{\boldsymbol{\kappa} \in \mathcal{K}} \mathbb{1}\left[\delta_m = d\right] \mathbb{1}\left[\kappa_m = k\right] \, y_{\boldsymbol{\delta}\boldsymbol{\kappa}}. \tag{3.9}$$

Marginally—i.e., not conditioned on its sub-counts—$y_{dk}^{(m)}$ is Poisson distributed:

$$y_{dk}^{(m)} \sim \text{Pois}\left(\sum_{\boldsymbol{\delta} \in \Delta} \sum_{\boldsymbol{\kappa} \in \mathcal{K}} \mathbb{1}\left[\delta_m = d\right] \mathbb{1}\left[\kappa_m = k\right] \, \mu_{\boldsymbol{\delta}\boldsymbol{\kappa}}\right). \tag{3.10}$$

We will return to this property later in this chapter when deriving conditionally-conjugate parameter updates. A common theme is that latent parameters are rendered independent of the data (and often conditionally independent of each other) when conditioned on aggregations of latent sources. Thus, posterior inference in APF—both MCMC and variational—hinges on inferring the latent sources. The complete conditional of the latent sources—i.e., $\boldsymbol{y}_{\boldsymbol{\delta}} \equiv (y_{\boldsymbol{\delta\kappa}})_{\boldsymbol{\kappa}\in\mathcal{K}}$—is multinomial:

$$\left((y_{\boldsymbol{\delta\kappa}})_{\boldsymbol{\kappa}\in\mathcal{K}} \mid - \right) \sim \text{Multinom}\left(y_{\boldsymbol{\delta}},\, \left(\frac{\mu_{\boldsymbol{\delta\kappa}}}{\sum_{\boldsymbol{\kappa}'\in\mathcal{K}} \mu_{\boldsymbol{\delta\kappa}'}} \right)_{\boldsymbol{\kappa}\in\mathcal{K}} \right). \tag{3.11}$$

The multinomial distribution is a multivariate count distribution [Johnson et al., 1997, Chapter 35] that conditions on the total count $y_{\boldsymbol{\delta}}$ and the relative—i.e., normalized—per-class rates. We may also use a generalization of the multinomial (Definition 3.3) that permits the following compact description with unnormalized rates $\boldsymbol{\mu}_{\boldsymbol{\delta}} \equiv (\mu_{\boldsymbol{\delta\kappa}})_{\boldsymbol{\kappa}\in\mathcal{K}}$:

$$\left(\boldsymbol{y}_{\boldsymbol{\delta}} \mid - \right) \sim \text{Multinom}\left(y_{\boldsymbol{\delta}},\, \boldsymbol{\mu}_{\boldsymbol{\delta}} \right). \tag{3.12}$$

This complete conditional follows from a well-known connection between the Poisson and multinomial distributions—i.e., *Poisson–multinomial thinning* (see Definition 3.4)—which was first referenced, in a footnote, by Fisher [1922] and later given more explicitly by Steel [1953].

A key property of the multinomial distribution is that the sub-counts are almost surely 0 if $y_{\boldsymbol{\delta}} = 0$. Standard algorithms for simulating multinomial random variables have $\mathcal{O}(|\mathcal{K}|)$ time complexity where $|\mathcal{K}|$ is the number of latent classes [Devroye, 2006, Chapter 11]. Thus, inference in APF models scales with $\mathcal{O}(|\Delta_+|\,|\mathcal{K}|)$ where $\Delta_+ \triangleq \{\boldsymbol{\delta} \in \Delta : y_{\boldsymbol{\delta}} > 0\}$ is the set of types associated with non-zero counts. It is common for high-dimensional discrete data to contain exponentially more types than non-zeros [Kunihama and Dunson, 2013]. This therefore represents a significant computational speedup over analogous non-APF decompositions that scale with $\mathcal{O}(|\Delta|\,|\mathcal{K}|)$.

Definition 3.3: Multinomial distribution (unnormalized rates)

A multinomial random variable $\boldsymbol{y} \sim \text{Multinom}\,(y_\cdot, \boldsymbol{\mu})$ is a K-dimensional count vector $\boldsymbol{y} \equiv (y_1, \ldots, y_K) \in \mathbb{N}_0^K$. The distribution is defined by the total count $y_\cdot \in \mathbb{N}_0$, a K-dimensional non-negative rate vector $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_K) \in \mathbb{R}_{0+}^K$ whose sum $\mu_\cdot \triangleq \sum_{k=1}^K \mu_k$ is positive $\mu_\cdot > 0$, and probability mass function:

$$\text{Multinom}\,(\boldsymbol{y};\, y_\cdot, \boldsymbol{\mu}) = \mathbb{1}\left[y_\cdot = \sum_{k=1}^K y_k\right] \frac{y_\cdot!}{\prod_{k=1}^K y_k!} \prod_{k=1}^K \left(\frac{\mu_k}{\mu_\cdot}\right)^{y_k}. \tag{3.13}$$

The standard definition assumes the rates are normalized—i.e., $\mu_\cdot = 1$. This definition is preferable for two reasons. First, it is faithful to algorithms for multinomial sampling that do not require a normalized rate vector. Second, the PMF under this representation can be easily manipulated to illuminate the multinomial's connection to the Poisson distribution (see Definition 3.4), which facilitates derivation of inference schemes in APF models:

$$\begin{aligned} \text{Multinom}\,(\boldsymbol{y};\, y_\cdot, \boldsymbol{\mu}) &= \mathbb{1}\left[y_\cdot = \sum_{k=1}^K y_k\right] \frac{y_\cdot!}{\mu_\cdot^{y_\cdot}} \prod_{k=1}^K \frac{\mu_k^{y_k}}{y_k!} \\ &= \mathbb{1}\left[y_\cdot = \sum_{k=1}^K y_k\right] \frac{\prod_{k=1}^K \text{Pois}\,(y_k;\, \mu_k)}{\text{Pois}\,(y_\cdot,\, \mu_\cdot)}. \end{aligned} \tag{3.14}$$

The expected value of a single element y_k can be defined as

$$\mathbb{E}\left[y_k \mid y_\cdot, \boldsymbol{\mu}\right] = \mu_k \frac{y_\cdot}{\mu_\cdot}. \tag{3.15}$$

An important property of the multinomial distribution that holds for any $\boldsymbol{\mu} \in \mathbb{R}_+^K$ is that all elements are zero $y_k = 0$ almost surely if $y_\cdot = 0$:

$$\text{Multinom}\,(\boldsymbol{y} = \mathbf{0};\, y_\cdot = 0, \boldsymbol{\mu}) = 1. \tag{3.16}$$

Definition 3.4: Poisson–multinomial thinning

Consider K independent Poisson random variables $y_k \sim \text{Pois}(\mu_k)$ and define their sum $y_\cdot \triangleq \sum_{k=1}^K y_k$ and the sum of their rates $\mu_\cdot \triangleq \sum_{k=1}^K \mu_k$. Then the joint distribution of the counts when further conditioned on their sum—i.e., $P(\boldsymbol{y} \mid y_\cdot, \boldsymbol{\mu})$—is the multinomial given in Definition 3.3. Combining this with Poisson additivity (Definition 3.2) further implies that the joint distribution of the counts and their sum can be factorized (by definition) as

$$P(\boldsymbol{y}, y_\cdot \mid \boldsymbol{\mu}) = P(y_\cdot \mid \boldsymbol{y})\, P(\boldsymbol{y} \mid \boldsymbol{\mu}) = \mathbb{1}\Big[y_\cdot = \sum_{k=1}^K y_k\Big] \prod_{k=1}^K \text{Pois}(y_k;\, \mu_k)\,, \tag{3.17}$$

or alternatively into the marginal distribution of their sum (which is Poisson) and the distribution conditioned on their sum (which is multinomial):

$$P(\boldsymbol{y}, y_\cdot \mid \boldsymbol{\mu}) = P(y_\cdot \mid \boldsymbol{\mu})\, P(\boldsymbol{y} \mid y_\cdot, \boldsymbol{\mu}) = \text{Pois}(y_\cdot;\, \mu_\cdot)\ \text{Multinom}(\boldsymbol{y};\, y_\cdot, \boldsymbol{\mu})\,. \tag{3.18}$$

Note the delta function in Eq. (3.17) is present in the multinomial PMF.

3.2.1 Thinning in CAVI

In the context of MCMC, the multinomial in Eq. (3.12) defines the complete conditional from which the latent sources are sampled. In the context of coordinate ascent variational inference (CAVI), the optimal surrogate distribution for the latent sources can be derived by plugging in the multinomial PMF into Eq. (2.6) [Bishop, 2006]; it's here where the multinomial PMF with unnormalized rates is helpful:

$$Q^*(\boldsymbol{y}_{\boldsymbol{\delta}}) \propto \mathbb{G}_{Q_{\setminus \boldsymbol{y}_{\boldsymbol{\delta}}}}\left[\text{Multinom}(\boldsymbol{y}_{\boldsymbol{\delta}};\, y_{\boldsymbol{\delta}},\, \boldsymbol{\mu}_{\boldsymbol{\delta}})\right] \tag{3.19}$$

$$\propto \mathbb{G}_{Q_{\setminus \boldsymbol{y}_{\boldsymbol{\delta}}}}\left[\mathbb{1}\Big[y_{\boldsymbol{\delta}} = \sum_{\boldsymbol{\kappa} \in \mathcal{K}} y_{\boldsymbol{\delta}\boldsymbol{\kappa}}\Big]\, \frac{y_{\boldsymbol{\delta}}!}{\mu_{\boldsymbol{\delta}}^{y_{\boldsymbol{\delta}}}} \prod_{\boldsymbol{\kappa} \in \mathcal{K}} \frac{\mu_{\boldsymbol{\delta}\boldsymbol{\kappa}}^{y_{\boldsymbol{\delta}\boldsymbol{\kappa}}}}{y_{\boldsymbol{\delta}\boldsymbol{\kappa}}!}\right]. \tag{3.20}$$

We can drop the terms μ^{y_δ} and $y_\delta!$ since neither depend on the sources $y_{\delta\kappa}$. The Dirac function can also be brought outside the expectation since it only depends on the observed data y_δ and the latent sources $y_{\delta\kappa}$, neither of which are governed by $Q_{\setminus y_\delta}$:

$$\propto \mathbb{1}\left[y_\delta = \sum_{\kappa\in\mathcal{K}} y_{\delta\kappa}\right] \mathbb{G}_{Q_{\setminus y_\delta}}\left[\prod_{\kappa\in\mathcal{K}} \frac{\mu_{\delta\kappa}^{y_{\delta\kappa}}}{y_{\delta\kappa}!}\right]. \tag{3.21}$$

The geometric expectation pushes into the product and we may equivalently substitute Q for $Q_{\setminus y_\delta}$ to simply notation:

$$\propto \mathbb{1}\left[y_\delta = \sum_{\kappa\in\mathcal{K}} y_{\delta\kappa}\right] \prod_{\kappa\in\mathcal{K}} \frac{\mathbb{G}_Q\left[\mu_{\delta\kappa}\right]^{y_{\delta\kappa}}}{y_{\delta\kappa}!}. \tag{3.22}$$

This is the unnormalized form of a multinomial distribution,

$$\propto \text{Multinom}\left(\boldsymbol{y}_\delta;\, y_\delta,\, \mathbb{G}_Q\left[\boldsymbol{\mu}_\delta\right]\right), \tag{3.23}$$

where the unnormalized rates of this multinomial $\mathbb{G}_Q\left[\boldsymbol{\mu}_\delta\right] \equiv \left(\mathbb{G}_Q\left[\mu_{\delta\kappa}\right]\right)_{\kappa\in\mathcal{K}}$ consist of the geometric expectation under the surrogate distribution of $\mu_{\delta\kappa}$. When $\mu_{\delta\kappa}$ is a product of latent variables whose surrogate distributions are gamma or Dirichlet, this expectation is available in closed form. The messages[1] this factor passes to other factors are of the form $\mathbb{E}_Q\left[y_{\delta\kappa}\right] = y_\delta \frac{\mathbb{G}_Q[\mu_{\delta\kappa}]}{\sum_{\kappa'\in\mathcal{K}} \mathbb{G}_Q[\mu_{\delta\kappa'}]}$ which needn't be computed for $y_\delta = 0$.

3.2.2 Thinning via allocation

We may equivalently represent the multinomial thinning step in Eq. (3.12) in terms of categorical indicators $\boldsymbol{z}_n \in \mathcal{K}$ that *allocate* the N tokens in the data set to latent classes. Recall that each observed token $\boldsymbol{e}_n \in \Delta$ is a multi-index.

[1] CAVI can be understood as a "message-passing algorithm". For more on this interpretation see Winn and Bishop [2005] and Blei et al. [2017].

$$\left(\boldsymbol{z}_n \mid -\right) \sim \text{Categorical}\left(\boldsymbol{\mu}_{\boldsymbol{e}_n}\right) \qquad \text{for } n = 1, \ldots, N \tag{3.24}$$

$$y_{\boldsymbol{\delta}\boldsymbol{\kappa}} \triangleq \sum_{n=1}^{N} \mathbb{1}\left[\boldsymbol{e}_n = \boldsymbol{\delta}\right] \mathbb{1}\left[\boldsymbol{z}_n = \boldsymbol{\kappa}\right] \qquad \text{for } \boldsymbol{\delta} \in \Delta$$

The equivalence between multinomial thinning and categorical allocation connects many existing models to Poisson factorization—any model that specifies a per-token categorical likelihood but assumes that tokens are conditionally independent (or *exchangeable*) admits a representation as a multinomial or conditionally Poisson model.

Definition 3.5: Categorical distribution (unnormalized rates)

A categorical random variable $\boldsymbol{z} \sim \text{Cat}(\boldsymbol{\mu})$ takes values $\boldsymbol{z} \in \mathcal{K}$ in a discrete set with K elements $\mathcal{K} = \{\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K\}$. The distribution is defined by a K-dimensional non-negative rate vector $\boldsymbol{\mu}$ (as in Definition 3.3) and PMF:

$$\text{Cat}(\boldsymbol{z}; \boldsymbol{\mu}) = \prod_{k=1}^{K} \left(\frac{\mu_k}{\mu_\cdot}\right)^{\mathbb{1}[\boldsymbol{z}=\boldsymbol{\kappa}_k]}. \tag{3.25}$$

A special case of the categorical distribution[a] is the one for which the set of outcomes is the set of integers that index them—i.e., $\mathcal{K} = \{1, \ldots, K\}$ and $\boldsymbol{\kappa}_k = k$. This definition applies to cases where outcomes are multi-indices—e.g., all possible undirected pairs of V countries: $\mathcal{K} = \{(1 \leftrightarrow 2), (1 \leftrightarrow 3), \ldots, (V \leftrightarrow V\text{–}1)\}$.

[a] The categorical (or "discrete") distribution is often conflated with a special case of the multinomial distribution for $y_\cdot = 1$. This special case has been called the "multinoulli distribution" by Murphy [2012, Section 2.3.2] who also conflates it with the categorical distribution. This conflation is the source of abuses in notation and software implementation errors—e.g., the sum of multinoulli random variables is a multinomial random variable (analogous to the Bernoulli–binomial relationship) however a sum of categorical variables is not. It is important to distinguish categorical and multinoulli random variables: the former takes values in $\mathcal{K}$ while the latter takes values in the set of K-dimensional "one-hot" vectors that are 1 at one level and 0 everywhere else.

Definition 3.6: Categorical allocation

A multinomial random variable $\boldsymbol{y} \sim \text{Multinom}\,(y_{\cdot}, \boldsymbol{\mu})$ can be represented in terms of auxiliary categorical random variables as,

$$\boldsymbol{e}_n \sim \text{Cat}(\boldsymbol{\mu}) \qquad \text{for } n = 1, \ldots, y_{\cdot} \tag{3.26}$$

$$y_{\boldsymbol{\kappa}} = \sum_{n=1}^{y_{\cdot}} \mathbb{1}\left[\boldsymbol{e}_n = \boldsymbol{\kappa}\right] \qquad \text{for } \boldsymbol{\kappa} \in \mathcal{K} \tag{3.27}$$

where the last step bin-counts the categorical variables into the entries of $\boldsymbol{y}$.

3.3 Missing data: masking, imputing, and marginalizing

Missingness manifests itself in multiple ways in discrete data. Zero-valued counts frequently conflate true non-occurrence with unobserved occurrence—such data sets are sometimes referred to as "presence-only data" [Pearce and Boyce, 2006]. Statistical literaure on contingency table analysis further distinguishes "sampling zeros"—i.e., zero-valued counts of possible but unobserved events—with "structural zeros"—i.e., counts of structurally impossible events that are zero-valued by definition [Agresti, 2003, Chapter 2.1.4, pp. 392].

3.3.1 Masking

To explicitly model structural zeros, we introduce a binary mask $\boldsymbol{B} \in \{0, 1\}^{L_1 \times \cdots \times L_M}$ which is the same size as the observed count tensor $\boldsymbol{Y} \in \mathbb{N}_0^{L_1 \times \cdots \times L_M}$. A zero-valued entry $y_{\boldsymbol{\delta}} = 0$ is a structural zero if $b_{\boldsymbol{\delta}} = 0$ and a sampling zero if $b_{\boldsymbol{\delta}} = 1$. The likelihood of allocative Poisson factorization using such a mask is

$$y_{\boldsymbol{\delta}} \sim \text{Pois}\left(b_{\boldsymbol{\delta}} \sum_{\boldsymbol{\kappa} \in \mathcal{K}} \mu_{\boldsymbol{\delta}\boldsymbol{\kappa}}\right), \tag{3.28}$$

where we appeal to the convention that a Poisson random variable is 0 almost surely if its rate is 0 (see Definition 3.1). The definition for Poisson factorization given previously in this chapter is consistent with this one—it simply assumes that there are no structural zeros (i.e., the mask is one everywhere).

We may treat the mask as observed or latent, depending on the application. For instance, in recommender systems, the "exposure" of a user to a particular item is sometimes explicitly modeled as a binary latent variable which deconflates structural and sampling zeros [Liang et al., 2016]. More generally, masked Poisson factorization corresponds to a "misrecorded" or "zero-inflated" Poisson model [Johnson et al., 2005, Chapter 4.10.3] when the mask is latent.

In this thesis, I only consider cases where the mask is observed. It is common in machine learning to create a heldout data set for the purpose of testing a model's out-of-sample prediction performance. In this case, the mask denotes which entries of the observed tensor have been held out.

3.3.2 Imputing

The simplest way to accommodate masked data during posterior inference in APF models is via imputation [Rubin, 1996] whereby the masked entries are treated as latent variables whose complete conditional is $\text{Pois}\left(y_{\boldsymbol{\delta}}; \mu_{\boldsymbol{\delta}}\right)$. In the context of MCMC, we would re-sample each heldout entry—i.e., $\left(y_{\boldsymbol{\delta}} \mid -\right) \sim \text{Pois}\left(\mu_{\boldsymbol{\delta}}\right)$. We may equivalently treat the latent sources $y_{\boldsymbol{\delta}\boldsymbol{\kappa}}$ as conditionally independent Poisson random variables—i.e., $\left(y_{\boldsymbol{\delta}\boldsymbol{\kappa}} \mid -\right) \sim \text{Pois}\left(\mu_{\boldsymbol{\delta}\boldsymbol{\kappa}}\right)$—since they are no longer conditioned on their sum. This latter representation is useful for incorporating missing data into mean-field variational inference, wherein each $y_{\boldsymbol{\delta}\boldsymbol{\kappa}}$ is given its own surrogate distribution,

$$Q^*(y_{\boldsymbol{\delta\kappa}}) \propto \mathbb{G}_{Q_{\setminus y_{\boldsymbol{\delta\kappa}}}} \left[\text{Pois}\left(y_{\boldsymbol{\delta\kappa}};\, \mu_{\boldsymbol{\delta\kappa}}\right)\right] \tag{3.29}$$

$$\propto \mathbb{G}_{Q_{\setminus y_{\boldsymbol{\delta\kappa}}}} \left[\frac{\mu_{\boldsymbol{\delta\kappa}}^{y_{\boldsymbol{\delta\kappa}}}}{y_{\boldsymbol{\delta\kappa}}!}\right] \tag{3.30}$$

$$\propto \frac{\left(\mathbb{G}_Q\left[\mu_{\boldsymbol{\delta\kappa}}\right]\right)^{y_{\boldsymbol{\delta\kappa}}}}{y_{\boldsymbol{\delta\kappa}}!} \tag{3.31}$$

$$\propto \text{Pois}\left(y_{\boldsymbol{\delta\kappa}};\, \mathbb{G}_Q\left[\mu_{\boldsymbol{\delta\kappa}}\right]\right), \tag{3.32}$$

where the geometric expectation $\mathbb{G}_Q\left[\mu_{\boldsymbol{\delta\kappa}}\right]$ is the same as those that appear in the optimal surrogate distribution for the latent sources in Eq. (3.23). The messages this factor passes to other factors take the form $\mathbb{E}_Q\left[y_{\boldsymbol{\delta\kappa}}\right] = \mathbb{G}_Q\left[\mu_{\boldsymbol{\delta\kappa}}\right]$ which collectively constitute an imputation of $y_{\boldsymbol{\delta}}$—i.e., $\mathbb{E}_Q\left[y_{\boldsymbol{\delta}}\right] = \sum_{\boldsymbol{\kappa}\in\mathcal{K}} \mathbb{G}_Q\left[\mu_{\boldsymbol{\delta\kappa}}\right]$. I illustrate this point in greater detail later in the chapter.

In both MCMC and VI, imputation scales with $\mathcal{O}\left(|\mathcal{K}| \cdot |1-\boldsymbol{B}|_1\right)$ where $|1-\boldsymbol{B}|_1$ is the number of heldout entries and $|\mathcal{K}|$ is the number of latent classes.

3.3.3 Marginalizing

Imputation augments the set of latent variables with the heldout data entries. If the number of heldout entries is large, imputation may significantly hamper the mixing or convergence rates of posterior inference. In that case, marginalizing out heldout entries may be preferable, if possible. However there is no free lunch: the computational cost of marginalization is greater than that of imputation by a factor of M—i.e., $\mathcal{O}\left(M \cdot |\mathcal{K}| \cdot |1-\boldsymbol{B}|_1\right)$—where M is the number of modes. The additional computational cost introduced by marginalization is somewhat obfuscated. Where it appears is during latent source aggregation. I illustrate this point in greater detail later in the chapter. Intuitively though, we can view imputation as updating the heldout entries once per interation while in marginalization, computing the expectation over the heldout entries during the parameter updates to the M modes is morally the same as re-sampling the heldout entries M different times per iteration.

3.4 Conjugate priors: gamma and Dirichlet

In this section, I give details about the two main choices of non-negative prior distribution for the latent parameters in APF: gamma (Definition 7.1) and Dirichlet Definition 3.9). Both have a conjugate relationship to a form of the likelihood—the gamma is conjugate to the Poisson (Definition 3.8), while the Dirichlet is conjugate to the multinomial (Definition 3.10). An important and widely overlooked consideration in choosing between these two is that marginalization of missing data is incompatible with multinomial–Dirichlet conjugacy—i.e., models based on Dirichlet-distributed parameters must impute missing data to exploit conjugacy for efficient parameter updates. I will illustrate this point in more detail after introducing some definitions and detailing how the two distributions are closely related (Definitions 3.13 and 3.14).

To illustrate their trade-offs, I will construct an APF model and derive its complete conditionals under gamma and Dirichlet priors. The model is based on the CP decomposition of a 3-mode count tensor $\boldsymbol{Y} \in \mathbb{N}_0^{V \times V \times A}$ of dyadic events counts, each of which $y_{i \xrightarrow{a} j}$ represents the number of times country i took action a towards country j. This model is a simpler version of the model in Chapter 4, which includes a 4^{th} mode (time step). We will not allow self-actions $i \xrightarrow{a} i$ and use the mask to denote them as structural zeros—i.e., $b_{i \xrightarrow{a} j} = 0$ for all $i = j$ and 1 otherwise. This model's likelihood is

$$y_{i \xrightarrow{a} j} \sim \text{Pois}\left(b_{i \xrightarrow{a} j} \sum_{k=1}^{K} \theta_{ik}^{(1)} \, \theta_{jk}^{(2)} \, \theta_{ak}^{(3)} \right), \tag{3.33}$$

where the non-negative latent parameters for the sender, receiver, and action modes are $\theta_{ik}^{(1)}$, $\theta_{jk}^{(2)}$, and $\theta_{ak}^{(3)}$ respectively.

3.4.1 Gamma priors

Assume the parameters of the first mode are gamma distributed:

$$\theta_{ik}^{(1)} \sim \Gamma(\alpha_0, \beta_0). \tag{3.34}$$

Definition 3.7: Gamma distribution

A gamma random variable $\theta \sim \Gamma(\alpha, \beta)$ takes values in the positive reals $\theta > 0$. Its distribution is defined by shape $\alpha > 0$ and rate $\beta > 0$ parameters and PDF:

$$\Gamma(\theta; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\theta\beta}. \tag{3.35}$$

The (arithmetic) expected value and variance are

$$\mathbb{E}[\theta] = \alpha\,\beta^{-1} \text{ and } \mathbb{V}[\theta] = \alpha\,\beta^{-2}. \tag{3.36}$$

The geometric expected value is defined in terms of the digamma function $\Psi(\cdot)$:

$$\mathbb{G}[\theta] = \exp\Big(\Psi(\alpha) - \ln\beta\Big). \tag{3.37}$$

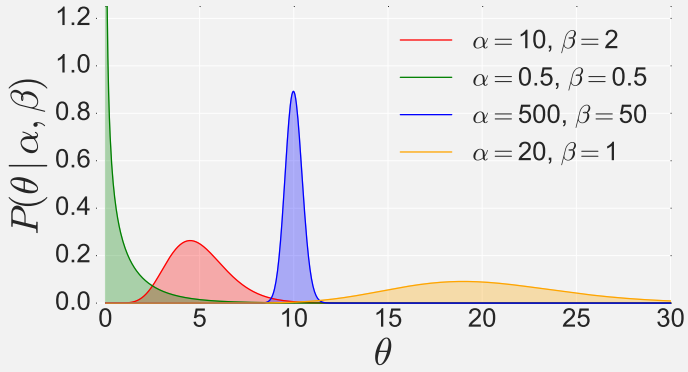


Figure 3.2: PDF of the gamma distribution for four combinations of the shape α and rate β.

Definition 3.8: Gamma–Poisson conjugacy

Consider a gamma random variable $\theta \sim \Gamma(\alpha, \beta)$ as in Definition 7.1 and N independent Poisson random variables $y_n \sim \text{Pois}(\theta\,\zeta_n)$ for $n = 1, \ldots, N$ each of whose rate is a product of θ and some constant $\zeta_n \geq 0$. Then the *posterior* or *inverse distribution* of θ—i.e., its distribution conditioned on the Poisson counts $\boldsymbol{y} \equiv (y_1, \ldots, y_N)$ and constants $\boldsymbol{\zeta} \equiv (\zeta_1, \ldots, \zeta_N)$ follows a gamma distribution that depends only on the sums $y_\cdot$ and $\zeta_\cdot$ of the counts and constants:

$$P(\theta \mid \boldsymbol{y}, \boldsymbol{\zeta}, \alpha, \beta) = P(\theta \mid y_\cdot, \zeta_\cdot, \alpha, \beta) = \Gamma(\theta; \alpha + y_\cdot,\ \beta + \zeta_\cdot) \tag{3.38}$$

This prior does not have a directly conjugate relationship to the likelihood in Eq. (3.33). However, this prior does have a conjugate relationship to the (marginal) distribution of the following *latent source aggregation*,

$$y_{ik}^{(1)} \triangleq \sum_{j=1}^{V} \sum_{a=1}^{A} y_{i \xrightarrow{a} j\, k}, \tag{3.39}$$

which represents the total number of dyadic events involving sender i that are allocated to component k. Marginally—i.e., not conditioned on $y_{i \xrightarrow{a} j}$—it is a Poisson random variable:

$$y_{ik}^{(1)} \sim \text{Pois}\left(\sum_{j=1}^{V} \sum_{a=1}^{A} b_{i \xrightarrow{a} j} \theta_{ik}^{(1)}\, \theta_{jk}^{(2)}\, \theta_{ak}^{(3)} \right). \tag{3.40}$$

We may pull $\theta_{ik}^{(1)}$ outside the sum and rewrite its rate as

$$y_{ik}^{(1)} \sim \text{Pois}\left(\theta_{ik}^{(1)} \zeta_{ik}^{(1)} \right), \tag{3.41}$$

where we've defined $\zeta_{ik}^{(1)}$ as the multilinear combination of the other parameters:

$$\zeta_{ik}^{(1)} \triangleq \sum_{j=1}^{V} \sum_{a=1}^{A} b_{i \xrightarrow{a} j}\, \theta_{jk}^{(2)}\, \theta_{ak}^{(3)}. \tag{3.42}$$

Eq. (3.41) describes the marginal distribution for the count of all events whose rates depend on $\theta_{ik}^{(1)}$—it is to this Poisson distribution that $\theta_{ik}^{(1)}$'s gamma prior is conjugate. We may apply Definition 3.8 to obtain $\theta_{ik}^{(1)}$'s complete conditional:

$$\left(\theta_{ik}^{(1)} \mid -\right) \sim \Gamma\left(\alpha_0 + y_{ik}^{(1)},\, \beta_0 + \zeta_{ik}^{(1)}\right). \tag{3.43}$$

In the context of MCMC, this complete conditional describes how to resample $\theta_{ik}^{(1)}$. For CAVI, the optimal surrogate distribution for $\theta_{ik}^{(1)}$ is then proportional to

$$Q^*\left(\theta_{ik}^{(1)}\right) \propto \mathbb{G}_{Q_{\setminus \theta_{ik}^{(1)}}}\left[\Gamma\left(\theta_{ik}^{(1)}; \alpha_0 + y_{ik}^{(1)},\, \beta_0 + \zeta_{ik}^{(1)}\right)\right]. \tag{3.44}$$

Plugging in only the terms in the gamma PDF that involve $\theta_{ik}^{(1)}$:

$$\propto \mathbb{G}_{Q_{\setminus \theta_{ik}^{(1)}}}\left[(\theta_{ik}^{(1)})^{\alpha_0 + y_{ik}^{(1)} - 1}\, e^{-(\beta_0 + \zeta_{ik}^{(1)})\theta_{ik}^{(1)}}\right]. \tag{3.45}$$

The geometric expectation pushes into the exponents and becomes an arithmetic expectation. We may also equivalently replace $Q_{\setminus \theta_{ik}^{(1)}}$ with Q:

$$\propto (\theta_{ik}^{(1)})^{\alpha_0 + \mathbb{E}_Q\left[y_{ik}^{(1)}\right] - 1}\, e^{-\left(\beta_0 + \mathbb{E}_Q\left[\zeta_{ik}^{(1)}\right]\right)\theta_{ik}^{(1)}}. \tag{3.46}$$

This is proportional to the following gamma distribution,

$$\propto \Gamma\left(\theta_{ik}^{(1)};\, \alpha_0 + \mathbb{E}_Q\left[y_{ik}^{(1)}\right],\, \beta_0 + \mathbb{E}_Q\left[\zeta_{ik}^{(1)}\right]\right), \tag{3.47}$$

where $\mathbb{E}_Q\left[y_{ik}^{(1)}\right]$ is the message from the multinomial factor described (in its general form) in Eq. (3.23) and the expectation in the rate parameter is equal to

$$\mathbb{E}_Q\left[\zeta_{ik}^{(1)}\right] = \sum_{j=1}^{V}\sum_{a=1}^{A} b_{i \xrightarrow{a} j}\, \mathbb{E}_Q\left[\theta_{jk}^{(2)}\right]\, \mathbb{E}_Q\left[\theta_{ak}^{(3)}\right], \tag{3.48}$$

since the mask is fixed and $\theta_{jk}^{(2)}, \theta_{ak}^{(3)}$ are independent under the surrogate distribution.

Note that if there are no structural zeros—i.e., if $b_{i \xrightarrow{a} j} = 1$ everywhere—the expression for the constant $\zeta_{ik}^{(1)}$ can be rewritten by pushing the sums in:

$$\zeta_{ik}^{(1)} = \left(\sum_{j=1}^{V} \theta_{jk}^{(2)} \right) \left(\sum_{a=1}^{A} \theta_{ak}^{(3)} \right). \tag{3.49}$$

In this case, we can compute $\zeta_{ik}^{(1)}$ by summing each parameter vector and multiplying the resultant scalars in $\mathcal{O}(V+A)$ time. However, if there are structural zeros, the presence of the mask term $b_{i \xrightarrow{a} j}$ prevents the sums from pushing in fully; we may rewrite Eq. (3.42) as

$$\zeta_{ik}^{(1)} = \sum_{j=1}^{V} \theta_{jk}^{(2)} \left(\sum_{a=1}^{A} b_{i \xrightarrow{a} j} \, \theta_{ak}^{(3)} \right), \tag{3.50}$$

which naïvely suggests that we compute $\zeta_{ik}^{(1)}$ as a bilinear combination of the $V \times A$ slice of mask indexed by the i^{th} sender $B_{i \xrightarrow{:} :}$ with the A- and V-length parameter vectors in $\mathcal{O}(VA)$ time. However, we can exploit the binary nature of the mask to rewrite this as

$$\zeta_{ik}^{(1)} = \left(\sum_{j=1}^{V} \theta_{jk}^{(2)} \right) \left(\sum_{a=1}^{A} \theta_{ak}^{(3)} \right) - \sum_{j=1}^{V} \sum_{a=1}^{A} \left(1 - b_{i \xrightarrow{a} j} \right) \theta_{jk}^{(2)} \, \theta_{ak}^{(3)}, \tag{3.51}$$

where the first term is the $\mathcal{O}(V+A)$ expression for the fully observed $\zeta_{ik}^{(1)}$ in Eq. (3.49) and the second is a correction term that subtracts off the bilinear terms for the entries of the mask that are 0. The second term represents the *additional* computational cost of marginalization which is $\mathcal{O}(|1 - B_{i \xrightarrow{:} :}|_1)$ where $|1 - B_{i \xrightarrow{:} :}|_1$ is the number of zeros in the mask slice; this number is at most VA, if *all* entries in the slice are structural zeros, but often a small fraction of that, in practice. The additional cost of marginalization in computing the analogous constants for all senders i and components k is then $\mathcal{O}\left(K \cdot \sum_{i=1}^{V} |1 - B_{i \xrightarrow{:} :}|_1 \right) = \mathcal{O}(K \cdot |1 - \boldsymbol{B}|_1)$.

More generally, if the parameters for all modes in an APF model are gamma-distributed, the additional cost of marginalization is $\mathcal{O}\left(M \cdot |\mathcal{K}| \cdot |1 - \boldsymbol{B}|_1 \right)$.

3.4.2 Dirichlet priors

Now assume that each column of the third parameter matrix—e.g., the k^{th} column $\tilde{\boldsymbol{\theta}}^{(3)}_{:k} \equiv (\tilde{\theta}^{(3)}_{1k}, \ldots, \tilde{\theta}^{(3)}_{Ak})$—is a Dirichlet random variable. Note I've added a tilde to the parameter to denote that the vector is normalized—i.e., $\sum_{a=1}^{A} \tilde{\theta}^{(3)}_{ak} = 1$.

$$\tilde{\boldsymbol{\theta}}^{(3)}_{:k} \sim \text{Dir}(\boldsymbol{\alpha}_0). \tag{3.52}$$

Definition 3.9: Dirichlet distribution

An N-dimensional Dirichlet random variable $\tilde{\boldsymbol{\theta}} \sim \text{Dir}\,(\alpha_1, \ldots, \alpha_N)$ takes values in the $(N-1)$-simplex—i.e., $\tilde{\boldsymbol{\theta}} \equiv (\tilde{\theta}_N, \ldots, \tilde{\theta}_N)$, $\tilde{\theta}_n \geq 0$, and $\sum_{n=1}^{N} \tilde{\theta}_n = 1$. Its distribution is defined by N shape parameters $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_N)$, $\alpha_n > 0$, and PDF:

$$\text{Dir}(\tilde{\boldsymbol{\theta}}; \boldsymbol{\alpha}) = \frac{\Gamma\,(\alpha_\cdot)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \tilde{\theta}_k{}^{\alpha_k - 1}. \tag{3.53}$$

The arithmetic and geometric expected values are:

$$\mathbb{E}\left[\tilde{\theta}_n \,|\, \boldsymbol{\alpha}\right] = \frac{\alpha_n}{\alpha_\cdot}, \tag{3.54}$$

$$\mathbb{G}\left[\tilde{\theta}_n \,|\, \boldsymbol{\alpha}\right] = \exp\Big(\psi(\alpha_n) - \psi(\alpha_\cdot)\Big). \tag{3.55}$$

where $\psi(\cdot)$ is the digamma function.

Definition 3.10: Dirichlet–multinomial conjugacy

Consider an N-dimensional Dirichlet random variable $\tilde{\boldsymbol{\theta}} \sim \text{Dir}(\boldsymbol{\alpha})$ as in Definition 7.1 and a multinomial random variable $\boldsymbol{y} \sim \text{Multinom}\left(y_\cdot, \tilde{\boldsymbol{\theta}}\right)$. Then the inverse distribution of $\tilde{\boldsymbol{\theta}}$—i.e., its distribution conditioned $\boldsymbol{y}$—is Dirichlet:

$$P(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{y}, \alpha) = \text{Dir}\left(\tilde{\boldsymbol{\theta}}; \alpha_1 + y_1, \ \ldots, \ \alpha_N + y_N\right). \tag{3.56}$$

Definition 3.11: Dirichlet–Poisson conjugacy

Consider an N-dimensional Dirichlet random variable $\tilde{\boldsymbol{\theta}} \sim \text{Dir}(\boldsymbol{\alpha})$ as in Definition 7.1 and N independent Poisson random variables $y_n \overset{\text{ind}}{\sim} \text{Pois}\left(\tilde{\theta}_n\, \zeta\right)$ where the constant $\zeta \geq 0$ is shared across all N. Then the inverse distribution of $\tilde{\boldsymbol{\theta}}$ is Dirichlet and independent of ζ:

$$P(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{y}, \zeta, \alpha) = P(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{y}, \alpha) = \text{Dir}\left(\tilde{\boldsymbol{\theta}}; \alpha_1 + y_1, \,\ldots,\, \alpha_N + y_N\right) \tag{3.57}$$

Proof: By Poisson–multinomial thinning, the joint distribution of the N Poisson random variables along with their sum $y_{\cdot}$ is equal to,

$$\mathbb{1}\left[y_{\cdot} = \sum_{n=1}^{N} y_n\right] \prod_{n=1}^{N} \text{Pois}\left(y_n;\, \tilde{\theta}_n\, \zeta\right) = \text{Pois}\left(y_{\cdot}; \zeta\right) \text{Multinom}\left(\boldsymbol{y};\, y_{\cdot}, \tilde{\boldsymbol{\theta}}\right) \tag{3.58}$$

where the Poisson rate for $y_{\cdot}$—i.e., $\sum_{n=1}^{N} \tilde{\theta}_n\, \zeta = \zeta(\sum_{n=1}^{N} \tilde{\theta}_n) = \zeta$—does not include contain any $\tilde{\theta}_n$ terms because they sum to 1 and the (normalized) multinomial rates—e.g., $\frac{\tilde{\theta}_n\, \zeta}{\sum_{n'=1}^{N} \tilde{\theta}_{n'}\, \zeta} = \frac{\tilde{\theta}_n}{\sum_{n'=1}^{N} \tilde{\theta}_{n'}} = \tilde{\theta}_n$—do not include ζ because it is constant across all N. Since the only dependence on $\tilde{\boldsymbol{\theta}}$ is through the multinomial term, the prior is conjugate via Dirichlet–multinomial conjugacy.

Note that this conjugate relationship holds *only if* ζ is shared by all N Poisson random variables—if the Poisson random variables have heterogeneous constants $\zeta_n \neq \zeta_{n'}$, the Dirichlet is not conjugate to their joint distribution.

This prior does not have a directly conjugate relationship to the Poisson likelihood in Eq. (3.33). However, consider the following latent source aggregation,

$$y^{(3)}_{ak} \triangleq \sum_{i=1}^{V} \sum_{j=1}^{V} y_{i \xrightarrow{a} j}, \tag{3.59}$$

which counts all events involving action type a allocated to component k. Its marginal distribution is Poisson,

$$y_{ak}^{(3)} \sim \text{Pois}\left(\sum_{i=1}^{V}\sum_{j=1}^{V} b_{i \xrightarrow{a} j}\, \theta_{ik}^{(1)}\, \theta_{jk}^{(2)}\, \tilde{\theta}_{ak}^{(3)}\right), \tag{3.60}$$

which may be rewritten as,

$$y_{ak}^{(3)} \sim \text{Pois}\left(\tilde{\theta}_{ak}^{(3)}\, \zeta_{ak}^{(3)}\right) \tag{3.61}$$

$$\zeta_{ak}^{(3)} \triangleq \sum_{i=1}^{V}\sum_{j=1}^{V} b_{i \xrightarrow{a} j}\, \theta_{ik}^{(1)}\, \theta_{jk}^{(2)}. \tag{3.62}$$

Consider all A of such latent sources aggregations $\boldsymbol{y}_{:k}^{(3)} \equiv (y_{1k}^{(3)}, \ldots, y_{Ak}^{(3)})$ for the k^{th} component. By Definition 3.11, the Dirichlet prior over $\tilde{\boldsymbol{\theta}}_{:k}$ is not conjugate to their joint distribution, due to the fact that their constants are heterogeneous $\zeta_{ak}^{(3)} \neq \zeta_{a'k}^{(3)}$. However, that heterogeneity is solely due to the presence of the mask. In the case where there are no heldout entries—i.e., $b_{i \xrightarrow{a} j} = 1$ everywhere—the form of the constants simplify to a form that is homogeneous across action types:

$$\zeta_{ak}^{(3)} = \zeta_{a'k}^{(3)} = \left(\sum_{i=1}^{V} \theta_{ik}^{(1)}\right)\left(\sum_{j=1}^{V} \theta_{jk}^{(2)}\right) \tag{3.63}$$

Thus, it is *only in the fully observed or imputed model* that we may apply Dirichlet–Poisson conjugacy to obtain the following complete conditional for $\tilde{\boldsymbol{\theta}}_{:k}$:

$$\left(\tilde{\boldsymbol{\theta}}_{:k} \mid -\right) \sim \text{Dir}\left(\alpha_{01} + y_{1k}^{(3)},\, \ldots,\, \alpha_{0A} + y_{Ak}^{(3)}\right) \tag{3.64}$$

3.4.3 Connection between the gamma and Dirichlet distributions

The gamma and Dirichlet distributions have a deep relationship. Understanding their relationship allows us to draw connections between seemingly disparate models—i.e., Dirichlet–categorical models and gamma–Poisson models. Moreover, their relationship is based on unique conditional independence properties, described by Lukacs [1955]'s theorem, that can be exploited for efficient and elegant posterior inference; I describe this in Definitions 3.13 and 3.14 (which depend on Definition 3.12).

Definition 3.12: Gamma additivity

The gamma distribution *with fixed rate* is closed under convolution. Consider K gamma random variables $\theta_k \sim \Gamma(\alpha_k, \beta)$ for $k = 1, \ldots, K$ with possibly heterogeneous shape parameters α_k but shared rate parameter β. Then the sum $\theta_.$ is marginally gamma distributed and depends only on the sum $\alpha_.$ of the shapes:

$$P(\theta_.; \boldsymbol{\alpha}, \beta) = P(\theta_.; \alpha_., \beta) = \Gamma\left(\theta_. \,;\, \alpha_., \, \beta\right). \tag{3.65}$$

The gamma distribution is *not* closed under convolution for varying rate β_n.

Definition 3.13: Gamma proportions are Dirichlet distributed

Consider K independent gamma random variables $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_K)$ with possibly heterogeneous shapes α_k but shared rate β—i.e., $\theta_k \overset{\text{ind}}{\sim} \Gamma(\alpha_k, \beta)$. Define their sum $\theta_.$ and *proportions* vector $\tilde{\boldsymbol{\theta}} \triangleq \boldsymbol{\theta}/\theta_.$ so that $\sum_{k=1}^{K} \tilde{\theta}_k = 1$. The proportions vector is marginally Dirichlet-distributed and independent of the rate β :

$$P(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\alpha}, \beta) = P(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\alpha}) = \text{Dir}\left(\boldsymbol{\alpha}\right). \tag{3.66}$$

A related fact is that the proportions are also independent of the sum $\theta_.$. This is only true of gamma random variables (Lukacs [1955], see Definition 3.14).

Definition 3.14: Sum–proportion independence of gammas

Consider K non-negative random variables $\theta_1, \ldots, \theta_K$ drawn from an arbitrary distribution and define their sum $\theta_\cdot$ and proportion vector $\tilde{\boldsymbol{\theta}}$ (as in Definition 3.13). Lukacs [1955] showed that the sum and proportion vector are independent random variables *if and only if* the original random variables are independently gamma-distributed with shared rate $\theta_k \overset{\text{ind}}{\sim} \Gamma(\alpha_k, \beta)$. Thus, the joint distribution of such gamma random variables along with their sum and normalized vector—i.e., $P(\boldsymbol{\theta}, \theta_\cdot, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\alpha}, \beta)$—can be factorized (by definition) as

$$P(\boldsymbol{\theta}, \theta_\cdot, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\alpha}, \beta) = P(\theta_\cdot, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta})\, P(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}, \beta) \tag{3.67}$$

$$= \mathbb{1}\left[\tilde{\boldsymbol{\theta}} = \frac{\boldsymbol{\theta}}{\theta_\cdot}\right] \mathbb{1}\left[\theta_\cdot = \sum_{k=1}^{K} \theta_k\right] \left[\prod_{k=1}^{K} \Gamma\left(\theta_k; \alpha_k, \beta\right)\right], \tag{3.68}$$

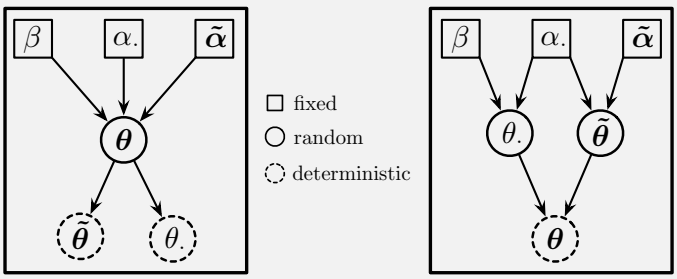
or alternatively by drawing the sum $\theta_\cdot$ (by gamma additivity) and the normalized vector $\tilde{\boldsymbol{\theta}}$ (as a Dirichlet random variable) which then determine $\boldsymbol{\theta}$:

$$P(\boldsymbol{\theta}, \theta_\cdot, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\alpha}, \beta) = P(\boldsymbol{\theta} \,|\, \theta_\cdot, \tilde{\boldsymbol{\theta}})\, P(\theta_\cdot, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\alpha}, \beta) \tag{3.69}$$

$$= P(\boldsymbol{\theta} \,|\, \theta_\cdot, \tilde{\boldsymbol{\theta}})\, P(\theta_\cdot \,|\, \boldsymbol{\alpha}, \beta)\, P(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\alpha}) \tag{3.70}$$

$$= \mathbb{1}\left[\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}\, \theta_\cdot\right] \Gamma\left(\theta_\cdot; \alpha_\cdot, \beta\right) \operatorname{Dir}(\tilde{\boldsymbol{\theta}}; \boldsymbol{\alpha}). \tag{3.71}$$

It is only for these specific parametric forms given here that the two graphical models below encode the same joint distribution.



3.5 Negative binomial magic

Definition 3.15: Negative binomial distribution

A negative binomial random variable $y \sim \mathrm{NB}\,(r,\, p)$ is a count $y \in \mathbb{N}_0$. Its distribution is defined by a *shape* $r \geq 0$ and *probability* $p \in (0, 1)$ parameter and PMF:

$$P(y \,|\, r, p) = \frac{\Gamma(r+y)}{y!\,\Gamma(r)}(1-p)^r\, p^y. \tag{3.72}$$

Its expected value, variance, and variance-to-mean ratio (VMR) are:

$$\mathbb{E}[y \,|\, r, p] = \frac{r\,p}{(1-p)}, \tag{3.73}$$

$$\mathbb{V}[y \,|\, r, p] = \frac{r\,p}{(1-p)^2}, \tag{3.74}$$

$$\mathrm{VMR}[y \,|\, r, p] = (1-p)^{-1}. \tag{3.75}$$

Since its VMR is always greater than 1, the negative binomial is *overdispersed*.

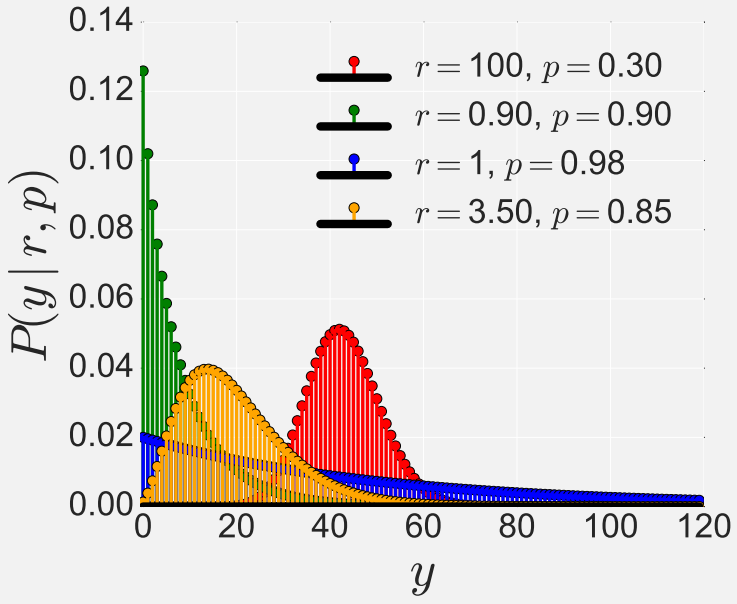


Figure 3.3: Probability mass function of the negative binomial distribution for four combinations of shape r and probability p parameter.

Recent auxiliary variable schemes based on augmentations of the negative binomial distribution—i.e., *augment-and-conquer* [Zhou and Carin, 2012]—have made posterior inference tractable in many new APF models, particularly those based on hierarchies of gamma and Dirichlet priors. These techniques are based on a handful of facts which I present here for later chapters to refer back to. I also attempt to provide intuition as to why these facts are special and how we might exploit them.

Definition 3.16: Negative binomial as a gamma–Poisson mixture

Consider a Poisson random variable $y \sim \text{Pois}(\theta\,\zeta)$ whose rate is a product of constant $\zeta > 0$ and $\theta \sim \Gamma(\alpha, \beta)$ which is a gamma random variable with shape α and *rate* β. The marginal distribution of y is negative binomial:

$$P(y \mid \zeta, \alpha, \beta) = \int \mathbf{d}\theta P(y \mid \theta, \zeta)\, P(\theta \mid \alpha, \beta) \tag{3.76}$$

$$= \int \mathbf{d}\theta\, \text{Pois}(y;\, \theta\,\zeta)\; \Gamma(\theta;\, \alpha, \beta) \tag{3.77}$$

$$= \text{NB}\left(y;\, \alpha,\, \frac{\zeta}{\zeta + \beta}\right). \tag{3.78}$$

Definition 3.17: Negative binomial as a compound Poisson

The negative binomial is a *compound Poisson distribution* [Adelson, 1966]—i.e., it can be constructed as the sum of a Poisson-distributed number of i.i.d. random variables. If $y \sim \text{SL}(\ell, p)$ is a sum–logarithmic random variable, as in Definition 3.19, that depends on fixed $p \in (0, 1)$, random $\ell \sim \text{Pois}\left(r\, \ln\left(\frac{1}{1-p}\right)\right)$ then its marginal distribution—i.e., not conditioned on ℓ—is negative binomial:

$$P(y \mid r, p) = \sum_{\ell=0}^{\infty} P(y \mid \ell, p)\, P(\ell \mid r, p) \tag{3.79}$$

$$= \sum_{\ell=0}^{\infty} \text{SL}(y;\, \ell, p)\, \text{Pois}\left(\ell;\, r\, \ln\left(\tfrac{1}{1-p}\right)\right) \tag{3.80}$$

$$= \text{NB}(y; r, p). \tag{3.81}$$

Definition 3.18: Chinese restaurant table (CRT) distribution

A Chinese restaurant table random variable [Zhou and Carin, 2015] $\ell \sim \mathrm{CRT}(y, r)$ is a bounded count $\ell \in \{0, \ldots, y\}$. Its distribution is defined by a *shape* $r \geq 0$ and count-valued *population* $y \in \mathbb{N}_0$ parameter and PMF,

$$P(\ell \,|\, r, y) = \frac{\Gamma(r)}{\Gamma(y+r)} |s(y, \ell)| \, r^{\ell}, \tag{3.82}$$

where $|\cdot|$ denotes absolute value and $s(\cdot, \cdot)$ are the Stirling numbers of the first kind [Johnson et al., 2005, pp. 12]. A CRT random variable can be generated as a sum $\ell \triangleq \sum_{i=1}^{y} b_i$ of independent Bernoulli random variables, each drawn:

$$b_i \overset{\text{ind}}{\sim} \mathrm{Bern}\left(\frac{r}{r+i-1}\right) \text{ for } i = 1, \ldots, y. \tag{3.83}$$

Thus if $y \geq 1$ then $\ell \geq 1$, almost surely, and if $y = 0$ then $\ell = 0$, almost surely.

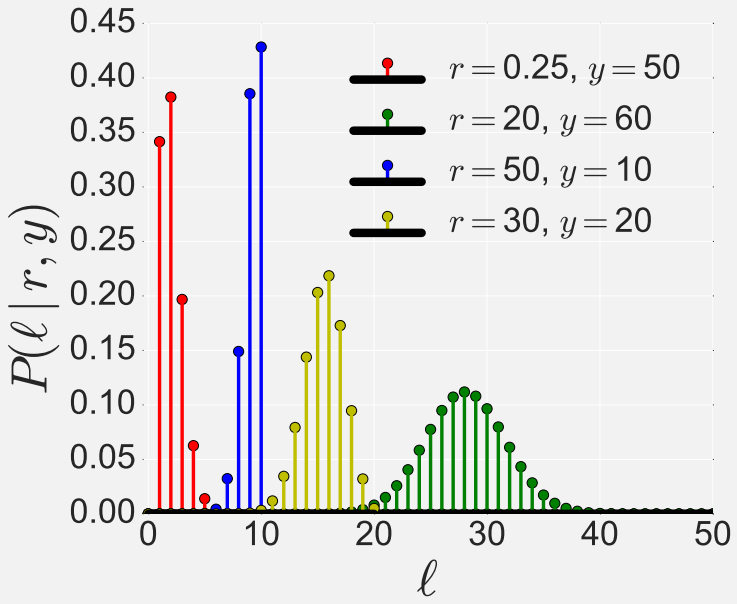


Figure 3.4: Probability mass function of the Chinese restaurant table distribution for four combinations of shape r and population y parameter.

Definition 3.19: Sum-Logarithmic (SumLog) distribution

A sum-logarithmic random variable [Zhou and Carin, 2015] $y \sim \mathrm{SL}(\ell, p)$ is a count $y \in \mathbb{N}_0$. Its distribution is defined by a count-valued *scale* parameter $\ell \in \mathbb{N}_0$, a *probability* parameter $p \in (0, 1)$, and PMF,

$$P(y \,|\, \ell, p) = \frac{p^y \, \ell! \, |s(y, \ell)|}{y! \left(\ln(\frac{1}{1-p})\right)^\ell}. \tag{3.84}$$

A sum-logarithmic random variable can be generated as a sum $y \triangleq \sum_{i=1}^{\ell} u_i$ of i.i.d. logarithmic random variables [Johnson et al., 2005, pp. 302], each drawn:

$$u_i \overset{\text{iid}}{\sim} \mathrm{Log}(p) \;\text{ for } i = 1, \ldots, \ell. \tag{3.85}$$

If $\ell \geq 1$ then $y \geq 1$, almost surely, and if $\ell = 0$ then $y = 0$. The logarithmic distribution—i.e., $\mathrm{Log}(y; p) = \frac{p^y}{y \ln(\frac{1}{1-p})}$ for $y \geq 1$—is a special case for $\ell = 1$.

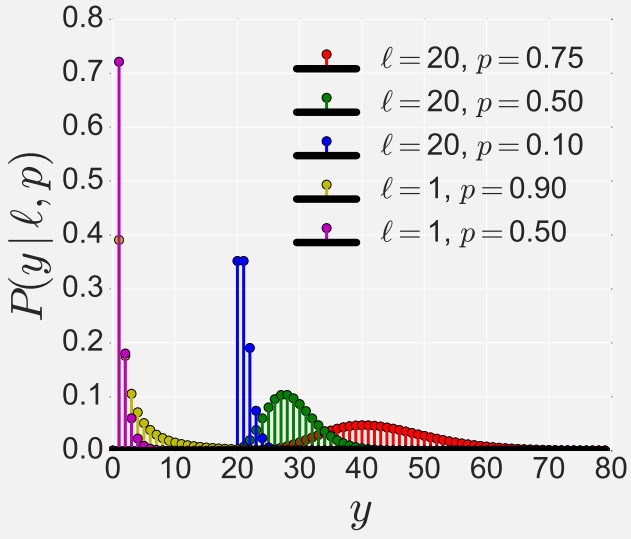


Figure 3.5: Probability mass function of the sum–logarithmic distribution for five combinations of scale ℓ and probability p parameter.

Definition 3.20: Magic bivariate count distribution

Consider a negative binomial under its compound Poisson construction (Definition 3.17) which is a sum–logarithmic random variable $y \sim \mathrm{SL}(\ell, p)$ whose scale parameter $\ell \sim \mathrm{Pois}\left(r \ln\left(\frac{1}{1-p}\right)\right)$ is Poisson-distributed. By definition, and by conditional probability, the bivariate distribution can be factorized in two ways:

$$P(y, \ell \,|\, r, p) = P(y \,|\, \ell, p)\, P(\ell \,|\, r, p) = \mathrm{SL}(y;\, \ell, p)\, \mathrm{Pois}\left(\ell;\, r \ln\left(\tfrac{1}{1-p}\right)\right) \tag{3.86}$$

$$= P(\ell \,|\, y, r, p) \sum_{\ell=0}^{\infty} P(y, \ell \,|\, r, p) = P(\ell \,|\, y, r, p)\, \mathrm{NB}\,(y;\, r,\, p)\,, \tag{3.87}$$

where $P(\ell \,|\, y, r, p)$ is the *inverse distribution* of ℓ—i.e., its distribution conditioned on y. Remarkably, this distribution does not depend on p and is a Chinese restaurant table distribution—i.e., $P(\ell \,|\, y, r, p) = P(\ell \,|\, y, r) = \mathrm{CRT}(\ell; y, r)$. Thus, we have the following magic[a] identity, proven by Zhou and Carin [2015]:

$$\mathrm{SL}(y;\, \ell, p)\, \mathrm{Pois}\left(\ell;\, r \ln\left(\tfrac{1}{1-p}\right)\right) \;=\; \mathrm{CRT}(\ell;\, y, r)\, \mathrm{NB}\,(y;\, r,\, p) \tag{3.88}$$

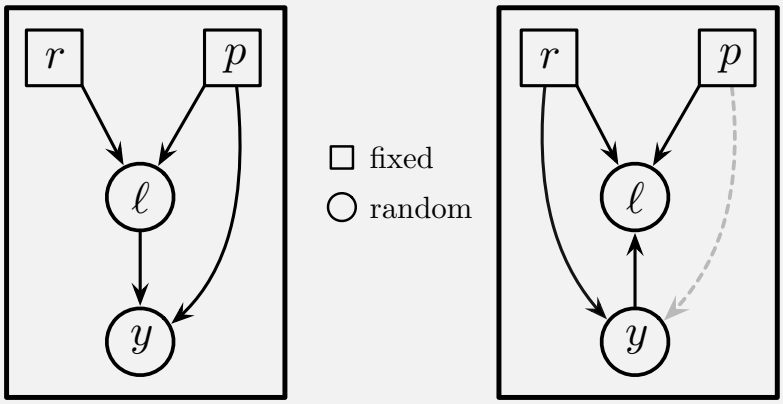


Figure 3.6: In general, these two graphical models only encode the same joint distribution when the shaded arrow is present. However, in the special case where the distributions take the parametric forms written above, these graphical models encode the same bivariate distribution without the shaded arrow.

[a] I'm using "magic" informally to mean unusual, interesting, and powerful—this fact enables tractable posterior inference in a vast array of models that were previously intractable.

3.6 Historical notes

Allocative Poisson factorization unifies convergent threads in statistics, signal-processing, and machine learning. I review some of those threads here. The purpose of this section is not to exhaustively catalogue all papers on Poisson factorization but rather to trace the roots of the main ideas.

3.6.1 Poisson and the Law(s) of Small Numbers

The Poisson distribution has long been understood as a natural and inevitable distribution for the counts of independent rare events. It was first derived as a limiting form of the binomial distribution in 1711 by Abraham de Moivre and again in 1837 by Siméon Denis Poisson, for whom the distribution is named [Johnson et al., 2005, Chapter 4.2]. Specifically, the probability mass function of a binomial random variable $y \sim \text{Binom}(n, p)$ converges to that of a Poisson random variable $y \sim \text{Pois}(\mu)$ as $n \in \mathbb{N}_0$ grows large $n \to \infty$ but the product of n and $p \in [0, 1]$ is fixed to $\mu \triangleq np$. This is often referred to as the "law of small numbers" (LSN) [Whitaker, 1914] since the probability of any individual event becomes small $p \to 0$ when μ is fixed in the limit.

The Law of Small Numbers[2] is the title of an 1898 book by Ladislaus Bortkiewicz who popularized the Poisson distribution to model counts of Prussian soldiers accidentally killed by horse kicks. Quine and Seneta [1987] point out that Bortkiewicz's use of "law of small numbers" does not refer to the binomial limit. Rather, it refers to a related property that a sample of V independent but heterogeneous Poisson random variables $y_v \overset{\text{ind}}{\sim} \text{Pois}(\mu_v)$ for $v \in \{1, \ldots, V\}$ will resemble a sample (of size V) of a homogeneous Poisson random variable $y_v \overset{\text{iid}}{\sim} \text{Pois}(\mu^*)$. Specifically, what Bortkiewicz called the "divergence coefficient"—i.e., the variance-to-mean ratio (VMR) $\frac{\hat{\sigma}^2}{\bar{y}}$ where $\hat{\sigma}^2$ is the sample variance and $\bar{y}$ is the sample mean—tends to 1 even when the

[2]Bortkiewicz may have intended the seeming reference to "law of large numbers", a phrase which was coined by Poisson in 1837 (https://en.wikipedia.org/wiki/Law_of_large_numbers).

rates $\mu_1, \ldots, \mu_V$ are heterogeneous. In expectation, heterogeneity produces *overdispersion*—i.e., VMR exceeding 1. Indeed, the presence of overdispersion via VMR is the standard test for heterogeneity in count data; it's sometimes called the "index of dispersion" [Cox and Lewis, 1966] or, when calculated on subsets of the data, the "Fano factor" [Fano, 1947]. However, Bortkiewicz showed that this test has low power—i.e., it will often fail to correctly identify heterogeneity—when the "field of experience" (i.e., the number of counts V) is small and/or the "scale of experience" (i.e., the scale of the rates μ_v) is small. Related to Bortkiewicz's LSN is the theorem of Le Cam et al. [1960] which establishes sharp bounds on the error induced by approximating the distribution of a sum of possibly dependent Bernoulli random variables—i.e., a Poissonian binomial random variable [Johnson et al., 2005, Chapter 3.12]—with a Poisson distribution. Harremoës et al. [2010] relate these properties back to the binomial limit via what they call the "laws of thin numbers".

An implication of Bortkiewicz's LSN is that a single Poisson distribution often fits a sample of heterogeneous Poisson random variables well. This suggests that APF model are inherently conservative—i.e., the Poisson likelihood assumption encodes an inductive bias towards parsimony. APF models use shared parameters to construct the (heterogeneous) rates of the observed Poisson-distributed counts; the closer those rates are to each other, the fewer shared parameters are needed to construct them.

3.6.2 Statistical analysis of contingency tables

There is a vast literature on the analysis of multivariate categorical data organized into contingency tables [Agresti, 2003]. A contingency table $\boldsymbol{Y}$—i.e., a count tensor (see Section 2.4)—contains the counts $y_{\boldsymbol{\delta}}$ of all possible (multivariate) event types $\boldsymbol{\delta} \equiv (\delta_1, \ldots, \delta_M) \in \Delta$. The goal of contingency table analysis is typically to estimate the probability table Π containing the probabilities $\pi_{\boldsymbol{\delta}}$ of each type, such that $\sum_{\delta \in \Delta} \pi_{\boldsymbol{\delta}} = 1$. Conditioned on the probability table, each event token is a cate-

gorical random variable $\boldsymbol{e}_n \sim \text{Cat}\left(\text{vec}(\Pi)\right)$. Equivalently, the (raveled) contingency table is multinomial $\text{vec}(\boldsymbol{Y}) \sim \text{Multinom}\left(N, \text{vec}(\Pi)\right)$ where N is the total number of observed tokens—i.e., the sum of all the cells in the table $N = \sum_{\delta \in \Delta} y_{\boldsymbol{\delta}}$.

N is often considered fixed (since it is always observed). However, treating N as a Poisson random variable induces a Poisson assumption on the cells $y_{\boldsymbol{\delta}}$ due to the relationship between the Poisson and multinomial (Definition 3.4). This connection is well-known in the contingency table literature—e.g., see Lauritzen [1996, pp. 69–71] for a discussion of Poisson versus multinomial "sampling schemes" as well as [Agresti, 2003, pp. 8–9, 39–40]. Under the multinomial sampling scheme we have $y_{\boldsymbol{\delta}}/N \approx \pi_{\boldsymbol{\delta}}$—thus, in an idealized setting, we needn't model the raw count *magnitudes* in order to estimate Π. However, in the presence of missingness or "censorship", the observed relative proportions $y_{\boldsymbol{\delta}}/N$ conflate the sampling probabilities $\pi_{\boldsymbol{\delta}}$ with the censorship process. Yvonne et al. [1975, Chapter 5] and Agresti [2003, pp. 392] refer to this issue as "sampling" versus "structural zeros" with respect to the zero-valued cells. Winship et al. [2002] discuss models for contingency tables involving masks that explicitly indicate missing cells. Approaches to maximum likelihood estimation for *incomplete multinomial distributions*—i.e., those with censored entries—are surveyed in Johnson et al. [1997, pp. 71-72]. Explicitly modeling the count magnitudes (e.g., by assuming a Poisson sampling scheme) may help propagate uncertainty about N to the estimation of Π. Murray and Reiter [2016] refer to the presence of "complex dependence" in contingency tables of survey response data as the main challenge in dealing with missing values. A related issue is correcting for *measurement error* in contingency tables [Manrique-Vallier and Reiter, 2017] which also motivates explicitly modeling magnitudes (not just relative proportions).

Estimation of the probability table Π was traditionally performed using log-linear models [Fienberg and Rinaldo, 2007] wherein $\log \pi_{\boldsymbol{\delta}}$ is modeled as a linear function of parameters indexed by $\delta_1, \ldots, \delta_M$ and interactions between them. Log-linear models

that consider interaction effects have many parameters; to reduce dimensionalilty, L_1-regularization is commonly employed to zero out parameters [Nardi et al., 2012].

An early alternative to the log-linear approach is known as *latent structure (or "class") analysis* [Lazarsfeld and Henry, 1968, McCutcheon, 1987]. This approach posits a set of latent variables—one for each observed event token $\boldsymbol{z}_n$—that index into a set of *latent classes* $\boldsymbol{z}_n \in \mathcal{K}$ and which render the tokens conditionally independent. Each latent class κ is associated with a simplistic representation of the probability table that makes strong independence assumptions—e.g., $\pi_{\boldsymbol{\delta\kappa}} \propto \prod_{m=1}^{M} \phi^{(m)}_{\delta_m \kappa_m}$. Marginalizing over the latent variables thus induces a mixture model over the cells of the probability table—e.g., $\pi_{\boldsymbol{\delta}} \propto \sum_{\kappa \in \mathcal{K}} \lambda_\kappa \prod_{m=1}^{M} \phi^{(m)}_{\delta_m \kappa_m}$. Since the number of latent classes is typically far fewer than the number of cells—i.e., $\mathcal{K} \ll \Delta$—these models have far fewer parameters than their log-linear counterparts and their parsimony obviates the need for aggressive regularization. Goodman [2002] details the history of the latent class approach and traces its origins all the way back to Peirce [1884].

Dunson and Xing [2009] and Bhattacharya and Dunson [2012] establish the equivalences between latent class analysis and tensor decomposition. Johndrow et al. [2017] further establishes a connection between them and shows tensor decomposition to be a form of regularized log-linear modeling. Both of these approaches are unified under the framework of probabilistic graphical models for discrete random variables [Lauritzen, 1996] which a recent preprint by Cemgil et al. [2019] shows in detail.

3.6.3 $\mathcal{I}$-divergence minimization

A contingency table is a special case of a multiway array or *tensor* whose elements are count-valued. The tensor decomposition literature has traditionally treated count-valued tensors simply as instances of *non-negative tensors*. There is a vast literature on non-negative tensor (NNT) decomposition [Cichocki et al., 2007, Kolda and Bader, 2009]. The traditional approach to tensor decomposition involves minimizing some

divergence between the observed tensor $\boldsymbol{Y}$ and its reconstruction $\boldsymbol{\mu}$. A commonly used divergence for the non-negative decomposition is known as the *generalized KL-divergence*, *unnormalized KL-divergence*, or $\mathcal{I}$*-divergence*. The $\mathcal{I}$-divergence is defined for two non-negative arrays as

$$\mathcal{I}\left(\boldsymbol{p}||\boldsymbol{q}\right) = \sum_{\boldsymbol{\delta}\in\Delta} p_{\boldsymbol{\delta}} \log\left(\frac{q_{\boldsymbol{\delta}}}{p_{\boldsymbol{\delta}}}\right) + \sum_{\boldsymbol{\delta}\in\Delta} p_{\boldsymbol{\delta}} - \sum_{\boldsymbol{\delta}\in\Delta} q_{\boldsymbol{\delta}}. \tag{3.89}$$

When $\boldsymbol{p}$ and $\boldsymbol{q}$ are probability vectors, the last two sums cancel out and the $\mathcal{I}$-divergence equals the (normalized) KL-divergence. When used for non-negative tensor decomposition, the $\mathcal{I}$-divergence is minimized with respect to the parameters governing the reconstruction rates $\mu_{\boldsymbol{\delta}}$. Thus, any terms only involving the data $y_{\boldsymbol{\delta}}$ are constants and can be dropped; doing so reveals that $\mathcal{I}$-divergence minimization is equivalent to maximization of the Poisson likelihood:

$$\begin{aligned}
\mathcal{I}(\boldsymbol{Y}\,||\,\boldsymbol{\mu}) &= \sum_{\boldsymbol{\delta}\in\Delta} y_{\boldsymbol{\delta}} \log\left(\frac{\mu_{\boldsymbol{\delta}}}{y_{\boldsymbol{\delta}}}\right) + \sum_{\boldsymbol{\delta}\in\Delta} y_{\boldsymbol{\delta}} - \sum_{\boldsymbol{\delta}\in\Delta} \mu_{\boldsymbol{\delta}} && (3.90)\\
&\propto \sum_{\boldsymbol{\delta}\in\Delta} y_{\boldsymbol{\delta}} \log \mu_{\boldsymbol{\delta}} - \sum_{\boldsymbol{\delta}\in\Delta} \mu_{\boldsymbol{\delta}} && (3.91)\\
&\propto -\sum_{\boldsymbol{\delta}\in\Delta} \log \text{ Pois}\left(y_{\boldsymbol{\delta}};\, \mu_{\boldsymbol{\delta}}\right) && (3.92)
\end{aligned}$$

Lee and Seung [1999] introduced the now widely-used multiplicative updates for non-negative *matrix* factorization (NMF) for both Euclidean distance and $\mathcal{I}$-divergence. Welling and Weber [2001] generalized the NMF multiplicative updates to CP decomposition of non-negative tensors while Kim and Choi [2007] generalized them to Tucker decomposition. The $\mathcal{I}$-divergence can be unified under different divergence families. Sra and Dhillon [2006] introduced NMF for general Bregman divergences while Cichocki et al. [2007] developed non-negative tensor decomposition for the family of β-divergences.

Direct minimization of the $\mathcal{I}$-divergence corresponds to maximum likelihood estimation in a Poisson factorization model. These algorithms may be unstable, particularly when the tensor is very sparse. Gonzalez and Zhang [2005] showed for Euclidean distance that Lee and Seung [1999]'s multiplicative updates may not converge to a local minimum. Finesso and Spreij [2006] established the *theoretical* convergence of the NMF updates for $\mathcal{I}$-divergence. However, Chi and Kolda [2012] later demonstrated that $\mathcal{I}$-divergence minimization may still not converge to a local minimum, in practice, due to numerical instability and parameters converging to "inadmissable zeros". Chi and Kolda [2012] propose an alternative algorithm for $\mathcal{I}$-divergence minimization that is more stable. Other heuristic solutions have been proposed, like Gillis and Glineur [2008], who clipped parameters above some small value $\epsilon > 0$. The instability in these algorithms motivate a subsequent line of probabilistic and Bayesian approaches that enjoy "implicit regularization" [Nakajima and Sugiyama, 2010] via priors.

3.6.4 Probabilistic and Bayesian Poisson factorization

The roots of Poisson factorization within the machine learning community are in topic models of document-by-word count matrices. Hofmann [1999] introduced probabilistic latent semantic analysis (pLSA) which consists of fitting the "aspect model" (previously coined by Hofmann et al. [1999]) with the EM algorithm [Dempster et al., 1977] to document-by-word matrices. The aspect model can be written as

$$P\left(\boldsymbol{e}_n = (d, v)\right) = \sum_{k=1}^{K} P(k)\, P(d \,|\, k)\, P(v \,|\, k), \tag{3.93}$$

which is the canonical form of latent class analysis and equivalent to the model of Saul and Pereira [1997] among others in the statistics literature on contingency tables.

Blei et al. [2001] introduced Latent Dirichlet allocation (LDA) by imposing Dirichlet priors over the document-topic factors of the aspect model—i.e., $\theta_{dk} \equiv P(k \,|\, d)$ where $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\cdots)$—and introduced a variational EM algorithm to fit it. Griffiths

[2002] further imposed Dirichlet priors over the word-topic factors—i.e., $\phi_{kv} \equiv P(v \mid k)$ where $\boldsymbol{\phi}_k \sim \text{Dir}(\cdots)$—and introduced the now widely-used collapsed Gibbs sampler (CGS) for inference of the topic indicators z_n. Minka and Lafferty [2002] introduced an expectation propagation algorithm [Minka, 2001] for the full model and Blei et al. [2003] introduced a mean-field variational inference algorithm. LDA was also concurrently introduced in the statistical genetics community by Pritchard et al. [2000].

The connections between pLSA, LDA, and NMF were quickly noticed. Buntine [2002] introduced multinomial PCA (mPCA) and showed that it unified the aspect model, LDA, and the (implicit) model of NMF with $\mathcal{I}$-divergence. Girolami and Kabán [2003] established the equivalence of pLSA and to fitting LDA with variational EM and uniform Dirichlet priors. Gaussier and Goutte [2005] established that pLSA solves the NMF objective—i.e., all the solutions to pLSA are local minima for NMF.

As in traditional contingency table analysis, LDA and pLSA are mainly models for the relative proportions of counts, not their raw magnitudes. Canny [2004] introduced the gamma–Poisson (GaP) model which was one of the first *explicit* instance of Poisson matrix factorization. The GaP model imposes independent gamma priors over the elements of one factor matrix and treats the other as fixed. Canny [2004] provides an EM-like algorithm and discusses its similarity to the multiplicative updates for NMF. Dunson and Herring [2005] introduced a Gibbs sampler for a Poisson matrix factorization model that includes covariates; equation 9 of their paper includes one of the first explicit references to the latent source representation (see Section 3.2). Buntine and Jakulin [2006] introduced Discrete Component Analysis which further generalized NMF, pLSA, LDA, mPCA, and GaP; Lemma 1 of their paper establishes the equivalence of LDA to gamma–Poisson models. Titsias [2008] introduced Bayesian non-parametric gamma–Poisson matrix factorization along with a Gibbs sampler. Cemgil [2009] introduced mean-field variational inference for Poisson factorization with gamma priors and discussed the algorithm's similarity to the

multiplicative updates of NMF. Paisley et al. [2014] introduced the corresponding algorithm for stochastic variational inference [Hoffman et al., 2013]. Gopalan et al. [2014a] introduced a variational algorithm for a Bayesian non-parametric Poisson MF and Gopalan et al. [2015] introduced a stochastic variational algorithm for Poisson MF with hierarchical priors over the rate parameters of the gamma priors.

Zhou et al. [2012] introduced a formal treatment of Bayesian non-parametric gamma–Poisson matrix factorization (or "Poisson Factor Analysis") in the language of completely random measures (CRMs) [Kingman, 1967, Jordan, 2010] which yields the beta–negative binomial process (BNBP). Virtually the same Bayesian non-parametric process was introduced concurrently by Broderick et al. [2011, 2015]. This model also goes beyond that of Titsias [2008] by imposing hierarchical priors over both the shape and rate of the gamma priors over factors. Zhou and Carin [2012] introduced a novel auxiliary variable scheme based on augmentation of the negative binomial distribution (discussed in Section 3.5) to perform closed-form inference for the shape of the gamma prior. Zhou and Carin [2015] expanded on this augmentation scheme and detailed its applications to different hierarchical models.

Finally, much of the Bayesian models for Poisson matrix factorization have been generalized to tensor decomposition. Mean-field variational inference for gamma–Poisson CP decomposition was introduced by Ermis and Cemgil [2014] (concurrently with Schein et al. [2014]). Although Ermis and Cemgil [2014] focus on CP decomposition, they employ the generalized coupled tensor factorization (GCTF) framework of Yılmaz et al. [2011] and their algorithm thus implies an algorithm for Tucker decomposition which Schein et al. [2016b] later introduced explicitly. Hu et al. [2015] introduced stochastic variational inference for Poisson CP decomposition with hierarchical priors. Many more instances of Bayesian Poisson tensor decompositions have since been introduced.

CHAPTER 4

BAYESIAN POISSON TENSOR FACTORIZATION FOR INFERRING MULTILATERAL RELATIONS FROM SPARSE DYADIC EVENT COUNTS

Over the past fifteen years, political scientists have engaged in an ongoing debate about using dyadic events to study inherently multilateral phenomena. This debate, as summarized by Stewart [2014], began with Green et al. [2001]'s demonstration that many regression analyses based on dyadic events were biased due to implausible independence assumptions. Researchers continue to expose such biases, e.g., [Erikson et al., 2014], and some have even advocated eschewing dyadic data on principle, calling instead for the development of multilateral event data sets [Poast, 2010]. Taking the opposite viewpoint—i.e., that dyadic events can be used to conduct meaningful analyses of multilateral phenomena—other researchers, beginning with Hoff and Ward [2004], have developed Bayesian latent factor regression models that explicitly model unobserved dependencies as occurring in some latent space, thereby controlling for their effects in analyses. This line of research has seen an increase in interest and activity over the past few years [Hoff et al., 2016, Stewart, 2014, Hoff, 2015].

This chapter proposes an APF model for measuring a particular kind of "complex dependence structure in international relations" [King, 2001] implicit in dyadic event data—i.e., *multilateral relations*. An example of an inferred multilateral relation is given in Fig. 4.1. This chapter is based on work published in KDD 2014 [Schein et al., 2015]. I have factored out many of the details originally presented in that paper into the preceding three chapters of this thesis. The example model in Section 3.4 with gamma priors is the 3-mode analogue of the one in this chapter and

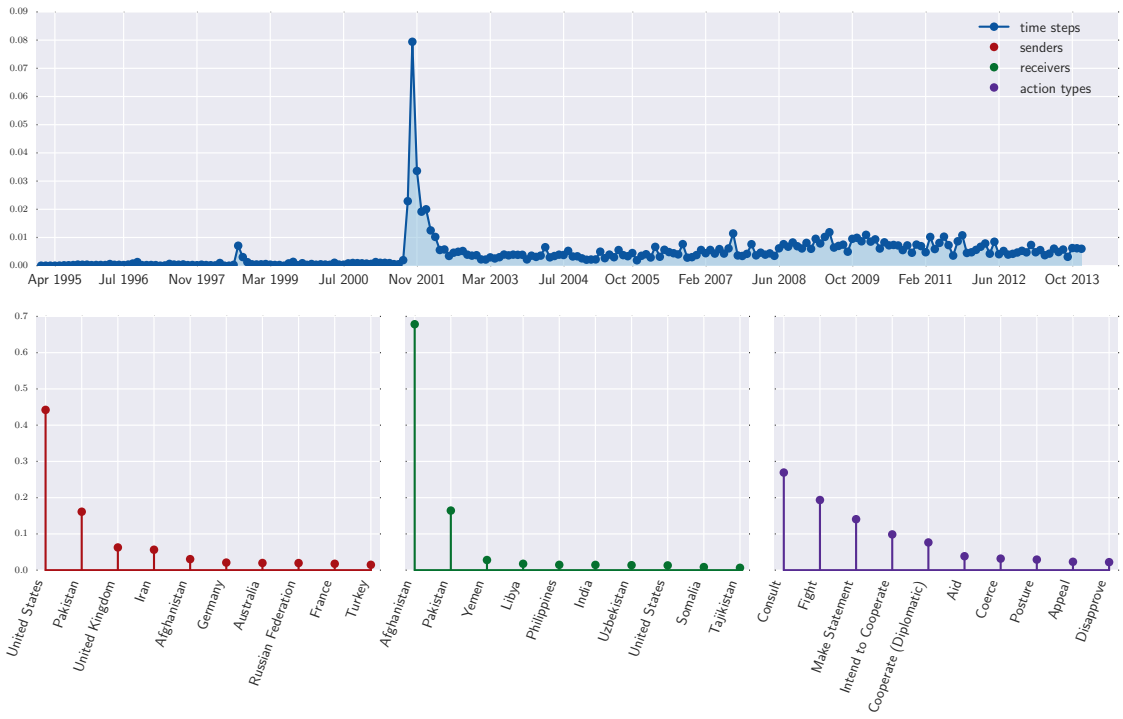


Figure 4.1: Our model infers latent classes that correspond to multilateral relations. Each class consists of four factor vectors summarizing sender, receiver, action-type, and time-step activity, respectively. Here we visualize a single class, for which we plot the top ten sender, receiver, and action-type factors sorted in decreasing order. We also plot the entire vector of time-step factors in chronological order. We found that the interpretation of each class was either immediately clear from our existing knowledge or easy to discover via a web search. This class was inferred from ICEWS data spanning 1995 through 2012 (with monthly time steps). It corresponds to events surrounding the US-led War on Terror following the September 11, 2001 attacks. The largest time-step factor is that of October 2001—the month during which the invasion of Afghanistan occurred. There is also a blip in August 1998, when the Clinton administration ordered missile attacks on terrorist bases in Afghanistan Wikipedia contributors [2018c].

the derivation of CAVI updates for that model apply to this one. This chapter thus omits derivations but includes more details and intuition about the CAVI updates including their close relationship to analogous maximum likelihood approaches [Lee and Seung, 1999]. This chapter also explores the geometric versus arithmetic expectations that appear in gamma–Poisson CAVI and advocates for the former to be used in constructing point estimates. Results are presented that demonstrate empirically that BPTF has better out-of-sample predictive performance than its maximum likelihood and non-Poisson counterparts. This chapter also explores the inferred latent parameter matrices and shows they capture structures that conform to and inform our knowledge of international affairs.

4.1 International relations dyadic event data

Over the past few years, researchers have created large data sets of dyadic events by automatically extracting them from Internet news archives. The largest of these data sets is the Global Database of Events, Location, and Tone (GDELT), introduced in 2013, which contains over a quarter of a billion events from 1979 to the present, and is updated with new events daily [Leetaru and Schrodt, 2013]. In parallel, government agencies (e.g., DARPA) and their contractors have also started to collect and analyze dyadic events, in order to forecast political instability and to develop early-warning systems [O'Brien, 2010]; Lockheed Martin publicly released the Integrated Crisis Early Warning System (ICEWS) database in early 2015. Ward et al. [2013] provide a comparison of GDELT and ICEWS.

GDELT and ICEWS use the CAMEO coding scheme [Gerner et al.]. A CAMEO-coded dyadic event consists of four pieces of information: a sender, a receiver, an action type, and a time stamp. An example of such an event (top) and a sentence from which it could have been extracted (bottom) is

(Turkey, Syria, *Fight*, 12/25/2014)

Dec. 25, 2014: "Turkish jets bombed targets in Syria."

CAMEO assumes that senders and receivers belong to a single set of actors, coded for their country of origin and sector (e.g., government or civilian) as well as other information (such as religion or ethnicity). CAMEO also assumes a hierarchy of action types, with the top level consisting of twenty basic action classes. These classes are loosely ranked based on sentiment from *Make Public Statement* to *Use Unconventional Mass Violence*. Each action class is subdivided into more specific actions; for example, *Make Public Statement* contains *Make Empathetic Comment*. When studying international relations using CAMEO-coded data, researchers commonly consider only the countries of origin as actors and only the twenty basic action classes as action types. In ICEWS, there are 249 unique country actors (which include non-universally recognized countries, such as Taiwan and Palestine); in GDELT, there are 223.

A data set of dyadic events can be aggregated into a 4-mode tensor $\boldsymbol{Y}$ of size $V \times V \times A \times T$, where V is the number of country actors and A is the number of action types, by aggregating the events into T time steps on the basis of their timestamps. Each element $y^{(t)}_{i \xrightarrow{a} j}$ of $\boldsymbol{Y}$ is a count of the number of actions of type a taken by country i toward country j during time step t. We experimented with various date ranges and time step granularities. For example, in one set of experiments, we used the entire ICEWS data set, spanning 1995 through 2012 (i.e., 18 years) with monthly time steps—i.e., a 249×249×20×216 tensor with 267,844,320 elements. In this tensor, only 0.54% of the elements (roughly 1.5 million elements) are non-zero. Moreover, these non-zero counts are highly dispersed with a variance-to-mean ratio (VMR) of 57. Any realistic model of such data must therefore be robust to sparsity and capable of representing high levels of dispersion.

4.2 Model: Bayesian Poisson Tensor Factorization

BPTF assumes each count is an independent Poisson random variable,

$$y^{(t)}_{i \xrightarrow{a} j} \sim \text{Pois}\left(\mu^{(t)}_{i \xrightarrow{a} j}\right) \tag{4.1}$$

$$\mu^{(t)}_{i \xrightarrow{a} j} \triangleq \sum_{k=1}^{K} \theta^{(1)}_{ki} \theta^{(2)}_{kj} \theta^{(3)}_{ka} \theta^{(4)}_{kt}, \tag{4.2}$$

which corresponds to a CP decomposition (see Section 2.3.3) of the observed tensor into four latent parameter matrices where $\Theta^{(1)}$ and $\Theta^{(2)}$ are both $K \times V$ matrices whose columns represent the embeddings of all V country actors as senders and receivers, respectively, $\Theta^{(3)}$ is a $K \times A$ matrix whose columns embed action types, and $\Theta^{(4)}$ is a $K \times T$ whose columns embed time steps. We assume independent gamma priors for all elements of the parameter matrices e.g.,

$$\theta^{(1)}_{ki} \sim \Gamma(\alpha_0, \alpha_0 \beta^{(1)}), \tag{4.3}$$

where α_0 is set to a small value and $\beta^{(m)}$ (one for each for the four modes) is inferred. This prior structure represents the tensor generalization of the prior structure for Bayesian Poisson *matrix* factorization introduced by Cemgil [2009] who showed that these gamma priors improve interpretability and prevent overfitting by promoting parameter shrinkage. Under this parameterization of the gamma distribution, where the rate parameter is the product of the shape parameter and $\beta^{(1)}$, the mean of the prior is completely determined by $\beta^{(1)}$ (since $\mathbb{E}[\theta^{(1)}_{ki}] = \frac{\alpha_0}{\alpha_0 \beta^{(1)}} = \frac{1}{\beta^{(1)}}$) Cemgil [2009], Liang et al. [2014]. The shape parameter α_0, which determines the shrinkage of the latent parameter matrices, can be set by the user; in our experiments, we use $\alpha_0 = 0.1$.

4.3 Variational inference

Given an observed tensor $\boldsymbol{Y}$, Bayesian inference of the latent factors involves "inverting" the generative process described in the previous section to obtain the

posterior distribution of the latent parameter matrices conditioned on $\boldsymbol{Y}$ and the model hyperparameters $\mathcal{H} \equiv \{\alpha_0, \beta^{(1)}, \beta^{(2)}, \beta^{(3)}, \beta^{(4)}\}$:

$$P\left(\Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}, \Theta^{(4)} \,|\, \mathbf{Y}, \mathcal{H}\right). \tag{4.4}$$

The posterior distribution for BPTF is analytically intractable and must be approximated. Variational inference turns the process of approximating the posterior distribution into an optimization algorithm. It involves first specifying a parametric family of distributions Q over the latent variables of interest, indexed by the values of a set of *variational parameters* $\mathcal{S}$. The functional form of Q is typically chosen so as to facilitate efficient optimization of $\mathcal{S}$. Here, we use a fully factorized *mean-field approximation* (see Section 2.2.2.2). As shown in Section 3.4 the optimal family for each factor—e.g., for $\theta^{(1)}_{ki}$—is,

$$Q(\theta^{(1)}_{ki}; \mathcal{S}^{(1)}_{ki}) = \Gamma\left(\theta^{(1)}_{ki}; \gamma^{(1)}_{ki}, \delta^{(1)}_{ki}\right), \tag{4.5}$$

where $\mathcal{S}^{(1)} \equiv \left(\left(\gamma^{(1)}_{ki}, \delta^{(1)}_{ki}\right)_{i=1}^{V}\right)_{k=1}^{K}$. The full set of variational parameters is thus $\mathcal{S} \equiv \{\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \mathcal{S}^{(3)}, \mathcal{S}^{(4)}\}$. This form of Q is similar to that used in Bayesian PMF [Cemgil, 2009, Paisley et al., 2014, Gopalan et al., 2015].

The variational parameters are then fit so as to yield the closest member of Q to the exact posterior. Specifically, the algorithm sets the values of $\mathcal{S}$ to those that minimize the KL divergence of the exact posterior from Q. It can be shown (e.g., see [Blei et al., 2017]) that these values are the same as those that maximize a lower bound on $P(\boldsymbol{Y} \,|\, \mathcal{H})$, known as the *evidence lower bound* (ELBO):

$$\mathcal{B}(\mathcal{S}) = \mathbb{E}_Q\left[\log\left(P(\boldsymbol{Y}, \Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}, \Theta^{(4)} \,|\, \mathcal{H})\right)\right] + H(Q),$$

where $H(Q)$ is the entropy of Q. When Q is a fully factorized approximation, finding values of $\mathcal{S}$ that maximize the ELBO can be achieved by performing coordinate ascent,

iteratively updating each variational parameter, while holding the others fixed, until convergence (defined by change in the ELBO). For $\gamma_{ki}^{(1)}$ and $\delta_{ki}^{(1)}$, the updates:

$$\gamma_{ki}^{(1)} := \alpha_0 + \sum_{j,a,t} y_{i \xrightarrow{a} j}^{(t)} \frac{\mathbb{G}_Q\left[\theta_{ki}^{(1)}\theta_{kj}^{(2)}\theta_{ka}^{(3)}\theta_{tk}^{(4)}\right]}{\sum_{k=1}^{K}\mathbb{G}_Q\left[\theta_{ki}^{(1)}\theta_{kj}^{(2)}\theta_{ka}^{(3)}\theta_{tk}^{(4)}\right]} \tag{4.6}$$

$$\delta_{ki}^{(1)} := \alpha_0\,\beta^{(1)} + \sum_{j,a,t}\mathbb{E}_Q\left[\theta_{kj}^{(2)}\theta_{ka}^{(3)}\theta_{tk}^{(4)}\right], \tag{4.7}$$

where $\mathbb{E}_Q[\cdot]$ and $\mathbb{G}_Q[\cdot] = \exp\left(\mathbb{E}_Q[\log(\cdot)]\right)$ denote arithmetic and geometric expectations. Since Q is fully factorized, each expectation of a product can be factorized into a product of individual expectations, which, e.g., for $\theta_{ki}^{(1)}$ are

$$\mathbb{E}_Q\left[\theta_{ki}^{(1)}\right] = \frac{\gamma_{ki}^{(1)}}{\delta_{ki}^{(1)}} \quad \text{and} \quad \mathbb{G}_Q\left[\theta_{ki}^{(1)}\right] = \frac{\exp\left(\Psi\left(\gamma_{ki}^{(1)}\right)\right)}{\delta_{ki}^{(1)}}, \tag{4.8}$$

where $\Psi(\cdot)$ is the digamma function. Each expectation—a sufficient statistic—can be cached to improve efficiency. Note that the summand in Eq. (4.6) need only be computed for those values of j, a, and t for which $y_{i \xrightarrow{a} j}^{(t)} > 0$; provided $\boldsymbol{Y}$ is very sparse, inference is efficient even for very large tensors.

The hyperparameters $\beta^{(1)}$, $\beta^{(2)}$, $\beta^{(3)}$, and $\beta^{(4)}$ can be optimized via an empirical Bayes method, in which each hyperparameter is iteratively updated along with the variational parameters according to the following update equation:

$$\beta^{(1)} := \left(\sum_{i,k}\mathbb{E}_Q\left[\theta_{ki}^{(1)}\right]\right)^{-1}. \tag{4.9}$$

Eqs. (4.6), (4.7) and (4.9) completely specify the variational inference algorithm for BPTF. Our Python implementation, which is intended to support arbitrary M-mode tensors in addition to the four-mode tensors described in this paper, is available for use under an open-source license[1].

[1] https://github.com/aschein/bptf

4.4 Predictive Analysis

This section compares BPTF's predictive performance to that of standard methods for non-negative tensor factorization involving maximum likelihood estimation.

Baselines: Non-Bayesian methods for CP decomposition find values of the latent parameter matrices that minimize some divergence of the observed tensor $\boldsymbol{Y}$ and its reconstruction $\boldsymbol{\mu}$. Researchers have proposed many divergences, but most often use Euclidean distance or generalized KL divergence (or $\mathcal{I}$-divergence), preferring the latter when the observed tensor consists of sparse counts. Generalized KL divergence is

$$\mathcal{I}(\boldsymbol{Y} \,||\, \boldsymbol{\mu}) = -\sum_{i,j,a,t} \left(y^{(t)}_{i \xrightarrow{a} j} \log \left(\mu^{(t)}_{i \xrightarrow{a} j} \right) - \mu^{(t)}_{i \xrightarrow{a} j} \right) + \mathcal{C}, \tag{4.10}$$

where constant $\mathcal{C} \equiv \sum_{i,j,a,t} \left(y^{(t)}_{i \xrightarrow{a} j} \log \left(y^{(t)}_{i \xrightarrow{a} j} \right) - y^{(t)}_{i \xrightarrow{a} j} \right)$ depends on the observed data only. The standard method for estimating the values of the latent factors involves multiplicative update equations, originally introduced for matrix factorization by Lee and Seung [1999] and later generalized to tensors by Welling and Weber [2001]. The multiplicative nature of these update equations acts as a non-negativity constraint on the factors which promotes interpretability and gives the algorithm its name: non-negative tensor factorization (NTF).

Some divergences also permit a probabilistic interpretation: finding values of the latent factors that minimize them is equivalent to maximum likelihood estimation of a probabilistic model. The log likelihood function of a Poisson tensor factorization model—$y^{(t)}_{i \xrightarrow{a} j} \sim \text{Pois}\left(y^{(t)}_{i \xrightarrow{a} j}; \mu^{(t)}_{i \xrightarrow{a} j} \right)$—is

$$\log \prod_{i,j,a,t} \text{Pois}\left(y^{(t)}_{i \xrightarrow{a} j}; \mu^{(t)}_{i \xrightarrow{a} j} \right) = \sum_{i,j,a,t} y^{(t)}_{i \xrightarrow{a} j} \log \left(\mu^{(t)}_{i \xrightarrow{a} j} \right) - \log \left(y^{(t)}_{i \xrightarrow{a} j}! \right) - \mu^{(t)}_{i \xrightarrow{a} j} \tag{4.11}$$

$$= \sum_{i,j,a,t} \left(y^{(t)}_{i \xrightarrow{a} j} \log \left(\mu^{(t)}_{i \xrightarrow{a} j} \right) - \mu^{(t)}_{i \xrightarrow{a} j} \right) + \mathcal{C}, \tag{4.12}$$

Table 4.1: Out-of-sample predictive performance for our model (BPTF) and non-negative tensor factorization with Euclidean distance (NTF-LS) and generalized KL divergence (NTF-KL or, equivalently, PTF). Each row contains the results of a single experiment. "I-top-25" means the experiment used data from ICEWS and we predicted the upper-left 25×25 portion of each test slice (and treated its complement as observed). "G-top-100^c" means the experiment used data from GDELT and we predicted the complement of the upper-left 100×100 portion of each test slice. For each experiment, we state the density (percentage of non-zero elements) and VMR (i.e., dispersion) of the unobserved portion of the test set. We report three types of error: mean absolute error (MAE), mean absolute error on non-zero elements (MAE-NZ), and Hamming loss on the zero elements (HAM-Z). All models achieved comparable scores when we predicted the sparser portion of each test slice (bottom four rows). BPTF significantly outperformed the other models when we predicted the denser 25×25 or 100×100 portion (top four rows).

			NTF-LS			NTF-KL (PTF)			BPTF		
	Density	VMR	MAE	-NZ	HAM-Z	MAE	-NZ	HAM-Z	MAE	-NZ	HAM-Z
I-top-25	0.1217	105.8755	34.4	217	0.271	8.37	56.7	0.138	**1.99**	**12.9**	**0.113**
G-top-25	0.2638	180.4143	52.5	167	0.549	15.5	53.7	0.327	**8.94**	**29.8**	**0.292**
I-top-100	0.0264	63.1118	29.8	979	0.0792	10.5	346	0.0333	**0.178**	**5.05**	**0.0142**
G-top-100	0.0588	111.8676	42.6	470	0.217	4	58.6	0.0926	**0.95**	**12.2**	**0.0682**
I-top-25^c	0.0021	8.6302	**0.00657**	**2.27**	**0.00023**	0.0148	2.72	0.00256	0.0104	2.31	0.00161
G-top-25^c	0.0060	20.4858	0.0435	4.4	**0.00474**	0.0606	4.9	0.00893	**0.0412**	**4.01**	0.00601
I-top-100^c	0.0004	4.4570	**0.000685**	1.63	**3.33e-07**	0.0011	**1.55**	5.43e-05	0.00109	1.56	4.97e-05
G-top-100^c	0.0015	9.9432	**0.00584**	3.23	**0.000112**	0.0084	**2.97**	0.00109	0.00803	3	0.000957

where constant $\mathcal{C} \equiv -\sum_{i,j,a,t} \log\left(y^{(t)}_{i \xrightarrow{a} j}!\right)$ depends on the observed data only. Since Eq. (4.10) is equal to the negative of Eq. (4.12) up to a constant, maximum likelihood estimation for Poisson tensor factorization is equivalent to minimizing the generalized KL divergence of $\boldsymbol{\mu}$ from $\boldsymbol{Y}$.

We compared the out-of-sample predictive performance of BPTF to that of non-negative tensor factorization with Euclidean distance (NTF-LS) and non-negative tensor factorization with generalized KL divergence (NTF-KL or, equivalently PTF).

Experimental design: Using both ICEWS and GDELT, we explored how well each model generalizes to out-of-sample data with varying degrees of sparsity and dispersion. For each data set—ICEWS or GDELT—we sorted the country actors by their overall activity (as both sender and receiver) so that the $V \times V$ sender–receiver slices of the observed tensor were denser toward the upper-left corner. Section 4.4 depicts this property. We then divided the observed tensor into a training set and a test set by randomly constructing an 80%–20% split of the time steps. We defined

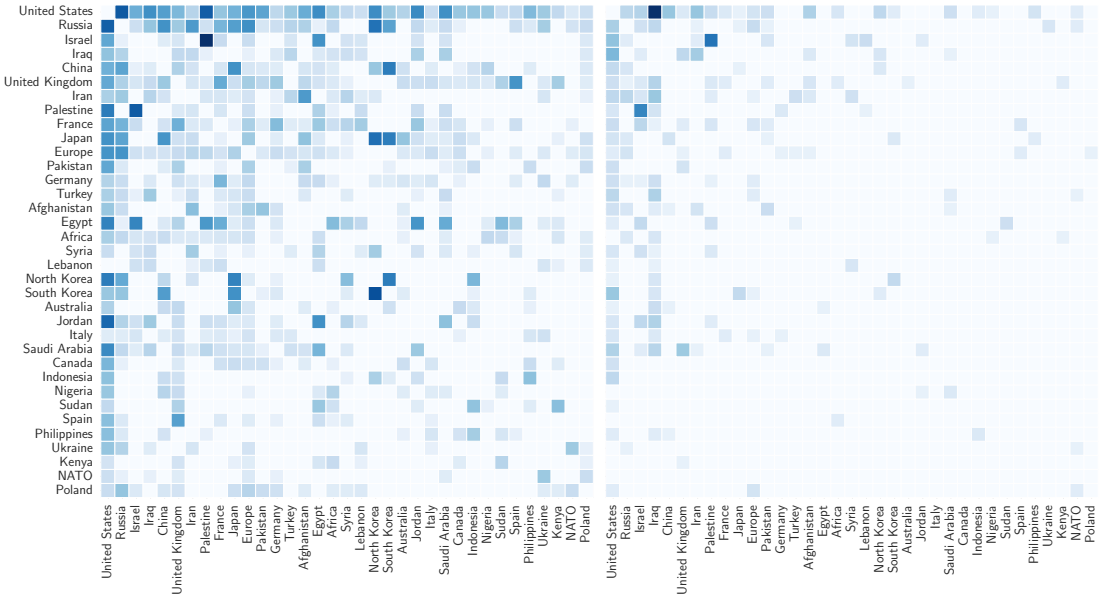


Figure 4.2: Sender–receiver slices from the GDELT tensor spanning 1990 through 2007, with monthly time steps (i.e., $T = 216$). Both slices correspond to $t = 151$ (July 2002). The left slice corresponds to *Intend to Cooperate*, while the right slice corresponds to *Threaten*. We sorted the country actors by their overall activity so that the slices were generally denser toward the upper-left corner; only the upper-left 35×35 portion of each slice is shown here. The three darkest elements (i.e., highest counts) in the second slice correspond to Israel → Palestine, Palestine → Israel, and US → Iraq.

training set $\boldsymbol{Y}^{\text{train}}$ to be the $V \times V \times A$ slices of $\boldsymbol{Y}$ indexed by the time steps in the 80% split and defined test set $\boldsymbol{Y}^{\text{test}}$ to be the $V \times V \times A$ slices indexed by the time steps in the 20% split. We compared the models' predictive performance in two scenarios, intended to test their abilities to handle different levels of sparsity and dispersion: one in which we treated the denser upper-left $V' \times V'$ (for some $V' < V$) portion of each test slice as observed at test time and predicted its complement, and one in which we observed the complement at test time and predicted the denser $V' \times V'$ portion.

In each setting, we used an experimental strategy analogous to *strong generalization* for collaborative filtering [Marlin, 2004]. During training, we fit each model to the fully observed training set. We then fixed the values of the variational parameters for the sender, receiver, and action-type parameter matrices (or direct point estimates

of the factors, for the non-Bayesian models) to those inferred from the training set. For each test slice, indexed by time step t, we used the observed upper-left $V' \times V'$ portion (or its complement) to infer variational parameters for (or direct point estimates of) its time-step factors $\{\theta^{(4)}_{kt}\}_{k=1}^{K}$. Finally, we reconstructed the missing portion of each test slice using $\mu^{(t)}_{i \xrightarrow{a} j}$. For the reconstruction step, we can obtain point estimates of the latent factors by taking their arithmetic expectations or their geometric expectations—i.e., $\mathbb{E}_Q\left[\mu^{(t)}_{i \xrightarrow{a} j}\right]$ or $\mathbb{G}_Q\left[\mu^{(t)}_{i \xrightarrow{a} j}\right]$. In this section, we report results obtained using geometric expectations only; we explain this choice in Section 4.6.

We used the entire ICEWS data set from 1995 through 2012 (i.e., 18 years)[2], with events aggregated into monthly time steps. The resultant tensor was of size 249×249×20×216. Since GDELT covers a larger date range (1979 to the present) than ICEWS, we therefore selected an 18-year subset of GDELT spanning 1990 through 2007, and aggregated events into monthly time steps to yield a tensor of size 223×223×20×216. Since we are interested in interactions between countries, we omitted self-actions so that the diagonal of each $V \times V$ sender–receiver slice was zero. Ranking the country actors by their overall activity (as both sender and receiver), the top four actors in the ICEWS tensor are USA, Russia, China, and Israel, while the top four actors in the GDELT tensor are USA, Russia, Israel, and Iraq. The GDELT tensor contains many more events than the ICEWS tensor (26 million events versus six million events). It is also much denser (1.6% of the elements are non-zero, as opposed to 0.54%) and exhibits a much higher level of dispersion (VMR of 100, as opposed to 57).

Summary of results: The out-of-sample predictive performance of each model is shown in Table 4.1. We experimented with several different values of K and found that all three models were insensitive to its value; we therefore report only those results obtained using $K = 50$. We computed three types of error: mean absolute error

[2]At the time these experiments were performed, this represented the entire ICEWS data set. ICEWS now contains events in subsequent years.

(MAE), mean absolute error on only non-zero elements (MAE-NZ), and Hamming loss on only the zero elements (HAM-Z). HAM-Z corresponds to the fraction of true zeros in the unobserved portion of the test set (i.e., elements for which $y^{(t)}_{i \xrightarrow{a} j} = 0$) whose reconstructions were (incorrectly) predicted as being greater than 0.5. For each data set, we generated three training–test splits, and averaged the error scores for each model across them. For each experiment included in Table 4.1, we display the density and dispersion of the corresponding test set. When we treated the dense upper-left $V' \times V'$ portion as observed at test time (and predicted its complement), all models performed comparably. In this scenario, NTF-LS consistently achieved the lowest MAE score and the lowest HAM-Z score, but not the lowest MAE-NZ score. This pattern suggests that NTF-LS overfits the sparsity of the training set: when the unobserved portion of the test set is much sparser than the training set (as it is in this scenario), NTF-LS achieves lower error scores by simply predicting many more zeros than NTF-KL (i.e., PTF) or BPTF. In the opposite scenario, when we observed the complement at test time and predicted the denser $V' \times V'$ portion, NTF-LS produced significantly worse predictions than the other models, and our model (BPTF) achieved the lowest MAE, MAE-NZ, and HAM-Z scores—in some cases by an order of magnitude over NTF-KL. These results suggest that in the presence of sparsity, BPTF is a much better model for the "interesting" portion of the tensor—i.e., the dense non-zero portion. This observation is consistent with previous work by Chi and Kolda [2012] which demonstrated that NTF can be unstable, particularly when the observed tensor is very sparse. In Section 4.6, we provide a detailed discussion comparing NTF and BPTF, and explain why BPTF overcomes the sparsity-related issues often suffered by NTF.

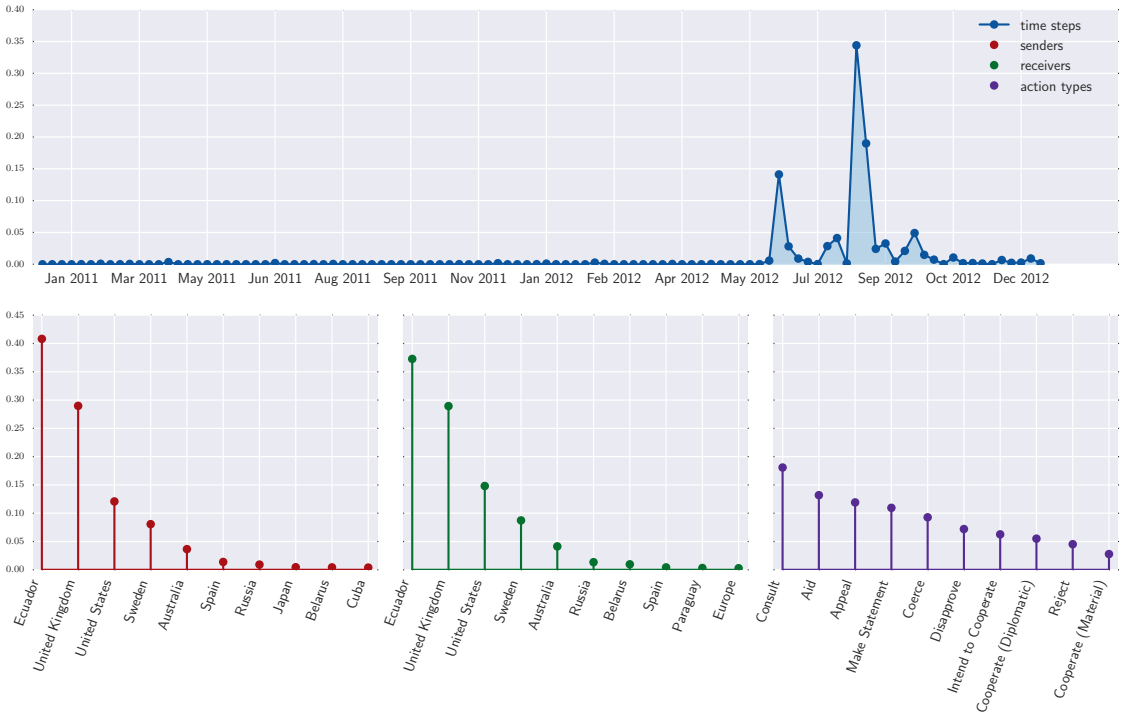


Figure 4.3: Julian Assange, editor-in-chief of WikiLeaks, sought asylum at the Ecuadorian embassy in the UK during June through August 2012. This component inferred from GDELT (2011 through 2012, with weekly time steps) had the sparsest time-step factor vector. We performed a web search for *ecuador UK sweden june 2012* to interpret this component.

4.5 Exploratory Analysis

In this section, we focus on the interpretability of the latent components inferred using our model. (Recall that each latent parameter matrix has K rows; a single index $k \in \{1, \ldots, K\}$ indexes a row in each matrix—$\left(\theta^{(1)}_{ki}\right)_{i=1}^{V}$, $\left(\theta^{(2)}_{kj}\right)_{j=1}^{V}$, $\left(\theta^{(3)}_{ka}\right)_{a=1}^{A}$, and $\left(\theta^{(4)}_{kt}\right)_{t=1}^{T}$—collectively known as a component.) We used our model to explore data from GDELT and ICEWS with several date ranges and time step granularities, including the 18-year, monthly-time-step tensors described in the previous section (treated here as fully observed).

When inferring parameter matrices from data that span a large date range (e.g., 18 years), we expect that the inferred components will correspond to multilateral

relations that persist or recur over time. We found that many of the components inferred from 18-year tensors summarize regional relations—i.e., multilateral relations that persist due to geographic proximity—similar to those found by Hoff [2015]. We further found a high correspondence between the regional components inferred from GDELT and the regional components inferred from ICEWS, despite the five-year difference in their date ranges. Fig. 4.4 illustrates this correspondence. We also found that components summarizing regional relations exhibited the least sparsity in their sender, receiver, and time-step factors. For example, the component depicted in Fig. 4.4 has near-uniform values for the top ten sender and receiver actors (all of whom are regional to Central Asia), while the time-step factors possess high activity throughout. In contrast, the time-step factors for the component shown in the second plot of Fig. 4.1 (i.e., the War on Terror) exhibit a major spike in October 2001. This component's sender and receiver factors also exhibit uneven activity over the top ten actors, with the US, Afghanistan, and Pakistan dominating.

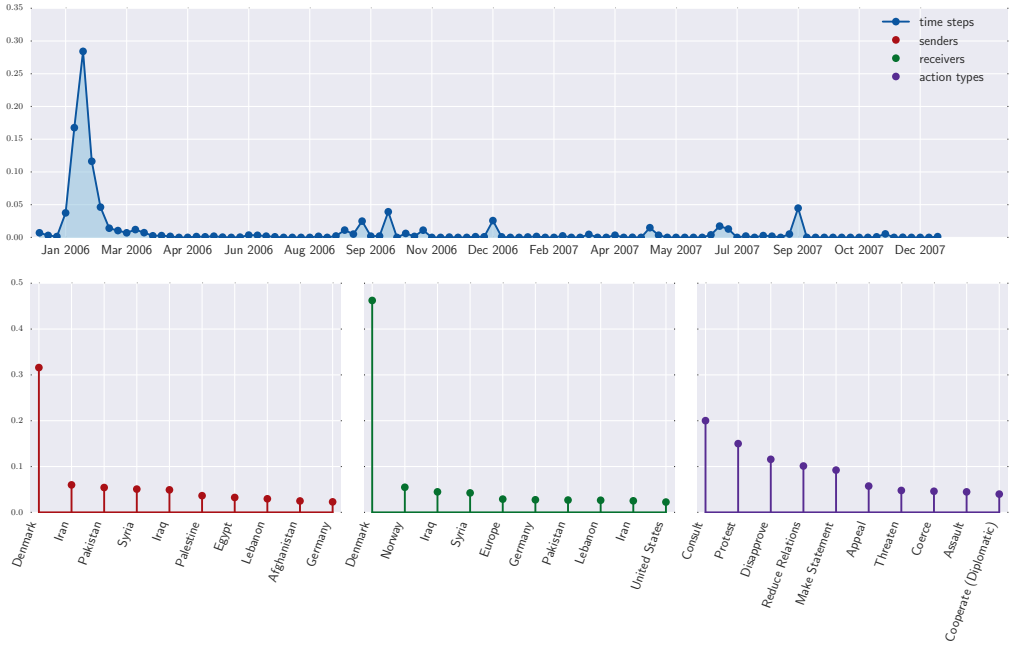
These "regional relations" components conform to our understanding of international affairs and foster confidence in BPTF as an exploratory analysis tool. However, for the same reason, they are also less interesting. To explore temporally localized multilateral relations—i.e., anomalous interaction patterns that do not simply reflect usual activity—we used our model to infer components from several subsets of GDELT and ICEWS, each spanning a two-year date range with weekly time steps. We ranked the inferred components by the sparsity of their time-step factors, measured using the Gini coefficient [Dorfman, 1979]. Ranking components by their Gini coefficients is a form of *anomaly detection*: components with high Gini coefficients have unequal time-step factor values—i.e., dramatic spikes. Fig. 4.3 shows the highest-ranked (i.e., most anomalous) component inferred from a subset of GDELT spanning 2011–2012. This component features an unusual group of top actors and a clear burst of activity around June 2012. To interpret this component, we performed a

Figure 4.4: Regional relations between Central Asian republics and regional superpowers, found in both GDELT (left; spanning 1990 through 2007, with monthly time steps) and ICEWS (right; spanning 1995 through 2012, with monthly time steps).

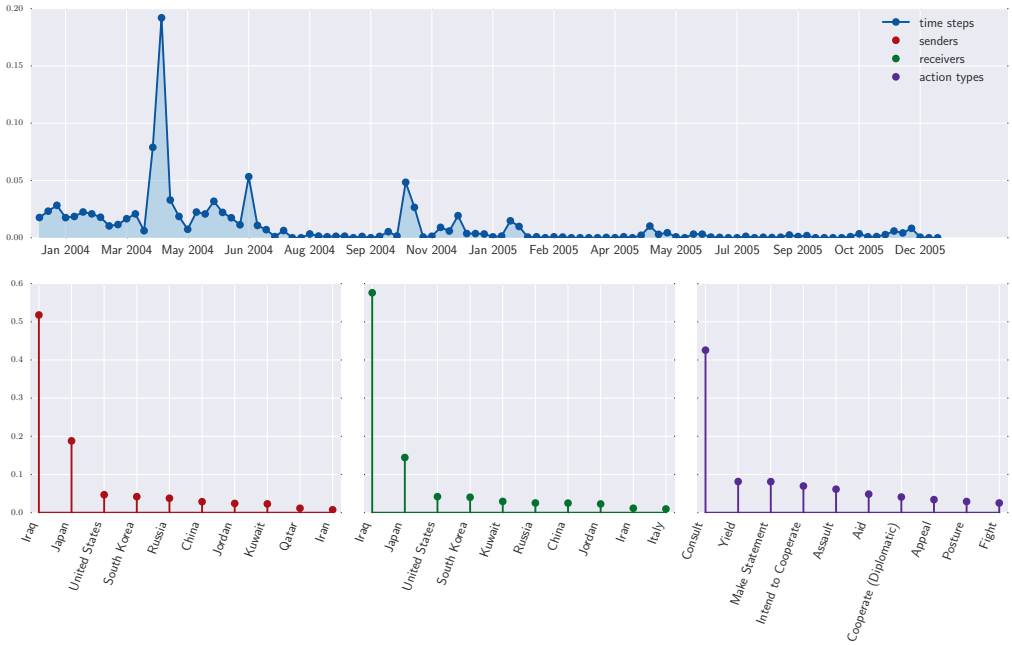
web search for *ecuador UK sweden june 2012* and found that the top hit was a Wikipedia page [Wikipedia contributors, 2018b] about Julian Assange, the editor-in-chief of the website WikiLeaks—an Australian national, wanted by the US and Sweden, who sought political asylum at the Ecuadorian embassy in the UK during June through August 2012. These countries are indeed the top actors for this component, while the time-step factors and top action types (i.e., *Consult*, *Aid*, and *Appeal*) track the dates and nature of the reported events. In general, we found that when our existing knowledge was insufficient to interpret an inferred component, performing a web search for the top two-to-four actors along with the top time step resulted in either a Wikipedia page or a news article that provided an explanation. We present further examples of the most anomalous components inferred from other two-year date ranges in Figs. 4.5a and 4.5b along with the web searches that we performed in order to interpret them.

Figure 4.5: Two anomalous components and their interpretations.

(a) Protests erupted in Islamic countries after a Danish newspaper published cartoons depicting the Prophet Muhammad Wikipedia contributors [2018e]. Denmark and Iran cut diplomatic ties in February 2006 after protesters attacked the Danish embassy in Tehran. This component inferred from GDELT (2006 through 2007, weekly time steps) had the second sparsest time-step factor vector. Web search: *denmark iran january 2006*.



(b) Three Japanese citizens were taken hostage in Iraq during April 2004 and a third was found murdered in October 2004 Wikipedia contributors [2018d]. This component inferred from GDELT (2004 through 2005, weekly time steps) had the sparsest time-step factor vector. We performed a web search for *japan iraq april 2004* to interpret this component.



4.6 Technical Discussion

Previous work on Bayesian Poisson matrix factorization (e.g., Cemgil [2009], Paisley et al. [2014], Gopalan et al. [2015]) presented update equations for the variational parameters in terms of auxiliary variables, known as *latent sources*. In contrast, we write the update equations for Bayesian Poisson tensor factorization in the form of Eqs. (4.6) and (4.7) in order to highlight their relationship to Lee and Seung [1999]'s multiplicative updates for non-negative tensor factorization—a parallel also drawn by Cemgil [2009]—and to show that our update equations suggest a new way of making out-of-sample predictions when using BPTF. In this section, we provide a discussion of these connections and their implications.

When performing NTF by minimizing the generalized KL divergence of reconstruction $\boldsymbol{\mu}$ from observed tensor $\boldsymbol{Y}$ (which is equivalent to MLE for PTF), the multiplicative update equation for, e.g., $\theta^{(1)}_{ki}$ is

$$\theta^{(1)}_{ki} := \theta^{(1)}_{ki} \sum_{j,a,t} \frac{\theta^{(2)}_{kj}\theta^{(3)}_{ka}\theta^{(4)}_{tk}}{\sum_{j,a,t}\theta^{(2)}_{kj}\theta^{(3)}_{ka}\theta^{(4)}_{tk}} \left(\frac{y^{(t)}_{i \xrightarrow{a} j}}{\mu^{(t)}_{i \xrightarrow{a} j}} \right). \tag{4.13}$$

These update equations sometimes converge to locally non-optimal values when the observed tensor is very sparse [Gonzalez and Zhang, 2005, Lin, 2007, Chi and Kolda, 2012]. This problem occurs when parameters are set to *inadmissible zeros*; the algorithm cannot recover from these values due to the multiplicative nature of the update equations. Several solutions have been proposed to correct this behavior when minimizing Euclidean distance. For example, Gillis and Glineur [2008] add a small constant ϵ to each factor to prevent them from ever becoming exactly zero. For KL divergence, Chi and Kolda [2012] proposed an algorithm—Alternating Poisson Regression—that "scooches" parameters away from zero more selectively (i.e., some parameters are still permitted to be zero).

In BPTF, point estimates of the latent parameters are not estimated directly. Instead, variational parameters for each factor, e.g., $\gamma_{ki}^{(1)}$ and $\delta_{ki}^{(1)}$ for factor $\theta_{ki}^{(1)}$, are estimated. These parameters then define a gamma distribution over the factor as in Eq. (4.5), thereby preserving uncertainty about its value. In practice, this approach solves the instability issues suffered by MLE methods, without any efficiency sacrifice. This assertion is supported empirically by the out-of-sample predictive performance results reported in Section 4.4, but can also be verified by comparing the form of the update in Eq. (4.13) with those of the updates in Eqs. (4.6) and (4.7). Specifically, if Eqs. (4.6) and (4.7) are substituted into the expression for the arithmetic expectation of a single latent factor, e.g., $\mathbb{E}\left[\theta_{ki}^{(1)}\right] = \frac{\gamma_{ki}^{(1)}}{\delta_{ki}^{(1)}}$, then the resultant update equation is very similar to the update in Eq. (4.13):

$$\mathbb{E}_Q\left[\theta_{ki}^{(1)}\right] := \frac{\alpha_0 + \sum_{j,a,t} \mathbb{G}_Q\left[\theta_{ki}^{(1)}\theta_{kj}^{(2)}\theta_{ka}^{(3)}\theta_{tk}^{(4)}\right] \frac{y_{i \xrightarrow{a} j}^{(t)}}{\mu_{i \xrightarrow{a} j}^{(t)}}}{\alpha_0\,\beta^{(1)} + \sum_{j,a,t} \mathbb{E}_Q\left[\theta_{kj}^{(2)}\theta_{ka}^{(3)}\theta_{kt}^{(4)}\right]},$$

where $\mu_{i \xrightarrow{a} j}^{(t)} \equiv \sum_{k=1}^{K} \mathbb{G}_Q\left[\theta_{ki}^{(1)}\theta_{kj}^{(2)}\theta_{ka}^{(3)}\theta_{tk}^{(4)}\right]$. Pulling $\mathbb{G}_Q\left[\theta_{ki}^{(1)}\right]$ outside the sum in the numerator and letting $\alpha_0 \to 0$, yields

$$\mathbb{E}_Q\left[\theta_{ki}^{(1)}\right] := \mathbb{G}_Q\left[\theta_{ki}^{(1)}\right] \frac{\sum_{j,a,t} \mathbb{G}_Q\left[\theta_{kj}^{(2)}\theta_{ka}^{(3)}\theta_{tk}^{(4)}\right] \frac{y_{i \xrightarrow{a} j}^{(t)}}{\mu_{i \xrightarrow{a} j}^{(t)}}}{\sum_{j,a,t} \mathbb{E}_Q\left[\theta_{kj}^{(2)}\theta_{ka}^{(3)}\theta_{kt}^{(4)}\right]},$$

which is exactly the form of Eq. (4.13), except that the point estimates of the factors are replaced with two kinds of expectation. This equation makes it clear that the properties that differentiate variational inference for BPTF from the multiplicative updates for PTF are 1) the hyperparameters α_0 and $\beta^{(1)}$ and 2) the use of arithmetic and geometric expectations of the factors instead of direct point estimates.

Since the hyperparameters provide a form of implicit correction, BPTF should not suffer from inadmissible zeros, unlike non-Bayesian PTF. It is also interesting to

Table 4.2: Predictive performance obtained using geometric and arithmetic expectations. (The experimental design was identical to that used to obtain the results in Table 4.1.) Using geometric expectations resulted in the same or better performance than that obtained using arithmetic expectations.

		BPTF-ARI		BPTF-GEO	
	Density	MAE	HAM-Z	MAE	HAM-Z
I-top-25	0.1217	2.03	0.121	**1.99**	**0.113**
G-top-25	0.2638	8.96	0.3	**8.94**	**0.292**
I-top-100	0.0264	0.197	0.0236	**0.178**	**0.0142**
G-top-100	0.0588	1	0.0857	**0.95**	**0.0682**
I-top-25^c	0.0021	0.0104	0.00163	0.0104	0.00161
G-top-25^c	0.0060	0.0414	0.00606	**0.0412**	**0.00601**
I-top100^c	0.0004	0.0011	5.03e-05	0.00109	4.97e-05
G-top100^c	0.0015	0.00804	0.000959	0.00803	0.000957

explore the contribution of the geometric expectations. The fact that each $\mu^{(t)}_{i \xrightarrow{a} j}$ is defined in terms of a geometric expectation suggests that when constructing point estimates of the latent factors from the variational distribution (e.g., for use in prediction), the geometric expectation is more appropriate than the arithmetic expectation (which is commonly used in Bayesian Poisson matrix factorization) since the inference algorithm is implicitly optimizing the reconstruction as defined in terms of geometric expectations of the factors.

To explore the practical differences between geometric and arithmetic expectations of the latent factors under the variational distribution, it is illustrative to consider the form of $\Gamma(\theta; a, b)$. Most relevantly, the gamma distribution is asymmetric, and its mean (i.e., its arithmetic expectation) is greater than its mode. When shape parameter $a \geq 1$, $\text{Mode}(\theta) = \frac{(a-1)}{b}$; when $a < 1$, the mode is undefined, but most of the distribution's probability mass is concentrated near zero—i.e., the PDF increases monotonically as $\theta \to 0$. In this scenario, the gamma distribution's heavy tail pulls the arithmetic mean away from zero and into a region of lower probability.

The geometric expectation is upper-bounded by the arithmetic expectation—i.e., $\mathbb{G}[\theta] = \frac{\exp(\Psi(a))}{b} \leq \frac{a}{b} = \mathbb{E}[\theta]$. Unlike the mode, it is well-defined for $a \in (0, 1)$ and

grows quadratically over this interval, since $\exp(\Psi(a)) \approx \frac{a^2}{2}$ for $a \in (0, 1)$; in contrast, the arithmetic expectation grows linearly over this interval. As a result, when $a < 1$, the geometric expectation yields point estimates that are much closer to zero than those obtained using the arithmetic expectation. When $a \geq 1$, $\exp(\Psi(a)) \approx a - 0.5$ and the geometric expectation is approximately equidistant between the arithmetic expectation and the mode—i.e., $\frac{a}{b} \geq \frac{a-0.5}{b} \geq \frac{a-1}{b}$. These properties are depicted in Fig. 4.6; the key point to take away from this figure is that when $a < 1$, the geometric expectation has a much more probable value than the arithmetic expectation, while when $a \geq 1$, the geometric and arithmetic expectations are very close. This observation suggests that the geometric expectation should yield similar or better point estimates of the latent factors than those obtained using the arithmetic expectation. In Table 4.2, we provide a comparison of the out-of-sample predictive performance for BPTF using arithmetic and geometric expectations. These results suggest that the performance obtained using geometric expectations is either the same as or better than the performance obtained instead using arithmetic expectations.

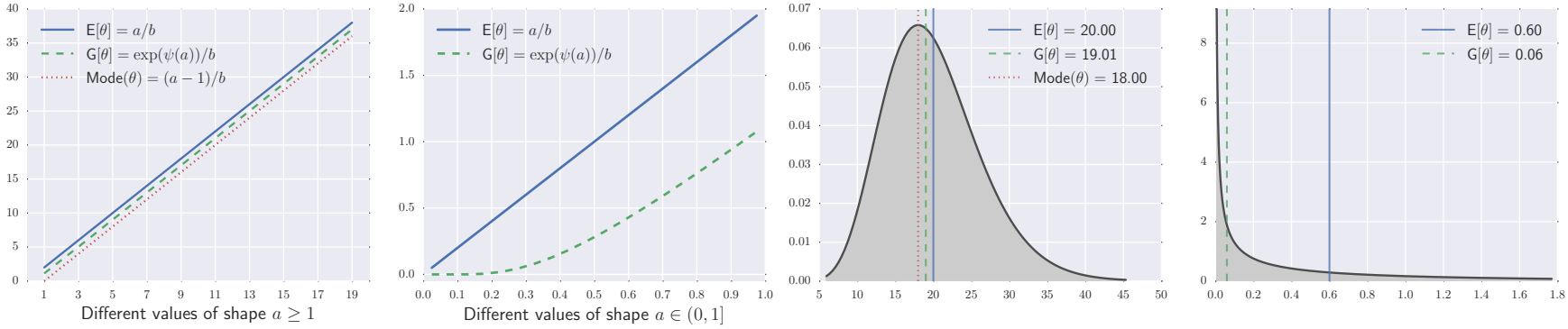


Figure 4.6: The mode, arithmetic expectation, and geometric expectation of a gamma-distributed random variable θ. *First*: The three quantities for different values of shape $a \geq 1$ (x axis) with rate $b = 0.5$. All three grow linearly with a and $\mathrm{E}[\theta] \geq \mathrm{G}[\theta] \geq \mathrm{Mode}(\theta)$. *Second*: Geometric and arithmetic expectations for different values of shape $a \in (0, 1)$, where the mode is undefined, with rate $b = 0.5$. $\mathrm{G}[\theta]$ grows more slowly than $\mathrm{E}[\theta]$. This property is most apparent when $a < 0.4$. *Third*: pdf of a gamma distribution with shape $a = 10$ and rate $b = 0.5$. The three quantities are shown as vertical lines. All three are close in the area of highest density, differing by about a half unit of inverse rate, i.e., $\frac{1}{2b} = 1$. *Fourth*: PDF of a gamma distribution with $a = 0.3$ and $b = 0.5$. The geometric and arithmetic expectations are shown as vertical lines (the mode is undefined). The two quantities differ greatly, with $\mathrm{G}[\theta]$ much closer to zero and in an area of higher density. If these expectations were used as point estimates to predict the presence or absence of a rare event—e.g., $y = 0$ if $\hat{\theta} < 0.5$; otherwise $y = 1$—they would yield different predictions.

CHAPTER 5

BAYESIAN POISSON TUCKER DECOMPOSITION FOR LEARNING THE STRUCTURE OF INTERNATIONAL RELATIONS

Like their inhabitants, countries interact with one another: they consult, negotiate, trade, threaten, and fight. These interactions are seldom uncoordinated. Rather, they are connected by a fabric of *overlapping communities*, such as security coalitions, treaties, trade cartels, and military alliances. For example, OPEC coordinates the petroleum export policies of its thirteen member countries, LAIA fosters trade among Latin American countries, and NATO guarantees collective defense against attacks by external parties. A single country can belong to multiple communities, reflecting its different identities. For example, Venezuela—an oil-producing country and a Latin American country—is a member of both OPEC and LAIA. When Venezuela interacts with other countries, it sometimes does so as an OPEC member and sometimes does so as a LAIA member. Countries engage in both *within-community* and *between-community* interactions. For example, when acting as an OPEC member, Venezuela consults with other OPEC countries, but trades with non-OPEC, oil-importing countries. Moreover, although Venezuela engages in between-community interactions when trading as an OPEC member, it engages in within-community interactions when trading as a LAIA member. To understand or predict how countries interact, we must account for their respective community memberships and how those communities influence their particular actions.

Models based on the CP decomposition, like the one presented in last chapter, require each latent class to jointly summarize information about senders, receivers,

actions, and time steps. This requirement conflates communities of countries with the inferred latent spaces of other modes—e.g.,"topics" of action types—potentially forcing each class to express redundant information. Moreover, by definition, these models cannot express between-community interactions and cannot express sender–receiver asymmetry without learning separate latent parameter matrices for senders and receivers. These limitations make it hard to interpret CP decomposition models as learning latent community memberships.

This chapter is based on work published by Schein et al. [2016b] which presents the APF analogue to Tucker decomposition—i.e., Bayesian Poisson Tucker decomposition (BPTD). The Tucker decomposition (see Section 2.3.4) decomposes a tensor, such as $\boldsymbol{Y}$, into latent parameter matrices that embed each dimension into its own space—e.g., senders and receivers into communities, actions into topics, and time steps into regime. By inferring separate latent structure for each mode, this model yields interpretable latent structure that aligns with well-known concepts in networks analysis when applied to dyadic event data. Exploratory results that demonstrate this are provided in Section 5.6. The Tucker decomposition also makes more efficient use of its parameters than the CP decomposition—Section 5.5 demonstrates empirically that BPTD has better out-of-sample predictive performance than the CP decomposition model presented the last Chapter 4 when both are granted the same number of parameters. BPTD also generalizes many existing block models from the networks community which Section 5.5 includes as baselines. Finally, the Tucker decomposition of the Poisson rate parameter further can be exploited to improve the computational complexity of the allocation step during posterior inference—this leads to an algorithm called *compositional allocation* described in Section 5.3.

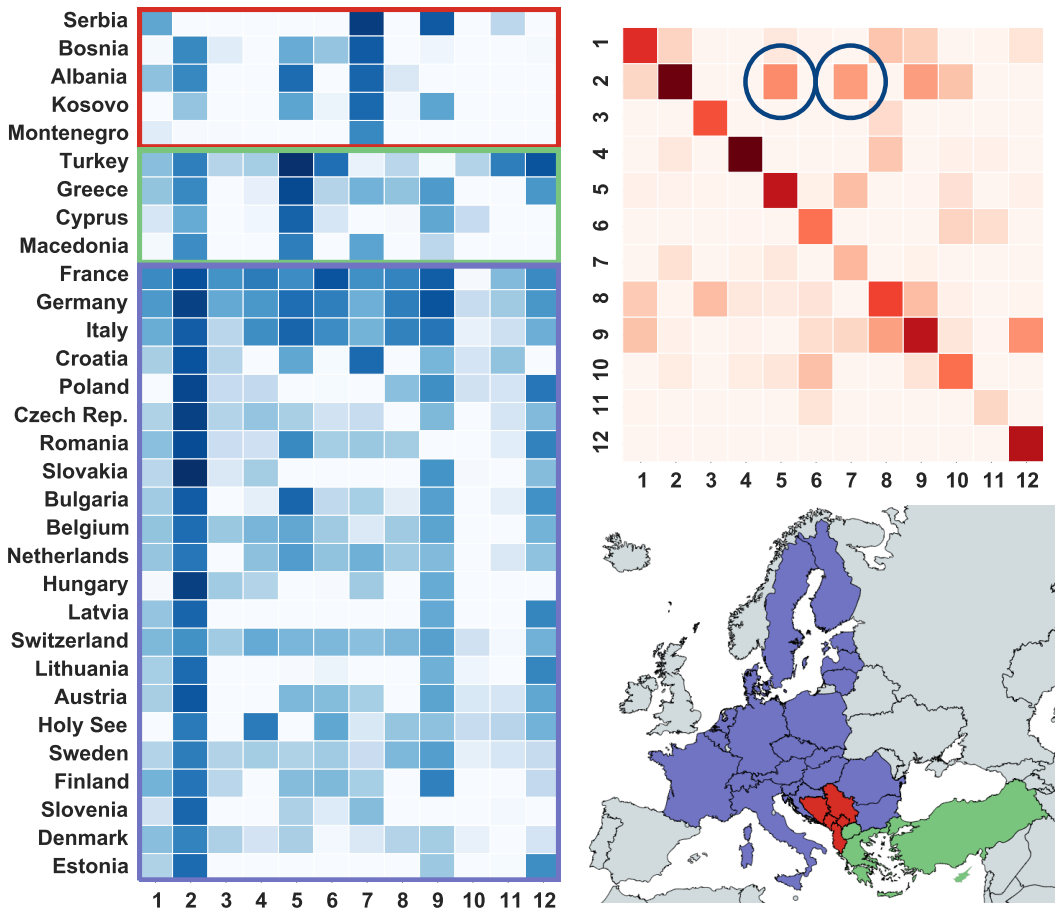


Figure 5.1: Latent structure learned by BPTD from country–country interaction events between 1995 and 2000. *Top right*: A community–community interaction network specific to a single topic of actions and temporal regime. The inferred topic placed most of its mass on the *Intend to Cooperate* and *Consult* actions, so this network represents cooperative community–community interactions. The two strongest between-community interactions (circled) are 2→5 and 2→7. *Left*: Each row depicts the overlapping community memberships for a single country. We show only those countries whose strongest community membership is to either community 2, 5, or 7. We ordered the countries accordingly. Countries strongly associated with community 7 are at highlighted in red; countries associated with community 5 are highlighted in green; and countries associated with community 2 are highlighted in purple. *Bottom right*: Each country is colored according to its strongest community membership. The latent communities have a very strong geographic interpretation.

5.1 Model: Bayesian Poisson Tucker Decomposition

BPTD models each element of the count tensor $\boldsymbol{Y}$ described in Section 4.1 as:

$$y_{i \xrightarrow{a} j}^{(t)} \sim \text{Pois}\left(\sum_{c=1}^{C} \psi_{ic} \sum_{d=1}^{C} \psi_{jd} \sum_{k=1}^{K} \phi_{ak} \sum_{r=1}^{R} \theta_r^{(t)} \lambda_{c \xrightarrow{k} d}^{(r)}\right), \tag{5.1}$$

where ψ_{ic} and ψ_{jd} capture the rates at which countries i and j participate in *communities* c and d, respectively; factor ϕ_{ak} captures the strength of association between action a and *topic* k; and $\theta_r^{(t)}$ captures how well *regime* r explains the events in time step t. We can collectively view the $V \times C$ country–community factors as a latent factor matrix Ψ, where the i^{th} row represents country i's community memberships. Similarly, we can view the $A \times K$ action–topic factors and the $T \times R$ time-step–regime factors as latent parameter matrices Φ and Θ, respectively. Factor $\lambda_{c \xrightarrow{k} d}^{(r)}$ captures the rate at which community c takes actions associated with topic k toward community d during regime r; these factors collectively form a core tensor $\boldsymbol{\Lambda}$ of size $C \times C \times K \times R$ that interacts communities, topics, and regimes. From this perspective, this model corresponds to a Tucker decomposition (see Section 2.3.4) of the observed four-mode tensor into the four parameter matrices and the core tensor

We assume the country–community factors are gamma-distributed,

$$\psi_{ic} \sim \Gamma(\alpha_i, \beta_i)\,, \tag{5.2}$$

where the shape and rate parameters α_i and β_i are specific to country i. We place an uninformative gamma prior over these shape and rate parameters: $\alpha_i, \beta_i \sim \Gamma(\epsilon_0, \epsilon_0)$. This hierarchical prior enables BPTD to express heterogeneity in the countries' rates of activity—e.g., we expect to observe more events involving China than Micronesia.

The action–topic and time-step–regime factors are also gamma-distributed; however, we assume that these factors are drawn from an uninformative gamma prior:

$$\phi_{ak}, \theta_r^{(t)} \sim \Gamma(\epsilon_0, \epsilon_0) \tag{5.3}$$

Because BPTD learns a single embedding of countries into communities, it preserves the traditional network-based notion of community membership. Sender–receiver asymmetry is then captured by the core tensor $\mathbf{\Lambda}$, which can be viewed as a compression of count tensor $\boldsymbol{Y}$. By allowing on-diagonal elements, which we denote by $\lambda^{(r)}_{c \circlearrowright k}$ and off-diagonal elements to be non-zero, the core tensor can represent both within- and between-community interactions. The elements of the core tensor are gamma-distributed,

$$\lambda^{(r)}_{c \circlearrowright k} \sim \Gamma\big(\eta^{\circlearrowright}_c \, \eta^{\leftrightarrow}_c \, \nu_k \, \rho_r, \delta\big) \tag{5.4}$$

$$\lambda^{(r)}_{c \xrightarrow{k} d} \sim \Gamma(\eta^{\leftrightarrow}_c \, \eta^{\leftrightarrow}_d \, \nu_k \, \rho_r, \delta) \;\text{ for } c \neq d. \tag{5.5}$$

Each community $c \in [C]$—where $[C] \equiv \{1, \ldots, C\}$—has two positive weights $\eta^{\circlearrowright}_c$ and $\eta^{\leftrightarrow}_c$ that capture its rates of within- and between-community interaction, respectively. Each topic $k \in [K]$ has a positive weight ν_k, while each regime $r \in [R]$ has a positive weight ρ_r. We place an uninformative prior over the within-community interaction rates and gamma shrinkage priors over the other weights: $\eta^{\circlearrowright}_c \sim \Gamma(\epsilon_0, \epsilon_0)$, $\eta^{\leftrightarrow}_c \sim \Gamma(\gamma_0 \,/\, C, \xi)$, $\nu_k \sim \Gamma(\gamma_0 \,/\, K, \xi)$, and $\rho_r \sim \Gamma(\gamma_0 \,/\, R, \xi)$. These priors bias BPTD toward inferring parsimonious latent structure. Finally, we assume that δ and ξ are drawn from an uninformative gamma prior: $\delta, \xi \sim \Gamma(\epsilon_0, \epsilon_0)$.

BPTD as a measurement model for international relations data. An advantage of the Tucker over the CP decomposition is the decoupling of the latent cardinalities of each factor matrix. This allows for the interpretation of each factor matrix as independently embedding the entities of that mode into their own latent space. As a measurement model, BPTD can be understood as measuring static structures in *dynamic multinetworks*—i.e., time-evolving networks with multiple edge types. These static structures are *communities* of actors, *topics* of action types, and network *regimes* which are characterized by the interactions of communities and topics. The presence of the core tensor in the Tucker decomposition also allows the

model to learn a *single* country–community embedding matrix Ψ while still expressing sender–receiver asymmetry in the data; in other words, the Poisson rates for the two counts $y^{(t)}_{i \xrightarrow{a} j}$ and $y^{(t)}_{j \xrightarrow{a} i}$ are not necessarily equal, so long as the core tensor is asymmetric in its community–community interactions—i.e., $\lambda^{(r)}_{c \xrightarrow{k} d} \neq \lambda^{(r)}_{d \xrightarrow{k} c}$. Contrast this with CP decomposition (Chapter 4), in which two separate parameter matrices that embed actors as senders and as receivers—i.e., $\Psi^{(\rightarrow)}$, $\Psi^{(\leftarrow)}$—are required to express sender–receiver asymmetry, thus straining their interpretation as measuring *communities*. For a given regime r and topic k the corresponding $C \times C$ slice of the core tensor $\Lambda^{(r)}_{: \xrightarrow{k} :}$ at what rates the communities interact with one another; we visualize one such slice along with country–community factor matrix Ψ in Fig. 5.1.

5.2 Connections to Previous Work

Poisson CP decomposition and latent class models: Recall that we can represent a data set of dyadic events as a set of N event tokens, where a single token $\boldsymbol{e}_n = (i \xrightarrow{a} j, t)$ indicates that sender country $i \in [V]$ took action $a \in [A]$ toward receiver country $j \in [V]$ during time step $t \in [T]$. DuBois and Smyth [2010] developed a model that assigns each event token (ignoring time steps) to one of Q latent classes, where each class $q \in [Q]$ is characterized by three categorical distributions—$\boldsymbol{\theta}^{\rightarrow}_q$ over senders, $\boldsymbol{\theta}^{\leftarrow}_q$ over receivers, and $\boldsymbol{\phi}_q$ over actions—i.e.,

$$P(\boldsymbol{e}_n = (i \xrightarrow{a} j, t) \mid z_n = q) = \theta^{(\rightarrow)}_{iq} \, \theta^{(\leftarrow)}_{jq} \, \phi_{aq}. \tag{5.6}$$

Due to the connection between categorical allocation and Poisson factorization (see Section 3.2.2), this model corresponds to the 3-mode CP decomposition model given as an example in Section 3.4; its four-mode generalization (including time step) presented in the last chapter can be rewritten (using this chapter's notation) as

$$y^{(t)}_{i \xrightarrow{a} j} \sim \text{Pois}\left(\sum_{q=1}^{Q} \theta^{(\rightarrow)}_{iq} \, \theta^{(\leftarrow)}_{jq} \, \phi_{aq} \, \psi_{tq} \right), \tag{5.7}$$

where the corresponding allocation step may be written,

$$P(\boldsymbol{e}_n = (i \xrightarrow{a} j, t) \mid z_n = q) \propto \theta_{iq}^{(\rightarrow)} \, \theta_{jq}^{(\leftarrow)} \, \phi_{aq} \, \psi_{tq}. \tag{5.8}$$

We may also add a per-class positive weight λ_q,

$$y_{i \xrightarrow{a} j}^{(t)} \sim \text{Pois}\left(\sum_{q=1}^{Q} \theta_{iq}^{(\rightarrow)} \, \theta_{jq}^{(\leftarrow)} \, \phi_{aq} \, \psi_{tq} \, \lambda_q. \right) \tag{5.9}$$

Tucker decomposition is equivalent to CP decomposition when the cardinalities of the latent dimensions are equal and the off-diagonal elements of the core tensor are zero. DuBois and Smyth's and Schein et al.'s models therefore constitute a constrained special case of BPTD that cannot capture dimension-specific structure, such as topics of actions or communities of countries that engage in between-community interactions.

Infinite relational models: The infinite relational model (IRM) of Kemp et al. [2006] also learns latent structure specific to each dimension of an M-mode tensor; however, unlike BPTD, the elements of this tensor are binary, indicating the presence or absence of the corresponding event type. The IRM therefore uses a Bernoulli likelihood. Schmidt and Mørup [2013] extended the IRM to model a tensor of event counts by replacing the Bernoulli likelihood with a Poisson likelihood (and gamma priors):

$$y_{i \xrightarrow{a} j}^{(t)} \sim \text{Pois}\left(\lambda_{z_i \xrightarrow{z_a} z_j}^{(z_t)} \right), \tag{5.10}$$

where $z_i, z_j \in [C]$ are the respective community assignments of countries i and j, $z_a \in [K]$ is the topic assignment of action a, and $z_t \in [R]$ is the regime assignment of time step t. This model, which we refer to as the gamma–Poisson IRM (GPIRM), allocates M-dimensional event types to M-dimensional latent classes—e.g., it allocates all tokens of type $(i \xrightarrow{a} j, t)$ to class $(z_i \xrightarrow{z_a} z_j, z_t)$.

The GPIRM is a special case of BPTD, in which the rows of the latent parameter matrices are constrained to be "one-hot" binary vectors—i.e., $\psi_{ic} = \mathbb{1}(z_i = c)$, $\psi_{jd} =$

$\mathbb{1}(z_j = d)$, $\phi_{ak} = \mathbb{1}(z_a = k)$, and $\theta_r^{(t)} = \mathbb{1}(z_t = r)$. With this constraint, the Poisson rate in Eq. (5.1) is equal to the Poisson rate in Eq. (5.10). Unlike BPTD, the GPIRM is a single-membership model. In addition, it cannot express heterogeneity in the countries' rates of activity. The latter limitation can be remedied by allowing θ_{iz_i} and θ_{jz_j} to be positive real numbers. We refer to this variant of the GPIRM as the degree-corrected GPIRM (DCGPIRM).

Stochastic block models: The IRM itself generalizes the stochastic block model (SBM) of Nowicki and Snijders [2001], which learns latent structure from binary networks. Although the SBM was originally specified using a Bernoulli likelihood, Karrer and Newman [2011] introduced an alternative that uses the Poisson likelihood:

$$y_{i \to j} \sim \text{Pois}\left(\sum_{c=1}^{C} \psi_{ic} \sum_{d=1}^{C} \psi_{jd}\, \lambda_{c \to d}\right), \tag{5.11}$$

where $\psi_{ic} = \mathbb{1}(z_i = c)$, $\theta_j = \mathbb{1}(z_j = d)$, and $\lambda_{c \to d}$ is a positive real number. Like the IRM and the GPIRM, the SBM is a single-membership model and cannot express heterogeneity in the countries' rates of activity. Airoldi et al. [2008] addressed the former limitation by letting $\psi_{ic} \in [0, 1]$ such that $\sum_{c=1}^{C} \psi_{ic} = 1$. Meanwhile, Karrer and Newman [2011] addressed the latter limitation by allowing both θ_{iz_i} and θ_{jz_j} to be positive real numbers, much like the DCGPIRM. Ball et al. [2011] simultaneously addressed both limitations by letting $\psi_{ic}, \psi_{jd} \geq 0$, but constrained $\lambda_{c \to d} = \lambda_{d \to c}$. Finally, Zhou [2015] extended Ball et al.'s model to be nonparametric and introduced the Poisson–Bernoulli distribution to link binary data to the Poisson likelihood in a principled fashion. In this model, the elements of the core matrix and their corresponding factors constitute a draw from a relational gamma process.

Non-Poisson Tucker decomposition: Researchers sometimes refer to the Poisson rate in Eq. (5.11) as being "bilinear" because it can equivalently be written as $\boldsymbol{\theta}_j \boldsymbol{\Lambda} \boldsymbol{\theta}_i^\top$. Nickel et al. [2012] introduced RESCAL—a non-probabilistic bilinear model for binary data that achieves state-of-the-art performance at relation extrac-

tion. Nickel et al. [2015] then introduced several extensions for extracting relations of different types. Bilinear models, such as RESCAL and its extensions, are all special cases (albeit non-probabilistic ones) of Tucker decomposition. Hoff et al. [2016] developed a model based on the Tucker decomposition for analyzing dyadic event data. This model uses a Gaussian likelihood and thus does not naturally yield an inference algorithm that takes advantage of the sparsity of the data. Finally, there are many other Tucker decomposition methods [Kolda and Bader, 2009] including nonparametric [Xu et al., 2012] and nonnegative variants [Kim and Choi, 2007, Mørup et al., 2008, Cichocki et al., 2009].

5.3 MCMC Inference

Complete conditionals for all latent variables are available in closed form. We use augment-and-conquer schemes (see Section 3.5) to obtain the complete conditionals for hierarchical gamma random variables—e.g., for $\eta_c^{\circlearrowleft}$ and $\eta_c^{\leftrightarrow}$. As usual, the complete conditionals depend on the latent sources—in this case, $y^{(tr)}_{ic \xrightarrow{ak} jd}$—which are re-sampled in the allocation step. The structure of the Tucker decomposition may be exploited to improve the complexity of this step, resulting in an algorithm called *compositional allocation*.

5.3.1 Compositional allocation

The latent source representation (see Section 3.2) of this model is:

$$y^{(tr)}_{ic \xrightarrow{ak} jd} \sim \text{Pois}\Big(\psi_{ic}\,\psi_{jd}\,\phi_{ak}\,\theta^{(t)}_r\,\lambda^{(r)}_{k \xrightarrow{} d}\Big) \tag{5.12}$$

$$y^{(t)}_{i \xrightarrow{a} j} = \sum_{c=1}^{C}\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{r=1}^{R} y^{(tr)}_{ic \xrightarrow{ak} jd}. \tag{5.13}$$

where the count for type $(i \xrightarrow{a} j, t)$ is thinned across latent classes. Unlike in CP decomposition, where latent classes are identified by a single index, here latent classes

are *compositional*—i.e., they are identified by a multi-index representing a unique combination of a sender community, receiver community, action topic, and regime e.g., $(c \xrightarrow{k} d, r)$. We can view each latent source in terms of its token representation,

$$y^{(tr)}_{ic \xrightarrow{ak} jd} = \sum_{n=1}^{N} \mathbb{1}\left[\boldsymbol{e}_n = (i \xrightarrow{a} j, t)\right] \mathbb{1}\left[\boldsymbol{z}_n = (c \xrightarrow{k} d, r)\right], \tag{5.14}$$

The complete conditional for each token's class assignment is categorical:

$$P\left(\boldsymbol{z}_n = (c \xrightarrow{k} d, r) \mid \boldsymbol{e}_n = (i \xrightarrow{a} j, t), -\right) \propto \psi_{ic}\, \psi_{jd}\, \phi_{ak}\, \theta^{(t)}_r\, \lambda^{(r)}_{c \xrightarrow{k} d} \tag{5.15}$$

where the main computational bottleneck is computing the normalizing constant:

$$Z^{(t)}_{i \xrightarrow{a} j} \triangleq \sum_{c=1}^{C} \sum_{d=1}^{C} \sum_{k=1}^{K} \sum_{r=1}^{R} \psi_{ic}\, \psi_{jd}\, \phi_{ak}\, \theta^{(t)}_r\, \lambda^{(r)}_{c \xrightarrow{k} d} \tag{5.16}$$

The naïve implementation enumerates all $C \cdot C \cdot K \cdot R$ summands—i.e., one for each latent class—and performs 4 multiplications for each. For a general M-mode Tucker decomposition, the naïve implementation thus involves $\mathcal{O}(M \prod_{m=1}^{M} L_m)$ operations where L_m is the cardinality of the m^{th} mode. However, since the Tucker decomposition is *multilinear*, we may push sums in to suggest a more efficient dynamic program[1] for computing the normalizing constant. Note that each sum is itself a normalizing constant for a different categorical probability:

$$Z^{(t)}_{i \xrightarrow{a} j} = \sum_{c=1}^{C} \psi_{ic} \underbrace{\sum_{d=1}^{C} \psi_{jd} \underbrace{\sum_{k=1}^{K} \theta_{ak} \underbrace{\sum_{r=1}^{R} \theta^{(t)}_r\, \lambda^{(r)}_{c \xrightarrow{k} d}}_{=Z^{(t)}_{c \xrightarrow{k} d}}}_{=Z^{(t)}_{c \xrightarrow{a} d}}}_{=Z^{(t)}_{c \xrightarrow{a} j}}, \tag{5.17}$$

[1] This algorithm can be understood as a form of "variable elimination" [Koller et al., 2009].

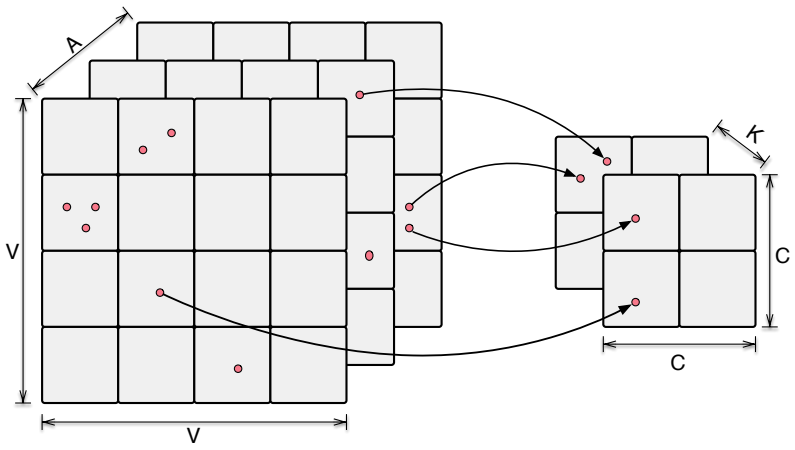


Figure 5.2: Compositional allocation. For clarity, we show the allocation process for a three-mode count tensor (ignoring time steps). Observed three-dimensional event tokens (left) are compositionally allocated to three-dimensional latent classes (right).

where for instance, $Z^{(t)}_{c \xrightarrow{a} j}$ is the normalizing constant for:

$$P\left(\boldsymbol{z}_{n\,2:} = (\xrightarrow{k} d, r) \,|\, z_{n1} = c, \boldsymbol{e}_n = = (i \xrightarrow{a} j, t)\right) = \frac{\psi_{jd}\, \phi_{ak}\, \theta^{(t)}_r\, \lambda^{(r)}_{c \xrightarrow{k} d}}{Z^{(t)}_{c \xrightarrow{a} j}} \tag{5.18}$$

where $\boldsymbol{z}_{n\,2:} \equiv (z_{n2}, z_{n3}, z_{n4})$ denotes the last three indices of the multi-index for the latent class assignment of the n^{th} token. This categorical probability is the probability of allocating the receiver community to d, the action type to topic k and the time step to regime r given that the sender was already allocated to community c. Compositional allocation is a dynamic program that caches the intermediate marginal probabilities along the way to computing the full probability in Eq. (5.17); it then allocates each per-mode indicator conditioned on the previous ones—i.e., $P(\boldsymbol{z}_n \,|\, -) = \prod_{m=1}^{M} P\left(z_{nm} \,|\, \boldsymbol{z}_{n:m}, -\right)$. In the general M-mode case, this reduces the computational complexity by a factor of M, down to $\mathcal{O}\left(\prod_{m=1}^{M} L_m\right)$. A graphical illustration of compositional allocation is provided in Fig. 5.2.

5.3.2 Complete conditionals

We assume that $\boldsymbol{Y}$ is partially observed and include a binary mask $\boldsymbol{B}$, where $b^{(t)}_{i \xrightarrow{a} j} = 0$ means that $y^{(t)}_{i \xrightarrow{a} j} = 0$ is heldout, not an observed zero. The complete conditionals for the first-level parameters—i.e., ψ_{ic}, ϕ_{ak}, $\theta^{(t)}_r$, and $\lambda^{(r)}_{c \xrightarrow{k} d}$—follow from previous derivations. These involve computing latent source aggregations and their corresponding constants.

Country–Community Factors:

$$
\begin{aligned}
y^{(\cdot)}_{ic \dot{\leftrightarrow} \cdot} &\triangleq \sum_{j \neq i}^{C} \sum_{d=1}^{A} \sum_{a=1}^{K} \sum_{k=1}^{T} \sum_{t=1}^{R} \sum_{r=1} \left(y^{(tr)}_{ic \xrightarrow{ak} dj} + y^{(tr)}_{jd \xrightarrow{ak} ci} \right) \\
\zeta_{ic} &\triangleq \sum_{j \neq i}^{A} \sum_{a=1}^{T} \sum_{t=1}^{C} \sum_{d=1} \psi_{jd} \sum_{k=1}^{K} \phi_{ak} \sum_{r=1}^{R} \theta^{(t)}_r \left(b^{(t)}_{i \xrightarrow{a} j} \lambda^{(r)}_{c \xrightarrow{k} d} + b^{(t)}_{j \xrightarrow{a} i} \lambda^{(r)}_{d \xrightarrow{k} c} \right) \\
(\psi_{ic} \mid -) &\sim \Gamma \left(\alpha_i + y^{(\cdot)}_{ic \dot{\leftrightarrow} \cdot}, \beta_i + \zeta_{ic} \right)
\end{aligned}
$$

Action–Topic Factors:

$$
\begin{aligned}
y^{(\cdot)}_{\cdot \overset{ak}{\leftrightarrow} \cdot} &\triangleq \sum_{i=1}^{V} \sum_{c=1}^{C} \sum_{j \neq i}^{C} \sum_{d=1}^{T} \sum_{t=1}^{R} \sum_{r=1} y^{(tr)}_{ic \xrightarrow{ak} dj} \\
\zeta_{ak} &\triangleq \sum_{i=1}^{V} \sum_{j \neq i}^{T} \sum_{t=1} b^{(t)}_{i \xrightarrow{a} j} \sum_{c=1}^{C} \psi_{ic} \sum_{d=1}^{C} \psi_{jd} \sum_{r=1}^{R} \theta^{(t)}_r \lambda^{(r)}_{c \xrightarrow{k} d} \\
(\phi_{ak} \mid -) &\sim \Gamma \left(\epsilon_0 + y^{(\cdot)}_{\cdot \overset{ak}{\leftrightarrow} \cdot}, \epsilon_0 + \zeta_{ak} \right)
\end{aligned}
$$

Time-Step–Regime Factors:

$$
\begin{aligned}
y^{(tr)}_{\cdot \dot{\rightarrow} \cdot} &\triangleq \sum_{i=1}^{V} \sum_{c=1}^{C} \sum_{j \neq i}^{C} \sum_{d=1}^{A} \sum_{a=1}^{K} \sum_{k=1} y^{(tr)}_{ic \xrightarrow{ak} dj} \\
\zeta_{tr} &\triangleq \sum_{i=1}^{V} \sum_{j \neq i}^{A} \sum_{a=1} b^{(t)}_{i \xrightarrow{a} j} \sum_{c=1}^{C} \psi_{ic} \sum_{d=1}^{C} \psi_{jd} \sum_{k=1}^{K} \phi_{ak} \lambda^{(r)}_{c \xrightarrow{k} d} \\
(\theta^{(t)}_r \mid -) &\sim \Gamma \left(\epsilon_0 + y^{(tr)}_{\cdot \dot{\rightarrow} \cdot}, \epsilon_0 + \zeta_{tr} \right)
\end{aligned}
$$

Diagonal Elements of the Core Tensor:

$$\begin{aligned}
\omega^{(r)}_{c\circlearrowright k} &\triangleq \eta^{\circlearrowright}_{c}\eta^{\leftrightarrow}_{c}\nu_k\rho_r \\
y^{(r)}_{c\circlearrowright k} &\triangleq \sum_{i=1}^{V}\sum_{j\neq i}\sum_{a=1}^{A}\sum_{t=1}^{T} y^{(tr)}_{ic\xrightarrow{ak}cj} \\
\zeta^{(r)}_{c\circlearrowright k} &\triangleq \sum_{i=1}^{V}\psi_{ic}\sum_{j\neq i}\theta_{jc}\sum_{a=1}^{A}\phi_{ak}\sum_{t=1}^{T}\theta^{(t)}_{r}\, b^{(t)}_{i\xrightarrow{a}j} \\
\left(\lambda^{(r)}_{c\circlearrowright k}\,|\,-\right) &\sim \Gamma\left(\omega^{(r)}_{c\circlearrowright k} + y^{(r)}_{c\circlearrowright k}, \delta + \zeta^{(r)}_{c\circlearrowright k}\right)
\end{aligned}$$

Off-Diagonal Elements of the Core Tensor:

$$\begin{aligned}
\omega^{(r)}_{c\xrightarrow{k}d} &\triangleq \eta^{\leftrightarrow}_{c}\eta^{\leftrightarrow}_{d}\nu_k\rho_r && c\neq d \\
y^{(r)}_{c\xrightarrow{k}d} &\triangleq \sum_{i=1}^{V}\sum_{j\neq i}\sum_{a=1}^{A}\sum_{t=1}^{T} y^{(tr)}_{ic\xrightarrow{ak}dj} && c\neq d \\
\zeta^{(r)}_{c\xrightarrow{k}d} &\triangleq \sum_{i=1}^{V}\psi_{ic}\sum_{j\neq i}\psi_{jd}\sum_{a=1}^{A}\phi_{ak}\sum_{t=1}^{T}\theta^{(t)}_{r}\, b^{(t)}_{i\xrightarrow{a}j} && c\neq d \\
\left(\lambda^{(r)}_{c\xrightarrow{k}d}\,|\,-\right) &\sim \Gamma\left(\omega^{(r)}_{c\xrightarrow{k}d} + + y^{(r)}_{c\xrightarrow{k}d}, \delta + \zeta^{(r)}_{c\xrightarrow{k}d}\right) && c\neq d
\end{aligned}$$

Complete conditionals for the hierarchical priors over gamma rate parameters are available via gamma–gamma conjugacy.

Per-Country Rate Parameters:

$$(\beta_i\,|\,-) \sim \Gamma\left(\epsilon_0 + C\alpha_i, \epsilon_0 + \sum_{c=1}^{C}\psi_{ic}\right)$$

Core Rate Parameter:

$$(\delta\,|\,-) \sim \Gamma\left(\epsilon_0 + \omega^{(\cdot)}_{\cdot\leftrightarrow\cdot}, \epsilon_0 + \lambda^{(\cdot)}_{\cdot\leftrightarrow\cdot}\right)$$

Weights Rate Parameter:

$$(\xi\,|\,-) \sim \Gamma\left(\epsilon_0 + 4\gamma_0, \epsilon_0 + \sum_{c=1}^{C}\eta^{\circlearrowright}_{c} + \sum_{c=1}^{C}\eta^{\leftrightarrow}_{c} + \sum_{k=1}^{K}\nu_k + \sum_{r=1}^{R}\rho_r\right)$$

The complete conditionals for the hierarchical priors over gamma shape parameters are available using augment-and-conquer schemes (see Section 3.5). These rely on the latent source aggregations and corresponding constants defined above and introduce auxiliary count random variables whose complete conditionals follow the Chinese Restaurant table distribution. Throughout this section, we use $g(x) \triangleq \ln(1+x)$.

Auxiliary Latent Country–Community Counts:

$$(\ell_{ic} \,|\, -) \sim \mathrm{CRT}\left(y^{(\cdot)}_{ic\leftrightarrow\cdot}, \alpha_i\right)$$

Per-Country Shape Parameters:

$$(\alpha_i \,|\, -) \sim \Gamma\left(\epsilon_0 + \sum_{c=1}^{C} \ell_{ic}, \epsilon_0 + \sum_{c=1}^{C} g\left(\zeta_{ic}\, {\beta_i}^{-1}\right)\right)$$

Diagonal Auxiliary Latent Core Counts:

$$\ell^{(r)}_{c\circlearrowright^k} \sim \mathrm{CRT}\left(y^{(r)}_{c\circlearrowright^k}, \omega^{(r)}_{c\circlearrowright^k}\right)$$

Off-Diagonal Auxiliary Latent Core Counts:

$$\ell^{(r)}_{\substack{k\\ c\longrightarrow d}} \sim \mathrm{CRT}\left(y^{(r)}_{\substack{k\\ c\longrightarrow d}}, \omega^{(r)}_{\substack{k\\ c\longrightarrow d}}\right) \qquad\qquad c \neq d$$

Within-Community Weights:

$$\ell^{(\cdot)}_{c\circlearrowright^\cdot} \triangleq \sum_{k=1}^{K}\sum_{r=1}^{R} \ell^{(r)}_{c\circlearrowright^k}$$

$$\zeta^{\circlearrowright}_c \triangleq \sum_{r=1}^{R} \rho_r \sum_{k=1}^{K} \nu_k \sum_{d\neq c} \eta^{\leftrightarrow}_d \left(g\left(\zeta^{(r)}_{\substack{k\\ c\longrightarrow d}}\, \delta^{-1}\right) + g\left(\zeta^{(r)}_{\substack{k\\ d\longrightarrow c}}\, \delta^{-1}\right)\right)$$

$$(\eta^{\circlearrowright}_c \,|\, -) \sim \Gamma\left(\frac{\gamma_0}{C} + \ell^{(\cdot)}_{c\circlearrowright^\cdot}, \xi + \zeta^{\circlearrowright}_c\right)$$

Between-Community Weights:

$$\ell^{(\cdot)}_{c\overset{\cdot}{\leftrightarrow}\cdot} \triangleq \ell^{(\cdot)}_{c\circlearrowright\cdot} + \sum_{d\neq c}\sum_{k=1}^{K}\sum_{r=1}^{R}\left(\ell^{(r)}_{c\overset{k}{\rightarrow}d} + \ell^{(r)}_{d\overset{k}{\rightarrow}c}\right)$$

$$\zeta^{\leftrightarrow}_{c} \triangleq \sum_{r=1}^{R}\rho_r\sum_{k=1}^{K}\nu_k\left[\eta^{\circlearrowright}_{c}\, g\left(\zeta^{(r)}_{c\circlearrowright k}\,\delta^{-1}\right) + \sum_{d\neq c}\eta^{\leftrightarrow}_{d}\left(g\left(\zeta^{(r)}_{c\overset{k}{\rightarrow}d}\,\delta^{-1}\right) + g\left(\zeta^{(r)}_{d\overset{k}{\rightarrow}c}\,\delta^{-1}\right)\right)\right]$$

$$(\eta^{\leftrightarrow}_{c}\,|\,-) \sim \Gamma\left(\frac{\gamma_0}{C} + \ell^{(\cdot)}_{c\overset{\cdot}{\leftrightarrow}\cdot}, \xi + \zeta^{\leftrightarrow}_{c}\right)$$

Topic Weights:

$$\ell^{(\cdot)}_{\cdot\overset{k}{\rightarrow}\cdot} \triangleq \sum_{c=1}^{C}\sum_{d=1}^{C}\sum_{r=1}^{R}\ell^{(r)}_{c\overset{k}{\rightarrow}d}$$

$$\zeta_{k} \triangleq \sum_{r=1}^{R}\rho_r\sum_{c=1}^{C}\eta^{\leftrightarrow}_{c}\left[\eta^{\circlearrowright}_{c}\, g\left(\zeta^{(r)}_{c\circlearrowright k}\,\delta^{-1}\right) + \sum_{d\neq c}\eta^{\leftrightarrow}_{d}\left(g\left(\zeta^{(r)}_{c\overset{k}{\rightarrow}d}\,\delta^{-1}\right) + g\left(\zeta^{(r)}_{d\overset{k}{\rightarrow}c}\,\delta^{-1}\right)\right)\right]$$

$$(\nu_{k}\,|\,-) \sim \Gamma\left(\frac{\gamma_0}{K} + \ell^{(\cdot)}_{\cdot\overset{k}{\rightarrow}\cdot}, \xi + \zeta_{k}\right)$$

Regime Weights:

$$\ell^{(r)}_{\cdot\overset{\cdot}{\rightarrow}\cdot} \triangleq \sum_{c=1}^{C}\sum_{d=1}^{C}\sum_{k=1}^{K}\ell^{(r)}_{c\overset{k}{\rightarrow}d}$$

$$\zeta_{r} \triangleq \sum_{k=1}^{K}\nu_k\sum_{c=1}^{C}\eta^{\leftrightarrow}_{c}\left[\eta^{\circlearrowright}_{c}\, g\left(\zeta^{(r)}_{c\circlearrowright k}\,\delta^{-1}\right) + \sum_{d\neq c}\eta^{\leftrightarrow}_{d}\left(g\left(\zeta^{(r)}_{c\overset{k}{\rightarrow}d}\,\delta^{-1}\right) + g\left(\zeta^{(r)}_{d\overset{k}{\rightarrow}c}\,\delta^{-1}\right)\right)\right]$$

$$(\rho_{r}\,|\,-) \sim \Gamma\left(\frac{\gamma_0}{R} + \ell^{(r)}_{\cdot\overset{\cdot}{\rightarrow}\cdot}, \xi + \zeta_{r}\right)$$

5.4 International relations dyadic event data

We used data from the Integrated Crisis Early Warning System (ICEWS) [Boschee et al., 2015] and the Global Database of Events, Language, and Tone (GDELT) [Leetaru and Schrodt, 2013]. ICEWS and GDELT both use the Conflict and Mediation Event Observations (CAMEO) hierarchy [Gerner et al.] for senders, receivers, and actions. The top level for actions, which we use in our analyses, consists of twenty action classes, roughly ranked according to their overall sentiment. For example, the most negative is 20—Use Unconventional Mass Violence. CAMEO further

divides these actions into the QuadClass scheme: Verbal Cooperation (actions 2–5), Material Cooperation (actions 6–7), Verbal Conflict (actions 8–16), and Material Conflict (16–20). The first action (1—Make Statement) is neutral.

5.5 Predictive Analysis

Baseline models: We compared BPTD's predictive performance to that of three baseline models, described in Section 5.2: 1) GPIRM, 2) DCGPIRM, and 3) the Bayesian Poisson tensor factorization (BPTF) model of Schein et al. [2015]. All three models use a Poisson likelihood and have the same two hyperparameters as BPTD—i.e., ϵ_0 and γ_0. We set $\epsilon_0 = 0.1$ and γ_0 so that $(\gamma_0 \,/\, C)^2 \,(\gamma_0 \,/\, K)\,(\gamma_0 \,/\, R) = 0.01$. This parameterization encourages shrinkage among the elements of the core tensor $\mathbf{\Lambda}$. We implemented an MCMC inference algorithm for each model.

GPIRM and DCGPIRM are both Tucker decomposition models and thus allocate events to four-dimensional latent classes. The cardinalities of these latent dimensions are the same as BPTD's—i.e., C, K, and R. In contrast, BPTF is a CP decomposition model and thus allocates events to one-dimensional latent classes. We set the cardinality of this dimension so that the total number of latent paramters in BPTF was equal to the total number of latent paramters in BPTD—i.e., $Q = \lceil \frac{(V \times C)+(A \times K)+(T \times R)+(C^2 \times K \times R)}{V+V+A+T+1} \rceil$. We could have alternatively set BPTF and BPTD to use the same number of latent classes—i.e., $Q = C^2 \times K \times R$—however, CP decomposition models tend to overfit when Q is large [Zhao et al., 2015]. Throughout our predictive experiments, we let $C = 25$, $K = 6$, and $R = 3$. These values were well-supported by the data, as we explain in Section 5.6.

Experimental setup: We constructed twelve different observed tensors—six from ICEWS and six from GDELT. Five of the six tensors for each source (ICEWS or GDELT) correspond to one-year time spans with monthly time steps, starting with 2004 and ending with 2008; the sixth corresponds to a five-year time span with

monthly time steps, spanning 1995–2000. We divided each tensor $\boldsymbol{Y}$ into a training tensor $\boldsymbol{Y}_{\text{train}} = \boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T-3)}$ and a test tensor $\boldsymbol{Y}_{\text{test}} = \boldsymbol{Y}^{(T-2)}, \boldsymbol{Y}^{(T-1)}, \boldsymbol{Y}^{(T)}$. We further divided each test tensor into a held-out portion and an observed portion via a binary mask. We experimented with two different masks: one that treats the elements involving the most active fifteen countries as the held-out portion and the remaining elements as the observed portion, and one that does the opposite. The first mask enabled us to evaluate the models' reconstructions of the densest (and arguably most interesting) portion of each test tensor, while the second mask enabled us to evaluate their reconstructions of its complement. Across the entire GDELT database, for example, the elements involving the most active fifteen countries—i.e., 6% of all 233 countries—account for 30% of the event tokens. Moreover, 40% of these elements are non-zero. These non-zero elements are highly dispersed, with a variance-to-mean ratio of 220. In contrast, only 0.7% of the elements involving the other countries are non-zero. These elements have a variance-to-mean ratio of 26.

For each combination of the four models, twelve tensors, and two masks, we ran 5,000 iterations of MCMC inference on the training tensor. We then clamped the country–community factors, the action–topic factors, and the core tensor and inferred the time-step–regime factors for the test tensor using its observed portion by running 1,000 iterations of MCMC inference. We saved every tenth sample after the first 500. We used each sample, along with the clamped country–community factors, the action–topic factors, and the core tensor, to compute the Poisson rate for each element in the held-out portion of the test tensor. Finally, we averaged these rates across samples and used each element's average rate—i.e., for each type $\boldsymbol{\delta} \in \Delta_{\text{heldout}}$ in the set of heldout indices, we estimate its rate by averaging over S samples of the time factors:

$$\mu_{\boldsymbol{\delta}} = \frac{1}{S} \sum_{s=1}^{S} \sum_{r=1}^{R} \left(\theta_r^{(\delta_4)}\right)^{(s)} \mu^{(r)}_{\delta_1 \xrightarrow{\delta_2} \delta_3} \tag{5.19}$$

where $\mu^{(r)}_{\delta_1 \xrightarrow{\delta_2} \delta_3} = \psi_{ic}\,\psi_{jd}\,\phi_{ak}\;\lambda^{(r)}_{c \xrightarrow{k} d}$ is clamped.

Given a point-estimate, averaged over the posterior, of each heldout entry's Poisson rate we compute the geometric mean of their likelihoods which is equivalent to inverse perplexity. Perplexity, is defined as,

$$\text{Perp}\left(\boldsymbol{Y}_{\text{heldout}}; \boldsymbol{\mu}_{\text{heldout}}\right) = \exp\left(-\frac{1}{|\Delta_{\text{heldout}}|} \sum_{\boldsymbol{\delta} \in \Delta_{\text{heldout}}} \log \text{Pois}\left(y_{\boldsymbol{\delta}}; \mu_{\boldsymbol{\delta}}\right)\right), \tag{5.20}$$

while its inverse is equal to the geometric mean of the Poisson heldout likelihoods:

$$\text{Perp}\left(\boldsymbol{Y}_{\text{heldout}}; \boldsymbol{\mu}_{\text{heldout}}\right)^{-1} = \exp\left(\frac{1}{|\Delta_{\text{heldout}}|} \sum_{\boldsymbol{\delta} \in \Delta_{\text{heldout}}} \log \text{Pois}\left(y_{\boldsymbol{\delta}}; \mu_{\boldsymbol{\delta}}\right)\right) \tag{5.21}$$

$$= \prod_{\boldsymbol{\delta} \in \Delta_{\text{heldout}}} \left[\text{Pois}\left(y_{\boldsymbol{\delta}}; \mu_{\boldsymbol{\delta}}\right)\right]^{1\,/\,|\Delta_{\text{heldout}}|}. \tag{5.22}$$

We chose this combination strategy to ensure that the models were penalized heavily for making poor predictions on the non-zero elements and were not rewarded excessively for making good predictions on the zero elements since the Poisson PMF penalizes underestimation more than overestimation. By clamping the country–community factors, the action–topic factors, and the core tensor after training, our experimental setup is analogous to that used to assess collaborative filtering models' strong generalization ability [Marlin, 2004].

Results: Figure 5.3 reports the results for each combination of the four models, twelve tensors, and two masks. The top row contains the results from the twelve experiments involving the first mask, where the elements involving the most active fifteen countries were treated as the held-out portion. BPTD outperformed the baselines significantly. BPTF performed better than BPTD in only one study. In general, the Tucker decomposition allows BPTD to learn richer latent structure that generalizes better to held-out data. The bottom row contains the results from the studies involving the second mask. The models' performance was closer in these studies,

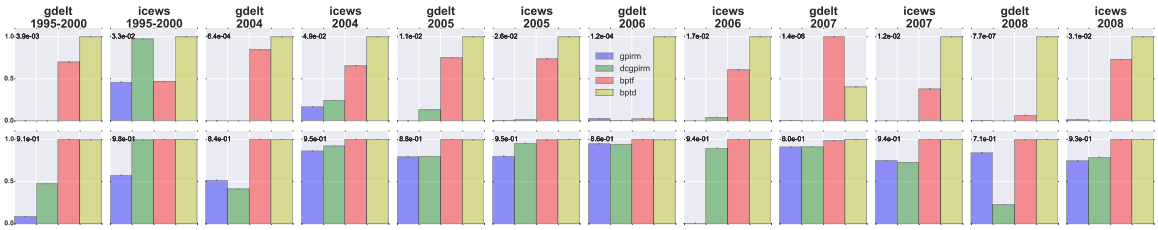


Figure 5.3: Predictive performance. Each plot shows the inverse perplexity (higher is better) for the four models: the GPIRM (blue), the DCGPIRM (green), BPTF (red), and BPTD (yellow). In the experiments depicted in the top row, we treated the elements involving the most active countries as the held-out portion; in the experiments depicted in the bottom row, we treated the remaining elements as the held-out portion. For ease of comparison, we scaled the inverse perplexities to lie between zero and one; we give the scales in the top-left corners of the plots. BPTD outperformed the baselines significantly when predicting the denser portion of each test tensor (top row).

probably because of the large proportion of easy-to-predict zero elements. BPTD and BPTF performed indistinguishably in these studies, and both models outperformed the GPIRM and the DCGPIRM. The single-membership nature of the GPIRM and the DCGPIRM prevents them from expressing high levels of heterogeneity in the countries' rates of activity. When the held-out elements were highly dispersed, these models sometimes made extremely inaccurate predictions.

5.6 Exploratory Analysis

We used a tensor of ICEWS events spanning 1995–2000, with monthly time steps, to explore the latent structure discovered by BPTD. We initially let $C = 50$, $K = 8$, and $R = 3$—i.e., $C \times C \times K \times R = 60,000$ latent classes—and used the shrinkage priors to infer the most appropriate numbers of communities, topics, and regimes. We found $C = 15$ communities and $K = 5$ topics with weights that were significantly greater than zero. We provide a plot of the community weights in Fig. 5.4. Although all three regimes had non-zero weights, one had a much larger weight than the other two. For comparison, Schein et al. [2015] used fifty latent classes to model the same data, while Hoff et al. [2016] used $C = 4$, $K = 4$, and $R = 4$ to model a similar tensor from GDELT.

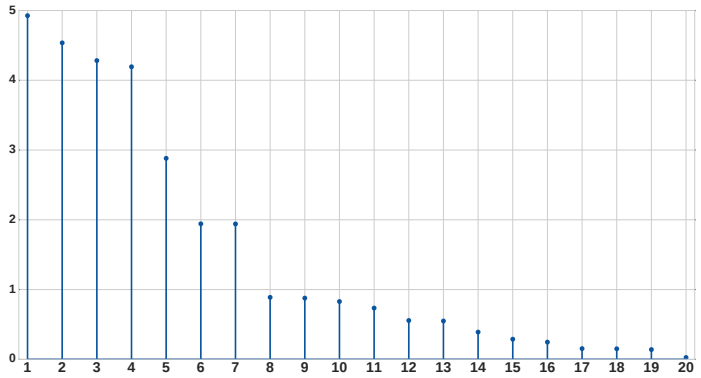


Figure 5.4: Top twenty inferred community weights $\eta_1^{\leftrightarrow}, \ldots, \eta_C^{\leftrightarrow}$.

Topics of actions: We show the inferred action–topic factors as a heatmap in the left subplot of Fig. 5.5. We ordered the topics by their weights $\nu_1, \ldots, \nu_K$, which we display above the heatmap. The inferred topics correspond very closely to CAMEO's QuadClass scheme. Moving from left to right, the topics place their mass on increasingly negative actions. Topics 1 and 2 place most of their mass on Verbal Cooperation actions; topic 3 places most of its mass on Material Cooperation actions and the neutral 1—Make Statement action; topic 4 places most of its mass on Verbal Conflict actions and the 1—Make Statement action; and topics 5 and 6 place their mass on Material Conflict actions.

Topic-partitioned community–community networks: In the right subplot of Fig. 5.5, we visualize the inferred community structure for topic $k = 1$ and the most active regime r. The bottom-left heatmap is the community–community interaction network $\mathbf{\Lambda}_k^{(r)}$. The top-left heatmap depicts the rate at which each country i acts as a sender in each community c—i.e., $\psi_{ic} \sum_{j=1}^{V} \sum_{d=1}^{C} \psi_{jd} \lambda^{(r)}_{c \xrightarrow{k} d}$. Similarly, the bottom-right heatmap depicts the rate at which each country acts as a receiver in each community. The top-right heatmap depicts the number of times each country i took an action associated with topic k toward each country j during regime r—i.e., $\sum_{c=1}^{C} \sum_{d=1}^{C} \sum_{a=1}^{A} \sum_{t=1}^{T} y^{(tr)}_{ic \xrightarrow{ak} jd}$. We grouped the countries by their strongest community memberships and ordered the communities by their within-community

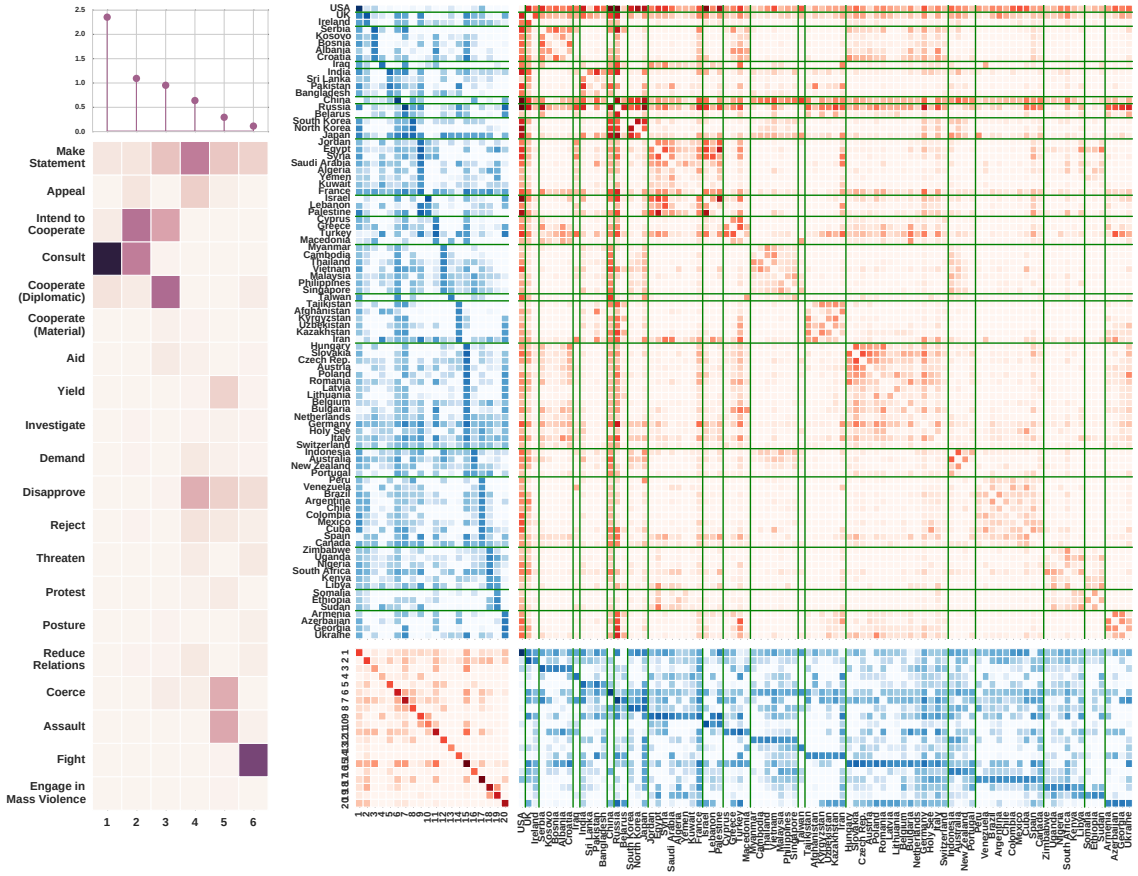


Figure 5.5: *Left:* Action–topic factors. The topics are ordered by $\nu_1, \ldots, \nu_K$ (above the heatmap). *Right:* Latent structure discovered by BPTD for topic $k = 1$ and the most active regime, including the community–community interaction network (bottom left), the rate at which each country acts as a sender (top left) and a receiver (bottom right) in each community, and the number of times each country i took an action associated with topic k toward each country j during regime r (top right). We show only the most active 100 countries.

interaction weights $\eta_1^{\circlearrowright}, \ldots, \eta_C^{\circlearrowright}$, from smallest to largest; the thin green lines separate the countries that are strongly associated with one community from the countries that are strongly associated with its adjacent communities.

Some communities contain only one or two strongly associated countries. For example, community 1 contains only the US, community 6 contains only China, and community 7 contains only Russia and Belarus. These communities mostly engage in between-community interaction. Other larger communities, such as communities

9 and 15, mostly engage in within-community interaction. Most communities have a strong geographic interpretation. Moving upward from the bottom, there are communities that correspond to Eastern Europe, East Africa, South-Central Africa, Latin America, Australasia, Central Europe, Central Asia, etc. The community–community interaction network summarizes the patterns in the top-right heatmap. This topic is dominated by the 4–Consult action, so the network is symmetric; the more negative topics have asymmetric community–community interaction networks. We therefore hypothesize that cooperation is an inherently reciprocal type of interaction. We provide visualizations for the other five topics figures 5.6– 5.11.

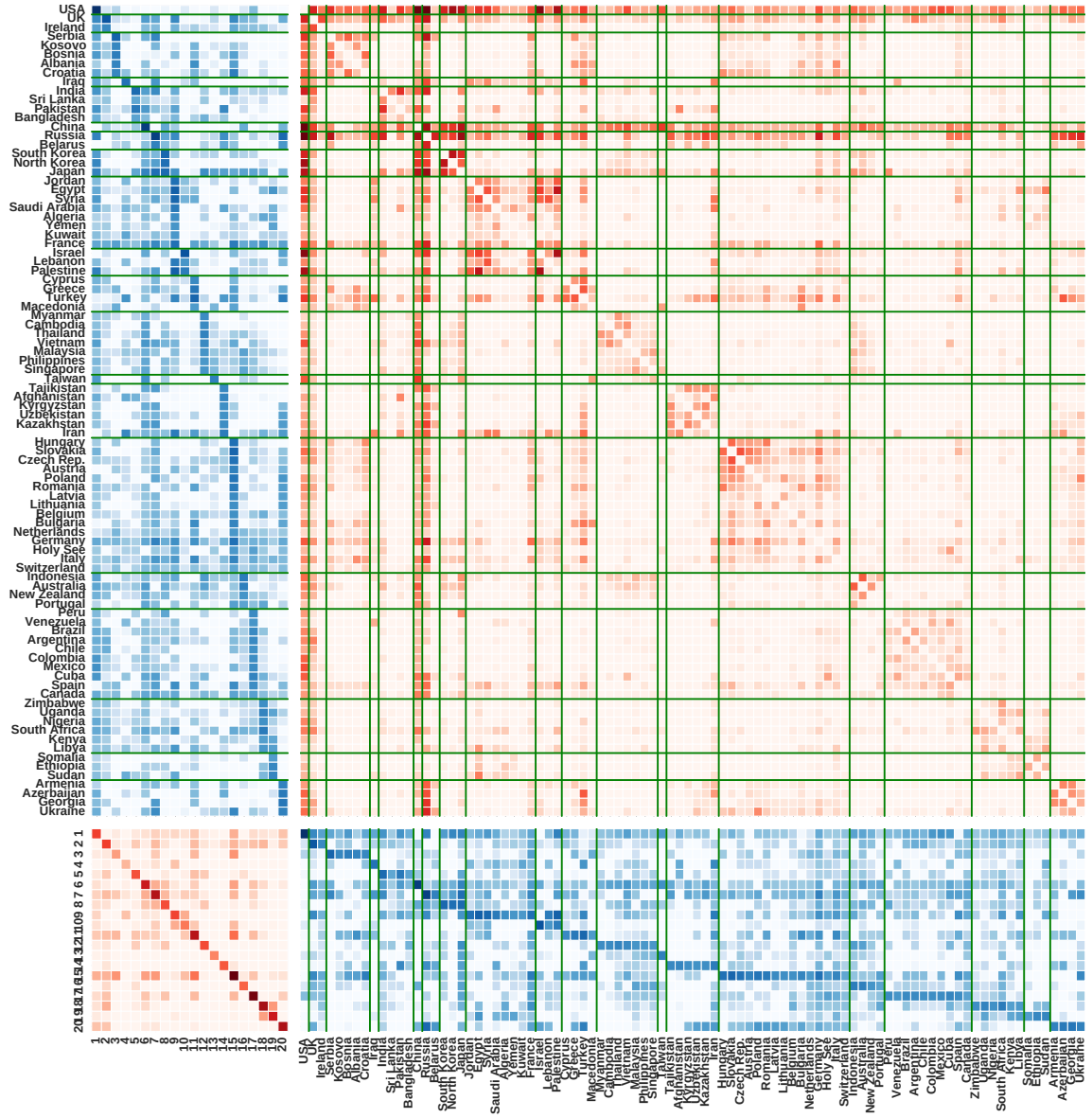


Figure 5.6: Latent structure discovered by BPTD for topic $k = 1$ (mostly Verbal Cooperation action types) and the most active regime, including the community–community interaction network (bottom left), the rate at which each country acts as a sender (top left) and a receiver (bottom right) in each community, and the number of times each country i took an action associated with topic k toward each country j during regime r (top right). We show only the most active 100 countries.

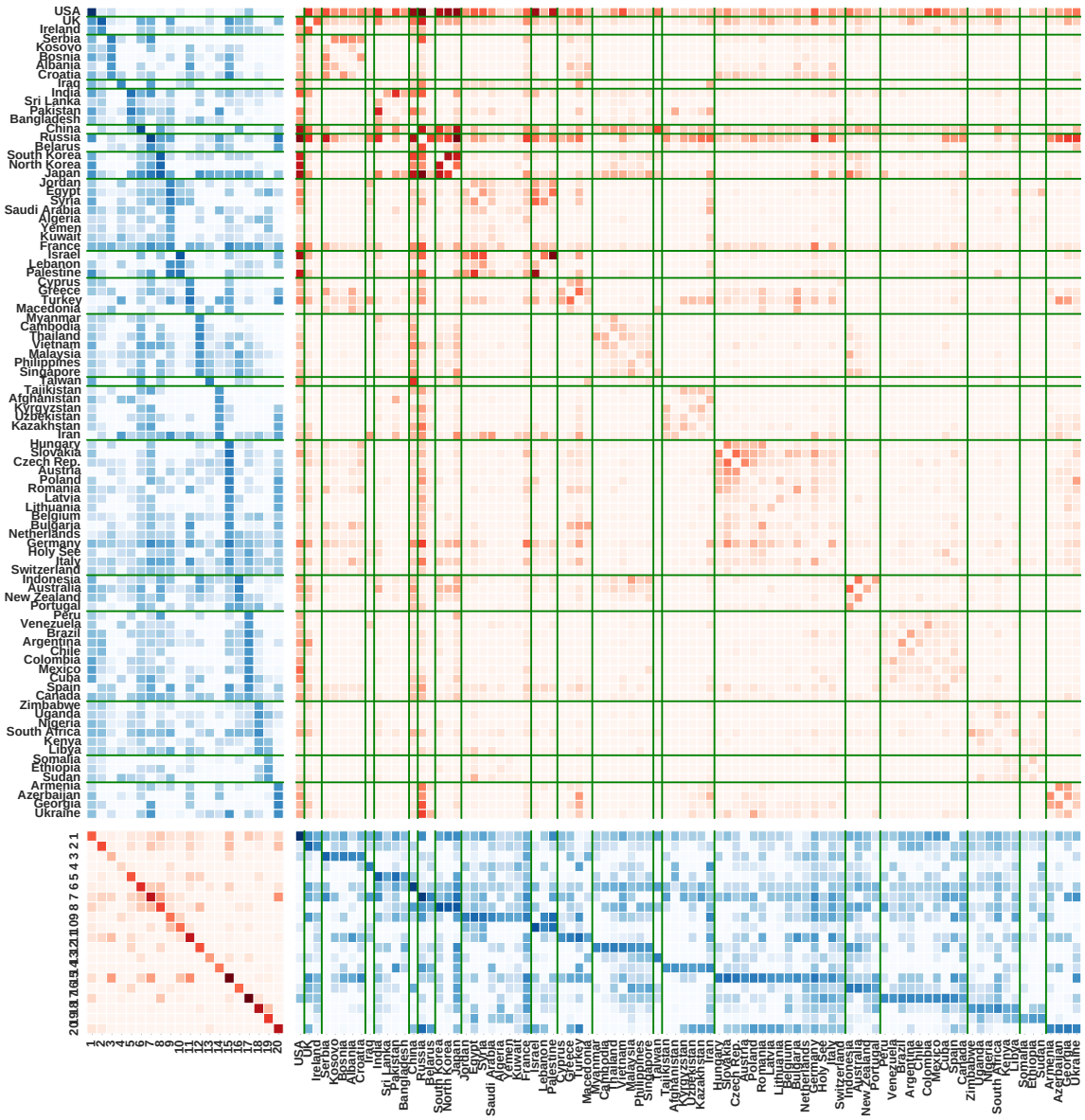


Figure 5.7: Latent structure for topic $k=2$ (Verbal Cooperation).

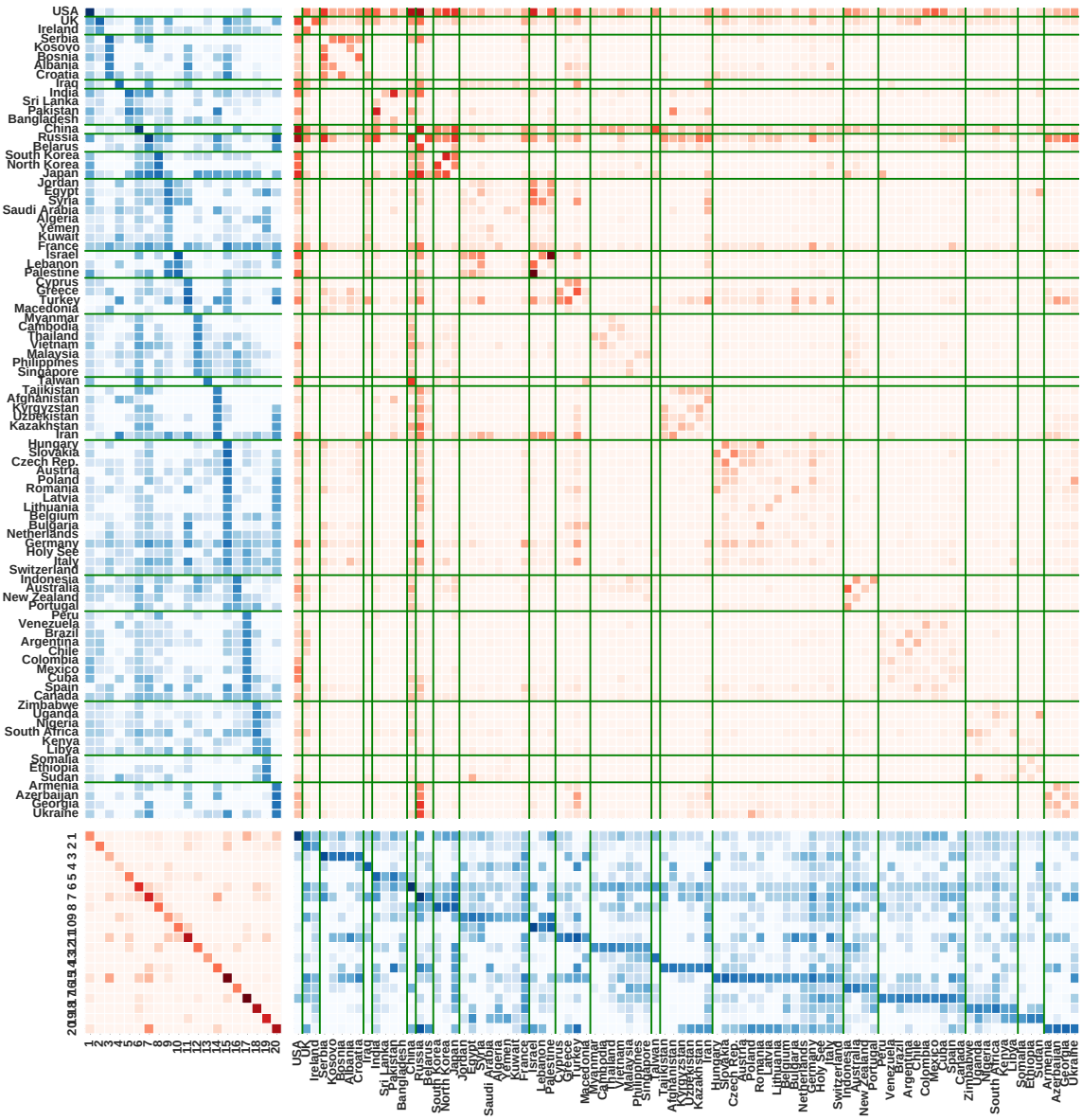


Figure 5.8: Latent structure for topic $k = 3$ (Material Cooperation).

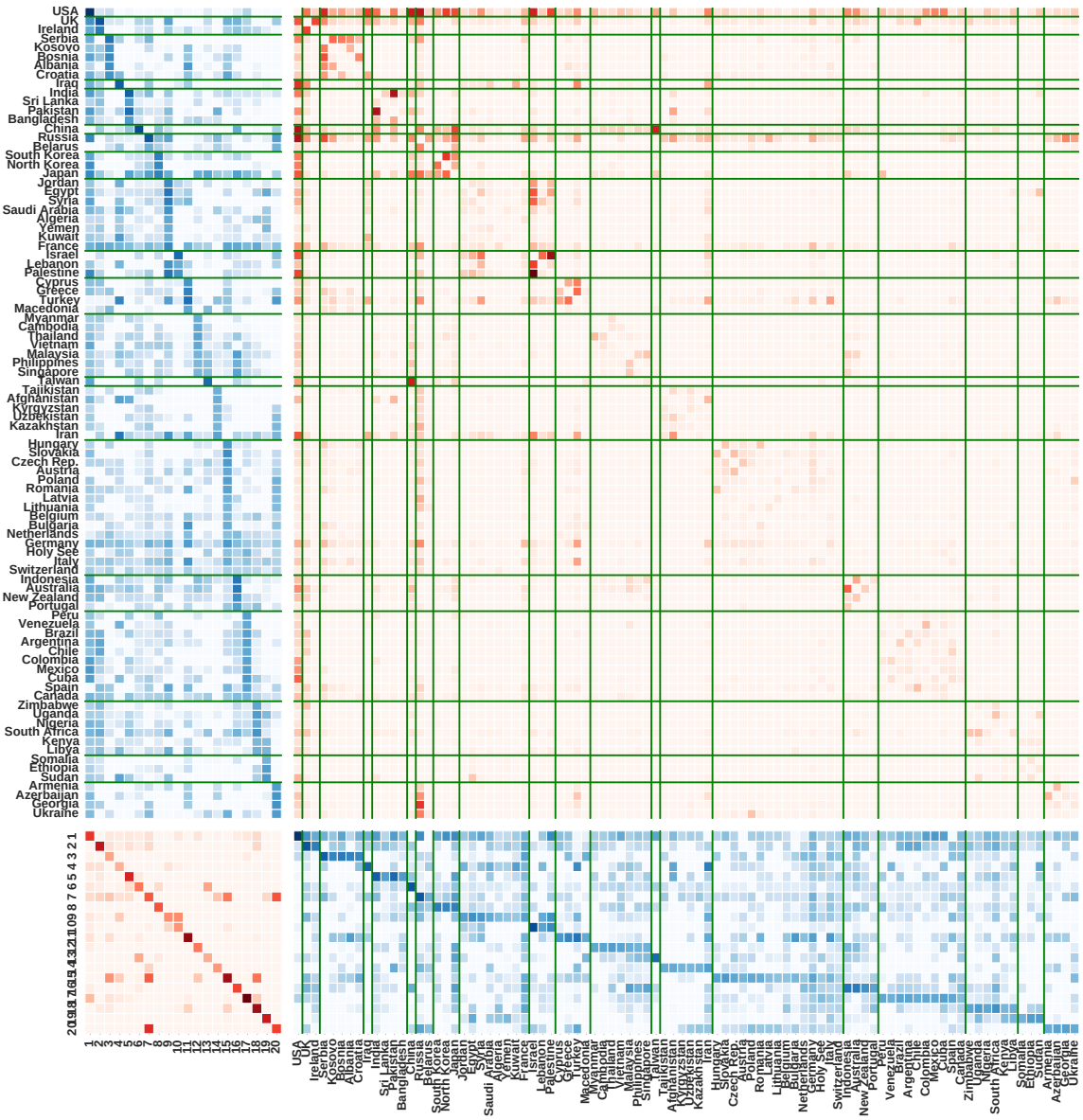


Figure 5.9: Latent structure for topic $k = 4$ (Verbal Conflict).

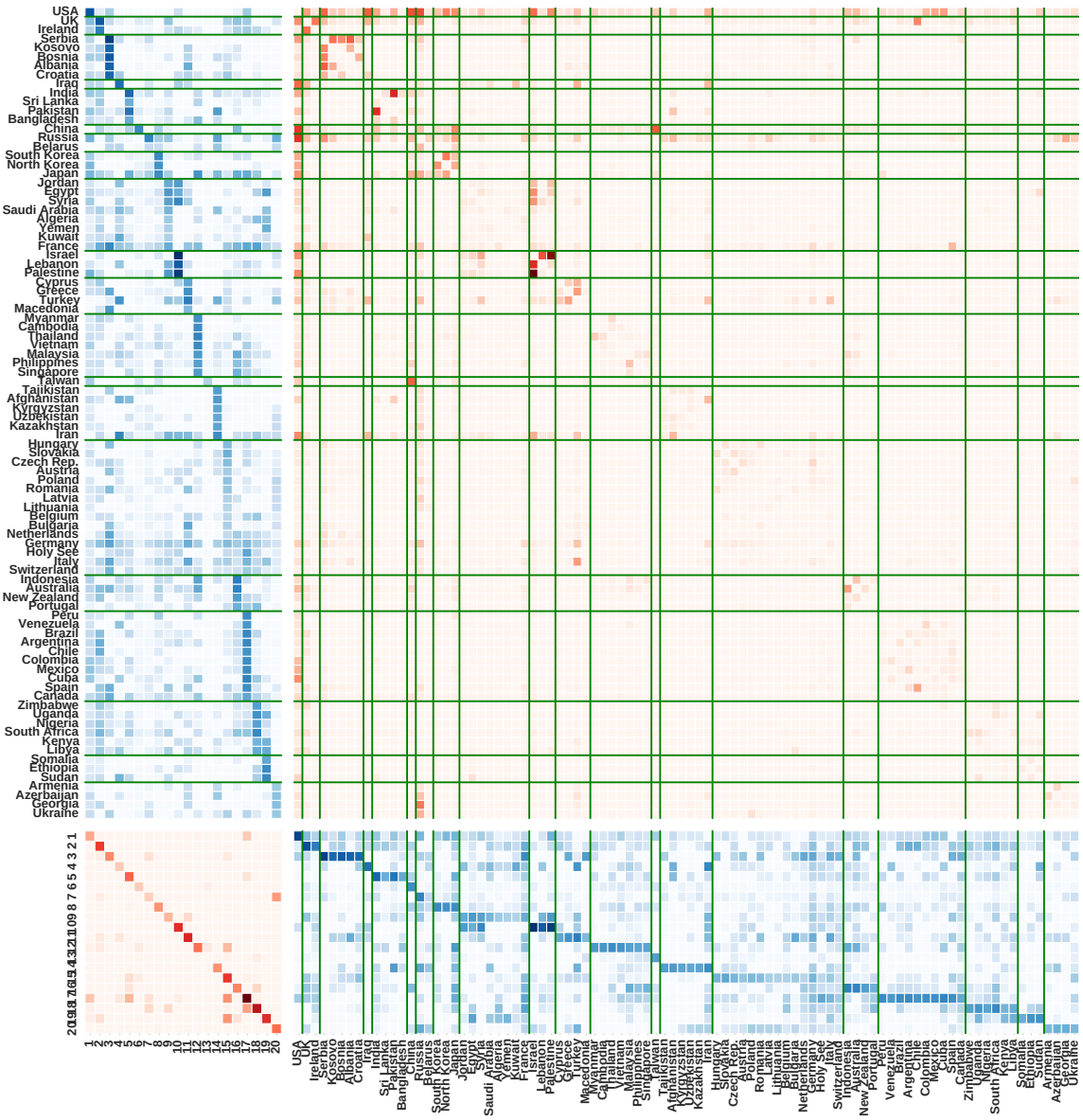


Figure 5.10: Latent structure for topic $k = 5$ (Material Conflict).

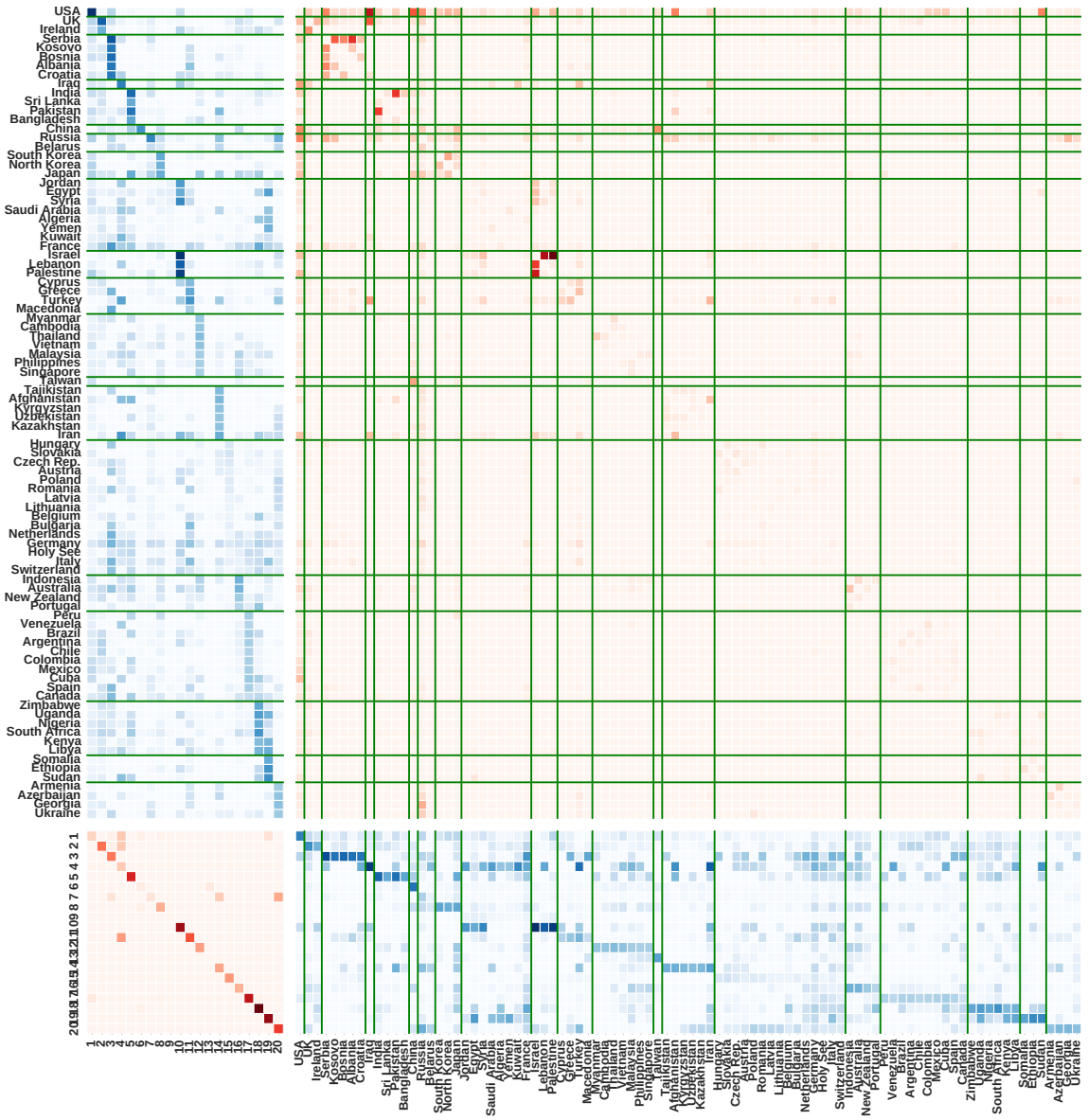


Figure 5.11: Latent structure for topic $k = 6$ (Material Conflict).

CHAPTER 6

POISSON–GAMMA DYNAMICAL SYSTEMS

Sequentially observed count vectors $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(T)}$ are the main object of study in many real-world applications, including text analysis, social network analysis, and recommender systems. Count data pose unique statistical and computational challenges when they are high-dimensional, sparse, and overdispersed, as is often the case in real-world applications. For example, when tracking counts of user interactions in a social network, only a tiny fraction of possible edges are ever active, exhibiting bursty periods of activity when they are. Models of such data should exploit this sparsity in order to scale to high dimensions and be robust to overdispersed temporal patterns. In addition to these characteristics, sequentially observed multivariate count data often exhibit complex dependencies within and across time steps. For example, scientific papers about one topic may encourage researchers to write papers about another related topic in the following year. Models of such data should therefore capture the topic structure of individual documents as well as the excitatory relationships between topics.

The linear dynamical system (LDS) is a widely used model for sequentially observed data, with many well-developed inference techniques based on the Kalman filter [Kalman, 1960, Ghahramani and Roweis, 1998]. The LDS assumes that each sequentially observed V-dimensional vector $\boldsymbol{r}^{(t)}$ is real valued and Gaussian distributed: $\boldsymbol{r}^{(t)} \sim \mathcal{N}(\Phi\, \boldsymbol{\theta}^{(t)}, \Sigma)$, where $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^K$ is a latent state, with K components, that is linked to the observed space via $\Phi \in \mathbb{R}^{V \times K}$. The LDS derives its expressive power from the way it assumes that the latent states evolve: $\boldsymbol{\theta}^{(t)} \sim \mathcal{N}(\Pi\, \boldsymbol{\theta}^{(t-1)}, \Delta)$, where

$\Pi \in \mathbb{R}^{K \times K}$ is a transition matrix that captures between-component dependencies across time steps. Although the LDS can be linked to non-real observations via the extended Kalman filter [Haykin, 2001], it cannot efficiently model real-world count data because inference scales at least linearly $\mathcal{O}(V)$ and sometimes cubically $\mathcal{O}((K+V)^3)$ with dimensionality of the data. [Ghahramani and Roweis, 1998].

Many previous approaches to modeling sequentially observed count data rely on the generalized linear modeling framework [McCullagh and Nelder, 1989] to link the observations to a latent Gaussian space—e.g., via the Poisson–lognormal link [Bulmer, 1974]. Researchers have used this construction to factorize sequentially observed count matrices under a Poisson likelihood, while modeling the temporal structure using well-studied Gaussian techniques [Blei and Lafferty, 2006, Charlin et al., 2015]. Most of these previous approaches assume a simple Gaussian state-space model—i.e., $\boldsymbol{\theta}^{(t)} \sim \mathcal{N}(\boldsymbol{\theta}^{(t-1)}, \Delta)$—that lacks the expressive transition structure of the LDS; one notable exception is the Poisson linear dynamical system [Macke et al., 2011]. In practice, these approaches exhibit prohibitive computational complexity in high dimensions, and the Gaussian assumption may fail to accommodate the burstiness often inherent to real-world count data [Kleinberg, 2003].

This chapter is based on work published at NIPS 2016 [Schein et al., 2016a] and presents the *Poisson–gamma dynamical system (PGDS)*—an APF analogue to the LDS. As an APF model, PGDS is robust to "bursty" data and posterior inference scales with only the number of non-zeros. The main challenge is in deriving tractable inference for the non-conjugate chains of gamma random variables, that are linked via their shape parameter. We develop a novel augment-and-conquer scheme (see Section 3.5) to obtain an elegant and efficient "backward filtering–forward sampling" algorithm. We also examine the way in which the dynamical gamma–Poisson construction propagates information and derive the model's steady state, which involves the Lambert W function [Corless et al., 1996]. Finally, we use the PGDS to analyze a

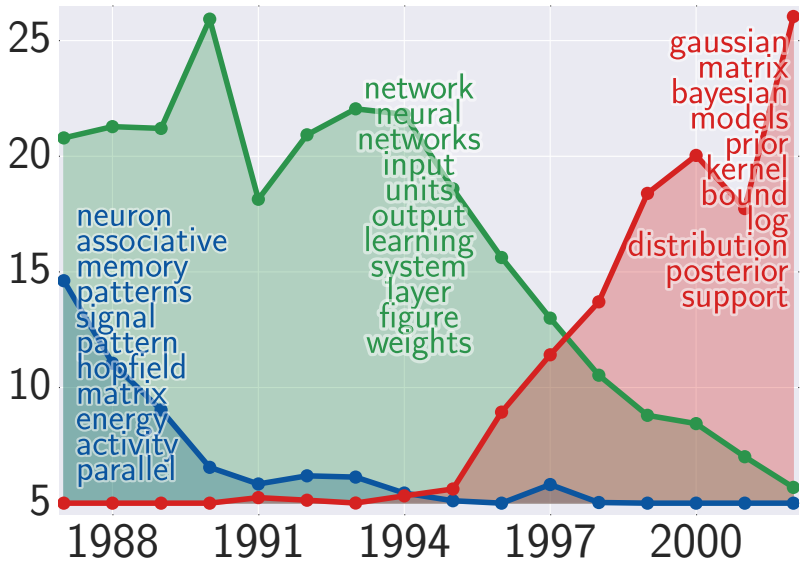


Figure 6.1: The time-step factors for three components inferred by PGDS from a corpus of NIPS papers. Each component is associated with a feature factor for each word type in the corpus; we list the words with the largest factors. The inferred structure tells a familiar story about the rise and fall of certain subfields of machine learning.

diverse range of real-world data sets, showing that it exhibits excellent predictive performance on smoothing and forecasting tasks and infers interpretable latent structure, an example of which is depicted in Fig. 6.1.

While the previous chapters in this thesis focused on modeling count tensors of dyadic events data, this chapter presents PGDS as a model for dynamic count matrices. This is mainly for ease of exposition and comparison to the Gaussian LDS, which is standardly specified for matrices. The generalization of PGDS to tensors is presented and applied to count tensors in Chapter 7.

6.1 Model: Poisson–Gamma Dynamical Systems

We can represent a data set of V-dimensional sequentially observed count vectors $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(T)}$ as a $V \times T$ count matrix Y. The PGDS likelihood corresponds to standard allocative Poisson matrix factorization, where a single count $y_v^{(t)}$ is modeled as

$$y_v^{(t)} \sim \text{Pois}\left(\rho^{(t)} \sum_{k=1}^{K} \tilde{\phi}_{vk}\, \theta_k^{(t)}\right), \tag{6.1}$$

where the positive latent factors $\tilde{\phi}_{vk}$ and $\theta_k^{(t)}$ represent the strength of feature v in component k and the strength of component k at time step t, respectively. The scaling factor $\rho^{(t)}$ captures the scale of the counts at time step t, and therefore obviates the need to rescale the data as a preprocessing step. We refer to the PGDS as *stationary* if $\rho^{(t)} = \rho$ for all t. The defining property of PGDS is how the latent states evolve:

$$\theta_k^{(t)} \sim \Gamma\left(\tau_0 \sum_{k_2=1}^{K} \pi_{kk_2}\, \theta_{k_2}^{(t-1)},\ \tau_0\right). \tag{6.2}$$

The PGDS is characterized by its expressive transition structure, which assumes that each time-step factor $\theta_k^{(t)}$ is drawn from a gamma distribution, whose shape parameter is a linear combination of the K factors at the previous time step. The latent transition weights $\pi_{11}, \ldots, \pi_{k_1k_2}, \ldots, \pi_{KK}$, which we can view collectively as a $K \times K$ transition matrix Π, capture the excitatory relationships between components. The vector $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_K^{(t)})$ has expected value $\mathbb{E}[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \Pi] = \Pi\, \boldsymbol{\theta}^{(t-1)}$ and is thus analogous to the latent state in linear dynamical systems [Kalman, 1960, Ghahramani and Roweis, 1998]. The concentration parameter τ_0 determines the variance of $\boldsymbol{\theta}^{(t)}$—specifically, $\text{Var}\,(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \Pi) = (\Pi\, \boldsymbol{\theta}^{(t-1)})\, \tau_0^{-1}$—without affecting its expected value.

To model the strength of each component, we introduce K component weights $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$ and place a shrinkage prior over them. We assume that the time-step factors and transition weights for component k are tied to its component weight ν_k. Specifically, we define the following structure:

$$\theta_k^{(1)} \sim \Gamma(\tau_0\, \nu_k, \tau_0) \tag{6.3}$$

$$\boldsymbol{\pi}_k \sim \text{Dir}(\nu_1\nu_k, \ldots, \xi\nu_k\ \ldots, \nu_K\nu_k) \tag{6.4}$$

$$\nu_k \sim \Gamma\left(\frac{\gamma_0}{K}, \beta\right), \tag{6.5}$$

where $\boldsymbol{\pi}_k = (\pi_{1k}, \ldots, \pi_{Kk})$ is the k^{th} column of Π. Because $\sum_{k_1=1}^{K} \pi_{k_1 k} = 1$, we can interpret $\pi_{k_1 k}$ as the probability of transitioning from component k to component k_1. (We note that interpreting Π as a stochastic transition matrix relates the PGDS to the discrete hidden Markov model.) For a fixed value of γ_0, increasing K will encourage many of the component weights to be small. A small value of ν_k will shrink $\theta_k^{(1)}$, as well as the transition weights in the k^{th} row of Π. Small values of the transition weights in the k^{th} row of Π therefore prevent component k from being excited by the other components and by itself. Specifically, because the shape parameter for the gamma prior over $\theta_k^{(t)}$ involves a linear combination of $\boldsymbol{\theta}^{(t-1)}$ and the transition weights in the k^{th} row of Π, small transition weights will result in a small shape parameter, shrinking $\theta_k^{(t)}$. Thus, the component weights play a critical role in the PGDS by enabling it to automatically turn off any unneeded capacity and avoid overfitting.

Finally, we place Dirichlet priors over the feature factors $\tilde{\boldsymbol{\phi}}_k = (\tilde{\phi}_{1k}, \ldots, \tilde{\phi}_{Vk})$,

$$\tilde{\boldsymbol{\phi}}_k \sim \text{Dir}(\eta_0, \ldots, \eta_0), \tag{6.6}$$

and draw the other parameters from a non-informative gamma prior:

$$\rho^{(t)}, \xi, \beta \sim \Gamma(\epsilon_0, \epsilon_0). \tag{6.7}$$

The PGDS has four positive hyperparameters to be set by the user: τ_0, γ_0, η_0, and ϵ_0.

6.2 MCMC Inference

PGDS is based on chaining gamma random variables via their shape parameter. This non-conjugate hierarchy makes deriving coordinate-ascent variational inference challenging. However, we can derive an augmentation scheme that facilitates an efficient Gibbs sampler. Our approach is similar to that used to develop Gibbs sampling algorithms for several other related models [Zhou and Carin, 2012, 2015, Acharya

et al., 2015, Zhou, 2015]; however, we extend this approach to handle the unique properties of the PGDS. The main technical challenge is sampling Θ from its conditional posterior, which does not have a closed form. We address this challenge by introducing a set of auxiliary variables. Under this augmented version of the model, marginalizing over Θ becomes tractable and its conditional posterior has a closed form. Moreover, by introducing these auxiliary variables and marginalizing over Θ, we obtain an alternative model specification that we can subsequently exploit to obtain closed-form conditional posteriors for Π, $\boldsymbol{\nu}$, and ξ. We marginalize over Θ by performing a "backward filtering" pass, starting with $\boldsymbol{\theta}^{(T)}$. We iteratively appeal to three definitions in order to do this:

1. *Poisson–multinomial thinning* as it appears in Definition 3.4,
2. the *magic bivariate count distribution* as it appears in Definition 3.20 and,
3. a reparameterized version of Definition 3.16 (*gamma–Poisson construction of the negative binomial*), given below.

Definition 6.1: Reparameterized gamma–Poisson mixture

Consider a Poisson random variable $y \sim \text{Pois}(\theta\, c)$ whose rate is a product of constant $c > 0$ and $\theta \sim \Gamma(a, b)$ which is a gamma random variable with shape a and *rate* b. The marginal distribution—i.e., $y \sim \text{NB}\left(a, \frac{c}{c+b}\right)$—is a negative binomial which can be equivalently expressed in a reparameterized form as:

$$y \sim \text{NB}\Big(a,\, g(\zeta)\Big) \tag{6.8}$$

$$\zeta \triangleq \ln\left(1 + \frac{c}{b}\right), \tag{6.9}$$

where $g(x) = 1 - \exp(-x)$ is the Bernoulli–Poisson link function [Zhou, 2015].

6.2.1 Marginalizing over Θ

We first re-express the Poisson likelihood in terms of the latent source: $y_v^{(t)} \equiv y_{v\cdot}^{(t)} = \sum_{k=1}^K y_{vk}^{(t)}$ and $y_{vk}^{(t)} \sim \text{Pois}\left(\rho^{(t)}\,\tilde{\phi}_{vk}\,\theta_k^{(t)}\right)$. We then define the aggregation $y_{\cdot k}^{(t)} \triangleq \sum_{v=1}^V y_{vk}^{(t)}$ which has marginal distribution $y_{\cdot k}^{(t)} \sim \text{Pois}(\rho^{(t)}\,\theta_k^{(t)})$ due to Poisson additivity and because $\sum_{v=1}^V \tilde{\phi}_{vk} = 1$. We start with the last state $\theta_k^{(T)}$ since no future gamma random variables depend on it; marginalizing over $\theta_k^{(T)}$ yields:

$$y_{\cdot k}^{(T)} \sim \text{NB}\left(\tau_0 \sum_{k_2=1}^K \pi_{kk_2}\,\theta_{k_2}^{(T-1)}, g(\zeta^{(T)})\right) \tag{6.10}$$

$$\zeta^{(T)} \triangleq \ln\left(1 + \frac{\rho^{(T)}}{\tau_0}\right). \tag{6.11}$$

Next, we marginalize out $\theta_k^{(T-1)}$. To do so, we first augment with an auxiliary variable:

$$l_k^{(T)} \sim \text{CRT}\left(y_{\cdot k}^{(T)}, \tau_0 \sum_{k_2=1}^K \pi_{kk_2}\,\theta_{k_2}^{(T-1)}\right). \tag{6.12}$$

We can then re-express the (magic) bivariate distribution over $y_{\cdot k}^{(T)}$ and $l_k^{(T)}$ as

$$y_{\cdot k}^{(T)} \sim \text{SumLog}\left(l_k^{(T)}, g(\zeta^{(T)}\right) \tag{6.13}$$

$$l_k^{(T)} \sim \text{Pois}\left(\zeta^{(T)}\,\tau_0 \sum_{k_2=1}^K \pi_{kk_2}\,\theta_{k_2}^{(T-1)}\right). \tag{6.14}$$

We are still unable to marginalize out $\theta_k^{(T-1)}$ because it is governed by the sum in the Poisson rate of $l_k^{(T)}$. However, we can re-express $l_k^{(T)}$ in terms of its latent sources,

$$l_k^{(T)} \equiv l_{k\cdot}^{(T)} = \sum_{k_2=1}^K l_{kk_2}^{(T)} \tag{6.15}$$

$$l_{kk_2}^{(T)} \sim \text{Pois}\left(\zeta^{(T)}\,\tau_0\,\pi_{kk_2}\,\theta_{k_2}^{(T-1)}\right), \tag{6.16}$$

and then define $l_{\cdot k}^{(T)} \triangleq \sum_{k_1=1}^K l_{k_1 k}^{(T)}$ whose marginal distribution is $l_{\cdot k}^{(T)} \sim \text{Pois}\left(\zeta^{(T)}\,\tau_0\,\theta_k^{(T-1)}\right)$. Crucially, this Poisson doesn't depend on transition weights because $\sum_{k_1=1}^K \pi_{k_1 k} = 1$.

The sources $(l_{1k}^{(T)}, \ldots, l_{Kk}^{(T)}) \sim \text{Mult}\left(l_{\cdot k}^{(T)}, (\pi_{1k}, \ldots, \pi_{kK})\right)$ are then conditionally multinomial when $l_{\cdot k}^{(T)}$ is marginally Poisson. Next, we define $m_k^{(T-1)} \triangleq y_{\cdot k}^{(T-1)} + l_{\cdot k}^{(T)}$, which summarizes all of the information about the data at time steps $T-1$ and T via $y_{\cdot k}^{(T-1)}$ and $l_{\cdot k}^{(T)}$, respectively. By Poisson additivity, $m_k^{(T-1)}$ is then marginally Poisson:

$$m_k^{(T-1)} \sim \text{Pois}\left(\theta_k^{(T-1)}\left(\rho^{(T-1)} + \zeta^{(T)}\,\tau_0\right)\right). \tag{6.17}$$

Combining this with the gamma prior in Eq. (6.2), marginalizing out $\theta_k^{(T-1)}$ yields

$$m_k^{(T-1)} \sim \text{NB}\left(\tau_0 \sum_{k_2=1}^{K} \pi_{kk_2}\,\theta_{k_2}^{(T-2)},\; g\big(\zeta^{(T-1)}\big)\right) \tag{6.18}$$

$$\zeta^{(T-1)} \triangleq \ln\left(1 + \frac{\rho^{(T-1)}}{\tau_0} + \zeta^{(T)}\right). \tag{6.19}$$

We then augment with $l_k^{(T-1)} \sim \text{CRT}\left(m_k^{(T-1)}, \tau_0 \sum_{k_2=1}^{K} \pi_{kk_2}\,\theta_{k_2}^{(T-2)}\right)$ and re-express the (magic) bivarite distribution of $l_k^{(T-1)}$ and $m_k^{(T-1)}$ as a Poisson and sum–logarithmic, similar to Eq. (6.13). This then allows us to marginalize over $\theta_k^{(T-2)}$ to obtain a negative binomial distribution. We can repeat the same process all the way back to $t = 1$, where marginalizing over $\theta_k^{(1)}$ yields

$$m_k^{(1)} \sim \text{NB}\left(\tau_0\,\nu_k,\; g(\zeta^{(1)})\right). \tag{6.20}$$

Note that just as $m_k^{(t)}$ summarizes all the information about the data at time steps $t, \ldots, T$, $\zeta^{(t)} = \ln\left(1 + \frac{\rho^{(t)}}{\tau_0} + \zeta^{(t+1)}\right)$ summarizes all the information about $\rho^{(t)}, \ldots, \rho^{(T)}$.

6.2.2 Steady state

Note that the expression $\zeta^{(t)} = \ln\left(1 + \frac{\rho^{(t)}}{\tau_0} + \zeta^{(t+1)}\right)$ constitutes a *backward pass*, that propagates information about $\rho^{(t)}, \ldots, \rho^{(T)}$ as we marginalize over Θ. In the case of the stationary PGDS—i.e., $\rho^{(t)} = \rho$—the backward pass has a fixed point.

Proposition 6.1: Steady state fixed point

The backward pass has a fixed point equal to

$$\zeta^\star = -\mathbb{W}_{-1}\left(-\exp\left(-1-\frac{\rho}{\tau_0}\right)\right) - 1 - \frac{\rho}{\tau_0}, \tag{6.21}$$

where $\mathbb{W}_{-1}$ is the lower real part of the Lambert W function [Corless et al., 1996].

Proof: If a fixed point exists, then it must satisfy the following equation:

$$\zeta^\star = \ln\left(1 + \tfrac{\rho}{\tau_0} + \zeta^\star\right) \tag{6.22}$$

$$\exp\left(\zeta^\star\right) = 1 + \tfrac{\rho}{\tau_0} + \zeta^\star \tag{6.23}$$

$$-1 = \left(-1 - \tfrac{\rho}{\tau_0} - \zeta^\star\right)\exp\left(-\zeta^\star\right) \tag{6.24}$$

$$-\exp\left(-1 - \tfrac{\rho}{\tau_0}\right) = \left(-1 - \tfrac{\rho}{\tau_0} - \zeta^\star\right)\exp\left(-\zeta^\star\right)\exp\left(-1 - \tfrac{\rho}{\tau_0}\right) \tag{6.25}$$

$$-\exp\left(-1 - \tfrac{\rho}{\tau_0}\right) = \left(-1 - \tfrac{\rho}{\tau_0} - \zeta^\star\right)\exp\left(-1 - \tfrac{\rho}{\tau_0} - \zeta^\star\right). \tag{6.26}$$

By definition of the Lambert W function, $y = x\,\exp(x)$ entails that $x = \mathbb{W}(y)$:

$$\left(-1 - \tfrac{\rho}{\tau_0} - \zeta^\star\right) = \mathbb{W}\left(-\exp\left(-1 - \tfrac{\rho}{\tau_0}\right)\right) \tag{6.27}$$

$$\zeta^\star = -\mathbb{W}\left(-\exp\left(-1 - \tfrac{\rho}{\tau_0}\right)\right) - 1 - \tfrac{\rho}{\tau_0}. \tag{6.28}$$

There are two branches of the Lambert W function. The lower branch decreases from $\mathbb{W}_{-1}\left(-\exp\left(-1\right)\right) = -1$ to $\mathbb{W}_{-1}\left(0\right) = -\infty$, while the principal branch increases from $\mathbb{W}_0\left(-\exp\left(-1\right)\right) = -1$ to $\mathbb{W}_0\left(0\right) = 0$ and beyond. Because $\zeta^\star$ must be positive, we therefore have $\zeta^\star = -\mathbb{W}_{-1}\left(-\exp\left(-1 - \frac{\rho}{\tau_0}\right)\right) - 1 - \frac{\rho}{\tau_0}$. ■

During inference, we perform the $\mathcal{O}(T)$ backward pass repeatedly. The existence of a fixed point means that we can assume the stationary PGDS is in its steady state and replace the backward pass with an $\mathcal{O}(1)$ computation of the fixed point ζ^*. To

make this assumption, we must also assume that $l_{\cdot k}^{(T+1)} \sim \text{Pois}(\zeta^\star \tau_0 \theta_k^{(T)})$ instead of $l_{\cdot k}^{(T+1)} = 0$. We note that an analogous steady-state approximation exists for the LDS and is routinely exploited to reduce computation [Rugh, 1995].

6.2.3 Alternative model specification

As we mentioned previously, introducing auxiliary variables and marginalizing over Θ enables us to define an alternative model specification that we can exploit to obtain closed-form conditional posteriors for Π, $\boldsymbol{\nu}$, and ξ. We provide part of its generative process here. Define $m_k^{(T)} = y_{\cdot k}^{(T)} + l_{\cdot k}^{(T+1)}$, where $l_{\cdot k}^{(T+1)} = 0$, and $\zeta^{(T+1)} = 0$.

$$l_{k\cdot}^{(1)} \sim \text{Pois}(\zeta^{(1)} \tau_0 \nu_k) \tag{6.29}$$

$$(l_{1k}^{(t)}, \ldots, l_{Kk}^{(t)}) \sim \text{Multinom}\left(l_{\cdot k}^{(t)}, (\pi_{1k}, \ldots, \pi_{Kk})\right) \text{ for } t > 1 \tag{6.30}$$

$$l_{k\cdot}^{(t)} = \sum_{k_2=1}^{K} l_{kk_2}^{(t)} \text{ for } t > 1 \tag{6.31}$$

$$m_k^{(t)} \sim \text{SumLog}\left(l_{k\cdot}^{(t)}, g(\zeta^{(t)})\right) \tag{6.32}$$

$$(y_{\cdot k}^{(t)}, l_{\cdot k}^{(t+1)}) \sim \text{Binom}\left(m_k^{(t)}, (\tfrac{\rho^{(t)}}{\rho^{(t)} + \zeta^{(t+1)}\tau_0}, \tfrac{\zeta^{(t+1)}\tau_0}{\rho^{(t)} + \zeta^{(t+1)}\tau_0})\right) \tag{6.33}$$

$$(y_{1k}^{(t)}, \ldots, y_{Vk}^{(t)}) \sim \text{Multinom}\left(y_{\cdot k}^{(t)}, (\tilde{\phi}_{1k}, \ldots, \tilde{\phi}_{Vk})\right). \tag{6.34}$$

6.2.4 Gibbs sampler and backwards filtering forwards sampling

Given Y and the hyperparameters, Gibbs sampling involves resampling each auxiliary variable or model parameter from its conditional posterior. Our algorithm involves a "backward filtering" pass and a "forward sampling" pass, which together form a "backward filtering–forward sampling" algorithm. We use $- \setminus \Theta^{(\geq t)}$ to denote everything excluding $\boldsymbol{\theta}^{(t)}, \ldots, \boldsymbol{\theta}^{(T)}$.

Sampling the auxiliary variables: This step is the "backward filtering" pass. For the stationary PGDS in its steady state, we first compute ζ^* and draw $(l_{\cdot k}^{(T+1)} \mid -) \sim$

$\text{Pois}(\zeta^\star \tau_0 \theta_k^{(T)})$. For the other variants of the model, we set $l_{\cdot k}^{(T+1)} = \zeta^{(T+1)} = 0$. Then, working backward from $t = T, \ldots, 2$, we draw

$$(l_{k\cdot}^{(t)} \mid - \setminus \Theta^{(\geq t)}) \sim \text{CRT}(y_{\cdot k}^{(t)} + l_{\cdot k}^{(t+1)}, \tau_0 \sum_{k_2=1}^{K} \pi_{kk_2} \theta_{k_2}^{(t-1)}) \text{ and} \tag{6.35}$$

$$(l_{k1}^{(t)}, \ldots, l_{kK}^{(t)} \mid - \setminus \Theta^{(\geq t)}) \sim \text{Multinom}\left(l_{k\cdot}^{(t)}, (\tfrac{\pi_{k1}\theta_1^{(t-1)}}{\sum_{k_2=1}^{K}\pi_{kk_2}\theta_{k_2}^{(t-1)}}, \ldots, \tfrac{\pi_{kK}\theta_K^{(t-1)}}{\sum_{k_2=1}^{K}\pi_{kk_2}\theta_{k_2}^{(t-1)}})\right). \tag{6.36}$$

After using equations 6.35 and 6.36 for all $k = 1, \ldots, K$, we then set $l_{\cdot k}^{(t)} = \sum_{k_1=1}^{K} l_{k_1 k}^{(t)}$. For the non-steady-state variants, we also set $\zeta^{(t)} = \ln(1 + \frac{\rho^{(t)}}{\tau_0} + \zeta^{(t+1)})$; for the steady-state variant, we set $\zeta^{(t)} = \zeta^*$.

Sampling Θ: We sample Θ from its conditional posterior by performing a "forward sampling" pass, starting with $\boldsymbol{\theta}^{(1)}$. Conditioned on the values of $l_{\cdot k}^{(2)}, \ldots, l_{\cdot k}^{(T+1)}$ and $\zeta^{(2)}, \ldots, \zeta^{(T+1)}$ obtained via the "backward filtering" pass, we sample forward from $t = 1, \ldots, T$, using the following equations:

$$(\theta_k^{(1)} \mid - \setminus \Theta) \sim \Gamma(y_{\cdot k}^{(1)} + l_{\cdot k}^{(2)} + \tau_0 \nu_k, \tau_0 + \rho^{(1)} + \zeta^{(2)} \tau_0) \text{ and} \tag{6.37}$$

$$(\theta_k^{(t)} \mid - \setminus \Theta^{(\geq t)}) \sim \Gamma(y_{\cdot k}^{(t)} + l_{\cdot k}^{(t+1)} + \tau_0 \sum_{k_2=1}^{K} \pi_{kk_2} \theta_{k_2}^{(t-1)}, \tau_0 + \rho^{(t)} + \zeta^{(t+1)} \tau_0). \tag{6.38}$$

Sampling Π: The alternative model specification, with Θ marginalized out, assumes that $(l_{1k}^{(t)}, \ldots, l_{Kk}^{(t)}) \sim \text{Multinom}\left(l_{\cdot k}^{(t)}, (\pi_{1k}, \ldots, \pi_{Kk})\right)$. Therefore, via Dirichlet–multinomial conjugacy,

$$(\boldsymbol{\pi}_k \mid - \setminus \Theta) \sim \text{Dir}(\nu_1\nu_k + \sum_{t=1}^{T} l_{1k}^{(t)}, \ldots, \xi\nu_k + \sum_{t=1}^{T} l_{kk}^{(t)}, \ldots, \nu_K\nu_k + \sum_{t=1}^{T} l_{Kk}^{(t)}). \tag{6.39}$$

Sampling Φ: By Dirichlet–multinomial conjugacy,

$$(\tilde{\boldsymbol{\phi}}_k \mid -) \sim \text{Dir}\left(\eta_0 + \sum_{t=1}^{T} y_{1k}^{(t)}, \ldots, \eta_0 + \sum_{t=1}^{T} y_{Vk}^{(t)}\right). \tag{6.40}$$

Sampling $\rho^{(1)}, \ldots, \rho^{(T)}$ or ρ (stationary): By gamma–Poisson conjugacy,

$$(\rho^{(t)} \mid -) \sim \Gamma\left(\epsilon_0 + \sum_{v=1}^{V} y_v^{(t)}, \epsilon_0 + \sum_{k=1}^{K} \theta_k^{(t)}\right) \tag{6.41}$$

$$(\rho \mid -) \sim \Gamma\left(\epsilon_0 + \sum_{t=1}^{T}\sum_{v=1}^{V} y_v^{(t)}, \epsilon_0 + \sum_{t=1}^{T}\sum_{k=1}^{K} \theta_k^{(t)}\right). \tag{6.42}$$

Sampling β: Via gamma–gamma conjugacy,

$$(\beta \mid -) \sim \Gamma\left(\epsilon_0 + \gamma_0, \epsilon_0 + \sum_{k=1}^{K} \nu_k\right). \tag{6.43}$$

Sampling $\boldsymbol{\nu}$ and ξ: We use the alternative model specification to obtain closed-form conditional posteriors for ν_k and ξ. First, we marginalize over $\boldsymbol{\pi}_k$ to obtain a Dirichlet–multinomial distribution. When augmented with a beta-distributed auxiliary variable, the Dirichlet–multinomial distribution is proportional to the negative binomial distribution [Zhou, 2018]. We draw such an auxiliary variable, which we use, along with negative binomial augmentation schemes, to derive closed-form conditional posteriors for ν_k and ξ. The derivation begins with the following line:

$$(l_{1k}^{(\cdot)}, \ldots, l_{kk}^{(\cdot)}, \ldots, l_{Kk}^{(\cdot)}) \sim \text{DirMult}(l_{\cdot k}^{(\cdot)}, (\nu_1\nu_k, \ldots, \xi\nu_k, \ldots, \nu_K\nu_k)), \tag{6.44}$$

where $l_{k_1k}^{(\cdot)} = \sum_{t=1}^{T} l_{k_1k}^{(t)}$ and $l_{\cdot k}^{(\cdot)} = \sum_{t=1}^{T}\sum_{k_1=1}^{K} l_{k_1k}^{(t)}$. As noted previously by Zhou [2018], when augmented with a beta-distributed auxiliary variable, the Dirichlet–multinomial distribution is proportional to the negative binomial distribution. We therefore draw a beta-distributed auxiliary variable:

$$q_k \sim \text{Beta}\left(l_{\cdot k}^{(\cdot)}, \nu_k\left(\xi + \sum_{k_1 \neq k} \nu_{k_1}\right)\right). \tag{6.45}$$

Conditioned on q_k, we then have

$$l_{kk}^{(\cdot)} \sim \text{NB}(\xi\,\nu_k, q_k) \text{ and } l_{k_1k}^{(\cdot)} \sim \text{NB}(\nu_{k_1}\nu_k, q_k), \tag{6.46}$$

for $k_1 \neq k$. Next, we introduce the following auxiliary variables:

$$h_{kk} \sim \text{CRT}(l_{kk}^{(\cdot)}, \xi\,\nu_k) \text{ and } h_{k_1k} \sim \text{CRT}(l_{k_1k}^{(\cdot)}, \nu_{k_1}\nu_k), \tag{6.47}$$

for $k_1 \neq k$. We can then re-express the joint distribution over the variables in equations 6.46 and 6.47 as

$$l_{kk}^{(\cdot)} \sim \text{SumLog}(h_{kk}, q_k) \text{ and } l_{k_1k}^{(\cdot)} \sim \text{SumLog}(h_{k_1k}, q_k) \tag{6.48}$$

$$h_{kk} \sim \text{Pois}\left(\xi\,\nu_k \ln\left(\tfrac{1}{1-q_k}\right)\right) \text{ and } h_{k_1k} \sim \text{Pois}\left(\nu_{k_1}\nu_k \ln\left(\tfrac{1}{1-q_k}\right)\right). \tag{6.49}$$

Then, via gamma–Poisson conjugacy,

$$(\xi \mid -\setminus \Theta, \boldsymbol{\pi}_k) \sim \Gamma\left(\frac{\gamma_0}{K} + \sum_{k=1}^{K} h_{kk},\ \beta + \sum_{k=1}^{K} \nu_k \ln\left(\tfrac{1}{1-q_k}\right)\right). \tag{6.50}$$

Next, because $l_{k\cdot}^{(1)} \sim \text{Pois}(\zeta^{(1)}\,\tau_0\,\nu_k)$ also depends on ν_k, we introduce

$$n_k \triangleq h_{kk} + \sum_{k_1 \neq k} h_{k_1k} + \sum_{k_2 \neq k} h_{kk_2} + l_{k\cdot}^{(1)}, \tag{6.51}$$

whose marginal distribution is Poisson—i.e., $n_k \sim \text{Pois}(\nu_k\,\omega_k)$—where ω_k is defined:

$$\omega_k \triangleq \ln\left(\tfrac{1}{1-q_k}\right)\left(\xi + \sum_{k_1 \neq k} \nu_{k_1}\right) + \sum_{k_2 \neq k} \ln\left(\tfrac{1}{1-q_{k_2}}\right)\nu_{k_2} + \zeta^{(1)}\tau_0. \tag{6.52}$$

Then, by gamma–Poisson conjugacy,

$$(\nu_k \mid -\setminus \Theta, \boldsymbol{\pi}_k) \sim \Gamma\left(\frac{\gamma_0}{\beta} + n_k, \beta + \omega_k\right). \tag{6.53}$$

6.3 Predictive Analysis

In this section, we compare the out-of-sample predictive performance of the PGDS to that of the LDS and that of gamma process dynamic Poisson factor analysis (GP-DPFA) [Acharya et al., 2015]. GP-DPFA models a single count in Y as $y_v^{(t)} \sim \text{Pois}\left(\sum_{k=1}^K \lambda_k \tilde{\phi}_{vk} \theta_k^{(t)}\right)$, where each component's time-step factors evolve as a simple gamma Markov chain, independently of those belonging to the other components: $\theta_k^{(t)} \sim \Gamma(\theta_k^{(t-1)}, c^{(t)})$. We consider the stationary variants of all three models.[1] We used five data sets, and tested each model on two time-series prediction tasks: *smoothing*—i.e., predicting $y_v^{(t)}$ given $y_v^{(1)}, \ldots, y_v^{(t-1)}, y_v^{(t+1)}, \ldots, y_v^{(T)}$—and *forecasting*—i.e., predicting $y_v^{(T+s)}$ given $y_v^{(1)}, \ldots, y_v^{(T)}$ for some $s \in \{1, 2, \ldots\}$ [Durbin and Koopman, 2012]. We provide brief descriptions of the data sets below before reporting results.

6.3.1 Data sets

Global Database of Events, Language, and Tone (GDELT): GDELT is an international relations dyadic events data set extracted from news corpora. We created five count matrices, one for each year from 2001 through 2005. We treated directed pairs of countries $i \rightarrow j$ as features and counted the number of events for each pair during each day. We discarded all pairs with fewer than twenty-five total events, leaving $T = 365$, around $V \approx 9,000$, and three to six million events for each matrix.

Integrated Crisis Early Warning System (ICEWS): ICEWS is another international relations event data set extracted from news corpora. It is more highly curated than GDELT and contains fewer events. We therefore treated undirected pairs of countries $i \leftrightarrow j$ as features. We created three count matrices, one for 2001–2003, one for 2004–2006, and one for 2007–2009. We counted the number of events for each pair during each three-day time step, and again discarded all pairs with fewer

[1] We used the `pykalman` Python library for the LDS and implemented GP-DPFA ourselves.

than twenty-five total events, leaving $T = 365$, around $V \approx 3,000$, and 1.3 to 1.5 million events for each matrix.

State-of-the-Union transcripts (SOTU): The SOTU corpus contains the text of the annual SOTU speech transcripts from 1790 through 2014. We created a single count matrix with one column per year. After discarding stopwords, we were left with $T = 225$, $V = 7,518$, and 656,949 tokens.

DBLP conference abstracts (DBLP): DBLP is a database of computer science research papers. We used the subset of this corpus that Acharya et al. used to evaluate GP-DPFA [Acharya et al., 2015]. This subset corresponds to a count matrix with $T = 14$ columns, $V = 1,771$ unique word types, and 13,431 tokens.

NIPS corpus (NIPS): The NIPS corpus contains the text of every NIPS conference paper from 1987 to 2003. We created a single count matrix with one column per year. We treated unique word types as features and discarded all stopwords, leaving $T = 17$, $V = 9,836$, and 3.1 million tokens.

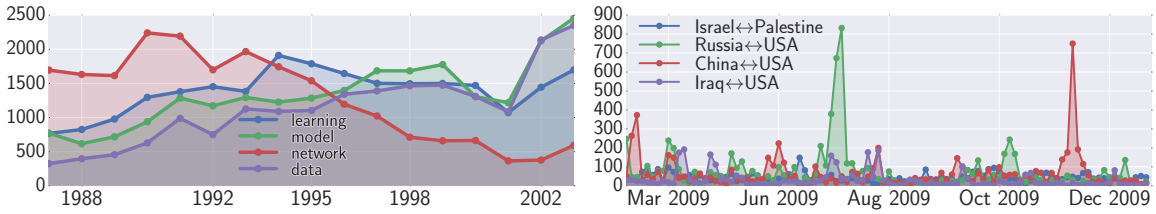


Figure 6.2: $y_v^{(t)}$ over time for the top four features in the NIPS (left) and ICEWS (right) data sets.

6.3.2 Experimental design

For each matrix, we created four masks indicating some randomly selected subset of columns to treat as held-out data. For the event count matrices, we held out six (non-contiguous) time steps between $t=2$ and $t=T-3$ to test the models' smoothing performance, as well as the last two time steps to test their forecasting performance.

The other matrices have fewer time steps. For the SOTU matrix, we therefore held out five time steps between $t=2$ and $t=T-2$, as well as $t=T$. For the NIPS and DBLP matrices, which contain substantially fewer time steps than the SOTU matrix, we held out three time steps between $t=2$ and $t=T-2$, as well as $t=T$.

For each matrix, mask, and model combination, we ran inference four times.[2] For the PGDS and GP-DPFA, we performed 6,000 Gibbs sampling iterations, imputing the missing counts from the "smoothing" columns at the same time as sampling the model parameters. We then discarded the first 4,000 samples and retained every hundredth sample thereafter. We used each of these samples to predict the missing counts from the "forecasting" columns. We then averaged the predictions over the samples. For the LDS, we ran EM to learn the model parameters. Then, given these parameter values, we used the Kalman filter and smoother [Kalman, 1960] to predict the held-out data. In practice, for all five data sets, V was too large for us to run inference for the LDS, which is at least $\mathcal{O}(V)$ [Ghahramani and Roweis, 1998], using all V features. We therefore report results from two independent sets of experiments: one comparing all three models using only the top $V = 1,000$ features for each data set, and one comparing the PGDS to just GP-DPFA using all the features. The first set of experiments is generous to the LDS because the Poisson distribution is well approximated by the Gaussian distribution when its mean is large.

6.3.3 Results

We used two error measures—mean relative error (MRE) and mean absolute error (MAE)—to compute the models' smoothing and forecasting scores for each matrix and mask combination. We then averaged these scores over the masks. For the data

[2]For the PGDS and GP-DPFA we used $K = 100$. For the PGDS, we set $\tau_0 = 1$, $\gamma_0 = 50$, $\eta_0 = \epsilon_0 = 0.1$. We set the hyperparameters of GP-DPFA to the values used by Acharya et al. [2015]. For the LDS, we used the default hyperparameters for `pykalman`, and report results for the best-performing value of $K \in \{5, 10, 25, 50\}$.

Table 6.1: Results for the smoothing ("S") and forecasting ("F") tasks. For both error measures, lower values are better. We also report the number of time steps T and the burstiness $\hat{B}$ of each data set.

				Mean Relative Error (MRE)			Mean Absolute Error (MAE)		
	T	$\hat{B}$	Task	PGDS	GP-DPFA	LDS	PGDS	GP-DPFA	LDS
GDELT	365	1.27	S	**2.335** ±0.19	2.951 ±0.32	3.493 ±0.53	9.366 ±2.19	**9.278** ±2.01	10.098 ±2.39
			F	**2.173** ±0.41	2.207 ±0.42	2.397 ±0.29	**7.002** ±1.43	7.095 ±1.67	7.047 ±1.25
ICEWS	365	1.10	S	**0.808** ±0.11	0.877 ±0.12	1.023 ±0.15	**2.867** ±0.56	2.872 ±0.56	3.104 ±0.60
			F	**0.743** ±0.17	0.792 ±0.17	0.937 ±0.31	**1.788** ±0.47	1.894 ±0.50	1.973 ±0.62
SOTU	225	1.45	S	**0.233** ±0.01	0.238 ±0.01	0.260 ±0.01	**0.408** ±0.01	0.414 ±0.01	0.448 ±0.00
			F	**0.171** ±0.00	0.173 ±0.00	0.225 ±0.01	0.323 ±0.00	**0.314** ±0.00	0.370 ±0.00
DBLP	14	1.64	S	0.417 ±0.03	0.422 ±0.05	**0.405** ±0.05	**0.771** ±0.03	0.782 ±0.06	0.831 ±0.01
			F	**0.322** ±0.00	0.323 ±0.00	0.369 ±0.06	0.747 ±0.01	**0.715** ±0.00	0.943 ±0.07
NIPS	17	0.33	S	0.415 ±0.07	**0.392** ±0.07	1.609 ±0.43	29.940 ±2.95	**28.138** ±3.08	108.378 ±15.44
			F	0.343 ±0.01	**0.312** ±0.00	0.642 ±0.14	62.839 ±0.37	**52.963** ±0.52	95.495 ±10.52

sets with multiple matrices, we also averaged the scores over the matrices. The two error measures differ as follows: MRE accommodates the scale of the data, while MAE does not. This is because relative error—which we define as $\frac{|y_v^{(t)} - \hat{y}_v^{(t)}|}{1 + y_v^{(t)}}$, where $y_v^{(t)}$ is the true count and $\hat{y}_v^{(t)}$ is the prediction—divides the absolute error by the true count and thus penalizes overpredictions more harshly than underpredictions. MRE is therefore an especially natural choice for data sets that are bursty—i.e., data sets that exhibit short periods of activity that far exceed their mean. Models that are robust to these kinds of overdispersed temporal patterns are less likely to make overpredictions following a burst, and are therefore rewarded accordingly by MRE.

In Table 6.1, we report the MRE and MAE scores for the experiments using the top $V = 1,000$ features. We also report the average burstiness of each data set. We define the burstiness of feature v in matrix Y to be $\hat{B}_v = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{|y_v^{(t+1)} - y_v^{(t)}|}{\hat{\mu}_v}$, where $\hat{\mu}_v = \frac{1}{T} \sum_{t=1}^{T} y_v^{(t)}$. For each data set, we calculated the burstiness of each feature in each matrix, and then averaged these values to obtain an average burstiness score $\hat{B}$. The PGDS outperformed the LDS and GP-DPFA on seven of the ten prediction

Table 6.2: Results for the smoothing ("S") and forecasting ("F") tasks using all the features. Lower values are better. We also report the number of time steps T and the burstiness $\hat{B}$ of each data set.

				Mean Relative Error		Mean Absolute Error	
	T	$\hat{B}$	Task	PGDS	GP-DPFA	PGDS	GP-DPFA
GDELT	365	1.71	S	**0.428** ±0.06	0.617 ±0.06	**1.491** ±0.22	1.599 ±0.21
			F	**0.432** ±0.09	0.494 ±0.08	**1.224** ±0.19	1.263 ±0.21
ICEWS	365	1.26	S	**0.334** ±0.02	0.372 ±0.01	**1.003** ±0.13	1.021 ±0.14
			F	**0.299** ±0.05	0.313 ±0.05	**0.646** ±0.13	0.673 ±0.14
SOTU	225	1.49	S	**0.216** ±0.00	0.226 ±0.00	**0.365** ±0.00	0.374 ±0.00
			F	0.172 ±0.00	**0.169** ±0.00	0.295 ±0.00	**0.289** ±0.00
DBLP	14	1.73	S	0.370 ±0.00	**0.356** ±0.00	0.604 ±0.00	**0.591** ±0.00
			F	**0.370** ±0.00	0.408 ±0.00	**0.778** ±0.00	0.790 ±0.00
NIPS	17	0.89	S	2.133 ±0.00	**1.199** ±0.00	9.375 ±0.00	**7.893** ±0.00
			F	1.173 ±0.00	**0.949** ±0.00	15.065 ±0.00	**12.445** ±0.00

tasks when we used MRE to measure the models' performance; when we used MAE, the PGDS outperformed the other models on five of the tasks.

In Table 6.2, we also report the results for the experiments comparing the PGDS to GP-DPFA using all the features. The superiority of the PGDS over GP-DPFA is even more pronounced in these results. We hypothesize that the difference between these models is related to the burstiness of the data. For both error measures, the only data set for which GP-DPFA outperformed the PGDS on both tasks was the NIPS data set. This data set has a substantially lower average burstiness score than the other data sets. We provide visual evidence in Fig. 6.2, where we display $y_v^{(t)}$ over time for the top four features in the NIPS and ICEWS data sets. For the former, the features evolve smoothly; for the latter, they exhibit bursts of activity.

6.3.4 Exploratory Analysis

We also explored the latent structure inferred by the PGDS. Because its parameters are positive, they are easy to interpret. In Fig. 6.1, we depict three components inferred from the NIPS data set. By examining the time-step factors and feature fac-

tors for these components, we see that they capture the decline of research on neural networks between 1987 and 2003, as well as the rise of Bayesian methods in machine learning. These patterns match our prior knowledge.

In Fig. 6.4, we depict the three components with the largest component weights inferred by the PGDS from a matrix of 2003 GDELT data. The top component is in blue, the second is in green, and the third is in red. For each component, we also list the sixteen features (directed pairs of countries) with the largest feature factors. The top component (blue) is most active in March and April, 2003. Its features involve USA, Iraq (IRQ), Great Britain (GBR), Turkey (TUR), and Iran (IRN), among others. This component corresponds to the 2003 invasion of Iraq. The second component (green) exhibits a noticeable increase in activity immediately after April, 2003. Its top features involve Israel (ISR), Palestine (PSE), USA, and Afghanistan (AFG). The third component exhibits a large burst of activity in August, 2003, but is otherwise inactive. Its top features involve North Korea (PRK), South Korea (KOR), Japan (JPN), China (CHN), Russia (RUS), and USA. This component corresponds to the six-party talks—a series of negotiations between these six countries for the purpose of dismantling North Korea's nuclear program. The first round of talks occurred during August 27–29, 2003.

In Fig. 6.3, we also show the component weights for the top ten components, along with the corresponding subset of the transition matrix Π. There are two components with weights greater than one: the components that are depicted in blue and green in Fig. 6.4. The transition weights in the corresponding rows of Π are also large, meaning that other components are likely to transition to them. As we mentioned previously, the GDELT data set was extracted from news corpora. Therefore, patterns in the data primarily reflect patterns in media coverage of international affairs. We therefore interpret the latent structure inferred by the PGDS in the following way: in 2003, the media briefly covered various major events, including the six-party talks,

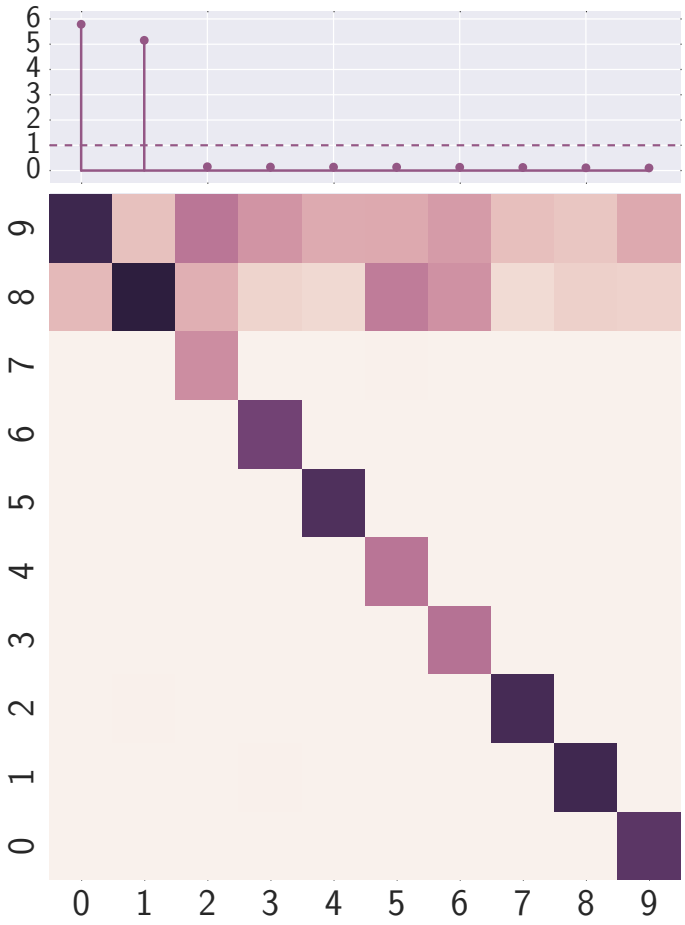


Figure 6.3: Transition structure inferred by the PGDS from the 2003 GDELT matrix. *Top:* The component weights for the top ten components; only two of the weights are greater than one. *Bottom:* Transition weights in the corresponding part of the transition matrix. All components are likely to transition to the top two components.

before quickly returning to a backdrop of the ongoing Iraq war and Israeli–Palestinian relations. By inferring the kind of transition structure depicted in Fig. 6.3, the PGDS is able to model persistent, long-term temporal patterns while accommodating the burstiness often inherent to real-world count data. This ability is what enables the PGDS to achieve superior predictive performance over the LDS and GP-DPFA.

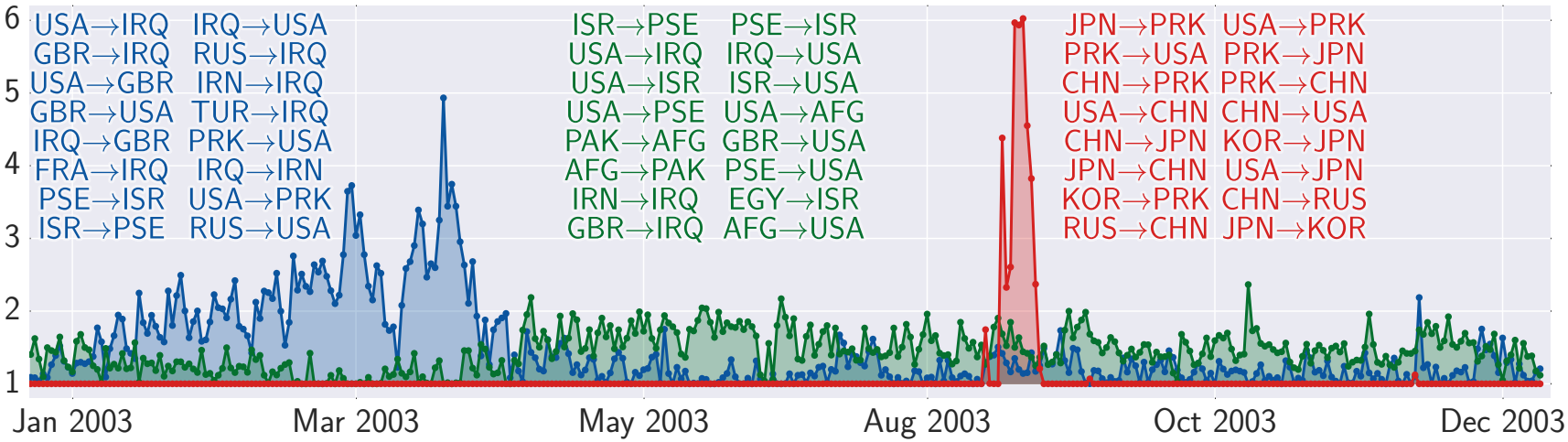


Figure 6.4: The time-step factors for the top three components inferred by the PGDS from the 2003 GDELT matrix. The top component is in blue, the second is in green, and the third is in red. For each component, we also list the features (directed pairs of countries) with the largest feature factors.

CHAPTER 7

POISSON-RANDOMIZED GAMMA DYNAMICAL SYSTEMS

This chapter presents the Poisson-randomized gamma dynamical system (PRGDS), a model for sequentially-observed counts. It is similar in structure to the Poisson–gamma dynamical system (PGDS) of Chapter 6—i.e., it is another APF analogue to linear dynamical systems. For ease of exposition and to facilitate comparison to the PGDS, I will first define the PRGDS for the special case of sequentially-observed count *vectors* $\boldsymbol{y}^{(t)} \in \mathbb{N}_0^V$. Later, I will generalize both PRGDS and PGDS to model sequentially-observed count tensors $\boldsymbol{Y}^{(t)} \in \mathbb{N}_0^{V \times V \times A}$ of dyadic event data and report a suite of experiments on such tensors where a variant of PRGDS either outperforms or equals PGDS's performance on smoothing or forecasting heldout time slices.

The PRGDS is a fundamentally more flexible model family than the PGDS. In Section 7.2.1.1, I detail some of the PGDS's limitations that motivate development of the PRGDS. In particular, the PRGDS is compatible with both gamma and Dirichlet priors over any subset of parameters while the PGDS is not. I compare multiple variants of the PRGDS, including ones using either gamma or Dirichlet priors, to the PGDS in out-of-sample predictive experiments on tensors of dyadic event data (Section 7.4); some variant of the PRGDS either equals or exceeds the smoothing and forecasting performance of the PGDS. Section 7.5 further provides a qualitative comparison of the latent structure inferred by different PRGDS variants and the PGDS. The PRGDS is defined in Section 7.2 and MCMC inference for it is given in Section 7.3. Unlike the PGDS, tractable posterior inference is available without any augmentation of the model.

The PRGDS is based on a novel modeling motif—i.e., the *gamma–Poisson–gamma chain*. Before defining the PRGDS, I provide a simple version of such a chain and describe the two marginal representations it yields. I then show how to perform tractable posterior inference in a gamma–Poisson–gamma chain by appealing to the lesser known Bessel distribution [Yuan and Kalbfleisch, 2000] and introducing a novel univariate discrete probability distribution—i.e., the *size-biased confluent hypergeometric (SCH) distribution*.

7.1 The gamma–Poisson–gamma chain

The PRGDS is based on a novel motif in probabilistic modeling—i.e., the *gamma–Poisson–gamma chain*—that is well-suited to constructing hierarchical priors in APF models. Consider the t^{th} step of simple gamma–Poisson–gamma chain:

$$\theta^{(t-1)} \sim \Gamma\left(\epsilon_0 + h^{(t-1)},\, \beta_0\right), \tag{7.1}$$

$$h^{(t)} \sim \text{Pois}\left(\theta^{(t-1)}\right), \tag{7.2}$$

$$\theta^{(t)} \sim \Gamma\left(\epsilon_0 + h^{(t)},\, \beta_0\right). \tag{7.3}$$

Such a chain alternates between continuous (gamma) and discrete (Poisson) states and features both a conjugate dependency—i.e., of $h^{(t)}$ on $\theta^{(t-1)}$—and a non-conjugate dependency—i.e., of $\theta^{(t)}$ on $h^{(t)}$. Due to the conjugate dependency, the complete conditional of $\theta^{(t-1)}$ is readily available as

$$\left(\theta^{(t-1)} \mid -\right) \sim \Gamma\left(\epsilon_0 + h^{(t-1)} + h^{(t)},\, \beta_0 + 1\right). \tag{7.4}$$

Related to that fact is the fact that all of the continuous states can be marginalized out jointly, yielding the following entirely discrete process:

$$h^{(t)} \sim \text{NB}\left(\epsilon_0 + h^{(t-1)},\, \tfrac{1}{1+\beta_0}\right). \tag{7.5}$$

Despite their non-conjugate dependency, the discrete states can also be marginalized out jointly—when $\epsilon_0 > 0$, doing so yields the following entirely continuous process:

$$\theta^{(t)} \sim \text{RG1}\left(\theta^{(t-1)},\, \epsilon_0,\, \beta_0\right). \tag{7.6}$$

Definition 7.1: Randomized gamma distribution of the first type

A randomized gamma random variable of the first type $\theta \sim \text{RG1}(\lambda, \alpha, \beta)$ [Makarov and Glew, 2010] takes values in the positive reals $\theta > 0$. Its distribution is defined by shape $\alpha > 0$, *hypershape* $\lambda > 0$, rate $\beta > 0$, and PDF:

$$\text{RG1}(\theta; \lambda, \alpha, \beta) = \beta \left(\sqrt{\frac{\beta\,\theta}{\lambda}}\right)^{\alpha-1} e^{-\lambda-\beta\,\theta}\, I_{\alpha-1}\left(2\sqrt{\lambda\,\beta\,\theta}\right), \tag{7.7}$$

where $I_v(a)$ is the modified Bessel function. As $\lambda \to 0$ the RG1 distribution becomes a gamma distribution with shape α and rate β. The *Poisson-randomized gamma distribution* of Zhou et al. [2016] is the limiting case when $\alpha \to 0$. The RG1 distribution is the marginal distribution of $\theta \sim \Gamma(\alpha+y,\ \beta)$ when $y \sim \text{Pois}(\lambda)$:

$$\text{RG1}(\theta; \lambda, \alpha, \beta) = \sum_{y=0}^{\infty} \Gamma(\theta;\, \alpha+y, \beta)\, \text{Pois}\,(y; \lambda)\,. \tag{7.8}$$

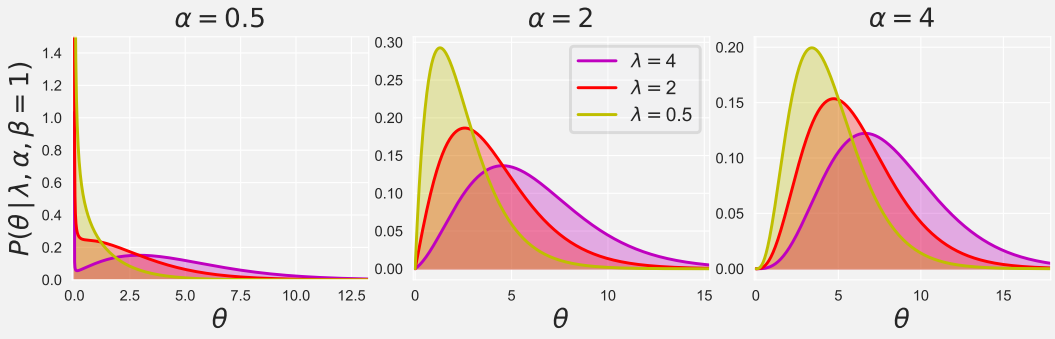


Figure 7.1: PDF of the first-type randomized gamma distribution for $\beta = 1$ and combinations of values for α and λ. As for the gamma distribution, the rate β simply rescales the axes. When $\alpha < 1$ (*left*), the distribution may be bimodal.

A consequence of the marginal distribution $P(\theta^{(t)} \,|\, \theta^{(t-1)}, \epsilon_0, \beta_0)$ being available in closed form is that the complete conditional for $h^{(t)}$ is also available in closed form:

$$\left(h^{(t)} \mid -\right) \sim \text{Bes}\left(\epsilon_0 - 1,\, 2\sqrt{\theta^{(t)}\,\beta_0\,\theta^{(t-1)}}\right) \tag{7.9}$$

Definition 7.2: Bessel distribution

A Bessel random variable $y \sim \text{Bes}\,(\nu,\, a)$ [Yuan and Kalbfleisch, 2000] is a count $y \in \mathbb{N}_0$. Its distribution is defined by *order* ν, *coordinate* a, and PMF:

$$\text{Bes}\,(y;\, \nu, a) = \frac{\left(\frac{a}{2}\right)^{2y+\nu}}{y!\,\Gamma(\nu+y+1)\,I_\nu(a)}, \tag{7.10}$$

where $I_\nu(a)$ is the modified Bessel function of the first kind. The expected value and variance are defined in terms of the *Bessel ratio*—i.e., $R(\nu, a) \triangleq \frac{I_{\nu+1}(a)}{I_\nu(a)}$:

$$\mathbb{E}\,[y \,|\, \nu, a] = \mu_{\nu,a} = \tfrac{1}{2}\,a\,R(\nu, a), \tag{7.11}$$

$$\mathbb{V}\,[y \,|\, \nu, a] = \mu_{\nu,a} + \mu_{\nu,a}^2\left(R(\nu+1, a) - R(\nu, a).\right) \tag{7.12}$$

The Bessel distribution is *underdispersed*—i.e., VMR is always less than one.

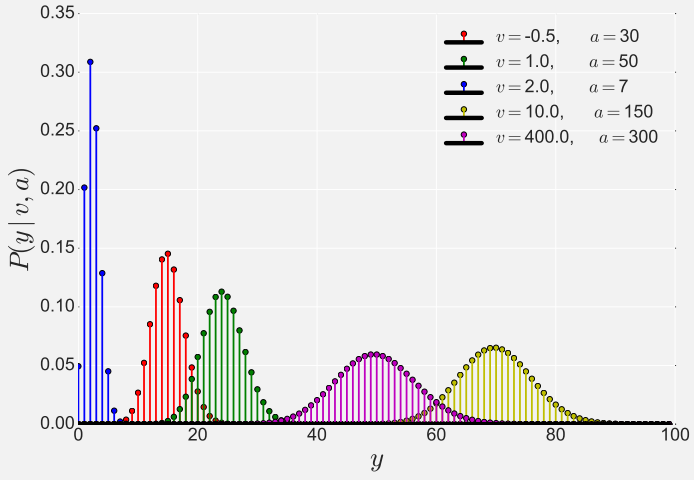


Figure 7.2: PMF of the Bessel distribution.

Definition 7.3: The Bessel as an inverse distribution

Consider a gamma random variable $\theta \sim \Gamma(\alpha+y,\, \beta)$ whose shape parameter is the sum of fixed $\alpha \geq 0$ and a Poisson random variable $y \sim \text{Pois}\,(\lambda)$. Then the inverse distribution of y is the Bessel distribution [Yuan and Kalbfleisch, 2000]:

$$P\left(y \,|\, \theta, \lambda, \alpha, \beta\right) = \text{Bes}\left(y; \alpha-1,\, 2\sqrt{\theta\,\beta\,\lambda}\right). \tag{7.13}$$

In the case where a subset of the discrete or continuous states are observed, sampling the unobserved states from Eqs. (7.4) and (7.9) constitutes a Gibbs sampler that is asymptotically guaranteed to generate samples of them from their exact posterior. We may also treat the entire chain as latent and link it to an observation model. Since the gamma distribution is conjugate to many distributions—e.g., Gaussian, gamma, and Poisson—it is natural to consider a hidden Markov model wherein observations are conditionally independent given the gamma latent states. The following observation model is relevant to this chapter:

$$y^{(t)} \sim \text{Pois}\left(\theta^{(t)}\right). \tag{7.14}$$

If we further assume this observation model, the complete conditional for the discrete states is the same as in Eq. (7.9) and the one for the continuous states becomes:

$$\left(\theta^{(t-1)} \,|\, -\right) \sim \Gamma\left(\epsilon_0+h^{(t-1)}+h^{(t)}+y^{(t)},\, \beta_0+2\right). \tag{7.15}$$

7.1.1 $\epsilon_0 = 0$

When the parameter ϵ_0 is non-zero, it guarantees the shape parameter of $\theta^{(t)}$ will be non-zero (i.e., when $h^{(t)}=0$). If we adopt the convention that the gamma distribution is a Dirac delta spike at zero when its shape parameter is zero, then setting $\epsilon=0$

allows for sparsity (i.e., exact zeros) among the continuous states. In the simple chain above, once a continuous state is set to zero, the chain enters an absorbing condition and dies—i.e., all discrete and continuous states thereafter will be zero, almost surely. However, in the more complex multivariate chains (e.g., those that the PRGDS is built on), the setting of $\epsilon = 0$ allows for sparsity in intermediate states without necessarily killing the chain. This property may be desirable for interpretability and may furthermore be useful for expressing burstiness. However, the absorbing condition means that the Gibbs sampler given above may be non-ergodic when $\epsilon_0 = 0$—i.e., $h^{(t)} = 0$ almost surely if $\theta^{(t)} = 0$ and vice versa; thus a Markov chain based on re-sampling each variable conditioned on the other would never leave this configuration.

We can construct an ergodic Markov by instead block-sampling the discrete and continuous states at each step t jointly—i.e.,

$$\left(\theta^{(t)}, h^{(t)} \,|-\right) \sim P(\theta^{(t)}, h^{(t)} \,|\, \theta^{(t-1)}, h^{(t+1)}, \epsilon_0, \beta_0), \tag{7.16}$$

which can be factorized as:

$$\left(h^{(t)} \,|\, -\backslash\theta^{(t)}\right) \sim P(h^{(t)} \,|\, \theta^{(t-1)}, h^{(t+1)}, \epsilon_0, \beta_0), \tag{7.17}$$

$$\left(\theta^{(t)} \,|\, -\right) \sim P(\theta^{(t)} \,|\, h^{(t)}, h^{(t+1)}, \epsilon_0, \beta_0), \tag{7.18}$$

where the $-\backslash x$ notation means "everything but x". In this factorization, the second line is identical to the complete conditional for $\theta^{(t)}$ given previously; the only unknown is the *incomplete* conditional $P(h^{(t)} \,|\, \theta^{(t-1)}, h^{(t+1)}, \epsilon_0, \beta_0)$. This unknown distribution is the inverse distribution of $h^{(t)}$ in the marginal model where $\epsilon_0 = 0$—i.e.:

$$h^{(t)} \sim \text{Pois}\left(\theta^{(t-1)}\right), \tag{7.19}$$

$$h^{(t+1)} \sim \text{NB}\left(h^{(t)}, p_0\right), \tag{7.20}$$

where I've defined $p_0 \triangleq \frac{1}{1+\beta_0}$ and note that $h^{(t+1)} = 0$, almost surely, if $h^{(t)} = 0$. When the negative binomial has a count-valued first parameter, it is referred to as the *Pascal distribution* and can be represented as the sum of i.i.d. geometric random variables [Johnson et al., 2005, Chapter 5]. The construction above thus defines a compound Poisson distribution [Adelson, 1966]. This construction has been studied previously—the distribution of $h^{(t+1)}$ with $h^{(t)}$ marginalized out is the *Polya–Aeppli distribution* [Johnson et al., 2005, Chapter 9.7], a special case of the *Poisson–Pascal distribution* [Katti and Gurland, 1961]:

$$P\left(h^{(t+1)} \,|\, \theta^{(t-1)}, p_0\right) = \text{Polya–Aeppli}\left(h^{(t+1)};\, \theta^{(t-1)},\, p_0\right). \tag{7.21}$$

Definition 7.4: Polya–Aeppli distribution

Consider a negative binomial random variable $m \sim \text{NB}(h, p)$ with fixed probability parameter $p \in (0, 1)$ and shape parameter that is Poisson-distributed $h \sim \text{Pois}(\theta)$. Then the marginal distribution of m is Polya–Aeppli:

$$P(m \,|\, \theta, p) = \text{Polya–Aeppli}(m;\, \theta,\, p) \tag{7.22}$$

$$= \begin{cases} e^{-p\theta} & \text{if } m = 0 \\ e^{-\theta}\, p^m\, (1-p)\, \theta\, {}_1\text{F}_1\,(m+1;\, 2;\, (1-p)\,\theta) & \text{otherwise} \end{cases} \tag{7.23}$$

where ${}_1\text{F}_1\,(a;\, b;\, z)$ is Kummer's confluent hypergeometric function.

This marginal is the normalizing constant for the desired incomplete conditional:

$$P(h^{(t)} \,|\, h^{(t+1)}, \theta^{(t-1)}, \epsilon_0, p_0) = \frac{P(h^{(t+1)}, h^{(t)} \,|\, \theta^{(t-1)}, \epsilon_0, p_0)}{P(h^{(t+1)} \,|\, \theta^{(t-1)}, \epsilon_0, p_0)} \tag{7.24}$$

$$= \frac{\text{NB}\left(h^{(t+1)};\, h^{(t)},\, p_0\right) \text{Pois}\left(h^{(t)};\, \theta^{(t-1)}\right)}{\text{Polya–Aeppli}\left(h^{(t+1)};\, \theta^{(t-1)},\, p_0\right)}. \tag{7.25}$$

Plugging in terms, we first note that if $h^{(t+1)}=0$, then:

$$P(h^{(t)} \mid h^{(t+1)}=0, \theta^{(t-1)}, \epsilon_0, p_0) = \frac{(1-p_0)^{h^{(t)}} \; \frac{(\theta^{(t-1)})^{h^{(t)}}}{h^{(t)}!} e^{-\theta^{(t-1)}}}{e^{-p_0\,\theta^{(t-1)}}} \tag{7.26}$$

$$= \frac{\left[(1-p_0)\,\theta^{(t-1)}\right]^{h^{(t)}}}{h^{(t)}!}\, e^{-(1-p_0)\theta^{(t-1)}} \tag{7.27}$$

$$= \text{Pois}\left(h^{(t)}; (1-p_0)\,\theta^{(t-1)}\right). \tag{7.28}$$

In words, the incomplete conditional of $h^{(t)}$ is a Poisson when the subsequent discrete state in the chain is zero $h^{(t+1)}=0$. When $h^{(t+1)} > 0$, we first note that $h^{(t)} > 0$, almost surely. Its distribution may be obtained by plugging in to Eq. (7.25). In this case, the resultant distribution is a two parameter discrete distribution that I define as the *size-biased confluent hypergeometric (SCH) distribution* in Definition 7.5:

$$P(h^{(t)} \mid h^{(t+1)} > 0, \theta^{(t-1)}, \epsilon_0, p_0) = \text{SCH}\left(h^{(t)};\, h^{(t+1)},\, (1-p_0)\,\theta^{(t-1)}\right) \tag{7.29}$$

The SCH distribution's PGF (given in Eq. (7.31)) nearly matches that of the *confluent hypergeometric distribution* [Johnson et al., 2005, Chapter 4.12.4]—i.e., $G(s) = \frac{{}_1\text{F}_1(a;b;sz)}{{}_1\text{F}_1(a;b;z)}$. The s in front of the SCH's PGF is the only difference. The class of *weighted distributions* [Johnson et al., 2005, Chapter 2.4.1] are those derived from re-weighting a probability mass function. Denote the PMF of a discrete variable H to be $P(H=h)$ and the distribution of the *recorded variable* H^* to be $P(H^*=h) = \frac{w(h)\,P(H=h)}{\sum_{h'=-\infty}^{\infty} w(h')\,P(H=h')}$ where $w(h) \geq 0$ is the weight for value h. In the special case where the weights are the identity function $w(h)=h$ the recorded variable is said to be *size-biased*. The PGF of a re-weighted confluent hypergeometric distribution, when the weights are $w(h)=h$, is equal to $G(s) = s\frac{{}_1\text{F}_1(a;b;sz)}{{}_1\text{F}_1(a;b;z)}$ which matches the PGF of the SCH distribution for $a=m+1$, $b=2$, and $z=\zeta$. Thus, the name of the SCH distribution is properly descriptive.

Definition 7.5: Size-biased confluent hypergeometric distribution

A size-biased confluent hypergeometric random variable $h \sim \mathrm{SCH}(m, \zeta)$ is a non-zero count $h \in \mathbb{N}$. Its distribution is defined by a non-zero count-valued *population* parameter $m \in \mathbb{N}$, a positive *rate* parameter $\zeta > 0$, and PMF:

$$\mathrm{SCH}(h;\, m, \zeta) = \frac{(m+h+1)!}{m!\, h!\, (h+1)!} \frac{\zeta^{h-1}}{{}_1\mathrm{F}_1\,(m+1;\, 2;\, \zeta)}, \tag{7.30}$$

where ${}_1\mathrm{F}_1\,(a; b; z)$ is Kummer's confluent hypergeometric function. The distribution may be equivalently defined by its probability generating function:

$$G(s) = \mathbb{E}\left[s^h \,|\, m, \zeta\right] = s \frac{{}_1\mathrm{F}_1\,(m+1; 2; s\zeta)}{{}_1\mathrm{F}_1\,(m+1; 2; \zeta)}. \tag{7.31}$$

The expected value is obtained by evaluating the partial derivative of the PGF—i.e., $G'(s) \equiv \frac{\partial}{\partial s} G(s)$—at one:

$$\mathbb{E}\left[h \,|\, m, \zeta\right] = G'(1) = 1 + \frac{\zeta(m+1)}{2} \frac{{}_1\mathrm{F}_1\,(m+2;\, 3;\, \zeta)}{{}_1\mathrm{F}_1\,(m+1;\, 2;\, \zeta)} \tag{7.32}$$

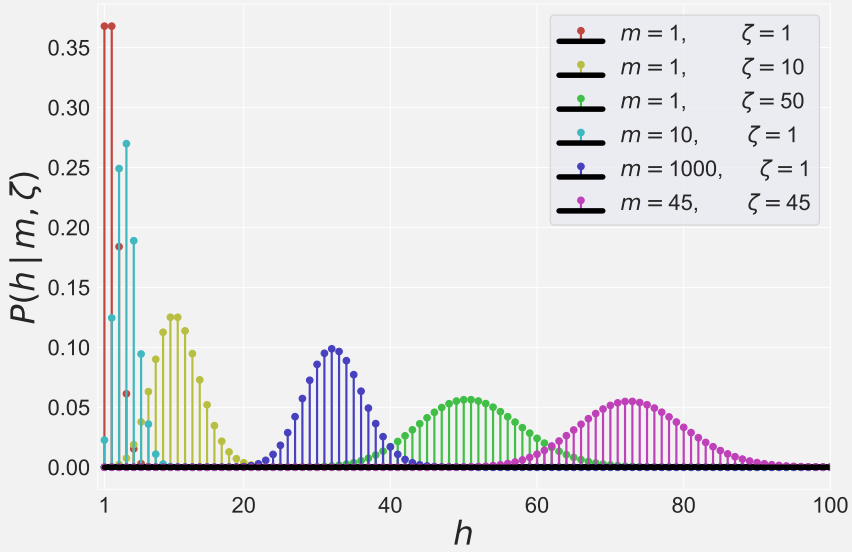


Figure 7.3: PMF of the size-biased confluent hypergeometric distribution.

7.2 Model: Poisson-randomized gamma dynamical systems

PRGDS makes the following generative assumption about that an element $y_v^{(t)}$ of a sequentially-observed count vector $\boldsymbol{y}^{(t)} \in \mathbb{N}_0^K$:

$$y_v^{(t)} \sim \text{Pois}\left(\rho^{(t)} \sum_{k=1}^{K} \lambda_k \, \phi_{kv} \, \theta_k^{(t)}\right). \tag{7.33}$$

This diverges from PGDS only in the inclusion of the per-component weight λ_k. What characterizes PRGDS is its assumption about how the latent states $\boldsymbol{\theta}^{(t)}$ evolve:

$$\theta_k^{(t)} \sim \Gamma\left(\epsilon_0 + h_{k\cdot}^{(t)}, \, \tau \, \beta^{(\theta)}\right), \tag{7.34}$$

$$h_{k\cdot}^{(t)} \sim \text{Pois}\left(\tau \sum_{k_2=1}^{K} \pi_{kk_2} \theta_{k_2}^{(t-1)}\right). \tag{7.35}$$

PRGDS posits an intermediate layer of count-valued discrete latent states $\boldsymbol{h}^{(t)} \in \mathbb{N}_0^K$ that sit between $\boldsymbol{\theta}^{(t-1)}$ and $\boldsymbol{\theta}^{(t)}$. While $\theta_k^{(t)}$ only depends on $h_{k\cdot}^{(t)}$, the Poisson rate of $h_k^{(t)}$ is a linear combination of all K components of the continuous latent state at the previous time step. Note that Eq. (7.35) conforms to the canonical form of APF; thus, $h_{k\cdot}^{(t)}$ has a latent source representation:

$$h_{k\cdot}^{(t)} \triangleq \sum_{k_2=1}^{K} h_{kk_2}^{(t)}, \tag{7.36}$$

$$h_{kk_2}^{(t)} \sim \text{Pois}\left(\tau \, \pi_{kk_2} \, \theta_{k2}^{(t-1)}\right). \tag{7.37}$$

At the first time step $t=1$, $h_{k\cdot}^{(1)}$ is drawn differently:

$$h_{k\cdot}^{(1)} \sim \text{Pois}\left(\tau \, \lambda_k\right), \tag{7.38}$$

where λ_k is the same as the one in Eq. (7.33).

7.2.1 Priors over Poisson rate parameters $\rho^{(t)}$, ϕ_{kv}, and λ_k

As with the PGDS, we refer to the the model as *stationary* if $\rho^{(t)} = \rho$ and assume the following gamma prior for $\rho^{(t)}$ (or ρ):

$$\rho^{(t)} \sim \Gamma(\alpha_0, \alpha_0). \tag{7.39}$$

The PRGDS is compatible with the full range of priors over ϕ_{kv}. In this chapter, I consider both independent gamma priors,

$$\phi_{kv} \sim \Gamma\left(\alpha_0,\, \alpha_0\, \beta^{(\phi)}\right), \tag{7.40}$$

and Dirichlet priors,

$$\tilde{\boldsymbol{\phi}}_k \sim \mathrm{Dir}\left(\boldsymbol{\alpha}_0 \mathbf{1}_V\right), \tag{7.41}$$

where $\boldsymbol{\alpha}_0 \mathbf{1}_V$ denotes the V-length vector $(\alpha_0, \ldots, \alpha_0)$ and the tilde notation emphasizes that the vector is normalized—i.e., $\sum_{v=1}^{V} \tilde{\phi}_{kv} = 1$.

The per-component weights λ_k are drawn from the following prior,

$$\lambda_k \sim \Gamma\left(\frac{\gamma_0}{K},\, \beta^{(\lambda)}\right), \tag{7.42}$$

which promotes shrinkage, particularly when K is set to a large value. This shrinkage prior is common—e.g., see the model of [Acharya et al., 2015]—and admits an interpretation as a truncated Bayesian non-parametric prior. Note that if λ_k is small, it shrinks the contribution of the k^{th} component to the Poisson rate of the data $y_v^{(t)}$ as well as to the latent discrete state at the first time step $h_k^{(1)}$.

The rate parameters are drawn from a non-informative gamma prior:

$$\beta^{(\phi)}, \beta^{(\lambda)} \sim \Gamma(\alpha_0, \alpha_0). \tag{7.43}$$

7.2.1.1 Limitations of PGDS that motivate PRGDS

The tractable posterior inference algorithm presented for PGDS, presented in the last chapter, is only available[1] when $\theta_k^{(t)}$ is the only component-specific variable—i.e., the only variable subscripted by k—in the Poisson rate of the latent source aggregation $y_k^{(t)} \sim \text{Pois}\left(\theta_k^{(t)} \cdots\right)$. This implies two constraints on PGDS:

1. PGDS cannot introduce per-components weights λ_k into the Poisson rate of $y_v^{(t)}$, unlike here in Eq. (7.33).

2. PGDS can only support priors over ϕ_{kv} that satisfy $\sum_{v=1}^{V} \phi_{kv} = 1$ (so that they sum out of the rate of the latent source aggregation $y_k^{(t)}$).

These are both non-trivial constraints. Per-component weights in the Poisson rate are frequently imposed—e.g., by Acharya et al. [2015]—to promote shrinkage among the components. This frequently obviates the need for cross-validation to select K and prevents overfitting when K is set to a large value. To promote shrinkage, PGDS instead relies on a hierarchical prior over the transition matrix that necessitates a complicated augmentation scheme for tractable inference.

Independent gamma priors over all parameter matrices allow APF models to express complex patterns of overdispersion, particularly in the tensor setting. Moreover, as discussed in Section 3.4, it is only with independent gamma priors that conjugacy can still be exploited when missing data is marginalized out instead of imputed.

Finally, the inference algorithm presented for the PGDS is also only compatible with Dirichlet priors over the columns of the transitions matrix Π. It is necessary that $\sum_{k_1=1}^{K} \pi_{k_1 k} = 1$ in Eq. (6.16) to recurse with the marginalization of the preceding latent state. The PRGDS, on the other hand, naturally yields posterior inference that is compatible with any non-negative matrix for Π.

---

[1] Specifically, if Eq. (6.11) is not the same for all components k, the alternative model specification in Section 6.2.3 no longer yields the multinomial step in Eq. (6.30) that is conjugate to the Dirichlet priors over the columns of the transition matrix $\boldsymbol{\pi}_k$.

7.2.2 Concentration parameter τ

The positive *concentration parameter* $\tau > 0$ serves a similar function in PRGDS as in PGDS—it controls the variance of the latent dynamics while only contributing weakly to its expected value by a factor of $\frac{\epsilon_0}{\beta^{(\theta)}}$:

$$\mathbb{E}\left[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \Pi, \tau, \beta^{(\theta)}, \epsilon_0\right] = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{h}^{(t)}, \tau, \beta^{(\theta)}, \epsilon_0\right] \mid \boldsymbol{\theta}^{(t-1)}, \Pi, \tau, \beta^{(\theta)}, \epsilon_0\right] \tag{7.44}$$

$$= \frac{\epsilon_0 + \mathbb{E}\left[\boldsymbol{h}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \Pi, \tau\right]}{\tau\, \beta^{(\theta)}} \tag{7.45}$$

$$= \frac{\epsilon_0}{\beta^{(\theta)}\tau} + \frac{\Pi\, \boldsymbol{\theta}^{(t-1)}}{\beta^{(\theta)}}. \tag{7.46}$$

When $\epsilon_0 = 0$ and $\beta^{(\theta)} = 1$, this matches the canonical form of linear dynamical systems:

$$\mathbb{E}\left[\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}, \Pi, \tau, \beta^{(\theta)} = 1, \epsilon_0 = 0\right] = \Pi\, \boldsymbol{\theta}^{(t-1)}. \tag{7.47}$$

Unlike in PGDS, a gamma prior over τ is conjugate to the distributions it appears in—i.e., to the gamma in Eq. (7.34) and the Poisson in Eq. (7.35). It is thus natural in PRGDS to infer τ as a latent variable, assuming the following non-informative prior:

$$\tau \sim \Gamma(\alpha_0,\, \alpha_0) \tag{7.48}$$

While the long-term expectation of the system—i.e., $\mathbb{E}\left[\boldsymbol{\theta}^{(\infty)} \mid \boldsymbol{\theta}^{(t)}, -\right]$—may diverge in the case of $\epsilon_0 > 0$, the parameter $\beta^{(\theta)}$ can temper the additive effect of ϵ_0 in the short term; it is thus useful to infer jointly alongside τ:

$$\beta^{(\theta)} \sim \Gamma(c_0, c_0). \tag{7.49}$$

We may set c_0 to a *large* value (e.g., $c_0 = 10$) to encode the prior belief that $\beta^{(\theta)} \approx 1$.

7.2.3 Marginalizing out $h_{k\cdot}^{(t)}$

Although $\theta_k^{(t)}$ has a non-conjugate dependence on $h_{k\cdot}^{(t)}$, we may still marginalize out $h_{k\cdot}^{(t)}$ in closed form; doing so yields the following entirely continuous process:

$$\theta_k^{(t)} \sim \mathrm{RG1}\left(\tau \sum_{k_2=1}^{K} \pi_{kk_2} \theta_{k_2}^{(t-1)},\, \epsilon_0,\, \tau\, \beta^{(\theta)}\right). \tag{7.50}$$

7.2.4 Marginalizing out $\theta_k^{(t)}$

Unlike the PGDS, this model does not assume any non-conjugate dependencies on the continuous latent states $\theta_k^{(t)}$. As a result, by appealing to the gamma–Poisson construction of the negative binomial (Definition 3.16) and Poisson–multinomial thinning (Definition 3.4) we can construct a representation of the model wherein all $\theta_k^{(t)}$ variables are marginalized out. Consider the latent source aggregation $y_k^{(t)}$:

$$y_k^{(t)} \sim \mathrm{Pois}\left(\theta_k^{(t)}\, \omega_k^{(t)}\right), \tag{7.51}$$

$$\omega_k^{(t)} \triangleq \rho^{(t)}\, \lambda_k \sum_{v=1}^{V} \phi_{kv}. \tag{7.52}$$

Consider also the latent source aggregation $h_{\cdot k}^{(t+1)} \triangleq \sum_{k_1=1}^{K} h_{kk_2}^{(t)}$:

$$h_{\cdot k}^{(t+1)} \sim \mathrm{Pois}\left(\theta_k^{(t)}\, \zeta_k^{(t)}\right), \tag{7.53}$$

$$\zeta_k^{(t+1)} \triangleq \tau \sum_{k_1=1}^{K} \pi_{k_1 k}. \tag{7.54}$$

Defining the sum of those two aggregations $m_k^{(t)} \triangleq y_k^{(t)} + h_{\cdot k}^{(t+1)}$ yields a Poisson random variable—i.e., $m_k^{(t)} \sim \mathrm{Pois}\left(\theta_k^{(t)}\, (\omega_k^{(t)} + \zeta_k^{(t+1)})\right)$—that depends exclusively on $\theta_k^{(t)}$; we may thus marginalize out $\theta_k^{(t)}$ to obtain the following generative process:

$$m_k^{(t)} \sim \mathrm{NB}\left(\epsilon_0 + h_{k\cdot}^{(t)},\ \frac{\omega_k^{(t)} + \zeta_k^{(t)}}{\omega_k^{(t)} + \zeta_k^{(t+1)} + \tau}\right), \tag{7.55}$$

$$\left(y_k^{(t)}, h_{\cdot k}^{(t+1)}\right) \sim \mathrm{Multinom}\left(m_k^{(t)},\ \left(\omega_k^{(t)}, \zeta_k^{(t+1)}\right)\right), \tag{7.56}$$

$$\left(h_{k_1 k}^{(t+1)}\right)_{k_1=1}^{K} \sim \mathrm{Multinom}\left(h_{\cdot k}^{(t+1)},\ (\pi_{k_1 k})_{k_1=1}^{K}\right). \tag{7.57}$$

Note that I am using the parameterization of the multinomial with *unnormalized* rates (Definition 3.3) and that the first can be equivalently represented as a binomial.

7.2.5 Priors over transition matrix Π

Marginalizing out $\theta_k^{(t)}$ yields the multinomial draw in Eq. (7.57) which would be conjugate to a Dirichlet prior over the k^{th} column of the transition matrix:

$$\boldsymbol{\pi}_k \sim \mathrm{Dir}(\boldsymbol{\rho}_0). \tag{7.58}$$

In the last chapter, the marginal representation of PGDS (Section 6.2.3) also yielded an analogous multinomial draw; Dirichlet priors over the columns of Π were thus natural. As mentioned above, the PRGDS is compatible with a broader class of priors over Π, including independent gamma priors—e.g.,

$$\pi_{k_1 k_2} \sim \Gamma(\alpha_0,\ \alpha_0\, \beta^{(\pi)}), \tag{7.59}$$

which would be conjugate to the Poisson distribution of $h_{k k_2}^{(t)}$ (under the original representation of the model) and do not constrain the sum of either the rows or columns of the transition matrix; they are thus are capable of expressing a broader range of dynamics. In all the experiments reported in this chapter, I impose Dirichlet priors so as to limit the differences between PGDS and PRGDS; future investigation of PRGDS with gamma-distributed transition elements is motivated. I note however that the Dirichlet priors defined in PGDS involve a hierarchy intended to impose shrinkage.

Inference in this construction requires complicated augment-and-conquer techniques. PRGDS, on the other hand, can impose shrinkage over components more directly than PGDS via the λ_k weights. Thus, the simple Dirichlet prior given above, which directly yields a closed-form complete conditional without further augmentation, may suffice.

7.2.6 Tensor generalization

The PRGDS has thus far been described for sequentially-observed count *vectors*. However, consider a sequentially-observed M-mode count tensor $\boldsymbol{Y}^{(t)} \in \mathbb{N}_0^{D_1 \times \cdots \times D_M}$. The PRGDS can be generalized to model this data via the CP decomposition by assuming that a single entry $y_{\boldsymbol{\delta}}^{(t)}$ of type $\boldsymbol{\delta} \equiv (\delta_1, \ldots \delta_M)$ at time t is drawn:

$$y_{\boldsymbol{\delta}}^{(t)} \sim \text{Pois}\left(\rho^{(t)} \sum_{k=1}^{K} \lambda_k \, \theta_k^{(t)} \prod_{m=1}^{M} \phi_{k\,\delta_m}^{(m)}\right), \tag{7.60}$$

where there are now M parameter matrices, each $\Phi^{(m)}$ of size $K \times D_m$. For the PRGDS, we may consider both conjugate priors, either independent gamma priors,

$$\phi_{k\,d_m}^{(m)} \sim \Gamma(\alpha_0,\, \alpha_0\, \beta^{(m)}) \;\text{ for } d_m \in [D_m], \tag{7.61}$$

or Dirichlet priors,

$$\tilde{\boldsymbol{\phi}}_k^{(m)} \sim \text{Dir}\left(\boldsymbol{\alpha}_0 \mathbf{1}_{D_m}\right). \tag{7.62}$$

The PGDS is more restricted in that it is incompatible with the per-component weights λ_k and is restricted to Dirichlet priors that guarantee $\sum_{d_m=1}^{D_m} \tilde{\phi}_{k d_m}^{(m)} = 1$.

Dyadic event data can be represented as a sequentially-observed 3-mode tensor $\boldsymbol{Y}^{(t)} \in \mathbb{N}_0^{V \times V \times A}$ where V is the number of countries and A is the number of action-types. PRGDS assumes the count $y^{(t)}_{i \xrightarrow{a} j}$ of type $i \xrightarrow{a} j$ at time t is drawn:

$$y^{(t)}_{i \xrightarrow{a} j} \sim \text{Pois}\left(\rho^{(t)} \sum_{k=1}^{K} \lambda_k \, \theta_k^{(t)} \phi_{ki}^{(1)} \phi_{kj}^{(2)} \phi_{ka}^{(3)}\right). \tag{7.63}$$

Finally, I note that generalizing both models to the Tucker decomposition is also possible; however, I only consider the CP version of both models in this chapter.

7.3 MCMC inference

In this section I rely on the identities presented in Section 7.1 to provide the complete conditionals for the variables with a non-standard relationship to others.

7.3.1 Sampling the discrete states when $\epsilon_0 > 0$

The complete conditional for the first discrete state is:

$$\left(h^{(1)}_{k\cdot} \mid -\right) \sim \text{Bessel}\left(\epsilon_0 - 1,\ 2\tau\sqrt{\beta^{(\theta)}\, \theta_k^{(1)}\, \lambda_k}\right). \tag{7.64}$$

And for those thereafter is $t = 2, \ldots, T$:

$$\left(h^{(t)}_{k\cdot} \mid -\right) \sim \text{Bessel}\left(\epsilon_0 - 1,\ 2\tau\sqrt{\beta^{(\theta)}\, \theta_k^{(t)} \sum_{k_2=1}^{K} \pi_{kk_2} \theta_{k_2}^{(t-1)}}\right), \tag{7.65}$$

$$\left(\left\{h^{(t)}_{kk_2}\right\}_{k_2=1}^{K} \mid -\right) \sim \text{Multinom}\left(h^{(t)}_{k\cdot},\ \left(\pi_{kk_2}\, \theta_{k_2}^{(t-1)}\right)_{k_2=1}^{K}\right). \tag{7.66}$$

Devroye [2002] provides four efficient rejection samplers for sampling from the Bessel distribution. I have open-sourced a fast Cython implementation of these algorithms.

7.3.2 Sampling the discrete states when $\epsilon_0 = 0$

The *incomplete* conditional for the first discrete state is:

$$\left(h_{k\cdot}^{(1)} \mid -\backslash\theta_k^{(1)}\right) \sim \text{SCH}\left(y_k^{(1)} + h_{\cdot k}^{(2)},\ \frac{\tau^2\,\lambda_k}{\omega_k^{(1)} + \zeta_k^{(1)} + \tau}\right), \tag{7.67}$$

where $\omega_k^{(1)}$ and $\zeta_k^{(1)}$ are defined in equations 7.52 and 7.54, respectively. The incomplete conditionals for those thereafter is $t = 2, \ldots, T$:

$$\left(h_{k\cdot}^{(t)} \mid -\backslash\theta_k^{(t)}\right) \sim \text{SCH}\left(y_k^{(t)} + h_{\cdot k}^{(t+1)},\ \frac{\tau^2 \sum_{k_2=1}^{K} \pi_{kk_2}\theta_{k_2}^{(t-1)}}{\omega_k^{(t)} + \zeta_k^{(t)} + \tau}\right). \tag{7.68}$$

7.3.3 Sampling the continuous states:

By gamma–Poisson conjugacy, we have:

$$\left(\theta_k^{(t)} \mid -\right) \sim \Gamma\left(\epsilon_0 + h_{k\cdot}^{(t)} + y_k^{(t)} + h_{\cdot k}^{(t+1)},\ \ \tau\,\beta^{(\theta)} + \omega_k^{(t)} + \zeta_k^{(t)}\right) \tag{7.69}$$

7.3.4 Sampling the concentration parameter:

By gamma–gamma and gamma–Poisson conjugacy, we have:

$$\left(\tau \mid -\right) \sim \Gamma\left(\alpha_0 + TK\epsilon_0 + 2\,h_{\cdot\cdot}^{(\cdot)},\ \alpha_0 + \lambda_\cdot + \beta^{(\theta)}\theta_\cdot^{(\cdot)} + \sum_{k=1}^{K}\sum_{t=2}^{T-1}\sum_{k_2=1}^{K}\pi_{kk_2}\theta_{k_2}^{(t-1)}\right). \tag{7.70}$$

7.3.5 Sampling the transition matrix:

By Dirichlet–multinomial conjugacy:

$$\left(\boldsymbol{\pi}_k \mid -\right) \sim \text{Dir}\left(\rho_{01} + h_{1k}^{(\cdot)}, \ldots, \rho_{0K} + h_{Kk}^{(\cdot)}\right). \tag{7.71}$$

7.3.6 Sampling the per-component weights:

By gamma–Poisson conjugacy:

$$\left(\lambda_k \mid -\right) \sim \Gamma\left(\frac{\gamma_0}{K} + h_{k\cdot}^{(1)} + \sum_{t=1}^{T} y_k^{(t)},\ \beta^{(\lambda)} + \tau + \left(\sum_{t=1}^{T}\rho^{(t)}\theta_k^{(t)}\right)\left(\sum_{v=1}^{V}\phi_{kv}\right)\right). \tag{7.72}$$

7.4 Predictive analysis

This section reports a suite of experiments that test the smoothing and forecasting performance of PGDS and PRGDS on tensors of dyadic event counts.

7.4.1 Data sets

In these experiments, I used two tensors of dyadic event counts, one based on events from the ICEWS data set [Boschee et al., 2015] and another based on events from the GDELT data set [Leetaru and Schrodt, 2013]. The tensor of ICEWS data includes all events from 1995 to 2013. I define a time step to be a calendar month; thus the number of time steps $T = 228$. ICEWS includes $V = 249$ unique country actors and $A = 20$ high-level action types. The full tensor is thus $\boldsymbol{Y}^{(\text{ICEWS})} \in \mathbb{N}_0^{228 \times 249 \times 249 \times 20}$. The total number of events (i.e., sum of all the entries) is 6,427,715. The empirical variance-to-mean ratio (VMR) of the counts is VMR= 57.

The tensor of GDELT data includes all events from 2003 to 2008. I again define a time step to be a calendar month; thus the number of time steps is $T = 223$. There are $V = 223$ unique country actors in the GDELT data set. The size of GDELT tensor is thus $\boldsymbol{Y}^{(\text{GDELT})} \in \mathbb{N}_0^{72 \times 223 \times 223 \times 20}$. GDELT events are more numerous than ICEWS; moreover GDELT data tends to be burstier due to less deduplication effort on the collection end. The number of events in this tensor is 13,722,933—which is more than twice the number in the ICEWS tensor that covers over three times the number of time steps—and the VMR is 142.

7.4.2 Models

In all of these experiments, I compare the tensor generalization of PGDS to four variants PRGDS—i.e., the four combinations of assuming gamma or Dirichlet priors over the parameter matrices and setting either $\epsilon_0 = 0$ or $\epsilon_0 = 1$. Note that the PGDS is only compatible with Dirichlet priors over the parameter matrices (and does have ϵ_0).

7.4.3 Experimental setup

For each tensor, I generated two masks, each of which randomly holds out six non-contiguous time steps $t \in [2, T-2]$; *all* counts $y^{(t)}_{i \xrightarrow{a} j}$ within a heldout time step t are masked. In addition, the last two time steps are always heldout. For each model, each data tensor, and each mask, I ran two independent chains of MCMC of 3,000 iterations, saving every 50^{th} sample after the first 1,000. For PGDS and the Dirichlet variants of PRGDS, the missing entries are imputed as latent variables every iteration while the gamma variants of PRGDS marginalize them out.

7.4.4 Performance metrics

MCMC inference returns a set of S samples of the latent variables from which the Poisson rate of every missing count—e.g., $y^{(t)}_{i \xrightarrow{a} j}$—can be computed to yield a set of S rates—i.e., $\left\{\mu^{(t)(s)}_{i \xrightarrow{a} j}\right\}_{s=1}^{S}$. These rates can be interpreted either as parameters to the heldout likelihood or as predictions of the heldout count (since, under the model, $\mathbb{E}[y^{(t)}_{i \xrightarrow{a} j}] = \mu^{(t)}_{i \xrightarrow{a} j}$). I evaluate the smoothing and forecasting performance of each model in two ways: 1) using the samples to estimate the (rescaled) posterior predictive probability of heldout data and 2) treating samples as predictions and evaluating their mean absolute scaled error (MASE), a standard metric for evaluating forecast error.

7.4.4.1 Rescaled posterior predictive probability (RPPP)f

Consider a set of multi-indices $\Delta^{(\text{miss})}$, each of which $\boldsymbol{\delta}$ corresponds to a heldout count $y^{(\text{miss})}_{\boldsymbol{\delta}}$. The posterior predictive probability of a single heldout count is

$$P\left(y^{(\text{miss})}_{\boldsymbol{\delta}} \,|\, \boldsymbol{Y}^{(\text{obs})}\right) = \int \mathbf{d}\mu_{\boldsymbol{\delta}}\, P\left(y^{(\text{miss})}_{\boldsymbol{\delta}} \,|\, \mu_{\boldsymbol{\delta}}\right)\, P\left(\mu_{\boldsymbol{\delta}} \,|\, \boldsymbol{Y}^{(\text{obs})}\right). \tag{7.73}$$

We can approximate this integral by averaging over samples drawn from the posterior:

$$\approx \frac{1}{S}\sum_{s=1}^{S} P\left(y^{(\text{miss})}_{\boldsymbol{\delta}} \,|\, \mu^{(s)}_{\boldsymbol{\delta}}\right). \tag{7.74}$$

The posterior predictive of all missing data entries can be approximated as

$$P\left(\boldsymbol{Y}^{(\text{miss})} \,|\, \boldsymbol{Y}^{(\text{obs})}\right) \approx \prod_{\boldsymbol{\delta} \in \Delta^{(\text{miss})}} \frac{1}{S} \sum_{s=1}^{S} P\left(y_{\boldsymbol{\delta}}^{(\text{miss})} \,|\, \mu_{\boldsymbol{\delta}}^{(s)}\right). \tag{7.75}$$

This quantity is a product of heldout likelihoods and is sensitive to the number of heldout entries; it is thus difficult to compare two different posterior predictive probabilities computed on different sized test sets. A more interpretable quantity can be obtained by rescaling according to the number of heldout entries $|\Delta^{(\text{miss})}|$:

$$\propto \left[\prod_{\boldsymbol{\delta} \in \Delta^{(\text{miss})}} \frac{1}{S} \sum_{s=1}^{S} P\left(y_{\boldsymbol{\delta}}^{(\text{miss})} \,|\, \mu_{\boldsymbol{\delta}}^{(s)}\right)\right]^{\frac{1}{|\Delta^{(\text{miss})}|}}. \tag{7.76}$$

This quantity is proportional to the posterior predictive but directly comparable across experiments with different numbers of missing entries. It defines the geometric mean of the heldout likelihoods and can be equivalently written as the exponentiated average heldout log-likelihood:

$$= \exp\left[\frac{1}{|\Delta^{(\text{miss})}|} \sum_{\boldsymbol{\delta} \in \Delta^{(\text{miss})}} \frac{1}{S} \sum_{s=1}^{S} \log P\left(y_{\boldsymbol{\delta}}^{(\text{miss})} \,|\, \mu_{\boldsymbol{\delta}}^{(s)}\right)\right]. \tag{7.77}$$

I also note that this quantity is equal to the inverse of *perplexity*, which is defined as:

$$\text{Perp}\left(\boldsymbol{Y}^{(\text{miss})} \,||\, \left\{\boldsymbol{\mu}^{(s)}\right\}_{s=1}^{S}\right) = \exp\left[-\frac{1}{|\Delta^{(\text{miss})}|} \sum_{\boldsymbol{\delta} \in \Delta^{(\text{miss})}} \frac{1}{S} \sum_{s=1}^{S} \log P\left(y_{\boldsymbol{\delta}}^{(\text{miss})} \,|\, \mu_{\boldsymbol{\delta}}^{(s)}\right)\right]. \tag{7.78}$$

7.4.4.2 Mean absolute scaled error (MASE)

A philosophically different approach is to treat each sample $\mu_{\boldsymbol{\delta}}^{(s)}$ as a point-estimate prediction of $y_{\boldsymbol{\delta}}^{(\text{miss})}$. In this case, we may average our predictions to approximate the

expected value of $y_{\boldsymbol{\delta}}^{(\text{miss})}$ under the posterior predictive:

$$\langle y_{\boldsymbol{\delta}}^{(\text{miss})} \rangle \triangleq \mathbb{E}\left[y_{\boldsymbol{\delta}} \,|\, \boldsymbol{Y}^{(\text{obs})} \right] \tag{7.79}$$

$$= \int \mathbf{d}\mu_{\boldsymbol{\delta}} \, P\left(\mu_{\boldsymbol{\delta}} \,|\, \boldsymbol{Y}^{(\text{obs})} \right) \, \mathbb{E}\left[y_{\boldsymbol{\delta}} \,|\, \mu_{\boldsymbol{\delta}} \right] \tag{7.80}$$

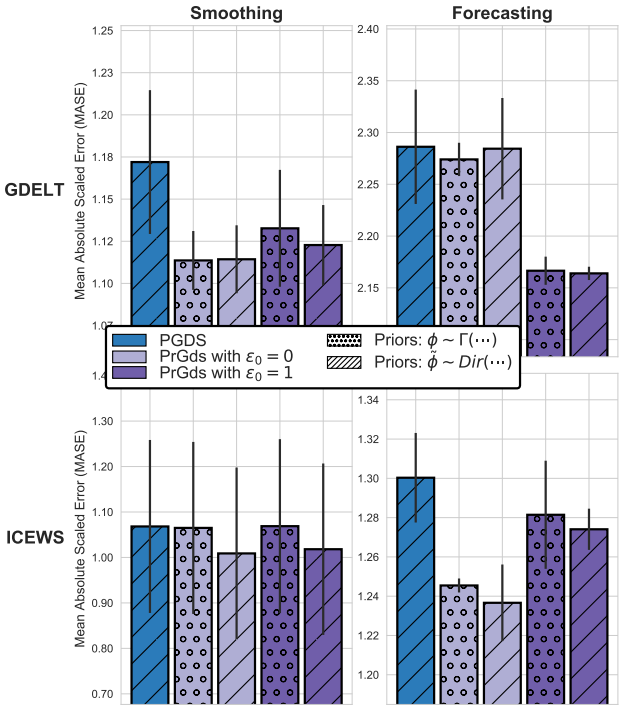
$$\approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}\left[y_{\boldsymbol{\delta}} \,|\, \mu_{\boldsymbol{\delta}} \right] \tag{7.81}$$

$$= \frac{1}{S} \sum_{s=1}^{S} \mu_{\boldsymbol{\delta}}^{(s)}. \tag{7.82}$$

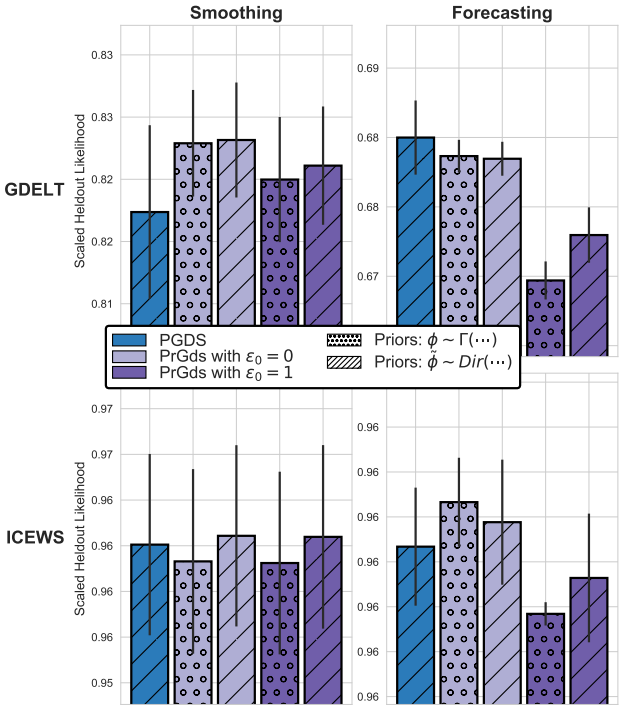
One reason to use the posterior to compute a point-estimate prediction is to be able to compare a Bayesian approach to non-Bayesian or non-probabilistic approaches. To do so, we define a error metric comparing the point estimate $\langle y_{\boldsymbol{\delta}}^{(\text{miss})} \rangle$ to the true value $y_{\boldsymbol{\delta}}^{(\text{miss})}$. In the context of time series, where predictions correspond to smoothing or forecasting, mean absolute scaled error (MASE) is standardly used to assess performance. To ease notation, I will drop the $^{(\text{miss})}$ superscript and write the data being predicted as $y_{\boldsymbol{\delta}}^{(t)}$, where I am now being explicit about the time index t. Hyndman and Koehler [2006] advocate for MASE and define it as

$$\text{MASE}\left(\boldsymbol{Y}^{(\text{miss})} \,||\, \langle \boldsymbol{Y}^{(\text{miss})} \rangle \right) = \frac{1}{|\Delta^{(\text{miss})}|} \sum_{(t,\boldsymbol{\delta}) \in \Delta^{(\text{miss})}} \frac{\left| y_{\boldsymbol{\delta}}^{(t)} - \langle y_{\boldsymbol{\delta}}^{(t)} \rangle \right|}{\frac{1}{T-1} \sum_{t'=2}^{T} \left| y_{\boldsymbol{\delta}}^{(t')} - y_{\boldsymbol{\delta}}^{(t'-1)} \right|}, \tag{7.83}$$

where the numerator is the absolute error and the denominator is the mean absolute error of the *lag-one naïve predictor* for the time-series of type $\boldsymbol{\delta}$. The lag-one naïve prediction of each count is simply the preceding count in the time series. Note that average error of the naïve predictor is computed over all ground truth data entries (not just those at missing time steps). MASE is thus a particular form of relative error that re-weights the absolute error according to the burstiness of the time series.



(a) Mean absolute scaled error (MASE) where *lower is better*.



(b) Rescaled posterior predictive probability (RPPP) where *higher is better*.

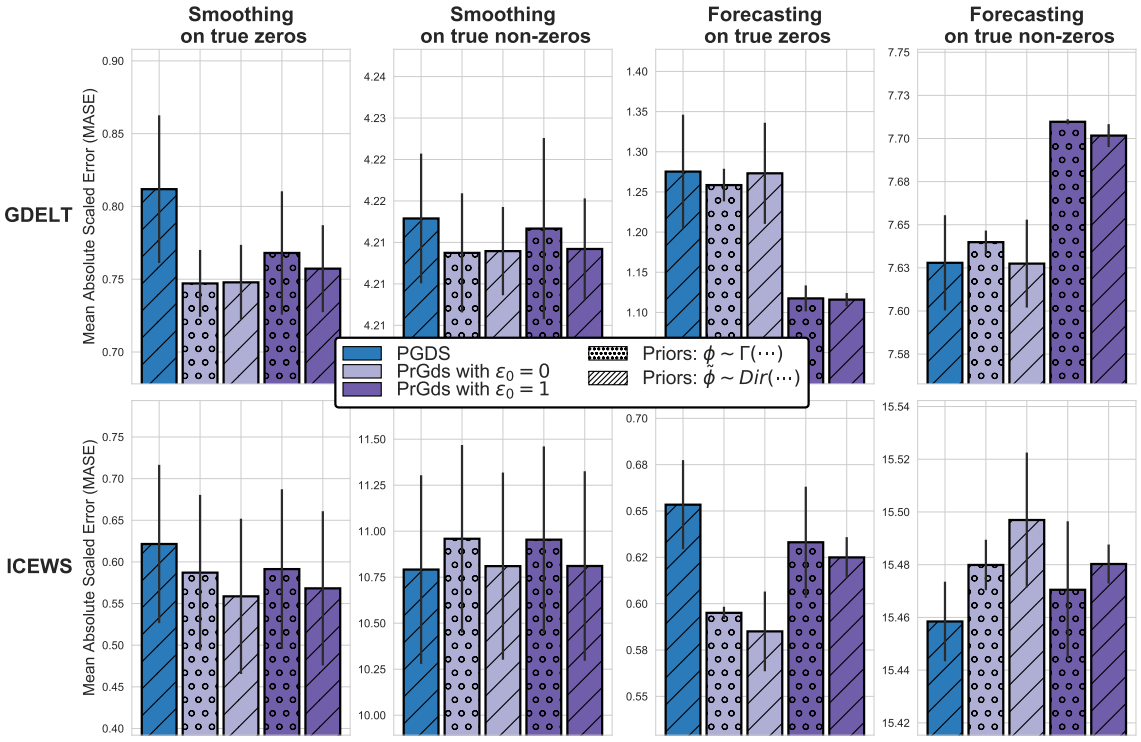
Figure 7.4: Smoothing and forecasting performance of PGDS and four variants of PRGDS on ICEWS and GDELT tensors. Performance is measured using two metrics.

7.4.5 Results

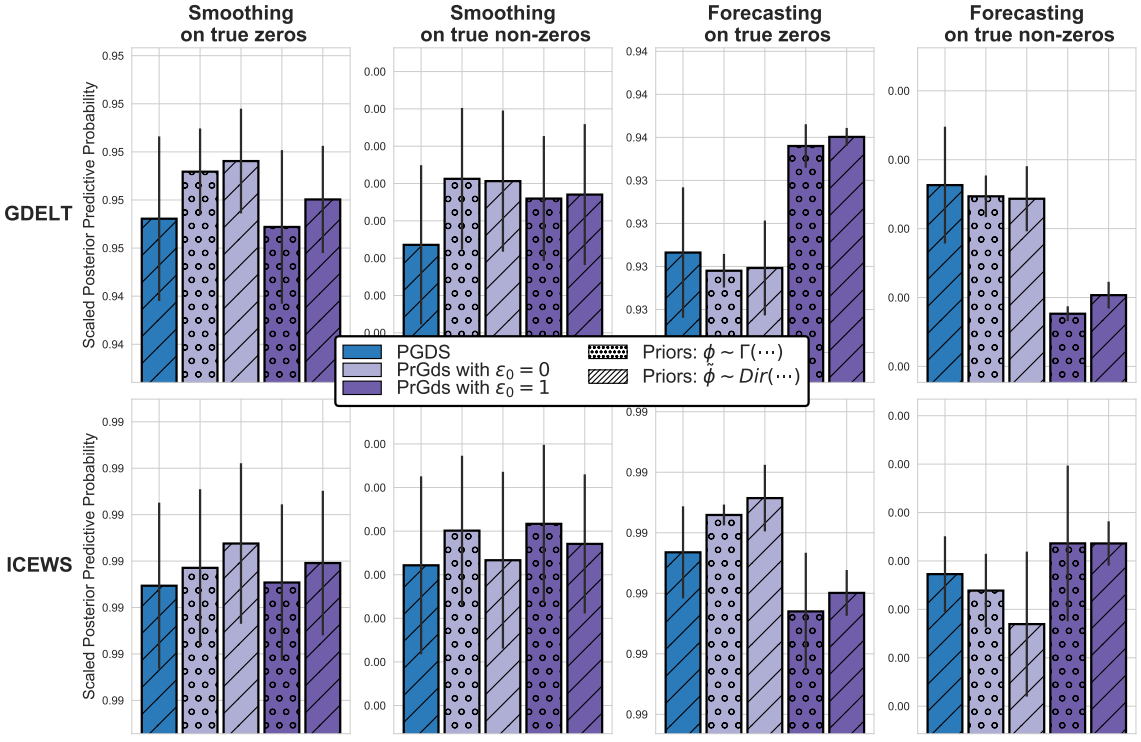
The smoothing and forecasting performance of the PGDS and four PRGDS variants on GDELT and ICEWS is displayed in Fig. 7.4. The top four plots measure MASE (lower is better) while the bottom four measure RPPP (higher is better). In each subplot, the leftmost blue bar corresponds to the PGDS while the rightmost four purple bars correspond to the PRGDS, where the darker shade denotes the $\epsilon_0 = 1$ variants and the lighter shade denotes $\epsilon = 0$. Each bar is textured to reflect the choice of prior: the diagonal dashes denote models with Dirichlet priors (including the PGDS) and circles denote gamma priors. Each MCMC chain, mask, and data set combination yields a single value for both MASE and RPPP. The barplots display each metric averaged across both per-model MCMC chains and both random masks, with error bars denoting the standard deviation.

As measured by both MASE and RPPP, there is no negligible difference in the four models' smoothing performance on ICEWS (i.e., the less bursty data set). The PGDS' smoothing performance on GDELT, for both metrics, is significantly worse than all four variants of the PRGDS, while the $\epsilon_0 = 0$ variants perform slightly better than the $\epsilon_0 = 1$ variants. The performance ranking is similar for forecasting results on ICEWS. For both metrics, the $\epsilon_0 = 0$ PRGDS variants perform the best out of all models. The PGDS performs worse than all four PRGDS variants on MASE. On RPPP, the PGDS is better than the PRGDS with $\epsilon_0 = 1$ and comparable to the PRGDS $\epsilon_0 = 0$ variant with Dirichlet priors, but worse than the one with gamma priors.

The forecasting results on GDELT tell a more complicated story—specifically, the performance ranking of the five models is *reversed* when measuring using MASE versus RPPP. On MASE, the $\epsilon_0 = 1$ PRGDS variants decisively outperform $\epsilon_0 = 0$ PRGDS variants and the PGDS, which all have comparable performance. However, when measuring performance with RPPP, the opposite is true: the $\epsilon_0 = 1$ PRGDS variants perform significantly worse than the other three models. Why might this be?



(a) Mean Absolute Scaled Error ***(lower* is better)** of the posterior expectation of heldout data.



(b) Scaled posterior predictive probability **(*higher* is better)** of heldout data.

Figure 7.5: Performance metrics are faceted by predicting true zeros versus non-zeros.

Fig. 7.5 displays the same results as in Fig. 7.4 but faceted by whether the true underlying count being predicted was zero or not; thus, the smoothing results column in Fig. 7.4 is here split into two columns—i.e., smoothing on true zeros and smoothing on true non-zeros. The same is true for the forecasting column. This faceting reveals the source of the discrepancy in performance ranking between MASE and RPPP. Specifically, the forecasting on GDELT subplots show that the $\epsilon_0 = 1$ PRGDS variants are better than all other models at forecasting non-zeros but are worse at forecasting zeros. RPPP is based on the Poisson PMF, which penalizes *underpredictions*—i.e., $\text{Pois}(x; y) < \text{Pois}(y; x)$ if $x > y$. MASE, like many relative error metrics, instead prioritizes accuracy on small values, and thus penalizes *overpredictions*. Thus, when the aggregate performance of a model is highly faceted across performance on forecasting zeros versus non-zeros, RPPP and MASE will tell different stories.

To summarize, by all metrics, and in all experiments, there was at least one variant of the PRGDS that equaled or exceeded the performance of the PGDS; in some cases, the PGDS performed significantly worse than all PRGDS variants. The $\epsilon_0 = 0$ versus $\epsilon_0 = 1$ seem to outperform the other in different contexts—with $\epsilon_0 = 1$ being better at forecasting non-zeros and $\epsilon_0 = 0$ being slightly better at smoothing in the burstier time series and significantly better at forecasting zeros. Any differences between gamma and Dirichlet-based models were washed out by the differences between the PGDS and PRGDS and between different ϵ_0 variants. In some plots, a small trend is evident—e.g., in the bottom row of Fig. 7.4a, the Dirichlet models are worse at smoothing non-zeros but better at smoothing zeros in ICEWS—-however, no consistent story emerges.

7.5 Exploratory analysis

In this section, I qualitatively compare the latent structure inferred by the different models. To do so, I aligned the inferred components of one model to another using the

Hungarian bipartite matching algorithm run on their inferred matrices of continuous states $\Theta \in \mathbb{R}_+^{K \times T}$. This algorithm finds an alignment of the per-component rows $\boldsymbol{\theta}_k \equiv (\theta_k^{(1)}, \ldots, \theta_k^{(T)})$. While this method does not account for other component-specific information—e.g., the top sender countries—I found it worked well. The k^{th} row provides a signature of when a component was active—due to the burstiness of the data, these signatures are sufficiently unique to allow for alignment based on them alone.

I interpreted components in the same way as in Chapter 4 where a component measures a *multilateral relation* and is described by when it is active $\boldsymbol{\theta}_k$ who the typical senders $\boldsymbol{\psi}_k^{(1)}$ and receivers $\boldsymbol{\psi}_k^{(2)}$ are, and what action types $\boldsymbol{\psi}_k^{(3)}$ are typically used. I visualize $\boldsymbol{\theta}_k$ in chronological order in the top panel (blue). The bottom left stem plot (red) displays the top values of sender parameters, in descending order. If fewer than ten senders account for more than 99% of the mass, I only display their names; otherwise, the top ten are given. The same is true for the middle (green) and right (purple) stem plots, corresponding to receivers and action types.

7.5.1 ICEWS 1995–2013 data

The inferred latent structure across on the ICEWS data set was remarkably aligned across different models. This qualitative observation corroborates the result in the predictive analysis that there were no noticeable differences between the models' smoothing performance. Figures 7.7– 7.9 each give an example of an aligned component that was inferred by the PGDS and two PRGDS variants for $\epsilon_0 = 0$ or $\epsilon_0 = 1$. Figures 7.7 and 7.8 display aligned components corresponding to the Kosovo War and the 2003 American invasion of Iraq, respectively. These were both major international events that consumed the media cycle in their respective time periods; we should thus expect for all models to measure them. Figures 7.9 and 7.10 display aligned components corresponding to more subtle sustained multilateral relations. The component in Fig. 7.9 measures the six-party talks [Wikipedia contributors,

2018f], a sustained period of negotiations about the North Korean nuclear weapons program; the negotiations took place in a series of summits between the six parties—i.e., North Korea, South Korea, United States, Russia, China, and Japan—from 2003 to 2011. Fig. 7.10 displays an aligned component that measures ongoing relations between Russia, Venezuela, and Iran.

I found only a small number of clearly misaligned components. These corresponded to cases in which the reference model inferred a component that no other model inferred. In both ICEWS and GDELT data, I found a pattern in the misalignments: the PRGDS variant with $\epsilon_0 = 0$ inferred components that featured particularly bursty dynamics—i.e., long periods of near-zero rates followed by sudden non-zero rates. An example of such a component is given in Fig. 7.6. In this component, South Sudan is both the top sender and top receiver. The continuous latent states are almost all (exactly) zero before a burst of sustained activity in the last 28 time steps, which correspond to the months of July 2011–October 2013. South Sudan gained its independence from the Republic of Sudan on July 9, 2011. Thus, South Sudan first became a possible sender or receiver in the time step of July 2011. Accordingly, 94% of the continuous latent states preceding July 2011 are exactly zero. Neither the PGDS nor the PRGDS with $\epsilon_0 = 1$ inferred a component in which South Sudan was within the top ten senders or receivers. These models instead allocated events involving South Sudan to less-specific components which featured non-zero activity across time. I speculate that the strong inductive bias towards sparsity uniquely allows the PRGDS with $\epsilon_0 = 0$ to identify components that measure substantively specific components whose activity is highly localized in time.

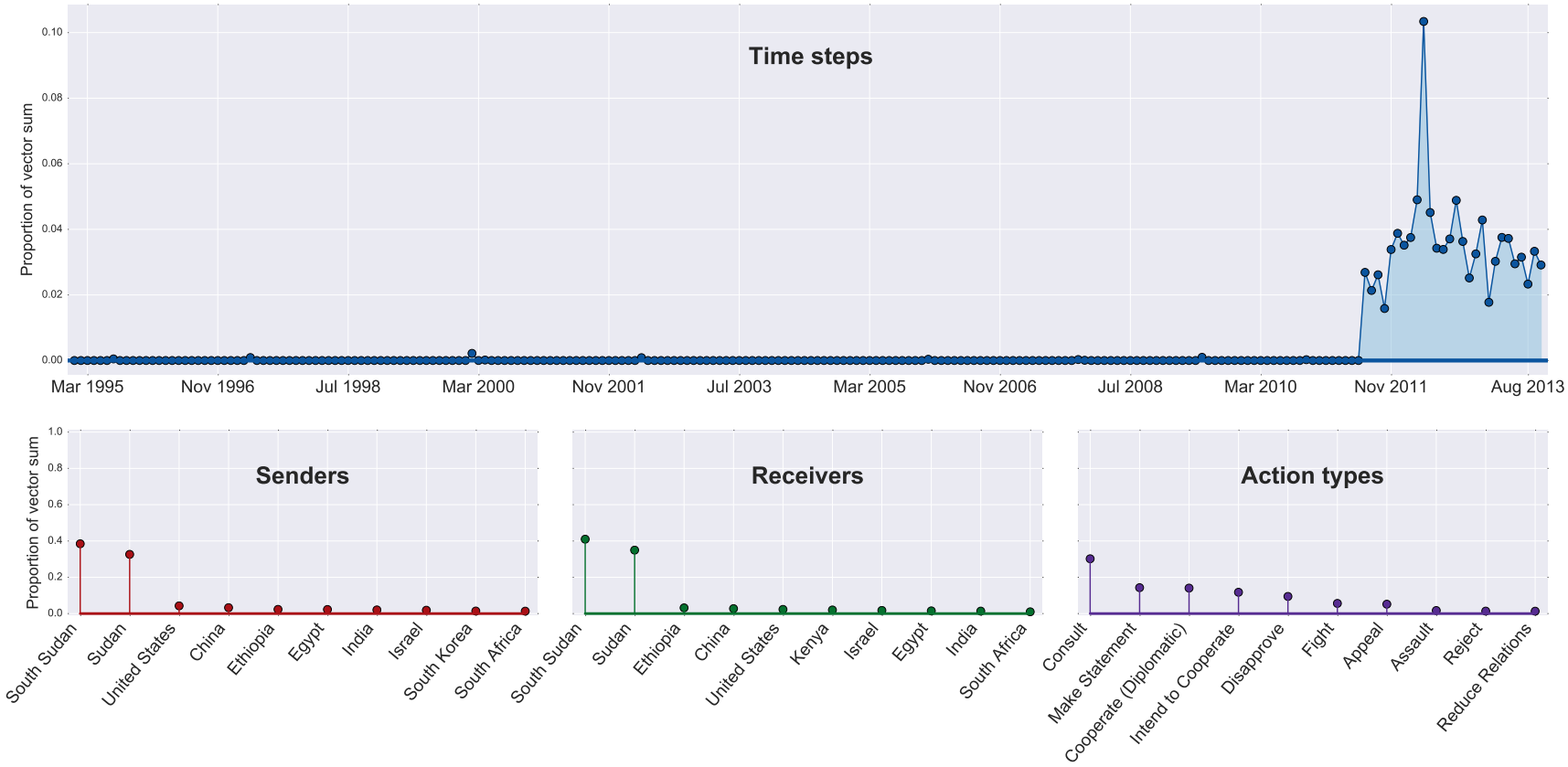
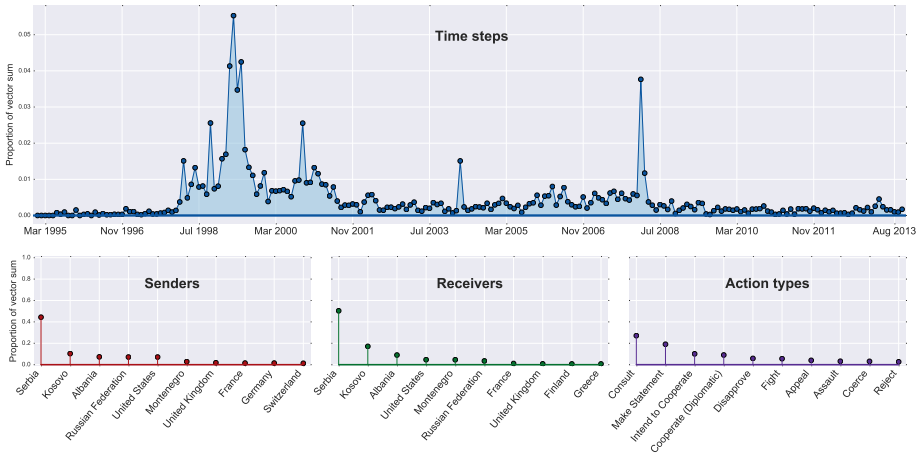
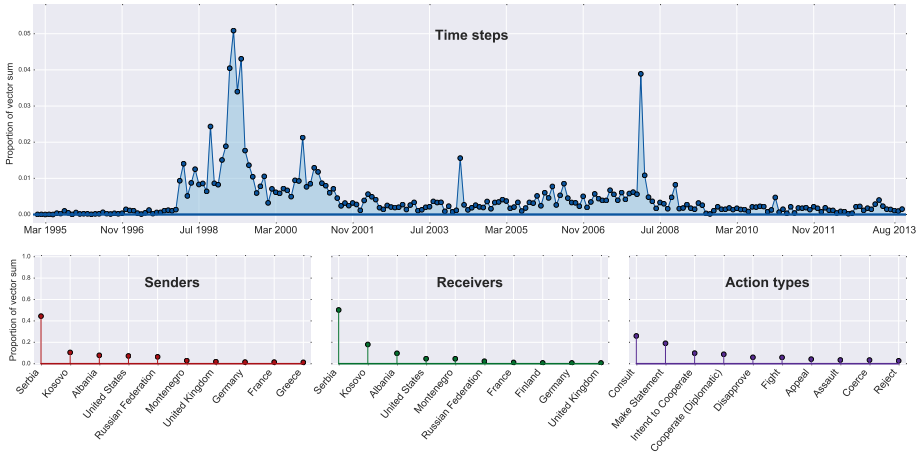


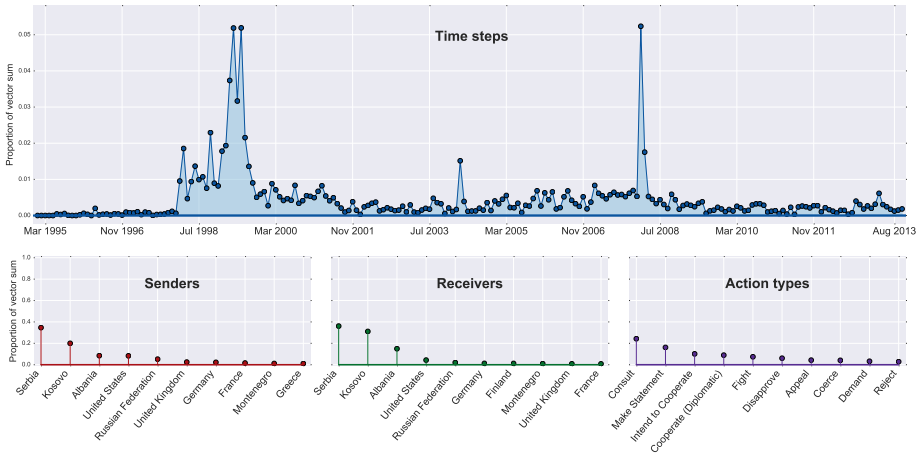
Figure 7.6: The PRGDS with $\epsilon_0 = 0$ was the only model to infer a component whose top sender and/or receiver was South Sudan, a country which did not exist until July 2011. 94% of the time steps (months) prior to July 2011 exhibit a latent state value of exactly zero $\theta_k^{(t)} = 0$. I speculate that the sparsity-inducing inductive bias of the $\epsilon = 0$ PRGDS variant allows it to measure more qualitatively specific components than the other models.



(a) Inferred by the PRGDS with $\epsilon_0 = 0$.

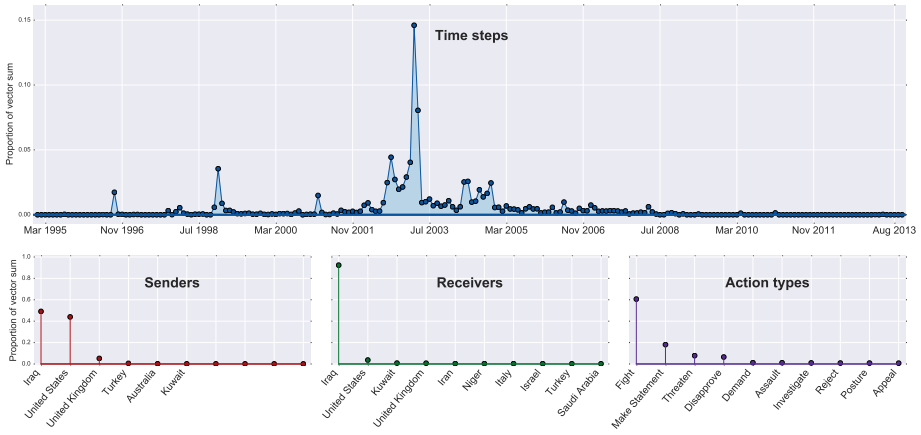


(b) Inferred by the PRGDS with $\epsilon_0 = 1$.

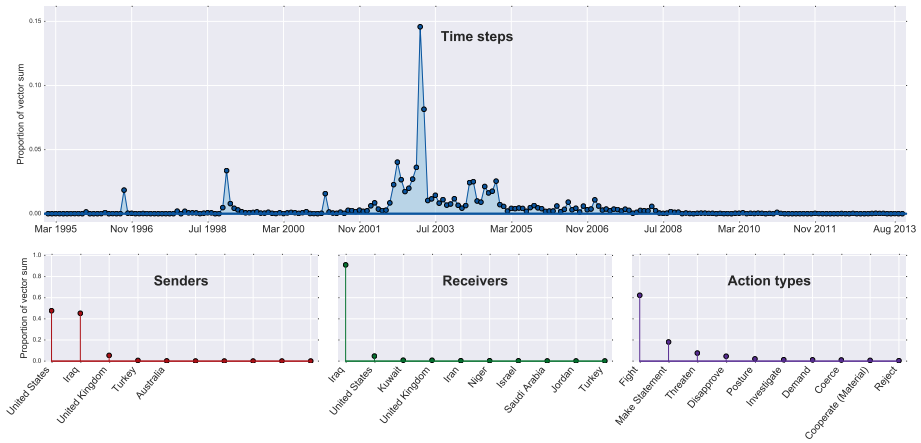


(c) Inferred by the PGDS.

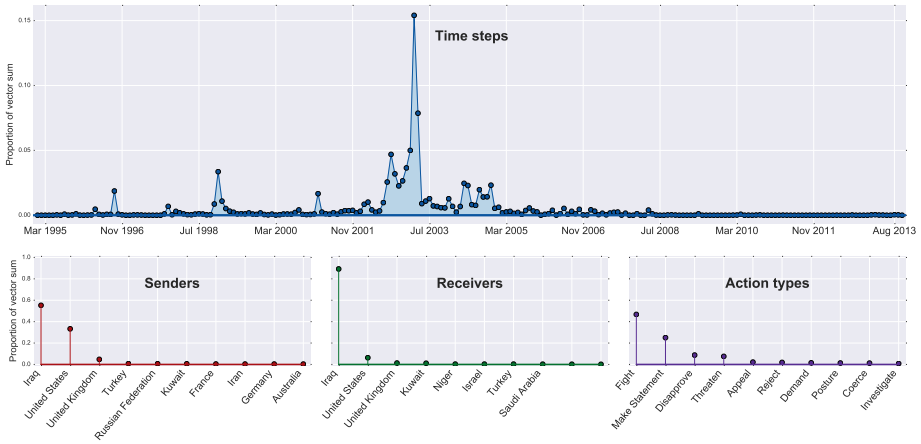
Figure 7.7: Kosovo War. A nearly identical component was inferred by all three models corresponding to the 1998 Kosovo War. The burst in February 2008 corresponds to Kosovo's declaration of independence from Serbia.



(a) Inferred by the PRGDS with $\epsilon_0 = 0$.

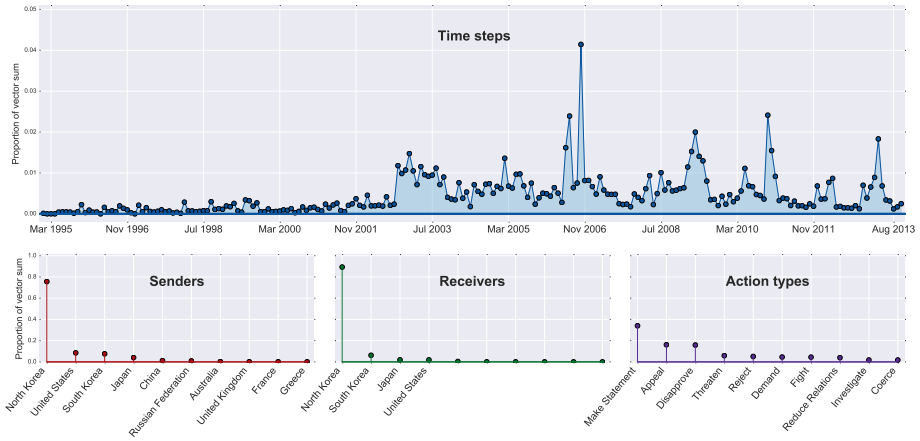


(b) Inferred by the PRGDS with $\epsilon_0 = 1$.

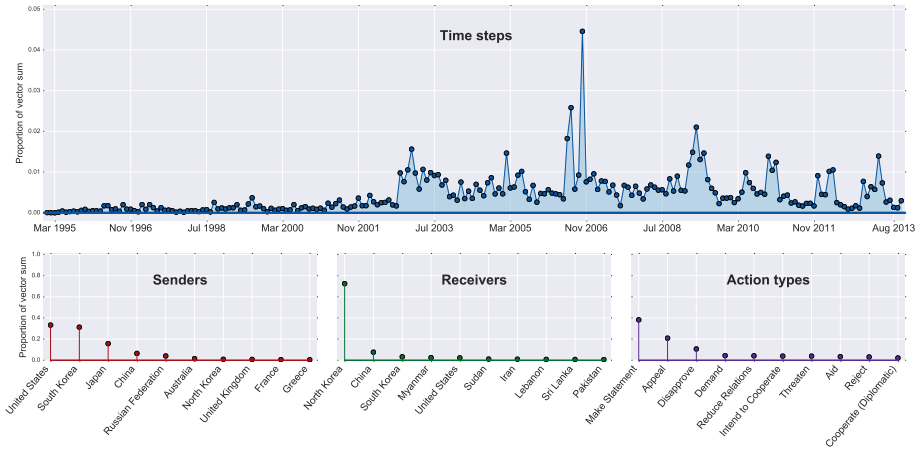


(c) Inferred by the PGDS.

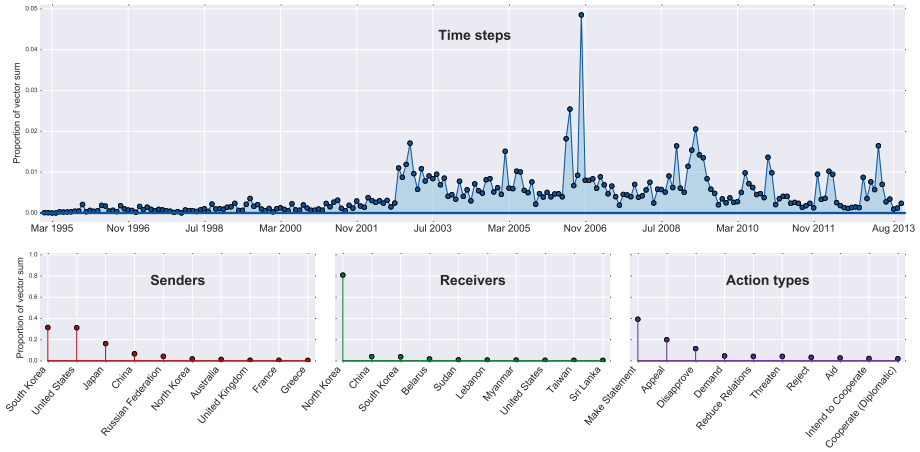
Figure 7.8: Second American invasion of Iraq and precursory strikes.



(a) Inferred by the PRGDS with $\epsilon_0 = 0$.

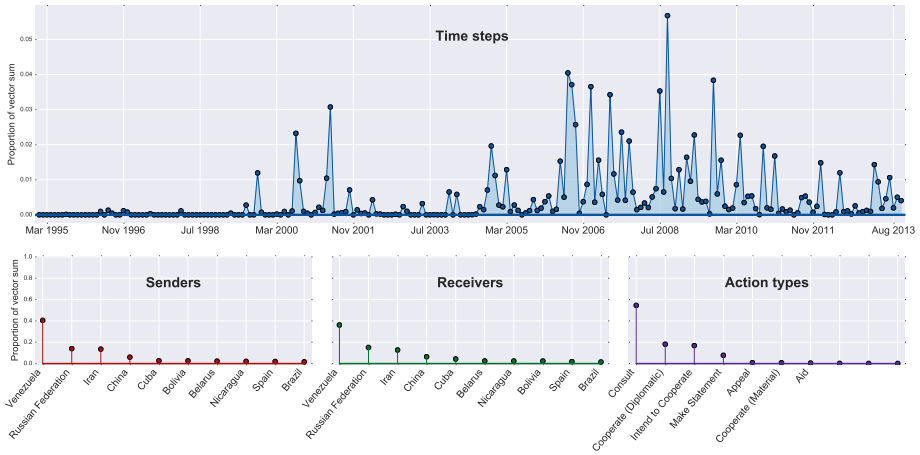


(b) Inferred by the PRGDS with $\epsilon_0 = 1$.

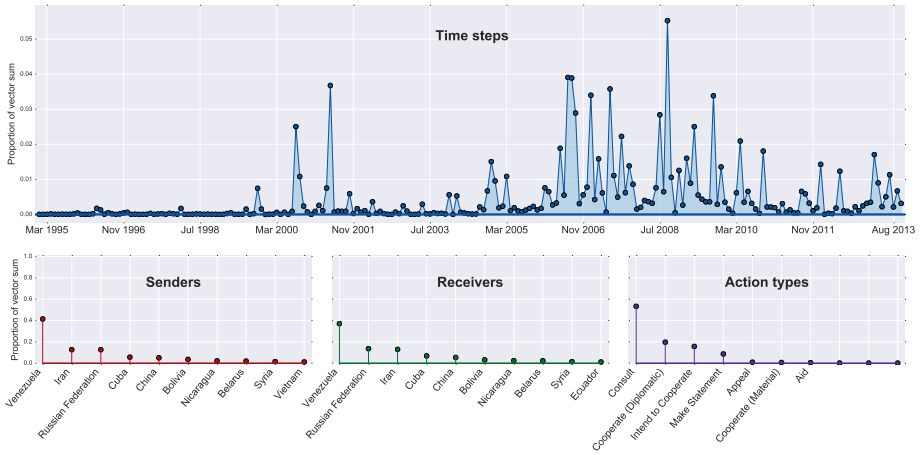


(c) Inferred by the PGDS.

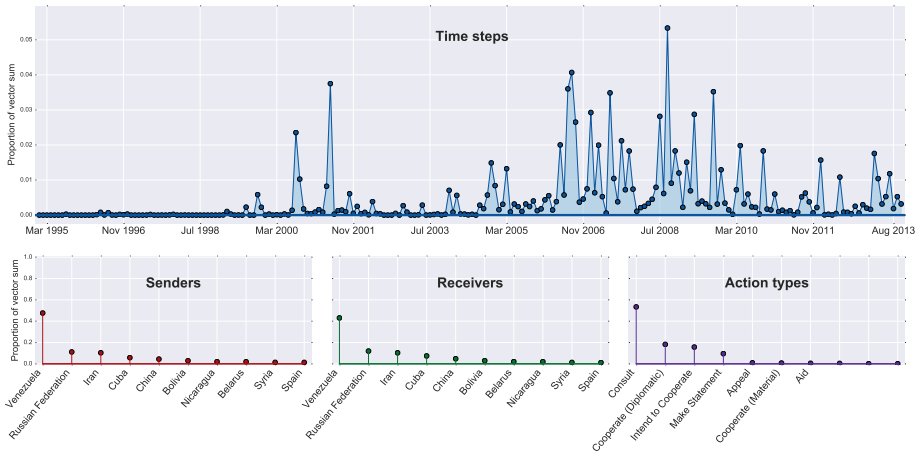
Figure 7.9: Six-party talks.



(a) Inferred by the PRGDS with $\epsilon_0 = 0$.



(b) Inferred by the PRGDS with $\epsilon_0 = 1$.



(c) Inferred by the PGDS.

Figure 7.10: Relations between Russia, Venezuela, and Iran.

7.5.2 GDELT 2003–2008 data

The different models' inferred latent structure was less similar to each other on the GDELT data than on the ICEWS data. This also corroborates the results in the predictive analysis section that showed greater variability in their smoothing performance on GDELT. Most of the aligned components are qualitatively similar—i.e., about the same set of events and relations—but the specific values of the parameters are more divergent. A representative example of a qualitatively well-aligned component is given in Fig. 7.12. This component is about Zimbabwe, particularly the Zimbabwean re-election of Robert Mugabe in 2008 which was criticized by the international community—notably, by the African Union, which appears as the the actor 'Africa'—for being rigged. The inferred structure of the PRGDS with $\epsilon_0 = 0$ is more similar to that of the PGDS than the PRGDS with $\epsilon_0 = 1$. The PRGDS with $\epsilon_0 = 1$ infers a less pronounced burst around the time of the election and features Zimbabwe as the second-most active participant behind the African Union.

I noticed a few instances in which the PRGDS with $\epsilon_0 = 1$ inferred a component that was qualitatively similar to its aligned counterpart inferred by the $\epsilon_0 = 0$ variant but featured smoother and less temporally localized latent states. An example of such a component is visualized in Fig. 7.13. Both PRGDS variants inferred a component that measures the 2006 East Timorese crisis [Wikipedia contributors, 2018a]. This conflict was primarily *within* East Timor, between elements of the military and government. However, an international military coalition lead by Australia, Malaysia, New Zealand, and Portugal ultimately intervened. The correspondence between the two PRGDS variants is not exact. The $\epsilon_0 = 0$ component is highly specific to the crisis. Its top sender and receiver is East Timor and the latent states are zero (or near zero) before the crisis and exhibit pronounce bursts during the months in which the crisis was active. The latent states inferred by the PRGDS with $\epsilon_0 = 1$ also exhibit bursts at the those times—however, the bursts are less pronounced and preceded by a steady

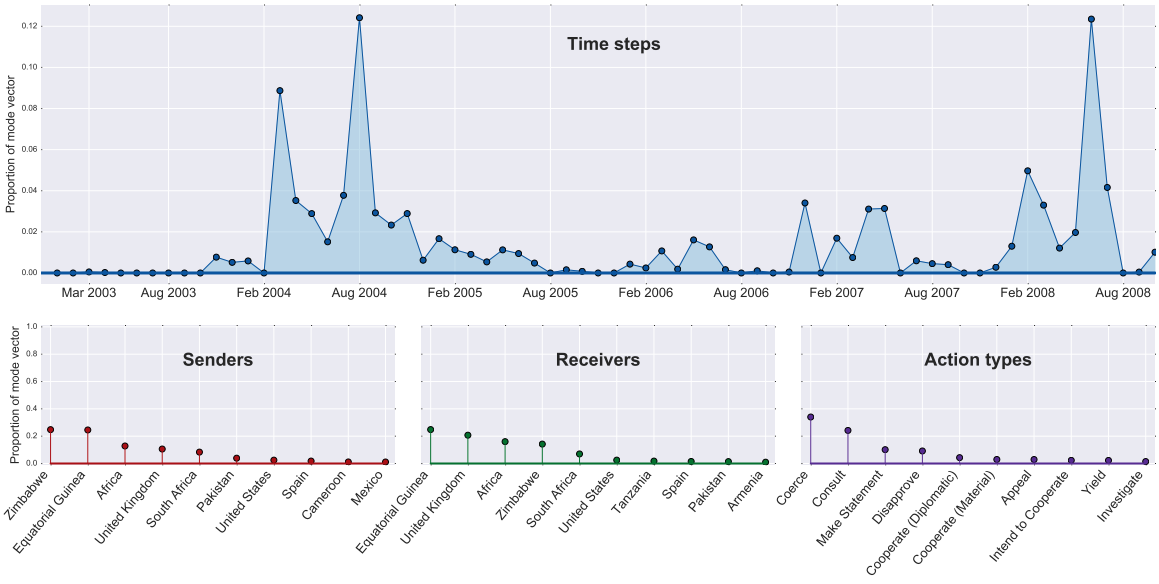


Figure 7.11: PRGDS with $\epsilon_0 = 0$ inferred a component specific to an attempted coup in Equatorial Guinea. The latent states burst at exactly the times given by a Reuters timeline (quoted in the main text) and are otherwise zero or near zero. No other model inferred a qualitatively similar component.

period of non-zero activity. As a consequence of being less temporally-localized to the crisis, the PRGDS with $\epsilon_0 = 1$ also infers sender, receiver, and action type parameters that are less specific to it, thus seeming to conflate the crisis with the surrounding relations between the countries involved in the crisis. The PGDS did not infer any qualitatively similar component.

As with the ICEWS results, there were a few instances in which the PRGDS with $\epsilon_0 = 0$ inferred a component that could not be qualitatively aligned to any inferred by the other models. These also tended to feature bursty and temporally localized latent states. An example of such a component is visualized in Fig. 7.11. This component precisely measures a 2004 coup attempt in Equatorial Guinea and the subsequent legal aftermath. The bursts in the inferred latent states exactly match the major events outlined in a Reuters timeline[2] which I quote with minimal paraphrasing below.

[2] https://www.reuters.com/article/us-equatorial-guinea-mann/timeline-coup-plot-case-in-equatorial-guinea-idUSL0747539320080707

March 7, 2004—Zimbabwe seizes U.S.-registered cargo plane carrying 64 suspected mercenaries and a cargo of military equipment.

March 8, 2004—About 15 suspected mercenaries are arrested in Equatorial Guinea in an investigation linked to the plane in Zimbabwe. Most of the suspects in both groups are South African.

March 16, 2004—Zimbabwe charges 70 suspected mercenaries with conspiring to murder Equatorial Guinea President Mbasogo.

August 23, 2004—Fourteen foreign suspected mercenaries and five Equatorial Guineans go on trial in Malabo, Equatorial Guinea.

August 25, 2004—South African police arrest Mark Thatcher, son of former British Prime Minister Margaret Thatcher, on suspicion of involvement in the plot. He is released from house arrest after posting 2 million rand (about $300,000) bail on September 3.

November 26, 2004—Equatorial Guinea court convicts 11 foreigners and two local men of charges stemming from the plot. Nick Du Toit, a South African alleged to have led an advance group of mercenaries, receives the stiffest sentence of 34 years' imprisonment.

January 13, 2005—Thatcher pleads guilty to a role in the plot under a plea bargain agreement allowing him to avoid jail by paying a 3 million rand fine and assisting South African authorities. He receives a four-year suspended jail sentence.

May 15, 2005—Zimbabwe frees 62 South Africans more than a year after they were arrested but the next day South Africa says it will charge them under its strict anti-mercenary laws.

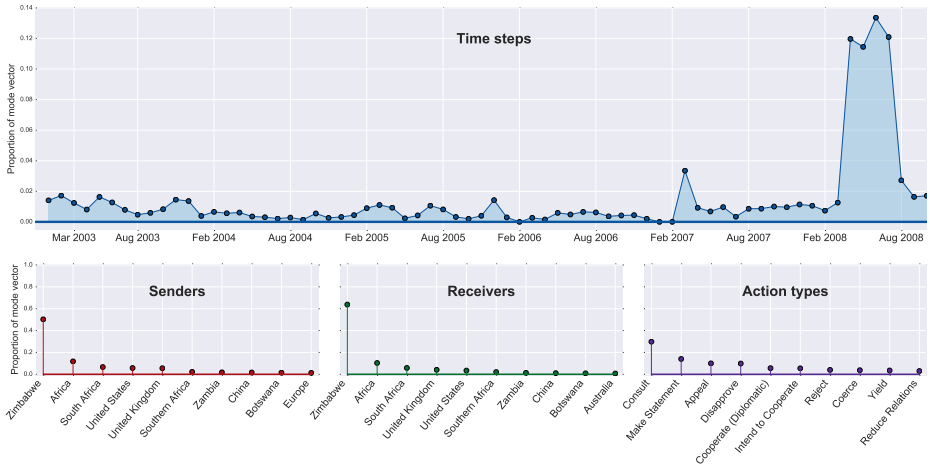
May 9, 2007—Zimbabwe agrees to extradite Mann to Equatorial Guinea.

January 30, 2008—Mann is deported to Equatorial Guinea from Zimbabwe to face coup plot charges after losing an appeal against extradition. He had served four years for buying weapons without a license.

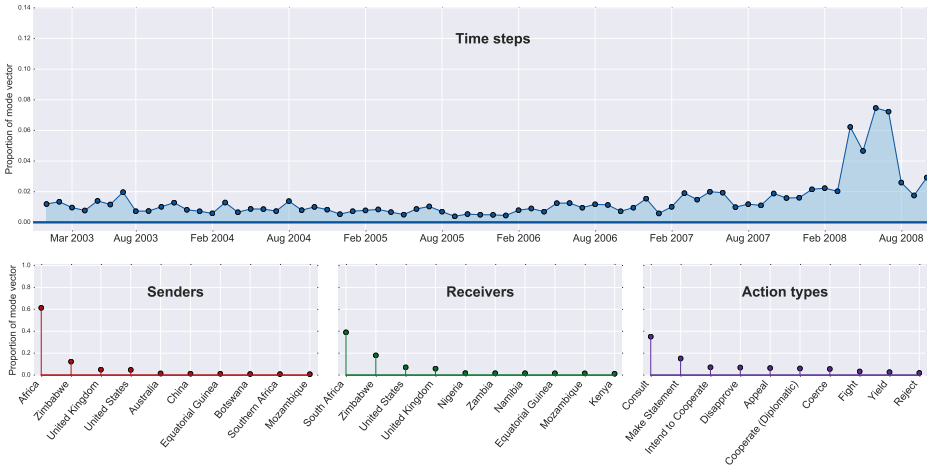
March 11, 2008—Mann says he plotted to oust Equatorial Guinea's president, but the scheme failed.

March 30, 2008—Guinea's public prosecutor says that Mann has testified that Thatcher knew all about the scheme to overthrow Obiang.

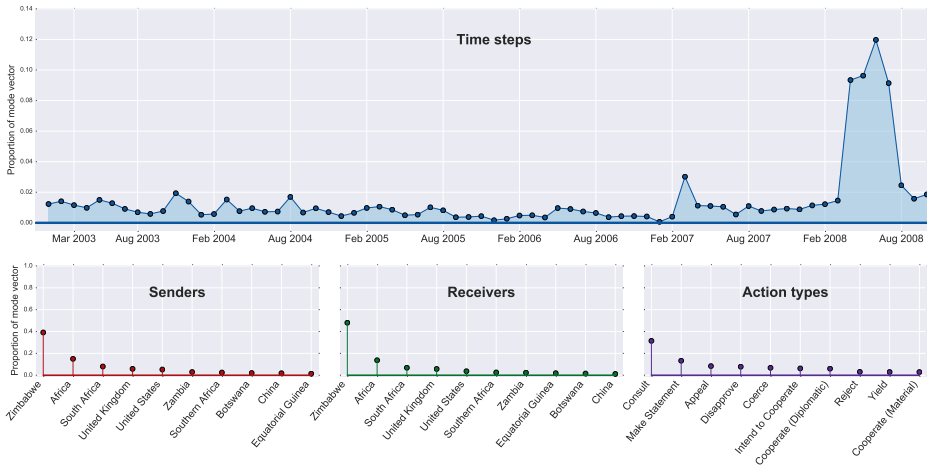
July 7, 2008—Mann sentenced to 34 years and four months in prison.



(a) Inferred by the PRGDS with $\epsilon_0 = 0$.

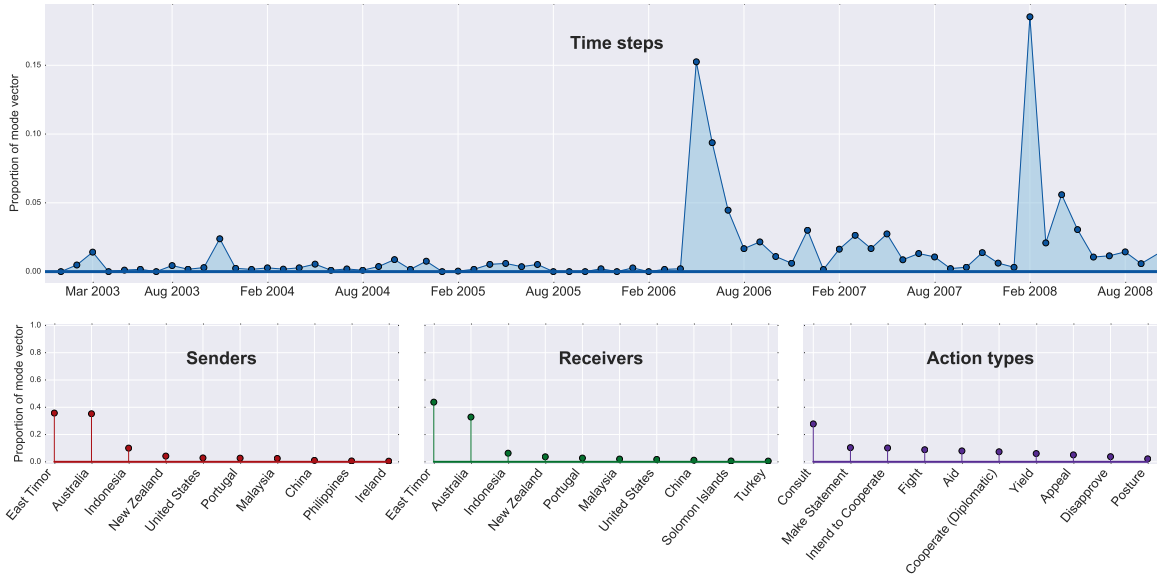


(b) Inferred by the PRGDS with $\epsilon_0 = 1$.

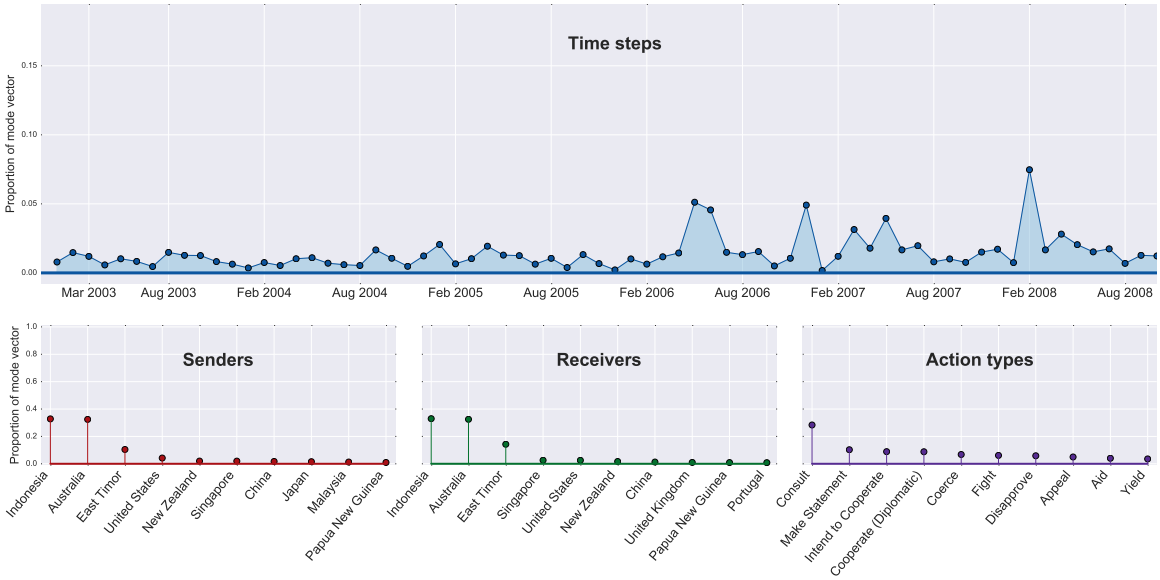


(c) Inferred by the PGDS.

Figure 7.12: 2008 Zimbabwean election.



(a) Inferred by the PRGDS with $\epsilon_0 = 0$.



(b) Inferred by the PRGDS with $\epsilon_0 = 1$.

Figure 7.13: 2006 East Timorese crisis. Both variants of the PRGDS inferred a component involving East Timor and the countries involved in the 2006 crisis. The PGDS did not infer any qualitatively similar component. The version inferred by the $\epsilon_0 = 0$ variant is more specific to the 2006 crisis.

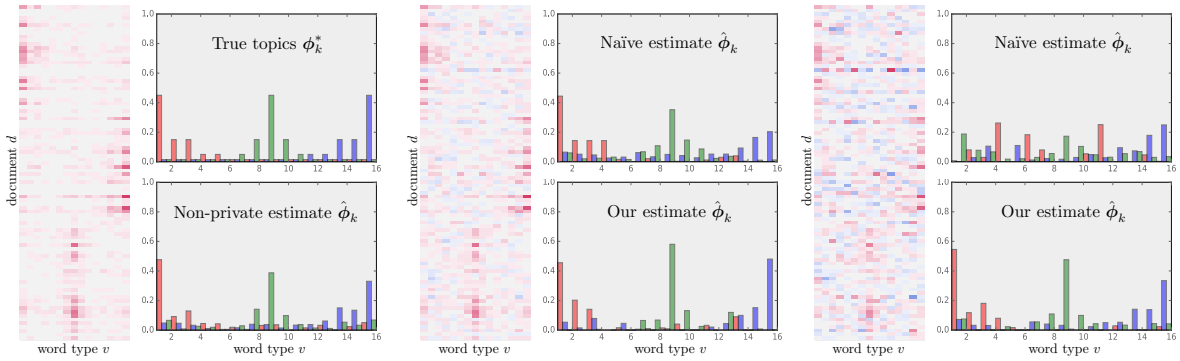
CHAPTER 8

LOCALLY PRIVATE BAYESIAN INFERENCE IN POISSON FACTORIZATION MODELS

Data from social processes often take the form of discrete observations (e.g., edges in a social network, word tokens in an email) and these observations often contain sensitive information about the people involved. As more aspects of social interaction are digitally recorded, the opportunities for social scientific insights grow; however, so too does the risk of unacceptable privacy violations. As a result, there is a growing need to develop privacy-preserving data analysis methods.

In practice, social scientists will be more likely to adopt these methods if doing so entails minimal change to their current methodology. Toward that end, under the framework of differential privacy [Dwork et al., 2006], this chapter (based on the preprint by Schein et al. [2018]) presents a method for privatizing Bayesian inference for Poisson factorization. The proposed method is general and modular, allowing social scientists to build on (instead of replace) their existing derivations and implementations of non-private Poisson factorization. To derive the method, we rely on a novel reinterpretation of the geometric mechanism [Ghosh et al., 2012], as well as a previously unknown general relationship between the Skellam [Skellam, 1946], Bessel [Yuan and Kalbfleisch, 2000], and Poisson distributions; these new results may be of independent interest in other contexts.

The proposed method satisfies a strong variant of differential privacy—i.e., local privacy—under which the sensitive data is privatized (or noised) via a randomized response method before inference. This ensures that no single centralized server need



(a) No noise. (b) Low noise. (c) High noise.

Figure 8.1: Topic recovery: proposed vs. the naïve approach. (a) We generated the non-privatized data synthetically so that the true topics were known. We then privatized the data using (b) a low noise level and (c) a high noise level. The heatmap in each subfigure visualizes the data, using red to denote positive counts and blue to denote negative counts. With a high noise level, the naïve approach overfits the noise and therefore fails to recover the true topics. We describe this experiment in more detail in Section 8.6.4.

ever store the non-privatized data—a condition that is non-negotiable in many real-world settings. The key challenge introduced by local privacy is how to infer the latent variables (including model parameters) given the privatized data. One option is a naïve approach, wherein inference proceeds as usual, treating the privatized data as if it were not privatized. In the context of maximum likelihood estimation, the naïve approach has been shown to exhibit pathologies when observations are discrete or count-valued; researchers have therefore advocated for treating the non-privatized observations as latent variables to be inferred [Yang et al., 2012, Karwa et al., 2014, Bernstein et al., 2017]. We embrace this approach and extend it to Bayesian inference, where the aim is to form the posterior distribution over the latent variables conditioned on the privatized data and the randomized response method; the proposed method is asymptotically guaranteed to draw samples from this posterior.

Section 8.6 reports two case studies applying the proposed method to 1) overlapping community detection in social networks and 2) topic modeling for text corpora. In order to formulate our local-privacy guarantees, we introduce and focus on

limited-precision local privacy—the local privacy analog of limited-precision differential privacy, originally proposed by Flood et al. [2013]. For each case study, a suite of experiments test the method's ability to form the posterior distribution over latent variables under different levels of noise. These experiments also demonstrate the utility of the method over the naïve approach for both case studies; an illustrative example is given in Fig. 8.1.

8.1 Differential privacy definitions

Differential privacy [Dwork et al., 2006] is a rigorous privacy criterion that guarantees that no single observation in a data set will have a significant influence on the information obtained by analyzing that data set.

Definition 8.1: Differential privacy

A randomized algorithm $\mathcal{A}(\cdot)$ satisfies ϵ-differential privacy if for all pairs of neighboring data sets Y and Y' that differ in only a single observation

$$P\left(\mathcal{A}(Y) \in \mathcal{S}\right) \leq e^{\epsilon} P\left(\mathcal{A}(Y') \in \mathcal{S}\right), \tag{8.1}$$

for all subsets $\mathcal{S}$ in the range of $\mathcal{A}(\cdot)$.

This work focuses on *local* differential privacy, hereby "local privacy". Under this criterion, the observations remain private from even the data analysis algorithm. The algorithm only sees *privatized* observations, often constructed by adding noise from specific distributions. The process of adding noise is known as randomized response—a reference to survey-sampling methods originally developed in the social sciences prior to the development of differential privacy [Warner, 1965]. Satisfying this criterion means one never needs to aggregate the true data in a single location.

Definition 8.2: Local differential privacy

A randomized response method $\mathcal{R}(\cdot)$ is ϵ-private if for any pair $y, y' \in \mathcal{Y}$

$$P\left(\mathcal{R}(y) \in \mathcal{S}\right) \leq e^{\epsilon} P\left(\mathcal{R}(y') \in \mathcal{S}\right), \tag{8.2}$$

for all subsets $\mathcal{S}$ in the range of $\mathcal{R}(\cdot)$. If a data analysis algorithm sees only the observations' ϵ-private responses, then the data analysis satisfies ϵ-local privacy.

The meaning of "observation" in Definitions 8.1 and 8.2 varies across applications. In the context of topic modeling, an observation is an individual document $y \in \mathbb{N}_0^V$ such that each single entry $y_v \in \mathbb{N}_0$ is the count of word tokens of type v in the document. To guarantee local privacy, the randomized response method has to satisfy the condition in Eq. (8.2) for any pair of observations. This typically involves adding noise that scales with the maximum difference between any pair of observations defined as $N^{(\max)} = \max_{y,y'} \|y - y'\|_1$. When the observations are documents, $N^{(\max)}$ can be prohibitively large and the amount of noise will overwhelm the signal in the data. This motivates the following alternative formulation of privacy.

While standard local privacy requires that arbitrarily different observations become indistinguishable under the randomized response method, this guarantee may be unnecessarily strong in some settings. For instance, suppose a user would like to hide only the fact that their email contains a handful of vulgar curse words. Then it is sufficient to have a randomized response method which guarantees that any two similar emails—one containing the vulgar curse words and the same email without them—will be rendered indistinguishable after randomization. In other words, this only requires the randomized response method to render small neighborhoods of possible observations indistinguishable. To operationalize this kind of guarantee, we generalize Definition 8.2 and define limited-precision local privacy (LPLP).

Definition 8.3: Limited-precision local privacy (LPLP)

For any positive integer N, we say that a randomized response method $\mathcal{R}(\cdot)$ is (N, ϵ)-private if for all pairs of observations $y, y' \in \mathcal{Y}$ such that $\|y - y'\|_1 \leq N$

$$P\left(\mathcal{R}(y) \in \mathcal{S}\right) \leq e^{\epsilon} P\left(\mathcal{R}(y') \in \mathcal{S}\right), \tag{8.3}$$

for all subsets $\mathcal{S}$ in the range of $\mathcal{R}(\cdot)$. If a data analysis algorithm sees only the observations' (N, ϵ)-private responses, then the data analysis itself satisfies (N, ϵ)-limited-precision local privacy. If $\|y\|_1 \leq N$ for all $y \in \mathcal{Y}$, then (N, ϵ)-limited-precision local privacy implies ϵ-local privacy.

LPLP is the local privacy analog to limited-precision differential privacy, originally proposed by Flood et al. [2013] and subsequently used to privatize analyses of geographic location data [Andrés et al., 2013] and financial network data [Papadimitriou et al., 2017]. Note that this is a *strict generalization* of local privacy. A randomized response method that satisfies LPLP adds noise which scales as a function of N and ϵ—thus the same method may be interpreted as being ϵ-private for a given setting of N or ϵ'-private for a different setting N'. Section 8.3 describes the geometric mechanism [Ghosh et al., 2012] and shows that it satisfies LPLP.

8.2 Private Bayesian inference

In Bayesian statistics, we begin with a probabilistic model $\mathcal{M}$ that relates observable variables Y to latent variables Z via a joint distribution $P_{\mathcal{M}}(Y, Z)$. The goal of inference is then to compute the posterior distribution $P_{\mathcal{M}}(Z \mid Y)$ over the latent variables conditioned on observed values of Y. The posterior is almost always analytically intractable and thus inference involves approximating it. The two most common methods of approximate Bayesian inference are variational inference, wherein we fit

the parameters of an approximating distribution $Q(Z \mid Y)$, and Markov chain Monte Carlo (MCMC), wherein we approximate the posterior with a set of samples $\{Z^{(s)}\}_{s=1}^{S}$ generated via a Markov chain whose stationary distribution is the exact posterior.

We can conceptualize Bayesian inference as a randomized algorithm $\mathcal{A}(\cdot)$ that returns an approximation to the posterior distribution $P_{\mathcal{M}}(Z \mid Y)$. In general $\mathcal{A}(\cdot)$ does not satisfy ϵ-differential privacy. However, if $\mathcal{A}(\cdot)$ is an MCMC algorithm that returns a single sample from the posterior, it guarantees privacy [Dimitrakakis et al., 2014, Wang et al., 2015, Foulds et al., 2016, Dimitrakakis et al., 2017]. Adding noise to posterior samples can also guarantee privacy [Zhang et al., 2016], though this set of noised samples $\{\tilde{Z}^{(s)}\}_{s=1}^{S}$ approximates some distribution $\tilde{P}_{\mathcal{M}}(Z \mid Y)$ that depends on ϵ and is different than the exact posterior (but close, in some sense, and equal when $\epsilon \to 0$). For specific models, we can also noise the transition kernel of the MCMC algorithm to construct a Markov chain whose stationary distribution is again not the exact posterior, but something close that guarantees privacy [Foulds et al., 2016]. We can also take an analogous approach to privatize variational inference, wherein we add noise to the sufficient statistics computed in each iteration [Park et al., 2016].

All of the aforementioned work on private Bayesian inference focuses on non-local privacy wherein the true data is conditioned on. Under *local privacy*, we must take a different approach. Let's formalize the general objective of Bayesian inference under local privacy. Given a generative model $\mathcal{M}$ for non-privatized data Y and latent variables Z with joint distribution $P_{\mathcal{M}}(Y, Z)$, we further assume a randomized response method $\mathcal{R}(\cdot)$ that generates privatized data sets: $\tilde{Y} \sim P_{\mathcal{R}}(\tilde{Y} \mid Y)$. The goal is then to approximate the *locally private posterior*:

$$\begin{aligned} P_{\mathcal{M},\mathcal{R}}(Z \mid \tilde{Y}) &= \mathbb{E}_{P_{\mathcal{M},\mathcal{R}}(Y \mid \tilde{Y})}\left[P_{\mathcal{M}}(Z \mid Y)\right] \\ &= \int P_{\mathcal{M}}(Z \mid Y)\, P_{\mathcal{M},\mathcal{R}}(Y \mid \tilde{Y})\, dY. \end{aligned} \tag{8.4}$$

This distribution correctly characterizes our uncertainty about the latent variables Z, conditioned on all of our observations and assumptions—i.e., the privatized data $\tilde{Y}$, the model $\mathcal{M}$, and the randomized response method $\mathcal{R}$. The expansion in Eq. (8.4) shows that this posterior implicitly treats the non-privatized data set Y as a latent variable and marginalizes over it using the mixing distribution $P_{\mathcal{M},\mathcal{R}}(Y \mid \tilde{Y})$ which is itself a posterior that characterizes our uncertainty about Y given all we observe. The central point here is that if we can generate samples from $P_{\mathcal{M},\mathcal{R}}(Y \mid \tilde{Y})$, we can use them to approximate the expectation in Eq. (8.4), assuming that we already have a method for approximating the non-private posterior $P_{\mathcal{M}}(Z \mid Y)$. In the context of MCMC, iteratively re-sampling values of the non-privatized data from their complete conditional—i.e., $Y^{(s)} \sim P_{\mathcal{M},\mathcal{R}}(Y \mid Z^{(s-1)}, \tilde{Y})$—and then re-sampling values of the latent variables—i.e., $Z^{(s)} \sim P_{\mathcal{M}}(Z \mid Y^{(s)})$—constitutes a Markov chain whose stationary distribution is $P_{\mathcal{M},\mathcal{R}}(Z, Y \mid \tilde{Y})$. In scenarios where we already have derivations and implementations for sampling from $P_{\mathcal{M}}(Z \mid Y)$, we need only be able to sample efficiently from $P_{\mathcal{M},\mathcal{R}}(Y \mid Z, \tilde{Y})$ in order to obtain a locally private Bayesian inference algorithm; whether we can do this efficiently depends on our choice of $\mathcal{M}$ and $\mathcal{R}$.

Note that the objective of Bayesian inference under local privacy, as defined in Eq. (8.4), is similar to that of Williams and McSherry [2010], who identify their key barrier to inference as being unable to analytically form the marginal likelihood that links the privatized data to Z:

$$P_{\mathcal{M},\mathcal{R}}(\tilde{Y} \mid Z) = \int P_{\mathcal{R}}(\tilde{Y} \mid Y)\, P_{\mathcal{M}}(Y \mid Z)\, dY. \tag{8.5}$$

In the next sections, we see that if $\mathcal{M}$ is a Poisson factorization model and $\mathcal{R}$ is the geometric mechanism, then we can form an augmented version of this marginal likelihood analytically and derive an MCMC algorithm that samples efficiently from the posterior in Eq. (8.4).

8.3 Locally private Poisson factorization

In this section, we will consider an existing randomized response method $\mathcal{R}$ that is natural for discrete data—i.e., the geometric mechanism [Ghosh et al., 2012]—and show how it combines naturally with $\mathcal{M}$—i.e., Poisson factorization—in an analytically tractable way. To do so we prove two theorems about $\mathcal{R}$: 1) that it is a mechanism for LPLP and 2) that it can be understood in terms of the Skellam distribution [Skellam, 1946]. We rely on the second theorem to show that our choices of $\mathcal{M}$ and $\mathcal{R}$ combine to yield a novel generative process for privatized count data which we exploit to derive efficient Bayesian inference.

8.3.1 Reinterpreting the geometric mechanism $\mathcal{R}$

The two most commonly used randomized response mechanisms in the privacy toolbox—the Gaussian and Laplace mechanisms—privatize observations by adding noise drawn from continuous real-valued distributions; they are thus unnatural choices for count data. Ghosh et al. [2012] introduced the geometric mechanism, which can be viewed as the discrete analog to the Laplace mechanism, and involves adding integer-valued noise $\tau \in \mathbb{Z}$ drawn from the two-sided geometric distribution $\tau \sim 2\text{Geo}(\alpha)$.

Definition 8.4: Two-sided geometric as the difference of geometrics

Define the difference $\tau \triangleq g_1 - g_2$ of two i.i.d geometric random variables $g_1 \sim \text{NB}(1, \alpha)$ and $g_2 \sim \text{NB}(1, \alpha)$ where $\alpha \in (0, 1)$ is the probability parameter and the geometric distribution is defined as a special case of the negative binomial distribution [Johnson et al., 2005]. Then τ is marginally a two-sided geometric random variable:

$$P(\tau \mid \alpha) = 2\text{Geo}(\tau; \alpha) \tag{8.6}$$

Definition 8.5: Two-sided geometric distribution

A two-sided geometric random variable $\tau \sim 2\text{Geo}(\alpha)$ is an integer $\tau \in \mathbb{Z}$. Its distribution is defined by probability parameter $\alpha \in (0, 1)$ and PMF:

$$2\text{Geo}(\tau; \alpha) = \frac{1-\alpha}{1+\alpha}\,\alpha^{|\tau|}. \tag{8.7}$$

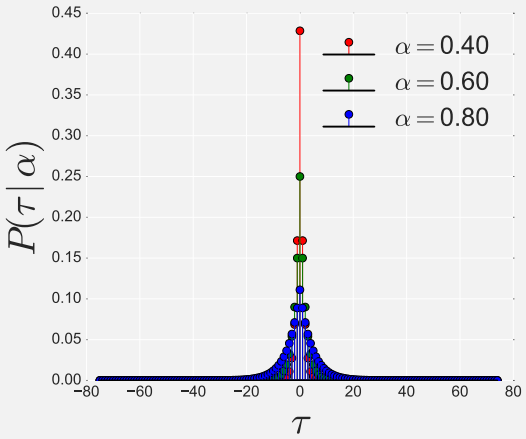


Figure 8.2: PMF of the two-sided geometric distribution for three values of α.

Theorem 8.1: Geometric mechanism as an LPLP mechanism

(*Proof in Section 8.8.1*) Let randomized response method $\mathcal{R}(\cdot)$ be the geometric mechanism with parameter $\alpha \in (0, 1)$. Then for any positive integer N, and any pair of observations $y, y' \in \mathcal{Y}$ such that $\|y - y'\|_1 \leq N$, $\mathcal{R}(\cdot)$ satisfies the limited-precision local privacy criterion in Eq. (8.3) with privacy level ϵ:

$$\epsilon = N \ln\left(\tfrac{1}{\alpha}\right). \tag{8.8}$$

Therefore, for any N, the geometric mechanism with parameter α is an (N, ϵ)-private randomized response method; if a data analysis algorithm sees only the (N, ϵ)-privatized observations, then the data analysis satisfies (N, ϵ)-LPLP.

Theorem 8.2: Geometric mechanism as a Skellam mechanism

(*Proof in Section 8.8.2*) A random variable $\tau \sim 2\text{Geo}(\alpha)$ can be generated as:

$$\lambda^{(+)}, \lambda^{(-)} \overset{\text{iid}}{\sim} \text{Exp}\left(\tfrac{\alpha}{1-\alpha}\right) \tag{8.9}$$

$$\tau \sim \text{Skel}\left(\lambda^{(+)}, \lambda^{(-)}\right) \tag{8.10}$$

where the exponential distribution is a special case of the gamma distribution—i.e., $\text{Exp}(\beta) \equiv \Gamma(1, \beta)$—and the Skellam distribution is defined below.

Definition 8.6: Skellam distribution

A Skellam random variable $\tau \sim \text{Skel}(\mu_1, \mu_2)$ is an integer $\tau \in \mathbb{Z}$. Its distribution is defined by two positive rate parameters parameters $\mu_1, \mu_2 > 0$ and PMF:

$$\text{Skel}(\tau; \mu_1, \mu_2) = e^{-(\mu_1+\mu_2)} \left(\frac{\mu_1}{\mu_2}\right)^{\frac{\tau}{2}} I_\tau\left(2\sqrt{\mu_1\mu_2}\right), \tag{8.11}$$

where $I_v(a)$ is the modified Bessel function of the first kind.

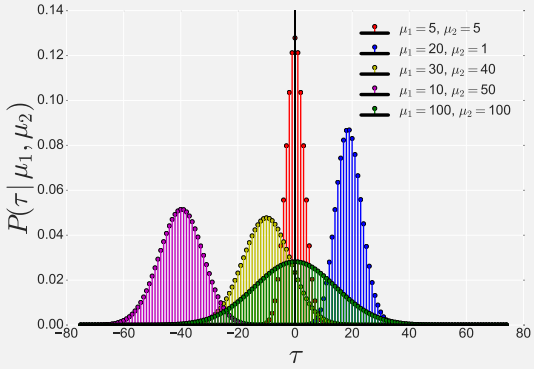


Figure 8.3: PMF of the Skellam distribution for different combinations of μ_1 and μ_2.

Definition 8.7: Skellam as the difference of independent Poissons

Define the difference $\tau \triangleq y_1 - y_2$ of two independent Poisson random variables $y_1 \sim \text{Pois}(\mu_1)$ and $y_2 \sim \text{Pois}(\mu_2)$. Then τ is marginally Skellam distributed:

$$P(\tau \mid \mu_1, \mu_2) = \text{Skel}(\tau;\ \mu_1, \mu_2) \tag{8.12}$$

8.4 Combining $\mathcal{M}$ and $\mathcal{R}$

Assume each observation is generated by $\mathcal{M}$ and then then privatized by $\mathcal{R}$—i.e.:

$$y_{\boldsymbol{\delta}} \sim \text{Pois}(\mu_{\boldsymbol{\delta}}), \tag{8.13}$$

$$\tau_{\boldsymbol{\delta}} \sim 2\text{Geo}(\alpha), \tag{8.14}$$

$$\tilde{y}_{\boldsymbol{\delta}}^{(\pm)} := y_{\boldsymbol{\delta}} + \tau_{\boldsymbol{\delta}}, \tag{8.15}$$

where $\tilde{y}_{\boldsymbol{\delta}}^{(\pm)}$ is the *privatized* observation which we superscript with $(\pm)$ to denote that it may be non-negative or negative since the additive noise $\tau_{\boldsymbol{\delta}} \in \mathbb{Z}$ may be negative.

Via Theorem 8.2, we can express the generative process for $\tilde{y}_{\boldsymbol{\delta}}^{(\pm)}$ in four equivalent ways, shown in Fig. 8.4, each of which provides a unique and necessary insight. Process 1 is a graphical representation of the generative process as defined thus far. Process 2 represents the two-sided geometric noise in terms of a pair of Poisson random variables with exponentially distributed rates; in so doing, it reveals the auxiliary variables that facilitate inference. Process 3 represents the sum of the true count and the positive component of the noise as a single Poisson random variable $\tilde{y}_{\boldsymbol{\delta}}^{(+)} = y_{\boldsymbol{\delta}} + g_{\boldsymbol{\delta}}^{(+)}$. Process 4 marginalizes out the remaining Poisson random variables to obtain a marginal representation $\tilde{y}_{\boldsymbol{\delta}}^{(\pm)}$ as a Skellam random variable with exponentially-randomized rates:

$$\lambda_{\boldsymbol{\delta}}^{(+)}, \lambda_{\boldsymbol{\delta}}^{(-)} \overset{\text{iid}}{\sim} \text{Exp}(\tfrac{\alpha}{1-\alpha}), \tag{8.16}$$

$$\tilde{y}_{\boldsymbol{\delta}}^{(\pm)} \sim \text{Skel}\left(\lambda_{\boldsymbol{\delta}}^{(+)} + \mu_{\boldsymbol{\delta}},\, \lambda_{\boldsymbol{\delta}}^{(-)}\right). \tag{8.17}$$

The derivation of these four processes follows from the two-sided geometric random variable being the difference of negative binomials (Definition 8.4), the representation of the negative binomial as a gamma–Poisson mixture (Definition 3.16), Poisson additivity (Definition 3.2), and the Skellam as the difference of Poissons (Definition 8.7).

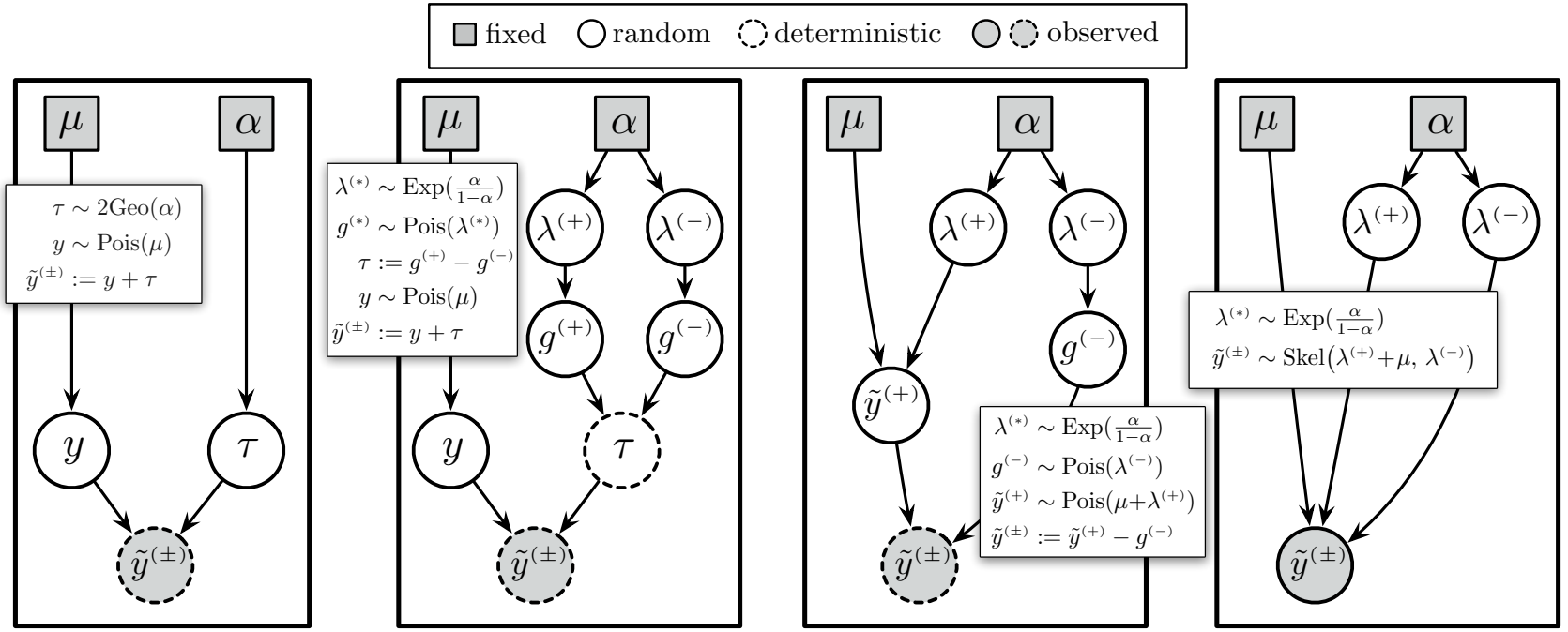


Figure 8.4: Four generative processes that yield the same marginal distributions $P(\tilde{y}^{(\pm)} \,|\, \mu, \alpha)$. Process 1 generates $y^{(\pm)}$ as the sum of an independent Poisson and two-sided geometric random variable. Process 2 augments the two-sided geometric random variable as the difference of two Poisson random variables with exponentially-randomized rates. Process 3 represents the sum of y and the additive geometric random variable $g^{(+)}$ as a single Poisson random variable $\tilde{y}^{(+)}$. Process 4 marginalizes out the Poisson random variables to yield a generative process for $\tilde{y}^{(\pm)}$ as a Skellam random variable with exponentially-randomized rates.

8.5 MCMC inference

Upon observing a privatized data set $\tilde{Y}^{(\pm)}$, the goal of a Bayesian agent is to approximate the locally private posterior. As explained in Section 8.2, to do so with MCMC, we need only be able to sample the true data $y_{\boldsymbol{\delta}}$ as a latent variable from its complete conditional $P_{\mathcal{M},\mathcal{R}}(y_{\boldsymbol{\delta}} \,|\, \tilde{y}_{\boldsymbol{\delta}}^{(\pm)}, \mu_{\boldsymbol{\delta}}, -)$. By assuming that the privatized observations $\tilde{y}_{\boldsymbol{\delta}}^{(\pm)}$ are conditionally independent Skellam random variables, as in Eq. (8.17), and we may exploit the following general theorem that relates the Skellam distribution to the Bessel distribution which also arose in the previous chapter (see Definition 7.2).

Theorem 8.3: The Skellam–Bessel relationship

(Proof in Section 8.8.3) Define the minimum $m \triangleq \min\{y_1, y_2\}$ and difference $\tau \triangleq y_1 - y_2$ of two Poisson random variables $y_1 \sim \text{Pois}(\lambda^{(+)})$ and $y_2 \sim \text{Pois}(\lambda^{(-)})$. Then m and τ are marginally (i.e., not conditioned on y_1, y_2) distributed as:

$$\tau \sim \text{Skel}(\lambda^{(+)}, \lambda^{(-)}), \tag{8.18}$$

$$m \sim \text{Bes}\left(|\tau|, 2\sqrt{\lambda^{(+)}\lambda^{(-)}}\right). \tag{8.19}$$

Theorem 8.3 means that we can generate two independent Poisson random variables by first generating their difference τ and then their minimum m. Since $\tau \triangleq y_1 - y_2$, if τ is greater than 0, then $y_2 = m$ must be the minimum and $y_1 = \tau - m$. In practice, this means that if we only get to observe the difference of two Poisson-distributed counts, we can still "recover" the counts by sampling a Bessel auxiliary variable.

Assuming that $\tilde{y}_{\boldsymbol{\delta}}^{(\pm)} \sim \text{Skel}(\lambda_{\boldsymbol{\delta}}^{(+)} + \mu_{\boldsymbol{\delta}}, \lambda_{\boldsymbol{\delta}}^{(-)})$ via Theorem 8.2, we first sample an auxiliary Bessel random variable $m_{\boldsymbol{\delta}}$:

$$\left(m_{\boldsymbol{\delta}} \,|\, -\right) \sim \text{Bes}\left(|\tilde{y}_{\boldsymbol{\delta}}^{(\pm)}|,\, 2\sqrt{(\lambda_{\boldsymbol{\delta}}^{(+)} + \mu_{\boldsymbol{\delta}})\lambda_{\boldsymbol{\delta}}^{(-)}}\right). \tag{8.20}$$

The Bessel distribution can be sampled from efficiently. I have open-sourced my Cython implementation of the rejection algorithms described by Devroye [2002].

By Theorem 8.3, $m_{\boldsymbol{\delta}}$ represents the minimum of two latent Poisson random variables whose difference equals the observed $\tilde{y}^{(\pm)}_{\boldsymbol{\delta}}$; these two latent counts are given explictly in process 3 of Fig. 8.4—i.e., $\tilde{y}^{(\pm)}_{\boldsymbol{\delta}} := \tilde{y}^{(+)}_{\boldsymbol{\delta}} - g^{(-)}_{\boldsymbol{\delta}}$ and thus $m_{\boldsymbol{\delta}} = \min\{\tilde{y}^{(+)}_{\boldsymbol{\delta}}, g^{(-)}_{\boldsymbol{\delta}}\}$. Given a sample of $m_{\boldsymbol{\delta}}$ and the observed value of $\tilde{y}^{(\pm)}_{\boldsymbol{\delta}}$, we can then compute $\tilde{y}^{(+)}_{\boldsymbol{\delta}}$, $g^{(-)}_{\boldsymbol{\delta}}$:

$$\begin{aligned} \tilde{y}^{(+)}_{\boldsymbol{\delta}} &:= m_{\boldsymbol{\delta}}, \;\; g^{(-)}_{\boldsymbol{\delta}} := \tilde{y}^{(+)}_{\boldsymbol{\delta}} - \tilde{y}^{(\pm)}_{\boldsymbol{\delta}} \;\text{ if } \tilde{y}^{(\pm)}_{\boldsymbol{\delta}} \leq 0 \\ g^{(-)}_{\boldsymbol{\delta}} &:= m_{\boldsymbol{\delta}}, \;\; \tilde{y}^{(+)}_{\boldsymbol{\delta}} := g^{(-)}_{\boldsymbol{\delta}} + \tilde{y}^{(\pm)}_{\boldsymbol{\delta}} \;\text{ otherwise.} \end{aligned} \tag{8.21}$$

Because $\tilde{y}^{(+)}_{\boldsymbol{\delta}} = y_{\boldsymbol{\delta}} + g^{(+)}_{\boldsymbol{\delta}}$ is itself the sum of two independent Poisson random variables, we can then sample $y_{\boldsymbol{\delta}}$ from its conditional posterior, which is a binomial distribution:

$$\left(y_{\boldsymbol{\delta}} \mid -\right) \sim \text{Binom}\left(\tilde{y}^{(+)}_{\boldsymbol{\delta}}, \frac{\mu_{\boldsymbol{\delta}}}{\mu_{\boldsymbol{\delta}} + \lambda^{(+)}_{\boldsymbol{\delta}}}\right). \tag{8.22}$$

Equations 8.20 through 8.22 sample the true underlying data $y_{\boldsymbol{\delta}}$ from $P_{\mathcal{M},\mathcal{R}}(y_{\boldsymbol{\delta}} \mid \tilde{y}^{(\pm)}_{\boldsymbol{\delta}}, \mu_{\boldsymbol{\delta}}, \boldsymbol{\lambda}_{\boldsymbol{\delta}})$. We may also re-sample the auxiliary variables $\lambda^{(+)}_{\boldsymbol{\delta}}, \lambda^{(-)}_{\boldsymbol{\delta}}$ from their complete conditional, which is a gamma distribution, by conjugacy:

$$\left(\lambda^{(*)}_{\boldsymbol{\delta}} \mid -\right) \sim \Gamma\left(1 + g^{(*)}_{\boldsymbol{\delta}}, \tfrac{\alpha}{1-\alpha} + 1\right). \tag{8.23}$$

Iteratively re-sampling $y_{\boldsymbol{\delta}}$ and $\boldsymbol{\lambda}_{\boldsymbol{\delta}}$ constitutes a chain whose stationary distribution over $y_{\boldsymbol{\delta}}$ is $P_{\mathcal{M},\mathcal{R}}(y_{\boldsymbol{\delta}} \mid \tilde{y}^{(\pm)}_{\boldsymbol{\delta}}, \mu_{\boldsymbol{\delta}})$, as desired. Conditioned on a sample of the underlying data set Y, we then re-sample the latent variables Z (that define the rates $\mu_{\boldsymbol{\delta}}$) from their complete conditionals, which match those in standard non-private Poisson factorization. Eqs. (8.20) to (8.23) along with non-private complete conditionals for Z thus define a privacy-preserving MCMC algorithm that is asymptotically guaranteed to sample from the locally private posterior $P_{\mathcal{M},\mathcal{R}}(Z \mid \tilde{Y}^{(\pm)})$ for any Poisson factorization model.

8.6 Case studies

This section reports two case studies applying the proposed method to 1) overlapping community detection in social networks and 2) topic modeling for text corpora. In each, we formulate natural local-privacy guarantees, ground them in examples, and demonstrate the proposed method on real and synthetic data.

8.6.1 Enron corpus data

For the real data experiments, we obtained count matrices derived from the Enron email corpus [Klimt and Yang, 2004]. For the community detection case study, we obtained a $V \times V$ adjacency matrix Y where y_{ij} is the number of emails sent from actor i to actor j. We included an actor if they sent at least one email and sent or received at least one hundred emails, yielding $V = 161$ actors. When an email included multiple recipients, we incremented the corresponding counts by one. For the topic modeling case study, we randomly selected $D = 10,000$ emails with at least 50 word tokens. We limit the vocabulary to $V = 10,000$ word types, selecting only the most frequent word types with document frequency less than 0.3. In both case studies, we privatize the data using the geometric mechanism under varying degrees of privacy and examine each method's ability to reconstruct the true underlying data.

8.6.2 Reference methods

We compare the performance of the proposed method to two references methods: 1) the non-private approach—i.e., standard Poisson factorization fit to the true underlying data, and 2) the naïve approach—i.e., standard Poisson factorization fit to the privatized data, as if it were the true data. The naïve approach must first truncate any negative counts $\tilde{y}^{(\pm)}_{\boldsymbol{\delta}} < 0$ to zero and thus implicitly uses the *truncated* geometric mechanism [Ghosh et al., 2012].

8.6.3 Performance measures

All methods generate a set of S samples of the latent variables using MCMC. We use these samples to approximate the posterior expectation of $y_{\boldsymbol{\delta}}$:

$$\hat{\mu}_{\boldsymbol{\delta}} = \frac{1}{S}\sum_{s=1}^{S} \mu_{\boldsymbol{\delta}}^{(s)} \approx \mathbb{E}_{P_{\mathcal{M},\mathcal{R}}(Z \mid \tilde{Y}^{(\pm)})}\left[\mu_{\boldsymbol{\delta}}\right]. \tag{8.24}$$

We then calculate the mean absolute error (MAE) of $\hat{\mu}_{\boldsymbol{\delta}}$ with respect to the true data $y_{\boldsymbol{\delta}}$. In the topic modeling case study, we also consider the interpretability of each method's inferred topics using coherence [Mimno et al., 2011] and NPMI [Bouma, 2009, Lau et al., 2014] which are standard within the topic modeling community.

8.6.4 Case study 1: Topic modeling

Topic models [Blei et al., 2003] are widely used in the social sciences [Ramage et al., 2009, Grimmer and Stewart, 2013, Mohr and Bogdanov, 2013, Roberts et al., 2013] for characterizing the high-level thematic structure of text corpora via latent "topics"—i.e., probability distributions over vocabulary items. In many settings, documents contain sensitive information (e.g., emails, survey responses) and individuals may be unwilling to share their data without formal privacy guarantees, such as those provided by differential privacy.

8.6.4.1 Limited-precision local privacy guarantees

In this scenario, a data set Y is a $D \times V$ count matrix, each element of which y_{dv} represents the count of type $v \in [V]$ in document $d \in [D]$. It's natural to consider each document $\boldsymbol{y}_d \equiv (y_{d1}, \ldots, y_{dV})$ to be a single "observation" we seek to privatize. Under LPLP, N determines the neighborhood of documents within ϵ-level local privacy applies. For instance, if $N = 4$, then two emails—one that contained four instances of a vulgar curse word and the same one that did not $\|\boldsymbol{y}_d - \boldsymbol{y}'_d\|_1 = 4$—would be rendered indistinguishable after privatization, assuming small ϵ.

8.6.4.2 Poisson factorization model

Gamma–Poisson matrix factorization is commonly used for topic modeling. Assume each element is drawn $y_{dv} \sim \text{Pois}\,(\mu_{dv})$ where $\mu_{dv} = \sum_{k=1}^{K} \theta_{dk}\,\phi_{kv}$. The factor θ_{dk} represents how much topic k is used in document d, while the factor ϕ_{kv} represents how much word type v is used in topic k. The set of latent variables is thus $Z = \{\Theta, \Phi\}$, where Θ and Φ are $D \times K$ and $K \times V$ non-negative matrices, respectively. It is standard to assume independent gamma priors over the factors—i.e., $\theta_{dk}, \phi_{kv} \sim \Gamma\,(a_0, b_0)$, where we set the shape and rate hyperparameters to $a_0 = 0.1$ and $b_0 = 1$.

8.6.4.3 Synthetic data experiments

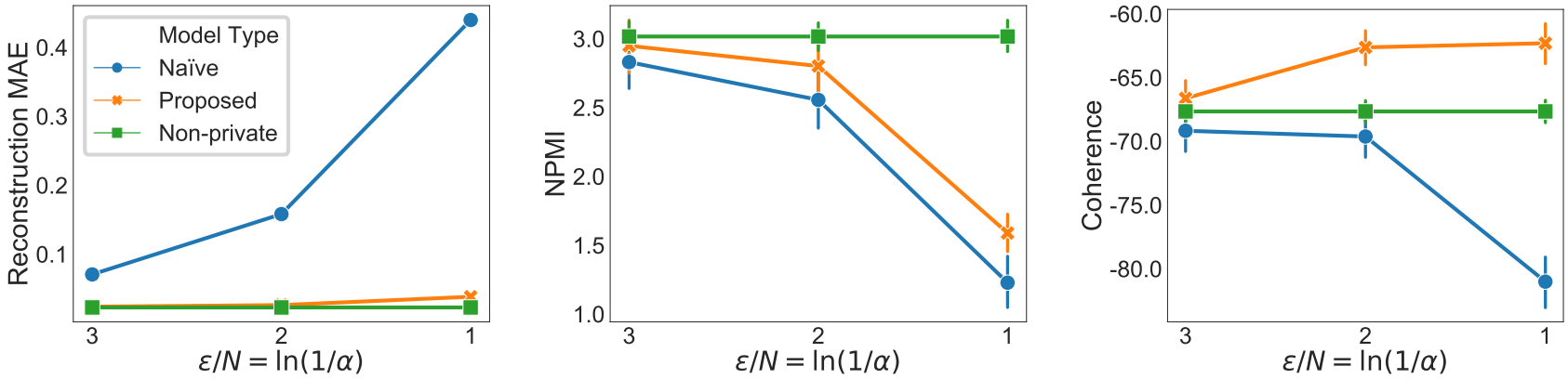
The proposed approach is more effective than the naïve approach at recovering the ground-truth topics Φ^* from synthetically-generated data. We generated a synthetic data set of $D = 90$ documents, with $K = 3$ topics and $V = 15$ word types. We set Φ^* so that the topics were well separated, with each putting the majority of its mass on five different word types. We also ensured that the documents were well separated into three equal groups of thirty, with each putting the majority of its mass on a different topic. We then sampled a data set $y^*_{dv} \sim \text{Pois}(\mu^*_{dv})$ where $\mu^*_{dv} = \sum_{k=1}^{K} \theta^*_{dk}\phi^*_{kv}$. We then generated a heterogeneously-noised data set by sampling the d^{th} document's noise level $\alpha_d \sim \text{Beta}\big(c\,\alpha_0,\, c\,(1-\alpha_0)\big)$ from a beta distribution with mean α_0 and concentration parameter $c = 10$ and then sampling $\tau_{dv} \sim 2\text{Geo}(\alpha_d)$ for each word type v. We repeated this for a small and large value of α_0. For each model, we ran 6,000 sampling iterations, saving every 25^{th} sample after the first 1,000. We selected $\hat{\Phi}$ to be from the posterior sample with the highest joint probability. Note that, due to label-switching, we cannot average the samples of Φ. Following Newman et al. [2009], we then aligned the topic indices of $\hat{\Phi}$ to Φ^* using the Hungarian bipartite matching algorithm. We visualize the results in Fig. 8.1 where we see that the naïve approach performs poorly at recovering the topics in the high noise case.

8.6.4.4 Enron corpus experiments

In these experiments, we use the document-by-term count matrix Y derived from the Enron email corpus. We consider three privacy levels $\epsilon/N \in \{3, 2, 1\}$ specified by the ratio of the privacy budget ϵ to the precision N. For each privacy level, we obtain five different privatized matrices, each by adding two-sided geometric noise with $\alpha = -\exp(\epsilon/N)$ independently to each element. We fit both privacy-preserving models—i.e., the proposed and the naïve approach—to all five privatized matrices for each privacy level. We also fit the non-private approach five independent times to the true matrix. We set $K = 50$ for all models. For every model and matrix, we perform 7,500 MCMC iterations, saving every 100^{th} sample of the latent variables $\Phi^{(s)}, \Theta^{(s)}$ after the first 2,500. We use the fifty saved samples to compute $\hat{\mu}_{dv} = \frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K}\theta_{dk}^{(s)}\phi_{kv}^{(s)}$.

We find that the proposed approach obtains both lower reconstruction error and higher quality topics than the naïve approach. For each model and matrix, we compute the mean absolute error (MAE) of the reconstruction rates $\hat{\mu}_{dv}$ with respect to the true underlying data y_{dv}. These results are visualized in the left subplot of Fig. 8.5 where we see that the proposed approach reconstructs the true data with nearly as low error as non-private inference (that fits to the true data) while the naïve approach has high error which increases dramatically as the noise increases.

To evaluate topic quality, we use two standard metrics—i.e., normalized pointwise mutual information (NPMI) [Lau et al., 2014] and topic coherence [Mimno et al., 2011]— applied to the 10 words with maximum weight for each sampled topic vector $\boldsymbol{\phi}_k^{(s)}$, using the true data as the reference corpus. For each method and privacy level, we average these values across samples. The center and right subplots of Fig. 8.5 visualize the NPMI and coherence results, respectively. The proposed approach obtains higher quality topics than the naïve approach, as measured by both metrics. As measured by coherence, the proposed approach even obtains higher quality topics than the non-private approach.



(a) Reconstruction error (*lower is better*): MAE of the estimated rates $\hat{\mu}_{dv}$ with respect to the true underlying data y_{dv}.

(b) Topic interpretability (*higher is better*) as measured by normalized pointwise mutual information (NPMI).

(c) Topic interpretability (*higher is better*) as measured by coherence.

Figure 8.5: The proposed approach obtains higher quality topics and lower reconstruction error than the naïve approach. When topic quality is measured using coherence (*right*), the proposed approach obtains higher quality topics than even the non-private method. Each plot compares the proposed, naïve, and non-private approaches for three increasing levels of noise (privacy) on the Enron corpus data; the non-private values are constant across privacy levels.

8.6.5 Case study 2: Overlapping community detection

Organizations often ask: are there missing connections between employees that, if present, would significantly reduce duplication of effort? Though social scientists may be able to draw insights based on employees' interactions, sharing such data risks privacy violations. Moreover, standard anonymization procedures can be reverse-engineered adversarially and thus do not provide privacy guarantees [Narayanan and Shmatikov, 2009]. In contrast, the formal privacy guarantees provided by differential privacy may be sufficient for employees to consent to sharing their data.

8.6.5.1 Limited-precision local privacy guarantees

In this setting, a data set Y is a $V \times V$ count matrix, where each element $y_{ij} \in \mathbb{Z}_+$ represents the number of interactions from actor $i \in [V]$ to actor $j \in [V]$. It is natural to consider each count y_{ij} to be a single "observation". Using the geometric mechanism with $\alpha = -\exp(\epsilon/N)$, if i interacted with j three times $y_{ij} = 3$ and $N = 3$, then an adversary would be unable to tell from $\tilde{y}_{ij}^{(\pm)}$ whether i had interacted with j at all, provided ϵ is sufficiently small. Note that if $y_{ij} \gg N$, then an adversary would be able to tell that i had interacted with j, though not the exact number of times.

8.6.5.2 Poisson factorization model

The mixed-membership stochastic block model for learning latent overlapping community structure in social networks [Ball et al., 2011, Gopalan and Blei, 2013, Zhou, 2015] is a special case of Poisson factorization where Y is a $V \times V$ count matrix, each element of which is drawn $y_{ij} \sim \text{Pois}(\mu_{ij})$ where $\mu_{ij} = \sum_{c=1}^{C} \sum_{d=1}^{C} \theta_{ic} \theta_{jd} \pi_{cd}$. The factors θ_{ic} and θ_{jd} represent how much actors i and j participate in communities c and d, respectively, while the factor π_{cd} represents how much actors in community c interact with actors in community d. The set of latent variables is thus $Z = \{\Theta, \Pi\}$ where Θ and Π are $V \times C$ and $C \times C$ non-negative matrices, respectively. We assume independent gamma priors over the factors—i.e., $\theta_{ic}, \pi_{cd} \overset{\text{iid}}{\sim} \Gamma(a_0, b_0)$ with $a_0 = 0.1$ and $b_0 = 1$.

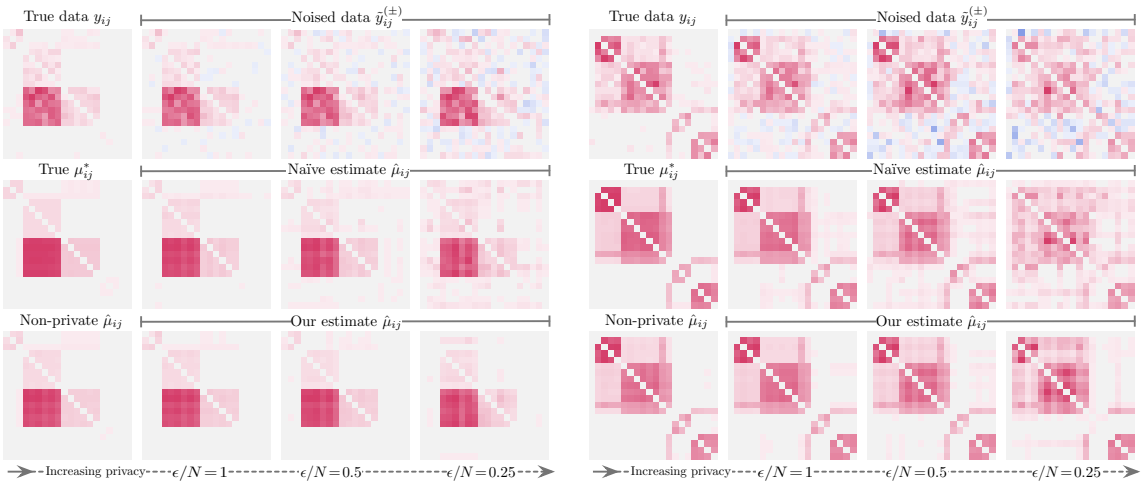


Figure 8.6: Block structure recovery: our method vs. the naïve approach. We generated the non-privatized data synthetically. We then privatized the data using three different levels of noise. The top row depicts the data, using red to denote positive observations and blue to denote negative observations. As privacy increases, the naïve approach overfits the noise and fails to recover the true $\mu_{ij}^\star$ values, predicting high values even for sparse parts of the matrix. In contrast, our method recovers the latent structure, even at high noise levels.

8.6.5.3 Synthetic data experiments

We generated social networks of $V = 20$ actors with $C = 5$ communities. We randomly generated the true parameters $\theta_{ic}^*, \pi_{cd}^* \sim \Gamma(a_0, b_0)$ with $a_0 = 0.01$ and $b_0 = 0.5$ to encourage sparsity; doing so exaggerates the block structure in the network. We then sampled a data set $y_{ij} \sim \text{Pois}(\mu_{ij}^*)$ and added noise $\tau_{ij} \sim 2\text{Geo}(\alpha)$ for three increasing values of α. In each trial, we set N to the empirical mean of the data $N := \hat{\mathbb{E}}[y_{ij}]$ and then set $\alpha := \exp(-\epsilon/N)$ for $\epsilon \in \{2.5, 1, 0.75\}$. For each model, we ran 8,500 MCMC iterations, saving every 25^{th} sample after the first 1,000 and using these samples to compute $\hat{\mu_{ij}}$. In Fig. 8.6, we visually compare the estimates of $\hat{\mu}_{ij}$ by our proposed method to those of the naïve and non-private approaches. The naïve approach overfits the noise, predicting high rates in sparse parts of the matrix. In contrast, the proposed approach reproduces the sparse block structure even under high noise.

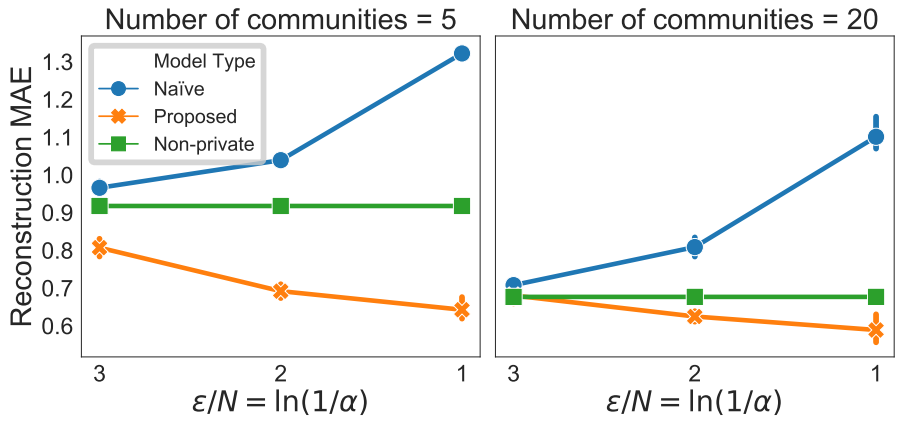
8.6.5.4 Enron corpus experiments

For the Enron data experiments, we follow the same experimental design outlined in the topic modeling case study; we repeat this experiment using three different numbers of communities $C \in \{5, 10, 20\}$. Each method is applied to five privatized matrices for three different privacy levels. We compute $\hat{\mu}_{ij} = \frac{1}{S}\sum_{s=1}^{S}\sum_{c=1}^{C}\sum_{d=1}^{C}\theta_{ic}^{(s)}\theta_{jd}^{(s)}\pi_{cd}^{(s)}$ from each run and measure reconstruction MAE with respect to the true underlying data y_{ij}. In these experiments, each method observes the entire matrix ($\tilde{Y}^{(\pm)}$ for the privacy-preerving methods and Y for the non-private method). Since missing link prediction is a common task in the networks community, we additionally run the same experiments but where a portion of the matrix is masked—specifically, we hold out all entries $\tilde{y}_{ij}^{(\pm)}$ (or y_{ij} for non-private) that involve any of the top 50 senders i or recipients j. We then compute $\hat{\mu}_{ij}$, as before, but only for the missing entries and compare heldout MAE across methods.

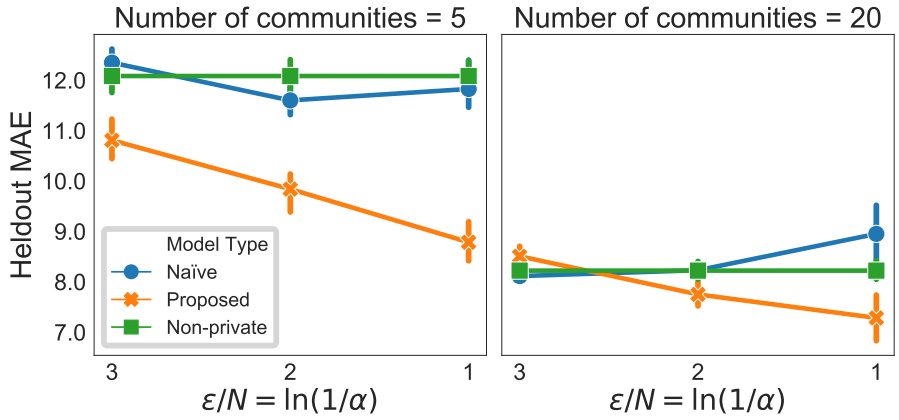
The results for $C \in \{5, 20\}$ are visualized in Fig. 8.7 (the results for $C = 10$ were similar). When reconstructing y_{ij} from observed $\tilde{y}_{ij}^{(\pm)}$, the proposed approach achieves lower error than the naïve approach and lower error than the non-private approach (which directly observes y_{ij}). Similarly, when predicting missing y_{ij}, the proposed approach achieves the lowest error in most settings.

8.7 Discussion

The proposed privacy-preserving MCMC method for Poisson factorization improves substantially over the commonly-used naïve approach. A suprising finding is that the proposed method was also often better at predicting the true y_{ij} from privatized $\tilde{y}_{ij}^{(\pm)}$ than even the non-private approach. Similarly, the the proposed approach inferred more coherent topics. These empirical findings are in fact consistent with known connections between privacy-preserving mechanisms and regularization [Chaudhuri and Monteleoni, 2009]. The proposed approach is able to ex-



(a) Reconstruction error: $\hat{\mu}_{ij}$ is estimated from a noised matrix $\tilde{Y}^{(\pm)}$ that is fully observed.



(b) Heldout link prediction: $\hat{\mu}_{ij}$ is estimated from a partially observed $\tilde{Y}^{(\pm)}$. MAE is calculated with respect to only the data y_{ij} for which the corresponding entries $\tilde{y}_{ij}^{(\pm)}$ were missing.

Figure 8.7: The proposed approach obtains lower error on both reconstruction (*top*) and heldout link prediction (*bottom*) than the naïve and even non-private approach.

plain natural dispersion in the true data as coming from the randomized response mechanism; it may thus be more robust—i.e., less susceptible to inferring spurious structure—than non-private Poisson factorization. Future application of the model $y_{\boldsymbol{\delta}} \sim \text{Skel}(\lambda_{\boldsymbol{\delta}}^{(+)} + \mu_{\boldsymbol{\delta}}, \lambda_{\boldsymbol{\delta}}^{(-)})$ as a robust alternative to Poisson factorization is thus motivated, as is a theoretical characterization of its regularizing properties.

8.8 Proofs

8.8.1 Proof of Theorem 8.1

It suffices to show that for any integer-valued vector $o \in \mathbb{Z}^D$, the following inequality holds for any pair of observations $y, y' \in \mathcal{Y} \subseteq \mathbb{Z}^D$ such that $\|y - y'\|_1 \leq N$:

$$\exp(-\epsilon) \leq \frac{P\left(\mathcal{R}(y) = o\right)}{P\left(\mathcal{R}(y') = o\right)} \leq \exp(\epsilon), \tag{8.25}$$

where $\epsilon = N \ln\left(\frac{1}{\alpha}\right)$.

Let ν denote a D-dimensional noise vector with elements drawn i.i.d from $\nu_d \sim 2\text{Geo}(\alpha)$. Then,

$$\begin{aligned}
\frac{P\left(\mathcal{R}(y) = o\right)}{P\left(\mathcal{R}(y') = o\right)} &= \frac{P(\nu = o - y)}{P(\nu = o - y')} && (8.26)\\
&= \frac{\prod_{d=1}^{D} \frac{1-\alpha}{1+\alpha}\, \alpha^{|o_d - y_d|}}{\prod_{d=1}^{D} \frac{1-\alpha}{1+\alpha}\, \alpha^{|o_d - y'_d|}} && (8.27)\\
&= \alpha^{\left(\sum_{d=1}^{D} |o_d - y_d| - |o_d - y'_d|\right)}. && (8.28)
\end{aligned}$$

By the triangle inequality, we also know that for each d,

$$-|y_d - y'_d| \leq |o_d - y_d| - |o_d - y'_d| \leq |y_d - y'_d|. \tag{8.29}$$

Therefore,

$$-\|y - y'\|_1 \leq \sum_{d=1}^{D} \left(|o_d - y_d| - |o_d - y'_d|\right) \leq \|y - y'\|_1. \tag{8.30}$$

It follows that

$$\alpha^{-N} \leq \frac{P\left(\mathcal{R}(y) = o\right)}{P\left(\mathcal{R}(y') = o\right)} \leq \alpha^{N}. \tag{8.31}$$

If $\epsilon = N \ln\left(\frac{1}{\alpha}\right)$, then we recover the bound in Eq. (8.25).

8.8.2 Proof of Theorem 8.2

By Definition 8.4, a two-sided geometric random variable $\tau \sim 2\text{Geo}(\alpha)$ can be generated as the difference of two i.i.d. geometric random variables:

$$g^{(+)}, g^{(-)} \overset{\text{iid}}{\sim} \text{NB}(1, \alpha), \tag{8.32}$$

$$\tau := g^{(+)} - g^{(-)}. \tag{8.33}$$

where the geometric distribution is a special case of the negative binomial distribution, with shape parameter equal to one [Johnson et al., 2005]. By Definition 3.16, the negative binomial distribution can be represented as a gamma–Poisson mixture distribution. We can therefore re-express equations 8.32– 8.33 as:

$$\lambda^{(+)}, \lambda^{(-)} \overset{\text{iid}}{\sim} \text{Gam}(1, \tfrac{\alpha}{1-\alpha}), \tag{8.34}$$

$$g^{(+)} \sim \text{Pois}(\lambda^{(+)}), \tag{8.35}$$

$$g^{(-)} \sim \text{Pois}(\lambda^{(-)}), \tag{8.36}$$

$$\tau := g^{(+)} - g^{(-)}. \tag{8.37}$$

Finally, the gamma distribution with shape parameter equal to one is equivalent to the exponential distribution, while the difference of two independent Poisson random variables is marginally Skellam distributed (Definition 8.7). We me thus rewrite equations 8.34– 8.37 as:

$$\lambda^{(+)}, \lambda^{(-)} \overset{\text{iid}}{\sim} \text{Exp}(\tfrac{\alpha}{1-\alpha}), \tag{8.38}$$

$$\tau \sim \text{Skel}\left(\lambda^{(+)}, \lambda^{(-)}\right) \tag{8.39}$$

8.8.3 Proof of Theorem 8.3

Consider the joint distribution of y_1 and y_2:

$$P(y_1, y_2) = \text{Pois}(y_1; \lambda^{(+)}) \, \text{Pois}(y_2; \lambda^{(-)}) \tag{8.40}$$

$$= \frac{(\lambda^{(+)})^{y_1}}{y_1!} \, e^{-\lambda^{(+)}} \, \frac{(\lambda^{(-)})^{y_2}}{y_2!} \, e^{-\lambda^{(-)}} \tag{8.41}$$

$$= \frac{(\sqrt{\lambda^{(+)}\lambda^{(-)}})^{y_1+y_2}}{y_1! \, y_2!} \, e^{-(\lambda^{(+)}+\lambda^{(-)})} \left(\frac{\lambda^{(+)}}{\lambda^{(-)}}\right)^{(y_1-y_2)/2}. \tag{8.42}$$

If $y_1 \geq y_2$, then

$$P(y_1, y_2) = \frac{(\sqrt{\lambda^{(+)}\lambda^{(-)}})^{y_1+y_2}}{I_{y_1-y_2}(2\sqrt{\lambda^{(+)}\lambda^{(-)}}) \, y_1! \, y_2!} \, e^{-(\lambda^{(+)}+\lambda^{(-)})} \left(\frac{\lambda^{(+)}}{\lambda^{(-)}}\right)^{(y_1-y_2)/2} I_{y_1-y_2}(2\sqrt{\lambda^{(+)}\lambda^{(-)}}) \tag{8.43}$$

$$= \text{Bes}\left(y_2;\, y_1 - y_2, 2\sqrt{\lambda^{(+)}\lambda^{(-)}}\right) \text{Skel}(y_1 - y_2; \lambda^{(+)}, \lambda^{(-)}); \tag{8.44}$$

otherwise

$$P(y_1, y_2) = \frac{(\sqrt{\lambda^{(+)}\lambda^{(-)}})^{y_1+y_2}}{I_{y_2-y_1}(2\sqrt{\lambda^{(+)}\lambda^{(-)}}) \, y_1! \, y_2!} \, e^{-(\lambda^{(+)}+\lambda^{(-)})} \left(\frac{\lambda^{(-)}}{\lambda^{(+)}}\right)^{(y_2-y_1)/2} I_{y_2-y_1}(2\sqrt{\lambda^{(+)}\lambda^{(-)}}) \tag{8.45}$$

$$= \text{Bes}\left(y_1;\, y_2 - y_1, 2\sqrt{\lambda^{(+)}\lambda^{(-)}}\right) \text{Skel}(y_2 - y_1; \lambda^{(-)}, \lambda^{(+)})$$

$$= \text{Bes}\left(y_1; -(y_1 - y_2), 2\sqrt{\lambda^{(+)}\lambda^{(-)}}\right) \text{Skel}(y_1 - y_2; \lambda^{(+)}, \lambda^{(-)}). \tag{8.46}$$

If

$$m := \min\{y_1, y_2\}, \;\; \tau := y_1 - y_2, \tag{8.47}$$

then

$$y_2 = m, \;\; y_1 = m + \tau \;\; \text{if } \tau \geq 0 \tag{8.48}$$

$$y_1 = m, \;\; y_2 = m - \tau \;\; \text{otherwise} \tag{8.49}$$

and

$$\left|\begin{array}{cc} \frac{\partial y_1}{\partial m} & \frac{\partial y_1}{\partial \tau} \\ \frac{\partial y_2}{\partial m} & \frac{\partial y_2}{\partial \tau} \end{array}\right| = \left|\begin{array}{cc} 1 & 1 \\ 1 & 0 \end{array}\right|^{\tau\geq 0} \left|\begin{array}{cc} 1 & 0 \\ 1 & -1 \end{array}\right|^{\tau<0} = 1, \tag{8.50}$$

so

$$\begin{aligned} P(m,\tau) &= P(y_1,y_2) \left|\begin{array}{cc} \frac{\partial y_1}{\partial m} & \frac{\partial y_1}{\partial \tau} \\ \frac{\partial y_2}{\partial m} & \frac{\partial y_2}{\partial \tau} \end{array}\right| \\ &= \mathrm{Bes}\left(m; |\tau|, 2\sqrt{\lambda^{(+)}\lambda^{(-)}}\right) \mathrm{Skel}(\tau; \lambda^{(+)}, \lambda^{(-)}). \end{aligned} \tag{8.51}$$

CHAPTER 9

CONCLUSIONS AND FUTURE DIRECTIONS

This thesis formalizes the framework of *allocative Poisson Factorization*. APF is a subset of Poisson factorization models within which posterior inference scales linearly with only the number of non-zero event counts (or, equivalently, the total number of event tokens) in the data. APF unifies and generalizes widely-used models in the statistics literature on contingency table analysis, non-negative tensor decomposition, and probabilistic models for discrete data in machine learning, among others. Many of the connections between these models have long been understood and unifying frameworks have been proposed [Buntine, 2002, Buntine and Jakulin, 2006, Dunson and Xing, 2009]. Connections and unifications have continued to be written about even in recent years [Johndrow et al., 2017, Cemgil et al., 2019].

A fair question to ask is: why APF? Why now? APF is defined by a single condition—i.e., that the latent Poisson rate $\mu_{\boldsymbol{\delta}}$ be a *multilinear* function of shared model parameters. This simple definition is both *descriptive*—i.e., it describes all the models for which there exists a latent source representation (Section 3.2) that allows the model to be written in terms of event counts or event tokens—as well as *prescriptive*—i.e., a researcher building a model need only obey that one condition to ensure that posterior inference scales well in the high-dimensional but sparse setting. This condition also highlights the main challenge in building and fitting APF models. The multilinear condition on the non-negative Poisson rate parameter means we can only impose non-negative prior distributions—e.g., gamma and Dirichlet—over the model parameters. This notably prevents the use of the Gaussian distribution, whose

analytic convenience researchers have traditionally relied on to construct complex hierarchical priors e.g., for time-series or networks models. Recent advances in auxiliary variable augmentation schemes—e.g., augment-and-conquer [Zhou and Carin, 2012] or Pólya–gamma augmentation [Polson et al., 2013]—have permitted efficient and analytically closed-form posterior inference in a wide array of non-conjugate models that were previously considered intractable. An example of such a model is given in Chapter 6, where the augment-and-conquer scheme is applied recursively in a dynamical system of gamma random variables tied through their shape parameter. Beyond augmentation schemes and conjugacy, there also exists a web of relationships among non-negative distributions that have thus far been under-exploited. An example of a useful but overlooked relationship is Definition 7.3—i.e., a Poisson prior over the shape parameter of a gamma distribution yields a Bessel distribution as its posterior [Yuan and Kalbfleisch, 2000]. Chapters 7 and 8 both exploit this fact to the construct non-conjugate yet analytically closed-form APF models. The relationship between the Skellam and Bessel distributions introduced by Theorem 8.3 is another example; this relationship is general and widely applicable beyond the model and application described in that chapter.

The work in this thesis prompts several promising future directions of research. Foremost among them, is the application of the APF framework to model non-Poisson distributed observations. The Poisson distribution can be linked to a wide range of other distributions—both discrete and continuous—via the compound Poisson construction [Adelson, 1966]. Basbug and Engelhardt [2016] draw on this fact to probabilistically link a latent count matrix to an observed (potentially non-count) one and then factorize it using Poisson matrix factorization. This idea can be generalized to all APF models and expressed as

$$y_{\boldsymbol{\delta}} \sim \text{Pois}\left(\mu_{\boldsymbol{\delta}}\right), \tag{9.1}$$

$$x_{\boldsymbol{\delta}} = \sum_{i=1}^{y_{\boldsymbol{\delta}}} x_{\boldsymbol{\delta}i}, \quad \text{where } x_{\boldsymbol{\delta}i} \sim f(\cdots). \tag{9.2}$$

where $y_{\boldsymbol{\delta}}$ is now latent, $x_{\boldsymbol{\delta}}$ is the observation, and $f(\cdots)$ is the emission distribution. One example of an emission distribution might be gamma—e.g., $f(\cdots) = \Gamma\left(\alpha_0, \beta_0\right)$—which would be appropriate to model positive real-valued data, as Zhou et al. [2016] does. For MCMC inference in compound Poisson models, we need only be able to efficiently sample $y_{\boldsymbol{\delta}}$ from its complete conditional—$\left(y_{\boldsymbol{\delta}} \,|\, -\right) \sim P(y_{\boldsymbol{\delta}} \,|\, x_{\boldsymbol{\delta}}, \mu_{\boldsymbol{\delta}}, -)$. With a sample of $y_{\boldsymbol{\delta}}$ in hand, we may treat it as observed and update the other parameters exactly as in standard APF. This idea is morally similar to that in Chapter 8, where the true count $y_{\boldsymbol{\delta}}$ must be re-sampled as a latent variable, conditioned on the noised (i.e., privatized) count $\tilde{y}_{\boldsymbol{\delta}}^{(\pm)}$.

A surprising result of Chapter 8 was that the private version of APF, which conditions on noised data and assumes a Skellam likelihood, sometimes outperforms non-private APF. One hypothesis is that this stems from the Poisson assumption being too restrictive for real-world count data: the private method outperforms standard APF due to its assumption of a more dispersed likelihood (Skellam) and does so *despite* conditioning on noised data. If this hypothesis is true, APF could be easily modified to assume an overdispersed negative binomial likelihood. The negative binomial can be constructed as a compound Poisson (Definition 3.17). Moreover, under this construction the complete conditional of $y_{\boldsymbol{\delta}}$ is available in closed form as the CRT (Definition 3.20). Thus, all that is needed to modify APF is to add a CRT sampling step. This approach is taken for count matrices by Zhou [2018].

APF can be extended to model data sets consisting of multiple tensors that overlap on different modes. This is sometimes referred to as "multi-view factorization" [Khan and Kaski, 2014]. Gopalan et al. [2014b] introduce "collaborative Poisson factorization" which factorizes two matrices that overlap on one mode. Yılmaz et al. [2011]

presents a general framework called "generalized coupled factorization" for multiple tensors. Beyond multiple tensors, we may also consider a data set of event tokens of arbitrary dimensionality that overlap on modes. Such a data set may not be able to be represented in terms of count tensors but would still be amenable to the token-based representation of APF. A future direction is an implementation of the APF framework that takes, as input a list of tokens of arbitrary dimensionality and automatically infers the size and cardinality of the latent parameters matrices.

This thesis has focused on batch inference—both MCMC and variational—that assumes the entire data set can be fit into memory. All of the algorithms presented here are based on closed-form complete conditionals. Many of them can be converted into stochastic variational algorithms [Hoffman et al., 2013] that stream over larger data sets. We may also use streaming MCMC techniques to scale some of the algorithms presented here—for instance, Guo et al. [2018] construct a deep version of the dynamical system in Chapter 6 and give a streaming MCMC algorithm. This thesis has also focused on uncollapsed Gibbs sampling, wherein the latent sources are re-sampled from a multinomial conditioned on samples of the parameters. There are many papers on fast CGS algorithms for LDA [Yao et al., 2009, Li et al., 2014, Yuan et al., 2015, Chen et al., 2016]; these ideas can be generalized to APF.

Finally, the gamma–Poisson–gamma motif introduced in Chapter 7 can be applied recursively to construct many new models—e.g., a belief network that alternates between Poisson and gamma layers. This motif can be thought of as an alternative to gamma hierarchies that require the augment-and-conquer scheme for inference. Moreover, when $\epsilon = 0$, the gamma–Poisson–gamma allows for true sparsity in both the discrete and continuous states. One promising application of this is as a prior over the core tensor in the Tucker decomposition model (BPTD) of Chapter 5. The allocation step in standard BPTD scales linearly with the number of latent classes—i.e., the number of cells in the core tensor. However, if the core tensor contained true

zeros, the allocation step would scale only with the number of non-zero cells (for the same reason it scales linearly with the non-zeros in the data). In general, promoting true sparsity among the parameters of APF models, and then exploiting that sparsity for more efficient computation is a promising future direction.

BIBLIOGRAPHY

A. Acharya, J. Ghosh, and M. Zhou. Nonparametric Bayesian factor analysis for dynamic count matrices. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

R. M. Adelson. Compound Poisson distributions. *OR*, 17(1):73–75, 1966.

Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

E. M. Airoldi, D. M. Blei, S. E. Feinberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

R Michael Alvarez. *Computational Social Science*. Cambridge University Press, 2016.

Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914. ACM, 2013.

Joseph Bafumi, Andrew Gelman, David K Park, and Noah Kaplan. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13(2):171–187, 2005.

B. Ball, B. Karrer, and M. E. J. Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3), 2011.

M. E. Basbug and B. Engelhardt. Hierarchical compound Poisson factorization. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Garrett Bernstein, Ryan McKenna, Tao Sun, Daniel Sheldon, Michael Hay, and Gerome Miklau. Differentially private learning of undirected graphical models using collective graphical models. *arXiv preprint arXiv:1706.04646*, 2017.

Anirban Bhattacharya and David B Dunson. Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107 (497):362–377, 2012.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer Science & Business Media, 2006.

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, 2006.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 601–608, 2001.

David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

E. Boschee, J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward. ICEWS coded event data. Harvard Dataverse, 2015. V10.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

Amber E Boydstun. *Making the news: Politics, the media, and agenda setting*. University of Chicago Press, 2013.

Tamara Broderick, Lester Mackey, John Paisley, and Michael I Jordan. Combinatorial clustering and the beta negative binomial process. *arXiv preprint arXiv:1111.1802*, 2011.

Tamara Broderick, Lester Mackey, John Paisley, and Michael I Jordan. Combinatorial clustering and the beta negative binomial process. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):290–306, 2015.

M. G. Bulmer. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, pages 101–110, 1974.

Wray Buntine. Variational extensions to EM and multinomial PCA. In *European Conference on Machine Learning*, pages 23–34. Springer, 2002.

Wray Buntine and Aleks Jakulin. Discrete component analysis. In *Subspace, Latent Structure and Feature Selection*, pages 1–33. Springer, 2006.

J. Canny. GaP: A factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129, 2004.

A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

Ali Taylan Cemgil, Mehmet Burak Kurutmaz, Sinan Yildirim, Melih Barsbey, and Umut Simsekli. Bayesian allocation model: Inference by sequential Monte Carlo for nonnegative tensor factorizations and topic models using Polya urns. *arXiv preprint arXiv:1903.04478*, 2019.

L. Charlin, R. Ranganath, J. McInerney, and D. M. Blei. Dynamic Poisson factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 155–162, 2015.

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, pages 289–296, 2009.

Jianfei Chen, Kaiwei Li, Jun Zhu, and Wenguang Chen. Warplda: a cache efficient o (1) algorithm for latent Dirichlet allocation. *Proceedings of the VLDB Endowment*, 9(10):744–755, 2016.

E. C. Chi and T. G. Kolda. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, 2012.

A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

Andrzej Cichocki, Rafal Zdunek, Seungjin Choi, Robert Plemmons, and Shun-Ichi Amari. Non-negative tensor factorization using alpha and beta divergences. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 3, pages III–1393. IEEE, 2007.

Clifford C Clogg and Leo A Goodman. Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, 79 (388):762–771, 1984.

R. Corless, G. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the LambertW function. *Advances in Computational Mathematics*, 5(1):329–359, 1996.

David Cox and P.A.W. Lewis. The statistical analysis of series of events. 1966.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Luc Devroye. Simulating Bessel random variables. *Statistics & Probability Letters*, 57(3):249–257, 2002.

Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin I. P. Rubinstein. Robust and private Bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 291–305, 2014.

Christos Dimitrakakis, Blaine Nelson, Zuhe Zhang, Aikateirni Mitrokotsa, and Benjamin Rubinstein. Differential privacy for Bayesian inference through posterior sampling. *Journal of Machine Learning Research*, 18(11):1–39, 2017.

Robert Dorfman. A formula for the Gini coefficient. *The Review of Economics and Statistics*, pages 146–149, 1979.

C. DuBois and P. Smyth. Modeling relational events via latent classes. In *Proceedings of the Sixteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 803–812, 2010.

D. B. Dunson and A. H. Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1):11–25, 2005.

David B Dunson and Chuanhua Xing. Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009.

J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2012.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, volume 3876, pages 265–284, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Robert S Erikson, Pablo M Pinto, and Kelly T Rader. Dyadic analysis in international relations: A cautionary tale. *Political Analysis*, 22(4):457–463, 2014.

Beyza Ermis and A Taylan Cemgil. A Bayesian tensor factorization model via variational inference for link prediction. *arXiv preprint arXiv:1409.8276*, 2014.

Ugo Fano. Ionization yield of radiations. ii. the fluctuations of the number of ions. *Physical Review*, 72(1):26, 1947.

Stephen E Fienberg and Alessandro Rinaldo. Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference*, 137(11):3430–3445, 2007.

Lorenzo Finesso and Peter Spreij. Nonnegative matrix factorization and I-divergence alternating minimization. *Linear Algebra and its Applications*, 416(2-3):270–287, 2006.

Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

Mark D Flood, Jonathan Katz, Stephen J Ong, and Adam D Smith. Cryptography and the economics of supervisory information: Balancing transparency and confidentiality. 2013.

James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving Bayesian data analysis. 2016.

Eric Gaussier and Cyril Goutte. Relation between PLSA and NMF and implications. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 601–602. ACM, 2005.

Andrew Gelman. What's the most important thing in statistics that's not in the textbooks? http://andrewgelman.com/2015/04/28/whats-important-thing-statistics-thats-not-textbooks/, 2015.

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

D. J. Gerner, P. A. Schrodt, R. Abu-Jabr, and Ö. Yilmaz. Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. Working paper.

Sean Gerrish and David M Blei. Predicting legislative roll calls from text. In *Proceedings of the 28th international conference on machine learning (icml-11)*, pages 489–496, 2011.

Z. Ghahramani and S. T. Roweis. Learning nonlinear dynamical systems using an EM algorithm. In *Advances in Neural Information Processing Systems*, pages 431–437, 1998.

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

Nicolas Gillis and François Glineur. Nonnegative factorization and the maximum edge biclique problem. *arXiv preprint arXiv:0810.4225*, 2008.

Mark Girolami and Ata Kabán. On an equivalence between PLSI and LDA. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 433–434. ACM, 2003.

K-I Goh and A-L Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002, 2008.

Edward F Gonzalez and Yin Zhang. Accelerating the lee-seung algorithm for non-negative matrix factorization. *Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*, pages 1–13, 2005.

Leo A Goodman. Latent class analysis: The empirical study of latent types, latent variables and latent structures. *Applied latent class analysis*, 2002.

P. Gopalan, F. J. R. Ruiz, R. Ranganath, and D. M. Blei. Bayesian nonparametric Poisson factorization for recommendation systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pages 275–283, 2014a.

P. Gopalan, J. Hofman, and D. Blei. Scalable recommendation with hierarchical Poisson factorization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 2015.

Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36): 14534–14539, 2013.

Prem K Gopalan, Laurent Charlin, and David Blei. Content-based recommendations with Poisson factorization. In *Advances in Neural Information Processing Systems*, pages 3176–3184, 2014b.

Donald P Green, Soo Yeon Kim, and David H Yoon. Dirty pool. *International Organization*, 55(2):441–468, 2001.

Tom Griffiths. Gibbs sampling in the generative model of latent Dirichlet allocation. 2002.

Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls fo automatic content analysis methods for political texts. *Political Analysis*, pages 1–31, 2013.

Dandan Guo, Bo Chen, Hao Zhang, and Mingyuan Zhou. Deep Poisson gamma dynamical systems. In *Advances in Neural Information Processing Systems*, pages 8442–8452, 2018.

David J Hand. Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 445–492, 1996.

Peter Harremoës, Oliver Johnson, and Ioannis Kontoyiannis. Thinning, entropy, and the law of thin numbers. *IEEE Transactions on Information Theory*, 56(9):4228–4244, 2010.

R. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

S. S. Haykin. *Kalman Filtering and Neural Networks*. 2001.

Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169, 2015.

Peter D Hoff and Michael D Ward. Modeling dependencies in international relations networks. *Political Analysis*, 12(2):160–175, 2004.

Peter D Hoff et al. Equivariant and scale-free Tucker decomposition models. *Bayesian Analysis*, 11(3):627–648, 2016.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

Thomas Hofmann, Jan Puzicha, and Michael I Jordan. Learning from dyadic data. In *Advances in neural information processing systems*, pages 466–472, 1999.

Daniel J Hopkins and Gary King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010.

C. Hu, P. Rai, C. Chen, M. Harding, and L. Carin. Scalable Bayesian non-negative tensor factorization for massive count data. In *Machine Learning and Knowledge Discovery in Databases*, volume 9285 of *Lecture Notes in Computer Science*, pages 53–70, 2015.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

James E Johndrow, Anirban Bhattacharya, and David B Dunson. Tensor decompositions and sparse log-linear models. *Annals of statistics*, 45(1):1, 2017.

N. Johnson, A. W. Kemp, and S. Kotz. *Univariate Discrete Distributions*. John Wiley & Sons, 2005.

Norman Lloyd Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Discrete multivariate distributions*, volume 165. Wiley New York, 1997.

M. I. Jordan. Hierarchical models, nested models, and completely random measures. In M.-H. Chen, D. K. Dey, P. Müller, D. Sun, and K. Ye, editors, *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*. Springer, 2010.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37 (2):183–233, 1999.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1), 2011.

Vishesh Karwa, Aleksandra B Slavković, and Pavel Krivitsky. Differentially private exponential random graphs. In *International Conference on Privacy in Statistical Databases*, pages 143–155. Springer, 2014.

SK Katti and John Gurland. The Poisson Pascal distribution. *Biometrics*, 17(4): 527–538, 1961.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, 2006.

Suleiman A Khan and Samuel Kaski. Bayesian multi-view tensor factorization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 656–671. Springer, 2014.

Y.-D. Kim and S. Choi. Nonnegative Tucker decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

Gary King. Proper nouns and methodological propriety: Pooling dyads in international relations data. *International Organization*, 55(2):497–507, 2001.

Gary King and Langche Zeng. Explaining rare events in international relations. *International Organization*, 55(3):693–715, 2001.

J. F. C Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21 (1):59–78, 1967.

J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

Bryan Klimt and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer, 2004.

T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Tsuyoshi Kunihama and David B Dunson. Bayesian modeling of temporal dependence in large sparse contingency tables. *Journal of the American Statistical Association*, 108(504):1324–1338, 2013.

Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, 2014.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Paul Felix Lazarsfeld and Neil W Henry. *Latent structure analysis*. Houghton Mifflin Co., 1968.

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

Lucien Le Cam et al. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, 1960.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

Seonjoo Lee, Pauline E Chugh, Haipeng Shen, R Eberle, and Dirk P Dittmer. Poisson factor models with applications to non-normalized microrna profiling. *Bioinformatics*, 29(9):1105–1111, 2013.

K. Leetaru and P. Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. Working paper, 2013.

Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. Reducing the sampling complexity of topic models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 891–900. ACM, 2014.

Dawen Liang, John William Paisley, Dan Ellis, et al. Codebook-based scalable music tagging with Poisson matrix factorization. In *ISMIR*, pages 167–172, 2014.

Dawen Liang, Laurent Charlin, James McInerney, and David M Blei. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 951–961. International World Wide Web Conferences Steering Committee, 2016.

Chih-Jen Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18(6):1589–1596, 2007.

Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421, 2014.

Eugene Lukacs. A characterization of the gamma distribution. *The Annals of Mathematical Statistics*, 26(2):319–324, 1955.

J. H. Macke, L. Buesing, J. P. Cunningham, B. M. Yu, K. V. Krishna, and M. Sahani. Empirical models of spiking in neural populations. In *Advances in Neural Information Processing Systems*, pages 1350–1358, 2011.

Roman N Makarov and Devin Glew. Exact simulation of Bessel diffusions. *Monte Carlo Methods and Applications*, 16(3-4):283–306, 2010.

Daniel Manrique-Vallier and Jerome P Reiter. Bayesian simultaneous edit and imputation for multivariate categorical data. *Journal of the American Statistical Association*, 112(520):1708–1719, 2017.

B. Marlin. Collaborative filtering: A machine learning perspective. Master's thesis, University of Toronto, 2004.

P. McCullagh and J. A. Nelder. *Generalized linear models*. 1989.

Allan L McCutcheon. *Latent class analysis*. Number 64. Sage, 1987.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359. Morgan Kaufmann Publishers Inc., 2002.

Thomas P Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

Marvin L Minsky. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI magazine*, 12(2):34–34, 1991.

John Mohr and Petko Bogdanov, editors. *Poetics: Topic Models and the Cultural Sciences*, volume 41. 2013.

M. Mørup, L. K. Hansen, and S. M. Arnfred. Algorithms for sparse nonnegative Tucker decompositions. *Neural Computation*, 20(8):2112–2131, 2008.

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Jared S Murray and Jerome P Reiter. Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479, 2016.

Shinichi Nakajima and Masashi Sugiyama. Implicit regularization in variational Bayesian matrix factorization. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 815–822. Omnipress, 2010.

Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, pages 173–187, 2009.

Yuval Nardi, Alessandro Rinaldo, et al. The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli*, 18(3):945–974, 2012.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10(Aug):1801–1828, 2009.

M. Nickel, V. Tresp, and H.-P. Kriegel. Factorizing YAGO: Scalable machine learning for linked data. In *Proceedings of the Twenty-First International World Wide Web Conference*, pages 271–280, 2012.

M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs: From multi-relational link prediction to automated knowledge graph construction. arXiv:1503.00759, 2015.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

Sean P O'Brien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International studies review*, 12(1):87–104, 2010.

John Paisley, David M. Blei, and Michael I. Jordan. Bayesian nonnegative matrix factorization with stochastic variational inference. In Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva, and Stephen E. Fienberg, editors, *Handbook of Mixed Membership Models and Their Applications*, pages 203–222. 2014.

Antonis Papadimitriou, Arjun Narayan, and Andreas Haeberlen. DStress: Efficient differentially private computations on distributed data. In *Proceedings of the Twelfth European Conference on Computer Systems*, pages 560–574. ACM, 2017.

Mijung Park, James Foulds, Kamalika Chaudhuri, and Max Welling. Private topic modeling. *arXiv:1609.04120*, 2016.

Jennie L Pearce and Mark S Boyce. Modelling distribution and abundance with presence-only data. *Journal of applied ecology*, 43(3):405–412, 2006.

Charles S Peirce. The numerical measure of the success of predictions. *Science*, 4 (93):453–454, 1884.

Paul Poast. (Mis)using dyadic data to analyze multilateral events. *Political Analysis*, 18(4):403–425, 2010.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

MP Quine and E Seneta. Bortkiewicz's data and the law of small numbers. *International Statistical Review/Revue Internationale de Statistique*, pages 173–181, 1987.

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D. Manning, and Daniel A. McFarland. Topic modeling for the social sciences. In *NIPS Workshop on Applications for Topic Models*, 2009.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. The structural topic model and applied social science. In *NIPS Workshop on Topic Models: Computation, Application, and Evaluation*, 2013.

Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

W. J. Rugh. *Linear System Theory*. Pearson, 1995.

Lawrence Saul and Fernando Pereira. Aggregate and mixed-order markov models for statistical language processing. In *Second Conference on Empirical Methods in Natural Language Processing*, 1997.

A. Schein, J. Paisley, D. M. Blei, and H. Wallach. Bayesian Poisson tensor factorization for inferrring multilateral relations from sparse dyadic event counts. In *Proceedings of the Twenty-First ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1045–1054, 2015.

Aaron Schein, John Paisley, David M Blei, and H Wallach. Inferring polyadic events with Poisson tensor factorization. In *Proceedings of the NIPS 2014 Workshop on" Networks: From Graphs to Rich Data*, 2014.

Aaron Schein, Hanna Wallach, and Mingyuan Zhou. Poisson–gamma dynamical systems. In *Advances in Neural Information Processing Systems 29*, pages 5005–5013, 2016a.

Aaron Schein, Mingyuan Zhou, David M. Blei, and Hanna Wallach. Bayesian Poisson Tucker decomposition for learning the structure of international relations. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016b.

Aaron Schein, Zhiwei Steven Wu, Alexandra Schofield, Mingyuan Zhou, and Hanna Wallach. Locally private Bayesian inference for count models. *arXiv preprint arXiv:1803.08471*, 2018.

M. N. Schmidt and M. Mørup. Nonparametric Bayesian modeling of complex networks: An introduction. *IEEE Signal Processing Magazine*, 30(3):110–128, 2013.

Philip A Schrodt. Event data in foreign policy analysis. *Foreign Policy Analysis: Continuity and Change in Its Second Generation*, pages 145–166, 1995.

Philip A Schrodt. Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research*, 51(2):287–300, 2014.

Galit Shmueli. To explain or to predict? *Statistical science*, pages 289–310, 2010.

Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

John G. Skellam. The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society, Series A (General)*, 109:296, 1946.

Suvrit Sra and Inderjit S Dhillon. Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in neural information processing systems*, pages 283–290, 2006.

Robert GD Steel. Relation between Poisson and multinomial distributions. 1953.

Brandon Stewart. Latent factor regressions for the social sciences. *Harvard University: Department of Government Job Market Paper*, 2014.

Michalis K. Titsias. The infinite gamma–Poisson feature model. In *Advances in Neural Information Processing Systems 21*, pages 1513–1520, 2008.

L. R. Tucker. The extension of factor analysis to three-dimensional matrices. In N. Frederiksen and H. Gulliksen, editors, *Contributions to Mathematical Psychology*. Holt, Rinehart and Winston, 1964.

Hanna Wallach. Research statement. https://people.cs.umass.edu/~wallach/research.pdf, 2012.

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: posterior sampling and stochastic gradient Monte Carlo. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2493–2502, 2015.

Michael D Ward, Andreas Beger, Josh Cutler, Matt Dickenson, Cassy Dorff, and Ben Radford. Comparing GDELT and ICEWS event data. *Analysis*, 21(1):267–97, 2013.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

M. Welling and M. Weber. Positive tensor factorization. *Pattern Recognition Letters*, 22(12):1255–1261, 2001.

Lucy Whitaker. On the Poisson law of small numbers. *Biometrika*, 10(1):36–71, 1914.

Thomas D Wickens. *Multiway contingency tables analysis for the social sciences*. Psychology Press, 2014.

Wikipedia contributors. 2006 East Timorese crisis — Wikipedia, the free encyclopedia, 2018a. URL https://en.wikipedia.org/w/index.php?title=2006_East_Timorese_crisis&oldid=859675634. [Online; accessed 14-December-2018].

Wikipedia contributors. Embassy of Ecuador, London — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Embassy_of_Ecuador,_London&oldid=835235962, 2018b. [Online; accessed 16-May-2018].

Wikipedia contributors. Operation Infinite Reach — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Operation_Infinite_Reach&oldid=839492999, 2018c. [Online; accessed 16-May-2018].

Wikipedia contributors. Japanese Iraq reconstruction and support group — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Japanese_Iraq_Reconstruction_and_Support_Group&oldid=837624918, 2018d. [Online; accessed 16-May-2018].

Wikipedia contributors. Jyllands-Posten Muhammad cartoons controversy — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Jyllands-Posten_Muhammad_cartoons_controversy&oldid=836434066, 2018e. [Online; accessed 16-May-2018].

Wikipedia contributors. Six-party talks — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Six-party_talks&oldid=841389786, 2018f. [Online; accessed 16-May-2018].

Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. In *Advances in Neural Information Processing Systems 23*, pages 2451–2459, 2010.

John Winn and Christopher M Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694, 2005.

Christopher Winship, Robert D Mare, and John Robert Warren. Latent class models for contingency tables with missing data. *Applied latent class analysis*, page 408, 2002.

Z. Xu, F. Yan, and Y. Qi. Infinite Tucker decomposition: Nonparametric Bayesian models for multiway data analysis. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, pages 1023–1030, 2012.

Xiaolin Yang, Stephen E Fienberg, and Alessandro Rinaldo. Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications. *Journal of Privacy and Confidentiality*, 4(1):5, 2012.

Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946. ACM, 2009.

Kenan Y Yılmaz, Ali T Cemgil, and Umut Simsekli. Generalised coupled tensor factorisation. In *Advances in neural information processing systems*, pages 2151–2159, 2011.

Jiho Yoo and Seungjin Choi. Probabilistic matrix tri-factorization. 2009.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. Lightlda: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1351–1361. International World Wide Web Conferences Steering Committee, 2015.

Lin Yuan and John D. Kalbfleisch. On the Bessel distribution and related problems. *Annals of the Institute of Statistical Mathematics*, 52(3):438–447, 2000.

M Yvonne, MM Bishop, Paul W Holland, and Stephen E Fienberg. *Discrete multivariate analysis: theory and practice*. MIT Press, 1975.

Zuhe Zhang, Benjamin I. P. Rubinstein, and Christos Dimitrakakis. On the differential privacy of Bayesian inference. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2365–2371, 2016.

Q. Zhao, L. Zhang, and A. Cichocki. Bayesian CP factorization of incomplete tensors with automatic rank determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1751–1763, 2015.

M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 1135–1143, 2015.

M. Zhou and L. Carin. Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems Twenty-Five*, pages 2546–2554, 2012.

M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015.

Mingyuan Zhou. Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis*, 2018.

Mingyuan Zhou, Lauren A Hannah, David B Dunson, and Lawrence Carin. Beta-negative binomial process and Poisson factor analysis. *Journal of Machine Learning Research*, 2012.

Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *The Journal of Machine Learning Research*, 17(1):5656–5699, 2016.