Models and Algorithms for Episodic Time-Series

Rafael Augusto Ferreira do Carmo

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Statistical Science

University College London

December 2018

I, Rafael Augusto Ferreira do Carmo, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This thesis is built on the idea of modeling episodes of multiple time series which can be briefly defined as multivariate time series whose individual dimensions vary in time and nature. This kind of data arises naturally when we observe repeatedly scenarios where collections of individual elements that may or may not take part in the collective observed behaviour. We illustrate the ideas constructed around this kind of data making use of datasets related to crowdfunding and video-on-demand. These datasets are prolonged periods of observation of these scenarios and provide natural examples to the ideas we develop. How to relate seemingly disconnected individual episodes and how to incorporate information from them into the general view of the multiple episodes is the main goal of this thesis. We focus on constructing this two-way flux so that even more complex models than the ones present in this work can be constructed using the proposed features. We describe models and algorithms that mix supervised and unsupervised tasks. Specifically, we construct models that connect Topic Models, unsupervised learning models that aim to summarize big corpora of texts with regression models on time series. We also discuss how summaries of past episodes may be helpfull in predicting future series of observations of same category.

Impact Statement

The ideas, expertise, analysis and insights present in this thesis can be put to a beneficial use both inside and outside academia. Inside academia, the studies on feature engineering topics for regression tasks can be extended to different scenarios in which this kind of latent variable is meaningful, leading to variations on the proposed models that can be directly explored and studied by undergraduate and master students willing to understand the proposed structure. Additionally, outside academia, the proposed models can be used in different crowdfunding portals in order to gain understanding of the market-level elements that drive donations to projects which then can be used in managing how to present open projects, helping creators understand what aspects are important for their projects to succeed and enhancing recommendation systems. In a more general sense, the studies proposed in this thesis for multiple episodic time series could guide future modeling of data whose statistical graphical models resemble the proposed models of this thesis.

Acknowledgements

This journey has been full of *episodes* of ups and downs and without the *multiple* help of different people in the different phases of it, I am totally sure I would not be able to successfully complete it. First and foremost, I am completely grateful by the supervision provided by Dr. Ricardo Silva. His openness to discussing every single doubt I have had, and they were plenty, no matter what period of the year or advancement of my studies where of invaluable importance to me and provided me with the tools to pursue a meaningful career in science. In a world where it is common for the relation supervisor-student to deteriorate, I am hopeful we will be able to continue working together in the near future. Also of very importance were my family. I own to my soul-mate Caroline Nobre probably a big piece of this work. Although thousands of kilometers far from London, she's managed to be integral part of my life thought this whole work, cheering me up in the difficulties and celebrating my achievements. Her care and love has brought me here. Mum and dad, my examples for life, have another piece of this work. They've worked hard to prepare me for this moment and I am sure they are proud of my and my sister's achievements. The Portuguese speaking gang, the Greeks, Italians, Chinese, Germans, Croatians and Tanzanians, all this micro-world that is the Statistical Department will be remembered as well, you guys made me laugh a lot.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

We are living in the age of data collection, where every single step of our digital lives is collected, stored, treated and used by entities that we even does not know they exist. Social networks make use of information related to one individual in order to construct models that can predict the behaviour of this individual's friends and even (seemingly) unrelated people. Search engines collect data on our quest for online shopping in order to try to advertise that brand new 4k TV you may have bought (but still gets the ad every single website you visit). Companies developing autonomous vehicles collect data on driving events of volunteers and employees in order to feed embed neural networks in cars that eventually will drive by themselves would be some examples of this massive collection of data that is part of our daily routine.

This scenario may make us concerned about privacy issues, about the limits on which third-parties can directly and indirectly get to know about our lives, but in the majority of times it performs two-way transactions that are beneficial to every part engaged in these transactions: individuals contribute with their individual actions, ideas and behaviors, to say a few, and in exchange get from the collection of other individuals contributions *trends* on important piece of news by discussions that show on the front page of online forums, the *discovery* of new music by the automatic examination of other people's "exquisite" taste for unknown bands, the *collective* action towards those in need via donations of money and goods that individually are not relevant but when made by a gigantic group of people (a crowd)

turns into something that is certainly relevant.

This thesis will revolve around ideas connected to such scenarios. We aim to describe, model and understand how what we observe in individuals in one scenario may affect the general view of all the individuals in this scenario and vice versa, how this general view can influence and be used in understanding the behavior of future individuals in this same scenario. What we call here individuals can be actors that willingly make part of a collection of actors (individuals) of passive elements which are composition of actions by other actors in the collection. Also, we try to understand these relations in unstable environments, in the sense that the set of individuals may vary in time and their active time - time in which they are part of the collection - may vary as well.

All these philosophical ideas are brought to real life in scenarios where these two-way transactions seem to be part of the dynamics of them, while it is unclear how they relate. We study the market of Crowdfunding[2], the ever alternative way of funding independent projects in several areas of knowledge, with its increasing number of projects, increasing number of individuals donating money and creating both niche products and ideas that are relevant to a greater public. We also study YouTube, the number one online video-on-demand platform, whose catalog of content can only be sorted and summarized automatically with models constructed on the behavior of people watching a only small fraction of this catalog.

We construct statistical models in these scenarios making use of Graphical Models[3] formalism, which allow us to easily simulate data and make use of generic algorithms such as the Expectation Maximization algorithm[4] in a very easy and straightforward way in order to make inference and estimation tasks in the statistical models proposed. We also rely heavily on Topic Models[1], a beautiful and important model that easily summarizes and explains in very low dimensional features the very high and sparse textual data that is found in large corpora of texts. Finally, given the discussed temporal and pervasive flow of information from individuals to the collection of them, we make use of classical time series models that are adapted to the nature of data we observe in crowdfunding and video-on-demand

datasets.

1.1 Thesis outline

The remainder of this thesis is constructed as follows: Chapter 2 deals with the main general concepts used in this thesis, presenting relevant literature to support the ideas built in the following chapters. It discusses relevant statistical models to the thesis such as topic models, time series and latent variable models and presents the Expectation-Maximization algorithm and its Variational Bayes variations that are used and adapted throughout the whole thesis.

Chapter 3 illustrates the main ideas concerning the connection this thesis makes to different elements of the literature present in chapter 2, constructing a latent state-space model that connect topics and time series through episodes of multiple time series, which is illustrated in a dataset composed of crowdfunding data.

Chapter 4 extends the work of the previous chapter in order to accommodate different elements of the market that were not taken into consideration in chapter 3 while presenting variations of the inference and estimation algorithm that can handle the new elements of the enhanced model. These elements allow us to have a better picture of individual and collective information related to crowdfunding projects.

Chapter 5 presents a different view of the elements related to time series and latent variables discussed in this work, where we aim to make use of individual episodes of time series as inputs that can impact a whole community of interests. We illustrate such ideas using a dataset of YouTube videos.

Chapter 6 summarizes the whole work and introduces future discussions and improvements to the present work and additional results and discussions are shown in Appendix. Throughout this thesis we going to make use of the following symbols:

Symbol	Explanation
$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$	Inner product of vectors x and y
$\boldsymbol{x}\otimes\boldsymbol{y}$	Outer product of vectors x and y
$\boldsymbol{x}\odot\boldsymbol{y}$	Element-wise products of vectors x and y, $\boldsymbol{z}=[x_1y_1,x_2,y_2,...]$
M'	Transposition of matrix M
$a..b$	Refers to an interval $[a,a+1,a+2,...,b]$
$M=[\boldsymbol{x},\boldsymbol{y}]$	Construction of matrix M by joining the column vectors x and y
$\sigma(x)$	sigma function $\sigma(x)={}^{1}/_{1+\exp(-x)}$

Table 1.1: List of Symbols

Chapter 2

Literature review

This chapter provides a comprehensive foundation on the models, methods and techniques that are important to the development of this thesis. Each component is explained in a self-contained session and the structure of the discussion in this chapter is taken to the following parts of the thesis, where the concepts here presented are used in the context of the applications studied.

In a Statistical Inference setting, we are usually faced with quantities of interest that involve integrating over a subset of random variables of the ones in hand. For instance, given the vector of random variables $\boldsymbol{z} = [\boldsymbol{x}, \boldsymbol{y}]$ composed by the concatenation of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ that we are using to model a problem and the vector $\boldsymbol{y} = [y_1.y_2, ..., y_n]$ is observable while $\boldsymbol{x} = [x_1, x_2, ..., x_m]$ is unobservable (latent), we may be interested in evaluating parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_o]$ related to this model in a maximum likelihood (ML) fashion. In order to do so, we are are required to evaluate the equation

$$\boldsymbol{\theta}^{\star} = \arg\max_{\boldsymbol{\theta}} \int_{\mathscr{X}} p(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x}$$

on which $\theta^{\star}$ is the ML estimate of the parameters of the model and $\mathscr{X}$ is the domain of integration (domain of the variables $\boldsymbol{x}$). On the other hand, in a maximum *a posteriori* (MAP) setting or full Bayesian setting, we must deal with different equations such as

$$\boldsymbol{\theta}^{\star} = \arg\max_{\boldsymbol{\theta}} \int_{\mathscr{X}} p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{x}$$

on which $p(\boldsymbol{\theta})$ refers to the prior distribution over $\boldsymbol{\theta}$. Finally, in a full Bayesian setting, we are interested in the posterior distribution of random variables of interest, which in this case would be $\boldsymbol{\theta}$ while marginalizing (integrating) over some possibly nuisance variables, leading to[1]

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\int_{\mathscr{X}} p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{x}}{\int_{\mathscr{X},\Theta} p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{x} d\boldsymbol{\theta}}$$

All these equations express some of the basic elements found in statistical modeling and for all them, we may face difficulties of different nature in the process of their evaluation. One common problem is the impossibility of evaluating the integrals via algebraic expressions, which can occur even in simple calculations such as

$$\int \log\left(\frac{1}{1+\exp-(\theta_0+\theta_1 x_1)}\right) p(\theta_0,\theta_1) d\theta_0 d\theta_1$$

which relates to a simple Logistic Regression with prior distribution on the parameters of the model[5].

So far we have talked about distributions over random variables but we have not tried to encode any conditional dependence structure among them, which is commonly done via Probabilistic Graphical Models[6] (PGM). PGMs make use of graph formalism to encode these conditional dependence structures. Graphs can be defined as mathematical elements describing pairwise relationships between entities. A graphs G is described as $\mathrm{G} = (\mathscr{V},\mathscr{E})$ where $\mathscr{V} = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes (entities) represented and $\mathscr{E} = \{(v_a, v_b) | v_a \in \mathscr{V} \wedge v_b \in \mathscr{V}\}$ is the set edges or of pairwise relationship between entities. These edges can be undirected, in which the relationship has no direction, or directed, in which the relationship has a direction from one vertex to the other, commonly from the first vertex to the second in the pair, in our case from v_a to v_b. In the PGM setting, random variables are the

[1] From now on we are skipping the domain of the integral wherever it is clear to the reader.

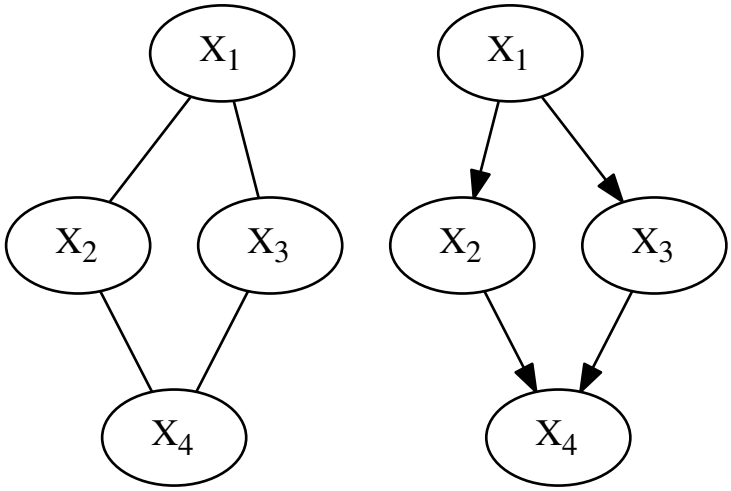


Figure 2.1: Markov Network (left) and Belief Network (right)

nodes and direct dependencies are represented by the edges.

Markov networks, also know as Markov Random Fields, represent distributions in undirected graphs via *cliques*, which are subsets of nodes such that there is an edge connecting every pair of nodes. Let us suppose we have four random variables for which we are constructing a graphical model. One possible Markov network representation is the one shown in the graph on the left of Figure 2.1. The joint probability distribution of these for variables represented by this graph could be written

$$p(x_1,x_2,x_3,x_4) = \frac{1}{Z}\phi(x_1,x_2)\phi(x_1,x_3)\phi(x_2,x_4)\phi(x_3,x_4)$$

where $Z = \int \phi(x_1,x_2)\phi(x_1,x_3)\phi(x_2,x_4)\phi(x_3,x_4)dx_1dx_2dx_3dx_4$ is the normalizing constant for the distribution, ensuring the it integrates to 1. Generally speaking, every ϕ function describes a form of coupling between the variables belonging to the clique it represents.

All the models constructed in this thesis are based in belief networks, also known as Bayesian networks or probabilistic directed acyclic graphical models,

which are models for representing sets of random variables and conditional dependencies among them via a directed acyclic graph, i.e., the modeler decides to directly encode a generative schema to the observation of the random variables belonging to the model. This kind of decision is usually made due to prior knowledge of the domain of the problem under study or to impose constraints that allow better interpretability of the models or ensure better computational costs to the inference and estimation processes.

Let us study the belief network described by the graph on the right of Figure 2.1. Its joint probability distribution can be written as

$$p(x_1,x_2,x_3,x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2,x_3)$$

where the joint is decomposed into different conditional probability distributions. This representation is more suitable for generative models, algorithms that describe the hypothetical natural process that generates data.

These two representations possess distinct features and belief networks can be transformed into Markov random fields via the moralization algorithm[6]. They are also the main representation of graphical models, but alternatives are also possible and provide suitable representations to different sets of problems. Mixed graphs construct joint probability distributions by making use of both directed and undirected connections between random variables. Cumulative distribution networks [7] take a different approach. While maintaining undirected graph representation, they encode the functions in the cumulative probability distribution functions space instead of the usual probability density space. When picturing these graphical models, it is common to use white nodes representing latent variables, grey nodes representing observed variables and labeled rectangles defining groups of variables that repeat in the model the number of times the label of the rectangle express.

2.1 Variational inference

Variational inference (VI) is a general deterministic approximation to intractable integrals or expectations which appear in these complex models[8, 9]. In this work,

we construct complex high-dimensional graphical models, for which direct Markov chain Monte Carlo (MCMC) is practical only in small instances of the proposed models. To avoid this enormous computational costs, we make use of VI procedures in exchange of cruder approximations to the distributions of interest. Specifically in this thesis, we make an explicit separation between model parameters $\boldsymbol{\theta}$ and latent random variables $\boldsymbol{x}$ so MCMC techniques would fit only as one component of the learning procedure, in a Monte Carlo Expectation Maximization fashion [10].

Let us return to the setting where our model of interest is composed of the random variables $\boldsymbol{x}$ which are latent, the observed random variables $\boldsymbol{y}$ and model parameters $\boldsymbol{\theta}$[2]. We are interested in estimating parameters based on the marginal likelihood of the observed variables $\boldsymbol{y}$ in a graphical model also containing the latent variables $\boldsymbol{x}$. By using Jensen's inequality, we can approximate such marginal via

$$\log p(\boldsymbol{y};\boldsymbol{\theta}) = \log \int p(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x} \geq$$

$$\int \log p(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta}) q(\boldsymbol{x}) d\boldsymbol{x} - \int q(\boldsymbol{x}) \log q(\boldsymbol{x}) d\boldsymbol{x}$$

$$\log p(\boldsymbol{y};\boldsymbol{\theta}) \geq \mathrm{E}[\log p(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})]_{q(\boldsymbol{x})} - \mathrm{E}[\log q(\boldsymbol{x})]_{q(\boldsymbol{x})}$$

$$= \mathrm{E}[\log p(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})]_{q(\boldsymbol{x})} + \mathrm{H}[q(\boldsymbol{x})]$$

where the construction of this lower bound is the root of variational inference methods. This bound is constructed using the auxiliary distribution $q(\boldsymbol{x})$, $\mathrm{E}[f(\boldsymbol{x})]_{q(\boldsymbol{x})}$ refers to the expected value of function $f(\boldsymbol{x})$ under the distribution $q(\boldsymbol{x})$ (we may use the shorthand version $< f(\boldsymbol{x}) >$ wherever it is clear which distribution we are taking the expectation from) and $\mathrm{H}[q(\boldsymbol{x})]$ refers to the differential entropy of the $q(\boldsymbol{x})$ distribution. This approximation is called the Evidence Lower Bound (ELBO) and provides an optimal approximation (in terms of KL-Divergence) to the desired log-marginal likelihood $\log p(\boldsymbol{y})$. Equality is achieved at $q(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y})$, which is

[2]So far, we are making use of *frequentist* terminology, making a clear distinction between random variables and model parameters but the discussion presented may be brought to the *Bayesian* world if we understand the latent variables as also the parameters of the model and the model parameters as hyperparameters that we may be interested in optimizing. We will follow the frequentist point of view but it is important to understand that they are similar in a broad sense

usually intractable to compute.

Given this bound, we must find a tractable representation of the distribution of interest. A tractable approximation to this posterior is obtained by forcing $q(\boldsymbol{x})$ to impose extra independence constraints than the ones implied by original graphical model. As we have freedom to choose $q(\boldsymbol{x})$ it can be shown that the problem turns into an optimization of the ELBO within a space of tractable distributions and if we split the set $\boldsymbol{x}$ of latent variables into disjoint sets $\boldsymbol{x} = [x_1, x_2, ..., x_n]$ where $x_i \cap x_j = \emptyset, \forall i \neq j$ and $n \leq |x|$, optimizing the ELBO regarding to $q(x_i)$ can be performed via coordinate ascent algorithm where every step is

$$\log q(x_i) \propto \mathrm{E}[\log p(\boldsymbol{y}, x_i, x_j)]_{q(\boldsymbol{x}_{i-})} \tag{2.1}$$

where $\boldsymbol{x}_{i-} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$ is the distribution of all the sets of variables in x but x_i. Depending on the complete distribution under study, the direct form of 2.1 provides a direct parametric form of $q(x_i)$ which can then be used as the variational distribution to the random variable x_i. Usually, this is not the case and one fixes the parametric form of the distributions, i.e. Gaussian and Dirichlet, and directly optimize their parameters via maximizing the ELBO. When picking the extra independence constraints to the latent variables, one can go extreme and enforce total independence $n = |x|$, which is called *mean-field* variational inference or impose less restrictive independence constraints, in a *structured* way [11].

2.1.1 Stochastic variational inference

Usually, when using VI procedures, one may develop model-based algorithms, taking into consideration all the unique elements of the graphical model proposed, the latent and observed variables and so on. This model dependent approach is effective although costly and generic software cannot be constructed in order to be reused to different models and instances.

In order to construct reusable and generic algorithms to handle different graphical models, several approaches make use of stochastic versions of variational inference in order to approximate distributions of interest. It may seem counter intuitive

to resort on sampling-based approaches to approximate distributions but stochastic variational inference and MCMC (and other sampling-based) methods are very different in nature. In short, MCMC methods construct a Markov chain whose equilibrium distribution is the distribution of interest, usually the posterior of the set of latent variables given the observed values, while stochastic variational inference departs from a given fixed-form q distribution and stochastically optimize its parameters through an interative process of sampling from q, evaluating the noisy gradient of the ELBO and optimizing the parameters of the distributions accordingly.

Let us return to the motivating model we are discussing from the beginning of the chapter, now assuming that, for some reason, we are not able to evaluate in closed-form the ELBO of the model and we define $q(x;\lambda)$ the variational distribution of the latent variables which is parameterized by λ, we can theoretically optimize λ via a first-order gradient descent method with α as learning rate and

$$\lambda_{i+1} = \lambda_i + \alpha \times \frac{\partial f(\lambda_i)}{\partial \lambda_i} \tag{2.2}$$

where

$$\begin{aligned} f(\lambda) &\approx \mathrm{E}[\log p(y,x;\theta)]_{q(x;\lambda)} + \mathrm{H}[q(x;\lambda)] \\ &= \frac{1}{N}\sum_{j=1}^{N} \left\{\log p(y,x^j;\theta) - \log q(x^j;\lambda)\right\} \text{ and } x^j \sim q(x;\lambda) \end{aligned}$$

being f a sampling approximation of the ELBO of the model. Unfortunately, the variance of this approximation is usually high making the simple first-order procedure unreliable. Different approaches try to tackle this problem in several different ways, be it Rao-Blackwalization or control variates [12], neural networks as estimators of the variables of the model [13], mini-batches of data [14] and reparametrization tricks [15]. In equation 2.2 we present a simple first-order stochastic gradient descent method but different approaches which adapt this basic equation are common and try to optimize the performance of such algorithm [16]

2.1.2 Variational expectation maximization

Throughout this thesis we are going to make use of Variational Expectation Maximization algorithms so it is appropriate to explain this approach in greater detail. Expectation maximization (EM) is a general algorithm designed as an iterative method to estimating parameters of a statistical model in a maximum likelihood fashion when there are latent random variables involved in the model [4]. The basic EM algorithm consists of alternating two separate steps, the called E-step which constructs the posterior distribution of the latent variables given the observations and the parameters in a fixed value and the called M-Step which optimizes the parameters of the model making use of the expected marginal log likelihood given the posterior distribution constructed in the previous E-Step.

In our explanation, let us use the same terminology used so far: let us assume a model composed of a vector $\boldsymbol{z} = [\boldsymbol{y}, \boldsymbol{x}]$ of random variables composed of the concatenation of the vectors $\boldsymbol{y} = [y_1.y_2, ..., y_n]$ of observable variables, $\boldsymbol{x} = [x_1, x_2, ..., x_m]$ of latent ones and the parameters vector $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_o]$. Let us also assume that evaluating $p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta})$ is not feasible and we must make use of variational distribution $q(\boldsymbol{x})$ in order to approximate this distribution of interest. Let us also assume that we impose some structure in this q distribution, i.e., we split the vector $\boldsymbol{x}$ into m' groups $q(x) = q(x^1)q(x^2)...q(x^{m'})$ which are composed of non-overlapping subsets of the elements of $\boldsymbol{x}$ - $x^1 = [x_1, x_3, x_5]$ for example -, $(m > m')$ such that the posterior distribution of each separate group is independent of each other while the variables within each group keep dependence among them.

Algorithm 1 Variational Expectation Maximization

Require: Initial guesses $\boldsymbol{\theta}^\star$ and $q(x)$
  while termination criteria not met **do**
    for $i = 1$ to m' **do** ▷ E-Step
      Update $q(\boldsymbol{x}^i) = \arg\max_{q(\boldsymbol{x})^i} \mathrm{E}[\log p(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}^\star)]_{q(\boldsymbol{x})} + \mathrm{H}[q(\boldsymbol{x})]$
    end for
    $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \mathrm{E}[\log p(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})]_{q(\boldsymbol{x})}$ ▷ M-Step
  end while
  return $\boldsymbol{\theta}^\star$ and $q(x)$

In this setting, the algorithm runs as expressed in Algorithm 1. We must pro-

vide initial guesses to both the variational distribution and the parameters of the model. Then we perform the E-Step in which every subset of latent variables has its distribution updated according to the optimization of the ELBO. This is done in a coordinate-ascent fashion until all the components of the distribution is updates. After that in the M-Step the parameters $\boldsymbol{\theta}$ of the model are updated though the optimization regarding only the expected log-likelihood of the model, given that the entropy part of the ELBO does not relate to the model parameters. The algorithm does not force any termination criteria but number of iterations or norm of the update of the model parameters are usual criteria. This procedure is guaranteed to achieve a *local maximum* of the log-likelihood function and must be run with several initial values for the parameters and distributions such that convergence is tested.

2.2 Topic models

Topic models (TM) are a class of mixture models for discrete data, where each mixture component describes a distribution over a possible set of discrete outcomes, it is "a branch" of latent Dirichlet allocation generative models[1], where each mixture component is itself random, following a Dirichlet prior. Topic models are generative statistical tools that allow sets of high dimensional observations (texts) to be explained by lower dimensional latent groups (topics). The idea behind this generative model in the context of text data is that topics define distributions over vocabulary, and texts are generated via a choice of topics proportions and words picked in the different topics. The generative process may be written as

where τ and α are model parameters on the Dirichlet priors of per-topic word distribution and per-document topic distributions, respectively.

Topics can be defined as sets of distributions over vocabularies. In this sense, a vocabulary is a set of individual unique words and a topic is a Dirichlet distribution placed over this set. Then, a text is composed by a mixture of words sampled from a set of topics, in an unstructured (bag-of-words) process [17]. An example of a text viewed in a Topic Model sense can be seen in Figure 2.2 (from the seminal paper

Algorithm 2 Basic Topic Model Generative Model

Require: model parameters $\boldsymbol{\tau}, \boldsymbol{\alpha}$
  for all Topic k **do**
    Sample $\boldsymbol{\beta}_k \sim$ Dirichlet$(\boldsymbol{\tau})$
  end for
  for all Text document p **do**
    Sample topic proportion $\boldsymbol{\theta}_p \sim$ Dirichlet$(\boldsymbol{\alpha})$
    for all Word w_i in the text **do**
      Sample topic allocation $z_{i,p} \sim$ Multinomial$(1, \boldsymbol{\theta_p})$
      Sample word $w_{i,p} \sim$ Multinomial$(1, \boldsymbol{\beta}_{z_{i,p}})$
    end for
  end for

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 2.2: Topics and a text - Source [1]

[1]).

The ability that Topic Models have to summarize big corpora has led to the development of several variations over the basic model. Dynamic Topic Models [18, 19], whose graphical description can be seen at Figure 2.3, are constructed to accommodate topic proportions and topic distributions that evolve over time. Such dependency is performed by transforming $\boldsymbol{\alpha}$ into a random variable and modeling both variables topics $\boldsymbol{\beta}$ and topic proportion prior $\boldsymbol{\alpha}$ as chains of multivariate

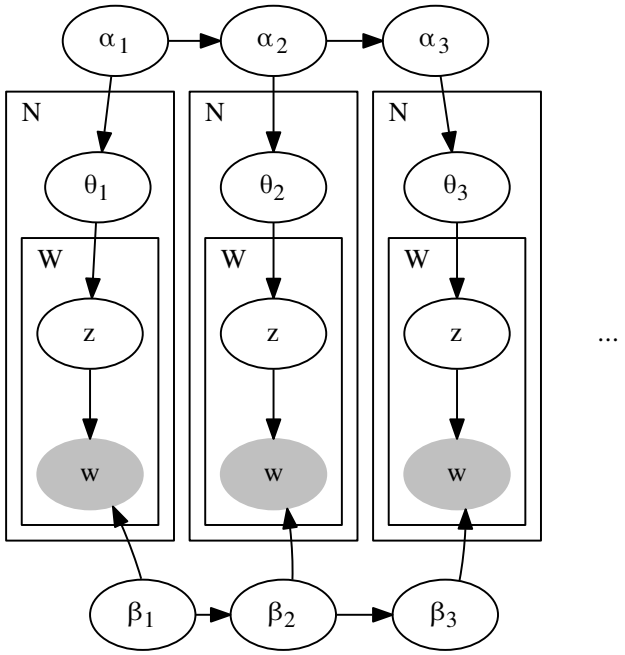


Figure 2.3: Dynamic Topic Model Graphical Description

normal distributions that evolve via a Markov process. Dynamic Topic models are meant to model sets of texts which rapidly change in vocabulary or which include data from a large spam of time ([18] uses 120 years of documents from *Science* magazine) .[20, 21] tries to relax the strict negative correlation between topics in a text by replacing the Dirichlet distribution these topic proportions are sampled by the logistic normal distribution, which then can express a much richer set of correlations. Also, other works [22, 23] try construct approximations to the learning procedure of topic models that are less expensive than full MCMC or variational inference while maintaining "good enough" (up to the practitioner) estimates.

All Topic Models variations discussed so far focus on constructing elements that allow flexible expression of quantities related to texts and topics but several different models try to add numerical observations based on the textual and topical variables of Topic Models. [24] constructs Supervised Topic Models which aim to model univariate numeric variables connected to texts such as movie ratings (pre-

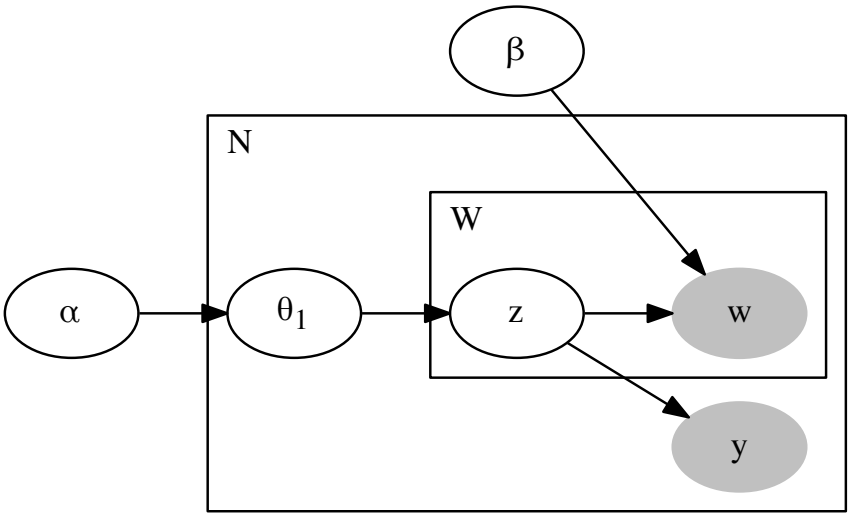


Figure 2.4: Supervised Topic Model Graphical Description

dicted from reviews) and site popularity [24]. It defines a vector $\bar{z}$ which is the empirical frequencies of topics in a given document to use as independent variables in a generalized linear regression step.

[25] adds numeric elements to Topic Models in a different sense, it tries to model the impact of scholarly documents in future documents via an adaptation of a basic Dynamic Topic model. In a DTM, topics (represented by the $\boldsymbol{\beta}$ variable in our discussion) of time $t+1$ depend on topics of time t. [25] adds more components in this dependence by adding regression components using document's words and topic allocations and also a random effect component, which it calls an *influence score*.

[26, 27] go in a different direction. Departing from basic Dynamic Topic Model representation as well, they connect numerical variables to the $\boldsymbol{\alpha}$ variables (instantaneous topic distribution priors) aiming to model a general relationship between texts existing in a given time-point and a general perception (numeric response) over a given subject. Both papers study the models using finance related datasets.

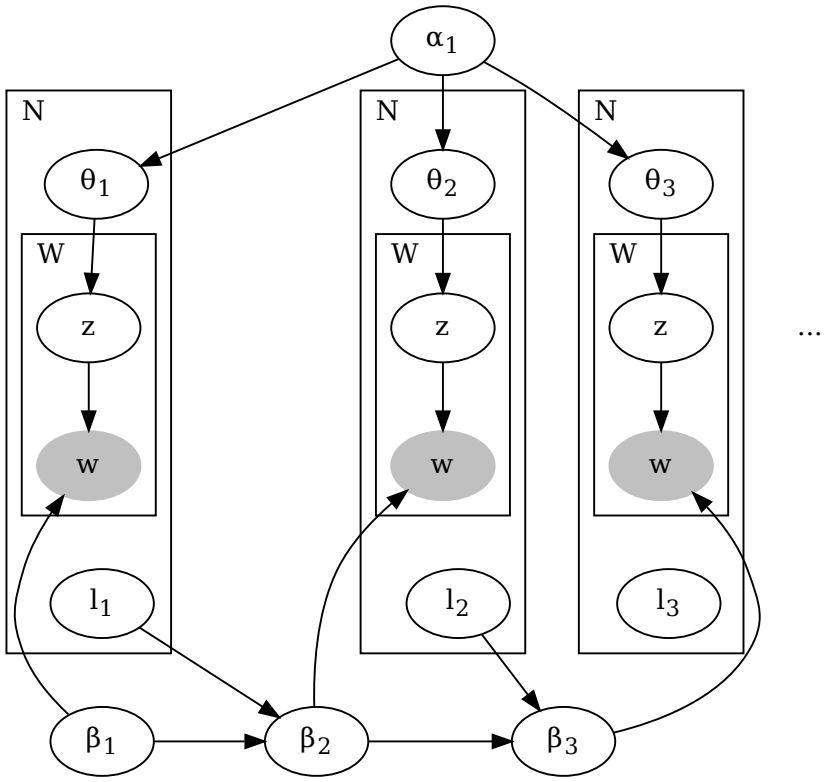


Figure 2.5: Document Influence Topic Model Graphical Description

[28] presents supervised topic models related to classification tasks and discuss concepts on the difficulties of learning topic models and doing classification simultaneously. While these works go in the direction of regressing numerical variables based on textual information, [29, 30] go in the opposite direction, making use of text meta-data as inputs that influence word distribution in texts but [30] does not follow Topic Models theory but construct similar low-dimensional representations.

2.3 Latent state-space models

Time-series Models are another important area to the development of this thesis. We are going to make use of Latent State-Space Models (LSSM) which are the workhorse of an enormous variety of models in different fields such as signal processing and econometric studies.

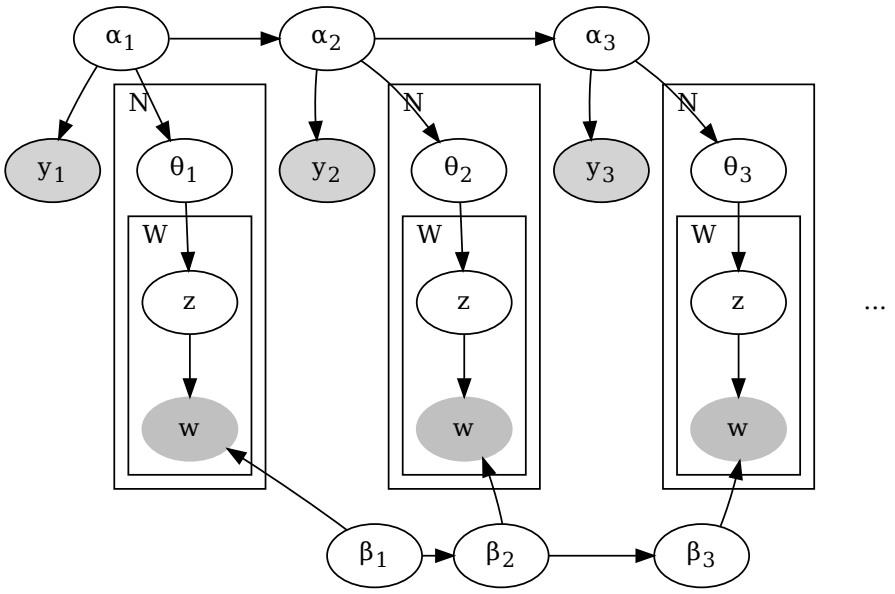


Figure 2.6: Supervised Dynamic Topic Model Graphical Description

LSSM provide a framework which assumes the observed sequence was generated from an underlying sequence of continuous latent states that possesses the Markov property. For a sequence of states $x_{1:T} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ in which every state is a vector $\boldsymbol{x}_i = [x_{i,1}, \ldots, x_{i,n}]$ composed of n elements and a set of observations $y_{1:T} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$ in which every observation is a vector $\boldsymbol{y}_i = [y_{i,1}, \ldots, y_{i,m}]$ composed of m elements (usually $m \geq n$), we may write its generative model as

Algorithm 3 Basic Latent State-Space Generative Model

Require: Model parameters θ
  Sample $x_1 \sim p(x_1; \theta)$
  Sample $y_1 \sim p(y_1|x_1; \theta)$
  for $t = 2$ to T **do**
    Sample $x_t \sim p(x_t|x_{t-1}; \theta)$
    Sample $y_t \sim p(y_t|x_t; \theta)$
  end for
  return $x_{1:T}$ and $y_{1:T}$

which provides conditional independence for the observations $y_{1:T} \perp\!\!\!\perp x_{1:T}$. One possible interpretation of this modeling is that it aims to construct a smoother process in x - compared to the one observed in y - that is capable of expressing rich sets

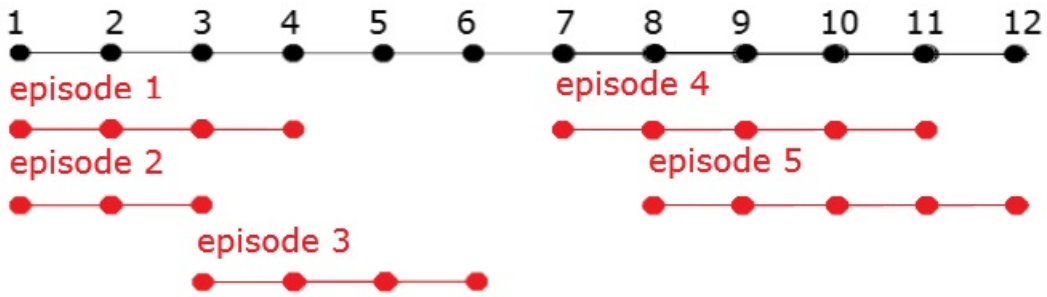


Figure 2.7: Example of multiple episodic time series

of observations in y. Whenever the parameterization for this system is fully Gaussian, the computation of quantities of interest such as the posterior distribution of the latent variables is facilitated via message parsing algorithms which constructs the distribution via a dynamic programming schema, that can be performed due to the sequential (tree-shaped) form of the belief network of the LSSM. Kalman filters [31] and its variations are some of the most traditional algorithms to dealing with this kind of system. As said previously, usually the observations at every time-point are of fixed size, i.e., at two different time-points y_i and y_j, both elements will have same dimensionality.

In the models developed in this thesis we make use of relaxations of these characteristics, via what we call episodic and multiple time series. Episodic time series can be defined as time series that, given a window of time on which we are observing a phenomena, occur in subsets of this window. Multiple time series [32] refer to the fact that in a given time-point, there might be a different number of time series occurring simultaneously. These concepts may be better explained by examining Figure 2.7. There, we observe a time window of 12 points in which there is the occurrence of 6 episodic time series. At different time points there is a variable number of multiple time series being observed, ranging from one time series up to three simultaneously. Each episode may be of different nature, like independent crowdfunding projects or episodes of the same nature, disciplines that occur in a course every term and whose length is variable, for example.

2.4 Auxiliary concepts

2.4.1 Sentiment Analysis

Sentiment analysis is the task of unveiling subjective ideas such as opinions, evaluations, and attitudes of people about specific elements [33] usually presented in textual format [34]. Taking as example a set of reviews of a product on a website, the task of sentiment analysis may be defined as the task of processing this corpus of texts, making use of Natural Language Processing (NLP) tools in order to extract linguistic resources such as the vocabulary and the structure of the text such that these elements are then used in extracting the sentiment expressed in the corpora. By sentiment, we can define the general perception (positive/negative - classification task) or a numerical value (0 to 5 stars - ordinal regression task) belonging to a whole document level or in each of its sentences [35]. Among the ideas presented in the literature, [36, 37] have deeper connections to the proposed work. Both pieces of works, while mainly focusing Twitter data, try to relate the sentiments expressed in tweets to different observations.

[36] constructs a time-varying sentiment score, which is a daily ratio of positive versus negative messages regarding one topic. Keywords are split into two different groups, positive $+$ and negative $-$) and the sentiment score of a topic k in a given day t is a positive value

$$x_{k,t} = \frac{\text{count}(+\text{words})}{\text{count}(-\text{words})}$$

assuming that the count of negative words is positive. Due to the daily volatility of this score, this amount is smoothed using a fixed-size moving average, and these smoothed observations are utilized in a linear model to regress the output of traditional public opinion polls.

[37] projects tweets into what it call *mood vector*, a six-dimensional vector of ratios by counting the number of terms of tweets that can represent six different dimensions of mood, named *Tension*, *Depression*, *Anger*, *Vigour*, *Fatigue*, and *Confusion*, as measured via the Profile of Mood State (POMS) psychometric in-

strument. Once again, due to the variability the number of tweets in a given day and consequently the score, they perform a z-score transformation to the sequential score observations and then perform a comparison of their dynamics to the happening of marked events during the studied period.

Works on sentiment analysis are an important area with different findings, and the field of understanding time-dynamics of sentiments is still an open problem [38, 39].

2.4.2 Information Diffusion

One field closed related to sentiment analysis is the field of information diffusion, which contemplates the idea of modeling the *the process by which a piece of information (knowledge) is spread and reaches individuals through interactions* [40], which requires the idea of senders, receivers and a media which allows these interactions.

Depending on the observability of these interactions, this diffusion can be classified into four different types [40]. When the network of connections between individuals is important for the process, the diffusion may be of the type *herd behavior* which occurs when all individuals observe other one's behavior, i.e., there is a complete graph or *information cascade*, in which individuals only observe their immediate neighbors. When dealing with unobserved/unexisting network, one may refer to *diffusion of innovations* in which only the total volume of the spread of an element is observed and *epidemic* in which one individual does not decide whether joining the process or not. Definitions may vary a little, but when dealing with online social networks specifically, the main elements are also existing [41].

It is of interest to model the dynamics in which information evolves. Works in this field usually set a reasonable window in which the diffusion might evolve and construct different functions to approximate the observed dynamics. In a continuous time setting, [42] proposes a flexible family of functions to model the propagation of news in social media, for cases in which one knows the number of participants of the network. Several different works try to model such evolution but also including the network (and its evolution) on which the information diffuses [43, 44, 45].

2.5 Summary

In this chapter we present the basic elements that compose the puzzle of this thesis: latent variables, topic models, episodic and multiple time series and auxiliary concepts, elements connected to sentiment analysis and information diffusion, that allow all of these areas to be consistently glued together. Our main interest in this work is to study the dynamic of sentiments and information attached to latent variables, such as topics, given the observation of multiple episodic time series whose observations of single episodes are apparently unrelated to each other. In order to perform inference and estimation in models constructed by these ideas, via graphical models - especially belief networks, we make use of Variational Inference, a reliable and fast inference - although not optimal in terms of accuracy - procedure for graphical models in the presence of latent random variables.

Chapter 3

Topics based latent state-space model for crowdfunding data

In this chapter, we showcase ideas related to collections of multiple time series. In order to do so, we present the market of crowdfunding and propose an algorithmic approach to the problem of modeling the amount of money donated to projects throughout time and assessing the general state of the market to these projects. Unlike existing methods, the proposed approach makes use of time-dependent latent features derived from the textual description of the projects as explanatory variables of project success. These features capture the current importance donors give to the different topics addressed by existing projects. The experiments on this paper show empirically the importance of inferring latent information in the regression model we use, improving its performance and making a clear contribution to the explanation of the observed data. The proposed approach connects topic models which model the descriptions of projects to state-space time-series models which describes the dynamics of donations to projects.

Online platforms such as Kickstarter and Indiegogo have amplified the range and impact of crowdfunding projects around the world. The removal of geographic barriers between independent entrepreneurs and a multitude of possible donors (the crowd) enables the funding of a larger range of possible projects compared to traditional markets, a novel kind of exchange that is still not fully understood. Such a market has gained much interest from the general public and the scientific com-

munity, which aims to understand the dynamics of these projects and to create tools that help creators to maximize the odds of success of their enterprises.

Crowdfunding in general and its specific features in Kickstarter are gaining significant interest from the Machine Learning, Statistics and Business research communities recently. Exploratory papers[46, 47] discuss and measure the correlation in between different descriptors of projects in the crowdfunding market and their likelihood of success. Several other papers try to understand the dynamics of the lifetime of a project and its final outcome. In the Machine Learning and Statistics literature, we can cite [48] constructs models using k-Nearest Neighbours and Markov Chains as tools for classifying projects outcomes, [49] makes use of decision trees to model the likelihood of new backers (donors) to projects, [50] uses features collected mainly from social media sources (Facebook and YouTube) in their Support Vector Machine model and others [51, 52, 51, 53] showcase the necessity of understanding and modeling this emerging market.

Kickstarter is one of the world's largest crowdfunding companies[1]. It works as a platform for both advertising and supporting independent projects. Crowdfunding is a growing option for entrepreneurs to gain access to resources they require to develop their projects, which is characterized by collecting small funding contributions from a large group of donors/investors. The donors then gain rewards and access to privileges on the early development of a product or a service.

Projects on Kickstarter are usually drafts of ideas of a product or a service which independent entrepreneurs do not have enough resources to develop. The entrepreneurs who create projects, called *creators*, have to set a period in which their projects will be available for donations. After this period, if the project gathers at least the amount requested, then the creator of the project receives the total money pledged. If not, then donations are sent back to the donors, here called *backers*. This works as a safety regulation for both creators and backers, by first making sure creators will have at least the amount of money they planned for their projects, and by mitigating the chance that backers waste their money on projects that will not be

[1] https://www.kickstarter.com/help/stats

finalized. There are however several other platforms with variations of this rule in the market[2].

Kickstarter allows for projects to be on their website by up to 60 days. Creators describe their projects textually, including images and videos, also providing a list of rewards to boost the possibility of high-value donations. Creators are allowed to keep updating their projects throughout the project's lifetime and afterward.

3.1 Model Definition

We assemble all the previous ideas to come up with a model that 1. predicts the amount of money a project will receive in a given time-window; 2. predicts the likelihood of success of a project within a few steps into the future; and 3. explain these predictions in terms of the "popularity" of particular topics as they evolve in time. A discrete-time process is used.

In order to achieve these, the model takes into consideration time-dependencies and latent factors related to the topics of the projects. Topics are inferred using topic models, and extra latent factors are introduced to account for the degree of attention a topic is receiving at any given time. We call these latent time-dependent factors "topic heats." The motivation for introducing these factors is illustrated in the context of movie projects as follows: there may be periods in which people are primarily interested in projects that involve cinema and environmental questions, but in other periods of time the mix could be cinema and politics. These "interests" are not directly recorded in the data, but we indirectly capture them by modeling on-going dependencies between the amount of money people donate to projects and the topics inferred from the (e.g. Kickstarter) web pages of the projects. These are direct analogies to the sentiment (positive/negative) and a suitable/interpretable measure for situational interest of a crowd to the topics describing projects. We are unable to observe any contagious that may bring more people to make donations and so we have a setting of diffusion of innovations.

In the following, let p index any particular project and let t index time. Given a

[2]http://marketingmoxie.biz/the-big-list-of-crowdfunding-sites/

predefined number K of topics $\{\beta_1, \ldots, \beta_K\}$, let $\theta_{p,k}$ be the corresponding k-th topic proportions of p, regardless of time, and $\alpha_{k,t}$ be the topic heat for topic k at time t. We denote as α_t the vector formed by $[\alpha_{1,t} \ldots \alpha_{K,t}]$. Let $z_{i,p}$ and $w_{i,p}$ be the topic allocation and word for position i in project p as in a standard topic model.

By following this process we have that the random variables "topic heats" are sampled via the sampling procedure

$$\boldsymbol{\alpha}_t \sim \text{Normal}(A\boldsymbol{\alpha}_{t-1}, I)$$

which defines a multivariate Normal first-order Markov chain with parameters A for the load matrix and I for the covariance matrix of the multivariate normal distribution. Based on Section 2.2, we remember that the topic proportion of each text is sampled via

$$\boldsymbol{\theta}_p \sim \text{Dirichlet}(\boldsymbol{\eta})$$

and along with $\boldsymbol{\alpha}_t$ variables, these elements compose a form of latent state-space model for the observations of y.

Finally, let $c_{p,t}$ and $y_{p,t}$ be, respectively, fixed covariates (such as the amount pledged by the project) and donations received (in e.g. dollars) for project p at time t. Projects start and end at different time-points, with the fixed covariates and the times of birth/death of a project assumed to be given instead of random.

Given all the elements $\boldsymbol{\alpha}_t$, $\boldsymbol{\theta}_p$ and the possibly time-dependent covariates $c_{p,t}$, the observed $y_{p,t}$ variables are sampled via a two-step hurdle model. First, an auxiliary $y^\star_{p,t}$ normal random variable is sampled. If the value of $y^\star_{p,t}$ is greater than zero, another normally distributed random variable is sampled and then $y_{p,t}$ value is defined. All this can be summarized in the generative algorithmic procedure 4 and diagram 3.1.

In Algorithm 4 all new symbols are model parameters, $\lambda_{y^\star}$, λ_y, $\lambda_{c^\star}$ and λ_c are vectors of parameters and $\rho_{y^\star}$ and ρ_y are intercept elements. By project active at time t, we mean any project p which is open to receiving donations at time-point t.

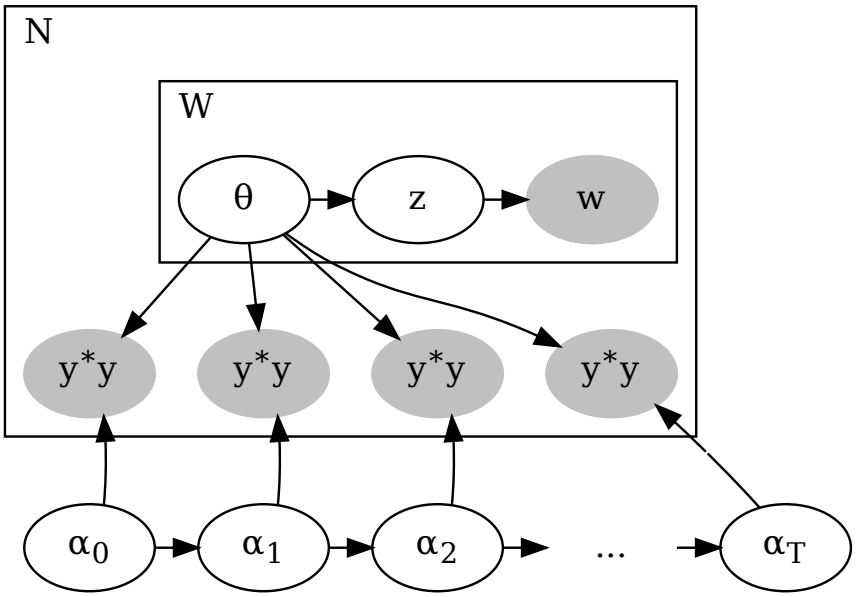


Figure 3.1: Graphical description of the proposed Model

Algorithm 4 Topics Based Latent State-Space Model for Crowdfunding

Require: model parameters $\boldsymbol{\tau}, \boldsymbol{\eta}, A$, all λ and ρ
  for all Topic k **do**
    Sample $\boldsymbol{\beta}_k \sim \text{Dirichlet}(\boldsymbol{\tau})$
  end for
  for all Project description p **do**
    Sample topic proportion $\boldsymbol{\theta}_p \sim \text{Dirichlet}(\boldsymbol{\eta})$
    for all Word slot i **do**
      Sample topic allocation $z_{i,p} \sim \text{Multinomial}(1, \boldsymbol{\theta_p})$
      Sample word $w_{i,p} \sim \text{Multinomial}(1, \boldsymbol{\beta}_{z_{i,p}})$
    end for
  end for
  for all time-point t **do**
    Sample $\alpha_t \sim \text{Normal}(A\alpha_{t-1}, I)$
    for all project p active at time t **do**
      evaluate $m^{\star}_{p,t} = \lambda^{\mathrm{T}}_{y^{\star}}(\theta_p \odot \alpha_t) + \rho^{\star}_y + \lambda^{\star\mathrm{T}}_c c_{p,t}$
      $n_{p,t} = \lambda^{\mathrm{T}}_y(\theta_p \odot \alpha_t) + \rho_y + \lambda^{\mathrm{T}}_c c_{p,t}$
      Sample $y_{p,t}$ according to the hurdle model definition with parameters $(m^{\star}_{p,t}, n_{p,t}, \delta_y)$
    end for
  end for

As said before, projects can last up to 60 days on Kickstarter and for different time-points, there will be a different number of projects running. Inference in our model means capturing this information of variable dimensionality at time t, reducing it to the fixed-size latent elements, and transferring such information across time.

Hurdle Model

Our definition of a hurdle model is based on a two-stage model that defines a distribution on non-negative variables. In our case, each variable Y is continuous for $Y > 0$ but with a positive probability for the event $Y = 0$. The mixture component that generates the choice between $Y = 0$ and $Y > 0$ is given by a model for Bernoulli outcomes based on the sign of a latent Gaussian variable. If the sign of the latent Gaussian is positive, this is followed by generating a numeric positive value following a log-Normal distribution:

$$y^\star \sim N(m^\star, 1)\ ,\ y = 0 \text{ if } y^\star \leq 0 \text{ else } \exp(z) \tag{3.1}$$

where $z \sim N(n, \delta)$. This model is going to be used to model the amount of money pledged for a given project p at time t. $m^\star$ and n are the location parameters of the Normal distributions and in this work they are going to be random variables defined accordingly. Specifically in this work, the hurdle model we propose is composed of the following variables and parameters. The first element $y^\star_{p,t}$ is sampled via

$$y^\star_{p,t} \sim \text{Normal}(\lambda^{\text{T}}_{y^\star}(\theta_p \odot \alpha_t) + \rho^\star_y + \lambda^{\star\text{T}}_c c_{p,t}, 1)$$

which is a normal distribution with fixed unitary variance and mean defined by the sum of a bias term $\rho^\star_y$, the attached covariate vector $c_{p,t}$ and their respective parameters $\lambda^\star_c$ plus the element-wise product of the project's topic proportion and the instantaneous topic heat values $\theta_p \odot \alpha_t$ times their respective parameters $\lambda_{y^\star}$. Provided that $y^\star_{p,t}$ sampled value is greater than zero, a new sampling for the $y_{p,t}$ value occurs via the equation

$$y_{p,t} \sim \text{Normal}(\lambda^{\text{T}}_y(\theta_p \odot \alpha_t) + \rho_y + \lambda^{\text{T}}_c c_{p,t}, \delta^2_y)$$

which stands for a different normal distribution with variance parameterized by δ_y^2 and similar construction to the mean element, having only the parameter set λ_y, ρ_y and λ_c being different.

To finish the definition of the model, let F be the full set of projects, N_p the length of the text description of project p, A_t the set of active projects at time t, and $1:T$ the whole history of observations. We then define the **complete log-likelihood** of the model as

$$\begin{aligned}\ell(\eta,\lambda,\rho,\delta) = \sum_{p\in F}[\log p(\theta_p;\eta)+ \\ \sum_{n=1}^{N_p}\log p(z_{p,n}|\theta_p)+\log p(w_{p,n}|z_{p,n})\Bigg]+ \\ \log p(\alpha_1)+\sum_{t=2}^{T}\log p(\alpha_t|\alpha_{t-1};\lambda_\alpha,\delta_\alpha)+ \\ \sum_{t=1}^{T}\sum_{p\in A_t}\log p(y_{p,t},y^\star_{p,t}|\theta_p,\alpha_t;\lambda_{y^\star},\rho_{y^\star},\lambda_{c^\star},\lambda_y,\rho_y,\lambda_c,\delta_y).\end{aligned}$$

This assumes and conditions on the idea that topics $\{\beta_1,\ldots,\beta_K\}$ have been predefined by first fitting the standard variational latent Dirichlet allocation algorithm of [1] which can either be done with the text of all projects or a separate set of projects, which was the solution used in this paper due to the availability of such separate set and preliminary tests that showed the need for such conditioning. Further details are discussed in the following sections.

We must complete the definition of the model by stating that the first α variables starting at time 1 follow $p(\alpha_1) = N(0,I)$ where I is the identity matrix. For each element k of α_t, we define the evolution of the independent chain $p(\alpha_{k,t}|\alpha_{k,t-1}) = N(\lambda_{\alpha,k}\times\alpha_{k,t-1},\delta_{\alpha,k})$ and, as shown, we also borrow the idea of time-varying elements but model the description of projects using traditional TM.

3.2 Inference and Estimation

Given the definition of the complete model and the characteristics of it, we turn our focus to defining the procedures for inference of the latent variables and estimation of the unknown parameters of the model.

The key quantity of interest is the posterior distribution of the latent variable, including topic heats α_t. Unfortunately this posterior is intractable to compute due to the non-linearity of the observation distribution in the time-series part of the model and to the Dirichlet structure of the TM. On the top of that, the parameters of the model are unknown and must be estimated from data. To obtain these quantities we develop a Variational Bayes Expectation-Maximization (VBEM) algorithm [54] in which a *structured* approximation to the posterior distribution is considered:

$$\log p(\theta, y^\star, \alpha, z|y, w) \approx q(\alpha_{1:T}) \prod_{p\in F} q(\theta_p) q(z_p) \prod_{t=1}^{T} \prod_{p\in A_t} q(y^\star_{p,t})$$

By doing so, we maintain the temporal dependency among the topic heats, preventing the loss of crucial temporal dependency of these latent variables. This structure and the Gaussianity of the explicit dependency of y and $y^\star$ on α allows us to perform exact (given the structure defined) forward-backward passes to infer the variational parameters of $q(\alpha)$ in a similar way to the Variational Kalman Smoother (VKM) algorithm [55].

We provide a thorough explanation of the VBEM algorithm starting by describing the more complicated E-Step and following the M-Step, which is straightforward to derive and makes use of expectations of the latent variables as replacements for their actual values. The explanation of the message-parsing algorithm is highly based on [55] but the differences in the model due to the existence of different random variables are stressed.

3.2.1 Topic heat variational distribution

In order to derive the variational distribution of the α variables, we must focus on equation 2.1. Bringing it to the model under scrutiny, we have that:

$$\log q(\alpha_{1:T}) \propto \mathrm{E}\left[\log p(\alpha_1) + \sum_{t=2}^{T} \log p(\alpha_t|\alpha_{t-1}) + \sum_{t=1}^{T}\sum_{p\in A_t} \log p(y_{p,t}, y^{\star}_{p,t}|\theta_p, \alpha_t)\right]_{q(\theta_p)q(y^{\star}_{p,t})} \tag{3.2}$$

The given approximation allows us to make use of forward-backward messages to calculate the marginal variational distributions $q(\alpha_t)$ and pairwise ones $q(\alpha_t, \alpha_{t-1})$, adapting the VKM algorithm, given that the only dependency and expectation taken in this equation takes place in the elements of $\log p(y_{p,t}, y^{\star}_{p,t}|\theta_p, \alpha_t)$, where expectations of the θ and $y^{\star}$ variational distributions are taken. We briefly explain the message parsing schema, focusing that the major differences of it to the algorithm presented in [55] are that instead of taking expectations with respect to the parameters of the model, we take expectations on the values of $y^{\star}$ and θ variables and the emission component of the model contains two parts.

Also, $\log p(y_{p,t}, y^{\star}_{p,t}|\theta_p, \alpha_t) = \log p(y^{\star}_{p,t}|\theta_p, \alpha_t)$ when $y^{\star}_{p,t} < 0$ and $y_{p,t} = 0$, namely $y_{p,t}$ is not random in this case. We make this clear so that we can perform the derivations without having to explicit this fact.

Although unorthodox, the following derivation general ideas in Variational Inference and Latent State-Space Models. Considering $\setminus\alpha$ the set of latent variables different than α, the best approximation to the sequence $\alpha_{1:T}$ as[3]

[3]usually the derivation takes into consideration a single latent variable but it is straightforward to prove its equivalence to sets of variables

$$\log q(\alpha_{1:T}) \propto \mathrm{E}\left[\log p(\alpha_{1:T}, y^{\star}_{1:T}, y_{1:T})\right]_{q(\setminus\alpha)}$$

$$\log q(\alpha_{1:T}) = \log p(\alpha_{1:T}) + \mathrm{E}\left[\sum_{t=1}^{T}\sum_{p\in A_T}\log p(y^{\star}_{p,t}, y_{p,t}|\theta_p, \alpha_t)\right]_{q(\setminus\alpha)} \tag{3.3}$$

$$\log q(\alpha_{1:T}) = \log p(\alpha_{1:T}) + \mathrm{E}\left[\sum_{t=1}^{T}\sum_{p\in A_t}\log p(y^{\star}_{p,t}|\theta_p, \alpha_t) + \sum_{t=1}^{T}\sum_{p\in A_{t+}}\log p(y_{p,t}|\theta_p, \alpha_t)\right]_{q(\setminus\alpha)}$$

where A_{t+} is the set of projects with non-zero $y_{p,u}$ observations. With this in hand, we can perform traditional message parsing algorithms in the log-space of the latent state-space model to capture the uni and bivariate marginals of interest $q(\alpha_t)$ and $q(\alpha_t, \alpha_{t+1})$. To derive the marginals, we can define

$$\log q(\alpha_t) = f(\alpha_t) + b(\alpha_t) + \text{const., where}$$

$$f(\alpha_t) = \int \log p(\alpha_t|\alpha_{t-1}) + \mathrm{E}\left[\log p(y^{\star}_{t-1}|\theta_{t-1}, \alpha_{t-1}) + \log p(y_{t-1}|\theta_{t-1}, \alpha_{t-1})\right]_{q(\setminus\alpha)} \mathrm{d}\alpha_{t-1} + \mathrm{E}\left[\log p(y^{\star}_t|\theta_t, \alpha_t) + \log p(y_t|\theta_t, \alpha_t)\right]_{q(\setminus\alpha)} \tag{3.4}$$

and

$$b(\alpha_t) = \int \log p(\alpha_{t+1}|\alpha_t) + \mathrm{E}\left[\sum_{u=t+1}^{T}\log p(y^{\star}_u|\theta_u, \alpha_u) + \sum_{u=t+1}^{T}\log p(y_u|\theta_u, \alpha_u)\right]_{q(\setminus\alpha)} \mathrm{d}\alpha_{t+1:T}$$

where we called $y^{\star}_t$ the sets $\{y^{\star}_{p,t} \forall p \in A_t\}$ to condense the notation (same for y_t and

$theta_t$). If we pay close attention, we can realize that the messages are basically constructed by the composition of three Normal distributions, $p(\alpha_{1:t-1}|y_{1:t-1})$, $p(\alpha_t|y_t)$ and $p(\alpha_{t+1:T}|y_{t+1:T})$ which are composed via $p(\alpha_t : \alpha_{t-1})$ and $p(\alpha_{t+1}|\alpha_t)$. These messages are in turn of same complexity of the Kalman Smoother recursive equations in the complete Gaussian setting. In the following subsections, we are going to develop these messages and distributions.

Forward messages

Given the definitions in 3.4, we can then develop the message-parsing algorithm. For the first time-point $t = 1$, the forward message can be seen as:

$$f(\alpha_1) = \log p(\alpha_1) + \mathrm{E}\left[\log p(y_1^\star|\theta_1,\alpha_1) + \log p(y_1|\theta_1,\alpha_1)\right]$$

$$\begin{aligned} f(\alpha_1) &= \log N(\alpha_1, 0, I) + \\ &\left[\sum_{p\in A_1} \log N(y^\star_{p,1}, m^\star_{p,1}, 1) + \sum_{p\in A_{1+}} \log N(y_{p,1}, n_{p,1}, \delta_y)\right]_{q(\setminus\alpha)} \end{aligned} \tag{3.5}$$

from which we can derive that

$$f(\alpha_1) = \log N(\alpha_1, \mu_1, \Sigma_1) \text{ where } \Sigma_1 = (A_1 + I)^{-1} \text{ and } \mu_1 = \Sigma_1 b_1$$

where the matrices A_1 and b_1, and also all other elements for general time-points t's are constructed:

$$\begin{aligned} A_t = (\lambda_y \otimes \lambda_y) \odot \sum_{p\in A_{t+}} &< \theta_p \otimes \theta_p >_{q(\theta_p)} / \delta_y^2 + \\ (\lambda_{y^\star} \otimes \lambda_{y^\star}) \odot \sum_{p\in A_t} &< \theta_p \otimes \theta_p >_{q(\theta_p)} \end{aligned} \tag{3.6}$$

and

$$
\begin{aligned}
b_t = \lambda_y \odot \sum_{p \in A_{t+}} < \theta_p > (y_{p,t} - (\lambda_c^T c_{p,t} + \rho_y) >)/\delta_y^2 + \\
\lambda_{y^\star} \odot \sum_{p \in A_t} < \theta_p > (< y_{p,t}^\star > - (\lambda_{c^\star}^T c_{p,t} + \rho_{y^\star}))
\end{aligned}
\tag{3.7}
$$

Having the base case of this recursion fixed, i.e., the elements of this message, the mean and covariance matrix, we can perform the general cases

$$
\begin{aligned}
f(\alpha_t) = \int f(\alpha_{t-1}) + \log p(\alpha_t | \alpha_{t-1}) d\alpha_{t-1} + \\
\mathrm{E}\left[\log p(y_t^\star | \theta_t, \alpha_t) + \log p(y_t | \theta_t, \alpha_t)\right]
\end{aligned}
$$

$$
\begin{aligned}
f(\alpha_t) = \int \log N(\alpha_{t-1}; \mu_1, \Sigma_1) + \log N(\alpha_t, \lambda_\alpha \alpha_{t-1}, I) d\alpha_{t-1} + \\
\left[\sum_{p \in A_t} \log N(y_{p,t}^\star, m_{p,t}^\star, 1) + \sum_{p \in A_{t+}} \log N(y_{p,t}, n_{p,t}, \delta_y) \right]_{q(\setminus \alpha)}
\end{aligned}
\tag{3.8}
$$

which in turn becomes

$$
f(\alpha_t) = \log N(\alpha_t; \mu_t, \Sigma_t) \text{ where}
$$

$$
\begin{aligned}
\Sigma_{t-1}^\star &= \left(\Sigma_{t-1}^{-1} + \lambda_\alpha^T \lambda_\alpha\right)^{-1} \\
\Sigma_t &= \left(A_t + I - \lambda_\alpha \Sigma_{t-1}^\star \lambda_\alpha^T\right)^{-1}
\end{aligned}
$$

$$
\mu_t = \Sigma_t \left(b_t + \lambda_\alpha \Sigma_{t-1}^\star \Sigma_{t-1}^{-1} \mu_{t-1}\right)
$$

This is the usual derivation of the VBKM as seen in the literature [55] and the basic difference is that the expectations of topic proportions θ are absorbed in the matrices A_t and vectors b_t.

Backward messages

The backward messages procedure follows the same scheme as previous equations and we follow the definition on 3.4. Starting with the definition that $b(\alpha_T) = 0$, we have

$$b(\alpha_t) = \int \log p(\alpha_{t+1}|\alpha_t) + \mathrm{E}\left[\sum_{u=t+1}^{T} \log p(y_u^\star|\theta_u, \alpha_u) + \sum_{u=t+1}^{T} \log p(y_u|\theta_u, \alpha_u)\right]_{q(\setminus\alpha)} \mathrm{d}\alpha_{t+1:T} \tag{3.9}$$

$$b(\alpha_t) = \int \log p(\alpha_{t+1}|\alpha_t) + b(\alpha_{t+1}) d\alpha_{t+1}$$

which give rises to:

$$\begin{aligned} \Psi_{t+1}^\star &= \left(A_{t+1} + I + \Psi_{t+1}^{-1}\right)^{-1} \\ \Psi_t &= \left(\lambda_\alpha^T \lambda_\alpha - \lambda_\alpha^T \Psi_{t+1}^\star \lambda_\alpha\right)^{-1} \end{aligned} \tag{3.10}$$

$$\eta_t = \Psi_t \lambda_\alpha^T \Psi_{t+1}^\star \left(b_{t+1} + \Psi_{t+1}^{-1} \eta_{t+1}\right)$$

and, by definition, $\Psi_T = 0$ and $\eta_T = 0$.

Marginals, Pairwise Covariance and Entropy

Given this message-parsing setting, the marginal distributions of α_t can be easily written as:

$$\begin{aligned} \Phi_t &= (\Sigma_t^{-1} + \Psi_t^{-1})^{-1} \\ \omega_t &= \Phi_t (\Sigma_t^{-1} \mu_t + \Psi_t^{-1} \eta_t)^{-1} \end{aligned} \tag{3.11}$$

The Markov structure of the α sequence allows us to calculate only the pair-

wise covariance matrices $\mathrm{Cov}[\alpha_t, \alpha_{t+1}]$ which are simply:

$$\begin{aligned} \Phi_{1,2} &= (I + \Psi_2^- 1)^- 1 \lambda^T (\Sigma_1^{-1} + \lambda^T \lambda - \lambda^T (I + \Psi_2^{-1})^{-1} \lambda)^{-1} \\ \Phi_{t,t+1} &= \Sigma_t^\star \lambda^T (A_t + I + \Psi_{t+1}^{-1} - \lambda \Sigma_t^\star \lambda^T)^{-1} \text{for } t > 1 \end{aligned} \tag{3.12}$$

Ending this derivation, we are able to explicit the entropy of the α distribution via:

$$q(\alpha_{1:T}) = q(\alpha_1) \prod_{t=2}^{T} q(\alpha_t | \alpha_{t-1}) = q(\alpha_1) \prod_{t=2}^{T} \frac{q(\alpha_t, \alpha_{t-1})}{q(\alpha_{t-1})} \tag{3.13}$$

of which we are only interested in the covariance matrices, which take the block matrix form of

$$\begin{pmatrix} \Phi_t & \Phi_{t,t+1} \\ \Phi'_{t,t+1} & \Phi_{t+1} \end{pmatrix}$$

making it easy to compute the entropy of the complete chain of latent variables.

3.2.2 Derivation of the other variables

The definition of the inference procedure for the auxiliary variables $q(y\star)$, $q(\theta)$ and $q(z)$ also need to be defined and they follow a straightforward and easy process. The Probit bit of the hurdle model we define in this word provides partial information about the states $y^\star_{p,t}$ given the observation of $y_{p,t}$. If $y_{p,t} = 0$, then $y^\star_{p,t}$ has got to be negative and it must be positive provided that $y_{p,t} > 0$. By joining this fact with equation 2.1 we observe that

$$\begin{aligned}
\log q(y^{\star}_{p,t}) &\propto \mathrm{E}\left[\log \mathbb{1}_{sign(y_{p,t})=sign(y^{\star}_{p,t})} N(y^{\star}_{p,t}, m_{p,t}, 1)\right] \\
q(y^{\star}_{p,t}) &\propto \mathbb{1}_{sign(y_{p,t})=sign(y^{\star}_{p,t})} N(y^{\star}_{p,t}, < m_{p,t} >, 1) \\
&= \begin{cases} rTN(y^{\star}_{p,t}, < m_{p,t} >, 1) \text{ if } y_{p,t} > 0 \\ lRN(y^{\star}_{p,t}, < m_{p,t} >, 1) \text{ if } y_{p,t} = 0 \end{cases}
\end{aligned} \tag{3.14}$$

where $\mathbb{1}$ is the indicator and *sign* is the signal function and rTN and lTN stand for right-truncated and left-truncated Normal distributions [56] (chapter 19), respectively. All of this is a direct derivation of Bayesian Probit Regression [57, 58].

We define the variational distributions of the topic proportions θ_p as Dirichlet with parameterization γ_p. In order to infer the γ variational parameters, we must optimize the fragment of the ELBO in which θ_p takes part. If we call V_p the time sequence in which project p is open, the objective function of the optimization procedure can be written as

$$q(\theta_p) = \frac{\prod_{i=1}^{K} \Gamma(\gamma_i)}{\Gamma(\sum_{i=1}^{K} \gamma_i)} \prod_{i=1}^{K} \theta_i^{\gamma_i - 1}$$

$$\begin{aligned}
\gamma_t = \arg\max_{\gamma} \sum_{n=1}^{N_p} &< \log p(z_{p,n}|\theta_p) >_{q(z_{p,n})q(\theta_p;\gamma)} + \\
\sum_{t \in V_p} &< \log p(y_{p,t}, y^{\star}_{p,t}|\theta_p, \alpha_t) >_{q(y^{\star}_{p,t})q(\alpha)q(\theta_p;\gamma)} - \\
&< \log q(\theta_p;\gamma) >_{q(\theta_p;\gamma)}
\end{aligned} \tag{3.15}$$

The distributions of the topic allocation variables z follow standard TM-based optimization. For a given project p and word w_n, if we parameterize q as a Multinomial distribution using ϕ, we can write the distribution of the n-th topic allocation as [1]

$$q(z_{p,n}) = \frac{\exp\{\log \beta_{w_n} + < \log \theta_p >\}}{\sum_{k=1}^{K} \exp\{\log \beta_{w_{n,k}} + < \log \theta_{p,k} >\}} \tag{3.16}$$

where k indexes the k-th element of the parameter vector.

3.2.3 M-Step

The derivation of the M-Step is straightforward. We start by defining the equation that optimizes λ_α parameters and follow by writing the equations that need to be optimized to update the other parameters.

Given the diagonal construction of the parameter matrix λ_α, each element of this diagonal matrix may be updated via:

$$\lambda_{\alpha_{i,i}} = \frac{\sum_{t=2}^{T} < \alpha_{t-1,(i,i)} \odot \alpha_{t,(i,i)} >}{\sum_{t=2}^{T} < \alpha_{t,(i,i)}^2 >} \tag{3.17}$$

The optimization of η follows

$$\eta = \arg\max_{\eta} (\eta - 1)^T \sum_{p \in F} < \log \theta_p > - |F| \left[\Gamma(\sum_{k=1}^{K} \eta_k) - \sum_{k=1}^{K} \Gamma(\eta_k) \right] \tag{3.18}$$

where Γ is the gamma function. To optimize $\lambda_\star$, $\rho_\star$ and $\lambda_{c^\star}$ we maximize

$$\lambda_\star, \rho_\star, \lambda_{c^\star} = \arg\max_{\lambda_\star, \rho_\star, \lambda_{c^\star}} \sum_{t=1}^{T} \sum_{p \in A_t} < \log p(y_{p,t}^\star | \theta_p, \alpha_t; \lambda_{y^\star}, \rho_{y^\star}, \lambda_{c^\star}) >_{q(\alpha_t) q(\theta_p) q(y_{p,t}^\star)} \tag{3.19}$$

and the procedure to optimize the parameters related to the y observations is analogous, one having to replace $y^\star$ elements for y and the set A_t for the set A_{t+}. These elements are optimized independently to δ_y, which is then optimized in closed form trivially.

3.2.3.1 Identifiability Issues

Due to the latent nature of the topic heats, their usage in the Hurdle part of the model turns out to be unidentifiable, unless we enforce constraints into the parameters domain. We enforce the parameters λ_y and $\lambda_{y^\star}$to be ≥ 0 by setting a constraint

in the optimization process. By doing so we define that the more "warm" a topic is, the more important it is to have great chunks of projects' definitions taken by that topic, and vice-versa, the "colder" a topic is at a given moment the less it is going to contribute for a project to obtain donations. Enforcing these restrictions is one of different ways to ensure identifiability and assure a meaningful model given that both these model parameters and the elements connected to them (topic proportion and topic heat) are latent unobserved quantities. Provided that these random variables were observed, it would not be necessary to enforce such constraints. A similar result could be achieved by enforcing the topic heat variables to be positive only (> 0) either by turning them into Log-Normal random variables or making use of the $\exp(\boldsymbol{\alpha})$ values in the expression shown in Algorithm 4. This solution is evaluated in Chapter 4. Connecting all the pieces of the model definition and inference and estimation procedures, the complete VBEM algorithm is present in Algorithm 5.

Algorithm 5 VBEM algorithm for learning model described in algorithm 4

initialize $q(\alpha)$, $q(\theta)$, $q(y^\star)$ and $q(z)$
initialize η, $\lambda_{y^\star}$, $\rho_{y^\star}$, $\lambda_{c^\star}$, λ_y, ρ_y, λ_c, δ_y
while not converged **do** ▷ VBEM convergence
    while not converged **do** ▷ VBE-Step convergence
        optimize $q(\alpha)$ according to 3.11 and 3.12
        optimize $q(\theta)$ according to 3.15 using L-BFGS-B
        optimize $q(y^\star)$ according to 3.14
        optimize $q(z)$ according to 3.16
    end while
    ▷ M-Step
    optimize η according to 3.18
    optimize λ_α according to 3.17
    optimize $\lambda_{y^\star}$, $\rho_{y^\star}$, $\lambda_{c^\star}$ according to 3.19
    optimize λ_y, ρ_y, λ_c, δ_y according to the adaptation to 3.19
end whilereturn all variational distributions and model parameters

For those elements in which there is no closed-form equation to perform the optimization, we make use of the L-BFGS-B algorithm, which is a quasi-Newton optimization method which allows the variables to be optimized to be constrained. In this work we implemented this solution using the Python language and the

Scipy/Numpy framework, along with a library of auto-differentiation. The EM algorithm guarantees only local-optima solutions and every single test was run different times until solutions converged to close-enough ELBO values.

3.3 Experiments and Results

For our experiments, we scraped a first dataset containing 100K projects from Kickstarter for which we used to construct the topics used in the modeling. We preprocessed the data and ended with 9086 different terms that stemmed, generating 2740 terms in total. These terms and these texts were used to construct the topics, which were then fed into the model and kept fixed. By doing so we condition the remaining of the model upon the expected values of the topics (β variables) as if they were observed. This procedure was defined and used after preliminary simulations in artificial data and runs using real data indicated the need for this sequential procedure and it provided a two-fold improvement in the learning process. Firstly, it improved the running time of the learning procedure greatly. Secondly, it was observed in the artificial data that the algorithm was unable to recover similar topics to the ones generated artificially due to what seemed to be an analogous characteristic to the one present in Generative Adversarial Networks (GAN's) [59] in which there is a generative and discriminative parts that compose a complete model. In our proposed model, the topics β played the role of the generative part and the y regression the discriminative elements. It was observed in these preliminary simulations that whenever the Variational EM algorithm was performed in the complete model that the construction of the generative topics and the utilization of them in the regression part (via the topic proportion variables) was not able to be performed satisfactorily specially due to the different nature of the unobserved components. Neither the topics constructed were not meaningful or similar to the artificial topics nor the generated regression let to stable estimations. It was not tested during the period of this work but it is expected that the feasibility of a joint scheme could be achieved had the learning procedure followed similar ideas to the learning schema proposed in [59].

The second and most important dataset was obtained throughout 7 contiguous months, from April 2014 to November 2014, in which we collected data of approximately 45K projects, which were collected regularly at every 12 hours to get snapshots of these projects. We collected only project-related features, such as *goal*, *duration*, *number of rewards* and textual description. We also constructed a time-varying feature which we call $\Delta_{p,t}$ that represents the scaling (unity-based $[0,1]$ normalization) of the duration of a project, e.g. a project p which starts at time-point 31 and ends at time-point 60 will have features $\Delta_{p,45} = 0.5$, $\Delta_{p,60} = 1$ and so on. This feature is added twice in the covariate set, one time in a square form, to simulate the U-shape format of the donations to projects observed in [60].

We evaluated the proposed model by separating the projects according to the categories defined by Kickstarter and by learning the model making use of half of the time-points and performing all the estimations on the projects that were active at this time cut. We fixed the number of topics K to 10 (picking the number of topics of a model is usually an ad-hoc task depending on the domain of the instances of the problem, although there are algorithms that automatically estimates an optimal number of topics [61]). We set the convergence criteria as the relative improvement of the ELBO, stopping the algorithm whenever a complete EM-Step does not improve the ELBO by 1% and set the same criteria plus a maximum of 50 iterations on the VBE-Step of the algorithm to maintain low computational costs.

Early experiments, not discussed in the following sections, tried to make the variational distributions of $\alpha_{1:T}$ variables independently in a mean-field way, i.e., $q(\alpha_{1:T}) = \prod_{t=1}^{T} q(\alpha_t)$, which turns the learning procedure to be less complex but resulted in subpar variational distributions and results. It was observed that the elements in $\log p(y_1^\star|\theta_1,\alpha_1)$ and $\log p(y_1|\theta_1,\alpha_1)$ densities placed much more information into the ELBO than the Markov dependency of the likelihood $\log p(\alpha_t|\alpha_{t-1})$ meaning that the α variables ended up working as "free parameters" that only provided more adjustment to the regression part of the ELBO while not giving any extra structure to the latent variables of meaningful information. By enforcing this dependency - which is a must in the current literature, the α elements did not pro-

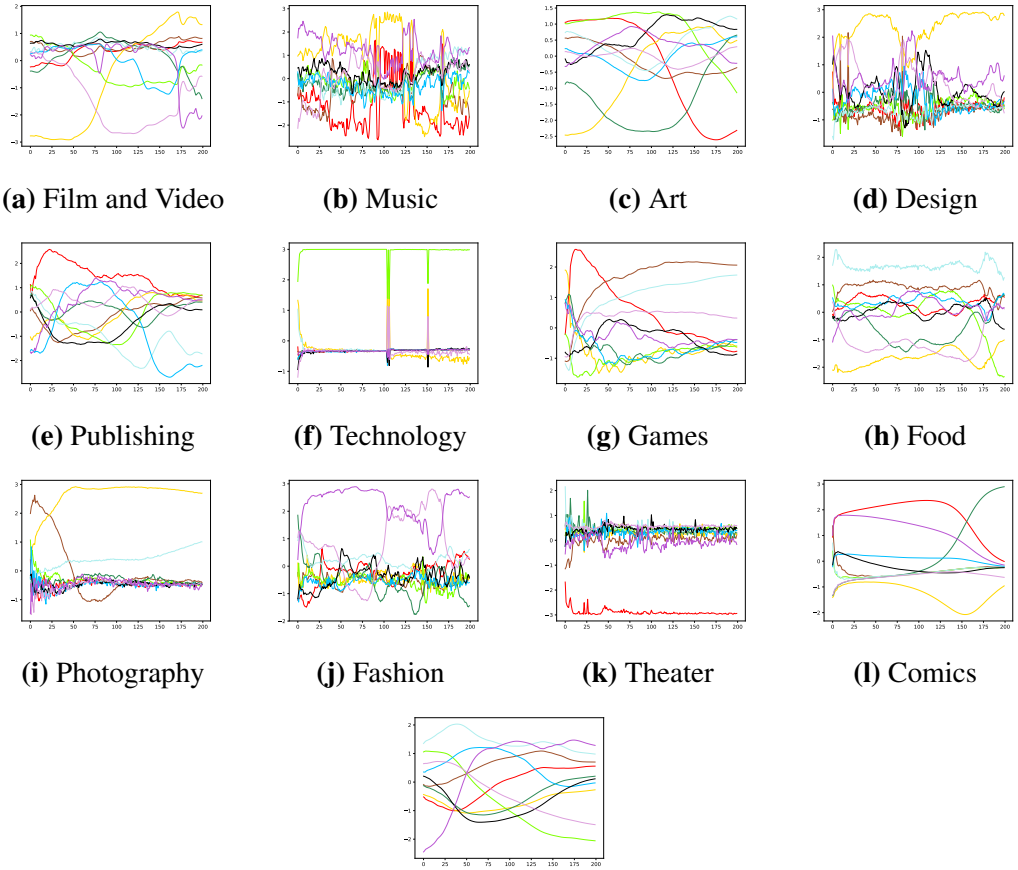
vide such tight adjustment but provided more structure to the data and by doing so, it was not necessarily acting as an adjustment to the y values of each time-point. Additionally, to this maintained structure the birth and death process characteristic of this data did not allow the topic heats to act as adjustments to the individual auto-correlation of y's of a given project.

3.3.1 Results

For each category, we present in 3.2 the scaled expected value of α given all donations (smoothing distribution) for every data point in the training dataset. We can interpret these graphical descriptions as follows: positive values for topic heats mean that projects containing a big chunk of text referring these topics will likely get more donations, while negative values for topic heats imply having big chunks of the descriptions devoted to these topics will negatively influence the likelihood of getting more donations.

With this understanding in hand, we observe some interesting relations in this figure. First of all, we observe a difference in the heat of the topics for each different category, which is a natural observation due to the diverse nature of these categories. For some categories, such as Art, Technology, Games, and Photography, there is a clear tendency of some topics having consistent more importance and others, while in Music and Comics there is a variability and change in the most important topics.

We then used the remaining part of the dataset to construct features for black-box algorithms in a test to understand if the topic heats variables add any sort of information to the individual projects. In order to do so, for those projects which started after the ending date existing in the training data, we estimated separately each topic proportion $\boldsymbol{\theta}_p$ using only the textual description of these projects. By doing so, we maintained a clear separation between the textual data and the numeric data that would be available only in the future, be it by design as in this testing procedure or if it was used by a manager in an online setting (data regarding donations coming every 12 hours). For the numeric data we performed forecast on the topic heat variables using a sliding window scheme. For a given time $t+1$, the expected values of α_{t+1} were estimated given the distribution of α_t and the ob-



(a) Film and Video **(b)** Music **(c)** Art **(d)** Design

(e) Publishing **(f)** Technology **(g)** Games **(h)** Food

(i) Photography **(j)** Fashion **(k)** Theater **(l)** Comics

(m) Dance

Figure 3.2: Scaled topic heat through time. (Best seen in color - Each color represents a specific topic)

served values $\boldsymbol{y}_t$. Then these expected values were used as inputs to estimate the numeric responses (donations) $\boldsymbol{y}_{t+1}$ and finally the distribution of topic heats α_{t+1} was updated given the $\boldsymbol{y}_{t+1}$ values in a filtering fashion. This process was repeated for every time-point t in the testing dataset.

Using the filtered distribution (*forward message* of the algorithm) for the $\boldsymbol{\alpha}$ variables of the remaining 200 time-points, we inputted along with the other previous covariates a new vector of covariates $< \boldsymbol{\theta}_p \odot \boldsymbol{\alpha}_t >$ in order to predict every observation $y_{p,t}$. For this regression task we use the original covariates in the linear regression model we call *baseline* (B) and add the new vector of covariates $< \boldsymbol{\theta}_p \odot \boldsymbol{\alpha}_t >$ in the setting we call *complete* (C). For each new time in the testing set, we update the filtered distribution of the $\boldsymbol{\alpha}_t$ variables, separately infer the distri-

	Film and Video	Music	Art	Design	
RMSE	473.8124	222.58390	140.93211	562.3657	
MAE	116.4483	74.92392	44.03222	176.6190	
RMSE	**472.6523**	**222.17493**	143.09378	**555.2242**	
MAE	**102.3413**	**69.29326**	**39.99157**	**124.7995**	
	Publishing	Technology	Games	Food	
RMSE	216.08103	368.8366	679.5713	270.90690	
MAE	48.39902	187.7912	227.7411	62.69218	
RMSE	216.13318	**354.1419**	**672.7153**	272.85313	
MAE	**46.77517**	**129.2110**	**204.0611**	**61.52649**	
	Photography	Fashion	Theater	Comics	Dance
RMSE	98.02709	189.13682	216.70607	120.64323	220.32350
MAE	34.82984	62.73192	89.88659	59.32592	93.79268
RMSE	**96.38986**	**188.72441**	**216.02921**	132.40438	222.0048
MAE	**26.34658**	**57.94165**	**84.96142**	65.82996	**93.5107**

Table 3.1: Average (over time) RMSE and MAE regression values for Linear Regression - Test set (White rows for baseline model and Grey rows for complete model)

bution of θ_p of the description of new projects and update the Linear Model adding the new data in the time-point in the training set and predicting the data in the following time point. This sort of short window (actually, the shortest possible) allow us to maintain a good estimate of the topic heat filtered distributions. A summary of the results is presented in Table 3.1.

As we can observe, when adding the information of the latent topic heats, the simple linear regression algorithm achieves better average results of RMSE and MAE in some of the categories. This provides empirical evidence that adding the topic heat information into black-box models may provide them valuable data to regression tasks even in the cases where there was not a clear structure evident in the topic heat visualization, such as in the categories of *Music* and *Design*. These results provide evidence that the α variables could be used as input for more complex prediction models which aim to model the amount of donations projects are going to receive in a short time-window.

The structure proposed in this model is similar to and could be used in common Multivariate Time Series instead of Collections of Episodic Time Series, it is basically a State-Space Model for which the latent space variables are common to all observed time-series no matter the dimension of these elements, be them vari-

able in time or not. In Chapter 04 and Chapter 05 we enhance the proposed model by adding dependence in episodes that occur simultaneously and by adding dependence on episodes that happened in the past and no longer exist, an element that can only be represented in episodic time series data. Alternative models to the Markov dependency of the proposed topic heat α process can also be utilized in this framework, depending on the nature of the observed data. We believe in this specific case this was the best possible model given that the number of active projects varied drastically throughout the observed period but its variation followed a smooth path of increase and decrease.

3.4 Summary

In this chapter we showcase a model for dealing with multiple episodic time series. This model constructs a structure that connects topic models to these multiple episodic time series and through this connection, a set of latent random variables work as explanatory variables to the collection of different behaviors exhibited in the time series elements. We chose the crowdfunding market as focus of our study given the nature of the data collected in this market. Results present in this chapter show empirically that the constructed features $\theta \odot \boldsymbol{\alpha}$ add information to both understanding the general state of the crowdfunding market as a whole, via the connection of the description of projects and the money they get donated and also add information to the task of predicting future donations to projects.

Chapter 4

Accommodating competition through composition of random variables

The discussion presented in chapter 3 showcases the construction of a model that connects the textual descriptions of crowdfunding projects and the amount of money they receive. This connection is constructed by connecting topic proportions, which stand for the textual description part, and random variables attaching sentiments to topics, i.e., how in vogue these topics are in a given time-point and though simple composition, topic proportions and topic sentiments, which we call *heat*, are the driving process of the amount of money received by projects.

The market constructed by crowdfunding sites are expressions of the main problem focused in this thesis, the problem of modelling collections of multiple episodic time series. Projects consists of description elements that can be summarized via latent low-dimensional variables, they occur for a fixed period of time, the number of projects existing in a given time-point is highly variable and the categorical separation existing in these sites indicates the existence of elements that persist throughout time and extend longer than the lifespan of any given project.

In this chapter, we expand the ideas presented in the previous chapter in order to accommodate other characteristics that the previous model could not handle properly. The proposed modifications allow the model to express the existence of competition among projects whose lifespan overlap and accommodate in a smooth *a posterior* estimation of the topic heats in the presence of very successful projects

while maintaining the state-space representation of topic heats. The modifications proposed add an extra layer of difficulty to the inference process and stochastic variational inference is used in order to overcome these difficulties.

4.1 Model Definition

The generative model proposed in chapter 3 builds up the numeric observations of multiple time series in a traditional way, at a given time-point t, we observe that $y_{p_1,t} \perp\!\!\!\perp y_{p_2,t} | \boldsymbol{\alpha}_t$ for every pair of projects (episodes) p_1 and p_2. This is a common independence assumption due to the nature of episodic time series, for which it is not straightforward to construct a covariance matrix that will both encode dependencies among the multiple time series and at the same time maintain a compact and easy to compute representation. Also, given that the $\boldsymbol{\alpha}$ random variables are latent and we make inference of its distribution *a posteriori* of observing the y values, the proposed model may overestimate the importance of some topics given the existence of extremely successful projects (those whose donations are of orders of magnitude higher than "the average" donation received by other projects), turning the existence of outliers (in terms of donations) the projects that are the most representative of the general state of topic heats, which is contrary to the aim of the model. In the proposed model we aim to model the process of *diffusion of information*, as described in Section 2.4.2, i.e., the process on which more people donate to different projects whose topics are similar resulting in an increase in volume (total of donations) to these projects. In order to construct solutions to these concerns, we try to addressed them via:

1. *Project quality* random variable: To every project, we attach a random latent variable that represents the individual quality of a given product. We aim to use these variables as elements that answer questions related to how projects compare to each other and how dependent of the general state of the market the donations it receives is. In order to do so, given the multiple episodic nature of the problem, project quality variables are then composed so that their interpretation and influence in the stochastic process of the model is

dependent on the collection of other projects that are simultaneously active at any given time-point t.

2. *Project quality X topic heat*: Given the existence of variables that aim to compare projects, we restructure how topics and projects relate to each other. In chapter 3 we compose topics and projects via the multiplication of topic heats times projects topic proportions in order to calculate a score that would be part of the evaluation for the donations of projects. In this chapter, this relation is also controlled by the relative importance of a project given all the others active simultaneously.

Algorithm 6 Topic Based Latent State-Space Model with competition

Require: model parameters $\boldsymbol{\tau}, \boldsymbol{\eta}, \mu_s, \delta_s, A, w_b, b_b, w_g, b_g, \delta_g$
  for all Topic k **do**
    Sample $\boldsymbol{\beta}_k \sim \text{Dirichlet}(\boldsymbol{\tau})$
  end for
  for all Project description p **do**
    Sample topic proportion $\boldsymbol{\theta}_p \sim \text{Dirichlet}(\boldsymbol{\eta})$
    for all Word slot i **do**
      Sample topic allocation $z_{i,p} \sim \text{Multinomial}(1, \boldsymbol{\theta_p})$
      Sample word $w_{i,p} \sim \text{Multinomial}(1, \boldsymbol{\beta}_{z_{i,p}})$
    end for
  end for
  for all Project p **do**
    Sample project quality $s_p \sim \text{Normal}(\mu_s, \delta_s^2)$
  end for
  for all time-point t **do**
    Sample $\alpha_t \sim \text{Normal}(A\alpha_{t-1}, I)$
    for all project p active ate time t **do**
      Sample $y_{p,t} \sim \text{Bernoulli}(\sigma(w_b v_{p,t} + b_b))\text{LogNormal}(w_g v_{p,t} + b_g, \delta_g^2)$
      where $v_{p,t} = (\theta_p^T \alpha_t)\pi(s_p, S_t)$ and
      $S_t = \{q |$ project q active in $t\}$
    end for
  end for

These modifications establish the generative model described in Algorithm 6, which maintains the textual-temporal structure of Algorithm 4 and adds the described modifications to the generative structure. These modifications add a clean and straightforward way to enhance the model to accommodate the characteristics

of the crowdfunding markets cited previously. The key element is the interaction between project quality variables and topic heats, which is done in the construction of $v_{p,t}$, which is

$$v_{p,t} = (\theta_p^T \alpha_t)\pi(s_p, S_t)$$
$$\text{where } \pi(s_p, S_t) = \frac{\log(1+\exp\{s_p\})}{\sum_{s_q \in S_t} \log(1+\exp\{s_q\})}$$

and controls the important properties to the model.

Firstly, the normalization procedure of the project quality variables, performed by the π function, addresses the concerns present in the *Project quality* discussion point. In this procedure, which is composed of two projection steps, one that brings the project quality values to the positive real space and other that brings these values to the simplex space, via the re-normalization of the previous value by the sum of all project quality values of active projects. This double-projection element constructs the relative project quality of projects given all the projects active simultaneously. In this implementation these variables are treated similar to a free parameter but are regularized by their prior distribution with fixed parameters. In order to project these variables to the space of positive real values, we make use of the softplus function due to it smoothness and slow growth, compared to the exponential function, for example. By doing so, we make a non-linear transformation of the project quality variable and add dependency among all of the projects active in a time point t without having to resort in pairwise comparison between all pairs of projects.

Secondly, the result of this re-normalization works as a multiplicative effect to the result of the inner-product $\theta_p' \alpha_t$, acting as a controller to the effect of the topic heats on the expected value of the amount of donations in time t for all active projects. Projects that are relatively more important than others will be more affected by the fluctuation of the market (topic heats) as a hole, stretching their expected value of donations the more important they are and the more the market is hot.

Joining these two elements, we construct a new hurdle model whose two parts

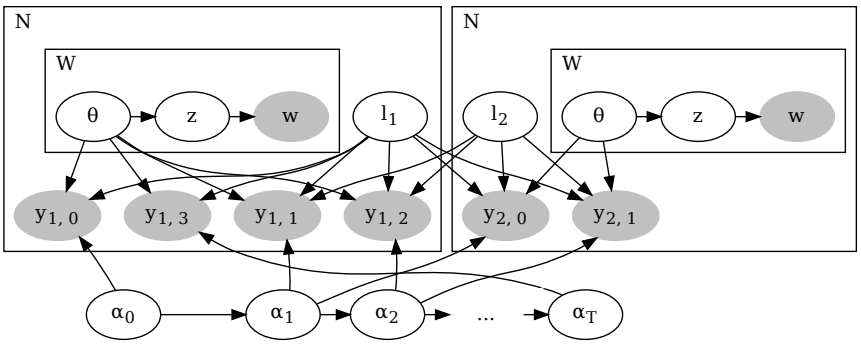


Figure 4.1: Simplified Graphical Model (to maintain readability) of the Generative Process shown in Algorithm 6

are defined by a Bernoulli trial times a sampled log-normal variable. Calling this Bernoulli trial $y^{\star}_{p,t}$, it is sampled via

$$y^{\star}_{p,t} \sim \text{Bernoulli}\left(\sigma\left(w_b(\theta_p^T \alpha_t)\frac{\log(1+\exp\{s_p\})}{\sum_{s_q \in S_t}\log(1+\exp\{s_q\})} + b_b\right)\right)$$

on which we apply the logistic function to the resulting value of the product between parameter w_b, the inner product of the θ_p and α_t variables and the normalized project quality value plus a bias term b_b. Calling this log-normal term $y^{+}_{p,t}$, its sampling procedure is defined as

$$y^{+}_{p,t} \sim \text{LogNormal}\left(w_g(\theta_p^T \alpha_t)\frac{\log(1+\exp\{s_p\})}{\sum_{s_q \in S_t}\log(1+\exp\{s_q\})} + b_g, \delta_g^2\right)$$

where in a similar manner of the previous step, the mean of the proposed distribution is composed by the product between parameter w_g, the inner product of the same latent variables θ_p and α_t and the normalized project quality value plus a bias term b_g - keep in mind that this sampling step is only required once the Bernoulli trial results in 1. All these elements can be best described visually in Figure 4.1.

4.2 Inference and Estimation

In the previous model, we were able to perform a closed-form model-based approach, on which, given the proposed joint distribution of the latent α variables, the observations y and functions of choice of the variational distributions of all the latent variables. In the current model, we are unable to perform such approach due to the impossibility of evaluating the ELBO of the model with the modifications in place. The use of the re-normalization step and log-likelihood of the Bernoulli distribution are the key elements that prevent this closed-form approach and we make use of stochastic variational inference to overcome this difficulty.

In general, inference in the current model follows the same principles of the inference procedure for model in Algorithm 4 present in section 3.2. We maintain the temporal structure of the $\boldsymbol{\alpha}_{1:T}$ variables while splitting the other variables as done in the previous chapter.

$$q(\alpha_{1:T}, \theta^P, l^P) = q(\alpha_{1:T}) \prod_p q(\theta_p) q(l_p) q(z_p)$$

where we are using the superscript ^{P} to represent the whole set of projects. In this setting, as discussed in Chapter 3, for the step of performing the optimization of the variational parameters of $q(\alpha_{1:T})$, we are faced with three different elements that composed make the message-parsing algorithm: $p(\alpha_{t-1}|y_{1:t-1})$, $p(\alpha_t|y_t)$ and $p(\alpha_{t+1}|y_{t+1:T})$, which are combined in order to construct the optimal parameters of $q(\alpha_{1:T})$. The message-parsing algorithm can be constructed in closed-form whenever these three distributions are Normal probability distributions but if we pay closer attention to the equation that constructs $p(\alpha_t|y_t)$, according to equation, we can see that

$$\log p(\alpha_t|y_t) \approx \mathrm{E}[\log p(y_t|\alpha_t,\theta_p,s_p,S_t)]_q(S_t) =$$
$$\sum_{p\in S_t \wedge y_{p,y}=0} \log \mathrm{Bernoulli}(0|\alpha_t,\theta_p,S_t;w_b,b_b) +$$
$$\sum_{p\in S_t \wedge y_{p,y}>0} \big\{\log \mathrm{Bernoulli}(1|\alpha_t,\theta_p,s_p,S_t;w_b,b_b)$$
$$+\log \mathrm{LogNormal}(y_{p,t}|\alpha_t,\theta_p,s_p,S_t;w_g,b_g)\big\}$$

which does not resemble any approximation to the Normal distribution and we cannot even evaluated this element in closed form due to the logistic function of the Bernoulli distribution density and the softplus function of the re-normalization of the project quality variables. Given that the other two elements of the message-parsing schema are kept the same, we overcome this difficulty by isolating the approximation of $p(\boldsymbol{\alpha}_t|y_t)$, parameterizing this approximation as $q(\boldsymbol{\alpha}_t^\star) = \mathrm{Normal}(\boldsymbol{\mu}_t,\Lambda_t)$[1] where $\Lambda_t = CC'$ and C is a lower diagonal matrix that simplifies the optimization process of the covariance matrix of this distribution. With all of that, we are able to separately stochastically optimize $q(\boldsymbol{\alpha}_t^\star)$ using

$$\frac{1}{N}\sum_{i=1} N \sum_{p\in S_t} \log p(y_{p,t}|\alpha_t^i,\theta_p^i,s_p^i,S_t^i) \text{ where}$$
$$\alpha_t \sim q(\alpha_t^\star)\ ,\ \theta_p^i \sim q(\theta_p) \text{ and } s_q \sim q(s_q) \forall \text{project } q \in S_t$$

as the f function of equation 2.2 and the superscript refers to the i-sample. With this approximation in hand, we plug it to the message-parsing algorithms as the approximation of $p(\boldsymbol{\alpha_t}|y_t)$ and proceed similarly as previously, as described the message-parsing section of Chapter 3. As initial values to $q(\alpha_t^\star) = \mathrm{Normal}(\mu_t,\Lambda_t)$ optimization, we make use of the current smoothed parameters of the α_t elements and perform the Cholesky decomposition of the covariance matrix of it in order to transform this matrix into a lower diagonal matrix that will be used in the stochastic

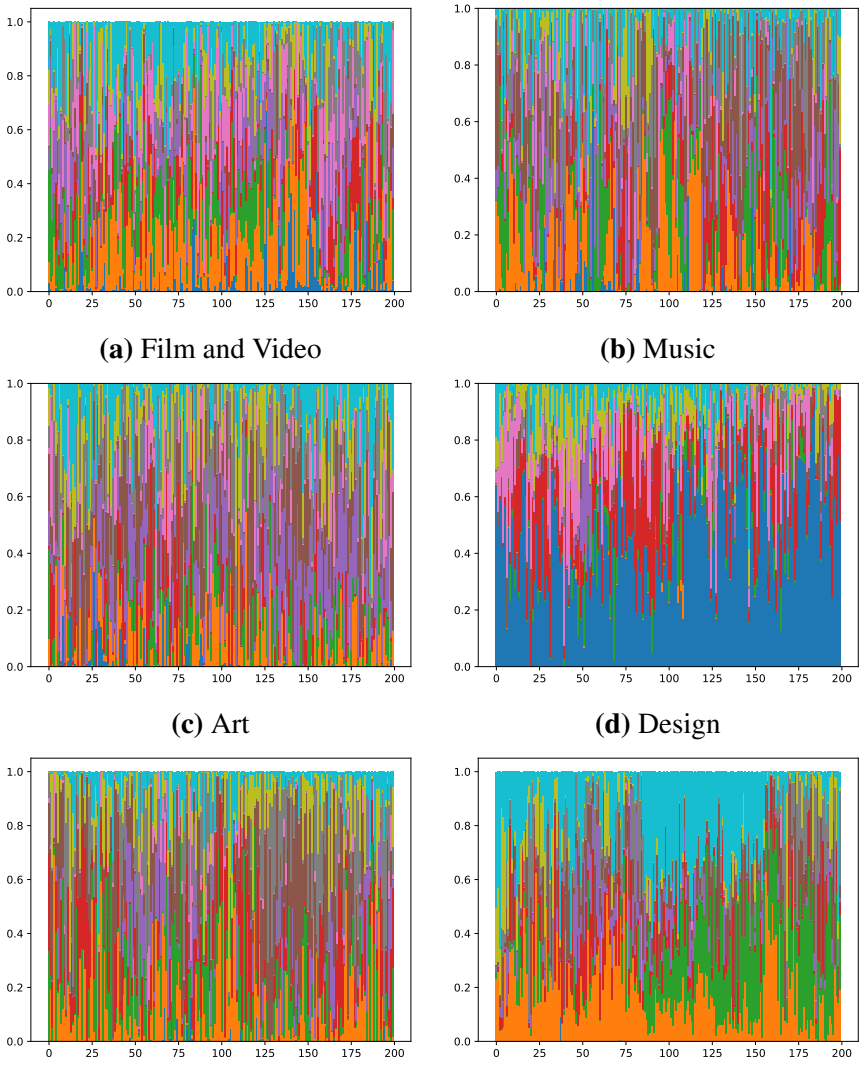
[1] Do not mistake this distribution and $q(\boldsymbol{\alpha}_t)$, which is the marginal distribution of $\boldsymbol{\alpha}_t$ after the message-parsing algorithm is run.

step.

4.3 Experiments and Results

The experiments performed in this section followed the same structure discussed in section 3.3. Given the similar nature of the proposed models, we focus on aspects that were not discussed in the previous chapter. We start by describing the relative importance of topics, in a similar way of the relative importance of projects. In Figures 4.2 and 4.3 we show the expected temporal relative importance of topics for every category on Kickstarter. These figures show the variation in time in which topics become more or lest important (comparing to the other topics) to a given category in a normalized form, i.e., relative to the interval 0-1 or 0% to 100%. So, for the given topic set and a time-point t , topic heats numerical values for time-point t are normalized and to the 0-1 scale.

An initial visual evaluation of these pictures shows the different relative importance every topic possesses to different categories of Kickstarter projects. This is an expected output given that it is natural to suppose that people donating to projects in the category Design pay attention to different topics when compared to people donating to the Music category and vice versa. For instance, it is observed in Figure 4.2 that Topic 09 - the one represented by the dark blue color, which discuss ideas on artwork, stories and means of publishing as seen in Table 4.1 is continuously a very important topic for projects which are placed in the Design category on Kickstarter. On the other side, this topic has barely no relative importance to the projects placed on the Music category, as the dark blue color is not seen in the figure regarding Music. For this category, Topic 08, which discuss elements regarding music production itself, placing record, album and studio among its top-10 words is continuously a relative important topic for the Music category. Additionally, it is observed for both categories and also the rest of the other ones that Topic 1, which is a topic regarding general ideas for all projects on Kickstarter is an important topic throughout the whole observed period. These elements are also seen in the concept of *Most Important Words* which is developed and presented further in this section.



(a) Film and Video **(b)** Music

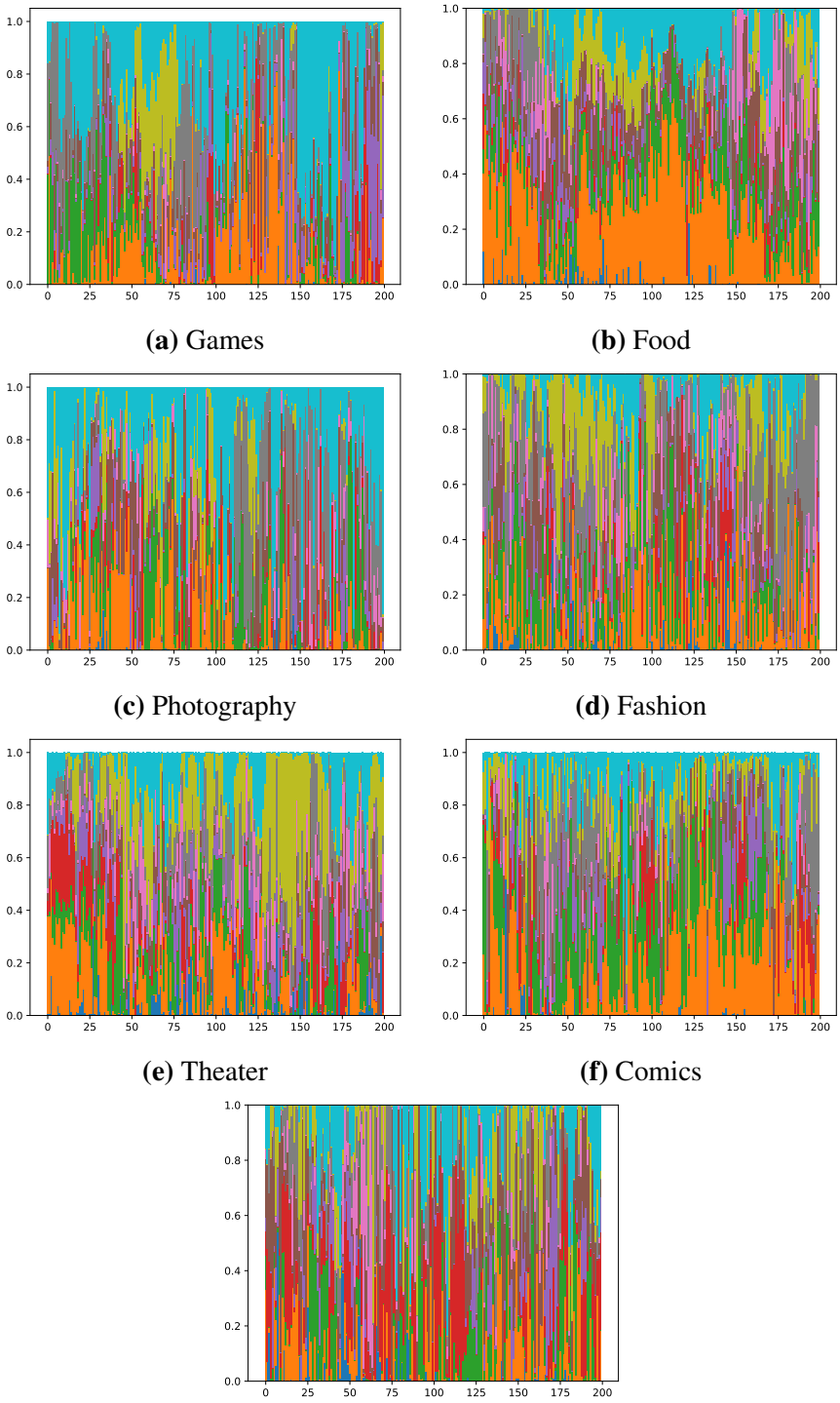
(c) Art **(d)** Design

(e) Publishing **(f)** Technology

Figure 4.2: Expected topic relative importance for 6 Kickstarter categories. (Best seen in color - Each color represents a specific topic)

In topic model literature, there is the concept of top words (TW), which are lists of words that are most probable for every topic. In table 4.1 we can observe the top 10 words of each of the 10 topics constructed in the models discussed in Chapter 03 and Chapter 04. This concept is important for defining topics and their nature given that topic labels are named by human analysis after the evaluation of these lists. Topics top words are important for eliciting how generic texts can be constructed and help explain the main characteristics/label of each topic.

Here, we expand this concept by construct time-varying *most important words*



(a) Games **(b)** Food

(c) Photography **(d)** Fashion

(e) Theater **(f)** Comics

(g) Dance

Figure 4.3: Expected topic relative importance for 7 Kickstarter categories. (Best seen in color - Each color represents a specific topic)

Topic 0	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**
help	updat	design	food	game
ask	last	product	help	card
question	use	shirt	busi	play
get	edt	reward	make	player
make	pm	us	local	get
challeng	app	make	ask	add
risk	devic	color	us	level
time	work	get	product	goal
want	product	ship	need	one
go	develop	custom	question	backer
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**
film	help	product	music	book
product	commun	use	record	print
work	challeng	design	album	art
produc	question	one	song	publish
stori	learn	manufactur	help	stori
show	creat	materi	studio	reward
us	ask	prototyp	ask	page
new	work	water	question	artist
festiv	fund	work	time	edit
director	need	need	make	work

Table 4.1: Top words - 10 Topics

(MIT). MITs are defined as the words that, in a given time-period, are the most important having in order to obtain higher amounts of donations. For a given topic k, $\beta^k = [\beta_1, \beta_2, ..., \beta_W]$, one constructs the top words of it by ordering the respective vector and extracting the index of the words with higher values. We then construct *most important words* by making a weighted sum of the β^k topics by their relative importance at a given time-point. Assuming that we store the topics in column-matrix $\beta = [\beta^1, \beta^2, ..., \beta^K]$ and their relative instantaneous importance a column-matrices $i = [i_1, i_2, ..., i_T]$ where $i_t = [i_{1,t}, i_{2,t}, ..., i_{K,t}]$ is the column vector of importance of all topics at time-point t, the MITs can be collect from the resulting vector

$$s = [s_1, s_2, .., s_T] = \beta^T \mathrm{i}'$$

Range 0 - 50	**Range** 50 - 100	**Range** 100 - 150	**Range** 150 - 200
us	us	us	use
time	use	one	one
use	time	time	time
one	one	product	product
game	ask	question	work
get	get	work	question
work	work	get	get
help	product	ask	ask
make	make	make	make
product	help	help	help

Table 4.2: 10 Most Important Words - Category Publishing

where $s_t = [s_1, s_2, ..., s_W]$ is the vector containing scores for the W words of the vocabulary at time t. We present the top words for periods 1-50, 51-100, 101-150 and 151-200 in tables 4.2 and 4.4 for the categories Publishing and Film and Video and the instantaneous most important words for time-points 25, 75, 125 and 175 in tables 4.3 and 4.5 for the same categories to illustrate the concept. Interestingly, the two tables with the averaged results show a certain degree of stability in the most important words and basically the same words repeat in both tables while the instantaneous tables show a complete different picture (doing the 26 tables for the 13 categories would be tiring to the reader but observe the same phenomena in different categories). This may show that while some elements catch the attention of the crowd at different time points, the general good practice for projects in different categories is to construct descriptions that thoroughly explain the aim of the project (that is, assuming that the most important words present in the tables reflect this idea). The Most Important Words concept is a byproduct of the proposed model and as such, it suffers the same problems encountered presented in Section 3.3. Fortunately, stable results for this are achieved by achieving stable results for the proposed model in this chapter.

In the model presented in this chapter, it is all important to discuss the posterior distribution of the project quality variables, given that they can be used as elements to try to explain the general behaviour of the crowdfunding market we study. Figures 4.4 and 4.5 show a scatter plot of all projects' project quality (l_p variables of

Time 25	**Time** 75	**Time** 125	**Time** 175
expos	challeng	credit	nearli
age	due	coffe	link
contact	school	click	becom
section	outsid	around	know
account	stay	nearli	goal
might	credit	shot	octob
school	finger	alon	soldier
citi	take	goal	product
challeng	around	octob	alon
finger	visit	adult	shot

Table 4.3: 10 Most Important Words - Category Publishing

Range 0 - 50	**Range** 50 - 100	**Range** 100 - 150	**Range** 150 - 200
question	one	use	question
use	use	time	game
time	time	question	time
one	get	one	one
ask	question	get	ask
get	work	ask	get
work	ask	work	work
product	product	product	product
make	make	make	make
help	help	help	help

Table 4.4: 10 Most Important Words - Category Film and Video

Time 25	**Time** 75	**Time** 125	**Time** 175
final	goal	escap	big
actor	good	secret	better
tri	novemb	fun	stay
novemb	enjoy	see	apart
huge	nearli	huge	bit
link	need	question	sun
next	link	extra	take
hope	huge	interview	secret
need	hope	take	see
page	page	shoot	shoot

Table 4.5: 10 Most Important Words - Category Film and Video

the model described in Algorithm 6) by the ratio of the donated money / requested by project (we exclude projects whose ratio > 5 to focus the graphs on the majority of projects).

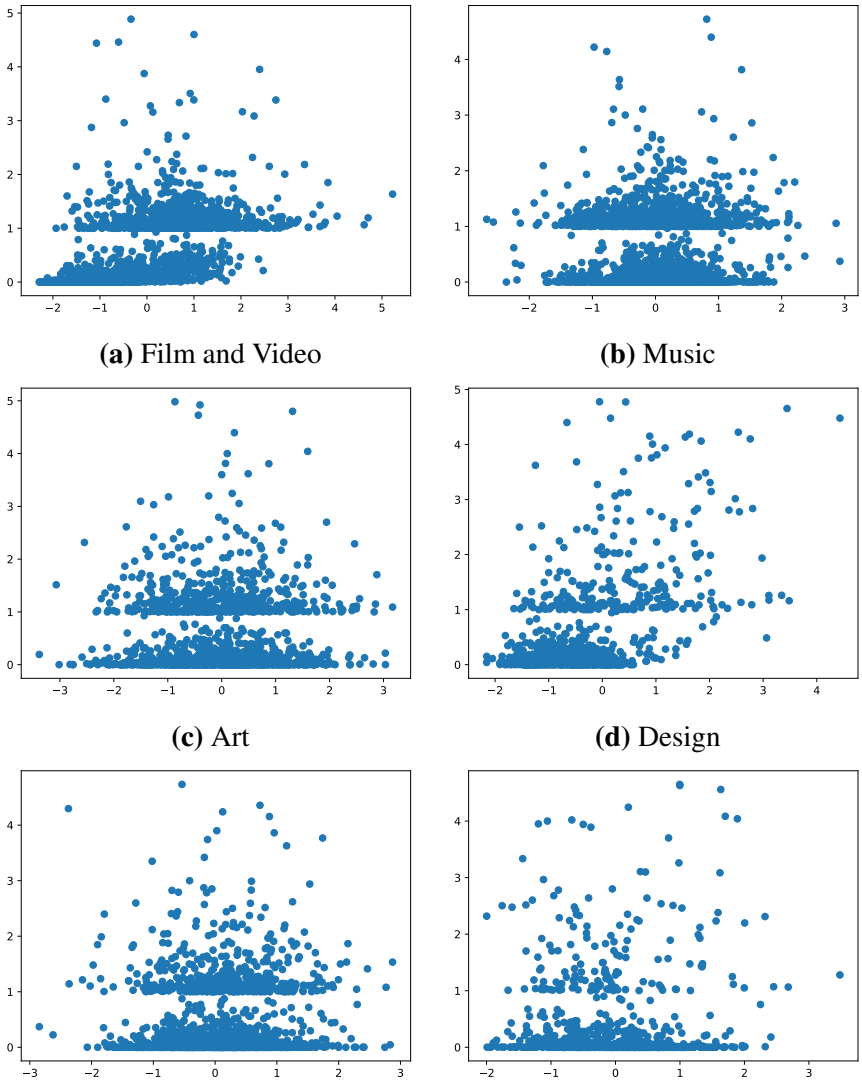
By examining these figures we observe interesting findings. In all figures, we may say there are big cluster of projects and the remaining projects dispersed in the graphs. These clusters located roughly at ratios 0 to 0.5 and just above 1 and are known in the literature of crowdfunding. Crowdfunding projects, in general, either fail by getting less than half of the money they ask or get just a little more than asked. Also, we expect that better projects are, in general, getting more money than worse ones. This trend is correctly captured by the inference of the l_p variables. We expect a positive connection in these graphs and that is what is shown, Design, Games and Comics being the clearest examples of this expressiveness. More interesting than these visual characteristics, it is important to try to understand the effect of the general market (topic heat) in the individual behaviour of projects.

Let us pick the general view provided by picture (**f**) in Figure 4.2. From around time-point 75, topic 2 (green) becomes relative important to the Technology category. In a painstaking process, we tried to find projects that are not great by themselves but which may have been influenced by this surge in the importance of this topic[2]. We select projects *South Paw Protector*[3], which starts at time-point 80, received more than 2x the amount required (100 USD) and whose $\mathrm{E}[l_p] = -0.5604590072130357$ and project *mBuino, a programmable mbed keychain*[4] which starts at time-point 127, received more than 5x the amount required (2000 USD) and whose $\mathrm{E}[l_p] = -0.5531430194260134$. Both projects have the topic 3, among their top 3 topics. Such analysis may lead to a better understanding of the general unobserved importance that donors give to different subjects (topics) presented in Projects descriptions and the effect of it in the amount of money they get throughout their lives (episodes).

[2]It is important to note that we are not making any causal assumption of the kind *surge in the market* $\rightarrow$ *these projects succeed because of that*, but we try to bring up elements that, upon further analysis can construct such claim

[3]https://www.kickstarter.com/projects/1199752982/south-paw-protector

[4]https://www.kickstarter.com/projects/1359959821/mbuino-a-programmable-mbed-keychain



(a) Film and Video

(b) Music

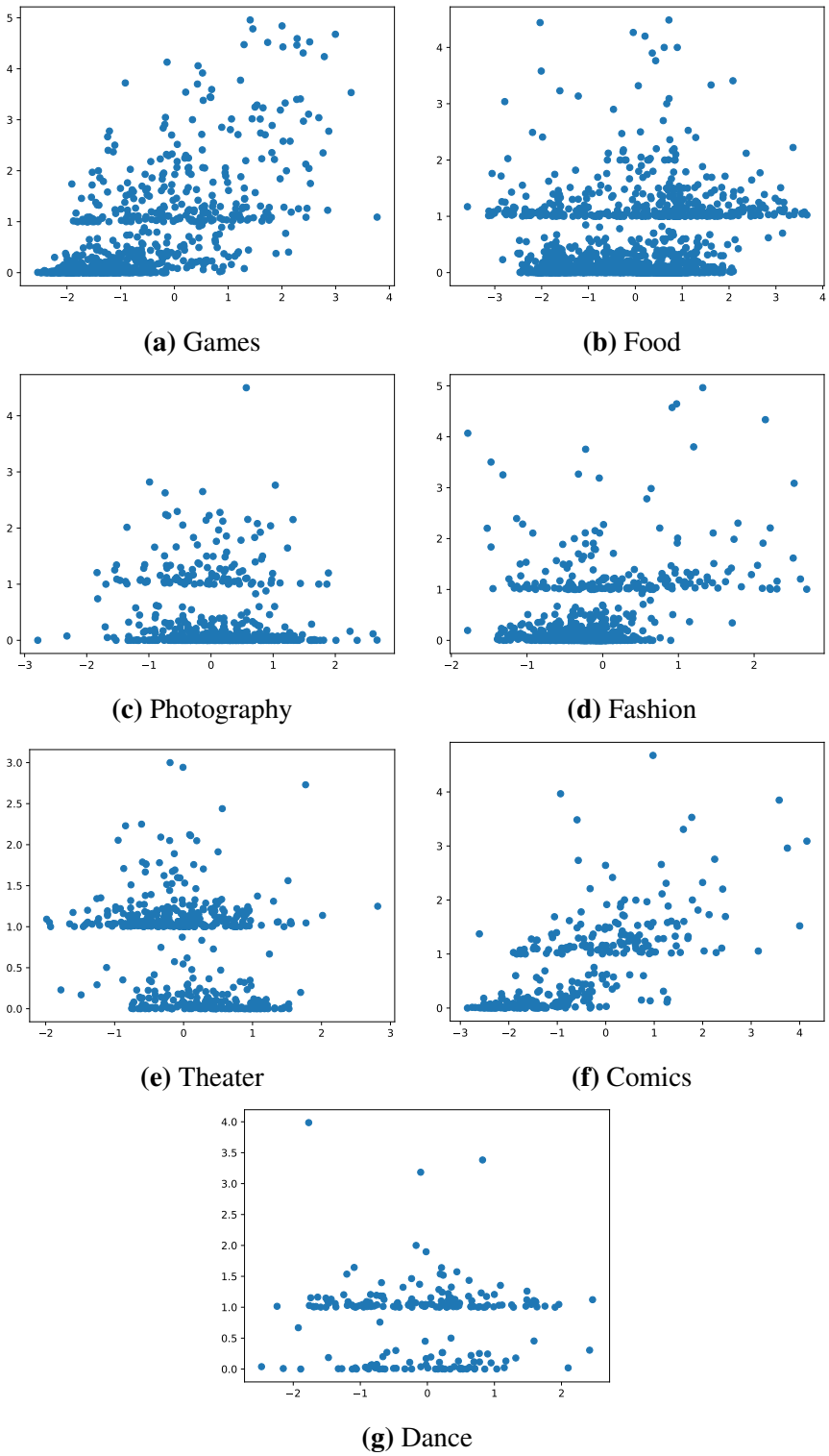
(c) Art

(d) Design

(e) Publishing

(f) Technology

Figure 4.4: Expected topic relative importance. X-axis = $E[l_p]$, Y-axis = $ pledged / $ requested



(a) Games **(b)** Food

(c) Photography **(d)** Fashion

(e) Theater **(f)** Comics

(g) Dance

Figure 4.5: Expected topic relative importance - Part 2. X-axis = $E[l_p]$, Y-axis = $ pledged / $ requested

Appendix A presents the same structure of results for a model containing 20 topics instead of 10 we present here.

4.4 Summary

In this chapter we propose an extension of the model discussed in chapter 03. This model accommodates a greater variety of elements that the previous one and is built on ideas that allow us to express and capture competition among projects, creating a model that ranks projects given their individual quality and that conditions the expected numeric variables in the relative quality of the projects. This gives us a clearer view of the general importance of topics and allow us to examine in greater detail their influence in the success of projects. Additionally, we also construct ideas that extend the idea of top words, giving us a view of the words that winning projects make use of.

Chapter 5

Influence of arbitrary number of episodes in the trending YouTube videos

So far, we have focused on applying the constructed ideas in the crowdfunding market. We now turn our focus on discussing elements of these ideas in the video on-demand market. YouTube and its competitors are now some of the most important entertainment tools for millennials, cable-cutters and digital natives.

Such markets and datasets are connected to some of the ideas present in previous chapter. Provided we collect data the way we described in Chapter 03 and make some reasonable assumptions, we are going to observe similar behaviors in YouTube: assuming that we can observe the list of videos (market) on YouTube in a daily-basis, we may observe a variation on the number of daily views of videos, new videos (episodes) are uploaded everyday, videos *die* - let us assume an arbitrary minimum number of views for videos and call them *dead* if they are bellow this threshold, they may return to the interest of the public given some kind of discussions brought by a third video and so on.

Up to this point in the thesis, we were interested in the information diffusion process of elements occurring at a given time-point. In Chapter 03 we constructed a set of latent variables that were able to represent the *diffusion of information* process of the collective interest of an unknown number of donors to the topics existing

in the description of crowdfunding projects. In Chapter 04 we improved upon this idea and through different elements in the model could accommodate the presence of outliers and inner qualities of projects that are unattached to their descriptions through topics. This kind of modeling had as hypothesis that there is a process by which information was passed to donors, be it by digital social networks or word of mouth, that ended up bringing more donations to a certain group of projects. In this chapter we are concerned to a different view on this process of information diffusion: are individual observations (in this case, videos) important enough so that they generate in the near future after their death a change in the usual balance between videos of different categories on the top trending videos of YouTube? This could be seen as a form of *contagion* between videos and also adds direct dependencies from past observations to the future elements of the latent space in the Bayesian Network of the proposed model, something that is missing in the models of Chapters 3 and 4.

With this in mind, we make use of a publicly available dataset on YouTube Trending Videos [1] to try to understand the effect of individual videos in the proportion of videos for each YouTube category. Contrary to the previous dataset related to crowdfunding, on this one we do not observe the multiple episodes but try to model how the summary of an important episode (a video that has been in the Trending Videos) can affect future elements in the Trending mixture. We aim to do it in a similar way to the ideas of propagating latent features through connections of networks[62] where past observations are summarized and are used as inputs to the next-step latent driver of the process of observations. In our case, videos stay in the YouTube Trending list for a variable amount of time and so we summarize this information into a fixed-size feature vector that is then fed into the model as explanatory variable for future observations of the Trending Videos list.

This chapter follows the same structure as previous ones, starting with the model definition, then passing to the inference and estimation procedures and finalizing with the experiments and results.

[1] https://www.kaggle.com/datasnaek/youtube-new

5.1 Model Definition

The generative model proposed in this section is present in Algorithm 7. The general idea is the following: we observe daily the proportion of videos of each category on YouTube in the Trending site. Videos may enter and leave the list in any arbitrary day. We then assume that a minority of videos is influential and, upon their departure of the Trending list, it will influence a new wave of videos in the same category, making the proportion of videos in the Trending list to have a different expected mix of videos of different categories.

Algorithm 7 Generative model for Influence of multiple videos in the Trending Videos of YouTube

Require: model parameters $w_c, w_n, \sigma_n, \mu_x, \Sigma_x$
  for all t in $1..T$ **do**
    for all category c in $1..C$ **do**
      Sample $b_{t,c} \sim \text{Bernoulli}(\sigma(\sum_{v \in \mathscr{D}^c_{t-1}} w_c^T c_v))\text{Normal}(\sum_{v \in \mathscr{D}^c_{t-1}} w_n^T c_v, \sigma_n^2)$
      where $\mathscr{D}^c_{t-1}$ is the set of videos of category c that
      left the Trending videos in $t-1$
    end for
    Make $b_t = [b_{t,1}, ..., b_{t_C}]$
    Sample $\boldsymbol{x_t} \sim \text{Normal}(\mu_x + b_t, \Sigma_x)$
    Sample $\boldsymbol{y}_t \sim \text{Multinomial}(n, \pi(_{x_t}))$
    where $\pi(x) = {}^{\exp(x)}/_{\sum \exp(x)}$
  end for
  return

It is straightforward to read the generative process shown in Algorithm 7 but we need to clarify two points: first, as constructed, the videos leaving the Trending list are only potentially influential to videos of the same category and secondly, we need to define c_v, the vector created just after a video leaves the list. Let us assume that there are two videos leaving the Trending list that are of category 01, one has been in the list for 10 days while the other for 2 only. The easiest way of transforming these individual videos whose observation differs a lot is by summarizing such individuals in features of fixed dimensionality. The collected dataset presents four covariates (views, likes, dislikes, comments) for every observation. With this in hand, we calculate the mean of these covariates and transform the resulting vector into the c_v covariate/variable, which is obviously of same dimensionality no matter how long

a video stayed in the Trending list. Also, the proposed model may seem to be an over-parameterized Multinomial-Logit regression model but we wanted to maintain the direct influence of videos in their own categories. In other words, for a given video v that has been in the Trending list for a period of n observations, we define c_v as

$$c_v = \frac{1}{n}\sum_{i=1}^{n}[\text{views}_i, \text{likes}_i, \text{dislikes}_i, \text{comments}_i]$$

where the summary of the period the video v has been in the Trending list is defined is the means of its views, likes, dislikes and comments throughout the period. With that in hand, for every category c and time-point t, we sample the variable $b_{t,c}$ which is the latent variable that defines if the videos on category c that left the Trending list at $t-1$ had influence on changing the mean number of videos of the same category existing in the list at t. This $b_{t,c}$ random variable is sampled a by a hurdle model defined via

$$b_{t,c} \sim \text{Bernoulli}(\sigma(\sum_{v \in \mathscr{D}^c_{t-1}} w_c^T c_v))\text{Normal}(\sum_{v \in \mathscr{D}^c_{t-1}} w_n^T c_v, \sigma_n^2)$$

where $\mathscr{D}^c_{t-1}$ is the set of videos of category c that left the Trending videos in $t-1$. This variable is then used to perturb the expected value of the number of videos of category c in the Trending videos list at time t, which is defined by the model parameter μ_x, a time-invariant vector parameter whose c-th component regards category c. Joining all $b_{t,c}$ elements into the vector b_t, the random variable x_t is sampled via

$$\boldsymbol{x_t} \sim \text{Normal}(\mu_x + b_t, \Sigma_x)$$

for which Σ_x is the Covariance Matrix for this multivariate normal distribution, another model parameters to be estimated (it is simplified in the experiments to the identity matrix). The sampled x_t elements are then normalized so that the mix of videos in the Trending list $\boldsymbol{y}_t$ is sampled using

$$\boldsymbol{y}_t \sim \text{Multinomial}(n, \pi(_{x_t})) \text{ where } \pi(x) = {}^{\exp(x)}/_{\sum \exp(x)}$$

where n is the fixed number of videos in the Trending list, usually 200. This is assumed to be given and fixed and is not object of study in this work.

5.2 Inference and Estimation

The learning procedure of this model follows the same guidelines of the procedures constructed in the previous chapters. The unknown quantities of the model are the random latent variables of the unormalized expected number of videos in each category $\boldsymbol{x}$ and the influence of past videos $b_{t,c}$ for each category and also the model parameters $w_c, w_n, \sigma_n, \mu_x, \Sigma_x$. We can not evaluate the posterior $p(x,b|y)$ in closed-form, neither we can evaluate the expected marginal $\mathrm{E}[p(y|x)]_{p(x,b|y)}$ in a vanilla EM algorithm, so we resort on the stochastic Variational Expectation maximization in order to perform the learning procedure of this model.

We start by constructing the structured variational distribution of the latent variables. The logistic-Normal step can be divided in two different elements, what we call $b^{\star}_{t,c}$ and $b_{t,c}$. By doing so, we construct the variational distributions of these elements as

$$q(\boldsymbol{x},b) = q(\boldsymbol{x}) \prod_{t=2}^{T} \prod_{p \in \lceil^{c}_{t-1} \forall c} q(b^{\star}_{t,c}) q(b_{t,c})$$

where we maintain the stochastic dependence between all the x_i elements and separate all the other random variables in the model. The parameterization for $q(x)$ in a similar way to $q(\alpha)$ in Chapter 4, by using a variational distribution of the form $q(x) = \mathrm{Normal}(m_x, \Lambda_x)$, where the variational parameter $\Lambda_x = CC'$ is constructed using the composition of a lower diagonal matrix C times its transpose C'. This construction allows for easy use of stochastic gradient descent and guarantees a suitable covariance matrix for the proposed multivariate normal parameterization. For the individual $b^{\star}_{t,c}$ variables we propose their variational distributions to take the form $q(b^{\star}_{t,c}) = \mathrm{Bernoulli}(\sigma(\mu^{\star}_{t,c}))$, where the variational parameter $\mu^{\star}_{t,c}$ is defined in the space of the Real line so that it is easily optimized by stochastic gradient descent as well. The remaining $b_{t,c}$ variables have their variational distributions defined via $q(b_{t,c}) = \mathrm{Normal}(l_{t,c}, \lambda_{t,c}))$.

Once again, we may resort on stochastic Variational Inference due to the step $\mathrm{E}[\log p(y_t|x)]$ for which we cannot make use of closed-form optimization to infer the variational parameters of the related latent distributions. Similar to how we did in Chapter 04, the function to be stochastically optimized is constructed as

$$\frac{1}{N}\sum_{i=1} N \log p(y_t|x_t^i) \text{ where}$$
$$x_t^i \sim \text{Normal}(m_x, \Lambda_x = CC')$$

where N is the number of samples to be simulated. The other elements of the ELBO are straightforward to write and can be computed in closed form, so the only noisy part of the computation of the ELBO is this expectation. The complete log-likelihood of the proposed model on which the ELBO is evaluated is

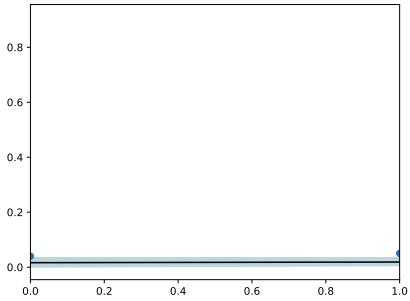
$$\log p(y,x,b) = \sum_{t=1}^{T}\left\{\sum_{c=1}^{C}\log p(b_{t,c}|\mathscr{D}_{t-1}^{c}) + \log p(\boldsymbol{x}_t|\boldsymbol{b}_t) + \log p(\boldsymbol{y}_t|\boldsymbol{x}_t)\right\} \tag{5.1}$$

for which we must keep in mind that the set $\mathscr{D}_{t-1}^{c}$ is actually making the conditioning of $b_{t,c}$ on the videos belonging to y_{t-1} and leaving the Trending list at that time-point.

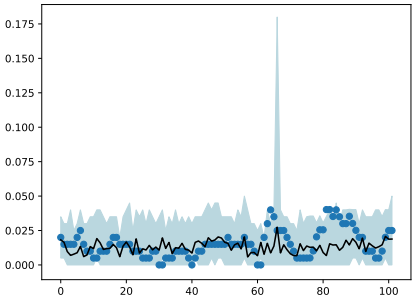
5.3 Experiments and Results

The public dataset collected for these experiments is composed of daily observations of the top 200 (it actually varies a little in very few observations, something that does not affect this model - this is taken as given, not estimated) videos in the Trending section of YouTube. These sections may vary by the location of the viewer and the data collected is related to the USA and other countries.

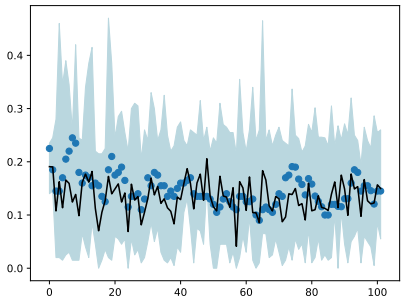
We then split the bases in half, 100 time-points for training and 100 for testing. All model parameters are estimated using the training set and the variational distributions of the latent variables are inferred using this part of the dataset. Following this procedure, we performed the inference of the latent variables of the remaining



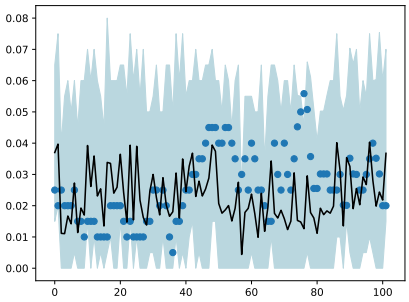
(a) Film & Animation



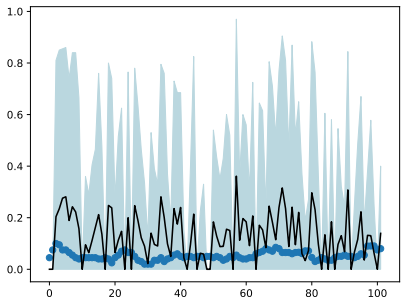
(b) Autos & Vehicles



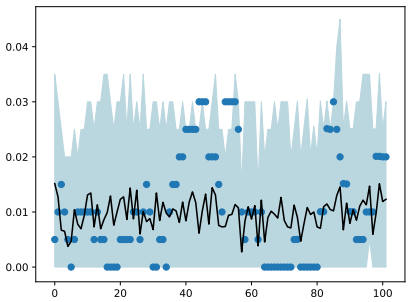
(c) Music



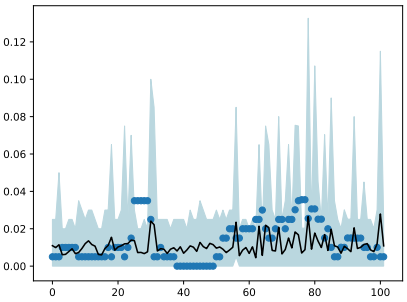
(d) Pets & Animals



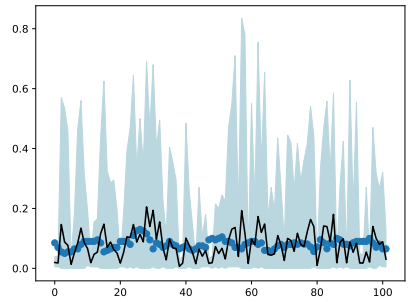
(e) Sports



(f) Short Movies



(g) Travel & Events



(h) Gaming

Figure 5.1: Observations (dots), Expected values (grey line) and 95% predictive interval - Training set Part 01

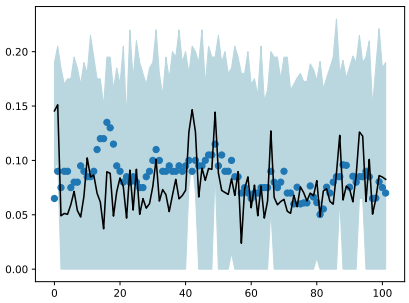
100 time-points through a sliding window of size 1. For each time-point t, distributions of all $b_{t,c}$ and $\boldsymbol{x}_t$ were evaluated and the distribution of $\boldsymbol{y}_t$ was derived. Then for time-point $t+1$, the observation of $\boldsymbol{y}_t$ was taken into consideration for the restart of the process regarding the latent variables and the distribution of the number of videos in the list at time t. In figures 5.1 and 5.2 we present the 95% predictive intervals for the proportion of videos in each category in the training dataset.

As expected, the predictive intervals for the training test comprise the majority of the observations and in some cases are relatively tight. Unfortunately, specially due to sampling bias, there are some problems in the fitting of the model. First, there is no video of the category Film & Animation observed in the dataset, so the data did not support any learning on the behavior of this specific category. In despite of that, some other categories such as Travel & Events have their data quite well adjusted.

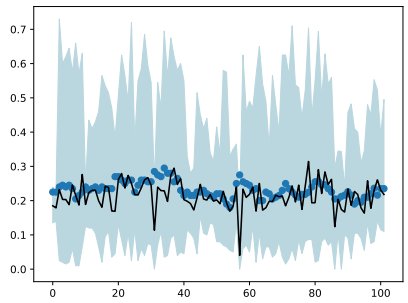
One important aspect of the proposed model is to look at distribution of the $b^{\star}_{t,c}$ latent variables. In this modeling, they were the key elements that "decided" if videos that left the Trending list at a past time-point were influential to the distribution of videos in the list at a future time-point. In Figure 5.3 we take a look at the expected values of $b^{\star}_{t,c}$ for the data in the US. In this histogram we observe that videos leaving the Trending list are lightly likely to influence the future list of videos in the short term.

Additionally, in Figure 5.4 we observe the expected values of $b^{\star}_{t,c} * b_{t,c}$ variables, i.e., the expected contribution of previous videos to the mean number of videos on each category in near future Trending lists. Surprisingly, it is observed that, although counter-intuitively, for some cases it was expected a negative contributions of previous videos to the future number of videos of the same category in the Trending list. On the other hand, it was estimated that the vast majority of time, the contributions were expected to be near zero, meaning that the equilibrium of the number of videos of every category in the Trending list is not usually changed drastically and are expected to remain stable over time.

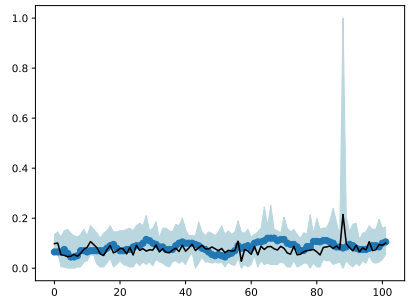
Although interesting, current results need further investigation in future work.



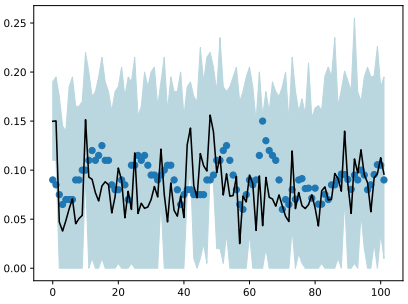
(a) Videobloggingn



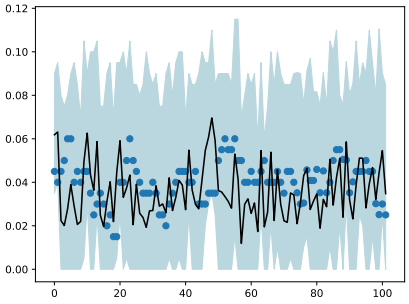
(b) People & Blogs



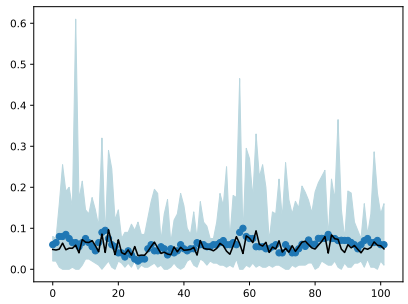
(c) Comedy



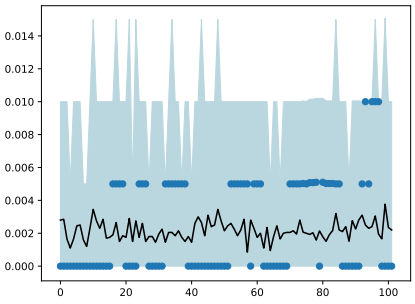
(d) Entertainment



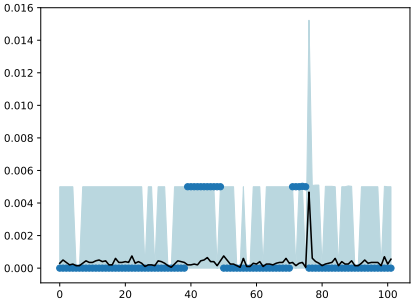
(e) News & Politics



(f) Howto & Style



(g) Education



(h) Science & Technology

Figure 5.2: Observations (dots), Expected values (grey line) and 95% predictive interval - Training set Part 02

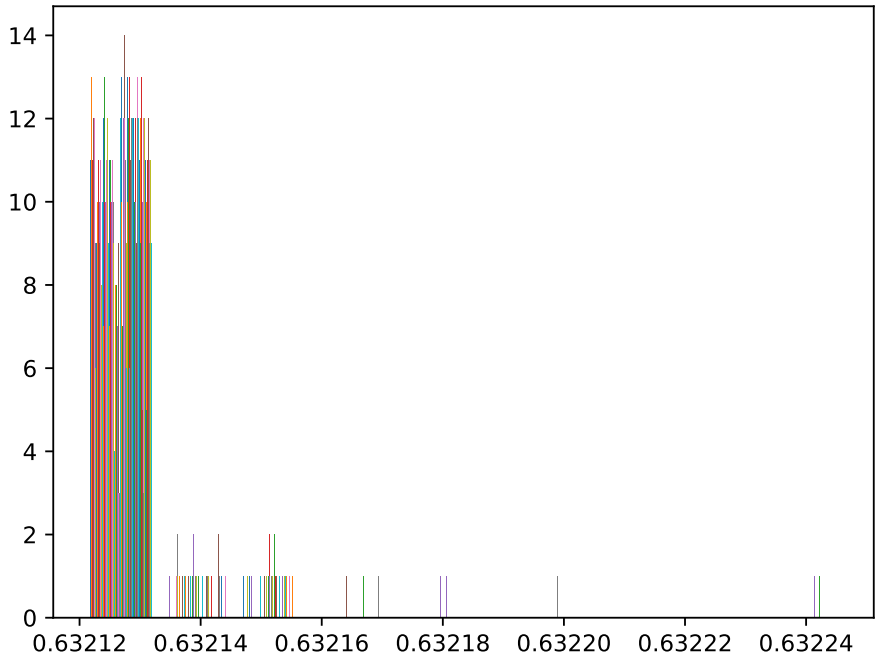


Figure 5.3: Histogram of the expected values of $b^{\star}_{t,c}$ variables - US Data

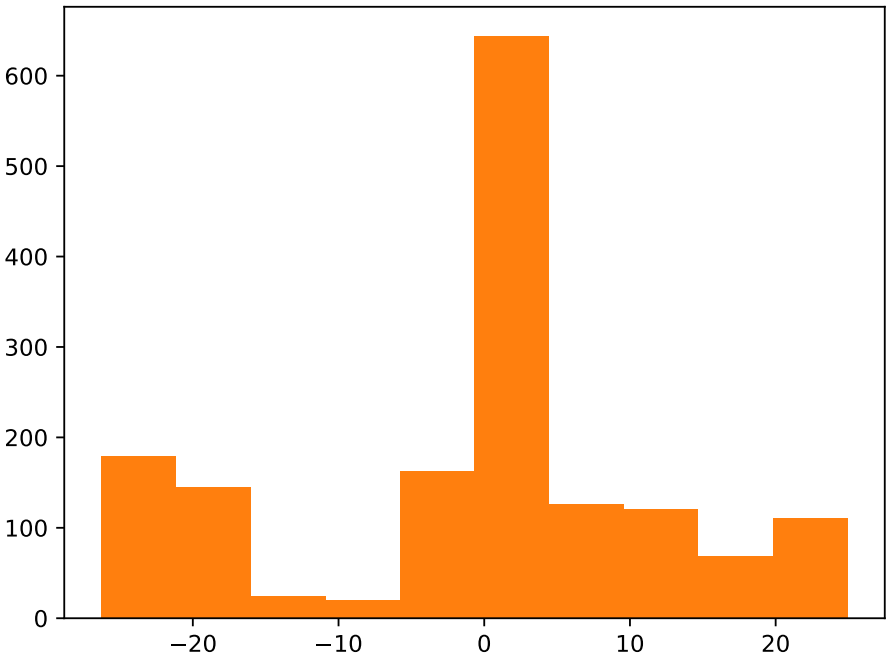


Figure 5.4: Histogram of the expected values of $b^{\star}_{t,c} * b_{t,c}$ variables - US Data

Further detailing of the data under scrutiny needs to be done in order to understand if the contributions of previous videos really occur or the model is adjusting the errors of the estimators via the latent variables. Current data did not provide means to these detailing but the small number of cases where it was expected some influence via $b^{\star}_{t,c} * b_{t,c}$ may help further investigation. In the Appendix we provide more results for the testing data but the general behaviour of the model remains. The adjustment of the model is less effective but this happens due to the simpler learning procedure.

5.4 Summary

In this chapter we proposed a model to encode the diffusion of information via contagion the moment after an episode occurs in the scenario of videos in the Trending list on YouTube. This contagion was meant to be enclosed to the videos in the same category as the episode that had just occurred. It adds a different element to the models studied in the previous chapters, which is the characteristic of *post mortem* contribution. This feedback mechanism has been studied in different settings [62] and may contribute to the enhancement of models containing episodic time series in general.

Chapter 6

General Conclusions

Throughout this thesis we have considered, discussed and constructed exploratory studies on information diffusion in multiple episodic time series. Here we are going to discuss the main characteristics of the models proposed in the thesis.

Firstly, we defined what multiple episodic time series are and discussed their existence and importance. When dealing with datasets composed of multiple episodes, it is important to construct a model that allows a flow (in a very generic definition) of information between past and future episodes, in a way that this flow of information is relevant to every different instant of observations, be it an instant in which there are few episodes occurring or many simultaneously.

Also in Chapter 2 we continued defining the elements we would use in the thesis. Topic Models were discussed in details given that some elements of its generative model, the topic proportion, would be part of the main feature space described in the work, in which we plugged both variables that allowed sentiment to be expressed by and information to be diffused through time. We also defined latent state-space models on which the temporal processes were defined and Variational Inference, the technique that allowed us to implement the graphical models we proposed and perform inference and estimation on them using reasonable computational resources.

In Chapter 3 we construct a first model that connects textual descriptions and numeric observations, via the random variables topic proportions and topic heats, which varies temporally and aims to measure positive sentiments people may have

towards topics. Via *diffusion of innovations*, projects that - probably by chance - created in the correct moment, enjoy higher likelihood of receiving donations. Also the features constructed by the topic proportions and topic heats may be used in a regression model trying to predict future amounts of donations to projects. This chapter showed some empirical evidence that 1 - we can connect these elements in a reasonable sense and 2 - they can be used as elements to describe both individual and multiple observations. The key element of this chapter is the simplicity of the composition between variables which allows us to get a straightforward explanation of the elements present in the composition. On the other hand, bringing the simplicity of the composition of the elements to the emission process (the observation of donations) may lead us to sub optimal results. Relaxing the assumptions made in this part of the model could possibly give us better explanatory and predictive results.

Chapter 4 continues the study proposed in Chapter 3 by allowing a more complex interaction between the elements of the model. Previous model assumed independence between numeric observations of episodes (projects) occurring simultaneously given the state of the topic heat and, which is a very strong assumption to make, given that donors in crowdfunding projects usually only donate to one project only. In order to accommodate this competition among projects, we construct a variant of the previous model by adding a new latent random variable which is used in an improved version of the composition of variables. Fortunately, the inference and estimation process is capable of capturing the idea of the added latent variable and, given that, we can explore different ideas in the dataset. Once again, on the other hand, relaxing the emission structure may allow us to improve in the results. The structure of the latent state-space topic heat seems to be a good choice for the flow of information that we aim to build but the very high variability of the projects is something to be addressed.

Chapter 5 takes a different approach to the diffusion of information process, aiming to model *post mortem* influence of episodes. This possibility of diffusion is characteristic of collections of episodic time series and can be used to enhance

the information contained in the latent space driving the process that generate the observed values for the collection under scrutiny.

6.1 Future work

After working in this thesis, I firmly believe that the construction of features that explain both individual and multiple observations simultaneously is of interest in several different areas. With the popularization of Deep Learning approaches [63], feature engineering has become an automatic task for models to tackle but I believe man-made features can still be interesting to work with. A clear example in the Educational Field, where the framework (pedagogues would call it differently) proposed by Paulo Freire[64], the patron of Brazilian education constructs, among many other things, a summary of human relations (not necessarily textually) in a low dimensional vector of latent sentiments and then construct future relations based on the previous summary elements. That is a good example of man-made feature whose usage is interesting enough to base further improvements in the actual models. This opens doors to applying the same ideas in scenarios where texts are not the main information source, but where well-defined psychological traits are the driving force of diverse episodes in different contexts.

Another area where future works may be of interest is in generating a flexible enough model for texts and numbers that accommodate a wider range of scenarios. All the models studied in this thesis have strong statistical assumptions and work in a very specific scenario. With the kind of interactions introduced in this work and the enormous variety of contexts the studied papers present, it may be of interest to construct a body of work that brigs all the scenarios under a similar construction and relax the strong assumptions the models make. With all the theory developed and good utilization in different datasets and scenarios, this kind of model may become basis upon more complex work can be made.

Another interesting line of research may be *post mortem* analysis and modeling, in which the episodes influence the driving process of multiple time series after their own episodes end, something that we tried to scratch in the model proposed

in Chapter 5. Crowdfunding projects are a good example of this kind of analysis. After they close for donations, they continue to exist if the succeed in getting the donations they required, being open for comments by donors and updates by creators. Given the lack of maturity the crowdfunding market has, a high number of projects fail to deliver what they promised[65] and modeling and understanding the reasons and effects of such failures may help in maturing the market. One again, all the difficulties of episodes and multiple time series arise in such environment.

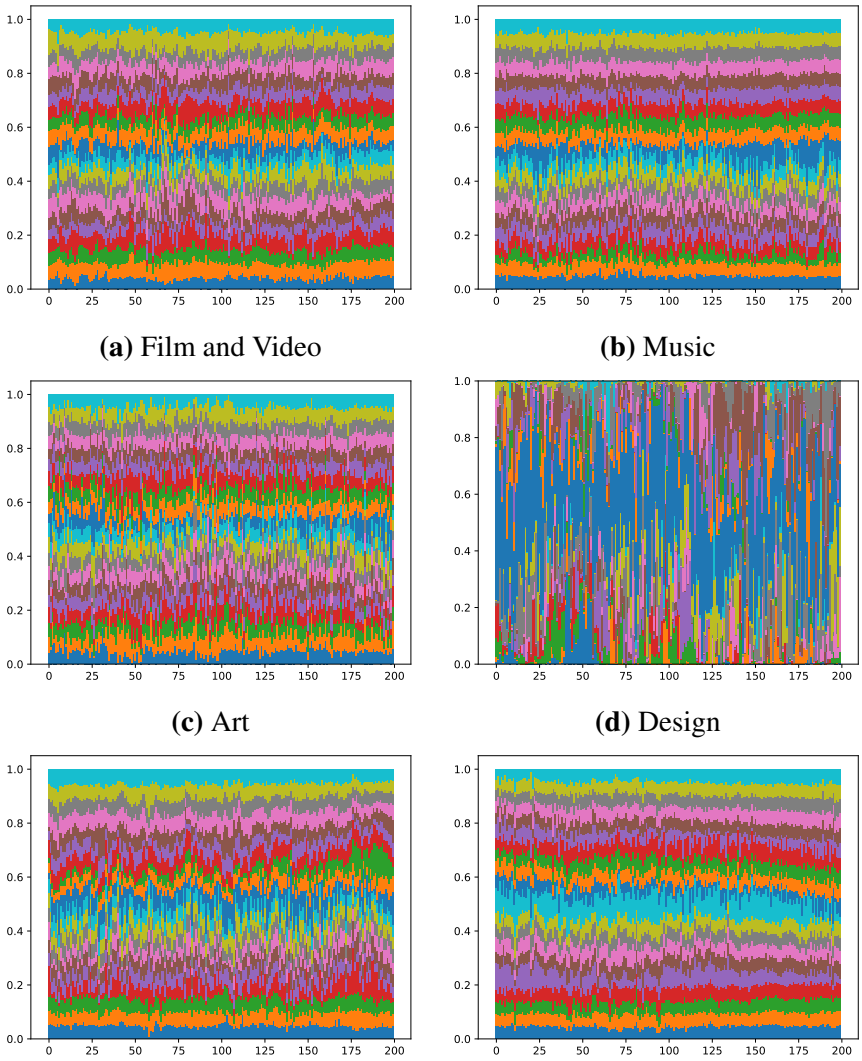
Models proposed in Chapters 3 and 4 mix supervised and unsupervised learning in a single generative models. This construction has given us many problems while learning the model parameters and latent variable distributions and we applied a "brute-force" solution to minimize these effects (which are already naturally present in the EM framework), so we run several different copies of the same instance of dataset, random restarting them in different areas of the parameters and hyperparameters spaces. This introduces an enormous computational burden to our algorithms and clever ways of training the models should be considered. One interesting element to solve this problem may be the use of a similar approach to the one present in Generative Adversarial Networks [59]. Note that we do not aim to construct yet another version of GAN's, but we aim to make use of the learning algorithm/approach it uses to train the network in our kind of models.

Appendix A

Additional results for chapter 04

We now explore the results of the same experiment shown in chapter 04 but now making use of 20 topics instead of 10. Although there are Latent Dirichlet algorithms in the literature that automatically estimate the number of topics that best describe a corpora, I do believe that in a real-world application, picking the number of topics in a model is a *management* task and, as such, depends on specialized knowledge. In general, topic models better fit (and overfit) the data they are trained on the more topics you allow in the model, so either one makes use of Information Criteria to evaluate the quality of models or make use of this prior specialized knowledge,

Figures A.1 and A.2 show the relative importance of the 20 topics in this model, tables A.1 and A.2 show the top 10 words in the model, tales A.3 and A.5 show the most important words categories Publishing and Film and Video and the instantaneous most important words for time-points 25, 75, 125 and 175 in tables A.4 and A.6 for the same categories.



(a) Film and Video

(b) Music

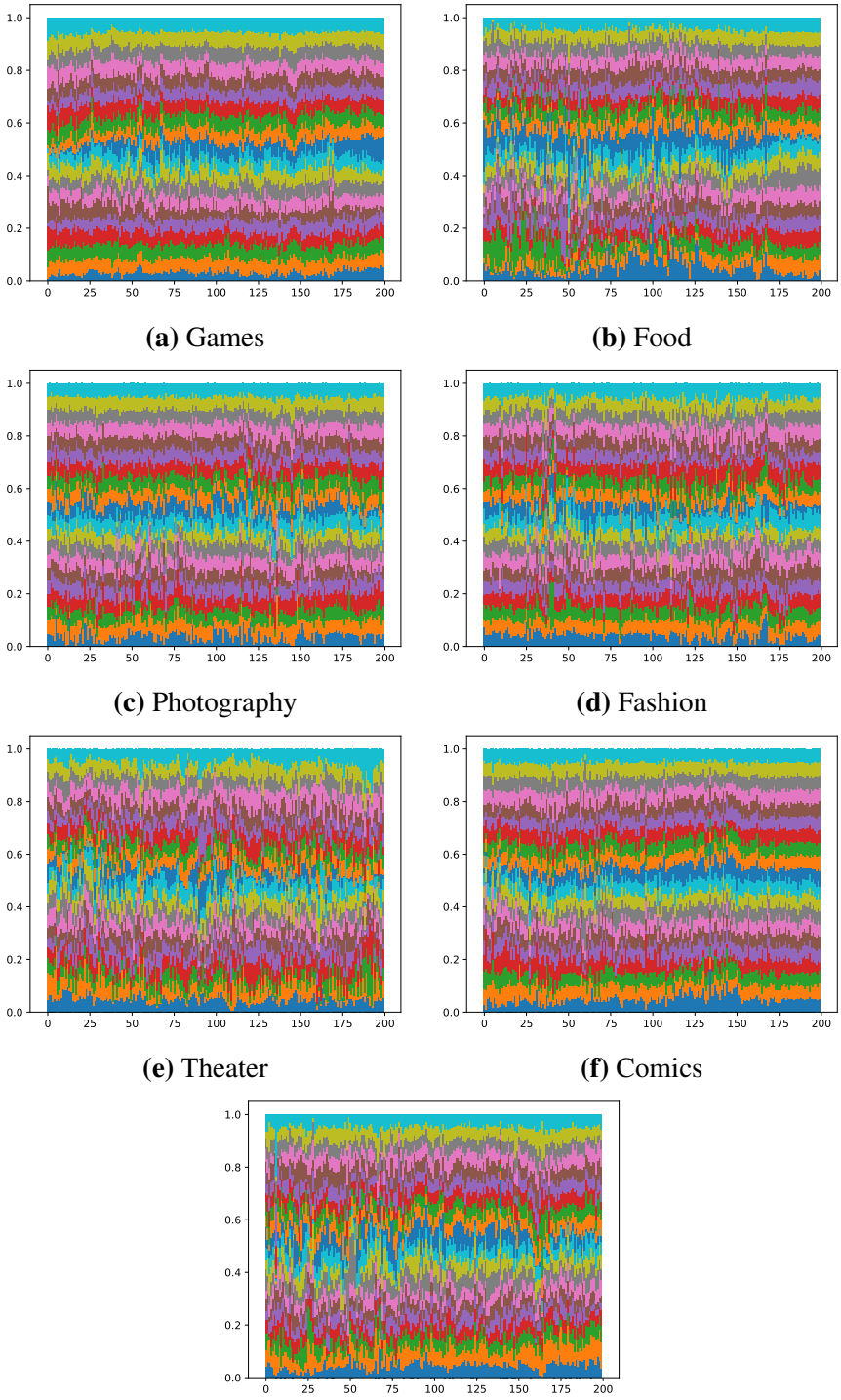
(c) Art

(d) Design

(e) Publishing

(f) Technology

Figure A.1: Expected topic relative importance - Part 01. (Best seen in color - Each color represents a specific topic)



(a) Games **(b)** Food

(c) Photography **(d)** Fashion

(e) Theater **(f)** Comics

(g) Dance

Figure A.2: Expected topic relative importance - Part 02. (Best seen in color - Each color represents a specific topic)

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
game	art	music	print	app
player	artist	record	reward	use
play	work	album	art	develop
charact	perform	song	backer	user
world	new	studio	page	devic
develop	show	band	comic	softwar
team	event	cd	pledg	phone
make	danc	releas	goal	work
new	commun	time	get	need
creat	citi	help	includ	video
Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
card	get	local	food	product
game	make	build	make	use
play	go	food	product	design
ship	like	commun	beer	manufactur
deck	want	busi	coffe	prototyp
add	one	need	recip	power
pledg	peopl	year	cook	test
set	time	locat	flavor	light
player	us	space	ingredi	work
backer	thing	help	dog	need

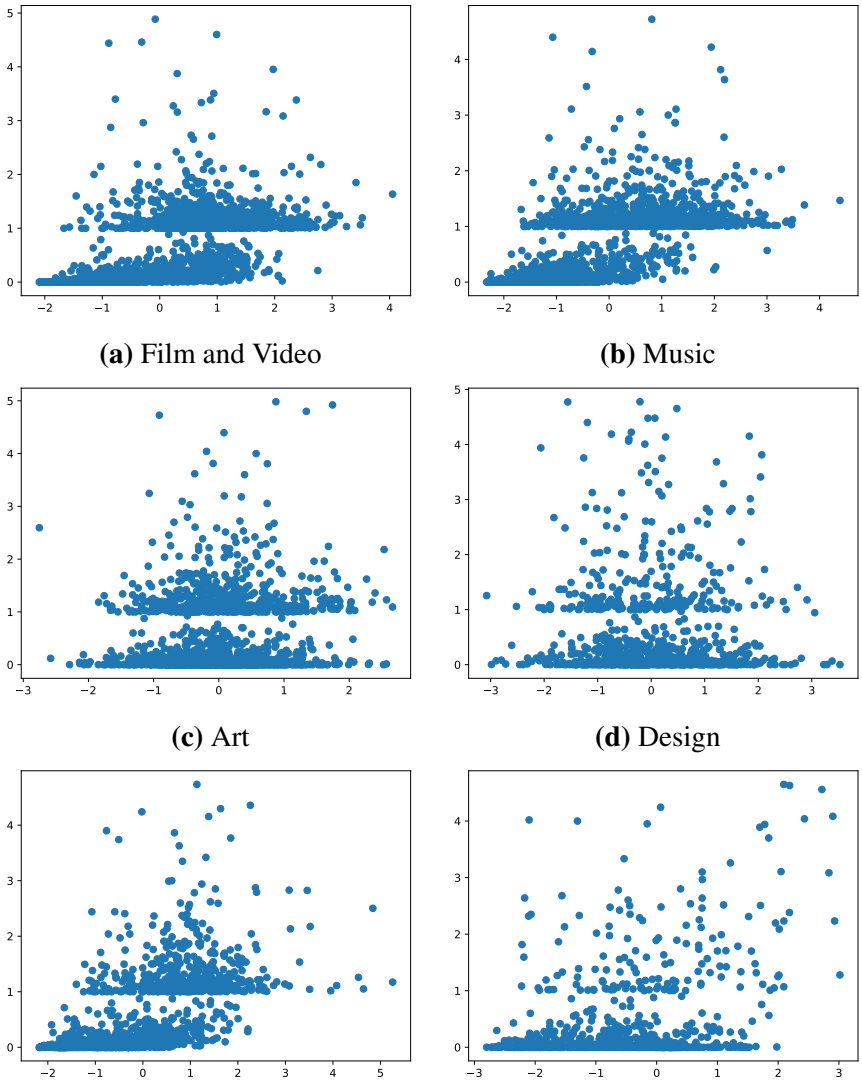
Table A.1: Top words - 20 topics. Part 1

Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
ask	student	book	updat	product
question	school	publish	last	market
help	educ	stori	edt	busi
challeng	commun	write	pm	compani
risk	learn	read	est	fund
learn	de	first	pledg	websit
creator	help	author	get	cost
info	world	children	us	us
directli	state	illustr	make	provid
account	program	work	want	success
Topic 15	Topic 16	Topic 17	Topic 18	Topic 19
life	piec	print	film	design
one	color	photo	product	product
time	design	imag	produc	shirt
stori	one	photograph	stori	size
peopl	use	x	work	bag
would	wood	camera	movi	color
love	reward	photographi	make	made
like	made	anim	crew	us
world	make	work	director	make
live	materi	paint	short	fit

Table A.2: Top words - 20 topics. Part 2

Range 0 - 50	**Range** 50 - 100	**Range** 100 - 150	**Range** 150 - 200
get	question	product	use
question	get	get	question
time	time	one	time
one	ask	use	ask
ask	one	time	get
use	use	question	one
product	product	work	work
work	work	ask	make
make	help	make	help
help	make	help	book

Table A.3: 10 Most Important Words - Category Publishing

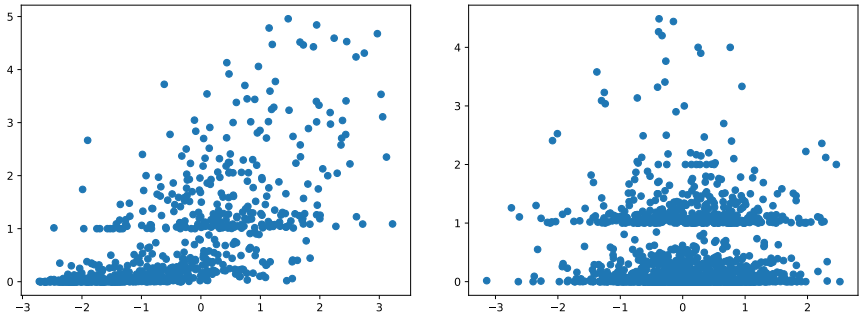


(a) Film and Video **(b)** Music

(c) Art **(d)** Design

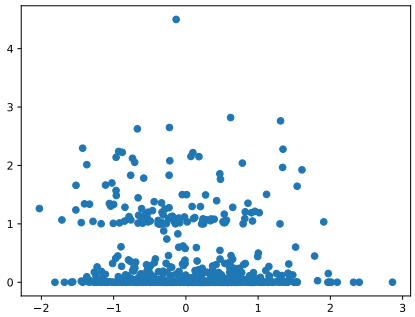
(e) Publishing **(f)** Technology

Figure A.3: Expected topic relative importance. X-axis = $E[l_p]$, Y-axis = $ pledged / $ requested

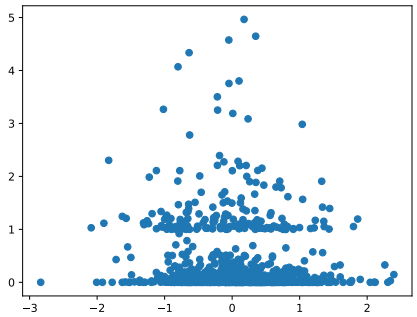


(a) Games

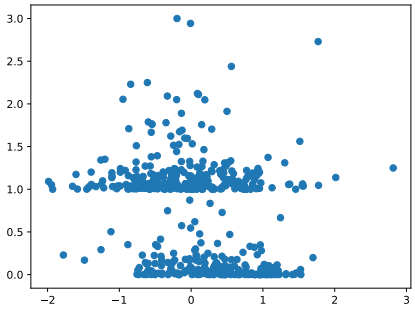
(b) Food



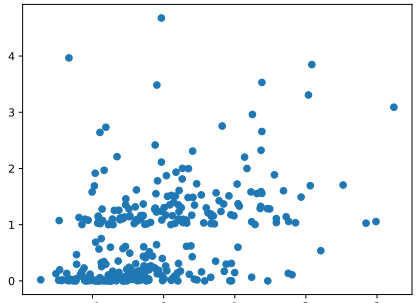
(c) Photography



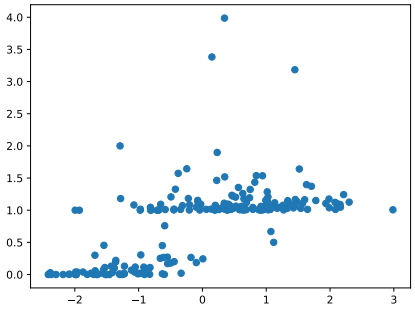
(d) Fashion



(e) Theater



(f) Comics



(g) Dance

Figure A.4: Expected topic relative importance - Part 2. X-axis = $E[l_p]$, Y-axis = $ pledged / $ requested

Time 25	**Time** 75	**Time** 125	**Time** 175
colleg	colleg	need	due
mine	edt	link	spring
actor	age	quick	book
credit	actor	shot	children
afford	advanc	adult	chang
death	afford	shoot	credit
advanc	death	star	offer
day	day	take	campaign
age	credit	secret	movi
decemb	decemb	see	much

Table A.4: 10 Most Important Words - Category Publishing

Range 0 - 50	**Range** 50 - 100	**Range** 100 - 150	**Range** 150 - 200
us	last	question	book
use	use	use	us
ask	product	ask	get
product	us	product	time
time	time	time	product
get	one	get	use
one	get	one	one
work	help	work	help
help	work	make	make
make	make	help	work

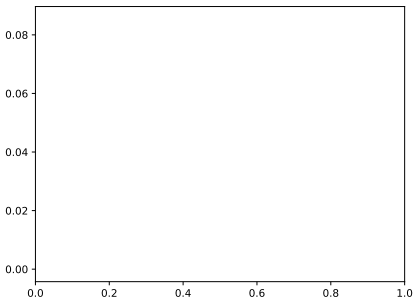
Table A.5: 10 Most Important Words - Category Film and Video

Time 25	**Time** 75	**Time** 125	**Time** 175
put	soldier	door	movi
take	tip	offer	forward
write	track	passion	pay
well	age	reach	much
job	today	section	two
art	aka	two	contact
aka	better	start	upon
internet	mysteri	self	age
today	dream	mysteri	page
better	offer	age	film

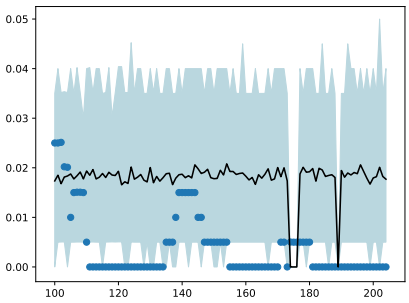
Table A.6: 10 Most Important Words - Category Film and Video

Appendix B

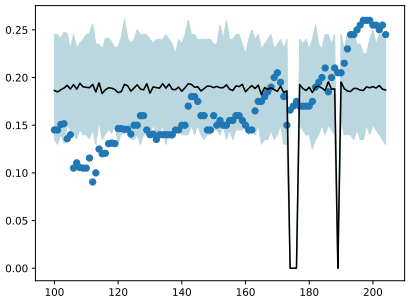
Additional results for chapter 05



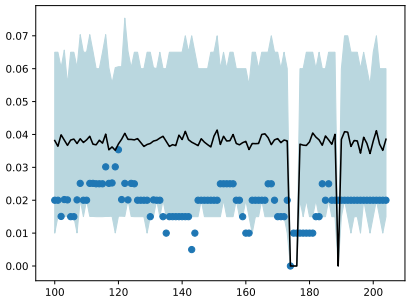
(a) Film & Animation



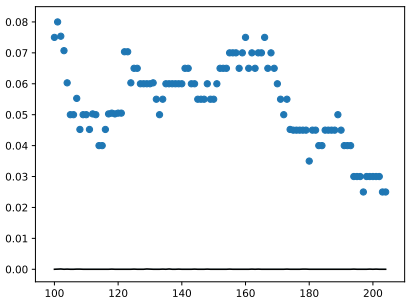
(b) Autos & Vehicles



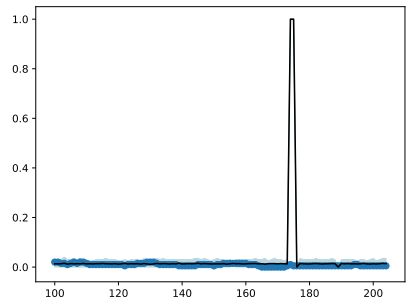
(c) Music



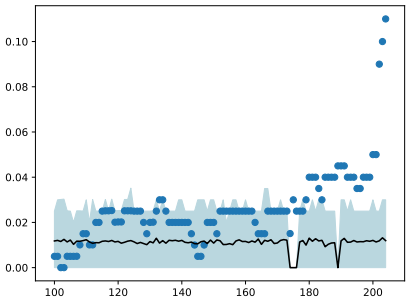
(d) Pets & Animals



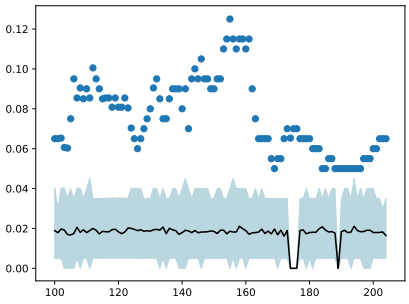
(e) Sports



(f) Short Movies

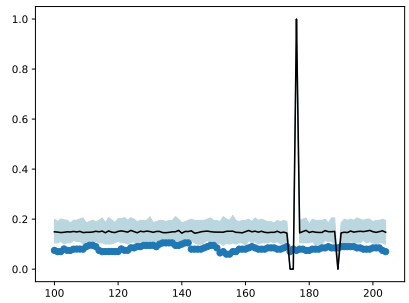


(g) Travel & Events

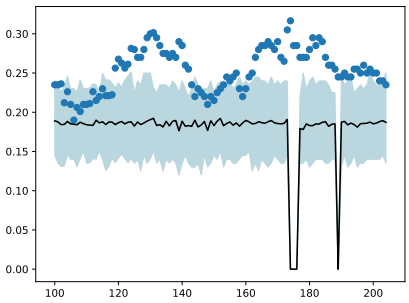


(h) Gaming

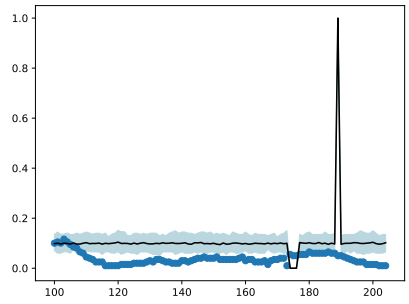
Figure B.1: Observations (dots), Expected values (grey line) and 95% predictive interval - Test set Part 01



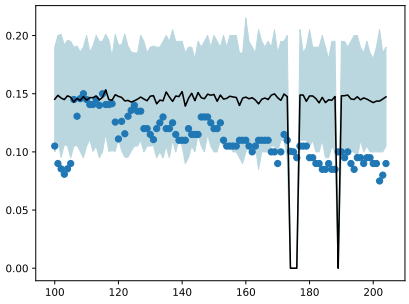
(a) Videobloggingn



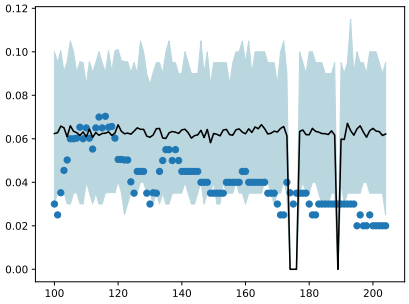
(b) People & Blogs



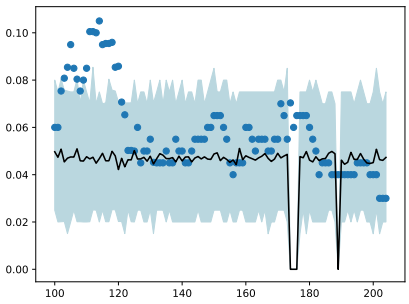
(c) Comedy



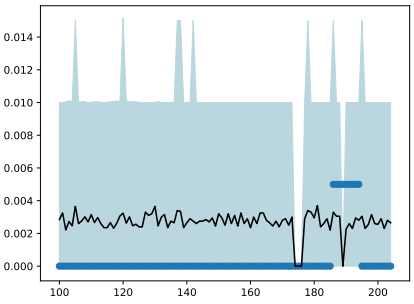
(d) Entertainment



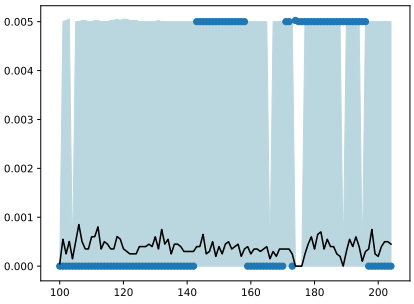
(e) News & Politics



(f) Howto & Style



(g) Education



(h) Science & Technology

Figure B.2: Observations (dots), Expected values (grey line) and 95% predictive interval - Test set Part 02

Bibliography

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[2] Elizabeth Gerber, Julie Hui, and Pei-Yi (Patricia Kuo. Crowdfunding: Why people are motivated to post and fund projects on crowdfunding platforms. In Computer Supported Cooperative Work, volume 10, 02 2012.

[3] Michael Irwin Jordan. Learning in graphical models. Springer Science & Business Media, 1998.

[4] Maya R. Gupta. Theory and Use of the EM Algorithm. Foundations and Trends® in Signal Processing, 4(3):223–296, 2010.

[5] Tommi S. Jaakkola and Michael I. Jordan. A variational approach to Bayesian logistic regression models and their extensions. AISTATS - International Conference on Artificial Intelligence and Statistics, 1996.

[6] Daphne Koller and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

[7] Jim C. Huang and Brendan J. Frey. Cumulative distribution networks and the derivative-sum-product algorithm: Models and inference for cumulative distribution functions on graphs. Journal of Machine Learning Research, 12:301–348, 2011.

[8] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. Introduction to variational methods for graphical models. Machine Learning, 37(2):183–233, 1999.

[9] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association, 112(518):859–877, 2017.

[10] Søren Feodor Nielsen et al. The stochastic em algorithm: estimation and asymptotic results. Bernoulli, 6(3):457–489, 2000.

[11] Ep Xing, Mi Jordan, and S Russell. A generalized mean field algorithm for variational inference in exponential families. UAI - Conference on Uncertainty in Artificial Intelligence, pages 583–591, 2003.

[12] Rajesh Ranganath, Sean M. Gerrish, and David M. Blei. Black Box Variational Inference. In AISTATS - International Conference on Artificial Intelligence and Statistics, 2014.

[13] Yin Cheng Ng, Pawel Chilinski, and Ricardo Silva. Scaling Factorial Hidden Markov Models: Stochastic Variational Inference without Messages. In NIPS - Advances in Neural Information Processing Systems, 2016.

[14] Nicholas J. Foti, Jason Xu, Dillon Laird, and Emily B. Fox. Stochastic variational inference for hidden Markov models. In NIPS - Advances in Neural Information Processing Systems, 2014.

[15] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In ICRL - International Conference on Learning Representations, 2014.

[16] Sebastian Ruder. An overview of gradient descent optimization algorithms. Technical report, Insight Research Centre for Data Analytics, 2016.

[17] Hanna M Wallach. Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning, pages 977–984. ACM, 2006.

[18] David M. Blei and John D. Lafferty. Dynamic Topic Models. In ICML - International Conference on Machine Learning, 2006.

[19] Chris Glynn, Surya T. Tokdar, David L. Banks, and Brian Howard. Bayesian Analysis of Dynamic Linear Topic Models. Bayesian Analysis, pages 1–28, 2018.

[20] David M. Blei and John D. Lafferty. Correlated Topic Models. In NIPS - Advances in Neural Information Processing Systems, pages 147–154, 2006.

[21] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. Scalable Inference for Logistic-Normal Topic Models. In NIPS - Advances in Neural Information Processing Systems, 2012.

[22] Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In ICML - International Conference on Machine Learning, 2012.

[23] Sanjeev Arora, Rong Ge, Frederic Koehler, Tengyu Ma, and Ankur Moitra. Provable Algorithms for Inference in Topic Models. In ICML - International Conference on Machine Learning, 2016.

[24] David M. Blei and Jon D. McAuliffe. Supervised Topic Models. In NIPS - Advances in Neural Information Processing Systems, pages 121–128, 2008.

[25] Sean M. Gerrish and David M. Blei. A Language-based Approach to Measuring Scholarly Impact. In ICML - International Conference on Machine Learning, 2010.

[26] Sungrae Park, Wonsung Lee, and Il-Chul Moon. Associative topic models with numerical time series. Information Processing & Management, 51(5):737–755, 2015.

[27] Sungrae Park, Wonsung Lee, and Il-Chul Moon. Supervised Dynamic Topic Models for Associative Topic Extraction with A Numerical Time Series. In CIKM - International Conference on Information and Knowledge Management, pages 49–54, 2015.

[28] Cheng Zhang and Hedvig Kjellström. How to Supervise Topic Models. Lecture Notes in Computer Science, 8927:500–515, 2015.

[29] Maxim Rabinovich and David M. Blei. The Inverse Regression Topic Model. In ICML - International Conference on Machine Learning, pages 199–207, 2014.

[30] Matt Taddy. Multinomial Inverse Regression for Text Analysis. Journal of the American Statistical Association, 108(503):755–770, 2013.

[31] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. Journal of basic Engineering, 82(1):35–45, 1960.

[32] Yunseong Hwang, Anh Tong, and Jaesik Choi. Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series. In ICML - International Conference on Machine Learning, 2016.

[33] Ronen Feldman. Techniques and applications for sentiment analysis. Communications of the ACM, 56(4):82, 2013.

[34] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2:1–135, 2008.

[35] Bing Liu. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1):1–167, 2012.

[36] Brendan OConnor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series Brendan. In ICWSM - International Conference on Weblogs and Social Media, 2010.

[37] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In ICWSM - International Conference on Weblogs and Social Media, 2011.

[38] CJ J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In AAAI Conference on Weblogs and Social Media, pages 216–225, 2014.

[39] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of Twitter sentiment analysis methods. ACM Computing Surveys, 49(2), 2016.

[40] Michael Dewing. Social Media : An Introduction. Technical report, Canada. Library of Parliament., 2012.

[41] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel a Zighed. Information Diffusion in Online Social Networks: A Survey. ACM SIGMOD Record, 42(2):17–28, 2013.

[42] Yasuko Matsubara, Christos Faloutsos, and B. Aditya Prakash. Rise and Fall Patterns of Information Diffusion : Model and Implications. In KDD - International Conference on Knowledge Discovery and Data Mining, 2012.

[43] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez Rodriguez, and Shuang Li. COEVOLVE : A Joint Point Process Model for Information Diffusion and Network Co-evolution. In NIPS - Advances in Neural Information Processing Systems, 2015.

[44] Yang Yang, Jie Tang, and Cane Wing-ki Leung. RAIN : Social Role-Aware Information Diffusion. In AAAI Conference on Weblogs and Social Media, pages 367–373, 2015.

[45] Manuel Gome-Rodrigues, Jure Leskovec, and Bernhard Schölkopf. Structure and Dynamics of Information Pathways in Online Media. In WSDM - ACM international conference on Web search and data mining, pages 23–32, 2013.

[46] Alexandra Moritz and Joern H Block. Crowdfunding: A literature review and research directions. In Crowdfunding in Europe, pages 25–53. Springer, 2016.

[47] Ethan Mollick. The dynamics of crowdfunding: An exploratory study. Journal of Business Venturing, 29(1):1–16, 2014.

[48] Vincent Etter, Matthias Grossglauser, and Patrick Thiran. Launch hard or go home!: predicting the success of kickstarter campaigns. In S. Muthu Muthukrishnan, Amr El Abbadi, and Balachander Krishnamurthy, editors, COSN, pages 177–182. ACM, 2013.

[49] Venkat Kuppuswamy and Barry L Bayus. Crowdfunding creative ideas: the dynamics of projects backers in kickstarter. SSRN Electronic Journal, 2013.

[50] Kevin Chen, Brock Jones, Isaac Kim, and Brooklyn Schlamp. Kickpredict: Predicting kickstarter success. Technical report, California Institute of Technology, 2013.

[51] Michael D Greenberg, Bryan Pardo, Karthic Hariharan, and Elizabeth Gerber. Crowdfunding support tools: predicting success & failure. In CHI'13 Extended Abstracts on Human Factors in Computing Systems, pages 1815–1820. ACM, 2013.

[52] Calvin Qiu. Issues in crowdfunding: Theoretical and empirical investigation on kickstarter. Available at SSRN 2345872, 2013.

[53] Rick Wash. The value of completing crowdfunding projects. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, ICWSM. The AAAI Press, 2013.

[54] Matthew J. Beal and Zoubin Ghahramani. The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. Bayesian Statistics, pages 453–463, 2003.

[55] Matthew J. Beal and Zoubin Ghahramani. The variational Kalman smoother. Technical report, Gatsby Computational Neuroscience Unit, University College London, 2001.

[56] William H. Greene. Econometric analysis. Pearson Education, 2003.

[57] Guido Consonni and Jean Michel Marin. Mean-field variational approximate Bayesian inference for latent variable models. Computational Statistics and Data Analysis, 52(2):790–798, 2007.

[58] Yuan Qi and Tommi S. Jaakkola. Parameter Expanded Variational Bayesian Methods. In NIPS - Advances in Neural Information Processing Systems, 2007.

[59] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[60] Venkat Kuppuswamy and Barry L Bayus. Crowdfunding creative ideas: The dynamics of project backers in kickstarter. A shorter version of this paper is in" The Economics of Crowdfunding: Startups, Portals, and Investor Behavior"-L. Hornuf and D. Cumming (eds.), 2017.

[61] Chong Wang, John Paisley, and David Blei. Online variational inference for the hierarchical dirichlet process. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 752–760, 2011.

[62] Creighton Heaukulani and Zoubin Ghahramani. Dynamic Probabilistic Models for Latent Feature Propagation in Social Networks. In ICML - International Conference on Machine Learning, volume 28, pages 275–283, 2013.

[63] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436, 2015.

[64] Paulo Freire. Pedagogy of the oppressed. Bloomsbury Publishing USA, 2018.

[65] Schiavone Francesco. Incompetence and managerial problems delaying reward delivery in crowdfunding. Journal of Innovation Economics & Management, 2017.