The Thesis Committee for Dylan Zachary Anderson
Certifies that this is the approved version of the following thesis:

Supervised Gamma Process Poisson Factorization

APPROVED BY

SUPERVISING COMMITTEE:

Joydeep Ghosh, Supervisor

Mingyuan Zhou

Supervised Gamma Process Poisson Factorization

by

Dylan Zachary Anderson, B.S.E.E.

Thesis

Presented to the Faculty of the Graduate School

of the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 2015

Supervised Gamma Process Poisson Factorization

by

Dylan Zachary Anderson, M.S.E.

The University of Texas at Austin, 2015

SUPERVISOR: Joydeep Ghosh

This thesis develops the supervised gamma process Poisson factorization (S-GPPF) framework, a novel supervised topic model for joint modeling of count matrices and document labels. S-GPPF is fully generative and nonparametric: document labels and count matrices are modeled under a unified probabilistic framework and the number of latent topics is controlled automatically via a gamma process prior. The framework provides for multi-class classification of documents using a generative max-margin classifier. Several recent data augmentation techniques are leveraged to provide for exact inference using a Gibbs sampling scheme.

The first portion of this thesis reviews supervised topic modeling and several key mathematical devices used in the formulation of S-GPPF. The thesis then introduces the S-GPPF generative model and derives the conditional posterior distributions of the latent variables for posterior inference via Gibbs sampling. The S-GPPF is shown to exhibit state-of-the-art performance for joint topic modeling and document classification on a dataset of conference abstracts, beating out competing supervised topic models. The unique properties of S-GPPF along with its competitive performance make it a novel contribution to supervised topic modeling.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis considers the problem of modeling text and other count data in a fully probabilistic framework. Furthermore, each observation within a dataset is assumed to have a single categorical response variable. The objective is to jointly perform dimensionality reduction and document class label prediction using the dimensionally-reduced space. This thesis describes the Supervised Gamma Process Poisson Factorization (S-GPPF), a fully probabilistic framework to jointly model count observations, latent factors, and observation class labels.

1.1 Problem Statement

Supervised topic modeling seeks to jointly perform dimensionality reduction into a latent topic space and predict document class labels. Some approaches provide supervision by labeling each document with its set of topics [Ramage et al., 2009, Rubin et al., 2012]. Other approaches [Mcauliffe and Blei, 2008, Zhu et al., 2009, Chang and Blei, 2009] assume that supervision is provided for a single *response variable* to be predicted for a given document. The response variable might be real-valued or categorical, and modeled by a normal, Poisson, Bernoulli, multinomial or other distribution (see Chang and Blei [2009] for details). Other works deal with supervision at both the topic and document level [Acharya et al., 2013]. Some examples of documents with response variables are essays with their grades, movie reviews with their numerical ratings, web pages with their number of hits over a certain period of time,

and documents with category labels.

Some supervised topic models [Mcauliffe and Blei, 2008, Chang and Blei, 2009] have found the categorical response variable difficult to model jointly with the latent topics as the resulting inference is intractable. *Maximum Entropy Discriminative LDA* (MedLDA, Zhu et al. [2009]) address this problem by solving two problems jointly: dimensionality reduction and max-margin classification using the features in the dimensionally-reduced space. MedLDA solves the inference problem via variational approximations. Though the update equations are simple, the approximation negatively affects the empirical performance. Additionally, the model has both discriminative and generative components combined in a unified framework, thereby limiting the choice of priors and model flexibility. MedLDA has been extended to Gibbs sampling based inference [Zhu et al., 2013] with a completely generative model. However, it does so at the cost of multi-class response variable modeling. The so-called Gibbs-MedLDA must make use of a one-versus-all (OVA) framework to extend its binary classification to the multi-class setting.

The problem addressed in this thesis is joint modeling of dimensionality reduction and a multi-class response variable. Furthermore, additional properties are imposed upon the model which contribute to its novelty: the model is restricted to be fully generative, non-parametric, and must have exact inference.

1.2 Contributions of this Thesis

This thesis addresses the supervised topic model problem by developing the supervised gamma process Poisson factorization framework. S-GPPF extends the Poisson factorization model put forth by Zhou et al. [2012] to the supervised setting. S-GPPF explicitly models multiclass document class responses, eliminating the need

Model	Nonparametric	Model Type	Inference	Multi-class
S-LDA[i]	✗	Gen[ii]	Variational	✓
MedLDA[iii]	✗	Disc+Gen[iv]	Variational	✓
NP-DSLDA[v]	✓	Disc+Gen	Variational	✓
GibbsMedLDA[vi]	✗	Gen	Gibbs	✗
S-GPPF	✓	Gen	Gibbs	✓

[i] Mcauliffe and Blei [2008] [ii] Generative [iii] Zhu et al. [2009] [iv] Discriminative+Generative [v] Acharya et al. [2013] [vi] Zhu et al. [2013]

Table 1.1: S-GPPF and Other Related Models

for a one-versus-all framework to extend a binary classification to the multiclass setting. Multiclass modeling improves classifier performance, particularly when there is a small amount of labeled data available as jointly modeling classes serves as an inductive bias in prediction of the other class labels, as in multi-task learning. S-GPPF is a fully generative model. This greatly expands the model flexibility in generalizing to new data and in providing interpretation for model predictions. S-GPFF is also a non-parametric model, automatically selecting the number of topics from the data. Finally, S-GPPF provides for exact inference via Gibbs sampling. No other supervised topic model provides a completely Bayesian formulation of a max-margin multiclass classification with an unbounded number of topics and closed form inference via Gibbs sampling. These properties are summarized in Table 1.1.

This thesis is organized as follows. In Chapter 2, relevant literature is reviewed and the mathematical machinery used to implement the framework is discussed. Chapter 3 describes the generative model and provides the Gibbs sampling update equations. It also describes running the model in separate training and testing phases and discusses parameter estimation. Chapter 4 uses S-GPPF to factor real data, and shows its empirical performance to be competitive to the state of the art. Finally, Chapter 5

concludes this thesis with a discussion of the work presented as well further avenues of research.

Chapter 2

Background

This chapter provides an overview of related literature and introduces the mathematical devices used to develop the model and its closed form updates for Gibbs sampling. Note that a proof for each lemma presented can be found in Appendix A or the relevant literature. The chapter is arranged as follows. Section 2.1 introduces the mathematical notations used throughout this thesis as well as some terminology to describe data within the S-GPPF framework. Section 2.2 introduces the broad subject of topic modeling and discusses prominent models for both supervised and unsupervised topic modeling. Section 2.3 provides key results that are necessary for the derivations of the conditional posterior sampling equations for the S-GPPF and introduces uncommon distributions used in the S-GPPF model. Finally, Section 2.4 discusses the formulation of max-margin classifiers and their multiclass extensions in fully generative frameworks.

2.1 Notation

A consistent mathematical notation is adopted throughout this document. Bold, upper [lower] case letters such as $\boldsymbol{A}$ [$\boldsymbol{b}$] denote matrices [vectors]. The element of a matrix $\boldsymbol{A}$ at row i, column j is denoted as a_{ij}. The set of real numbers is denoted as $\mathbb{R}$. The set of numbers $\{x \in \mathbb{R} : x > 0\}$ is denoted $\mathbb{R}_+$. The set of positive integers $\{0, 1, 2, \cdots\}$ is denoted as $\mathbb{Z}_+$. $\boldsymbol{I}_K$ is used to denote a $K \times K$ sized identity matrix. A shorthand for the summation over an axis of the elements of the matrix is represented

as $\sum_k x_{dk} = x_{\cdot k}$ where the index of the axis that is summed over is replaced with a dot. The notation $x|\cdots$ is used to denote the random variable x given all other variables in the model. The script letter $\mathcal{N}$ is used to denote the normal distribution.

Throughout this thesis, data is referred to within the context of text corpora. The smallest unit of discrete data is the *word*, which is an item from a *vocabulary* set indexed by $\{1, \cdots, V\}$. A *document* is a sequence of words. A *corpus* is a set of documents indexed by $\{1, \cdots, D\}$. Using text terms for describing data is useful for providing intuition behind modeling state variables. However, S-GPPF is not limited to textual data; the model is readily applied to supervised factorization of generic count matrices.

2.2 Topic Modeling

Topic modeling can be viewed as unsupervised dimensionality reduction and clustering of documents in a lower dimensional latent space. Topic modeling posits that underlying a corpus is a set of latent topics. From Blei et al. [2003], "each word is generated from a single topic, and different words in a document may be generated from different topics." Each topic is a distribution over words and each document is in turn a distribution over topics. The key insight is that given a document about a particular topic or set of topics, the document should predominantly feature words related to those topics. Informally, topics represent the underlying thematic content of a document.

Underpinning the theory of topic modeling is an assumption of exchangeability of documents and words. In other words, the ordering of words within a document and documents within a corpus is inconsequential [Aldous, 1985]. This is clearly a simplifying assumption: the order of words in a sentence is paramount to understanding.

The exchageability assumption makes modeling and inference much more straightforward and tractable. To introduce some dependency on the ordering of words, larger discrete units of data such as n-grams can be just as easily modeled under topic modeling. From the De Finetti et al. [1990] representation theorem, a collection of exchangeable random variables has a representation as a mixture distribution. In general, the mixture can be infinite, naturally lending to nonparametric Bayesian methods. This mixture representation motivates the most prolific topic model, latent Dirichlet allocation (LDA, Blei et al. [2003]), which is described in detail in the next section.

2.2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation [Blei et al., 2003] treats documents as a mixture of topics, which in turn are defined by a distribution over a set of words. LDA assumes the following generative process for each document from a corpus:

1. Draw the length of a document $N \sim \text{Poisson}\,(\xi)$.
2. Draw the document's distribution over topics $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$.
3. For each word w_n in the document:
   (a) Draw the topic $z_n \sim \text{Multinomial}(\boldsymbol{\theta})$.
   (b) Draw the word w_n from $p(w_n|z_n, \beta)$, a multinomial probability distribution.

The corresponding plate model is shown in Figure 2.1.

In its original formulation, LDA can be viewed as a purely-unsupervised form of dimensionality reduction and clustering of documents in the topic space, although several extensions of LDA have subsequently incorporated some sort of supervision. Two of these extensions are described in sections 2.2.2 and 2.2.3.

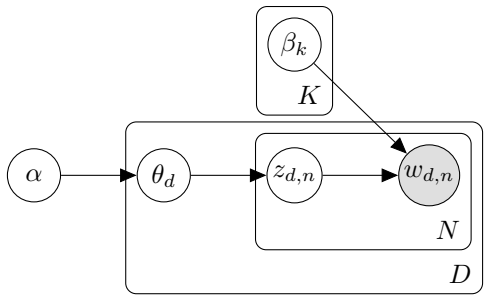


Figure 2.1: Plate model for Latent Dirichlet Allocation

2.2.2 Supervised Latent Dirichlet Allocation

Supervised latent Dirichlet allocation (sLDA, Mcauliffe and Blei [2008]) extends LDA to include a single response variable for each document. Document responses are easily incorporated by using the topic-word distributions (z) to regress onto a response variable. The model also incorporates the regression coefficients in the probabilistic framework. The document responses are linked to their regression coefficients via a generalized linear model (GLM) framework. The leads to the following addition to the generative process of LDA:

- Draw response variable $y|\boldsymbol{Z}, \boldsymbol{\eta}, \delta \sim GLM(\bar{\boldsymbol{z}}, \boldsymbol{\eta}, \delta)$

where $\bar{\boldsymbol{z}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{z}_n$ is the mean of $\boldsymbol{z}$ for each document. The corresponding plate model is shown in Figure 2.2. There are several features lacking from sLDA that are present in S-GPPF. First, sLDA is a parametric model: the number of topics K must be specified *apriori*. In practice, the number of topics must be determined using cross-validation as too many topics will lead to several junk topics that provide poor predictive information and do not represent thematic structure. The other major drawback of sLDA is that inference is via an expectation-maximization (EM) approximation to the maximum likelihood. EM provides an inexact estimate of model parameters, and is subject to local maxima.

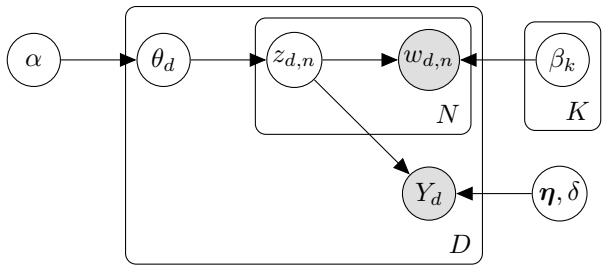


Figure 2.2: Plate model for supervised latent Dirichlet allocation

2.2.3 Maximum Entropy Discriminant Latent Dirichlet Allocation

Maximum entropy discriminant latent Dirichlet allocation (MedLDA, Zhu et al. [2009]) differs from sLDA in that it optimizes a joint objective function that represents a combination of max-margin learning and a Bayesian topic model. Topics learned in MedLDA not only cluster the data but are learned in an optimal max-margin sense: the latent topics are well suited for use as predictive features. MedLDA provides inference via variational methods, which are detrimental to predictive performance. Additionally, the model has both discriminative and generative components combined under a single unified framework, limiting model flexibility. MedLDA is also a parametric model, requiring the number of topics be specified *apriori*. MedLDA has empirically shown very good performance, and is generally considered state of the art for class prediction in supervised topic modeling.

Further development of MedLDA has led to the so-called Gibbs MedLDA [Zhu et al., 2013]. Gibbs MedLDA employs a Gibbs sampling based inference framework on a completely generative model instead of using variational approximations by using various ideas from Gibbs-based classifiers. However, it does this at the cost of native multiclass classification, requiring a one-versus-all framework to extend binary predictions to the multiclass setting.

2.2.4 Poisson Count Matrix Factorization

An alternative to an LDA-based topic model is to factorize the corpus (which is represented by a document $\times$ word count matrix) using a latent variable model called Poisson factor analysis (PFA, Zhou et al. [2012]). The matrix $\boldsymbol{X} \in \mathbb{Z}_+^{D \times V}$ has a Poisson likelihood over the observed counts

$$\boldsymbol{X} \sim \text{Poisson}\left(\boldsymbol{\Theta}\boldsymbol{\Phi}\right), \tag{2.1}$$

where $\boldsymbol{\Phi} \in \mathbb{R}_+^{K \times V}$ is the factor loading matrix or dictionary, $\boldsymbol{\Theta} \in \mathbb{R}_+^{D \times K}$ is the factor score matrix.

PFA offers two major advantages over classical matrix factorization models that rely on Gaussian observation models [Mnih and Salakhutdinov, 2007]. Gaussian-based factorizations require intricate strategies to mitigate the effects of zeros in settings where zeros represent unobserved entries [Hu et al., 2008]. In contrast, PFA models zeros as the result of finite resources [Gopalan et al., 2013]. This can be easily seen by rewriting the factorization as a two level model: first draw a budget given by $\boldsymbol{x}_{d.}$, which is Poisson-distributed according to the likelihood, then allocate the budget onto individual columns following a multinomial distribution (see lemma 2.3.4 for details). This allows PFA to explain zeros as partially due to a lack of resources. The other advantage of PFA is that it need only iterate over non-zero elements. In a latent variable model each element is represented by a summation of latent elements:

$$x_{dw} = \sum_k x_{dwk}$$

Under a Gaussian likelihood and with $x_{dw} = 0$, each of the latent values must also be sampled because x_{dwk} can be both positive and negative. On the other hand, with a Poisson likelihood as in PFA, $x_{dw} = 0 \implies x_{dwk} = 0 \; \forall k$ since $x_{dwk} \geq 0$.

PFA represents a very general framework for factorization of count matrices. A wide variety of algorithms can all be posed as PFA by placing different prior distributions on $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$. For example, non-negative matrix factorization [Lee and Seung, 2001, Cemgil, 2009], with the objective to minimize the Kullback-Leibler divergence between $\boldsymbol{X}$ and its factorization $\boldsymbol{\Phi}\boldsymbol{\Theta}$ is PFA solved with maximum likelihood estimation. Imposing Dirichlet priors on both the columns of $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ makes LDA equivalent to PFA in terms of both block Gibbs sampling and variational inference. Placing gamma priors on $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ leads to the gamma-Poisson model [Canny, 2004, Titsias, 2008, Gopalan et al., 2014]. This flexibility makes PFA easy to extend to non-parametric factorizations by careful prior selection. A family of negative binomial (NB) processes, such as the beta-NB [Zhou et al., 2012, Broderick et al., 2015] and gamma-NB processes [Zhou et al., 2012, Zhou and Carin, 2015], impose different gamma priors on $\boldsymbol{\Theta}$. Marginalizing over $\boldsymbol{\Theta}$ explains the latent counts using a gamma-Poisson construction of the negative binomial distribution. For example, the beta-NB process imposes $\theta_{tk} \sim \text{Gamma}\left(r_t, p_k/(1-p_k)\right)$, where $\{p_k\}_{1,\infty}$ are the weights of the countably infinite atoms of the beta process [Hjort, 1990], and the gamma-NB process imposes $\theta_{tk} \sim \text{Gamma}\left(r_k, p_t/(1-p_t)\right)$, where $\{r_k\}_{1,\infty}$ are the weights of the countably infinite atoms of the gamma process. Both the beta- and gamma- NB process PFAs allow the number of latent factors, K, to grow without limits [Hjort, 1990].

As its name implies, S-GPPF uses a gamma process prior to control the number of latent factors. The unsupervised gamma process Poisson factorization [Zhou et al., 2012] has been extended to other problem settings, including dynamic count matrices where columns of the observed count matrix represent observations over time [Acharya et al., 2015] and network modeling where user network information is observed in addition to the count matrix [Zhou, 2015]. A plate model for the gamma process

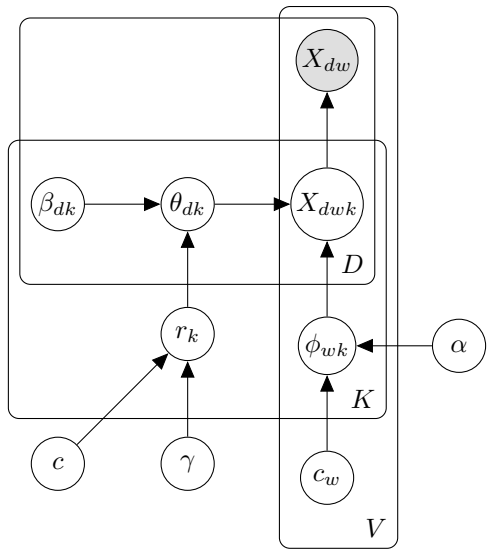


Figure 2.3: Plate model for Gamma Process Poisson Factorization

Poisson factorization is shown in Figure 2.3.

2.3 Useful Distributions and Results

This section describes several distributions and processes used in the modeling framework of S-GPPF. Several properties are presented in the form of lemmas, the proofs of which can be found in Appendix A.

2.3.1 Gamma Distribution

Throughout this thesis, a random variable $x \sim \text{Gamma}\,(a, b)$ has probability density function $p(x) = \frac{1}{\Gamma(a)b^a} x^{a-1} \exp\left(-\frac{x}{b}\right)$. This is the shape-scale parameterization of the Gamma distribution with shape $a > 0$ and scale $b > 0$.

2.3.2 Gamma Process

The gamma process [Ferguson, 1973, Wolpert et al., 2011] $G \sim \text{GaP}(c, G_0)$ is a stochastic process whose realizations are random measures: it is a probability

distribution over measures. This is called a completely random measure [Kingman, 1967, 1992]. Realizations are drawn from the product space $\mathbb{R}_+ \times \Omega$. The gamma process is parameterized with concentration parameter c and a finite and continuous base measure G_0 over a complete separable metric space Ω, such that $G(A_i) \sim \text{Gamma}(G_0(A_i), 1/c)$ are independent gamma random variables for disjoint partition $\{A_i\}_i$ of Ω. The Lévy measure of the gamma process can be expressed as $\nu(drd\omega) = r^{-1}e^{-cr}drG_0(d\omega)$. Since the Poisson intensity $\nu^+ = \nu(\mathbb{R}_+ \times \Omega) = \infty$ and $\int_{\mathbb{R}_+\times\Omega} r\nu(drd\omega)$ is finite, following Wolpert et al. [2011], a draw from the gamma process consists of countably infinite atoms, which can be expressed as:

$$G = \sum_{k=1}^{\infty} r_k \delta_{\omega_k},\ (r_k, \omega_k) \overset{iid}{\sim} \pi(drd\omega), \pi(drd\omega)\nu^+ \equiv \nu(drd\omega). \tag{2.2}$$

Imposing a gamma process prior on topics, assigns weights corresponding to the atoms of the process to the topics. This leads to the number of active topics being discovered automatically, rather than specified *apriori*.

2.3.3 Conjugate Prior Distributions

For computational convenience, many of the modeling assumptions are designed using conjugate prior distributions. Some results are presented here in the form of lemmas for ease of deriving the conditional posterior equations in Section 3.2.

Lemma 2.3.1. If $\lambda \sim \text{Gamma}(r, 1/c), x_i \sim \text{Poisson}(m_i\lambda)$, then $\lambda|\{x_i\} \sim \text{Gamma}\left(r + \sum_i x_i, 1/(c + \sum_i m_i)\right)$.

Lemma 2.3.2. If $r_i \sim \text{Gamma}(a_i, 1/b)\ \forall i \in \{1, 2, \cdots, K\}, b \sim \text{Gamma}(c, 1/d)$, then $b|\{r_i\} \sim \text{Gamma}\left(\sum_{i=1}^{K} a_i + c, 1/(\sum_{i=1}^{K} r_i + d)\right)$.

Lemma 2.3.3. If $z_i \sim \mathcal{N}(\mu_i, \sigma^{-1})\ \forall i \in \{1, 2, \cdots, K\}, \sigma \sim \text{Gamma}(a, 1/b)$, then $\sigma|\{z_i\} \sim \text{Gamma}\left(a + K/2, 1/(b + \sum_{i=1}^{K}(z_i - \mu_i)^2/2\right)$.

Lemma 2.3.4. Let $x_k \sim \text{Pois}(\zeta_k)\ \forall k,\ X = \sum_{k=1}^{K} x_k,\ \zeta = \sum_{k=1}^{K} \zeta_k$. If $(y_1, \cdots, y_K) \sim \text{mult}(X; \zeta_1/\zeta, \cdots, \zeta_K/\zeta)$, then the following holds:

$$p(x_1, \cdots, x_K) = p(y_1, \cdots, y_K; X).$$

2.3.4 Negative Binomial Distribution

The negative binomial (NB) distribution $m \sim \text{NB}(r, p)$ has probability mass function $\Pr(M = m) = \frac{\Gamma(m+r)}{m!\Gamma(r)} p^m (1-p)^r$ for $m \in \mathbb{Z}$. NB variables can be constructed via augmentation into a gamma-Poisson construction as $m \sim \text{Pois}(\lambda),\ \lambda \sim \text{Gamma}(r, p/(1-p))$, where the gamma distribution is parameterized by its shape r and scale $p/(1-p)$. This construction can be extended via the following lemma

Lemma 2.3.5. If $\lambda \sim \text{Gamma}(r, 1/c), x_i \sim \text{Poisson}(m_i \lambda)$, then $x = \sum_i x_i \sim \text{NB}(r, p)$, where $p = \frac{\sum_i m_i}{c + \sum_i m_i}$.

The Negative Binomial can also be augmented under a compound Poisson representation [Zhou et al., 2012, Zhou and Carin, 2012] as $m = \sum_{t=1}^{l} u_t, u_t \overset{iid}{\sim} \text{Log}(p), l \sim \text{Pois}(-r\ln(1-p))$, where $u \sim \text{Log}(p)$ is the logarithmic distribution [Johnson et al., 2005]. The two different constructions are shown graphically in Figure 2.4, and they lead to the following lemma:

Lemma 2.3.6. [Zhou et al., 2012] If $m \sim NB(r, p)$ is represented under its compound Poisson representation, then the conditional posterior of l given m and r has PMF:

$$Pr(l = j | m, r) = \frac{\Gamma(r)}{\Gamma(m+r)} |s(m, j)| r^j,\ j = 0, 1, \cdots, m,$$

where $|s(m, j)|$ are unsigned Stirling numbers of the first kind. We denote this conditional posterior as $l | m, r \sim CRT(m, r)$, a Chinese restaurant table (CRT) count random variable, which can be generated via $l = \sum_{n=1}^{m} z_n, z_n \sim \text{Bernoulli}(r/(n-1+r))$.

(a) Gamma-Poisson Construction (b) Compound Poisson Construction

Figure 2.4: Alternative constructions of the Negative Binomial distribution

This lemma allows leads to the next lemma, which provides closed form sampling of the gamma shape parameter via CRT data augmentation in the gamma-gamma-Poisson framework.

Lemma 2.3.7. If $r_1 \sim \text{Gamma}(a, 1/b), r_2 \sim \text{Gamma}(r_1, 1/d)$, $x_i \sim \text{Poisson}(m_i r_2)\ \forall i$, then $r_1|\{x_i\} \sim \text{Gamma}(a + \ell, 1/(b - \log(1-p)))$ where $\ell \sim \text{CRT}(\sum_i x_i, r_1), p = \sum_i m_i/(d + \sum_i m_i)\ \forall i$.

2.3.5 Pólya-Gamma Distribution

A random variable X has a Pólya-Gamma distribution [Polson et al., 2011] with parameters $b > 0$ and $c \in \mathbb{R}$, denoted $X \sim PG(b, c)$, if $X \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/4\pi^2}$, where $g_k \sim \text{Gamma}(b, 1)$'s are independent Gamma random variables, and where $\overset{D}{=}$ indicates equality in distribution. This leads to the following lemma:

Lemma 2.3.8. [Polson et al., 2013] If $\omega \sim \text{PG}(b, 0)$, then

$$\frac{\exp(\psi)^a}{(1 + \exp(\psi))^b} = 2^{-b}\exp((a - b/2)\psi) \int_0^{\infty} \exp(-\omega\psi^2/2)p(\omega)d\omega \tag{2.3}$$

$$\omega|\psi \sim \text{PG}(b, \psi) \tag{2.4}$$

Pólya-Gamma random variables are used for augmentation in sampling the state variables that link the count matrix factorization to the document class labels in the S-GPPF model.

2.4 Hinge Loss and Location-mixture of Normals

The support vector machine (SVM) seeks to find a classification function $f(x)$ by solving the following regularized learning problem:

$$\arg\min_{f(\boldsymbol{x})} \ \gamma \sum_{n=1}^{N} (1 - y_n f(\boldsymbol{x}_n))_+ + R(f(\boldsymbol{x})), \tag{2.5}$$

where $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ is the set of N observation tuples, $\boldsymbol{x}_n \in \mathbb{R}$ is a feature vector and $y_n \in \{-1, 1\}$ is the corresponding label for observation n. $(1 - y_n f(\boldsymbol{x}_n))_+$ is the hinge loss, $R(f(\boldsymbol{x}))$ is a regularization term that controls the complexity of $f(\boldsymbol{x})$, and γ is a tuning parameter controlling the trade-off between error penalization and the complexity of the classification function. The decision boundary is defined as $\{\boldsymbol{x} : f(\boldsymbol{x}) = 0\}$ and $\text{sign}(f(\boldsymbol{x}))$ is the decision rule, classifying x as either -1 or 1.

Polson et al. [2011] showed that for the linear classifier $f(\boldsymbol{x}) = \langle \boldsymbol{\eta}, \boldsymbol{x} \rangle$, minimizing Eq. (2.5) is equivalent to estimating the mode of the pseudo-posterior of $\boldsymbol{\eta}$:

$$p(\boldsymbol{\eta}|\boldsymbol{X}, \boldsymbol{Y}, \gamma) \propto \prod_{n=1}^{N} L(y_n|\boldsymbol{x}_n, \boldsymbol{\eta}, \gamma) p(\boldsymbol{\eta}), \tag{2.6}$$

where $\boldsymbol{Y} = (y_1, \cdots, y_N)$, $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$, $L(y_n|\boldsymbol{x}_n, \boldsymbol{\eta}, \gamma)$ is the pseudo-likelihood function, and $p(\boldsymbol{\eta})$ is the prior distribution for the vector of coefficients $\boldsymbol{\eta}$. This can be seen by taking the exponential of the negative of Eq. (2.5), where the likelihood is given by the hinge loss and the regularization term is given by the prior on $\boldsymbol{\eta}$. Placing a normal distribution prior on $\boldsymbol{\eta}$ is akin to L2 regularization. Estimating the mode of Eq. (2.6) is equivalent to finding the corresponding minimum of the loss in Eq. (2.5) since they are related under a monotonic transform. Lemma 2.4.1 enables one to solve the optimization problem in Eq. (2.6) using closed form Gibbs sampling updates via data augmentation.

Lemma 2.4.1. [Polson et al., 2011] If $u \sim \mathcal{N}(\mu, \sigma^2)$, one can show that:

$$\exp\left(-2u_+\right) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} \exp\left(-\frac{(\lambda+u)^2}{2\lambda}\right) d\lambda \tag{2.7}$$

where

$$p(\lambda^{-1}|u) \sim \mathcal{IG}(\|u\|^{-1}, 1) \tag{2.8}$$

$$p(u|\lambda) \sim \mathcal{N}(\mu', \sigma'^2) \tag{2.9}$$

where $\mu' = \frac{\lambda(\mu-\sigma^2)}{(\lambda+\sigma^2)}$, and $\sigma'^2 = \frac{\lambda\sigma^2}{(\lambda+\sigma^2)}$ and $\mathcal{IG}$ denotes the inverse-gaussian distribution.

2.4.1 Formulation of Multiclass SVM

SVM has also been extended to solve multiclass problems [Weston and Watkins, 1998, Crammer and Singer, 2002, Lee et al., 2004]. Tewari and Bartlett [2007] describes the theoretical consistency of different formulations of multiclass SVMs. S-GPPF uses the formulation of Lee et al. [2004] where the discriminant function for multiclass SVM is defined as $f_y(\boldsymbol{x}) = \langle \boldsymbol{\eta}_y, \boldsymbol{x} \rangle$, $\boldsymbol{\eta}_y$ being the weight vector corresponding to the class label $y \in \{1, \cdots, M\}$. The regularized risk minimization problem is given as follows:

$$\min_{\boldsymbol{\eta}} \frac{\lambda}{2}\|\boldsymbol{\eta}\|^2 + \sum_{d=1}^{D} \sum_{y \neq y_d} \left(\langle \boldsymbol{\eta}_y, \boldsymbol{x} \rangle + \frac{1}{(M-1)} \right)_+ \tag{2.10}$$

such that $\sum_{y=1}^{M} \langle \boldsymbol{\eta}_y, \boldsymbol{x} \rangle = 0 \ \forall d$. This formulation of multiclass SVM is amenable to tractable inference when the features (*i.e.* $\boldsymbol{x}$'s) are latent rather than directly observed as in a hierarchical model: for example, if they represent the assignment of documents to topics in a topic model framework.

Chapter 3

Supervised Gamma Process Poisson Factorization

This chapter describes the supervised gamma process Poisson factorization, which is the main contribution of this thesis. As previously discussed, the novelty of S-GPPF stems from several properties it simultaneously enjoys. First, the model is non-parametric; the number of topics present in a corpus are determined automatically through the use of a gamma process prior on topic weights. Second, the model provides for multiclass classification directly through the use of a multiclass max-margin formulation. Third, the model is fully generative, capturing the relationships between documents, words, topics, and class labels under a completely probabilistic framework. Finally, the model provides for closed form, exact inference through the use of data augmentation and Gibbs sampling. This chapter is structured as follows. Section 3.1 describes the generative process and the latent variables of the model. Section 3.2 details the inference procedure using Gibbs sampling. Section 3.3 describes training the S-GPPF model using documents with known class labels and Section 3.4 details using a trained S-GPPF model to predict unknown class labels.

3.1 Generative Process

Consider a corpus of D documents with a vocabulary of size V. The corpus is partitioned into two sets of documents: those with observed class labels (denoted as the training set) and those without observed class labels (denoted as the testing set). Each document label takes one of M possible values, and the notation y_d denotes the

label for the dth document. The document $\times$ word count matrix, $\boldsymbol{X}$, is decomposed as a product of two latent matrices ($\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$) and a topic strength vector ($\boldsymbol{r}$) under a Poisson likelihood. The generative process is described below, and the corresponding plate model describing the family of distributions of which S-GPPF is a member is shown in Figure 3.1.

For the dth document, sample $\theta_{dk} \sim \text{Gamma}\left(\tau_d, 1/\exp\left(-\beta_{dk}\right)\right)\ \forall k$, where $\tau_d \sim \text{Gamma}\left(c_0, 1/d_0\right)$, $\beta_{dk} \sim \mathcal{N}\left(0, \alpha_{dk}^{-1}\right)$, and $\alpha_{dk} \sim \text{Gamma}\left(g_0, 1/h_0\right)$. θ_{dk} represents the affinity of the dth document to the kth topic. Ideally, θ_{dk} would be used as feature to predict document class labels. However it is not tractable to do so when using a multi-class max-margin classifier. Instead, β_{dk} is sampled as the per-document features for classification. From the properties of Gamma distribution, we have $\mathbb{E}(\theta_{dk}) = \mathbb{E}(\tau_d \exp\left(\beta_{dk}\right))$ and hence any change in θ_{dk} gets reflected in β_{dk} monotonically under a logarithmic transformation.

For the kth topic, sample $\boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\xi})$ where $\boldsymbol{\phi}_k = \left(\phi_{wk}\right)_{w=1}^{V}$ and $\boldsymbol{\xi}$ is a $V-$dimensional parameter. Each ϕ_{wk} maps the affinity of word w onto topic k. A Dirichlet prior is used instead of a hierarchical gamma structure to improve model identifiability.

The strength of the kth topic is sampled as $r_k \sim \text{Gamma}\left(\gamma_0/K, 1/\exp\left(-\zeta_k\right)\right)$, where $\gamma_0 \sim \text{Gamma}\left(a_0, 1/b_0\right)$, $\zeta_k \sim \mathcal{N}\left(0, \nu_k^{-1}\right)$, and $\nu_k \sim \text{Gamma}\left(u_0, 1/v_0\right)$. This places a gamma process prior on the topic strengths, approximating an infinite number of topics with a finite number K; only a small number of topics will be appreciably larger than zero due to the stick-breaking construction of the gamma process. Ideally, the r_k's would be used directly as part of the classification weights but as in the case of θ_{dk} this is intractable under a max-margin classifier. Instead, the ζ_k's are used, which are monotonically proportional to the r_k's under expectation. The ζ_k's represent the strength of the topics for the linear classifier.

The count corresponding to the dth document and the wth word is sampled as $x_{dw} \sim \text{Poisson}\left(\sum_k \theta_{dk}\phi_{wk}r_k\right)$. Alternatively, due to the property of Poisson distribution, one may write $x_{dw} = \sum_k x_{dwk}$, where $x_{dwk} \sim \text{Poisson}\left(\theta_{dk}\phi_{wk}r_k\right)\ \forall k$. Each latent count variable represents the contributions of the kth topic onto the $(d, w)^{\text{th}}$ entry in the document $\times$ word matrix.

The class label $y_d \in \{1, 2, \cdots, M\}$ for the dth document is calculated using a multiclass max-margin classifier. From the work of Lee et al. [2004], a pseudo-likelihood of the class label y_d is defined as:

$$q(y_d|\cdots) = \exp\left(-\sum_{y \neq y_d}\left(z_{yd} + \frac{1}{(M-1)}\right)_+\right) \tag{3.1}$$

where $z_{yd} \sim \mathcal{N}\left(\sum_k \eta_{yk}\beta_{dk}\zeta_k, \sigma^{-1}\right)$, $\sigma \sim \text{Gamma}\left(s_0, 1/t_0\right)$ and $\sum_{y=1}^{M} z_{yd} = 0$. $\boldsymbol{\eta}_y = (\eta_{yk})_{k=1}^{K}$ is the set of weights corresponding to the yth class and is generated as $\eta_{yk} \sim \mathcal{N}\left(0, \epsilon_k^{-1}\right)$, $\epsilon_k \sim \text{Gamma}\left(e_0, 1/f_0\right)$. η_{yk} represent the classifier weights for the kth topic to the yth class label. This is the same formulation of the constrained multiclass SVM as described in Section 2.4.1 with the auxiliary variables $\{z_{yd}\}$ introduced to provide closed form inference of the auxiliary variables $\{\beta_{dK}\}$ and $\{\zeta_k\}$ via a data augmentation scheme.

3.2 Inference

Given the S-GPPF sampling model and a corpus, the main problem is posterior inference: finding the distribution of the model parameters given the observed data. Since S-GPPF seeks to also predict test set class labels, the unobserved class labels are also considered as unknown parameters. One method for estimating the posterior is through the use of Markov chain Monte Carlo Methods (MCMC, Hoff [2009],

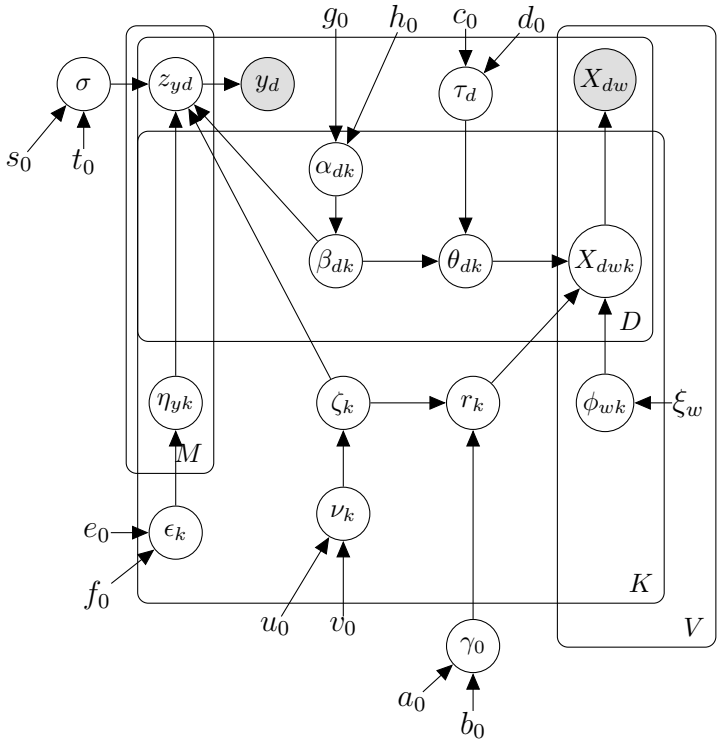


Figure 3.1: Plate Diagram of Supervised GPPF

Gamerman and Lopes [2006]). MCMC methods describe Markov chains that are easy to sample from and whose invariant distributions are the target posterior. Then the samples from the Markov chain are also distributed according to the posterior, and the distribution can be estimated via Monte Carlo integration. For a Markov chain to converge to its invariant distribution regardless of its initialization, it must be irreducible, invariant, and aperiodic [Cosma and Evers, 2010].

Hierarchical Bayesian models such as S-GPPF naturally lend themselves to Gibbs sampling [Cosma and Evers, 2010, Hoff, 2009, Gamerman and Lopes, 2006]. In Gibbs sampling, the conditional posterior distributions for each parameter are sampled progressively one by one (there are other Gibbs sampling schemes, such as random scan but systematic scanning is considered here for simplicity). Parameters are drawn using the most recent samples of all of the other parameters. It has been shown that Gibbs sampling schemes are invariant [Cosma and Evers, 2010], so to show that a Gibbs sampling scheme is valid for posterior inference it must be shown to be aperiodic and irreducible.

The proposed Gibbs sampling scheme for S-GPPF is easily shown to have both of these properties. With the exception of the latent count variables (x_{dwk}), all of the parameters are drawn either from gamma or normal distributions (excluding the augmented variables, since they are marginalized out). As such, each variable has positive probability mass on its entire state-space (either $\mathbb{R}$, $\mathbb{R}_+$, or $\mathbb{R}^K$). This means that regardless of the current state of the chain, there is some positive (although it may be very small) probability to move to any other valid state. Therefore, the Gibbs sampling scheme forms an irreducible Markov chain. Furthermore, since the entire state space has positive probability mass, there is positive mass on staying in the same state. This implies that the chain is also aperiodic and therefore the Gibbs

sampling scheme is valid for posterior inference.

Enumerated below are the conditional posterior distributions for each of the model parameters. Full derivations of the sampling equations are provided in Appendix B.

Sampling of $(x_{dwk})_{k=1}^{K}$

$$(x_{dwk})_{k=1}^{K} | \cdots \sim \text{mult}\left(\left(\frac{r_k \theta_{dk} \phi_{wk}}{\sum_{k=1}^{K} r_k \theta_{dk} \phi_{wk}}\right)_{k=1}^{K}; x_{dw}\right) \tag{3.2}$$

Sampling of θ_{dk}

$$\theta_{dk} | \cdots \sim \text{Gamma}\left(\tau_d + x_{d.k}, 1/\left(\exp\left(-\beta_{dk}\right) + r_k\right)\right) \tag{3.3}$$

Sampling of τ_d

$$l_{dk} | \cdots \sim \text{CRT}\left(x_{d.k}, \tau_d\right) \tag{3.4}$$

$$\tau_d | \cdots \sim \text{Gamma}\left(c_0 + \sum_k l_{dk}, 1/(d_0 - \sum_k \log(1 - p_{dk}))\right), \tag{3.5}$$

where $p_{dk} = \frac{r_k}{\exp(-\beta_{dk}) + r_k}$.

Sampling of $\boldsymbol{\phi}_k$

$$\boldsymbol{\phi}_k | \cdots \sim \text{Dir}\left(\xi_1 + x_{.1k}, \cdots, \xi_V + x_{.Vk}\right) \tag{3.6}$$

Sampling of r_k

$$r_k|\cdots \sim \text{Gamma}\left(\gamma_0/K + x_{..k}, 1/\left(\exp\left(-\zeta_k\right) + \theta_{.k}\right)\right) \tag{3.7}$$

Sampling of γ_0

$$l_k|\cdots \sim \text{CRT}\left(x_{..k}, \gamma_0/K\right) \tag{3.8}$$

$$\gamma_0|\cdots \sim \text{Gamma}\left(a_0 + \sum_k l_k, 1/(b_0 - \frac{1}{K}\sum_k \log(1-p_k))\right), \tag{3.9}$$

where $p_k = \frac{\theta_{.k}}{\exp(-\zeta_k)+\theta_{.k}}$.

Sampling of σ

$$\sigma|\cdots \sim \text{Gamma}\left(s_0 + \frac{MD}{2}, 1/t_0'\right), \tag{3.10}$$

where $t_0' = \left(t_0 + \sum_{y,d} \frac{(z_{yd} - \sum_k \eta_{yk}\beta_{dk}\zeta_k)^2}{2}\right)$.

Sampling of ϵ_k

$$\epsilon_k|\cdots \sim \text{Gamma}\left(e_0 + \frac{M}{2}, 1/\left(\sum_y \frac{\eta_{yk}^2}{2} + f_0\right)\right) \tag{3.11}$$

Sampling of ν_k

$$\nu_k|\cdots \sim \text{Gamma}\left(u_0 + \frac{1}{2}, 1/\left(\frac{\zeta_k^2}{2} + v_0\right)\right) \tag{3.12}$$

Sampling of α_{dk}

$$\alpha_{dk}|\cdots \sim \text{Gamma}\left(g_0 + \frac{1}{2}, 1/\left(\frac{\beta_{dk}^2}{2} + h_0\right)\right) \tag{3.13}$$

Sampling of z_{yd}

Since $\sum_{y=1}^{M} z_{yd} = 0$, we need only sample $\{z_{yd}\}_{y \neq y_d}$ for each document d and assign $z_{y_d d} = -\sum_{y \neq y_d} z_{yd}$. Then for $y \neq y_d$:

$$\gamma_{yd}|\cdots \sim \mathcal{IG}\left(\left|\frac{z_{yd} + \frac{1}{M-1}}{2}\right|^{-1}, 1\right) \tag{3.14}$$

$$z_{yd}|\cdots \sim \mathcal{N}\left(\mu', \sigma'^2\right) \tag{3.15}$$

where $\sigma'^2 = \frac{\gamma_{yd}}{\gamma_{yd}\sigma + 1/4}$ and $\mu' = \sigma'^2\left(\sigma \sum_k \beta_{dk}\eta_{yk}\zeta_k - \frac{1}{4\gamma_{yd}(M-1)} - \frac{1}{2}\right)$, and $\mathcal{IG}$ is used to denote the inverse-Gaussian distribution.

Sampling of $\boldsymbol{\eta}_y$

$$\boldsymbol{\eta}_y|\cdots \sim \mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right), \tag{3.16}$$

where $\boldsymbol{\Sigma}_y^{-1} = \left[\boldsymbol{\alpha}_1 \boldsymbol{I}_K + \sigma\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\sum_d \left(\boldsymbol{\beta}_d'\boldsymbol{\beta}_d\right)\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\right]$ and
$\boldsymbol{\mu}_y = \boldsymbol{\Sigma}_y\left(\sigma\boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\sum_d z_{yd}\boldsymbol{\beta}_d'\right)$

Sampling of $\boldsymbol{\beta}_d$

$$\omega_{dk}|\cdots \sim \text{PG}(x_{d.k} + \tau_d, \beta_{dk} + \log(r_k)) \tag{3.17}$$

$$\boldsymbol{\beta}_d|\cdots \sim \mathcal{N}\left(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d\right), \tag{3.18}$$

where $\boldsymbol{\Sigma}_d^{-1} = \left[(\boldsymbol{\alpha}_d \boldsymbol{I}_K) + (\boldsymbol{\omega}_d \boldsymbol{I}_K) + \sigma(\boldsymbol{\zeta}\boldsymbol{I}_K) \sum_y \left[\boldsymbol{\eta}_y' \boldsymbol{\eta}_y\right] (\boldsymbol{\zeta}\boldsymbol{I}_K)\right]$,
$\boldsymbol{\mu}_d = \left[\sigma \sum_y \left[z_{yd}\boldsymbol{\eta}_y\right] (\boldsymbol{\zeta}\boldsymbol{I}_K) - \boldsymbol{\omega}_d(\log(\boldsymbol{r})\boldsymbol{I}_K) + \frac{\boldsymbol{\nu}_d}{2}\right] \boldsymbol{\Sigma}_d$, and $\boldsymbol{\nu}_d = \{x_{d.k} - \tau_d\}_{k=1}^K$.

Sampling of $\boldsymbol{\zeta}$

$$\omega_k | \cdots \sim \mathrm{PG}(x_{..k} + \gamma_0/K, \log(\theta_{.k}) + \zeta_k) \tag{3.19}$$

$$\boldsymbol{\zeta} | \cdots \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3.20}$$

where $\boldsymbol{\Sigma}^{-1} = \left[(\boldsymbol{\alpha}_2 \boldsymbol{I}_K) + (\boldsymbol{\omega}\boldsymbol{I}_K) + \sigma \sum_{y,d} (\boldsymbol{\beta}_d \boldsymbol{I}_K) \boldsymbol{\eta}_y' \boldsymbol{\eta}_y (\boldsymbol{\beta}_d \boldsymbol{I}_K)\right]$,
$\boldsymbol{\mu} = \left[\sigma \sum_{y,d} z_{yd} \boldsymbol{\eta}_y (\boldsymbol{\beta}_d \boldsymbol{I}_K) - \log(\boldsymbol{\theta})(\boldsymbol{\omega}\boldsymbol{I}_K) + \boldsymbol{\lambda}/2\right] \boldsymbol{\Sigma}$, $\boldsymbol{\lambda} = \{x_{..k} - \gamma_0/K\}_{k=1}^K$, and
$\boldsymbol{\theta} = \{\theta_{.k}\}_{k=1}^K$.

3.3 Training Phase

In the training phase, the document class labels are observed and all model variables are sampled in a Gibbs sampling scheme. The implementation used in this thesis forms parameter estimates after the training phase by computing the sample average as a Monte Carlo approximation of the posterior mean. This provides a point estimate for parameters that can be easily used for sampling in the test phase.

It is worth noting that since topics are related to both observed counts and class labels only through inner products, the resulting posterior distribution is not identifiable: the exact index, k, for a specific topic will vary between runs of the Markov chain. This has important ramifications for propagating multiple samples instead of single point estimates of parameters for the test phase. It is certainly possible to run separate chains for each sample, but care must be taken when concatenating the

samples: results should only be combined at the level of the observed data, i.e. it is invalid to combine the samples of θ_{dk} from multiple chains as draws from the same posterior but it is valid to combine $\sum_k \theta_{dk} r_k \phi_{wk}$ across chains as all being drawn from the same posterior.

The experiments presented in Chapter 4 use only point estimates of the posterior mean from the training phase. This makes the implementation simpler and eases the computational burden since only a single chain is run in test phase, but it is a simplification. In fact, using a point estimate gives up some of the advantage gained by using a fully generative framework with exact inference as an estimate of the full posterior distribution is no longer available. Incorporating better estimates between training and test phases provides an opportunity for improvement upon the work presented in this thesis.

3.4 Test Phase

The goal in the training phase it to estimate the model parameters that are not specific to any document. The test phase then uses these estimates to sample the posterior of the document specific parameters and uses the result to predict the unknown class labels. In test phase, the posterior mean estimates from training for all of the parameters that not document specific are held fixed. Gibbs sampling is then run on only those variables that are document specific. This means that for each test document d, only $\{\theta_{dk}\}_{k=1}^{K}$,$\{x_{dwk}\}_{k=1,w=1}^{K,V}$, $\boldsymbol{\beta}_d$, and $\{\alpha_{dk}\}_{k=1}^{K}$ are sampled. Since the class labels are unknown, $\boldsymbol{\beta}_d$ is sampled without the influence of z_{yd}, σ, and y_d. To estimate the class labels, z_{yd} is estimated by its mean, given by $\sum_k \beta_{dk}\eta_{yk}\zeta_k$. The class labels are then estimated by maximizing the likelihood given in Eq. (3.1).

Chapter 4

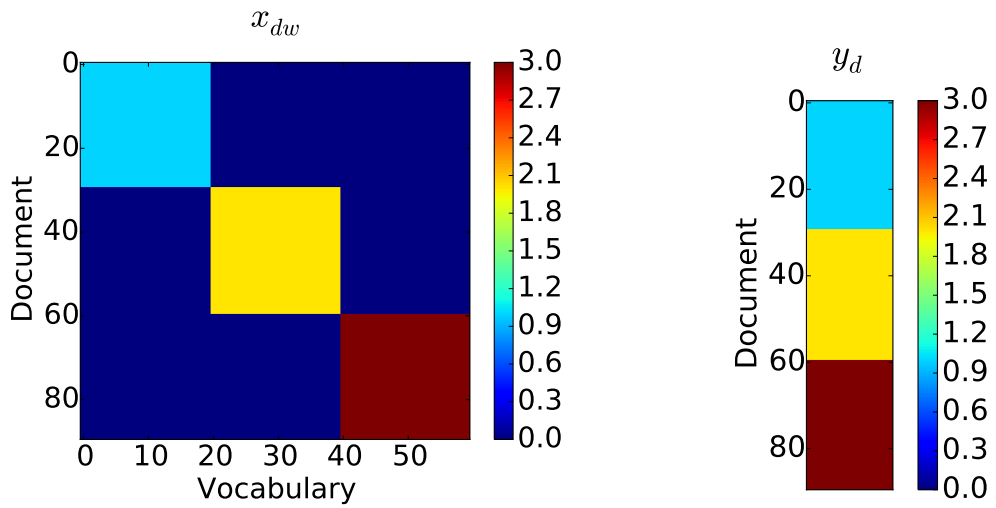
Experiment Analysis

This section describes using the S-GPPF model on two different datasets: a synthetic dataset used to explore and visualize the latent variables of the model and a corpus of abstracts from several conferences to compare the performance of S-GPPF against competing supervised topic models. S-GPPF is shown to have state-of-the-art levels of performance for classification of document labels.

4.1 Factorization of Synthetic Data

This section considers a synthetic corpus to gain intuition into the model parameters. The data has 90 documents and 60 words arranged in a block diagonal structure. The upper third of the block diagonal are assigned a count value of one, the middle third are assigned a count value of two, and the final third block diagonal are assigned a value of three. All other values in the document $\times$ word matrix are set to zero. Accordingly, the first third of the documents are assigned a class label of one, the second third are assigned a class label of two, and the final third are assigned a class label of three. The document $\times$ word matrix and class labels are shown in Fig 4.1. Sampling is run for 1,000 iterations, with the first 500 discarded and the last 500 averaged to generate point estimates of the posterior mean for the parameters. These estimates are what is displayed throughout this section.

The S-GPPF accurately models the synthetic data. To illustrate this, the original document $\times$ word count matrix and document class labels are estimated from the



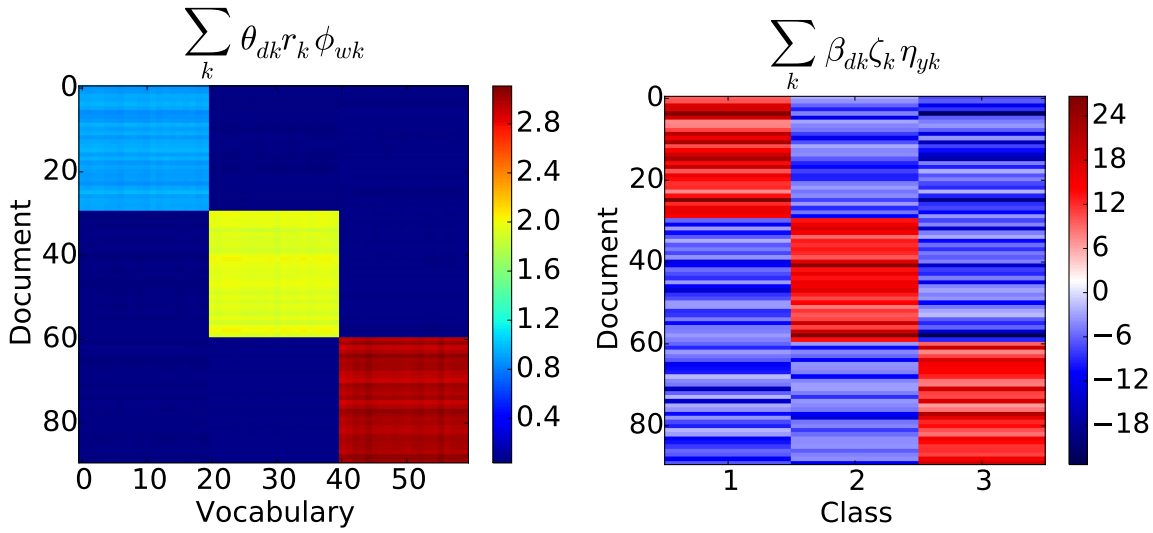
(a) Document × Vocabulary count matrix.

(b) Document class labels.

Figure 4.1: Synthetic block diagonal data

model parameter estimates. Since the data is assumed to be drawn from a Poisson likelihood, the count matrix is estimated as $\hat{x}_{dw} = \sum_{k=1}^{K} \theta_{dk} r_k \phi_{wk}$. Similarly, the z_{yd} parameters are estimated by their mean, $\hat{z}_{yd} = \sum_k \beta_{dk} \zeta_k \eta_{yk}$. The class labels can then be estimated by Eq. (3.1). These parameter estimates are shown in Fig 4.2.

The next thing to evaluate is whether the Markov chain has converged to the stationary distribution and if the non-parametric process is working as expected. An ad-hoc method for assessing the convergence of Gibbs sampling is to examine trajectory plots of parameter samples. Such trajectory plots are generated for the strength of the topics, given by r_k in Figure 4.3. The red region on the plot highlights the "burn-in" iterations. The trajectory plots are indicative of a chain which has reached its invariant distribution. From the clearly defined block diagonal structure in the input data, three distinct topics should emerge from the data. The gamma process prior on r_k allows for the automatic discovery of the number of topics. In these tests, the number of topics is artificially set to ten, but the estimate for r_k



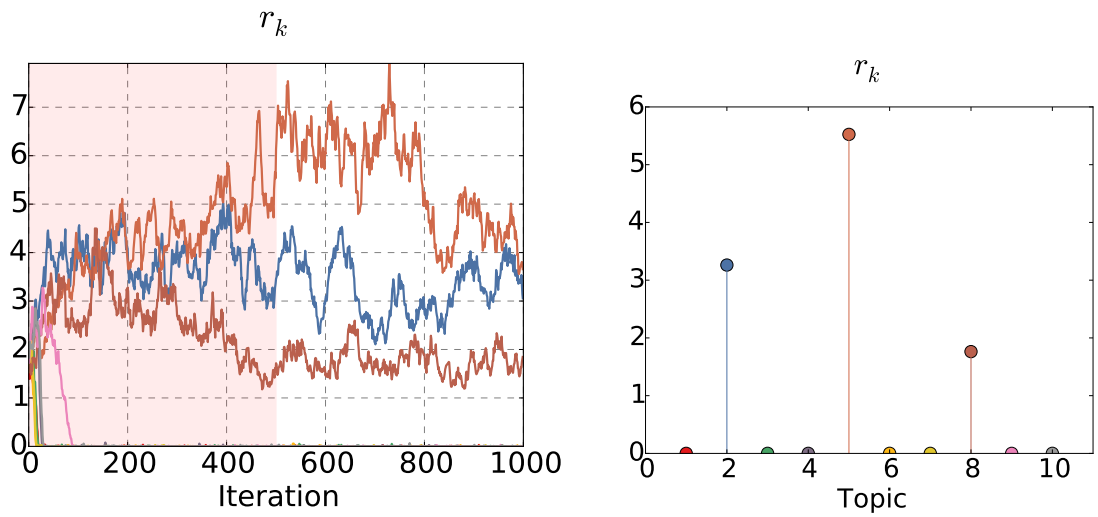
(a) Estimate of Document × Vocabulary count matrix from model parameters.

(b) Estimate of $\boldsymbol{z}$ matrix from the model parameters.

Figure 4.2: Reconstructed block diagonal $\boldsymbol{X}$ and $\boldsymbol{z}$ matrices

(see Figure 4.3) assigns significant weight to only three topics, which is the desired behavior in non-parametric modeling. A parametric model (such as LDA) assigns weights for all ten topics.

To further explore and understand the latent topic space of S-GPPF, it is useful to examine the mappings from documents and words to latent topics. Given the block diagonal input structure, each third of documents should map onto its own topic. The affinity between document d and topic k is computed as $\theta_{dk}\sqrt{r_k}$. This quantity can be thought as the degree to which the content of document d is from topic k. Likewise, each third of the vocabulary should map onto a single topic, and that topic should be the same as the corresponding third of documents (due to the block diagonal input structure). The affinity between word w and topic k is computed as $\phi_{wk}\sqrt{r_k}$. This quantity gives the relative weighting for a word w onto the topics. The document- and word- topic affinities are shown in Figure 4.4, and have the expected structure.



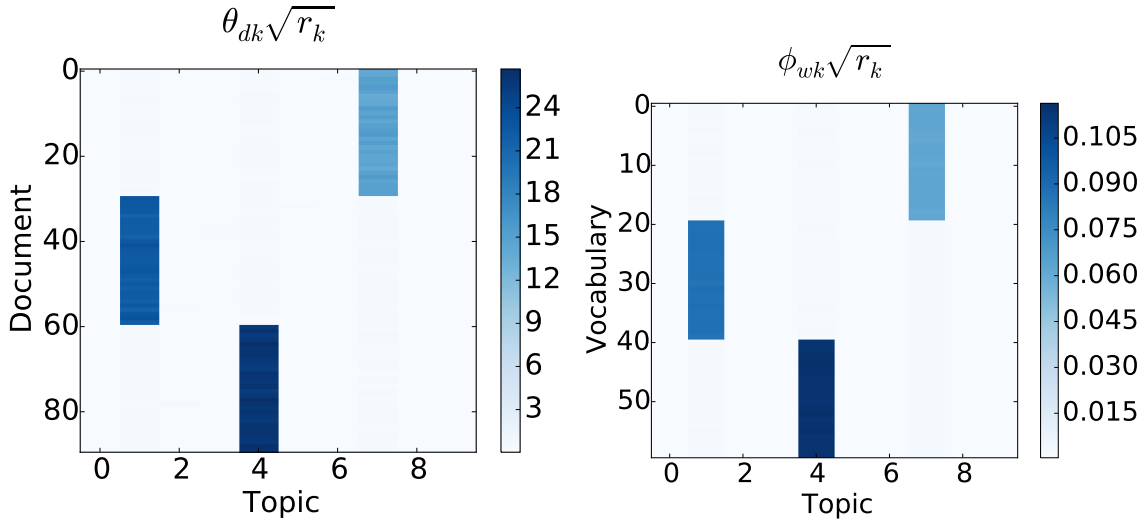
(a) Trajectory plot of active topics, the red highlight indicates burn-in iterations

(b) Active topics

Figure 4.3: Active topics in synthetic data

Note that the product of these matrices leads to the estimate for x_{dw}, as shown in Figure 4.2.

Also of value to explore are the higher level document and class mappings that are used for class label prediction. Recall that using the document-to-topic mappings given by $\theta_{dk}\sqrt{r_k}$ directly as the classification feature matrix is intractable under the multiclass max-margin formulation used in S-GPPF. Instead, β_{dk} and ζ_k are introduced, which are logarithmically proportional under expectation to θ_{dk} and r_k, respectively. Ideally, the higher level document-to-topic affinities given by β_{dk} should have similar structure as θ_{dk}. The classification weights to assign class label y from topic topic k is given by $\eta_{yk}\zeta_k$. Under the block-diagonal structure of the data, the class label of each third of documents should map onto the same latent topic as its corresponding documents. These quantities are shown in Figure 4.5, and have the desired structure. Note that the product of these matrices leads to the estimate for



(a) Document-to-topic affinity

(b) Vocabulary-to-topic affinity

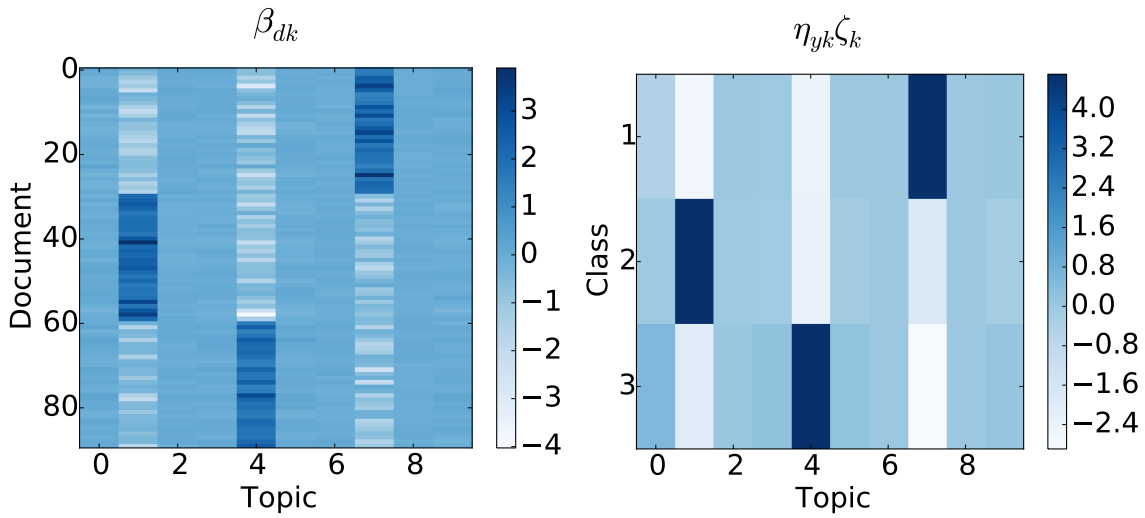
Figure 4.4: Document- and word- topic affinities in synthetic data

z_{yd}, as shown in Figure 4.2.

4.2 ACM Conference Abstracts Classification

The ACM conference abstracts text corpus described by Acharya et al. [2013] consists of abstracts collected from four data mining related conferences and two VLSI conferences. The data mining conferences are Knowledge Discovery and Data Mining (KDD), the International Conference on Machine Learning (ICML), the Special Interest Group on Information Retrieval (SIGIR), and the International World Wide Web conference (WWW). The two VLSI conferences are the International Symposium on Physical Design (ISPD), and the Design Automation Conference (DAC). A total of 5,755 abstracts were collected. The documents in this dataset are abstracts and the class labels are the conferences in which each abstract appeared.

The abstracts are preprocessed as follows: each abstract is converted to a count vector under a bag-of-words assumption. Bag-of-words assumes that the specific



(a) Higher level document-to-topic affinity

(b) Class-to-topic affinity

Figure 4.5: Class regression parameters in synthetic data

ordering of words within a text document is unimportant (this is not really the case: consider any paragraph in this document as an example). Hence each document is represented as a "bag" of the words found in the document. Raw text is tokenized using the Natural Language Tool Kit (NLTK, Loper and Bird [2002]) word tokenizer with punctuation and numeric words stripped from the resulting tokens. Tokens are stemmed using a Porter stemmer [Porter, 2001]. The set of English stop words provided by NLTK are removed from the resulting tokens. Additionally, rare corpus words (words that appear in less than 1% of all documents) and corpus specific stop words (words that appear in more than 50% of all documents) are also removed. After preprocessing, the vocabulary size is 971 words. A histogram of document lengths (in words) is provided in Figure 4.6. The black vertical lines indicate the edges of bins used to compute classification accuracy vs. document length. The document frequency of words and the number of document per length bin are shown in Figure 4.7.

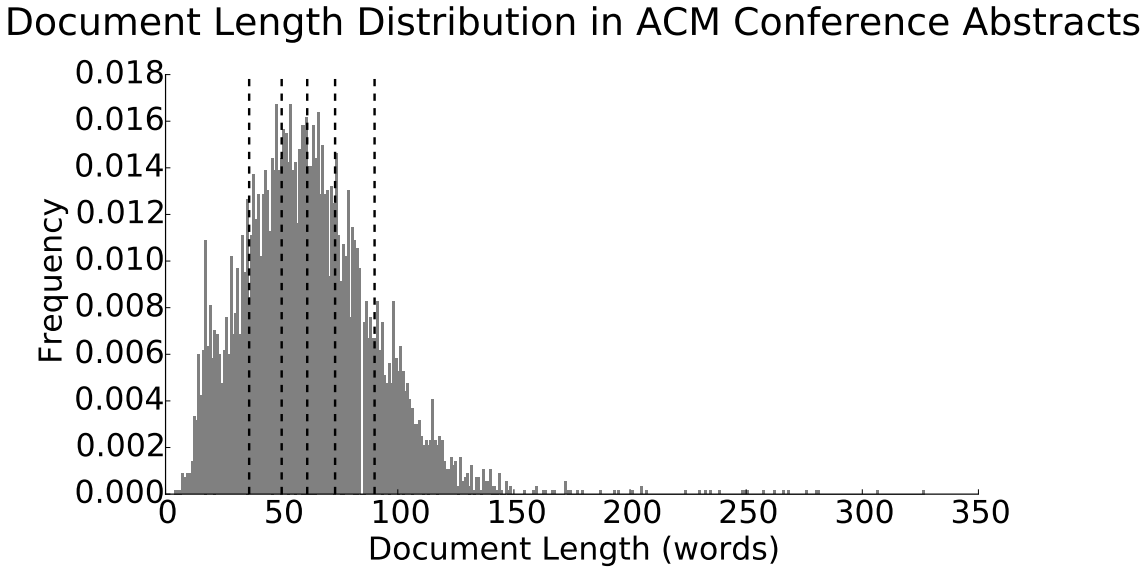
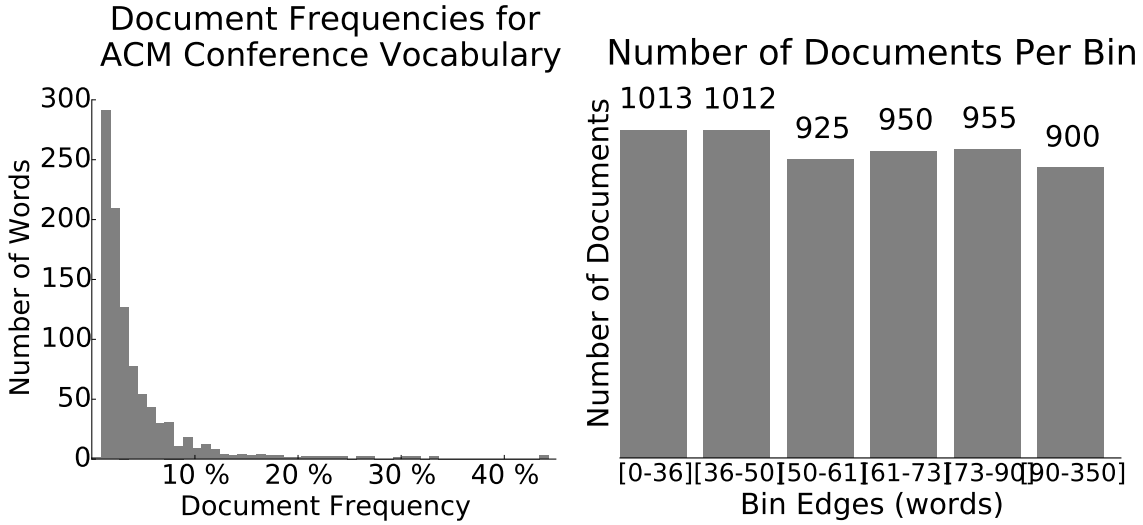


Figure 4.6: Distribution of abstract lengths (in words) of ACM conference data. Black lines denote the bins used for computing classification accuracy vs doc length (see Figure 4.9)



(a) Document frequencies of words in ACM conference dataset vocabulary after preprocessing.

(b) Number of documents assigned to each document length bin.

Figure 4.7: Dataset statistics for ACM conference abstracts.

The following models are also run on the ACM conference abstracts dataset with identical preprocessing to compare against S-GPPF. All model parameters were selected using a standard 10-fold cross validation on the entire dataset.

- Maximum entropy discrimination latent Dirichlet allocation (MedLDA, Zhu et al. [2009]). Model parameters are the number of topics, $K = 50$, and the max-margin penalty factor, $C = 30$. This model is intended as a strong baseline as it jointly models both the class labels and count matrix, and is generally considered state-of-the-art in supervised topic modeling.

- Latent Dirichlet allocation [Blei et al., 2003] with support vector machine (LDA + SVM). LDA is fit on the entire dataset, and the resulting document-topic matrix (the $\boldsymbol{\theta}$ parameter from LDA) is used as the feature matrix for a linear support vector classifier. Model parameters are the number of topics, $K = 50$, and the standard linear SVM tuning parameters with margin penalty $C = 1.0$. This model is intended as a weak baseline, as topics and class labels are learned in a disjoint manner.

Tests are run as follows. Twenty five independent splits of the data into equal test and train sets are generated, maintaining class proportions to be the same as in the whole dataset. Results are aggregated across independent splits, computing the mean and standard deviation (the standard deviation is represented in performance plots as error bars). Within this framework, accuracy is compared against two different variables: amount of training data and document length. The model is sampled for two thousand burnin and two thousand collection iterations for each independent run of the model.

The first test compares classification accuracy vs. the amount of training data. For each train/test split, only a portion of the training data is available to the models. The training data is subsampled to maintain relative class proportions in 10% increments up to 100% of the training data. The results of this test are shown in Figure 4.8. Since S-GPPF is a fully generative model, it outperforms the state of the art model, MedLDA, by a large margin when there is limited training data available. By using fully probabilistic priors, S-GPPF better generalizes to unseen data than discriminative models. The LDA + SVM performs worse than S-GPPF and MedLDA, which both jointly learn the topics and labels. This is exaggerated for small amounts of training data since the topics learned in the disjoint model are not well suited to the classification task.

The second test compares classification accuracy vs. document length. All of the training data is available in these tests. The documents are binned in to equal volume bins (see Figure 4.6 for the bin placement and Figure 4.7 for the bin volume). Classification accuracy is then computed for each of the bins. The results of the test are shown in Figure 4.9. These plots show that S-GPPF uniformly outperforms competing supervised topic models for widely varying document lengths, a useful property for modeling the long tail often found in real-world data [Gopalan et al., 2013].

The topics learned in S-GPPF are easily interpreted. To visualize this, a single run of S-GPPF is considered using all of the available training data. For each conference, the document-to-topic weights ($\theta_{dk}\sqrt{r_k}$) are summed along the document axis and the topic with the greatest weight for each conference is considered. Each topic can be viewed as a distribution over words, given by ϕ_{wk}. The top ten words (by weight) for the top topic of each conference are shown in Table 4.1. The data mining conferences

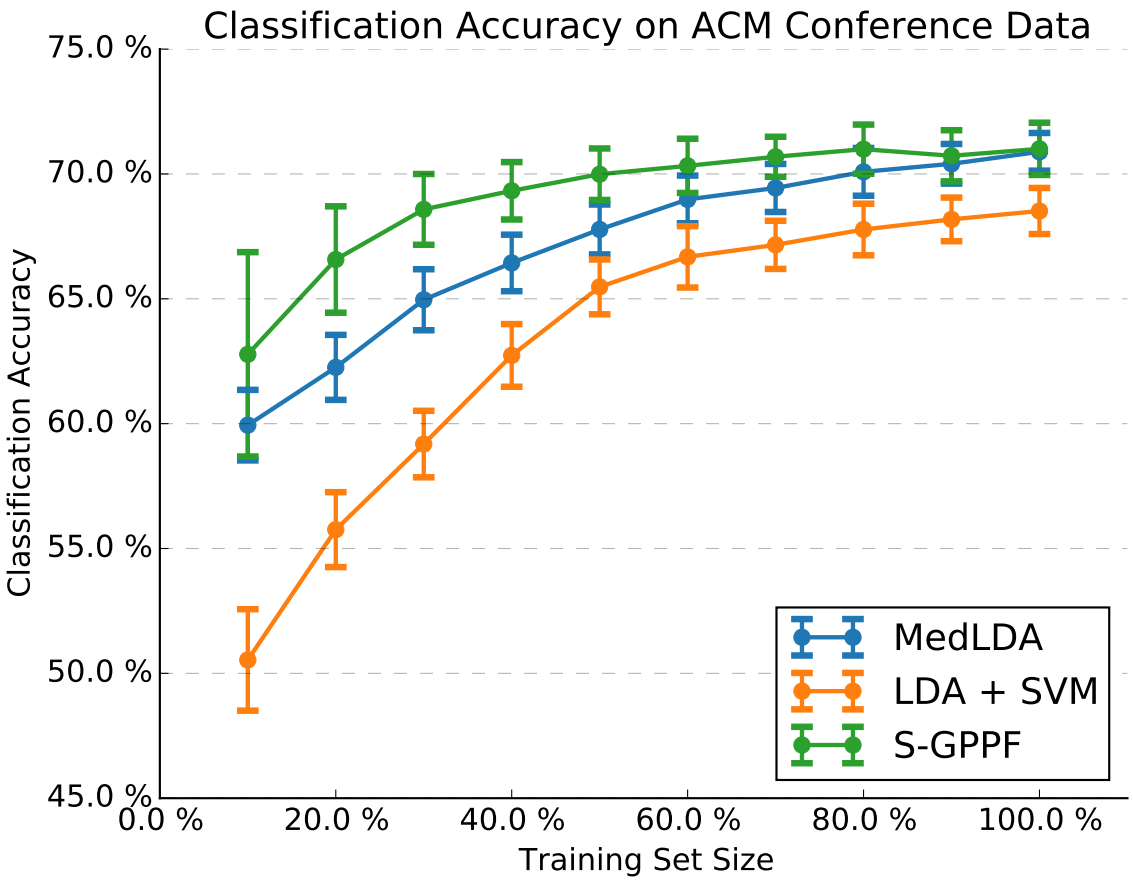


Figure 4.8: ACM conference abstracts classification accuracy vs. percent of training data observed

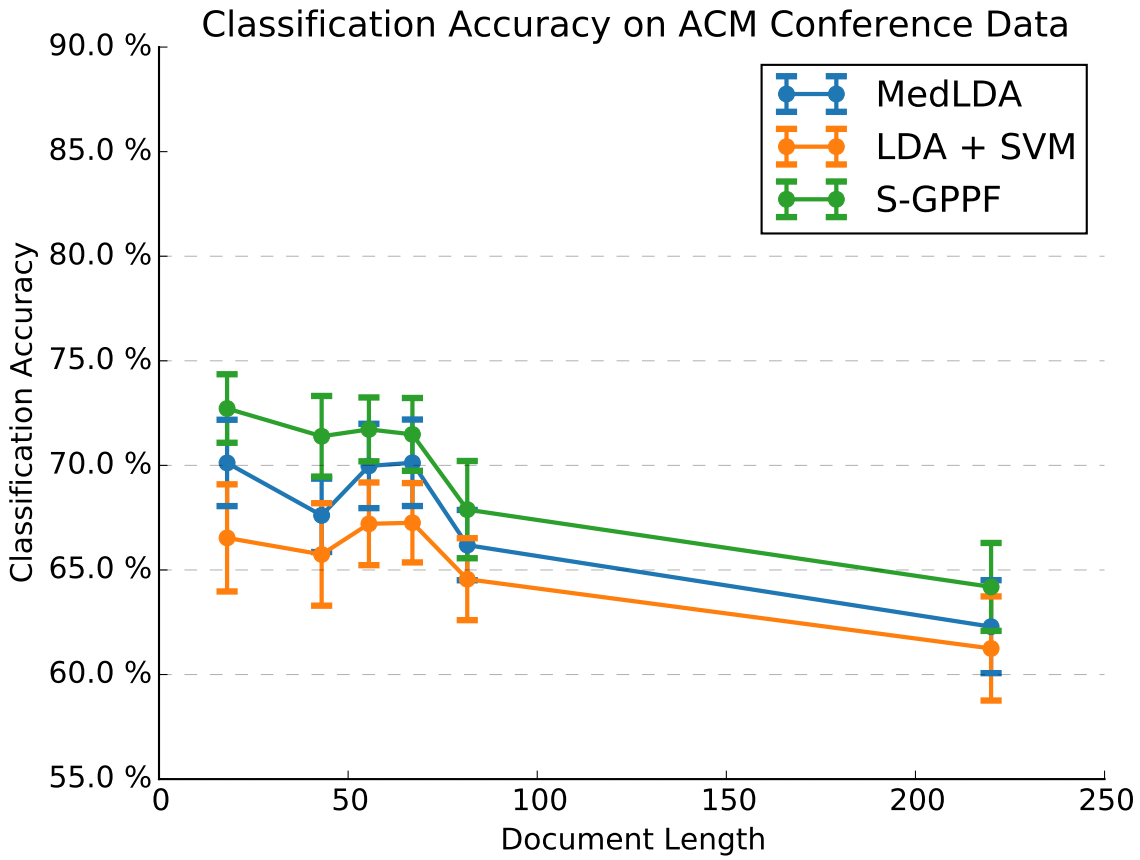


Figure 4.9: ACM conference abstracts classification accuracy vs. document length in words

KDD	ICML	SIGIR	WWW	ISPD	DAC
deal	algorithm	dynamic	server	plan	system
mixture	outperform	return	weight	router	device
algorithm	minimum	query	appropriate	device	appropriate
people	propose	test	provide	algorithm	throughput
growth	embed	baseline	system	chain	life
propose	baseline	insert	difficulty	baseline	simulate
approximate	gather	minimum	utility	throughput	methodology
set	retrieval	reliable	deal	retrieval	period
differentiable	configure	period	baseline	outperform	arising
minimum	set	approximate	internet	worst	process
recall	solve	independent	procedure	learn	technology
decision	architecture	propose	determine	busy	solution
retrieval	context	textual	ir	view	hidden
strong	approximate	latter	supply	intelligent	partial

Table 4.1: Top topic for each conference

(KDD, ICML, SIGIR, and WWW) feature words related to data mining: minimum, query, approximate, etc. The VLSI conferences (ISPD and DAC) prominently feature words related to the VLSI field: router, throughput, device, etc. This indicates that the Poisson factorization captures the low dimensional topic space known to exist in the data.

Chapter 5

Conclusion

S-GPPF represents a novel supervised topic model due to its unique properties. S-GPPF addresses the supervised topic modeling in a fully probabilistic framework. It directly models multiclass document labels and automatically selects the number of latent topics present in a corpus. Furthermore, S-GPPF provides for exact inference by using several data augmentation techniques for closed form Gibbs sampling updates. S-GPPF is shown to provide state-of-the-art levels of performance, simultaneously learning clearly interpretable features and document class labels. Because of its fully probabilistic framework, S-GPPF gains further advantage over competing models when the amount of training data available is limited.

This chapter is structured as follows. Section 5.1 provides suggestions on implementing and running S-GPPF in practical applications. Section 5.2 concludes this thesis with a discussion of further avenues of research and provides final thoughts.

5.1 Practical Suggestions

S-GPPF is a good choice for document classification when the resulting classifier should use clearly interpretable features in its decisions. The latent topic representation makes the S-GPPF model easy to interpret, as documents and class labels are represented as mixtures of topics which are in turn mixtures of words. The fully generative nature of S-GPPF makes the class-decision rules easy to understand as they are given by probability distributions. S-GPPF also gains clear advantage when the

number of topics present in a corpus is not easily discernible or known *apriori*: extra topics are automatically pulled to zero to avoid junk topics. This means that S-GPPF can be used to estimate the number of clusters in a labeled corpus. Additionally, S-GPPF gains substantial computational benefits by not requiring a cross-validation over the number of topics.

The implementation of S-GPPF used in this thesis leads to several suggestions regarding efficient computation and numeric stability in practice. This implementation makes use of the Armadillo C++ library for linear algebra [Sanderson, 2015] and the GNU Scientific Library [Galassi et al., 2010] for random number generation. The conditional posterior sampling requires the inversion of $K \times K$ precision matrices for multivariate normal sampling. Recognizing that precision matrices are positive semi-definite allows for more stable and faster methods of inversion such as through the use of the Cholesky decomposition, which is provided by the Armadillo library.

Samplers for several of the distributions found in the conditional posterior updates are not readily available in common libraries: namely the Chinese restaurant table count, Pólya-Gamma, and inverse-Gaussian random variables. The Chinese restaurant table can be implemented as a sum of Bernoulli draws with varying parameters [Zhou et al., 2012] and inverse-Gaussian sampling can be implemented using a Gaussian sampler [Michael et al., 1976]. Polson et al. [2013] describe an efficient sampler for Pólya-Gamma random variables, but it requires integer valued parameters. S-GPPF makes use of Pólya-Gamma random variables with real-valued parameters, requiring a truncation of an infinite sum of gamma distributed random variables. Such a sum works wells for most parameter values, but becomes unstable for small floating point values of parameters. The parameter values required for S-GPPF are very small due to the sparsity in latent topics. The implementation used in this thesis corrects for

this floating point bias by scaling the result of the Pólya-Gamma sampler to ensure that the first moment of the resulting samples maintains the expected value.

Other variables in the model are subject to floating point errors when they become sufficiently small. Specifically, r_k and θ_{dk} (since they are sparse in topic space and are used under a logarithmic transform in the sampling of other variables) and α_{dk}, ϵ_k, and ν_k (since they represent precisions). Since the values need only be small compared to the active topic values, a minimum value clamp can be used to prevent numerical instabilities. The implementation used here clamps these variables to greater than 10^{-10}.

5.2 Discussion and Future Work

There is room for improvement in the classification performance of S-GPPF due to the multi-class max-margin formulation used. The formulation requires the introduction of the latent z_{yd} random variables, which add Gaussian noise to the classification features. The z_{yd}'s are needed due to the sum-to-zero constraint of Eq. (2.5). This constraint makes it intractable to directly link $\boldsymbol{\eta}$, $\boldsymbol{\beta}$, and $\boldsymbol{\zeta}$ to the class labels. Adopting a different multi-class max-margin formulation or data augmentation strategy that does not require the introduction of $\boldsymbol{z}$ should lead to an increase in classifier performance. Additionally, the current formulation does not fit intercepts. Adding an intercept term to the learned parameters could also boost classifier performance.

The multi-class nature of S-GPPF can be extended to a multi-task setting. The latent variables are already formulated in a multi-task framework: different class labels serve as an inductive bias for inferring document labels. The existing classifier structure suggests easy extensions to problems where a vector of binary labels are observed for each document rather than a single multi-class response variable. In such

a problem, a document may belong to multiple classes: for instance, a movie may belong to multiple genres. The framework can also be extended to the active learning setting similar to Acharya et al. [2013]. In the active learning setting, classification labels are very expensive to obtain. The modeling process formulates queries of the training examples for which class labels would be the most informative in modeling.

S-GPPF fits neatly into a group of several models which extend non-parametric PFA to problems in which information in addition to the count matrix is observed. In S-GPPF, document class labels are also observed. Acharya et al. [2015] extends PFA to modeling count matrices when the columns are temporally related. Zhou [2015] jointly models count matrices along with network side information. There is potential to unify these modeling extensions under a single, broad non-parametric PFA framework. Such a framework could jointly model count matrices with the side information available on a case-by-case basis.

This thesis developed and presented the supervised gamma process Poisson factorization model. S-GPPF represents a novel supervised topic model; it is fully generative and nonparametric, allows for multi-class classification, and provides for exact inference via Gibbs sampling. S-GPPF is shown to outperform MedLDA and other competing topic models for classification. S-GPPF fits neatly into a framework of extending the broad class of algorithms that can be unified under Poisson factorization to including additional side information.

Appendices

Appendix A

Proofs of Lemmas

Proof of lemma 2.3.1. The proof follows directly from application of Bayes' rule

$$
\begin{aligned}
p(\lambda|\{x_i\}) &\propto p(\lambda|r,c)\prod_i p(x_i|m_i,\lambda) \\
&\propto \lambda^{r-1}\exp\left(-\lambda c\right)\prod_i \lambda^{x_i}\exp\left(-m_i\lambda\right) \\
&= \lambda^{r+\sum_i x_i - 1}\exp\left(-\lambda\left(c+\sum_i m_i\right)\right) \\
\implies \lambda|\{x_i\} &\sim \text{Gamma}\left(r+\sum_i x_i, 1/\left(c+\sum_i m_i\right)\right)
\end{aligned}
$$

□

Proof of lemma 2.3.2. This lemma follows directly from application of Bayes' rule

$$
\begin{aligned}
p(b|\{r_i\}) &\propto p(b|c,d)\prod_i p(r_i|a_i,b) \\
&= \text{Gamma}\left(b;c,1/d\right)\prod_i \text{Gamma}\left(r_i;a_i,1/b\right) \\
&\propto b^{c-1}\exp\left(-bd\right)\prod_i b^{a_i}\exp\left(-r_i b\right) \\
&= b^{c+\sum_i a_i - 1}\exp\left(-b\left(d+\sum_i r_i\right)\right) \\
\implies b|\{r_i\} &\sim \text{Gamma}\left(c+\sum_i a_i, 1/\left(d+\sum_i r_i\right)\right)
\end{aligned}
$$

□

Proof of lemma 2.3.3. This lemma follows directly from application of Bayes' rule

$$
\begin{aligned}
p(\sigma|\{z_i\}) &\propto p(\sigma|a,b)\prod_{i=1}^{K} p(z_i|\mu_i,\sigma) \\
&= \text{Gamma}\left(\sigma; a, 1/b\right)\prod_{i=1}^{K}\mathcal{N}\left(z_i;\mu_i,\sigma^{-1}\right) \\
&\propto \sigma^{a-1}\exp\left(-\sigma b\right)\prod_{i=1}^{K}\sigma^{1/2}\exp\left(-\sigma\frac{\left(z_i-\mu_i\right)^2}{2}\right) \\
&= \sigma^{a+K/2-1}\exp\left(-\sigma\left(b+\sum_{i=1}^{K}\frac{\left(z_i-\mu_i\right)^2}{2}\right)\right) \\
\implies \sigma|\{z_i\} &\sim \text{Gamma}\left(a+K/2, 1/\left(b+\sum_{i=1}^{K}\frac{\left(z_i-\mu_i\right)^2}{2}\right)\right)
\end{aligned}
$$

□

Proof of lemma 2.3.4. The joint distribution for $\{y_k\}_{k=1}^{K}$ is given as

$$
\begin{aligned}
p(y_1,\cdots,y_K;X) &= X!\prod_{k=1}^{K}\frac{(\zeta_i/\zeta)^{y_k}}{y_k!} \\
&= \frac{X!}{\zeta^X}\prod_{k=1}^{K}\frac{\zeta_k^{y_k}}{y_k!} \\
&\propto 1\left[\sum_{k=1}^{K} y_i = X\right]\prod_{k=1}^{K}\frac{\zeta_k^{y_k}}{y_k!}
\end{aligned}
$$

which has the same form as $p(x_1,\cdots,x_K)$ since $\sum_{k=1}^{K} x_k = X$ by construction. □

Proof of lemma 2.3.5. Since a summation of Poisson random variables is also a Poisson, $x \sim \text{Poisson}\left(\lambda\sum_i m_i\right)$. Integrating out λ gives the following form for the

distribution of x:

$$
\begin{aligned}
p(x) &= \int_0^\infty \text{Gamma}\left(\lambda; r, 1/c\right) \text{Poisson}\left(x; \lambda \sum_i m_i\right) d\lambda \\
&= \int_0^\infty \frac{\lambda^{r-1}\exp\left(-\lambda c\right)}{c^r \Gamma(r)} \frac{\left(\lambda \sum_i m_i\right)^x}{x!} \exp\left(-\lambda \sum_i m_i\right) d\lambda \\
&= \frac{\left(\sum_i m_i\right)^x c^r}{x!\Gamma(r)} \int_0^\infty \lambda^{r+x-1} \exp\left(-\lambda\left(c + \sum_i m_i\right)\right) d\lambda \\
&= \frac{\left(\sum_i m_i\right)^x c^r}{x!\Gamma(r)} \frac{\Gamma(r+x)}{\left(c + \sum_i m_i\right)^{r+x}} \\
&= \frac{\Gamma(r+x)}{x!\Gamma(r)} \left(\frac{\sum_i m_i}{c + \sum_i m_i}\right)^x \left(1 - \frac{\sum_i m_i}{c + \sum_i m_i}\right)^r \\
\implies x &\sim \text{NB}\left(r, \frac{\sum_i m_i}{c + \sum_i m_i}\right)
\end{aligned}
$$

□

Proof of lemma 2.3.6. The derivation of this lemma is found in Zhou et al. [2012]. □

Proof of lemma 2.3.7. Since the summation of Poisson random variables is also Poisson, $x = \sum_i x_i \sim \text{Poisson}\left(r_2 \sum_i m_i\right)$. From lemma 2.3.5, this implies $x \sim \text{NB}(r_1, p)$ where $p = \frac{\sum_i m_i}{d + \sum_i m_i}$. Then from the compound Poisson construction of the Negative Binomial, $l \sim \text{Poisson}\left(-r_1 \log(1-p)\right)$. From lemma 2.3.6, $l|x, r_1 \sim \text{CRT}(x, r_1)$. Finally, by the gamma-Poisson conjugacy (lemma 2.3.1),

$r_1|l, \cdots \sim \text{Gamma}\left(r + l, 1/(b - \log(1-p)\right)$, which is the desired result. □

Proof of lemma 2.3.8. The derivation of this lemma is found in Polson et al. [2013].

□

Appendix B

Conditional Posterior Derivations

Sampling of $(x_{dwk})_{k=1}^{K}$

The conditional posterior of $(x_{dwk})_{k=1}^{K}$ follows directly from the multinomial-Poisson distribution equivalence lemma 2.3.4.

$$\begin{aligned} p(x_{dw1}, \cdots, x_{dwK}| \cdots) &\propto \prod_k p(x_{dwk}|r_k, \theta_{dk}, \phi_{wk}) \\ &= \prod_k \text{Poisson}\left(x_{dwk}; r_k\theta_{dk}\theta_{dk}\right) \\ x_{dw} &= \sum_k x_{dwk} \end{aligned}$$

$$(x_{dwk})_{k=1}^{K}| \cdots \sim \text{mult}\left(\left(\frac{r_k\theta_{dk}\phi_{wk}}{\sum_{k=1}^{K} r_k\theta_{dk}\phi_{wk}}\right)_{k=1}^{K}; x_{dw}\right) \tag{B.1}$$

Sampling of θ_{dk}

The conditional posterior of θ_{dk} follows directly from the gamma-Poisson conjugacy lemma 2.3.1.

$$\begin{aligned} p(\theta_{dk}| \cdots) &\propto p(\theta_{dk}|\beta_{dk}, \tau_d)p(x_{d.k}|r_k, \theta_{dk}) \\ &= \text{Gamma}\left(\theta_{dk}; \tau_d, \exp\left(\beta_{dk}\right)\right) \text{Poisson}\left(x_{d.k}; r_k\theta_{dk}\right) \end{aligned}$$

$$\theta_{dk}| \cdots \sim \text{Gamma}\left(\tau_d + x_{d.k}, 1/\left(\exp\left(-\beta_{dk}\right) + r_k\right)\right) \tag{B.2}$$

Sampling of τ_d

The conditional posterior of τ_d follows by repeated application of the CRT augmentation lemma 2.3.7. Introduce $l_{dk} \sim \text{Poisson}\left(-\tau_d \ln\left(1 - p_{dk}\right)\right)$ where $p_{dk} = \frac{\sum_k r_k}{\exp(-\beta_{dk}) + r_k}$. The posterior is then found by application of the gamma-Poisson conjugacy lemma 2.3.1. Then

$$l_{dk}| \cdots \sim \text{CRT}\left(x_{d.k}, \tau_d\right) \tag{B.3}$$

$$\tau_d| \cdots \sim \text{Gamma}\left(c_0 + \sum_k l_{dk}, 1/(d_0 - \sum_k \log(1 - p_{dk}))\right) \tag{B.4}$$

Sampling of $\boldsymbol{\phi}_k$

$$p(\boldsymbol{\phi}_k| \cdots) \propto p(\boldsymbol{\phi}_k|\boldsymbol{\xi}) \prod p(x_{.wk}|r_k, \theta_{.k}, \phi_{wk})$$

$$\boldsymbol{\phi}_k| \sim \text{Dir}\left(\xi_1 + x_{.1k}, \cdots, \xi_V + x_{.Vk}\right) \tag{B.5}$$

Sampling of r_k

The conditional posterior of r_k follows directly from the gamma-Poisson conjugacy lemma 2.3.1.

$$\begin{aligned} p(r_k| \cdots) &\propto p(r_k|\gamma_0, \zeta_k) p(x_{..k}|r_k, \theta_{.k}) \\ &= \text{Gamma}\left(r_k; \gamma_0/K, \exp\left(\zeta_k\right)\right) \text{Poisson}\left(x_{..k}; r_k \theta_{.k}\right) \end{aligned}$$

$$r_k| \sim \text{Gamma}\left(\gamma_0/K + x_{..k}, 1/\left(\exp\left(-\zeta_k\right) + \theta_{.k}\right)\right) \tag{B.6}$$

Sampling of γ_0

The conditional posterior of γ_0 follows by repeated application of the CRT augmentation lemma 2.3.7. Introduce $l_k \sim \text{Poisson}\left(-\gamma_0/K \ln\left(1 - p_k\right)\right)$ where $p_k =$

$\frac{\theta_{.k}}{\exp(-\zeta_k)+\theta_{.k}}$. The posterior is then found by application of the gamma-Poisson conjugacy lemma 2.3.1. Then

$$l_k|\cdots \sim \text{CRT}\left(x_{..k}, \gamma_0/K\right) \tag{B.7}$$

$$\gamma_0|\cdots \sim \text{Gamma}\left(a_0 + \sum_k l_k, 1/(b_0 - \frac{1}{K}\sum_k \log(1-p_k))\right) \tag{B.8}$$

Sampling of σ

The conditional posterior of σ follows directly from the gamma-normal conjugacy lemma 2.3.3.

$$\begin{aligned} p(\sigma|\cdots) &\propto p(\sigma|s_0,t_0)\prod_{y,d} p(z_{yd}|\sigma,\boldsymbol{\beta}_d,\boldsymbol{\eta}_y,\boldsymbol{\zeta}) \\ &= \text{Gamma}\left(\sigma; s_0, 1/t_0\right)\prod_{y,d}\mathcal{N}\left(z_{yd}; \sum_k \beta_{dk}\eta_{yk}\zeta_k, 1/\sigma\right) \\ \sigma|\cdots &\sim \text{Gamma}\left(s_0 + \frac{MD}{2}, 1/t_0'\right), \end{aligned} \tag{B.9}$$

where $t_0' = \left(t_0 + \sum_{y,d}\frac{(z_{yd} - \sum_k \eta_{yk}\beta_{dk}\zeta_k)^2}{2}\right)$.

Sampling of ϵ_k

The conditional posterior of ϵ_k follows directly from the gamma-normal conjugacy lemma 2.3.3.

$$\begin{aligned} p(\epsilon_k|\cdots) &\propto p(\epsilon_k|e_0,f_0)\prod_y p(\eta_{yk}|\epsilon_k) \\ &= \text{Gamma}\left(\epsilon_k; e_0, 1/f_0\right)\prod_y \mathcal{N}\left(\eta_{yk}; 0, \epsilon_k^{-1}\right) \\ \epsilon_k|\cdots &\sim \text{Gamma}\left(e_0 + \frac{M}{2}, 1/\left(\sum_y \frac{\eta_{yk}^2}{2} + f_0\right)\right) \end{aligned} \tag{B.10}$$

Sampling of ν_k

The conditional posterior of ν_k follows directly from the gamma-normal conjugacy lemma 2.3.3.

$$\begin{aligned} p(\nu_k|\cdots) &\propto p(\nu_k|u_0, v_0)p(\zeta_k|\nu_k) \\ &= \text{Gamma}\left(\nu_k; u_0, 1/v_0\right)\mathcal{N}\left(\zeta_k; 0, \nu_k^{-1}\right) \\ \nu_k| &\sim \text{Gamma}\left(u_0 + \frac{1}{2}, 1/\left(\frac{\zeta_k^2}{2} + v_0\right)\right) \end{aligned} \tag{B.11}$$

Sampling of α_{dk}

The conditional posterior of α_{dk} follows directly from the gamma-normal conjugacy lemma 2.3.3.

$$\begin{aligned} p(\alpha_{dk}|\cdots) &\propto p(\alpha_{dk}|g_0, h_0)p(\beta_{dk}|\alpha_{dk}) \\ &= \text{Gamma}\left(\alpha_{dk}; g_0, 1/h_0\right)\mathcal{N}\left(\beta_{dk}; 0, \alpha_{dk}^{-1}\right) \\ \alpha_{dk}|\cdots &\sim \text{Gamma}\left(g_0 + \frac{1}{2}, 1/\left(\frac{\beta_{dk}^2}{2} + h_0\right)\right) \end{aligned} \tag{B.12}$$

Sampling of z_{yd}

Since $\sum_{y=1}^{M} z_{yd} = 0$, only $\{z_{yd}\}_{y \neq y_d}$ need be sampled for each document d; assign $z_{y_d d} = -\sum_{y \neq y_d} z_{yd}$. Then for $y \neq y_d$:

$$\begin{aligned} p(z_{yd}|\cdots) &\propto p(z_{yd}|\boldsymbol{\beta}_d, \boldsymbol{\eta}_y, \boldsymbol{\zeta}, \sigma)q(y_d|\cdots) \\ &\propto \mathcal{N}\left(z_{yd}; \sum_k \beta_{dk}\eta_{yk}\zeta_k, \sigma^{-1}\right)\exp\left(-\left(z_{yd} + \frac{1}{M-1}\right)_+\right) \end{aligned}$$

To handle the second term in this expression, the inverse-Gaussian data augmentation of lemma 2.4.1 is used. Introduce γ_{yd}, where

$$\gamma_{yd}|\cdots \sim \mathcal{IG}\left(\left|\frac{z_{yd} + \frac{1}{M-1}}{2}\right|^{-1}, 1\right) \tag{B.13}$$

Then from the SVM data-augmentation strategy we have

$$p(z_{yd}|\cdots) \propto \exp\left(-\frac{\sigma}{2}\left(z_{yd} - \sum_k \beta_{dk}\eta_{yk}\zeta_k\right)^2\right) \exp\left(-\frac{1}{2\gamma_{yd}}\left(\frac{z_{yd} + \frac{1}{M-1}}{2} + \gamma_{yd}\right)^2\right)$$

$$\propto \exp\Bigg(-\frac{1}{2}\Bigg[\sigma\left(z_{yd}^2 - 2z_{yd}\sum_k \beta_{dk}\eta_{yk}\zeta_k\right) + \frac{1}{\gamma_{yd}}\left(\frac{1}{4}\left(z_{yd} + \frac{1}{M-1}\right)^2 + \gamma_{yd}\left(z_{yd} + \frac{1}{M-1}\right)\right)\Bigg]\Bigg)$$

Considering just the argument under the $\exp\left(-\frac{1}{2}(\cdots)\right)$, since that is the form of a normal distribution:

$$\sigma\left(z_{yd}^2 - 2z_{yd}\sum_k \beta_{dk}\eta_{yk}\zeta_k\right) + \frac{1}{\gamma_{yd}}\left(\frac{1}{4}\left(z_{yd}^2 + \frac{2z_{yd}}{M-1}\right) + \gamma_{yd}z_{yd}\right)$$

$$= z_{yd}^2\left(\sigma + \frac{1}{4\gamma_{yd}}\right) - 2z_{yd}\left(\sigma\sum_k \beta_{dk}\eta_{yk}\zeta_k - \frac{1}{4\gamma_{yd}(M-1)} - \frac{1}{2}\right)$$

Comparing this to the P.D.F. of a normal distribution we get

$$z_{yd}|\cdots \sim \mathcal{N}\left(\mu', \sigma'^2\right) \tag{B.14}$$

where $\sigma'^2 = \frac{\gamma_{yd}}{\gamma_{yd}\sigma + 1/4}$ and $\mu' = \sigma'^2\left(\sigma\sum_k \beta_{dk}\eta_{yk}\zeta_k - \frac{1}{4\gamma_{yd}(M-1)} - \frac{1}{2}\right)$.

Sampling of $\boldsymbol{\eta}_y$

The conditional posterior distribution has the form of a multivariate-normal distribution.

$$p(\boldsymbol{\eta}_y|\cdots) \propto p(\boldsymbol{\eta}_y|\boldsymbol{\alpha}_1)\prod_d p(z_{yd}|\boldsymbol{\beta}_d, \boldsymbol{\eta}_y, \boldsymbol{\zeta})$$

$$= \mathcal{N}\left(\boldsymbol{\eta}_y; 0, \boldsymbol{\alpha}_1^{-1}\boldsymbol{I}_K\right)\prod_d \mathcal{N}\left(z_{yd}; \boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\beta}_d', \sigma^{-1}\right)$$

$$\propto \exp\left(-\frac{1}{2}\boldsymbol{\eta}_y(\boldsymbol{\alpha}_1\boldsymbol{I}_K)\boldsymbol{\eta}_y'\right)\prod_d \exp\left(-\frac{\sigma}{2}\left(z_{yd} - \boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'\right)^2\right)$$

$$= \exp\left(-\frac{1}{2}\left(\boldsymbol{\eta}_y(\boldsymbol{\alpha}_1\boldsymbol{I}_K)\boldsymbol{\eta}_y' + \sigma\sum_d\left(z_{yd} - \boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'\right)^2\right)\right)$$

Considering just the argument under the $\exp\left(-\frac{1}{2}\left(\cdots\right)\right)$, since that is the form of a normal distribution:

$$\begin{aligned}
&\boldsymbol{\eta}_y(\boldsymbol{\alpha}_1\boldsymbol{I}_K)\boldsymbol{\eta}_y' + \sigma\sum_d\left((\boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\beta}_d')(\boldsymbol{\beta}_d\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\eta}_y') - 2z_{yd}\boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'\right)\\
&= \boldsymbol{\eta}_y(\boldsymbol{\alpha}_1\boldsymbol{I}_K)\boldsymbol{\eta}_y' + \sigma\left(\boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\sum_d\left(\boldsymbol{\beta}_d'\boldsymbol{\beta}_d\right)\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\eta}_y' - 2\sum_d z_{yd}\boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'\right)\\
&= \boldsymbol{\eta}_y\left[\boldsymbol{\alpha}_1\boldsymbol{I}_K + \sigma\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\sum_d\left(\boldsymbol{\beta}_d'\boldsymbol{\beta}_d\right)\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\right]\boldsymbol{\eta}_y' - 2\sigma\boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\sum_d z_{yd}\boldsymbol{\beta}_d'
\end{aligned}$$

Comparing this expression with that of the multivariate-normal P.D.F. the expression for the conditional posterior is

$$\boldsymbol{\eta}_y|\cdots \sim \mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right), \tag{B.15}$$

where $\boldsymbol{\Sigma}_y^{-1} = \left[\boldsymbol{\alpha}_1\boldsymbol{I}_K + \sigma\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\sum_d\left(\boldsymbol{\beta}_d'\boldsymbol{\beta}_d\right)\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\right]$ and
$\boldsymbol{\mu}_y = \boldsymbol{\Sigma}_y\left(\sigma\boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\sum_d z_{yd}\boldsymbol{\beta}_d'\right)$

Sampling of $\boldsymbol{\beta}_d$

The conditional posterior distribution has the form of a multivariate-normal distribution. First, integrate out θ_{dk} as described as follows: $x_{d.k} \sim \text{Poisson}\left(r_k\theta_{dk}\right)$ (note that $\phi_{.k} = 1 \quad \forall k$) where $\theta_{dk} \sim \text{Gamma}\left(\tau_d, \exp\left(\beta_{dk}\right)\right)$. Therefore, by the gamma-Poisson construction of the Negative Binomial described in Section 2.3.4, $x_{d.k} \sim \text{NB}\left(\tau_d, p_{dk}\right)$ where $p_{dk} = \frac{r_k}{\exp(-\beta_{dk})+r_k}$. Then the posterior of $\boldsymbol{\beta}_d$ is given as follows:

$$\begin{aligned}
p(\boldsymbol{\beta}_d|\cdots) &\propto p(\boldsymbol{\beta}_d|\boldsymbol{\alpha}_d)\prod_k p(x_{d.k}|\tau_d, r_k, \beta_{dk})\prod_y p(z_{yd}|\boldsymbol{\eta}_y, \boldsymbol{\beta}_d, \boldsymbol{\zeta}, \sigma)\\
&= \mathcal{N}\left(\boldsymbol{\beta}_d; 0, \boldsymbol{\alpha}_d\boldsymbol{I}_K\right)\prod_k \text{NB}\left(x_{d.k}; \tau_d, p_{dk}\right)\prod_y \mathcal{N}\left(z_{yd}; \boldsymbol{\eta}_y\cdot\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\beta}_d', \sigma^{-1}\right)
\end{aligned}$$

The second term can be manipulated as follows:

$$
\begin{aligned}
p(x_{d.k}|\tau_d, r_k, \beta_{dk}) &= \mathrm{NB}\left(x_{d.k}; \tau_d, p_{dk}\right) \\
&\propto p_{dk}^{x_{d.k}}(1-p_{dk})^{\tau_d} \\
&= \left(\frac{r_k}{\exp\left(-\beta_{dk}\right)+r_k}\right)^{x_{d.k}} \left(1-\frac{r_k}{\exp\left(-\beta_{dk}\right)+r_k}\right)^{\tau_d} \\
&= \frac{r_k^{x_{d.k}}\exp\left(-\beta_{dk}\right)^{\tau_d}}{\exp\left(-\beta_{dk}\right)+r_k} \\
&= \frac{\left(r_k\exp\left(\beta_{dk}\right)\right)^{x_{d.k}}}{\left(r_k\exp\left(\beta_{dk}\right)+1\right)^{x_{d.k}+\tau_d}} \\
&= \frac{\exp\left(\psi_{dk}\right)^{x_{d.k}}}{\left(\exp\left(\psi_{dk}\right)+1\right)^{\kappa_{dk}}},
\end{aligned}
$$

where $\psi_{dk} = \beta_{dk} + \log(r_k)$ and $\kappa_{dk} = x_{d.k} + \tau_d$. Then the expression for the conditional posterior becomes

$$
\begin{aligned}
&p(\boldsymbol{\beta}_d|\cdots) \propto \\
&\exp\left(-\frac{1}{2}\boldsymbol{\beta}_d\left(\boldsymbol{\alpha}_d\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'\right)\prod_k \frac{\exp\left(\psi_{dk}\right)^{x_{d.k}}}{\left(\exp\left(\psi_{dk}\right)+1\right)^{\kappa_{dk}}}\prod_y \exp\left(-\frac{\sigma}{2}\left(z_{yd}-\boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'\right)^2\right)
\end{aligned}
$$

To handle the second factor, the Pólya-Gamma augmentation lemma 2.3.8 is used. Introduce $\omega_{dk} \sim \mathrm{PG}(\kappa_{dk}, 0)$, and apply lemma 2.3.8, for which the corresponding factor becomes proportional to the following:

$$
\exp\left(\frac{(x_{d.k}-\tau_d)\psi_{dk}}{2}\right)\int_0^\infty \exp\left(-\omega_{dk}\psi_{dk}^2/2\right)p(\omega_{dk})d\omega_{dk}
$$

By lemma 2.3.8, the posterior update of the augmented Pólya-Gamma variable is given by:

$$
\omega_{dk}|\cdots \sim \mathrm{PG}(\kappa_{dk}, \psi_{dk}). \tag{B.16}
$$

This leads to a posterior given by

$$
\begin{aligned}
&p(\boldsymbol{\beta}_d|\cdots) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}_d\left(\boldsymbol{\alpha}_d\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'\right)\prod_k \exp\left((x_{d.k}-\tau_d)/2\psi_{dk}-\omega_{dk}\psi_{dk}^2/2\right)\cdot \\
&\prod_y \exp\left(-\frac{\sigma}{2}\left(z_{yd}-\boldsymbol{\eta}_y\left(\boldsymbol{\zeta}\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'\right)^2\right) \\
&\propto \exp\Bigg(-\frac{1}{2}\boldsymbol{\beta}_d\left(\boldsymbol{\alpha}_d\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'+\sum_k\left[(x_{d.k}-\tau_d)/2\beta_{dk}-\omega_{dk}\left(\beta_{dk}^2+2\beta_{dk}\log(r_k)\right)/2\right]- \\
&\qquad\frac{\sigma}{2}\sum_y\left[(\boldsymbol{\eta}_y(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\beta}_d')(\boldsymbol{\eta}_y(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\beta}_d')-2z_{yd}(\boldsymbol{\eta}_y(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\beta}_d')\right]\Bigg)
\end{aligned}
$$

Let $\boldsymbol{\nu}_d=\{x_{d.k}-\tau_d\}_{k=1}^K$. Consider just the argument under the $\exp\left(-\frac{1}{2}\left(\cdots\right)\right)$, since that is the form of a normal distribution:

$$
\begin{aligned}
&\boldsymbol{\beta}_d\left(\boldsymbol{\alpha}_d\boldsymbol{I}_K\right)\boldsymbol{\beta}_d'-\boldsymbol{\nu}_d\boldsymbol{\beta}_d'+\boldsymbol{\beta}_d(\boldsymbol{\omega}_d\boldsymbol{I}_K)\boldsymbol{\beta}_d'+2\boldsymbol{\omega}_d(\log(\boldsymbol{r})\boldsymbol{I}_K)\boldsymbol{\beta}_d'+ \\
&\sigma\sum_y\left[\boldsymbol{\beta}_d(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\eta}_y'\boldsymbol{\eta}_y(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\beta}_d'-2z_{yd}\boldsymbol{\eta}_y(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\beta}_d'\right] \\
&=\boldsymbol{\beta}_d\left[(\boldsymbol{\alpha}_d\boldsymbol{I}_K)+(\boldsymbol{\omega}_d\boldsymbol{I}_K)+\sigma(\boldsymbol{\zeta}\boldsymbol{I}_K)\sum_y\left[\boldsymbol{\eta}_y'\boldsymbol{\eta}_y\right](\boldsymbol{\zeta}\boldsymbol{I}_K)\right]\boldsymbol{\beta}_d'- \\
&\quad 2\left[\sigma\sum_y\left[z_{yd}\boldsymbol{\eta}_y\right](\boldsymbol{\zeta}\boldsymbol{I}_K)-\boldsymbol{\omega}_d(\log(\boldsymbol{r})\boldsymbol{I}_K)+\frac{\boldsymbol{\nu}_d}{2}\right]\boldsymbol{\beta}_d'
\end{aligned}
$$

Comparing this expression with that of the multivariate-normal P.D.F. the expression for the conditional posterior is

$$
\boldsymbol{\beta}_d|\cdots\sim\mathcal{N}\left(\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d\right), \tag{B.17}
$$

where $\boldsymbol{\Sigma}_d^{-1}=\left[(\boldsymbol{\alpha}_d\boldsymbol{I}_K)+(\boldsymbol{\omega}_d\boldsymbol{I}_K)+\sigma(\boldsymbol{\zeta}\boldsymbol{I}_K)\sum_y\left[\boldsymbol{\eta}_y'\boldsymbol{\eta}_y\right](\boldsymbol{\zeta}\boldsymbol{I}_K)\right]$ and $\boldsymbol{\mu}_d=\left[\sigma\sum_y\left[z_{yd}\boldsymbol{\eta}_y\right](\boldsymbol{\zeta}\boldsymbol{I}_K)-\boldsymbol{\omega}_d(\log(\boldsymbol{r})\boldsymbol{I}_K)+\frac{\boldsymbol{\nu}_d}{2}\right]\boldsymbol{\Sigma}_d$.

Sampling of $\boldsymbol{\zeta}$

The derivation of the conditional posterior for $\boldsymbol{\zeta}$ is very similar to that of $\boldsymbol{\beta}_d$. The conditional posterior distribution has the same form as a multivariate-normal distribution. Integrate out r_k as follows. Since $x_{..k} \sim \text{Poisson}\left(r_k \theta_{.k}\right)$ and $r_k \sim \text{Gamma}\left(\gamma_0/K, \exp\left(\zeta_k\right)\right)$, then from the gamma-Poisson construction of the negative binomial described in Section 2.3.4, $x_{..k} \sim \text{NB}\left(\gamma_0/K, p_k\right)$, where $p_k = \frac{\theta_{.k}}{exp - \zeta_k + \theta_{.k}}$. This term in the posterior has the form:

$$
\begin{aligned}
p(x_{..k}|\cdots) &= \text{NB}\left(\gamma_0/K, p_k\right) \\
&\propto p_k^{x_{..k}} \left(1 - p_k\right)^{\gamma_0/K} \\
&= \left(\frac{\theta_{.k}}{\exp\left(-\zeta_k\right) + \theta_{.k}}\right)^{x_{..k}} \left(1 - \frac{\theta_{.k}}{\exp\left(-\zeta_k\right) + \theta_{.k}}\right)^{\gamma_0/K} \\
&= \frac{\left(\theta_{.k} \exp\left(\zeta_k\right)\right)^{x_{..k}}}{\left(\theta_{.k} \exp\left(\zeta_k\right) + 1\right)^{x_{..k} + \gamma_0/K}} \\
&= \frac{\exp\left(\psi_k\right)^{x_{..k}}}{\left(\exp\left(\psi_k\right) + 1\right)^{\kappa_k}},
\end{aligned}
$$

where $\psi_k = \log(\theta_{.k}) + \zeta_k$ and $\kappa_k = x_{..k} + \gamma_0/K$. To handle this factor in the posterior the Pólya-Gamma augmentation lemma 2.3.8 is used. Introduce $\omega_k \sim \text{PG}(\kappa_k, 0)$, and apply lemma 2.3.8, for which the corresponding factor becomes proportional to the following:

$$
\exp\left(\frac{(x_{..k} - \gamma_0/K)\psi_k}{2}\right) \int_0^\infty \exp\left(-\omega_k \psi_k^2/2\right) p(\omega_k) d\omega_k
$$

By lemma 2.3.8, the posterior update of the augmented Pólya-Gamma variable is given by:

$$
\omega_k | \cdots \sim \text{PG}(\kappa_k, \psi_k). \tag{B.18}
$$

Combining this with the prior and z_{yd} contributed likelihood, the posterior has the form:

$$
\begin{aligned}
&p(\boldsymbol{\zeta}|\cdots) \propto \\
&\exp\left(-\frac{1}{2}\boldsymbol{\zeta}(\boldsymbol{\alpha}_2\boldsymbol{I}_K)\boldsymbol{\zeta}'\right)\prod_k \exp\left((x_{..k}-\gamma_0/K)/2\zeta_k - \omega_k\left(log(\theta_{.k})+\zeta_k\right)^2/2\right)\cdot \\
&\quad\prod_{y,d}\exp\left(-\frac{\sigma}{2}\left(z_{yd}-\boldsymbol{\eta}_y(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\beta}_d'\right)^2\right) \\
&= \exp\Bigg(-\frac{1}{2}\Bigg[\boldsymbol{\zeta}(\boldsymbol{\alpha}_2\boldsymbol{I}_K)\boldsymbol{\zeta}' - \sum_k\left[(x_{..k}-\gamma_0/K)\zeta_k - \omega_k\left(\zeta_k^2+2\log(\theta_{.k})\zeta_k\right)\right]+ \\
&\quad\sigma\sum_{y,d}\left(\boldsymbol{\eta}_y(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\beta}_d'\boldsymbol{\eta}_y(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\beta}_d' - 2z_{yd}\boldsymbol{\eta}_y(\boldsymbol{\zeta}\boldsymbol{I}_K)\boldsymbol{\beta}_d'\right)\Bigg]\Bigg)
\end{aligned}
$$

Let $\boldsymbol{\lambda} = \{x_{..k}-\gamma_0/K\}_{k=1}^K$ and let $\boldsymbol{\theta} = \{\theta_{.k}\}_{k=1}^K$. Once again, consider just the argument under the $\exp\left(-\frac{1}{2}(\cdots)\right)$, since that is the form of a normal distribution:

$$
\begin{aligned}
&\boldsymbol{\zeta}(\boldsymbol{\alpha}_2\boldsymbol{I}_K)\boldsymbol{\zeta}' - \boldsymbol{\lambda}\boldsymbol{\zeta}' + \boldsymbol{\zeta}(\boldsymbol{\omega}\boldsymbol{I}_K)\boldsymbol{\zeta}' + 2\log(\boldsymbol{\theta})(\boldsymbol{\omega}\boldsymbol{I}_K)\boldsymbol{\zeta}' + \\
&\quad\sigma\sum_{y,d}\left[\boldsymbol{\zeta}(\boldsymbol{\beta}_d\boldsymbol{I}_K)\boldsymbol{\eta}_y'\boldsymbol{\eta}_y(\boldsymbol{\beta}_d\boldsymbol{I}_K)\boldsymbol{\zeta}' - 2z_{yd}\boldsymbol{\eta}_y(\boldsymbol{\beta}_d\boldsymbol{I}_K)\boldsymbol{\zeta}'\right] \\
&= \boldsymbol{\zeta}\left[(\boldsymbol{\alpha}_2\boldsymbol{I}_K)+(\boldsymbol{\omega}\boldsymbol{I}_K)+\sigma\sum_{y,d}(\boldsymbol{\beta}_d\boldsymbol{I}_K)\boldsymbol{\eta}_y'\boldsymbol{\eta}_y(\boldsymbol{\beta}_d\boldsymbol{I}_K)\right]\boldsymbol{\zeta}' - \\
&\quad 2\left[\sigma\sum_{y,d}z_{yd}\boldsymbol{\eta}_y(\boldsymbol{\beta}_d\boldsymbol{I}_K) - \log(\boldsymbol{\theta})(\boldsymbol{\omega}\boldsymbol{I}_K)+\boldsymbol{\lambda}/2\right]\boldsymbol{\zeta}'
\end{aligned}
$$

Comparing this expression with that of the multivariate-normal P.D.F. the expression for the conditional posterior is

$$
\boldsymbol{\zeta}|\cdots \sim \mathcal{N}\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right), \tag{B.19}
$$

where $\boldsymbol{\Sigma}^{-1} = \left[(\boldsymbol{\alpha}_2\boldsymbol{I}_K)+(\boldsymbol{\omega}\boldsymbol{I}_K)+\sigma\sum_{y,d}(\boldsymbol{\beta}_d\boldsymbol{I}_K)\boldsymbol{\eta}_y'\boldsymbol{\eta}_y(\boldsymbol{\beta}_d\boldsymbol{I}_K)\right]$ and $\boldsymbol{\mu} = \left[\sigma\sum_{y,d}z_{yd}\boldsymbol{\eta}_y(\boldsymbol{\beta}_d\boldsymbol{I}_K) - \log(\boldsymbol{\theta})(\boldsymbol{\omega}\boldsymbol{I}_K)+\boldsymbol{\lambda}/2\right]\boldsymbol{\Sigma}$.

Bibliography

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

Timothy N. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012. ISSN 0885-6125. doi: 10.1007/s10994-011-5272-5. URL `http://dx.doi.org/10.1007/s10994-011-5272-5`.

Jon D. Mcauliffe and David M. Blei. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Inc., 2008. URL `http://papers.nips.cc/paper/3328-supervised-topic-models.pdf`.

Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1257–1264. ACM, 2009.

Jonathan Chang and David M Blei. Relational topic models for document networks. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, pages 81–88, 2009.

Ayan Acharya, Raymond J Mooney, and Joydeep Ghosh. Active multitask learning using both latent and supervised shared topics.

A Acharya, A Rawal, RJ Mooney, and ER Hruschka. Using both supervised and latent shared topics for multitask learning. *ECML PKDD, Part II, LNAI*, 8189: 369–384, 2013.

Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. Gibbs max-margin topic models with fast sampling algorithms. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 124–132, 2013.

Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. Beta-negative binomial process and poisson factor analysis. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 1462–1471, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of machine Learning research*, 3:993–1022, 2003.

David J Aldous. *Exchangeability and related topics*. Springer, 1985.

Bruno De Finetti, Antonio Machi, and Adrian Smith. *Theory of probability: a critical introductory treatment*. Wiley New York, 1990.

Andriy Mnih and Ruslan Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2007.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. IEEE, 2008.

Prem Gopalan, Jake M Hofman, and David M Blei. Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*, 2013.

Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001. URL `http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf`.

Ali Taylan Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009, 2009.

John 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. ACM, 2004.

Michalis K. Titsias. The infinite gamma-poisson feature model. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1513–1520. Curran Associates, Inc., 2008. URL `http://papers.nips.cc/paper/3309-the-infinite-gamma-poisson-feature-model.pdf`.

Prem Gopalan, Francisco JR Ruiz, Rajesh Ranganath, and David M Blei. Bayesian nonparametric poisson factorization for recommendation systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 275–283, 2014.

T. Broderick, L. Mackey, J. Paisley, and M.I. Jordan. Combinatorial clustering and the beta negative binomial process. *Pattern Analysis and Machine Intelligence,*

IEEE Transactions on, 37(2):290–306, Feb 2015. ISSN 0162-8828. doi: 10.1109/TPAMI.2014.2318721.

M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):307–320, Feb 2015. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.211.

Nils Lid Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. Nonparametric bayesian factor analysis for dynamic count matrices. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, San Diego, CA, May 2015. JMLR W&CP.

Mingyuan Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, San Diego, CA, May 2015. JMLR W&CP.

Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

Robert L Wolpert, Merlise A Clyde, Chong Tu, et al. Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *The Annals of Statistics*, 39(4):1916–1962, 2011.

John Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21 (1):59–78, 1967.

John Frank Charles Kingman. *Poisson processes*, volume 3. Oxford university press, 1992.

Mingyuan Zhou and Lawrence Carin. Augment-and-conquer negative binomial processes. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2546–2554. Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4677-augment-and-conquer-negative-binomial-processes.pdf`.

Norman L Johnson, Adrienne W Kemp, and Samuel Kotz. *Univariate discrete distributions*, volume 444. John Wiley & Sons, 2005.

Nicholas G Polson, Steven L Scott, et al. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

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.

Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Department of Computer Science, Royal Holloway, University of London, May 1998.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *The Journal of Machine Learning Research*, 8:1007–1025, 2007.

Peter D Hoff. *A first course in Bayesian statistical methods*. Springer, 2009.

Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006.

Ioana A Cosma and Ludger Evers. Markov chains and monte carlo methods. *African Institute for Mathematical Sciences, Cape Town*, 2010.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118108.1118117. URL `http://dx.doi.org/10.3115/1118108.1118117`.

Martin F Porter. Snowball: A language for stemming algorithms, 2001.

Conrad Sanderson. Armadillo: C++ linear algebra library, version 4.650. `http://arma.sourceforge.net/`, 2015.

M Galassi et al. Gnu scientific library reference manual , isbn 0954612078. *Library available online at http://www. gnu. org/software/gsl*, 2010.

John R Michael, William R Schucany, and Roy W Haas. Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2):88–90, 1976.