Multiset Model Selection and Averaging, and Interactive Storytelling

Dipayan Maiti

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Scotland C. Leman, Chair
Naren Ramakrishnan
Inyoung Kim
Leanna House
Chris North
Eric Smith

August 6, 2012
Blacksburg, Virginia

Keywords: Bayesian model selection, Visual analytics, Topic modeling

Multiset Model Selection and Averaging, and INTERACTIVE STORYTELLING

Dipayan Maiti

(ABSTRACT)

The Multiset Sampler [Leman et al., 2009] has previously been deployed and developed for efficient sampling from complex stochastic processes. We extend the sampler and the surrounding theory to model selection problems. In such problems efficient exploration of the model space becomes a challenge since independent and ad-hoc proposals might not be able to jointly propose multiple parameter sets which correctly explain a new proposed model. In order to overcome this we propose a multiset on the model space to enable efficient exploration of multiple model modes with almost no tuning. The Multiset Model Selection (MSMS) framework is based on independent priors for the parameters and model indicators on variables. We show that posterior model probabilities can be easily obtained from multiset averaged posterior model probabilities in MSMS. We also obtain typical Bayesian model averaged estimates for the parameters from MSMS. We apply our algorithm to linear regression where it allows easy moves between parameter modes of different models, and in probit regression where it allows jumps between widely varying model specific covariance structures in the latent space of a hierarchical model.

The *Storytelling algorithm* [Kumar et al., 2006] constructs stories by discovering and connecting latent connections between documents in a network. Such automated algorithms often do not agree with user's mental map of the data. Hence systems that incorporate *feedback* through visual interaction from the user are of immediate importance. We propose a visual analytic framework in which such interactions are naturally incorporated in to the existing *Storytelling* algorithm through a redefinition of the latent topic space used in the similarity measure of the network. The document network can be explored using the newly learned normalized topic weights for each document. Hence our algorithm augments the limitations of human sensemaking capabilities in large document networks by providing a collaborative framework between the underlying model and the user. Our formulation of the problem is a supervised topic modeling problem where the supervision is based on relationships imposed by the user as a set of inequalities derived from tolerances on edge costs from inverse shortest path problem. We show a probabilistic modeling of the relationships based on auxiliary variables and propose a Gibbs sampling based strategy. We provide detailed results from a simulated data and the Atlantic Storm data set.

Acknowledgments

As with any adventure, a PhD has three distinct phases. There is the initial excitement of confronting known unknowns, and unknown unknowns; a phase of utter disillusionment due to conflicts between puerile expectations and the reality of independent research; and finally, an emphatic feeling when everything falls in place for one common but selfish goal of writing a dissertation. In each of these phases, I have been fortunate to have individuals who have taken the roles of friends, philosophers and guides.

First and foremost, I would like to thank my advisor, Dr. Scotland Leman, for moulding my philosophy of statistical thinking. His focus on taking time to frame a problem for me, encouraging approaches from multiple angles, being extremely detailed in checking consistency of my solutions, and having patience in my mistakes and progress, will be lessons that will hold me in stead in the future and beyond research.

To Dr. Chris North, I owe my appreciation for the single most important question in interactive visual analytics: *Why did the user do it?*. This recurring theme in discussions with him has greatly improved my understanding of consumption of analytics by users. I would like to thank Dr. Naren Ramakrishnan for exposing me to algorithms at the convergence of Bayesian statistics and data mining. The realization that the efficacy of a statistical algorithm, however sophisticated it might be, rests to a large degree on its speed and space constraints, was due to interactions with him and his research team. I would like to thank Dr. Leanna House and Dr. Inyoung Kim, for their role as teachers and critics; for exposing me to a wide range of Bayesian models, and at the same time helping me hone my modeling and writing skills. My exposure to Bayesian methodology owes much to their courses. I would like to thank Dr. Eric Smith for important inputs on my proposal, and for having provided every possible departmental help throughout my stay here as a graduate student. I would also like to express my gratitude to the members of the committee who ensured that no financial or hardware constraints came in the way of my research work.

I would like to thank Dr. Eric Vance for his support and guidance throughout my association with the Laboratory for Interdisciplinary Statistical Analysis. His insistence on valuing inter-personal and intra-personal skills in statistical consulting, helped shape my interactions with other researchers.

The quality of any discourse depends on long drawn discussions. A singular characteristic of the graduate school experience is the prolonged interaction with a few like minded graduate students. To this list I would like to add Ciro Velasco-Cruz and Alex Endert. In moments of frustration they have given me hope, in times of boredom we have had stimulating discussions, and in views about life and research, I have often been corrected or vindicated.

Beyond academics, I would like to thank a small group of friends, whom I had unknowingly made my own family in Blacksburg. My dissertation would not have been worth much, had it not been for the precious times I spent with these wonderful individuals; they include Prasun, Jaideep, Suchi, Somik, Atanu, Sai, Ravi, Naro and Sandeep (in the order we got introduced).

Last but not the least, I would like thank my parents and brother for being an endless source of inspiration and encouragement. My dissertation is the result of numerous sacrifices they have made in their lives. To them, I dedicate my dissertation.

Preface

The last couple of years have seen a veritable explosion of articles in popular media related to *big data* or the *data deluge* phenomenon. Without resorting to hard restrictions in terms of size or scope of data and with a view to keeping the definition flexible with fast evolving technologies, *big data* has typically been referred to in literature as data which is not amenable to existing methods of storage or analysis. Special reports from the McKinsey Quarterly [Manyika et al., 2011] and The Economist [eco, 2010] have dealt with this phenomenon in detail in quick succession. It is projected that global data will grow at an annualized rate of almost 40%. A major component of this data is 'multi-dimensional, multi-source, time-varying' unstructured data - this includes data from digital libraries, news sources, claim files, corporate emails and customer opinions from company-wide intranets, biomedical and health information systems, security and defense databases etc. As witnesses to conception of the *social-network* generation, where more than 30 billion pieces of social media content is added to Facebook alone in a month, it has hardly been a leap of faith for most companies to recognize data as an important asset in their functioning. The focus on productivity based on large databases can be explained on one hand by the cheap storage costs and on the other hand by the sharp increase in computing power that supports more sophisticated algorithms. This shift in the focus on productivity based on harnessing data has however been gradual. Owing to the convergence of the aforementioned factors, it is only in recent times that mining of information from databases has been recognized as an active area of interdisciplinary research from the fields of computing, neuroscience, psychology, mathematics, statistics and economics.

Large data immediately translates to large number of parameters in models to be used for inference; hence feature selection becomes a huge challenge. For our first problem, we propose the *Multiset Model Selection* algorithm in a Bayesian model selection setting, that allows efficient selection of multiple viable models. Although we elaborate the strengths of our model selection algorithm with simulated data, our theoretical foundation easily extends to models of any size. Our next problem proposes a solution for text data mining with specific applications to path discovery problems in document networks. We propose an interactive method to *learn* about topics underlying corpus, by incorporating user interaction in to a visual analytic system – the algorithm is termed as INTERACTIVE STORYTELLING.

Multiset Model Selection

The Multiset Sampler [Leman et al., 2009] has previously been deployed and developed for efficient sampling from complex stochastic processes. We extend the sampler and the surrounding theory to model selection problems. In such problems efficient exploration of the model space becomes a challenge since independent and ad-hoc proposals might not be able to jointly propose multiple parameter sets which correctly explain a new proposed model. In order to overcome this we propose a multiset on the model space to enable efficient exploration of multiple model modes with almost no tuning. The Multiset

Model Selection (MSMS) framework is based on independent priors for the parameters and model indicators on variables. We show that posterior model probabilities can be easily obtained from multiset averaged posterior model probabilities in MSMS. We also obtain typical Bayesian model averaged estimates for the parameters from MSMS. We apply our algorithm to linear regression where it allows easy moves between parameter modes of different models, and in probit regression where it allows jumps between widely varying model specific covariance structures in the latent space of a hierarchical model.

INTERACTIVE STORYTELLING Algorithm

The Storytelling algorithm [Kumar et al., 2006] constructs stories by discovering and connecting latent connections between documents in a network. Such automated algorithms often do not agree with user's mental map of the data. Hence systems that incorporate *feedback* through visual interaction from the user are of immediate importance. We propose a visual analytic framework in which such interactions are naturally incorporated in to the existing *Storytelling* algorithm through a redefinition of the latent topic space used in the similarity measure of the network. The document network can be explored using the newly learned normalized topic weights for each document. Hence our algorithm augments the limitations of human sensemaking capabilities in large document networks by providing a collaborative framework between the underlying model and the user.

The general question that both the algorithms try to answer is : *How do we select features or underlying latent variables appropriately in large data sets?*. The two approaches vary in that the former is devoid of user feedback, while the basis of the latter is feedback. In Part I we discuss Multiset Model Selection and in Part II we discuss INTERACTIVE STORYTELLING.

Contents

List of Figures

List of Tables

Part I

Multiset Model Selection

Chapter 1

Introduction

Model selection is an ubiquitous problem in statistics. Problems like variable subset selection in linear and non-linear regression, either with continuous or polychotomous responses (George and McCulloch [1993], Albert and Chib [1993], Vila et al. [2000]), selection of categorical variables in contingency tables using log-linear models (Albert [1995], Dahinden et al. [2007]), identification of mixtures and change points (Green [1995], Richardson and Green [1997]), can all be formulated as a model selection problem. Inference using Bayesian model selection is of value to the researcher, if he is interested in selecting the model that *best* explains some observed data or in including model uncertainty in his decisions. In the former case, known as model selection, the researcher obtains parameter estimates corresponding to the model with the highest posterior probability. In the latter case, known as model averaging, the researcher weighs the model specific parameters by their posterior model probabilities. Details of model selection and averaging is discussed in Hoeting et al. [1999].

Model selection is typically done using comparative measures such as Akaike Information Criteria, Bayes Information Criteria (BIC) or Bayes factors. For our purposes here, we primarily focus on Bayes factors which can be computed using various methods: it is estimated directly using the marginal data likelihood, using samples from the posterior distribution of model parameters obtained from an MCMC procedure when posterior density is known [Chib, 1995], or by a harmonic average of the data likelihoods based on MCMC samples [Newton and Raftery, 1994], Laplace approximation [Tierney et al., 1989], BIC [Kass and Raftery, 1995], or importance sampling [Geweke, 1989]. For large model spaces, stochastic exploration of the model space is the only recourse to the researcher. Methods exploring the model space using MCMC include Gibbs sampling based strategies like Stochastic Search Variable Selection (SSVS) [George and McCulloch, 1993], the Bayes factor based framework using the continuous-discrete mixture prior [Geweke, 1996], the pseudo-prior based framework [Kuo and Mallick, 1998] and the reversible jump algorithm [Green, 1995] that allows trans-dimensional jumps.

1.1 Challenges in Model Selection and Multiset Model Selection

Irrespective of whether the model space is explored using a transdimensional algorithm like reversible jump or constant-dimensional algorithms with the mixture priors, the main challenge is efficient exploration of the model space. For across-model jumps in all these algorithms, the proposal has to ensure that the parameter values for the proposed model are in the vicinity of a high likelihood region of the proposed model. Devising proposals in Euclidean space is simpler because of the inherent concept of distance and neighborhood. For a target density in the Euclidean space, the proposal variance is chosen such that a desired proportion of the proposed moves are accepted – large proposal variance causes too many rejections and small proposal variance causes most moves to be accepted which lead to poor mixing. But in model selection algorithms the similarity between two models and consequently the neighborhood of a model is a very tricky concept, and hence an alternate understanding of proposal mechanism or perhaps a re-formulation of the target density, is required.

Figure 1.1: Model jump from M_2 and M_1 where the marginal for β_1 under the two models are close. Dashed distributions are marginals for model M_1. Such a jump will be accepted often.

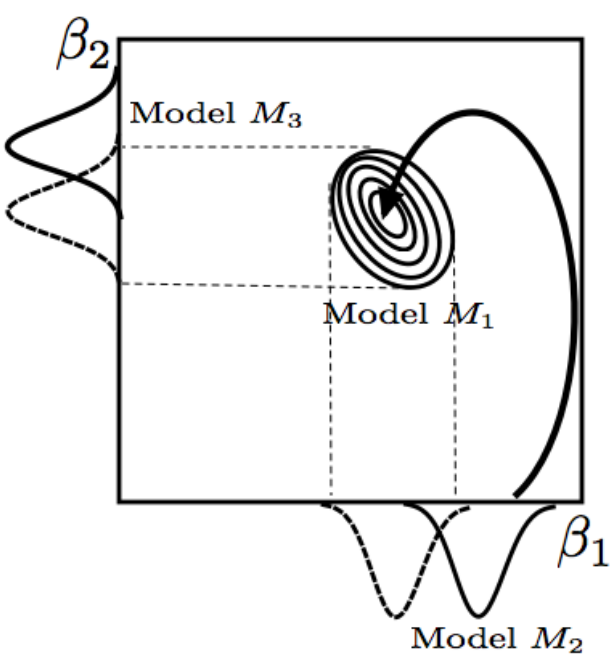
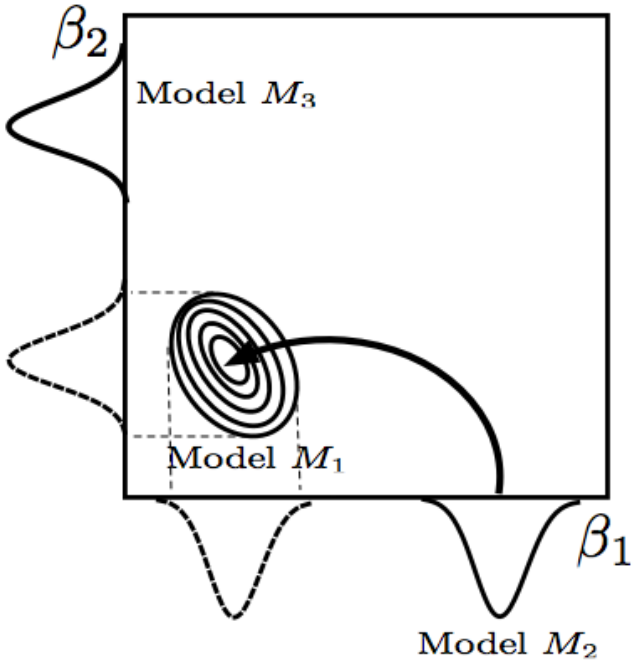


Figure 1.2: Model jump from M_2 and M_1 where the marginal for β_1 under the two models are far apart. Dashed distributions are marginals for model M_1. Such a jump will almost always be rejected.



Consider for example, a model space composed of the full model M_1 (having parameters

(β_1, β_2)) and reduced models M_2 and M_3 with parameters β_1 and β_2 only respectively, as shown in Figures 1.1 and 1.2. The posterior model probabilities are such, that all the models are of interest to the researcher. In such a scenario, we want to make an across-model jump from M_2 to M_1. In Figure 1.1, since the distribution of β_1 under model M_2 and the marginal distribution of β_1 under model M_1 are close to each other, we can easily make a jump from model M_2 to M_1 by proposing β_1 locally for M_1 (we assume β_2 being proposed independently). However, if the marginal distribution of β_1 under model M_1 lies far from distribution of β_1 under model M_2 as in Figure 1.2, such a jump becomes almost impossible, even not getting in to the much harder problem of proposing around the right vicinity of β_2 in model M_1. In general, jumps between two nested models will be rejected, if the marginal distributions of a parameter shared between the two models are very different and a local proposal for the parameter is used. The alternative solution is an extensive tuning step.

We address the problem of efficient exploration of the model space by extending the multiset sampler [Leman et al., 2009] to model selection. Instead of sampling from a target density $f(x, y)$, Leman's sampler sampled from a new target density, $f^*(x, \{y_1, \ldots, y_K\}) \propto f(x, y_1) + \cdots + f(x, y_K)$, defined on a state space, $(x, \{y_1, \ldots, y_K\})$. Here, $\{.,.\}$ is a multiset of size K. Similarly in multiset model selection (MSMS), instead of sampling from the original space of the model and their respective parameter subspaces, $(\boldsymbol{\beta^{(k)}}, M_k)$, we sample from a modified density which is defined on $(\boldsymbol{\beta}, \{M_1, M_2\})$, the product space of multiset over models and the complete set of parameters i.e. we define a multiset on the model space while the space of model parameters remain untouched. Here and in subsequent examples, we use a multiset of size two, although the general approach can be extended to multisets of size more than two. The new target density on this space, from which we intend to sample from is given by,

$$f^*(\boldsymbol{\beta}, \{M_1, M_2\}|\boldsymbol{Y}) \propto f(\boldsymbol{\beta}, M_1|\boldsymbol{Y}) + f(\boldsymbol{\beta}, M_2|\boldsymbol{Y}), \tag{1.1}$$

where,

$$f(\boldsymbol{\beta}, M_k|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}|\boldsymbol{\beta}, M_k)\pi(\boldsymbol{\beta}|M_k)\pi(M_k)}{p(\boldsymbol{Y})}.$$

In a linear regression setting, consider a multiset $\{M_1, M_1\}$ at a step of the sampling procedure and $\boldsymbol{\beta}$ in the high probability region for M_1. Let M_1^* be a model which has a high posterior probability but the high probability region for the parameters in M_1^*, $\boldsymbol{\beta^*}$, is not easily reachable using a simple proposal from the current value of $\boldsymbol{\beta}$. In fact, $\boldsymbol{\beta}$ is such a low probability region for M_1^* that any move relying on the ratio $f(\boldsymbol{\beta}, M_1^*|\boldsymbol{Y})/f(\boldsymbol{\beta}, M_1|\boldsymbol{Y})$ will be rejected. However, in MSMS the acceptance ratio is a function of $(f(\boldsymbol{\beta}, M_1^*|\boldsymbol{Y}) + f(\boldsymbol{\beta}, M_1|\boldsymbol{Y}))/(f(\boldsymbol{\beta}, M_1|\boldsymbol{Y}) + f(\boldsymbol{\beta}, M_1|\boldsymbol{Y}))$ (ignoring the ratio of the proposals) and hence a move to the multiset $\{M_1, M_1^*\}$ is accepted with probability close to 0.5. The sampling step for $f(\boldsymbol{\beta}|\{M_1, M_1^*\}, \boldsymbol{Y})$ ensures that $\boldsymbol{\beta}$ is sampled from the high prob-

ablity region for M_1^*. Hence the multiset allows easy moves between parameter modes of different models without requiring much tuning for the proposal for $\boldsymbol{\beta}$. Similarly, in a probit regression setting with auxiliary variable $\boldsymbol{Z}$ (Albert and Chib [1993]) and models denoted by $\boldsymbol{\gamma}$, the MSMS acceptance ratio for a proposed move from $\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(1)}}\}$ to $\{\boldsymbol{\gamma^{(1)*}}, \boldsymbol{\gamma^{(1)}}\}$ depends on the ratio $(f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)*}}, \boldsymbol{Y}) + f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)}}, \boldsymbol{Y}))/\ (f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)}}, \boldsymbol{Y}) + f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)}}, \boldsymbol{Y}))$. We show in Section 5 that the only difference between $f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)}}, \boldsymbol{Y})$ and $f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)*}}, \boldsymbol{Y})$ at a Gibb's step is in their covariance structures. Hence our algorithm easily allows jumps between widely varying model specific covariance structures in the latent space of a probit hierarchical model. By contrast, most model selection algorithms require extensive tuning for exploration of all high probability models. Alternate automatic proposal strategies also exist, but they are harder to implement and often complicated. The greatest strength of MSMS as a model selection problem is that such a simple reformulation allows efficient exploration of the model space with almost no tuning. In spite of using an augmented version of the model space in MSMS, the multiset model probabilities (to be explained in Section 3) are ordered according to their posterior model probabilities. The posterior distribution of the model parameters averaged over multisets in MSMS is also preserved i.e. they are the same as typical Bayesian model averaged parameter distributions.

In Section 2, we give a short overview of Bayesian model selection and motivation for model averaging. We give a short overview of current strategies for efficient exploration of the model space to compare different proposal strategies to our algorithm. We also discuss the prior framework in *best subset selection algorithms* in a regression setting which we use in formulating MSMS. In Section 3, we define the probability density over a multiset, formulate the MSMS algorithm for continuous data in a linear regression setting, and discuss the benefits of the algorithm along with the simplicity in its implementation. In Section 4, we show how standard proposal strategies fail on a toy dataset, but MSMS performs much better in comparison. In our examples, we do not use model averaged predictive ability as a criteria to compare the performances of our methodology with standard strategies; instead, we use the total variation distance between the estimated posterior probability distribution over the model space and the expected model probability distribution. We end this section with a discussion on why MSMS is able to navigate the state space more efficiently than other strategies and a note on the inference about multiset averaged parameter estimates. In Section 5 we apply the algorithm to binary data using a probit regression model, provide a mixture formulation of the Gibbs step and hence further insight in to the MSMS, and end with a conclusion.

Chapter 2

Bayesian Model Selection & Averaging

Consider $\boldsymbol{\mathcal{M}} = (M_1, \ldots, M_K)$, a class of models under consideration. Model $M_k, k \in (1, \ldots, K)$ is associated with a parameter subspace $\boldsymbol{\theta_k} \in \boldsymbol{\Theta}$. $\boldsymbol{\Theta}$, represented by $\cup_{k=1}^{K} \boldsymbol{\theta_k}$ is the parameter space encompassing the parameter sub-spaces of all the models in $\boldsymbol{\mathcal{M}}$. Bayesian inference in a model selection problem is a coherent framework that is based on the joint posterior distribution over the space of models and model-specific parameters, $(\boldsymbol{\theta_k}, M_k)$, given the data. Let $\pi(M_k)$ be the prior probability that data was generated from model M_k and $\pi(\boldsymbol{\theta_k}|M_k)$ be the prior distribution over the parameter sub-space $\boldsymbol{\theta_k}$ for model M_k. For a sampling distribution $p(\boldsymbol{Y}|\boldsymbol{\theta_k}, M_k)$, the posterior model probability is,

$$p(M_k|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|M_k)\pi(M_k)}{\sum_{l=1}^{K} p(\boldsymbol{Y}|M_l)\pi(M_l)} = \frac{\int_{\boldsymbol{\theta_k}} p(\boldsymbol{Y}|\boldsymbol{\theta_k}, M_k)\pi(\boldsymbol{\theta_k}|M_k)d\boldsymbol{\theta_k}\pi(M_k)}{p(\boldsymbol{Y})}. \tag{2.1}$$

The marginal likelihood conditional on a model also follows from the identity,

$$\frac{1}{p(\boldsymbol{Y}|M_k)} = \int \frac{p(\boldsymbol{\theta_k}|M_k, \boldsymbol{Y})}{p(\boldsymbol{Y}|\boldsymbol{\theta_k}, M_k)} d\boldsymbol{\theta_k}.$$

The researcher might choose the model based on the highest posterior model probability or compare posterior odds of two competing models,

$$\frac{p(M_k|\boldsymbol{Y})}{p(M_l|\boldsymbol{Y})} = \frac{p(\boldsymbol{Y}|M_k)}{p(\boldsymbol{Y}|M_l)} \times \frac{\pi(M_k)}{\pi(M_l)} = BF(M_k|M_l) \times \frac{\pi(M_k)}{\pi(M_l)}. \tag{2.2}$$

The term $BF(M_k|M_l)$ is known as the Bayes factor of model M_k compared to model M_l. Alternately, when the Bayes factors are known, they can be used to calculate the posterior model probability,

$$p(M_k|\boldsymbol{Y}) = \frac{1}{1 + \sum_{l=1, l\neq k}^{K} BF(M_l|M_k) \times \frac{\pi(M_l)}{\pi(M_k)}}.$$

Note that the Bayes factor requires that we know the integrated likelihood $\mathcal{C}_k$ for all the models $M_k \in \boldsymbol{\mathcal{M}}$ where, $\int_{\boldsymbol{Y}} \mathcal{L}(\boldsymbol{Y}|\boldsymbol{\theta_k}, M_k)d\boldsymbol{Y} = 1/\mathcal{C}_k$.

It might be required to incorporate model uncertainty in some situations. Not only is it more desirable for better estimation of uncertainty for our quantities of interest, [Madigan and Raftery, 1993] show that Bayesian model averaged predictive ability based on the logarithmic scoring rule is better than the predictive ability of any individual model in the model space:

$$-\mathbb{E}(log(p(\boldsymbol{Y_f}|\boldsymbol{Y}))) \leq -\mathbb{E}(log(p(\boldsymbol{Y_f}|M_k, \boldsymbol{Y}))), \forall M_k \in \boldsymbol{\mathcal{M}}$$

where the model averaged predictive distribution for a future value $\boldsymbol{Y_f}$ is

$$p(\boldsymbol{Y_f}|\boldsymbol{Y}) = \sum_{k=1}^{K} p(\boldsymbol{Y_f}|M_k, \boldsymbol{Y})p(M_k|\boldsymbol{Y}) \tag{2.3}$$

$$p(\boldsymbol{Y_f}|M_k, \boldsymbol{Y}) = \int_{\boldsymbol{\theta_k}} p(\boldsymbol{Y_f}|\boldsymbol{\theta_k}, M_k, \boldsymbol{Y})p(\boldsymbol{\theta_k}|M_k, \boldsymbol{Y})d\boldsymbol{\theta_k} \tag{2.4}$$

and the expectation is taken over the joint space of $\boldsymbol{Y_f}$ and $\boldsymbol{\mathcal{M}}$.

Similarly, if we want to minimize squared error loss in our prediction, the best prediction is the posterior expected value,

$$\mathbb{E}(\boldsymbol{Y}_f|\boldsymbol{Y}) = \sum_{k=1}^{K} \mathbb{E}(\boldsymbol{Y_f}|M_k, \boldsymbol{Y})p(M_k|\boldsymbol{Y})$$

2.1 Strategies for Exploration of Model Space

In reversible jump MCMC, various strategies for automatic design of proposals have been suggested for efficient exploration of the model space. Broadly these fall into two categories dealing with two different aspects of the jump – *centring* which refers to the mean of the proposed parameter subspace and *scaling* which refers to the variance or covariance structure of the proposed parameter subspace. Green [2003] suggested that for small sized model selection problems, information from random-walk Metropolis samplers for the models can be used to design the mapping $g(.)$ such that proposed parameters are in the vicinity of their model specific posterior modes. An alternative approach also suggested by Green [2003] is based on moment matching between the original state and the

proposed state in a transdimensional jump. Brooks et al. [2003b] gives the most recent definitive work for generating automatic proposals for the reversible jump algorithm under some fixed canonical mapping $g(.)$. The implementations of these automatic proposals will be discussed later for a simple example to allude to some of the challenges that still remain. Another approach is *adaptive MCMC* where the proposal is tuned automatically and adaptively as the sampler runs such that a new proposal depends on the history of the parameter space covered under that model. Hastie [2005] and Green and Mira [1999] provide some methods in this direction. The saturated space approach due to Brooks et al. [2003b] also augments the state space of the Markov chain with auxiliary variables such that the exploration is over a constant dimensional space and hence attempts to circumvent the problem of difficult cross-dimensional model jumps. Another approach to allow better mixing for model selection is based on changing the target density as in transdimensional simulated annealing [Brooks et al., 2003a] where an *annealed* penalized likelihood criteria is optimized for model selection. Jasra et al. [2007] extended the parallel tempering algorithm for reversible jump MCMC by simulating a *population* of MCMC chains that interact via various crossover moves. In the sense that we modify the target density, our approach is closest to methods like transdimensional simulated annealing.

Chapter 3

Multiset Model Selection

3.1 Concept of a Multiset

Let $\mathcal{M}$ be a set of possible models. A multiset of size K on $\mathcal{M}$ is a *bag* of K elements from the set $\mathcal{M}$. For $\mathcal{M} = (M_1, M_2, M_3, M_4)$ and $K = 2$, the possible elements of the multiset are $\{M_1, M_1\}$, $\{M_1, M_2\}$, $\{M_1, M_3\}$, $\{M_1, M_4\}$, $\{M_2, M_2\}$, $\{M_2, M_3\}$, $\{M_2, M_4\}$, $\{M_3, M_3\}$, $\{M_3, M_4\}$ and $\{M_4, M_4\}$. Denote the set of all possible multisets by $\mathcal{S}$ and the size of $\mathcal{S}$ by T. Here $T = 10$. For the general case of a model space $\mathcal{M}$=$(M_1, \ldots, M_N)$ with N models, the cardinality of $\mathcal{S}$= $\{s_1, \ldots, s_T\}$, the set of multisets of size K on the model space $\mathcal{M}$ is given by $T = \binom{N+K-1}{K}$. Any $s \in \mathcal{S}$, will be denoted by $\uplus_{i=1}^{K} M_i$. If the multiset $s \in \mathcal{S}$ contains $\delta(s)$ distinct models denoted by M_1,M_2,. . . ,$M_{\delta(s)}$ with multiplicities a_1,a_2,. . . ,$a_{\delta(s)}$ respectively, this denotation can be further simplified as

$$\uplus_{i=1}^{K} M_i = \uplus_{j=1}^{\delta(s)} {M_j}^{a_j}.$$

As a special case, if the multiplicity of a specific model M_1 in the multiset is known to be a, it will be denoted by

$$M_1^a \uplus_{j=1}^{\delta(s-M_1)} M_j^{a_j},$$

where $\delta(s - M_1)$ are the number of distinct models in the multiset s, excluding M_1. Since for most intensive model selection methods, the number of models N is extremely large (2^p models with p covariates and intercept), we assume the $N >> K$. The true marginal probability for model $M_i \in \mathcal{M}$ is given by $p(M_i|\boldsymbol{Y})$. The probability on a multiset $s \in \mathcal{S}$ is denoted by $p(s|\boldsymbol{Y})$. The *multiset probability for model* M_i is,

$$p^*(M_i|\boldsymbol{Y}) = \sum_{s=M_i \uplus_{j=1}^{\delta(s-M_i)} M_j} \frac{1}{K} p(s|\boldsymbol{Y}) + \sum_{s=M_i^2 \uplus_{j=1}^{\delta(s-M_i)} M_j} \frac{2}{K} p(s|\boldsymbol{Y}) + \cdots + \sum_{s=M_i^K \uplus_{j=1}^{\delta(s-M_i)} M_j} \frac{K}{K} p(s|\boldsymbol{Y}) \quad (3.1)$$

3.2 Prior Structures for Model Selection

While some of the model selection ideas in Section **??** were proposed for subset selection in a linear regression setting, their general approaches can easily be expanded to a broader class of model selection problems. For problems with nested models, proposing a new model (with its new parameter vector) is equivalent to proposing a new vector of indicator variables where each indicator represents a covariate. In a regression setting, this can be accomplished by using a mixture proposal with a point masses at zero and perhaps a normal distribution as the two components of a proposed parameter. Hence model selection in these scenarios is not exactly an exploration of a transdimensional model space – it is an exploration of the parameter space of the *full model* where at any point of the MCMC sampling step, some indicators are *switched on* and the remaining are *switched off*. Godsill [2001] discusses model selection algorithms and unifies them in a single framework.

A model selection algorithm relies on a prior being defined on the space of model and the parameters. For best subset selection in a linear regression setting using SSVS, the prior on θ_i is expressed as a mixture normal distribution using a latent variable γ_i. The latent variable $\gamma_i = 0$, if θ_i is not in the model and $\gamma_i = 1$, if θ_i is in the model. The hierarchical prior for θ_i is

$$\theta_i|\gamma_i \sim (1-\gamma_i)N(0,\tau_i^2) + \gamma_i N(0, c_i\tau_i^2), \text{ and } \gamma_i|p_i \sim Bernoulli(p_i)$$

If $\gamma_i = 0$, the prior variance τ_i^2 is chosen so small that the posterior estimate of the effect size after shrinkage towards zero can effectively be assumed to be of no practical importance. The constant $c_i > 1$ is chosen such that when $\gamma_i = 1$ i.e. when θ_i is in the model, the flatness of the prior on θ_i allows the posterior estimate of θ_i to be dictated by the data. The prior probability that the ith predictor is in the model is p_i. The choice of c_i and τ_i^2 can be motivated by data from previous experiments or semi-automatic approaches as suggested in George and McCulloch [1993]. Samples generated from the full conditional distributions for θ_i and γ_i are used to construct the posterior joint distribution over the model and parameter space. Geweke [1996] also uses a mixture prior for θ_i with a point mass at zero, and a normal prior, as two components of the mixture,

$$\theta_i|\gamma_i \sim (1-\gamma_i)\delta_0 + \gamma_i N(0,\sigma_i^2), \text{ and } \gamma_i|p_i \sim Bernoulli(p_i)$$

The typical Gibbs step of sampling from the full conditional for θ_i is replaced by a sampling step based on the conditional Bayes factor. Based on the conditional Bayes factor, θ_i is sampled from a point mass at zero (signifying that θ_i is not in the model) or the appropriate conditional distribution of θ_i at that step of the Gibbs sampler. Such a formulation is equivalent to a Metropolis-Hastings step using a continuous-discrete mixture proposal which has a point mass at zero as one component and a normal proposal as another com-

ponent (with certain restrictions for boundary models like the *intercept only* model or the full model).

3.3 Prior Structure for Multiset Model Selection

In this article, we focus throughout on proper priors for model-specific parameters $\boldsymbol{\theta_k}$. We consider a finite model space $\boldsymbol{\mathcal{M}}$ so that the prior on the model space is also proper. Based on the prior framework in Kuo and Mallick [1998], we introduce some notation since this will be used for the remainder of the article. Each model M_k is represented by a vector of $0-1$ indicator variables $\boldsymbol{\gamma^{(k)}} = (\gamma_1, \ldots, \gamma_p)$ for p covariates. We specify independent priors on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ with $\beta_j \sim N({\beta^0}_j = 0, \sigma_j^2)$ and $\gamma_j \sim Bernoulli(p_j), j = 1, \ldots, p$. The above prior on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\gamma}$ can also be formulated in terms of $\boldsymbol{\vartheta}$ where $\boldsymbol{\vartheta} = (\gamma_1\beta_1, \ldots, \gamma_p\beta_p)^T$ with the linear predictor $\boldsymbol{X\beta}$ replaced by $\boldsymbol{X\vartheta}$, and $\vartheta_j = \beta_j$ if $\gamma_j = 1$, and $\vartheta_j = 0$ if $\gamma_j = 0$, independent of the value of β_j. Hence ϑ_j is the effect size corresponding to the j^{th} covariate when it is in the model. The unknown variance σ^2 has an Inverse Gamma prior. The dimension of $\boldsymbol{\beta}$ is the number of possible covariates in the full model, denoted by p. $\boldsymbol{\beta^{(k)}}$ is said to be in model M_k while $\boldsymbol{\beta^{(-k)}}$ is said to be out of model M_k. Our model selection procedure can be thought of sampling from a high dimensional $\boldsymbol{\beta}$, a sub-vector $\boldsymbol{\beta^{(k)}}$ of which is sampled from an appropriate posterior distribution depending on a model M_k, while the remaining parameters represented by $\boldsymbol{\beta^{(-k)}}$, are sampled from the prior (or equivalently, the corresponding effect sizes are equated to zero). Note that sampling from the prior for $\boldsymbol{\beta^{(-k)}}$ is equivalent to failing to update the prior since the likelihood under model M_k has no information about $\boldsymbol{\beta^{(-k)}}$.

Under such a parameterization, model selection algorithms in linear regression (and also in generalized linear models, contingency tables) are not strictly trans-dimensional – the dimension of the model is $always$ the dimension of the full model i.e. p, with indicators corresponding to $\boldsymbol{\beta^{(-k)}}$ being merely *switched off* using indicator variables. The target density to sample from is:

$$\begin{aligned} p(\boldsymbol{\beta}, M_k|\boldsymbol{Y}) &\propto p(\boldsymbol{Y}|\boldsymbol{\beta}, M_k)\pi(\boldsymbol{\beta})\pi(M_k) \propto p(\boldsymbol{Y}|\boldsymbol{\beta^{(k)}}, M_k)\pi(\boldsymbol{\beta^{(k)}})\pi(\boldsymbol{\beta^{(-k)}})\pi(M_k) \\ &\propto p(\boldsymbol{\beta^{(k)}}|M_k, \boldsymbol{Y})p(M_k|\boldsymbol{Y})\pi(\boldsymbol{\beta^{(-k)}}). \end{aligned} \tag{3.2}$$

Under this prior framework,

$$\begin{aligned} p(M_k|\boldsymbol{Y}) &= \int f(\boldsymbol{\beta}, M_k|\boldsymbol{Y})d\boldsymbol{\beta} \\ &= (1/p(\boldsymbol{Y})) \int f(\boldsymbol{Y}|\boldsymbol{\beta^{(k)}}, M_k)\pi(\boldsymbol{\beta^{(k)}}|M_k)d\boldsymbol{\beta^{(k)}} \times \int \pi(\boldsymbol{\beta^{(-k)}}|M_k)d\boldsymbol{\beta^{(-k)}}\pi(M_k) \\ &= (1/p(\boldsymbol{Y}))\, f(\boldsymbol{Y}|M_k)\pi(M_k). \end{aligned} \tag{3.3}$$

In MSMS, we define a multiset on the model space i.e. the space of the model indicator vector $\boldsymbol{\gamma}$. Hence an equivalent representation of the multiset $\{M_1, M_2\}$ of size two, is $\{(\gamma_1^{(1)}, \ldots, \gamma_p^{(1)}), (\gamma_1^{(2)}, \ldots, \gamma_p^{(2)})\}$.

3.4 General Formulation of the Multiset Model Selection Algorithm

By Equation 1.1, our new target density for $K = 2$ is defined by $f^*(\boldsymbol{\beta}, (M_1, M_2)|\boldsymbol{Y}) \propto f(\boldsymbol{\beta}, M_1|\boldsymbol{Y}) + f(\boldsymbol{\beta}, M_2|\boldsymbol{Y})$, where the constant of proportionality with N possible models is given by:

$$\begin{aligned} C^* &= \left(\sum_{\{M_i, M_j\} \in \mathcal{S}} \int f(\boldsymbol{\beta}, M_i|\boldsymbol{Y}) d\boldsymbol{\beta} + \int f(\boldsymbol{\beta}, M_j|\boldsymbol{Y}) d\boldsymbol{\beta} \right)^{-1} = \left(\sum_{\{M_i, M_j\} \in \mathcal{S}} (p(M_i|\boldsymbol{Y}) + p(M_j|\boldsymbol{Y})) \right)^{-1} \\ &= (N+1)^{-1}. \end{aligned} \tag{3.4}$$

Based on our earlier example of $\mathcal{M} = (M_1, M_2, M_3, M_4)$, $C^* = 1/5$. Similarly, for $K = 3$ and $N = 4$, $C^* = 1/15$ and for $K = 4$ and $N = 4$, $C^* = 1/70$. The apparent simplicity of this target density however has four important benefits –

- Allows a more efficient exploration of the model space, facilitating between-model jumps inspite of $\boldsymbol{\beta}$ being in a low likelihood region for at least one model in the multiset,
- Posterior model probability $p(M_k|\boldsymbol{Y})$ can be extracted from the multiset based posterior model probabilities $p^*(M_j|\boldsymbol{Y}), M_j \in \mathcal{M}$,
- Ordering of the models based on their posterior probabilities is invariant to sampling from the new target density i.e. $p(M_k|\boldsymbol{Y}) \leq p(M_j|\boldsymbol{Y}) \Leftrightarrow p^*(M_k|\boldsymbol{Y}) \leq p^*(M_j|\boldsymbol{Y}), M_j, M_k \in \mathcal{M}$, and,
- Multiset averaged posterior distribution of the model parameters has an easy interpretation within our prior framework.

The remainder of this section discusses these points in detail.

3.5 Sampling for Multiset Model Selection

To sample from the target density given by Equation 1.1 we use Gibbs sampling and sample iteratively from the full conditional distributions $f(\boldsymbol{\beta}|\{M_1, M_2\}, \boldsymbol{Y})$ and $f(\{M_1, M_2\}|\boldsymbol{\beta}, \boldsymbol{Y})$.

The full conditional distribution for $\boldsymbol{\beta}$ is a mixture distribution with mixture weights proportional to the posterior probabilities of the models in the multiset:

$$f(\boldsymbol{\beta}|\{M_1, M_2\}, \boldsymbol{Y}) \propto f(\boldsymbol{\beta^{(1)}}|\boldsymbol{Y}, M_1)p(M_1|\boldsymbol{Y})\pi(\boldsymbol{\beta^{(-1)}}) + f(\boldsymbol{\beta^{(2)}}|\boldsymbol{Y}, M_2)p(M_2|\boldsymbol{Y})\pi(\boldsymbol{\beta^{(-2)}}).$$

While model specific posterior distribution of the parameters can be obtained in some situations, posterior model probabilities are unknown and hence a Metropolis step is used to sample from $f(\boldsymbol{\beta}|\{M_1, M_2\}, \boldsymbol{Y})$. In our implementation of the algorithm we use a dependent proposal $q(\boldsymbol{\beta^*}|\boldsymbol{\beta})$ for $\boldsymbol{\beta}$. Note that such a formulation allows for a sub-vector of $\boldsymbol{\beta}$ to evolve although it is in neither of the models M_1 and M_2. The evolution for this sub-vector is based solely on the prior. The Metropolis-Hastings ratio for a move for $\boldsymbol{\beta}$ is given by:

$$\alpha_{\boldsymbol{\beta}} = \frac{f(\boldsymbol{\beta}^*, M_1|\boldsymbol{Y}) + f(\boldsymbol{\beta}^*, M_2|\boldsymbol{Y})}{f(\boldsymbol{\beta}, M_1|\boldsymbol{Y}) + f(\boldsymbol{\beta}, M_2|\boldsymbol{Y})} \times \frac{q(\boldsymbol{\beta^*}|\boldsymbol{\beta})}{q(\boldsymbol{\beta}|\boldsymbol{\beta}^*)}.$$

To sample from $f(\{M_1, M_2\}|\boldsymbol{\beta}, \boldsymbol{Y})$ we use a Metropolis step after proposing a multiset $\{M_1^*, M_2\}$ using a proposal $q(\{M_1, M_2\} \rightarrow \{M_1^*, M_2\})$. The Metropolis-Hastings ratio is given by:

$$\alpha_S = \frac{f(\boldsymbol{\beta}, M_1^*|\boldsymbol{Y}) + f(\boldsymbol{\beta}, M_2|\boldsymbol{Y})}{f(\boldsymbol{\beta}, M_1|\boldsymbol{Y}) + f(\boldsymbol{\beta}, M_2|\boldsymbol{Y})} \times \frac{q(\{M_1^*, M_2\} \rightarrow \{M_1, M_2\})}{q(\{M_1, M_2\} \rightarrow \{M_1^*, M_2\})}.$$

While in general the proposed model M_1^* can be any model from the model space, for high dimensional models, local model moves are usually chosen. In such cases, $M_1^* = (\gamma_1^{(1)}, \ldots, \gamma_j^{(1)*}, \ldots, \gamma_p^{(1)})$ where a change in only the jth indicator is proposed. For our examples we use symmetric proposals throughout unless otherwise mentioned.

3.6 Extracting True Model Probabilities from Multiset Model Probabilities

Once multiset samples have been obtained from the Gibbs sampler, true posterior model probabilities can be extracted from posterior multiset probabilities. The approach is similar to what is described in Leman et al. [2009]. We describe it for $K = 2$ and then provide the general formula. Consider the model space $\mathcal{M} = (M_1, M_2, M_3, M_4)$ described above for the linear regression problem of $\boldsymbol{Y}$ versus two covariates and an intercept. From Equation 3.1, the posterior model probability from multiset samples for model M_1 (also called

the multiset model probability for M_1) is,

$$\begin{aligned} p^*(M_1|\boldsymbol{Y}) &= p^*(\{M_1, M_1\}|\boldsymbol{Y}) + \frac{1}{2}p^*(\{M_1, M_2\}|\boldsymbol{Y}) + \frac{1}{2}p^*(\{M_1, M_3\}|\boldsymbol{Y}) + \frac{1}{2}p^*(\{M_1, M_4\}|\boldsymbol{Y}) \\ &= C^*(3p(M_1|\boldsymbol{Y}) + \frac{1}{2}), \end{aligned}$$

since each of $p^*(\{M_1, M_i\}|\boldsymbol{Y}) = C^*(p(M_1|\boldsymbol{Y}) + p(M_i|\boldsymbol{Y})), i = 1, \ldots, 4$. If $p(M_1|\boldsymbol{Y})$ is extremely small, it is *inflated* to a value of $C^*/2$ and a very large value of $p(M_1|\boldsymbol{Y})$ is *shrunk* by a factor of $3C^*$ (with an additive effect of $C^*/2$). Note that $C^* = 5$ based on Equation 3.4. Similarly, for $K = 3$ and $N = 4$, $p^*(M_1|\boldsymbol{Y}) = \frac{1}{15}(7p(M_1|\boldsymbol{Y}) + 2)$, and for $K = 4$ and $N = 4$, $p^*(M_1|\boldsymbol{Y}) = \frac{1}{35}(14p(M_1|\boldsymbol{Y}) + \frac{21}{4})$. Hence multiset based posterior model probabilities are a flatter version of the true model probabilities. As a corollary, the ordering of the models based on their posterior model probabilities does not change under multiset model selection. While for small or medium dimensional problems such an inverse calculation is possible, for extremely high dimensional problems, this might be a prohibitive approach due to the expensive computing involved. However, this result guarantees that the top few models based on multiset model probabilities are the same as the ones based on posterior model probabilities. For a model space $\mathcal{M}$=$(M_1, \ldots, M_N)$ with N models and $N >> K$, the posterior model probability from multiset samples is

$$p^*(M_i|\boldsymbol{Y}) = p(M_i|\boldsymbol{Y}) \sum_{a=1}^{K} \frac{a(N-K)}{K(N-1)} \binom{N+(K-a)-1}{K-a} + \sum_{a=1}^{K} \frac{\binom{N+(K-a)-1}{K-a} \times a(K-a)}{(N-1)K}, i = 1, \ldots, N.$$

For a very large N and $K = 2$, $p^*(M_i|\boldsymbol{Y})/p^*(M_j|\boldsymbol{Y}) \approx p(M_i|\boldsymbol{Y})/p(M_j|\boldsymbol{Y})$ i.e. with equal prior weight on all models, the ratio of the posterior model probabilities obtained from the multiset samples is approximately equal to their Bayes factor.

3.7 Multiset Averaged $\boldsymbol{\beta}$ Estimates

The posterior distribution of $\boldsymbol{\beta}$ marginalized over the space of multiset is a mixture distribution with mixture components defined by model specific posterior $\boldsymbol{\beta}$ estimates and weights defined by posterior model probabilities (using Equation 3.2) :

$$\begin{aligned} f^*(\boldsymbol{\beta}|\boldsymbol{Y}) &\propto \sum_{\{M_i, M_j\} \in \mathcal{S}} f(\boldsymbol{\beta}, M_i|\boldsymbol{Y}) + f(\boldsymbol{\beta}, M_j|\boldsymbol{Y}) \propto \sum_{M_i \in \mathcal{M}} f(\boldsymbol{\beta}, M_i|\boldsymbol{Y}) \\ &\propto \sum_{M_i \in \mathcal{M}} f(\boldsymbol{\beta}^{(i)}|\boldsymbol{Y}, M_i)\pi(\boldsymbol{\beta}^{(-i)})p(M_i|\boldsymbol{Y}) = f(\boldsymbol{\beta}|\boldsymbol{Y}). \end{aligned}$$

It should be noted that for each model M_i, the sub-vector of $\boldsymbol{\beta}$ not in the model is sampled from its prior with the weight defined by the posterior model probability. Hence under

our chosen prior framework, the posterior distribution of $\boldsymbol{\beta}$ from MSMS is exactly the same as the Bayesian model averaged distribution of $\boldsymbol{\beta}$.

Chapter 4

A Linear Regression Example

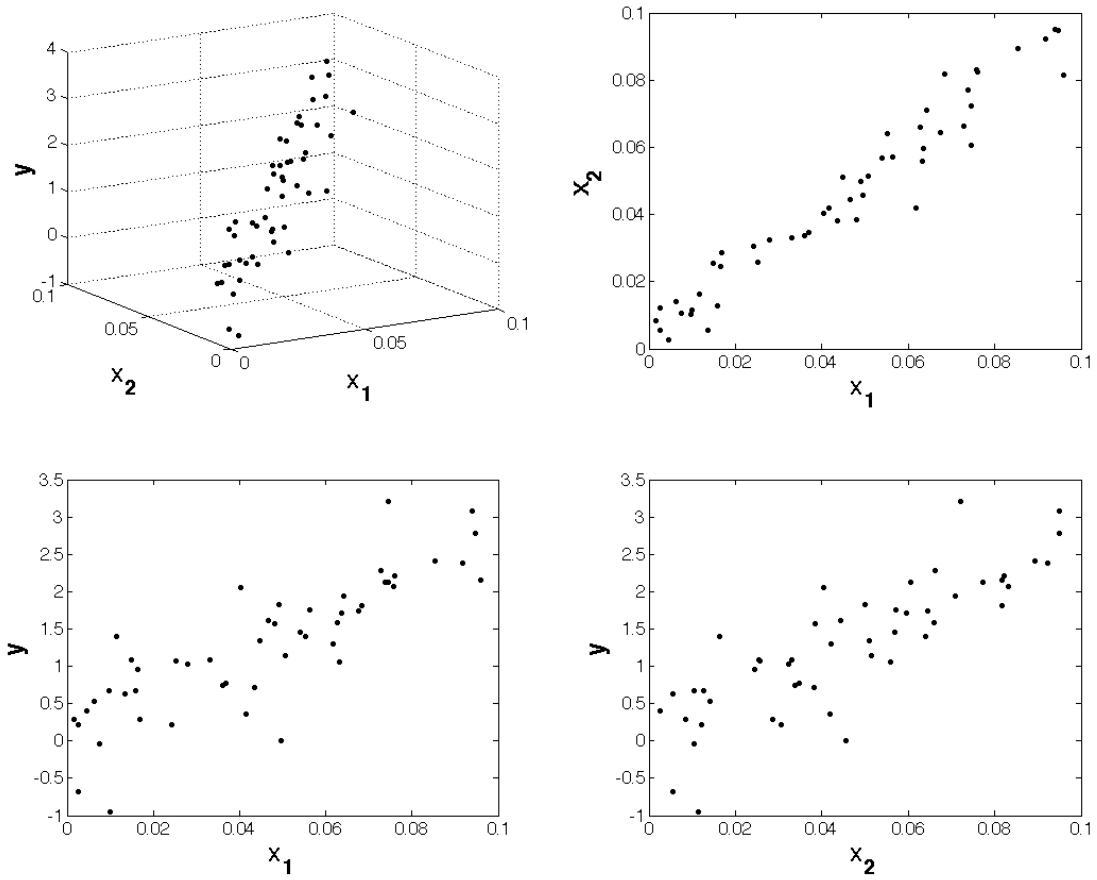
Consider the model selection problem in linear regression with two possible predictors $\mathbf{x_1}$ and $\mathbf{x_2}$ and perhaps an unknown intercept, since it is the simplest of model selection problems. The error is assumed to be independent Gaussian with mean zero and unknown variance σ^2. Our model space consists of the models $M_i, i = 1, 2, 3, 4$ with corresponding parameter sub-spaces $(\boldsymbol{\beta^{(i)}}, \sigma^2)$. Notationally, $\boldsymbol{\beta^{(1)}} = (\beta_{10}, \beta_{11}, \beta_{12})$, $\boldsymbol{\beta^{(2)}} = (\beta_{20}, \beta_{21}\}$, $\boldsymbol{\beta^{(3)}} = (\beta_{30}, \beta_{32})$ and $\boldsymbol{\beta^{(4)}} = (\beta_{40})$. β_{ij} corresponds to the jth covariate in the ith model and a missing β_{ij} implies that the jth covariate is *not in* the ith model. For our simulated data, maximum likelihood estimates of model specific parameters and the marginal probability of the data conditional on the model, are given below. The calculation for $P(\boldsymbol{Y}|M)$ is based on Gaussian priors for $\boldsymbol{\beta^{(i)}}$ with mean equal to their maximum likelihood estimates and a prior variance of 15. The prior for σ^2 is a Gaussian distribution centered and left truncated at zero with a variance of 0.5.

Table 4.1: Model specific maximum likelihood estimates and model probabilities for simulated data

Model	$\beta_{0,}$	$\beta_{1,}$	$\beta_{2,}$	σ^2	$P(\boldsymbol{Y}\|M)$	$P(M\|\boldsymbol{Y})$
M_1	0.0047	13.2227	14.6431	0.2443	1.8360×10^{-18}	0.4049
M_2	0.0622	26.8236	-	0.2483	1.3932×10^{-18}	0.3073
M_3	-0.0142	-	28.0717	0.2473	1.3051×10^{-18}	0.2878
M_4	1.2375	-	-	0.8160	2.465×10^{-31}	≈ 0

M_1 is the *full* model and models M_2 and M_3 will be referred to as *reduced* models. It is evident from Table 4.1 that there is a relationship between the effect sizes for $\mathbf{x_1}$ and $\mathbf{x_2}$ in the full and reduced models. The slopes in the full model M_1 are close to half of the slopes in any of the reduced models M_2 and M_3. As explained in Section **??**, across-model jumps from M_1 to M_2 will almost surely be rejected, if β_{11} is proposed locally in the

Figure 4.1: Visual summary of 50 datapoints simulated for linear regression. a) 3-d plot of $\boldsymbol{Y}$ versus $\mathbf{x_1}$ and $\mathbf{x_2}$ b) Plot of $\mathbf{x_1}$ and $\mathbf{x_2}$ c) Plot of $\mathbf{y}$ and $\mathbf{x_1}$ d) Plot of $\mathbf{y}$ and $\mathbf{x_2}$.



neighborhood of 13 (which is to be expected under a simple dependent proposal scheme). As shown in Table 4.2, the sampler ends up spending over 70% of its time in M_1 under automatic proposal strategies. One might argue that a larger proposal variance when proposing β_{11} in model M_1 might solve the issue, but this is the exhausting tuning step that a an intelligent sampling strategy should attempt to avoid.

A sampling strategy that allows us to incorporate model specific parameter estimates in designing proposals will be the best, but in its absence, any prior information about the relationships between the model specific parameters should be exploited. We compare results of the reversible jump algorithm where the researcher has information about the relationship between model specific modes for a subset of the models, to a simpler model selection algorithm with a *non-informative* proposal structure. A non-informative proposal does not make use of any prior relationships between the model specific modes. We show, that the informative proposal provides more accurate posterior model probabilities than the latter algorithm but still suffers from convergence to the true posterior distribution.

When such information is missing, alternatives are automatic proposal strategies due to Green [2003] and Brooks et al. [2003b] (details of which will be provided later). We show that both the cases do not provide satisfactory approaches, especially when the researcher does not want to spent too much time in tuning the sampler. Under such scenarios, we show that multiset model selection performs better than any of the alternatives.

4.1 Proposal Strategies in Model Selection

We first devise the Reversible-Jump Metropolis-Hastings sampler [Green, 1995] and allude to the fact that unless the structure of βs is not known a priori in the models, an ad-hoc proposal scheme might not be able to sample efficiently from the model space. To propose models we use the transition matrix P below. We allow transitions from the

$P =$

	M_1	M_2	M_3	M_4
M_1	1/3	1/3	1/3	0
M_2	1/3	1/3	0	1/3
M_3	1/3	0	1/3	1/3
M_4	0	1/3	1/3	1/3

$full$ model M_1 to any of the $reduced$ models through a $death$ process by removing any of the two available parameters in the $full$ model. When in any of the $reduced$ models, we allow transition to 1) the $full$ model through a $birth$ process by introducing the missing parameter in the model and a transition to 2) the *intercept only* model through the *death* process. The transition matrix also allows for a finite probability that the model remains unchanged after transition. Adding of a new parameter using only the *birth* process applies to the *intercept only* model.

The prior information that the researcher has about the relationships of the βs between the models, enables him to propose parameters around their model specific modes; specifically for our example, the researcher knows that the slope parameter for any of the reduced models can be split equally between the slope parameters of the full model. Hence the birth process involves splitting the value of the existing slope parameter equally between the two new slope parameters of the full model. The death process involves adding the two slope parameters of the full model in to a single slope parameter for the reduced model.

Using notation in Green [1995], the transition $M_2 \rightarrow M_1$ involves sampling $u_{(1)}$ from a proposal function $q_{(M_2 \rightarrow M_1)}\ (u_{(1)}|\beta_{20}, \beta_{21})$ to match dimensions and subsequently defining a bijective mapping $g(.)$ from $(\boldsymbol{\beta^{(2)}}, u_{(1)}) = (\beta_{20}, \beta_{21}, u_{(1)})$ to $\boldsymbol{\beta^{(1)}} = (\beta_{10}, \beta_{11}, \beta_{12})$ which in

our case we define as,

$$\begin{aligned}\beta_{10} &= \beta_{20} \\ \beta_{11} &= \frac{1}{2}\beta_{21} + u_{(1)} \\ \beta_{12} &= \frac{1}{2}\beta_{21} - u_{(1)}.\end{aligned} \tag{4.1}$$

The Jacobian of the transformation based on the mapping in (4.1) is given by,

$$\left|\frac{\partial(\beta_{10},\beta_{11},\beta_{12})}{\partial(\beta_{20},\beta_{21},u_{(1)})}\right| = \begin{vmatrix} \frac{\partial\beta_{10}}{\partial\beta_{20}} & \frac{\partial\beta_{10}}{\partial\beta_{21}} & \frac{\partial\beta_{10}}{\partial u_{(1)}} \\ \frac{\partial\beta_{11}}{\partial\beta_{20}} & \frac{\partial\beta_{11}}{\partial\beta_{21}} & \frac{\partial\beta_{11}}{\partial u_{(1)}} \\ \frac{\partial\beta_{12}}{\partial\beta_{20}} & \frac{\partial\beta_{12}}{\partial\beta_{21}} & \frac{\partial\beta_{12}}{\partial u_{(1)}} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 1 \\ 0 & \frac{1}{2} & -1 \end{vmatrix} = 1.$$

Ideally in situations like these where the relationship between parameters of the current and proposed model is known, it makes more sense to keep the proposal variance for $u_{(1)}$ small. Note that the mapping $g(.)$ specifically uses the prior information that the researcher has about the relationship between the slope parameters in the $full$ and the $reduced$ model; the *dimension matching* step which involves sampling $u_{(1)}$ and the bijective mapping together essentially ensure that β_{11} (and β_{12}) is proposed around a mean of $\frac{1}{2}\beta_{21}$ with some random noise specified by $u_{(1)}$. This increases the probability of accepting between-model jumps since under the *appropriate* mapping $g(.)$, the proposed parameter will be in the proximity of the local mode for the proposed model. The acceptance ratio $\alpha_{(b)}$ for the birth $type$ transition $M_2 \rightarrow M_1$ is given by,

$$\alpha_{(b)} = \frac{P(\boldsymbol{Y}|\boldsymbol{\beta^{(1)}}, M_1)\pi(\boldsymbol{\beta^{(1)}}|M_1)\pi(M_1)P(M_1 \rightarrow M_2)}{P(\boldsymbol{Y}|\boldsymbol{\beta^{(2)}}, M_2)\pi(\boldsymbol{\beta^{(2)}}|M_2)\pi(M_2)P(M_2 \rightarrow M_1)q_{(M_2\rightarrow M_1)}(u_{(1)}|\beta_{20},\beta_{21})} \times \left|\frac{\partial(\beta_{10},\beta_{11},\beta_{12})}{\partial(\beta_{20},\beta_{21},u_{(1)})}\right|, \tag{4.2}$$

where $P(\boldsymbol{Y}|\boldsymbol{\beta^{(i)}}, M_i)$ is the sampling density of the data given the model and corresponding parameter, $\pi(.)$ the respective priors on the parameter sub-spaces and the model space, and P the transition matrix specified above. Similarly for a death $type$ transition $M_1 \rightarrow M_2$, we propose parameters $(\boldsymbol{\beta^{(2)}}, u^{(2)}) = (\beta_{20}, \beta_{21}, u_{(2)})$ for the reduced model from the full model parameters $\boldsymbol{\beta^{(1)}} = (\beta_{10}, \beta_{11}, \beta_{12})$ deterministically using the mapping $g^{-1}(.)$ given by:

$$\begin{aligned}\beta_{20} &= \beta_{10} \\ \beta_{21} &= \beta_{11} + \beta_{12} \\ u_{(2)} &= \frac{1}{2}(\beta_{11} - \beta_{12}).\end{aligned} \tag{4.3}$$

For the move 'in the other direction' we use the proposal function $q_{(M_2\rightarrow M_1)}(u_{(2)}|\beta_{20},\beta_{21})$.

The acceptance ratio $\alpha_{(d)}$ for the death *type* transition $M_1 \rightarrow M_2$ is given by,

$$\alpha_{(d)} = \frac{P(\boldsymbol{Y}|\boldsymbol{\beta^{(2)}}, M_2)\pi(\boldsymbol{\beta^{(2)}}|M_2)\pi(M_2)P(M_2 \rightarrow M_1)q_{(M_2 \rightarrow M_1)}(u_{(2)}|\beta_{20}, \beta_{21})}{P(\boldsymbol{Y}|\boldsymbol{\beta^{(1)}}, M_1)\pi(\boldsymbol{\beta^{(1)}}|M_1)\pi(M_1)P(M_1 \rightarrow M_2)} \times \left|\frac{\partial(\beta_{20}, \beta_{21}, u_{(2)})}{\partial(\beta_{10}, \beta_{11}, \beta_{12})}\right|, \tag{4.4}$$

where $u_{(2)} = \frac{1}{2}(\beta_{11} - \beta_{12})$ and the Jacobian of transformation is given by,

$$\left|\frac{\partial(\beta_{20}, \beta_{21}, u_{(2)})}{\partial(\beta_{10}, \beta_{11}, \beta_{12})}\right| = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = 1$$

based on the mapping defined in (4.3). As expected this is the inverse of the Jacobian for the mapping $g(.)$. However the researcher has no prior information about the relationship between the parameter values in the *intercept only* model M_4 and a model with a single parameter like M_2 and M_3. Hence, for a birth type transition from M_4 to M_2 or M_3 we propose a new parameter $u_{(3)}$ using an independent proposal and choose the identity function as the transformation $g(.)$. For a move proposed move from M_4 to M_2,

$$\begin{aligned} \beta_{20} &= \beta_{40} \\ \beta_{21} &= u_{(3)}. \end{aligned}$$

The Jacobian of the transformation is one and the acceptance ratio will be of the same form as given in equation (4.2). Apart from the $death$ and the $birth$ type moves, we also allow an $update$ move in which the model does not change but the parameters in the model are updated. For an update move $M_1 \rightarrow M_1$, where the parameters $\boldsymbol{\beta^{(1)}} = (\beta_{10}, \beta_{11}, \beta_{12})$ are updated to $\boldsymbol{\beta^{*(1)}} = (\beta^*_{10}, \beta^*_{11}, \beta^*_{12})$, the dimension matching step in Green's algorithm involves proposing $(\beta^*_{10}, \beta^*_{11}, \beta^*_{12})$ using a proposal $q_{(M_1 \rightarrow M_1)}(\beta^*_{10}, \beta^*_{11}, \beta^*_{12}|\beta_{10}, \beta_{11}, \beta_{12})$ and an identity map between $(\beta_{10}, \beta_{11}, \beta_{12})$ and $(\beta^*_{10}, \beta^*_{11}, \beta^*_{12})$.The Jacobian of transformation $\left|\frac{\partial(\beta^*_{10}, \beta^*_{11}, \beta^*_{12})}{\partial(\beta_{10}, \beta_{11}, \beta_{12})}\right|$ is obviously one in this case. The acceptance ratio $\alpha_{(e)}$ for the update *type* transition $M_1 \rightarrow M_1$ is given by,

$$\alpha_{(e)} = \frac{\mathcal{L}(\boldsymbol{\beta^{*(1)}}, M_1|Y)\pi(\boldsymbol{\beta^{*(1)}}|M_1)\pi(M_1)q_{(M_1 \rightarrow M_1)}(\beta_{10}, \beta_{11}, \beta_{12}|\beta^*_{10}, \beta^*_{11}, \beta^*_{12})}{\mathcal{L}(\boldsymbol{\beta^{(1)}}, M_1|Y)\pi(\boldsymbol{\beta^{(1)}}|M_1)\pi(M_1)q_{(M_1 \rightarrow M_1)}(\beta^*_{10}, \beta^*_{11}, \beta^*_{12}|\beta_{10}, \beta_{11}, \beta_{12})} \tag{4.5}$$

However, the above approach is infeasible when the researcher does not have prior information about the relationship of the βs in different models or when the number of models is so large that each model can not be analyzed marginally to obtain any insights about the 'local' modes of the parameters in each model. In such a scenario, the reversible jump algorithm for the transition $M_2 \rightarrow M_1$ involves sampling $u_{(1)}$ from a proposal function $q_{(M_2 \rightarrow M_1)}(u^{(1)}|\beta_{20}, \beta_{21})$ to match dimensions as before, but the bijective mapping $g(.)$ from $(\boldsymbol{\beta^{(2)}}, u_{(1)}) = (\beta_{20}, \beta_{21}, u_{(1)})$ to $\boldsymbol{\beta^{(1)}} = (\beta_{10}, \beta_{11}, \beta_{12})$ does not take into consideration any

prior knowledge about the relationship of the βs between the reduced and the full model. A straight forward $g(.)$ in this case might be,

$$\begin{aligned}\beta_{10} &= \beta_{20} \\ \beta_{11} &= \beta_{21} \\ \beta_{12} &= u_{(1)}.\end{aligned} \tag{4.6}$$

Maximum likelihood based estimates for the full model might however be utilized if known, and in that case our mapping $g(.)$ becomes,

$$\begin{aligned}\beta_{10} &= \beta_{20} \\ \beta_{11} &= \beta_{21} \\ \beta_{12} &= {\beta_2}^{(ML)} + u_{(1)},\end{aligned} \tag{4.7}$$

where $({\beta_0}^{(ML)}, {\beta_1}^{(ML)}, {\beta_2}^{(ML)})$ are the maximum likelihood estimates for the full model. In both the cases the Jacobian of transformation is one. Except for the mapping $g(.)$, the equations for $\alpha_{(b)}$,$\alpha_{(d)}$ and $\alpha_{(e)}$ remains the same as above.

One idea discussed in Green [2003], formulates the mapping $g(.)$ in reversible jump in terms of posterior covariances of $\boldsymbol{\beta^{(k)}}$, such that under some assumptions of the transition matrix P (probability of proposing a model jump from $\boldsymbol{M_k}$ to $\boldsymbol{M_l}$) and the form of posterior target densities $p(\boldsymbol{\beta^{(k)}}|\boldsymbol{Y}, M_k)$, the chain of MCMC samples automatically satisfy detailed balance. We will assume for our example, the lower triangular matrix $B_k = chol((\boldsymbol{X_k}^{\boldsymbol{T}}\boldsymbol{X_k})^{-1})$ and our interpretation of R will be a tuning parameter (scalar) in the proposal $q(u)$ used for scaling. In our simulated data, the estimated noise for each model is comparable and hence it makes sense for a scalar value of R. If the noise fluctuates highly between models, such a choice of R is not appropriate. For model specific means for $\boldsymbol{\beta^{(k)}}$ in the mapping $g(.)$, we will use the maximum likelihood estimate, $\boldsymbol{\mu_k} = ({\boldsymbol{X_k}}^T\boldsymbol{X_k})^{-1}{\boldsymbol{X_k}}^T\boldsymbol{Y}$. Hence the mapping is given by:

$$\boldsymbol{\beta}^{(l)} = \begin{cases} \boldsymbol{\mu}_l + B_l \left[R B_k^{-1} \left(\boldsymbol{\beta}^{(k)} - \mu_k \right) \right]_1^{n_l} & \text{if } n_l < n_k \\ \boldsymbol{\mu}_l + B_l R B_k^{-1} \left(\boldsymbol{\beta}^{(k)} - \mu_k \right) & \text{if } n_l = n_k \\ \boldsymbol{\mu}_l + B_l R \begin{bmatrix} B_k^{-1} \left(\boldsymbol{\beta}^{(k)} - \mu_k \right) \\ u \end{bmatrix} & \text{if } n_l > n_k, \end{cases} \tag{4.8}$$

and the acceptance ratio is given by:

$$\alpha = \frac{p(\boldsymbol{\beta^{(l)}}, M_l|\boldsymbol{Y})P(M_l \to M_k)\,|B_l|}{p(\boldsymbol{\beta^{(k)}}, M_k|\boldsymbol{Y})P(M_k \to M_l)\,|B_k|} \times \begin{cases} q(u) & \text{if } n_l < n_k \\ 1 & \text{if } n_l = n_k \\ q(u)^{-1} & \text{if } n_l > n_k \end{cases} .$$

In a second automatic proposal strategy based on Brooks et al. [2003b], the *centring* of the proposed parameter is based on *weak-identifiablity*. Hence for our linear regression case, the new parameter is proposed around zero, and the mapping $g(.)$ is an identity map. The *scaling* is based on the zeroth order method, in which the proposal variance ${\sigma_q}^2$ is chosen such that the probability of accepting a move from the current state to the proposed centered state is one. The acceptance probability for a birth move from $(\boldsymbol{\beta^{(k)}}, M_k)$ to $(\boldsymbol{\beta^{(l)}} = (\boldsymbol{\beta^{(k)}}, 0), M_l)$ is given by,

$$\alpha_{(b)} = \frac{P(\boldsymbol{Y}|[\boldsymbol{\beta^{(k)}}, 0], M_l)\pi(\boldsymbol{\beta^{(l)}}|M_l)\pi(M_l)P(M_l \to M_k)}{P(\boldsymbol{Y}|\boldsymbol{\beta^{(k)}}, M_k)\pi(\boldsymbol{\beta^{(k)}}|M_k)\pi(M_k)P(M_k \to M_l)q_{(M_l \to M_k)}(0|{\sigma_q}^2)} \times \left|\frac{\partial(\boldsymbol{\beta^{(l)}})}{\partial(\boldsymbol{\beta^{(k)}}, u)}\right|_{u=0} .$$

Under an independent Gaussian prior centered around zero for the parameters, a Gaussian proposal $q(.)$, and identity mapping (Jacobian equals unity), the proposal variance equals the prior variance if $\alpha_{(b)} = 1$.

4.2 Comparing Results from Proposals Strategies and Multiset Model Selection

For our implementation of the reversible jump algorithm with informative proposal, $q_{(M_i \to M_j)}(.|\boldsymbol{\beta^{(i)}})$ is an independent Gaussian proposal $N(.|0, \sigma_q^2 = 1)$ for any choice of M_i and M_j $(i \neq j)$, $q_{(M_i \to M_i)}(.|\boldsymbol{\beta^{(i)}})$ for the update move is a dependent multivariate Gaussian proposal $MN(.|\boldsymbol{\beta^{(i)}}, {\sigma_q}^2 I_{n_i} = (1)I_{n_i})$ and the prior $\pi(\boldsymbol{\beta^{(i)}}|M_i)$ is a multivariate Gaussian $MN(.|\boldsymbol{0}, \sigma_\beta^2 I_{n_i} = (100)I_{n_i})$, where n_i is the dimensionality of model M_i. Apart from the unknown effects sizes, we also need to estimate the unknown variance σ^2, of the Gaussian error term in the model. We use for $q(.|\sigma^2)$, $N^+(.|\sigma^2, 0.1)$, a Gaussian proposal centered at the previous value of σ^2 with a variance of 0.1 and left truncated at zero. The prior for σ^2 is a Gaussian distribution centered and left truncated at zero with a variance of 1. The results of the reversible jump algorithm with different proposal structures as discussed above is in Table 4.2. The results are based on ten independent runs of the algorithm, each run up to $100,000$ iterations. The ordering of the posterior model probabilities obtained from the reversible jump algorithm using the informative proposal is as expected although the exact values indicate that the sampler was occasionally stuck in M_1. The model specific MAP estimates are close to the maximum likelihood estimates of the parameters in the

Table 4.2: MAP estimates of model specific parameters from Reversible jump algorithm

Proposal structure	Model	$\hat{\beta}_{0,}$	$\hat{\beta}_{1,}$	$\hat{\beta}_{2,}$	$\hat{\sigma}^2$	$\hat{p}(M\|\boldsymbol{Y})$	Total variation distance (Kendal's τ correlation) ($\hat{p}(M\|\boldsymbol{Y})$ vs $p(M\|\boldsymbol{Y})$)
Reversible jump with informative proposal	M_1	0.0322	9.3843	18.1791	0.2533	0.7177	0.3129 (1)
	M_2	0.1533	24.9108	-	0.2738	0.1520	
	M_3	0.0567	-	26.6709	0.2691	0.1303	
	M_4	1.3936	-	-	1.3440	0.0001	
Reversible jump with simple proposal	M_1	0.0085	13.8284	13.9884	0.2121	0.5594	0.1552 (0.6667)
	M_2	0.0702	26.6968	-	0.2161	0.1552	
	M_3	-0.0101	-	28.0141	0.2164	0.2847	
	M_4	0.1259	-	-	0.4445	0.0006	
Reversible jump with automatic proposal ($R = 1$)	M_1	0.0653	11.6410	15.0493	0.2556	0.7286	0.3238 (0.6667)
	M_2	0.1249	25.1940	-	0.2647	0.1134	
	M_3	0.1730	-	23.9664	0.2626	0.1579	
	M_4	-	-	-	-	0	

reduced models. For the model selection algorithm without an informative proposal, we use a Metropolised Gibbs sampler to sample iteratively from the full conditional distributions of $\boldsymbol{\beta}$ and M_k. Our proposal for $\boldsymbol{\beta}$ is a dependent multivariate Gaussian proposal $MN(.|\boldsymbol{\beta}, {\sigma_q}^2 I_p = (1)I_p)$ and the proposal for σ^2 is a truncated Gaussian proposal $N^+(.|\sigma^2, 0.1)$. The priors are the same as before. While the model specific estimates are very close to the maximum likelihood estimates, the posterior model probabilities are not ordered based on their expected marginal likelihoods given the model in Table 4.1. Hence the Kendal's τ correlation coefficient between the ordering of the models based on non-informative proposal and the true ordering is less than one (0.6667). In this case, for most of the runs the sampler gets stuck in the full model. Although the posterior model probabilities using Green's strategy of automatic proposal ($R = 1$) are ranked based on their true model probabilities, the sampler gets stuck in the full model.

The results of MSMS on our simulated data with $K = 2, 3, 4$ is summarized in Table 4.3 with the prior variance for $\beta_j = 100$. For every value of K, the results are based on ten independent runs of MMS, each with $100,000$ iterations. To propose a new multiset, we first randomly choose one element of the multiset (i.e. a specific model) and then randomly propose to change it to any of the existing models. Hence our multiset proposal is symmetric. The proposal for β_j and σ^2 remain unchanged from the implementation of model selection in our earlier section. The posterior model probabilities are ranked based on their true model probabilities. In terms of total variation distance, MSMS performs better than reversible jump with informative, simple and automatic proposals. Although in Table 4.2 the total variation distance between the estimated posterior model probability distribution using a simple proposal and the true model probability distribution, is the least, the model probabilities are not in the expected order and hence Kendal's τ distance

Table 4.3: Model specific estimates from Multiset Model Selection

Multiset Size	Model	*Estimated* Multiset Model Probabilities ($\hat{p}(M_i^*\|\boldsymbol{Y})$)	True Multiset Model Probabilities ($p(M_i^*\|\boldsymbol{Y})$)	*Estimated* Model Probabilities ($\hat{p}(M_i\|\boldsymbol{Y})$)	Total variation distance (Kendal's τ correlation) ($\hat{p}(M\|\boldsymbol{Y})$ vs $p(M\|\boldsymbol{Y})$)
$K=2$	M_1	0.3329	0.3429	0.3882	0.0566 (1)
	M_2	0.2935	0.2844	0.3225	
	M_3	0.2488	0.2727	0.2480	
	M_4	0.1249	0.1000	0.0415	
$K=3$	M_1	0.3095	0.3223	0.3775	0.1671 (1)
	M_2	0.2732	0.2767	0.2997	
	M_3	0.2502	0.2676	0.2504	
	M_4	0.1671	0.1333	0.0724	
$K=4$	M_1	0.3373	0.3120	0.4682	0.1603 (1)
	M_2	0.2424	0.2729	0.2310	
	M_3	0.2315	0.2651	0.2038	
	M_4	0.1818	0.1500	0.0970	

is non-zero. The Kendal's τ correlation between the estimated ordering of posterior model probabilities from MSMS and the true model probabilities is one for all values of $K = 2, 3, 4$.

4.3 Discussion

In the full model, β_{11} and β_{12} are highly negatively correlated (≈ -0.9) in their posterior distribution. Our spherical proposal in the *update* step causes slow convergence and does not mix well. The model also tends to get stuck in the full model due to this posterior correlation structure. The acceptance ratio $\alpha_{(d)}$ for the death $type$ transition $M_1 \rightarrow M_2$ is given by,

$$\alpha_{(d)} = \frac{P(\boldsymbol{Y}|\boldsymbol{\beta^{(2)}}, M_2)\pi(\boldsymbol{\beta^{(2)}}|M_2)\pi(M_2)P(M_2 \rightarrow M_1)q_{(M_2 \rightarrow M_1)}(u_{(2)}|\beta_{20}, \beta_{21})}{P(\boldsymbol{Y}|\boldsymbol{\beta^{(1)}}, M_1)\pi(\boldsymbol{\beta^{(1)}}|M_1)\pi(M_1)P(M_1 \rightarrow M_2)} \times \left|\frac{\partial(\beta_{20}, \beta_{21}, u_{(2)})}{\partial(\beta_{10}, \beta_{11}, \beta_{12})}\right|, \tag{4.9}$$

where $u_{(2)} = \frac{1}{2}(\beta_{11} - \beta_{12})$ and the Jacobian of transformation, $\left|\frac{\partial(\beta_{20}, \beta_{21}, u_{(2)})}{\partial(\beta_{10}, \beta_{11}, \beta_{12})}\right|$ is one based on the mapping defined in (4.3). As specified in equation (4.9), $\alpha_{(d)}$ depends on the posterior correlation between β_{11} and β_{12}. A high negative correlation between them results in a large value for $u_{(2)}$. Under a small proposal variance for $q_{(M_2 \rightarrow M_1)}(.)$, this results in extremely low model jump probabilities. It is a pointer to how difficult it is to devise good model jump proposals. For the automatic proposal strategy, the fact that jumps to

same or smaller sized models is deterministic, severely affects the performance of this proposal. The sampler not only gets stuck in M_1, but also does not maintain the ranking in posterior model probabilities. Such a mapping makes sense only if the posterior means μ's, can be roughly be estimated as in our linear regression case. In these scenarios, a small value of R would suffice, but if μ's are unknown (say assumed to be zeros), a large value of R has to be considered to bring in an unknown but larger effect size in the proposed model.

Although the multiset averaged posterior distribution of the model parameters clearly identify the model specific modes, it has much lighter tails than expected – some mass at the tails of the posterior distribution of any β_j is not explored by the sampler under minimal or no tuning. This is an artifact of our prior specification for $\boldsymbol{\beta}$ and the indicator variables. For models in which β_j does not exist, the posterior contribution of β_j will be only from the prior. Hence the tails of the multiset averaged posterior distribution for β_j will be dominated by the prior variance for any β_j. For a very small proposal variance (0.1 in our implementations), mixing will be too slow to capture the tail characteristics. Different proposal variances is a possible alternative, but our goal from the beginning has been a methods that is simple to implement and needs almost zero tuning.

Our primary conclusion from this example is that prior information about model specific parameter values help us devise intelligent proposals. But even in small dimensional problems, mere relationships between model specific parameters are not enough for good mixing in model selection algorithms. Mutliset model selection with its superior performance, is a simpler alternative that allows a more efficient exploration of the model space without getting $stuck$ in local modes and simultaneously allows sampling from varying model specific correlation structures.

Chapter 5

Multiset Model Selection with Binary Responses

We extend MSMS to models with binary responses, specifically probit regression using a multiset on the model indicators in the linear predictor. Section 5.0.1 briefly discusses the model selection problem using the prior framework of Kuo and Mallick [1998] and the probit model formulation due to Albert and Chib [1993], the multiset version of this model selection problem and its extension using the composition sampler due to Holmes and Held [2006] (joint update of regression coefficients and auxiliary variables). We also provide a better understanding of why our algorithm explores the model space more efficiently and show how model specific estimators can be obtained easily from the MCMC samples of our algorithm without any extra effort. We also discuss possible numerical challenges which might arise in some cases. In Section 5.1, using simulated data, we show that MMS helps us evaluate correct posterior model probabilities in a simple case where a ad-hoc proposal gets stuck in a model.

5.0.1 Bayesian Formulation of Probit Model Using Multiset

Multiset Model Selection Framework in Probit Model

Using our previous prior framework, hierarchical probit model is as follows:

$$\begin{aligned} y_i|z_i &= \begin{cases} 1, \text{if } z_i \geq 0 \\ 0, \text{if } z_i < 0 \end{cases} \\ z_i|\boldsymbol{\vartheta} &\sim f(z_i|\boldsymbol{\vartheta}) \equiv N(\boldsymbol{x_i^T}\boldsymbol{\vartheta}, 1), i = 1, \ldots, n \\ \beta_j &\sim \pi(\beta_j) \equiv N({\beta^0}_j, \sigma_j^2), j = 1, \ldots, p \\ \boldsymbol{\gamma} &\propto \pi(\boldsymbol{\gamma}) \equiv \prod_{j=1}^{p} {p_j}^{\gamma_j}(1-p_j)^{(1-\gamma_j)}. \end{aligned} \tag{5.1}$$

Let $\boldsymbol{I(Z,Y)} = \prod_{i=1}^{n}\{y_i I(z_i \geq 0) + (1-y_i)I(z_i < 0))\}$ be the joint indicator function, $\boldsymbol{X}^{(\gamma)}$ and $\boldsymbol{\beta}^{(\gamma)}$ are the subsets of the design matrix and the covariate vector respectively with columns corresponding to only $non-zero$ parameters in model $\boldsymbol{\gamma}$, $\boldsymbol{D}^{(\gamma)}$ is the prior diagonal covariance matrix with size equal to the number of $non-zero$ parameters in model $\boldsymbol{\gamma}$. Under a uniform prior on the model space, the posterior model probability for model $\boldsymbol{\gamma}$ is proportional to the area under the multivariate truncated Gaussian distribution on $\boldsymbol{Z}$, $N(\boldsymbol{Z}|\boldsymbol{0}_n, I_n + \boldsymbol{X}^{(\gamma)}\boldsymbol{D}^{(\gamma)}\boldsymbol{X}^{(\gamma)^T})\boldsymbol{I(Z,Y)}$. We will always assume that the mean for the prior on $\boldsymbol{\beta}$ is zero i.e. $\boldsymbol{\beta}^0 = \boldsymbol{0}_p$. To obtain the model probabilities and the parameter estimates via MCMC we need the full conditional distributions for $\boldsymbol{\beta}$, $\boldsymbol{Z}$ and model indicator $\boldsymbol{\gamma}$. The first two can be easily obtained in closed form using model (5.1). From the hierarchy it is evident that the full conditional distribution for $\boldsymbol{\beta}^{(\gamma)}$, given model $\boldsymbol{\gamma}$, is the posterior distribution for $\boldsymbol{\beta}^{(\gamma)}$ in the linear regression of $\boldsymbol{Z}$ versus $\boldsymbol{X}^{(\gamma)}$ with $N(\boldsymbol{0}, \boldsymbol{I}_n)$ distributed error,

$$f(\boldsymbol{\beta}^{(\gamma)}|\boldsymbol{Z}, \boldsymbol{\gamma}) = N((\boldsymbol{D}^{(\gamma)^{-1}} + \boldsymbol{X}^{(\gamma)^T}\boldsymbol{X}^{(\gamma)})^{-1}\boldsymbol{X}^{(\gamma)^T}\boldsymbol{Z}, (\boldsymbol{D}^{(\gamma)^{-1}} + \boldsymbol{X}^{(\gamma)^T}\boldsymbol{X}^{(\gamma)})^{-1}),$$

and $\boldsymbol{\beta}^{(-\gamma)}$ will be sampled from its Gaussian prior. The full conditional distribution for each z_i is a truncated normal distribution with mean $\boldsymbol{x_i^T}\boldsymbol{\vartheta}$ and variance 1, where the left or right truncation at zero is conditional on whether the data y_i is one or zero respectively. To allow better mixing in the presence of strong posterior correlation between $\boldsymbol{\beta}$ and $\boldsymbol{Z}$, Holmes and Held [2006] suggested a composite sampler to replace the separate sampling steps for $\boldsymbol{\beta}$ and $\boldsymbol{Z}$ with a joint sampling step as $f(\boldsymbol{\beta}, \boldsymbol{Z}|\boldsymbol{Y}) = f(\boldsymbol{\beta}|\boldsymbol{Z})f(\boldsymbol{Z}|\boldsymbol{Y})$.

The multiset version of the probit regression model using auxiliary variable approach involves specifying a multiset on the model indicator space $\boldsymbol{\gamma}$. For a multiset of size two, a sample from the multiset space is denoted by $\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}$. As before, the space of the parameters $\boldsymbol{\beta}$ in the linear predictor does not have a multiset on it. Hence the modified

target distribution is given by,

$$\begin{aligned} f^*(\boldsymbol{\beta}, \{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Z}|\boldsymbol{Y}) &\propto f(\boldsymbol{\beta}, \boldsymbol{\gamma^{(1)}}, \boldsymbol{Z}|\boldsymbol{Y}) + f(\boldsymbol{\beta}, \boldsymbol{\gamma^{(2)}}, \boldsymbol{Z}|\boldsymbol{Y}) \\ &\propto f(\boldsymbol{\beta^{(1)}}, \boldsymbol{\beta^{(-1)}}, \boldsymbol{\gamma^{(1)}}, \boldsymbol{Z}|\boldsymbol{Y}) + f(\boldsymbol{\beta^{(2)}}, \boldsymbol{\beta^{(-2)}}, \boldsymbol{\gamma^{(2)}}, \boldsymbol{Z}|\boldsymbol{Y}), \end{aligned} \tag{5.2}$$

where $\boldsymbol{\beta^{(i)}}$ and and $\boldsymbol{\beta^{(-i)}}$ correspond to the sub-vectors of $\boldsymbol{\beta}$ that is *in* model $\boldsymbol{\gamma^{(i)}}$ and *out* of model $\boldsymbol{\gamma^{(i)}}$ respectively.

Sampling Strategy

Although Metropolized steps can be used for each of the Gibbs sampling steps, we propose a method inspired by the joint updating of $\boldsymbol{\beta}$ and $\boldsymbol{Z}$ in Holmes and Held [2006]. The joint sampling from the full conditional for $\boldsymbol{\beta}$ and $\boldsymbol{Z}$ uses the following factorization,

$$\begin{aligned} & f^*(\boldsymbol{\beta}, \boldsymbol{Z}|\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y}) \\ = \; & f^*(\boldsymbol{\beta}|\boldsymbol{Z}, \{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y}) f^*(\boldsymbol{Z}|\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y}) \\ = \; & f^*(\boldsymbol{\beta}|\boldsymbol{Z}, \{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y}) \int f^*(\boldsymbol{\beta}, \boldsymbol{Z}|\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y}) d\boldsymbol{\beta} \\ \propto \; & f^*(\boldsymbol{\beta}|\boldsymbol{Z}, \{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y}) \int \left(f(\boldsymbol{\beta}, \boldsymbol{\gamma^{(1)}}, \boldsymbol{Z}|\boldsymbol{Y}) + f(\boldsymbol{\beta}, \boldsymbol{\gamma^{(2)}}, \boldsymbol{Z}|\boldsymbol{Y}) \right) d\boldsymbol{\beta} \\ \propto \; & f^*(\boldsymbol{\beta}|\boldsymbol{Z}, \{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y}) \left(\int f(\boldsymbol{\beta^{(1)}}, \boldsymbol{Z}|\boldsymbol{Y}) d\boldsymbol{\beta^{(1)}} + \int f(\boldsymbol{\beta^{(2)}}, \boldsymbol{Z}|\boldsymbol{Y}) d\boldsymbol{\beta^{(2)}} \right) \\ \propto \; & f^*(\boldsymbol{\beta}|\boldsymbol{Z}, \{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y}) \left(\pi(\boldsymbol{\gamma^{(1)}}) f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)}}, \boldsymbol{Y}) + \pi(\boldsymbol{\gamma^{(2)}}) f(\boldsymbol{Z}|\boldsymbol{\gamma^{(2)}}, \boldsymbol{Y}) \right), \end{aligned} \tag{5.3}$$

where the last but one step follows from the fact that the proper priors on the parameters that are not in $\boldsymbol{\gamma^{(i)}}$ integrate to one and $f(\boldsymbol{Z}|\boldsymbol{\gamma^{(i)}}, \boldsymbol{Y})$ is the marginal distribution of $\boldsymbol{Z}$ with $\boldsymbol{\gamma^{(i)}}$ as the model. Hence, under a uniform prior on the model space, sampling from $f^*(\boldsymbol{Z}|\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y})$ is equivalent to sampling from a two component mixture of truncated normal distributions in $\boldsymbol{Z}$ with equal weights,

$$\begin{aligned} f^*(\boldsymbol{Z}|\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y}) &\sim N(\boldsymbol{Z}|\boldsymbol{0}_n, I_n + \boldsymbol{X^{(1)}} \boldsymbol{D}^{(1)} \boldsymbol{X^{(1)}}^T) \boldsymbol{I(Z, Y)} \\ &+ N(\boldsymbol{Z}|\boldsymbol{0}_n, I_n + \boldsymbol{X^{(2)}} \boldsymbol{D}^{(2)} \boldsymbol{X^{(2)}}^T) \boldsymbol{I(Z, Y)} \end{aligned} \tag{5.4}$$

The joint distribution of $f^*(\boldsymbol{\beta}, (\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}), \boldsymbol{Z}|\boldsymbol{Y})$ in Equation (5.2) also indicates that $f^*(\boldsymbol{\beta}|\boldsymbol{Z}, \{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y})$ will be a two component mixture of normal densities with weights w_3 and

w_4 given by,

$$\begin{aligned} f^*(\boldsymbol{\beta}|\boldsymbol{Z},\{\boldsymbol{\gamma^{(1)}},\boldsymbol{\gamma^{(2)}}\},\boldsymbol{Y}) &\sim w_3 N(\boldsymbol{\beta}^{(1)}|\hat{\boldsymbol{\beta}}^{(\mathbf{1})},\hat{\boldsymbol{\Sigma}}^{(\mathbf{1})}) \times \pi(\boldsymbol{\beta}^{(-1)}) + w_4 N(\boldsymbol{\beta}^{(2)}|\hat{\boldsymbol{\beta}}^{(\mathbf{2})},\hat{\boldsymbol{\Sigma}}^{(\mathbf{2})}) \times \pi(\boldsymbol{\beta}^{(-2)}) \\ w_3 &\propto \pi(\boldsymbol{\gamma^{(1)}}) \frac{|\boldsymbol{S^{(1)}}|^{\frac{1}{2}}}{|\boldsymbol{D}^{(1)}|^{\frac{1}{2}}} e^{-\frac{1}{2}\widehat{\boldsymbol{r^{(1)}}}^{\boldsymbol{T}}\widehat{\boldsymbol{r^{(1)}}}} \\ w_4 &\propto \pi(\boldsymbol{\gamma^{(2)}}) \frac{|\boldsymbol{S^{(2)}}|^{\frac{1}{2}}}{|\boldsymbol{D}^{(2)}|^{\frac{1}{2}}} e^{-\frac{1}{2}\widehat{\boldsymbol{r^{(2)}}}^{\boldsymbol{T}}\widehat{\boldsymbol{r^{(2)}}}} \end{aligned} \tag{5.5}$$

where $\boldsymbol{S^{(i)}} = (\boldsymbol{X^{(i)^T}}\boldsymbol{X^{(i)}} + \boldsymbol{D}^{(i)^{-1}})^{-1}$, $\pi(\boldsymbol{\beta^{(-i)}})$ is the prior over the parameters not in model $\boldsymbol{\gamma^{(i)}}$, $\widehat{\boldsymbol{\beta^{(i)}}}$ is the MAP estimator for $\boldsymbol{\beta^{(i)}}$ with posterior covariance $\widehat{\boldsymbol{\Sigma^{(i)}}}$ and $\widehat{r^{(i)}} = \boldsymbol{Z} - \boldsymbol{X^{(i)}}\widehat{\boldsymbol{\beta^{(i)}}}$. To sample from the truncated normal distribution $N(\boldsymbol{Z}|\mathbf{0}_n, I_n + \boldsymbol{X^{(1)}}\boldsymbol{D}^{(1)}\boldsymbol{X^{(1)^T}})\boldsymbol{I(Z,Y)}$, we maintain the iterative Gibb's sampling framework based on a faster algorithm proposed by Holmes and Held [2006] (results based on Henderson and Searle [1981]).

The Metropolis-Hastings ratio for a proposed move from $\{\boldsymbol{\gamma^{(1)}},\boldsymbol{\gamma^{(2)}}\}$ to $\{\boldsymbol{\gamma^{(1)*}},\boldsymbol{\gamma^{(2)}}\}$ is obtained by marginalizing the target density over $\boldsymbol{\beta}$ (see Lee and et al. [2003]). For a proposal distribution given by $q(\{\boldsymbol{\gamma^{(1)}},\boldsymbol{\gamma^{(2)}}\} \rightarrow \{\boldsymbol{\gamma^{(1)*}},\boldsymbol{\gamma^{(2)}}\})$, the MH ratio is given by,

$$\alpha_{\boldsymbol{\gamma}} = min\left(1, \frac{\pi(\boldsymbol{\gamma^{(1)*}})f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)*}},\boldsymbol{Y}) + \pi(\boldsymbol{\gamma^{(2)}})f(\boldsymbol{Z}|\boldsymbol{\gamma^{(2)}},\boldsymbol{Y})}{\pi(\boldsymbol{\gamma^{(1)}})f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)}},\boldsymbol{Y}) + \pi(\boldsymbol{\gamma^{(2)}})f(\boldsymbol{Z}|\boldsymbol{\gamma^{(2)}},\boldsymbol{Y})} \times \frac{q(\{\boldsymbol{\gamma^{(1)*}},\boldsymbol{\gamma^{(2)}}\} \rightarrow \{\boldsymbol{\gamma^{(1)}},\boldsymbol{\gamma^{(2)}}\})}{q(\{\boldsymbol{\gamma^{(1)}},\boldsymbol{\gamma^{(2)}}\} \rightarrow \{\boldsymbol{\gamma^{(1)*}},\boldsymbol{\gamma^{(2)}}\})}\right) \tag{5.6}$$

Equation 5.4 provides us insight in to the sampling step for $f^*(\boldsymbol{Z}|\{\boldsymbol{\gamma^{(1)}},\boldsymbol{\gamma^{(2)}}\},\boldsymbol{Y})$. Recall that the posterior model probability is proportional to the area of the multivariate Gaussian distribution of $\boldsymbol{Z}$ in the region defined by the truncation $\boldsymbol{I(Z,Y)}$. Hence, a posteriori the domain of $\boldsymbol{Z}$ is an $n - dimensional$ Euclidean hyper-quadrant satisfying $\boldsymbol{I(Z,Y)}$. Since we have modified our target distribution to Equation 5.2, the posterior probability of a multiset is proportional to the sum of the areas of the truncated multivariate Gaussians corresponding to the two models in the multiset. The marginal likelihood of $\boldsymbol{Y}$ due to a specific probit model is difficult to calculate. The hierarchical model using auxiliary variable enables calculation of the integral in a relatively straight forward way using a Gibbs sampler. The Gibbs steps in our multiset approach provides us a tool for integrating $\boldsymbol{Z}$ over mixtures (corresponding to models in the multiset) of truncated Gaussian distributions. Sampling for $f^*(\boldsymbol{\beta}|\boldsymbol{Z},\{\boldsymbol{\gamma^{(1)}},\boldsymbol{\gamma^{(2)}}\},\boldsymbol{Y})$ using Equation 5.5 can be thought of sampling from the posterior distribution of $\boldsymbol{\beta}$ weighted by the posterior model probabilities of $\boldsymbol{\gamma^{(1)}}$ and $\boldsymbol{\gamma^{(2)}}$ in a linear regression of $\boldsymbol{Z}$ versus $\boldsymbol{X}$ with independent $N(0,1)$ errors. This is obvious when the weights w_3 and w_4 are compared with the linear regression model probabilities.

Benefits of the Sampling Strategy

If the mixture components in the sampling steps for $f^*(\boldsymbol{Z}|\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y})$ and $f^*(\boldsymbol{\beta}|\boldsymbol{Z}, \{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y})$ are from the same model, say $\boldsymbol{\gamma^{(1)}}$, then the joint sample is from the posterior distribution given model $\boldsymbol{\gamma^{(1)}}$ i.e. $f(\boldsymbol{\beta}, \boldsymbol{Z}|\boldsymbol{\gamma^{(1)}}, \boldsymbol{Y})$. Such samples can be stored to obtain model specific estimates. Using the *basic marginal likelihood identity*, we calculate the marginal likelihood of $\boldsymbol{Y}$ given $\boldsymbol{\gamma^{(1)}}$ using the approach due to Chib [1995],

$$p(\boldsymbol{Y}|\boldsymbol{\gamma^{(1)}}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\beta^*}, \boldsymbol{\gamma^{(1)}})\pi(\boldsymbol{\beta^*}|\boldsymbol{\gamma^{(1)}})}{p(\boldsymbol{\beta^*}|\boldsymbol{Y}, \boldsymbol{\gamma^{(1)}})} = \frac{p(\boldsymbol{Y}|\boldsymbol{\beta^*}, \boldsymbol{\gamma^{(1)}})\pi(\boldsymbol{\beta^*}|\boldsymbol{\gamma^{(1)}})}{\int p(\boldsymbol{\beta^*}|\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{\gamma^{(1)}})p(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\gamma^{(1)}})d\boldsymbol{Z}} \tag{5.7}$$

where $\boldsymbol{\beta^*}$ is in a region of high posterior density for $\boldsymbol{\beta}$ corresponding to model $\boldsymbol{\gamma^{(1)}}$. The integral in Equation 5.7 can easily be evaluated using the Monte Carlo approximation,

$$f(\boldsymbol{\beta^*}|\boldsymbol{Y}, \boldsymbol{\gamma^{(1)}}) = \frac{1}{G} \sum_{\substack{g=1 \\ \boldsymbol{Z^{(g)}} \sim f(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\gamma^{(1)}})}}^{G} f(\boldsymbol{\beta^*}|\boldsymbol{Z^{(g)}}, \boldsymbol{\gamma^{(1)}}, \boldsymbol{Y}) \tag{5.8}$$

5.1 An Example

5.1.1 Data

Table 5.1: Model specific maximum likelihood estimates and model probabilities for simulated data. Models $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_3$ are the two top models, but a move from the low probability model $\boldsymbol{\gamma}_2$, to $\boldsymbol{\gamma}_3$ is almost impossbile.

Model	$\hat{\boldsymbol{\beta}}_{\mathbf{0}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{1}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{2}}$	$P(\boldsymbol{Y}\|\gamma)$, $(\sigma^2 = 100)$	$P(\gamma\|\boldsymbol{Y})$
γ_0	-0.0515	0	0	1.211×10^{-32}	≈ 0
γ_1	-0.0143	7.8371	0	9.6824×10^{-6}	0.4807
γ_2	-0.1571	0	34.6711	1.7062×10^{-13}	≈ 0
γ_3	0.1616	11.9172	1.7929	10.459×10^{-6}	0.5193

Consider $n = 100$ data points simulated using two predictors $\mathbf{x}_1$ and $\mathbf{x}_2$ and an intercept ($p = 3$). The model specific parameter estimates and marginal likelihood of the data are given below (data available on request). As suggested by the marginal likelihood of the data, models $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_3$ have almost equal posterior model probabilities under a uniform prior on the model space. The posterior probability of model $\boldsymbol{\gamma}_2$ is almost zero. In spite of such well defined model probabilities in a small dimensional

problem, it is extremely difficult to explore the model space because of the model specific parameter estimates. A move from model γ_2 to γ_3 has to ensure that when the new parameter β_1 comes in to the model, the value of β_2 has to be simultaneously proposed around a small value of 1.7929. A dependent proposal centered at the last value for β_2 will fail to be accepted as it will almost always propose a value around 34.67. A comparison of the log-likelihoods, $\mathcal{L}(\boldsymbol{\beta} = (-0.1571, 0, 34.6711)|\boldsymbol{Y}) = -16.39$ versus $\mathcal{L}(\boldsymbol{\beta} = (0.1616, 11.9172, 34.6711)|\boldsymbol{Y}) = -6.70$, suggests that even if β_0 and β_1 are proposed (perhaps using independent proposals) around acceptable values for model γ_3, the difference in the log-likelihoods is such that such a move will be accepted once in about $16,000$ tries. If model specific modes for the parameters are known a priori then proposals that make use of this information can be designed. Unfortunately, for model selection with a large number of covariates, this becomes a problem since it would demand knowing model specific parameter estimates for a large number of models. In such cases, rather than designing complicated proposal strategies, we focus on changing the target distribution using a multiset that surprisingly provides a quick and easy method of exploring the model space efficiently.

5.1.2 Results

In our example, a simple model selection algorithm implemented using the point mass at zero prior due to Kuo and Mallick [1998] with prior covariance $D = 100\boldsymbol{I}_p$ and the joint updating scheme of $\boldsymbol{\beta}$ and $\boldsymbol{Z}$ gets 'stuck' in γ_2, either when the sampler is started from γ_2 or when it reaches γ_2 in course of its exploration of the model space. Multiset model selection for a probit model solves this algorithmic issue in two ways; firstly, by marginalizing over the model specific parameters $\boldsymbol{\beta}$, we sample iteratively in our Gibbs sampler from $f^*(\boldsymbol{Z}|\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y})$ and $f^*(\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}|\boldsymbol{Z}, \boldsymbol{Y})$ and secondly, if model specific estimates are *required* by the researcher, then a sampling step for $f^*(\boldsymbol{\beta}|\boldsymbol{Z}, \{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\}, \boldsymbol{Y})$ is accommodated. Averaged over ten independent runs of the multiset model selection algorithm, multiset based posterior model probabilities are given in table (5.2). The prior on the model space is uniform i.e. $p = 0.5$ in Equation 5.1 and the prior covariance matrix over $\boldsymbol{\beta}$ is $D = 100\boldsymbol{I}_p$. Our proposal $q(\{\boldsymbol{\gamma^{(1)}}, \boldsymbol{\gamma^{(2)}}\} \rightarrow \{\boldsymbol{\gamma^{(1)*}}, \boldsymbol{\gamma^{(2)}}\})$, proposes *no* change in the multiset with probability 0.999 and with probability 0.001 proposes to change any one of the models $\boldsymbol{\gamma^{(1)}}$ or $\boldsymbol{\gamma^{(2)}}$ randomly to one of the remaining models. Estimates from each run is based on $200,000$ MCMC iterations. Designing a *sticky* proposal as above allows the parameters to converge to high likelihood regions for the specific models in the multiset. As expected, the model probabilities from multiset are a *flatter* version of the true model probabilities (refer to section (3.6) for details), but it preserves their ordering. In Figure 5.1, we plot the boxplots for the logarithm of inverse Bayes Factors for all possible model pairs, from the 10 independent runs of the Multiset Model Selection algorithm and using Equations 5.7 and 5.8. Models γ_1 and γ_3 have the same posterior model probabilties based on the graphic. The comparative *weights of evidence* of any other model pair also

suggest that models γ_1 and γ_3 are the only high probability models in our model space. It is important to note that this conclusion (and calculation of the logarithm of Bayes Factors) was reached without resorting to the previous analytical framework for extracting true model probabilities from multiset model probabilities (section 3.6).

Table 5.2: Model specific estimates from Multiset Model Selection for Probit Model

Model	True Model Probabilities	*Estimated* Model Probabilities (Kuo & Mallick's Prior)	*Estimated* Model Probabilities (MSMS)	*Estimated* Multiset Model Probabilities	True Multiset Model Probabilities
γ_0	≈ 0	≈ 0	0.0550	0.1330	0.1000
γ_1	0.4808	0.4278	0.4095	0.3457	0.3885
γ_2	≈ 0	0.4998	0.1122	0.1673	0.1000
γ_3	0.5191	0.0722	0.4235	0.3541	0.4115

5.2 Discussion

We now focus our attention on how the multiset approach helps an efficient exploration of the model space. As described in section (5.0.1), multiset model selection integrates a mixture of truncated Gaussian distributions over $\boldsymbol{Z}$. Each model specific truncated Gaussian distribution, $f(\boldsymbol{Z}|\boldsymbol{\gamma}, \boldsymbol{Y})$, is centered at zero and share the same domain but characterized by a covariance structure $I_n + \boldsymbol{X}^{(\gamma)}\boldsymbol{D}^{(\gamma)}\boldsymbol{X}^{(\gamma)^T}$. The covariance structure determines how *diffuse* the distribution is as well as the direction of *maximal variation* or *ridge*. Marginalized over the multiset space, $\boldsymbol{Z}$ has multiple ridges emanating from a single mode at zero.

Consider contour plots based on $f(\boldsymbol{Z}|\boldsymbol{\gamma}, \boldsymbol{Y})$, of z_{52} versus z_{58}, marginalized over the remaining z for models $\boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$ in Figure 5.2. Let the '$\times$' denote the location of $\boldsymbol{Z}$ at a step of the Gibbs sampler with $\boldsymbol{\gamma}_2$ as the current state of the model i.e. $\boldsymbol{\gamma^{(1)}} = \boldsymbol{\gamma}_2$. If $\boldsymbol{\gamma}_3$ is the proposed model, i.e. $\boldsymbol{\gamma^{(1)*}} = \boldsymbol{\gamma}_3$, the ratio of $f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)*}}, \boldsymbol{Y})$ to $f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)}}, \boldsymbol{Y})$ is so small that the probability of accepting such a model update move is almost nil. However for a multiset with $\boldsymbol{\gamma^{(2)}} = \boldsymbol{\gamma}_2$, the ratio of $f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)*}}, \boldsymbol{Y}) + f(\boldsymbol{Z}|\boldsymbol{\gamma^{(2)}}, \boldsymbol{Y})$ to $f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)}}, \boldsymbol{Y}) + f(\boldsymbol{Z}|\boldsymbol{\gamma^{(2)}}, \boldsymbol{Y})$ is close to 0.5. This is exactly the MH ratio in Equation 5.6 under a flat prior on the model space and a symmetric proposal for multiset. Once $\boldsymbol{\gamma}_3$ is brought in as an element of the multiset, Equation 5.4 samples with equal weights from $f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)*}}, \boldsymbol{Y})$ and $f(\boldsymbol{Z}|\boldsymbol{\gamma^{(1)}}, \boldsymbol{Y})$. Hence, the multiset approach does *not* explicitly facilitate moves from one mode to the other in the parameter space. Rather, it enables moves between two *ridges* in the mixture components of the posterior distribution of $\boldsymbol{Z}$.

Figure 5.1: Boxplots of $log(BF(\boldsymbol{\gamma}_i|\boldsymbol{\gamma}_j)^{-1})$ for all possible model combinations, from 10 independent runs of the Multiset Model Selection algorithm. As expected, models $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_3$ have comparable *weights of evidence*.

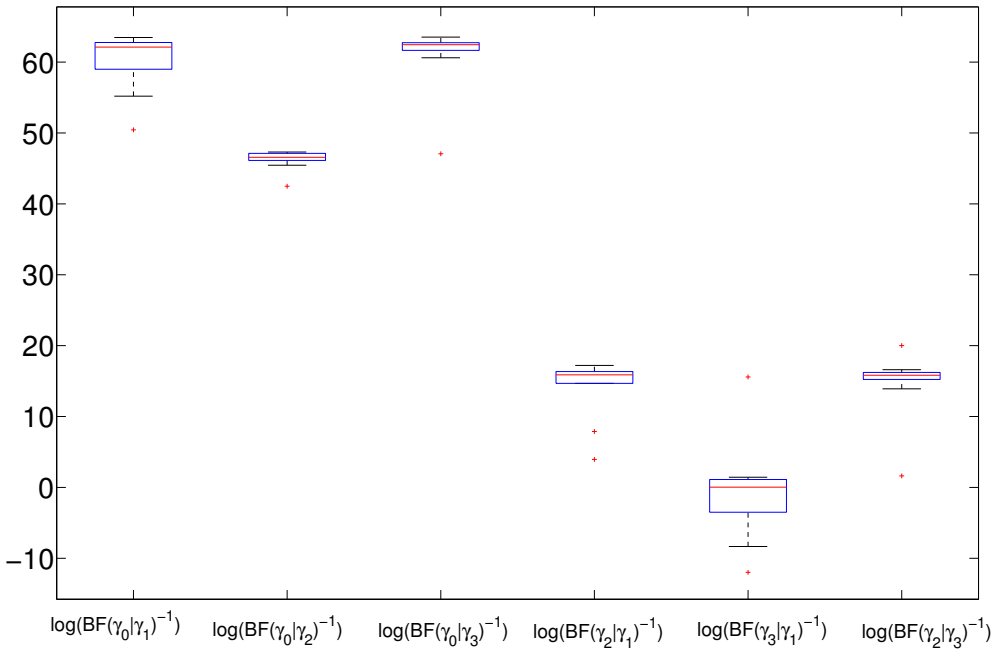
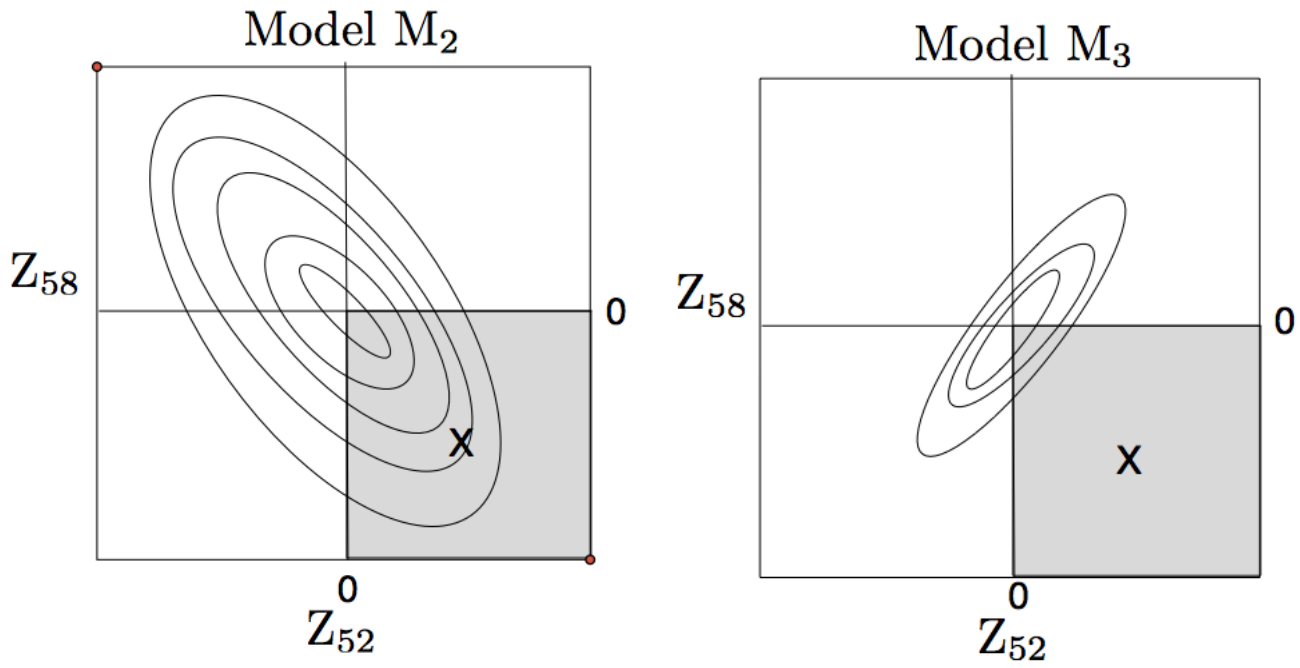


Figure 5.2: Contour plot of z_{52} versus z_{58} marginalized over the remaining z for models M_2 and M_3. Shaded portion shows the quadrant of the Z space defined by $\boldsymbol{I}(\boldsymbol{X}, \boldsymbol{Y})$. $y_{52} = 1$, $y_{58} = 0$.



Chapter 6

Conclusion

We have proposed the multiset model selection algorithm (MSMS) by extending the multiset sampler to model selection. We define a multiset on the space of models but not on the space of the model parameters. While automatic proposal strategies exist for efficient exploration of the model space, we have shown that they might still require some tuning. Although adaptive methods perform better, their general formulation and implementation is tricky and complicated. Our formulation of MSMS as a general model selection algorithm, is extremely easy to implement and explores the model space more efficiently than some of the automatic proposal strategies without any tuning. In linear regression, MSMS explores the model space by facilitating across-model jumps where the shared parameter(s) between two models might have extremely different values. In model selection using probit regression, it allows across-model jumps where model specific covariance structure of the auxiliary variable might be very different. The multiset averaged model probabilities are ordered exactly as the Bayesian posterior model probabilities and the the latter can be extracted from the former by a simple inverse calculation. The posterior distribution of the model parameters averaged over the space of multisets is the same as the Bayesian model averaged parameter distribution.

Part II

INTERACTIVE STORYTELLING - Bringing User Interaction to Path Discovery in Document Networks

Chapter 7

Introduction

The *Storytelling* algorithm [Kumar et al., 2006, Hossain et al., 2010] constructs a path or a sequence of documents, henceforth called a *story* in a document network with an appropriate distance function defined on it. The *story* is created by starting with a given document and then discovering and connecting seemingly unrelated documents along the way before finally ending in another specified document. Throughout this process documents are brought into the path based on the criteria of generating the shortest path between the starting and ending document in the document network. The original *Storytelling* algorithm was proposed as a methodology to navigate through a graph of redescriptors defined over a set of objects and a collection of subsets over these objects. The similarity between two redescriptions was defined as the Jaccard distance between the two sets defined by the redescriptions. An $A^*Search$ algorithm was used for the path search.

The *Storytelling algorithm* has been applied to intelligence data [Hossain et al., 2011] to aid the user in *foraging* information from a corpus containing news clips or text snippets. In this case, *Storytelling* helps us understand how one document is related to a non-adjacent document via intermediate inter-document relationships. It does so by connecting two documents if they have overlapping term-vectors, and creating a sequence of documents along the shortest path, from a pre-specified starting document to an ending document . An edge in the graph naturally signifies an inter-document relationship via the overlapping terms in the document-pair that define the edge. From a term-vector representation of document space, two documents are *unrelated* if the inter-document distance is greater than a certain threshold or due to a total absence of any common terms between them. The resulting *story* is supposed to span over key characters, places and events that point to one or more plausible scenarios. In intelligence analysis these *stories* refer to viable terrorist threats.

In our algorithm, a vector of normalized topic weights generated by (supervised) Latent Dirichlet Allocation (LDA) for every document induces a distance between every

document-pair. Documents which share similar topic distributions, are closer while ones that are very different in their topic distributions are farther apart. This distance information, coupled with the edge information, is used to generate a network – documents as nodes, with an edge between two nodes as defined above. Our focus is on the user's guidance when the progression of the initial *story* is not based on his expected hypotheses. The guidance from the user is in terms of favoring a certain path over other alternate paths. Such guidance from the user will henceforth be termed as *feedback*. In such scenarios, the user uses his subject matter expertise and semantic associations to accept, modify or reject the hypotheses contained in the initial *story*.

The original algorithm [Hossain et al., 2010] had parameters at the user's disposal to control the adjacency matrix of the documents via constraints on clique size and the maximum nearest neighbor edge length - visually this translates to impacting the length of the story for the user. Clique size guides the path through regions of desired levels of connectivity and the thresholding distance controls the radius of the local neighborhood of a document in the graph. We propose to incorporate user feedback injected strictly from the cognitive domain; a domain in which the user is comfortable with interacting with the documents. The feedback will be only via a sequence of documents that the user expects to be in the *story*, and completely shields the user from interacting at the parameter level of the model. While the idea of interactive visual analytics is not new, the *Storytelling* algorithm and our approach of bringing in interaction is unique. Hence, in the following section, we briefly show an example of our INTERACTIVE STORYTELLING algorithm applied to an intelligence dataset, following which we provide a detailed outline of our article.

7.1 An Example of INTERACTIVE STORYTELLING

For our example, we use the Atlantic Storm dataset [Huges, 2005, Hossain et al., 2011] developed at the Joint Military Intelligence College. The intelligence analyst dealing with the problem will henceforth be referred to as the user. As in a typical *Storytelling* paradigm, an user identifies two documents that are unrelated in as much as their general theme is concerned but which are perhaps connected as a coherent *story* via a sequence of intermediate documents. The first document that has caught the user's attention reads as follows:

> Document (CIA06) – (a) Saeed Hasham [aliases: Hamid Qatada, Yasir Salman]. Member of central staff, Al Qaeda; headed Farooq training camp in Khost area; name appears on many planning documents for insurgent activities in 2002 at Jorm, Taloqan, and the Salang tunnel; received special notice from Osama bin Laden on 22 September 2002 for his planning of successful insurgent actions. (b) Khalid Taha [no recorded aliases], an

> Egyptian who headed the Spin Shaga camp in Pakhtia province for Kashmiri militants; specializes in chemical weapons. (c) Fahd al Badawi [no recorded aliases]; a Saudi with a medical degree from University of Cairo, Egypt having a specialty in immunology; on medical staff at Farooq training camp in 2001; awarded for successfully treating Osama bin Laden for a bacterial infection in July, 2000.

and the second document mentions a veiled threat about a shipment of antibiotics:

> Document (NSA16) – Intercepted phone call on 28 September, 2004 from Havana, Cuba to Santo Domingo, Dominican Republic. The Havana call originated at 45 Desague St. in Havana listed in the name Jose Escalante; the destination was at 65 Ave. San Martin in Santo Domingo, listed in the name Carlos Morales. The caller says [in Spanish]: "Hello Carlos, I'm calling just to tell you that Arze will have the medical supplies to you by November 21st. You know who to give them to?" The reply is: " Yes, some guy named Sufaat. He also gets the antibiotics that I just heard about from the Arab. They will also be here by November 21st. Where is all this stuff going?" The caller says : "I don't know anything about any antibiotics. I guess we are not supposed to know, but my guess is that they will go to our friends in Columbia."

The *Storytelling* algorithm under the absence of any interaction provides the user with the following *story*, which connects documents $CIA06$ and $NSA16$. This *story* is the shortest path between these two documents in the network.

> Document (CIA06) – (a) Saeed Hasham [aliases: Hamid Qatada, Yasir Salman]. Member of central staff, Al Qaeda; headed Farooq training camp in Khost area; name appears on many planning documents for insurgent activities in 2002 at Jorm, Taloqan, and the Salang tunnel; received special notice from Osama bin Laden on 22 September 2002 for his planning of successful insurgent actions. (b) Khalid Taha [no recorded aliases], an Egyptian who headed the Spin Shaga camp in Pakhtia province for Kashmiri militants; specializes in chemical weapons. (c) Fahd al Badawi [no recorded aliases]; a Saudi with a medical degree from University of Cairo, Egypt having a specialty in immunology; on medical staff at Farooq training camp in 2001; awarded for successfully treating Osama bin Laden for a bacterial infection in July, 2000.

Document (CIA20) – A paid source in Chitral said he was treated on about 2 October, 2003 for a broken arm by a Dr. Badawi in a refugee camp outside Chitral. Asked how he knew it was a Dr. Badawi, the source replied that another refugee had told him who this doctor was.

Document (NSA09) – Intercepted phone call from Casablanca, Morocco to Santo Domingo, Dominican Republic on 23 October, 2003. The call originated from a phone listed at the Holland Orange Shipping Lines in Casablanca; the recipient of the call was at a residence at 65 Avenue San Martin in Santo Domingo. During the conversation the caller states [in Arabic]: "You will receive a shipment of antibiotics at a time we will announce to you. At about the same time you will receive a shipment of medical supplies. We will let you know where these items will be going."

Document (NSA16) – Intercepted phone call on 28 September, 2004 from Havana, Cuba to Santo Domingo, Dominican Republic. The Havana call originated at 45 Desague St. in Havana listed in the name Jose Escalante; the destination was at 65 Ave. San Martin in Santo Domingo, listed in the name Carlos Morales. The caller says [in Spanish]: "Hello Carlos, I'm calling just to tell you that Arze will have the medical supplies to you by November 21st. You know who to give them to?" The reply is: "Yes, some guy named Sufaat. He also gets the antibiotics that I just heard about from the Arab. They will also be here by November 21st. Where is all this stuff going?" The caller says: "I don't know anything about any antibiotics. I guess we are not supposed to know, but my guess is that they will go to our friends in Columbia."

In terms of bag of words, the story progression is given in Table 7.1. The terms common to two consecutive documents will be known as *transitive terms*; transitive terms are responsible for connections between two documents.

In our example, the user believes that the threat is related to ***Al Qaeda*** activities in the ***Farooq training camp***. To that effect, he identifies $CIA08$ to be an important document in the threat. This document might have information about other connections via persons, dates, organizations, and places, to a credible threat. He also believes that document $NSA09$ ties up *appropriately* with the ending document. Hence the user specifies the sequence of documents, $CIA08$ and $NSA09$, that *should* be in the *story*. Our final goal is to provide a new *story* that takes into account this feedback and represents the documents

Table 7.1: Bag of words for documents and transitive terms for connections between documents, in initial story ($CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA16$) from the Storytelling algorithm.

Document	Bag of words	Transitive terms
CIA06	action,alias,badawi,bin,degre,document,egypt,fahd, farooq,hasham,head,insurg,khost,laden,milit,osama,pakhtia, plan,provinc,salman,special,staff,success,treat,yasir	
↓		badawi, treat
CIA20	arm,badawi,doctor,octob,outsid,paid,refuge,told,treat	
↓		octob
NSA09	avenu,convers,list,martin,morocco,octob,orang,san,shipment	
↓		list,martin,san
NSA16	arz,call,destin,don,friendget,guess,heard,hello,just list,martin,moral,san,stuff,sufaat,suppos	

in the light of newly discovered underlying topics. The process of identifying these new set of topics induces a different proximity structure between the documents, which in turn provides a story which is hopefully consistent with the user's view of how the threat might have progressed. After incorporating the feedback in the algorithm, the new story is as follows:

Document (CIA06) – (a) Saeed Hasham [aliases: Hamid Qatada, Yasir Salman]. Member of central staff, Al Qaeda; headed Farooq training camp in Khost area; name appears on many planning documents for insurgent activities in 2002 at Jorm, Taloqan, and the Salang tunnel; received special notice from Osama bin Laden on 22 September 2002 for his planning of successful insurgent actions. (b) Khalid Taha [no recorded aliases], an Egyptian who headed the Spin Shaga camp in Pakhtia province for Kashmiri militants; specializes in chemical weapons. (c) Fahd al Badawi [no recorded aliases]; a Saudi with a medical degree from University of Cairo, Egypt having a specialty in immunology; on medical staff at Farooq training camp in 2001; awarded for successfully treating Osama bin Laden for a bacterial infection in July, 2000.

Document (CIA08) – Mohamed al Omari is a Saudi who was apprehended by Pakistani police in Peshawar on 12 January 2003. Omari was badly wounded at the time, saying that he was shot while attempting to desert from an Al Qaeda insurgent group that had taken refuge near Parachinar on the Afghanistan –Pakistan border. Omari was placed in a hospital and, after

his recovery, he was encouraged to talk about his Al Qaeda activities. Among the items of information Omari revealed was that he had been at the Farooq training facility in 2001 before the American presence in Afghanistan. He was shown photos of various persons and asked if he could identify any of them. One photo, that had been taken in Peshawar on 2 January, 2003, showed a group of four men. Omari said he knew three of them since they had been with him at the Farooq training camp. One man he identified as Dr. Badawi, the second he identified as Mamdouh al Hazmi, and the third he identified only as Hasham; Omari could not remember his first name. Omari said that all three were members of Al Qaeda.

Document (DIA01) – Interrogation of Abdul Ahmed Nasser, a Saudi born October 1967 in Abba. Captured 3 February, 04 at Baglan. Entered fight against Russians in Chechnya in 1997; went to Afghanistan 1999; provided chemical weapons training to Taliban/Al Qaeda members. Very talkative and says he has former students now in the USA who will shortly have lots of cocktails for "American alcoholic bastards."

Document (NSA09) – Intercepted phone call from Casablanca, Morocco to Santo Domingo, Dominican Republic on 23 October, 2003. The call originated from a phone listed at the Holland Orange Shipping Lines in Casablanca; the recipient of the call was at a residence at 65 Avenue San Martin in Santo Domingo. During the conversation the caller states [in Arabic]: "You will receive a shipment of antibiotics at a time we will announce to you. At about the same time you will receive a shipment of medical supplies. We will let you know where these items will be going."

Document (NSA16) – Intercepted phone call on 28 September, 2004 from Havana, Cuba to Santo Domingo, Dominican Republic. The Havana call originated at 45 Desague St. in Havana listed in the name Jose Escalante; the destination was at 65 Ave. San Martin in Santo Domingo, listed in the name Carlos Morales. The caller says [in Spanish]: "Hello Carlos, I'm calling just to tell you that Arze will have the medical supplies to you by November 21st. You know who to give them to?" The reply is: " Yes, some guy named Sufaat. He also gets the antibiotics that I just heard about from the Arab. They will also be here by November 21st. Where is all this stuff going?" The caller says

: "I don't know anything about any antibiotics. I guess we are not supposed to know, but my guess is that they will go to our friends in Columbia."

Table 7.2: Bag of words and transitive terms for final story after incorporating feedback ($CIA06 \rightarrow CIA08 \rightarrow DIA01 \rightarrow NSA09 \rightarrow NSA16$) from the INTERACTIVE STORYTELLING algorithm.

Document	Bag of words	Transitive terms
CIA06	action,alias,badawi,bin,degre,document,egypt,fahd, farooq,hasham,head,insurg,khost,laden,milit,osama,pakhtia, plan,provinc,salman,special,staff,success,treat,yasir	
↓		badawi,farooq,insurg
CIA08	badawi,border,facil,farooq,group,hasham,hazmi, insurg,mamdouh,moham,omari,presenc,qaeda,refug rememb,second,show,shown,talk,wound	
↓		qaeda,talk
DIA01	abdul,ahm,captur,chemic,enter,februari,fight, interrog,member,nasser,octob,qaeda,shortli,student, taliban,talk,usa,went	
↓		octob
NSA09	avenu,convers,list,martin,morocco,octob,orang,san,shipment	
↓		list,martin,san
NSA16	arz,call,destin,don,friendget,guess,heard,hello,just list,martin,moral,san,stuff,sufaat,suppos	

The transitive terms for the new *story* is given in Table 7.2. Firstly we notice that the new story now incorporates documents which the user wanted to include in the *story* in a specific order. Secondly, in addition to the two specified documents which were identified by the user as feedback, document *DIA01* also shows up in the *story* using the transitive term ***Al Qaeda***. Based on this *story*, ***Abdul Ahmed Nasser*** might be a person of interest to the user. While the algorithm has picked up on clues from the user with respect to which terms might be more important in the story, it has not simply *stitched* together a new story by merely forcing it to pass through the feedback documents (although is also a possibility as we will see in a later example). This is clearly a strength for the algorithm as it *learns* from the feedback and provides a story that it considers more appropriate based on some criterion (the shortest path criterion in this case).

The structure of our article is as follows. In Section 8.2 we give a brief overview of topic modeling and the $A^*Search$ algorithm. Section 9 discusses the *V2PI* framework [Leman et al., 2010] in short and clarifies a few salient features about the INTERACTIVE STORYTELLING algorithm, namely, that it is *supervised, user feedback-based, constraint-specific* and

partly a reformulation of the *inverse shortest path problem*. Section 10 discusses the components of the algorithm – the distance metric inducing similarity between documents; formulation of our problem as the solution to a supervised topic modeling problem; interpretation of relationships imposed by the user as a set of inequalities derived from tolerances on edge costs in an inverse shortest path problem; probabilistic modeling of the relationships based on auxiliary variables; and, finally a Gibbs sampling based strategy. In Section 11, we discuss in detailed results from a simulated data and the Atlantic Storm data set. Finally, we conclude with possible directions for future work.

Chapter 8

Related Work

In this Section we discuss related work in topic modeling and give a brief overview of the $A^*Search$ algorithm applied to our *Storytelling* algorithm.

8.1 Approaches to User Feedback Based Topic Modeling

Topic modeling using LDA represents a document as a mixture distribution over a set of topics. Each topic is a distribution over words in the dictionary. Representing a document by the vector of normalized topic weights allows us to embed a document in the reduced dimensional simplex space of the topics. For T predefined number of topics, this space is a T dimensional simplex. The proximity induced by this embedding can then be used to understand how *close* a pair of documents are, or if one path in the document network is shorter than another. For details on different types of topic models, please refer to Blei et al. [2003], Steyvers and Griffiths [2007].

Recent development of topic modeling has progressed along three lines:

- **Using richer information from document content** – Rosen-Zvi et al. [2004] incorporated author and content to model topics, Wallach [2005] moved beyond bag of words to include ordering of words, Chemudugunta et al. [2006, 2008] extended the topic model to include specific words, or previously known concepts, along with latent topics in calculating the proximity between documents, Andrzejewski et al. [2009], Hu et al. [2011] incorporated domain knowledge in topic modeling using Dirichlet forest priors, Blei and Lafferty [2005] introduced topic models with correlated topics;
- **Using labels or external data for documents to create supervised or semi-supervised topic models** - Blei and McAuliffe [2007] modeled observed responses for a docu-

ment using a linear model in terms of the latent topic variables, Lu and Zhai [2008] gave a semi-supervised topic modeling approach to incorporate expert opinions in modeling underlying topics for text based opinion data, Ramage et al. [2009] constrained topics in LDA by defining a one-to-one correspondence between the latent topics and user tags;

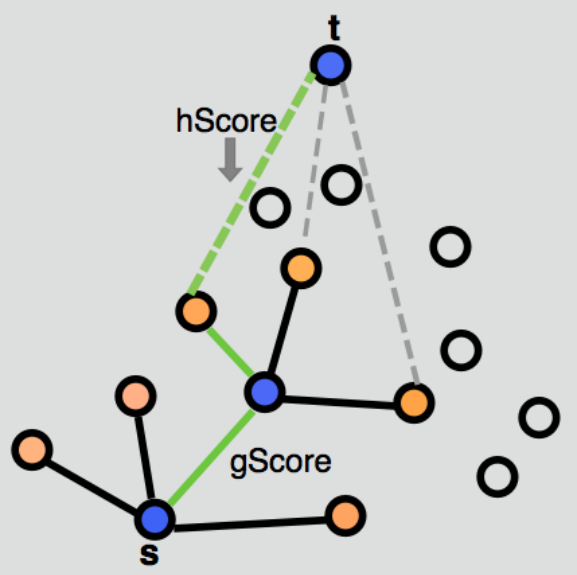
- **Developing text data mining platforms in a visual analytic framework** – Wei et al. [2010] visualized the evolution of topics over time in a corpus, Crossno et al. [2011] visually compared document contents based on different topic modeling algorithms, Mei et al. [2008] represented documents in the topic space but with connections as in a graph communities to show inter-document connectivity and community clustering simultaneously in a visual interface.

The interactive topic modeling framework spawned by current research typically involves automatic constraint generation, or generating constraints by explicitly soliciting feedback from the expert at the word or term level. In our framework the user does not interact directly with the parameters of the model which include prior correlation structures or hyperparameter specification in the hierarchy, clique size and maximum edge length threshold in the existing formulation of *Storytelling*, etc. As shown in the earlier example, instead he interacts in the visual domain to inject his feedback directly about the relationships between *data points i.e. documents*, into the underlying model – he does *not* provide any explicit feedback relative to the of terms in the documents.

8.2 $A^*Search$ Applied to *Storytelling*

The $A^*Search$ [Hart et al., 1968] algorithm provides a formal heuristic approach for finding the minimum cost path between two nodes in a graph. The path generation in *Storytelling* and INTERACTIVE STORYTELLING algorithm is based on $A^*Search$ (pseudocode below). $fScore$ is the criteria based on which the minimum cost path is generated in $A^*Search$. The $fScore$ for a node is expressed as: $fScore = gScore + hScore$. The $gScore$ for a node is the cost of the shortest path from s to the node, based on the graph traversed by $A^*Search$ (Figure 8.1). $hScore$ is the cost of the heuristic path from the node to t. For our algorithm, $hScore$ is the cost of the direct edge from the node to t. An admissible heuristic is one which does not overestimate the cost of the actual path it is estimating. Our version of $hScore$ is admissible. Under an admissible heuristic, the path generated by $A^*Search$ is indeed the shortest. Figure 8.2 shows the generation of a typical *story* using $A^*Search$. The algorithm starts from the blue node s and expands the graph locally around the node with the minimum $fScore$ to its set of neighbors (the orange nodes). The node with the minimum $fScore$ is identified at every step of $A^*Search$ thereafter, prior to expanding it again locally. The algorithm continues till the goal t is reached or terminates if a path from s to t does not exist.

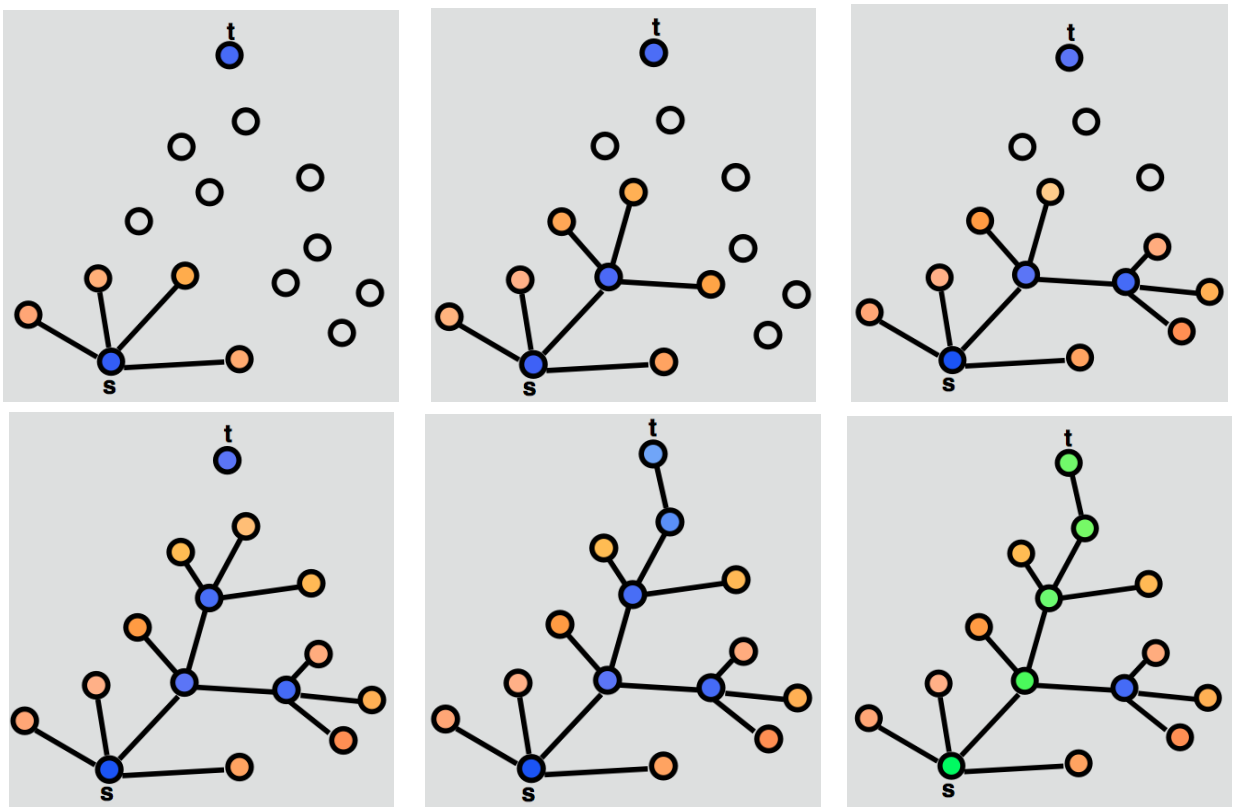
Figure 8.1: The document graph generated at a step of $A^*Search$. The cost of the solid green path is $gScore$, and the cost of the dotted green path is $hScore$. The dotted lines are heuristic edges between an open node and the goal t.



```
Initialize open list of nodes, O, to s.
Initialize closed list of nodes, C, to empty.

while (t has not been reached) {
        Select the node with lowest fScore

        if (selected node is t) {
                Generate Story
        }
        else {
                Move selected node to C and expand node to
                   nearest neighbors.

                for (each neighbor) {
                        if (neighbor ∈ C and current gScore is
                           lower) {
                                Update neighbor with new lower
                                   gScore
                                Reassign predecessor of neighbor
```

```
                                to current node
                        }
                        else if (neighbor ∈ 𝒪 and new gScore is
                           lower) {
                                Update neighbor with new lower
                                   gScore
                                Reassign predecessor of neighbor
                                   to current node
                        }
                        else neighbor ∉ O or ∉ C {
                                add the neighbor to O and set
                                   its gScore
                        }
                }
        }
}
```

Figure 8.2: Step-by-step $A^*Search$ on a document network showing the generation of a *story*. *Story* from s to t is denoted by the connected sequence of green documents. The blue node (except for s and t) is the node with the minimum $fScore$ at every stage. The set of blue nodes form the set of *Closed* set. The orange nodes are the local neighbors of the node with the minimum $fScore$ at every stage. They are the set of *Open* nodes.



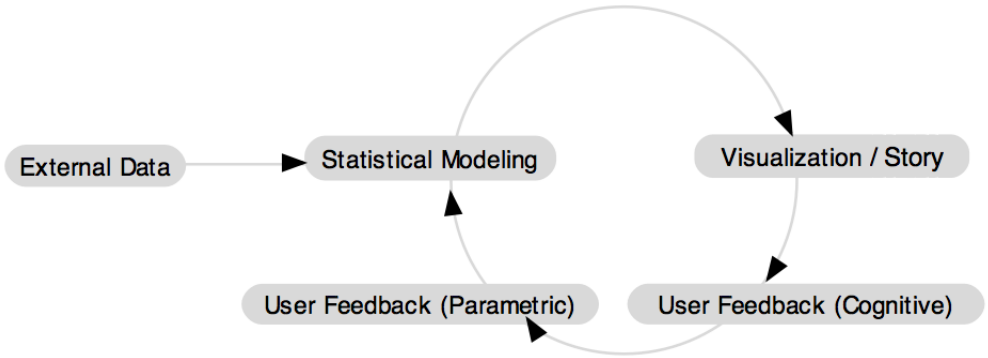
Chapter 9

General Framework of INTERACTIVE STORYTELLING

The preceding example shows that the knowledge discovery process in document networks is complicated – the conclusions about the data are as important as the path to reach those conclusions. It is further complicated in large document networks due to inherent limits to human cognition when it comes to making sense of large amounts of data – short term memory can process only as many as seven distinct pieces of information [Miller, 1956]. A viable recourse has been to devise visualization methodologies that glean important pieces of information from a large dataset and then allow the user to delve in to parts of data in more detail that he deems interesting, by interacting with the visualization. Thus, incorporating visualization provides a visual metaphor to the knowledge discovery model – e.g. proximity in spatial layout of data points automatically refer to similarity, edges in networks possibly refer to hidden relationships and scatterplots make trends visible. Add to that the plethora of human interactions that are possible on a visual display, and we have a rich framework for analyzing massive datasets. [Keim et al., 2008] also point out, that a visual framework simplifies our understanding of *analyzing our analyses* for future reference – hence replicating an atypical analysis or collaborating between multiple users and across different levels of abstraction become easier.

Analyst's Workspace [Hossain et al., 2011], Interactive Principal Component Analysis [Jeong et al., 2009], Interactive Multi-Dimensional Scaling [Buja et al., 2008] are all examples of such visual systems of knowledge discovery. The interactions here can be termed as *parameter level interaction* – the user directly modifies the parameters of the model and updates the visualization. The new visualization provides the user with a template to update his existing set of hypotheses. The challenge in implementing this visual analytic framework is that the responsibility of understanding the effect of parameters on the final visualization falls on the user – the user or the user should be well versed in the mathematical formulation of the problem to modify the underlying parameters.

Figure 9.1: The Visual To Parametric Interaction (V2PI) Framework



The INTERACTIVE STORYTELLING algorithm takes its cue from *V2PI* in that we deviate from the current methodology of parameter level interaction and propose that the user solely focus on interacting with the documents rather than the terms in the document – this type of interaction is termed *object level interaction* [Endert et al., 2011]. Since a user is more comfortable with just interacting with the data visualization and validating hypotheses, rather than having to understand the technicalities involved in the model, we believe this has broader appeal and acceptability for the users. In addition, our framework allows us to employ more sophisticated algorithms to model the data and at the same time enhance its usefulness or predictive ability based on the user's inputs. The typical steps in an interactive Storytelling algorithm based on such a framework will be as follows:

I) Algorithm provides a visualization based on initial latent topics. In INTERACTIVE STORYTELLING, this visualization is based on the shortest path in the document network between the the starting document (s) and an ending document (t), as defined by the user.

II) User interacts with observations i.e. documents in the two-dimensional visualization (reads, searches, highlights terms etc.) to inject feedback based on his semantic reasoning of the data. This feedback is in terms of documents that the user prefers in the *story* and is known as *cognitive feedback*. The user defined *story* is denoted by P^*. The user is completely shielded from the path searching and visualization algorithms.

III) P^* is not the shortest path from s to t in initial topic space. We define a mapping from the user's observation-level interaction to a mental model of reasoning; e.g. bringing documents closer implies that they are similar, or as in our case, we interpret the user's feedback as P^* being shortest over all paths from s to t in some (as of yet undiscovered) topic space. This is known as *parametric feedback*. We formulate a

system of inequalities (or relationships) denoted by $\mathfrak{R}$, where the cost (length) of the *story* P^* is constrained to be smaller than other alternate stories. The parameters of the model are re-estimated under these constraints.

IV) System regenerates a new story and an updated visualization based on the new topic space which is consistent with the system of inequalities $\mathfrak{R}$. The process continues iteratively, as does sensemaking, for the duration of the analytic process.

We interpret user feedback as relationships amongst documents that the user wants to enforce. In the sense that these relationships form user generated responses over the corpus, our algorithm is a *supervised* algorithm. As will be discussed later, these relationships induce constraints on edge costs in the document graph. Since edge costs depend on the topic space in which the documents are embedded, our algorithm is *constrained* topic modeling. The constraints on the edge costs are in terms of costs associated with P^* and a competing *story*). In the sense that we obtain edge costs given thet P^* is the shortest path between s and t, our algorithm is a formulation of the *inverse shortest path problem* in the topic space. In the next Section, we discuss these aspects of the algorithm in more detail.

Chapter 10

INTERACTIVE STORYTELLING Algorithm

Any notion of distance or similarity between two documents induces an abstract network structure amongst documents in a corpus. The notion of distance is based on terms, or semantic associations between terms, or as in our case, underlying topics in a document pair. Filtering out non-relevant terms minimizes computational overhead and reduces model complexity; this is the simplest form of dimension reduction. After appropriate filtering of terms in the corpus using stopwords or criteria like *idf* or *gini index*, the next step in creating a document network is defining the distance metric between two documents.

There have been semi-supervised or supervised metric learning approaches that *learn* about the appropriate distance metric in the term space, to provide meaningful visualizations to the user [Xing et al., 2002, Bilenko et al., 2004, Hoi et al., 2006]. However, it has been our experience, that for the more complicated user defined constraints that we intend to entertain in our framework, a metric learning method that merely reweighs the features in the distance metric does not yield satisfactory results. Often, the constraints are unsatisfiable. For example, any edge cost (based on Manhattan or squared Euclidean distance, etc) of the form $\sum_{k=1}^{T} w_k \Delta_k$, where w_k is the weight associated with the kth term, and Δ_k is the distance contribution from the kth term, does not capture weights associated with co-occurrence of terms. In our introductory example, the algorithm was able to discriminate between the occurrence of the term ***badawi*** with and without ***Al Qaeda***. When the term ***badawi*** occurs with the term ***Al Qaeda*** (and perhaps other terms) in a document, it induces a certain neighborhood structure and hence a preference for proximity towards a certain kind of documents (refer to transitive terms in Table 7.1). This preference for proximity towards documents is different when ***badawi*** occurs with term ***octob*** in a document (refer to transitive terms in Table 7.2). After incorporating feedback, the topic space in which the documents are embedded is such that document $CIA20$ is now moved further away from $CIA06$, while document $CIA08$ is brought closer to $CIA06$. This facet of any term in the corpus can not be captured using using a linearly weighted

distance metric. Using a positive definite matrix as a weight matrix for the term space is another option. But such a weighing scheme converts our optimization problem into optimizing sums and differences of quadratic functions, which is computationally cumbersome.

An attractive proposition for further dimension reduction is modeling a document as a mixture of topics. Probabilistic topic modeling not only provides insights into documents in terms of underlying topics, but in our case it also provides an intuitive distance metric that is more malleable to satisfying user defined constraints. Shortly, we will elaborate how the topic information between two documents can be incorporated in a geometric distance calculation. For a brief overview of term-weighting schemes and distance measures please refer to Noreault et al. [1981].

10.1 Probabilistic Topic Models

In probabilistic topic models, each document is associated with a mixture distribution over topics. The mixture weights over topics is a signature for the document. Each topic in the document is associated with a Dirichlet distribution over the unique terms (the vocabulary) of the corpus. To sample a term from a document (henceforth called a token), we first sample the topic for the token using a multinomial distribution; the parameters of the multinomial distribution is the set of mixture weights over the topics. After choosing the topic, a term is sampled from a multinomial distribution over the vocabulary; the parameters of the multinomial distribution over the vocabulary is a signature for the corresponding topic.

Notationally, the set of N tokens in the corpus is given by $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_N\}$, the set of document indices for the tokens is $\boldsymbol{d} = \{d_1, \ldots, d_N\}$, and the latent topic assignments for tokens is given by $\boldsymbol{z} = \{z_1, \ldots, z_N\}$. There are M dictionary terms and Q documents in the corpus. The combined data of the tokenized terms and their document indices is the term-document data denoted by $\mathfrak{D}$. The mixture weights over T topics for document d_i is given by $\theta^{(d_i)} = (\theta_1^{(d_i)}, \ldots, \theta_T^{(d_i)})$. To sample token η_i (in document d_i) we first sample the latent topic z_i for the token using the distribution $p(z_i = j) = \theta_j^{(d_i)}$. The jth topic is represented by a multinomial distribution over the M terms with parameter $\phi^{(j)}$. Hence, the term is then sampled from a distribution given by $p(\eta_i | z_i = j) = \phi_{\eta_i}^{(j)}$. We also assume conjugate Dirichlet prior distributions on the parameters $\theta^{(d_i)}$ and $\phi^{(j)}$. The hierarchical model is given by:

$$\begin{aligned}
\eta_i|z_i, \phi^{(z_i)} &\sim Discrete(\phi^{(z_i)}), \ldots \text{sample term.} \\
\phi &\sim Dirichlet(\beta) \\
z_i|\theta^{(d_i)} &\sim Discrete(\theta^{(d_i)}), \ldots \text{sample topic.} \\
\theta &\sim Dirichlet(\alpha),
\end{aligned}$$

with the complete generative model so far given by,

$$p(\boldsymbol{\eta}, \boldsymbol{z}, \boldsymbol{\phi}, \boldsymbol{\theta}|\alpha, \beta, \boldsymbol{d}) \propto \left(\prod_{i=1}^{N} \phi_{\eta_i}^{(z_i)} \theta_{z_i}^{(d_i)}\right) \left(\prod_{j=1}^{D} p(\theta^{(d_j)}|\alpha)\right) \left(\prod_{t=1}^{T} p(\phi^{(t)}|\beta)\right)$$

For given hyperparameters α and β, token to document mapping $\boldsymbol{d}$, and observed term-document frequency $\mathfrak{D}$, the goal is to estimate ϕ, the distribution over the vocabulary for every topic, and θ, the distribution over topics for every document. The MCMC sampling strategy using a Gibb's sampler to estimate ϕ and θ will be discussed in Section 10.8.

10.2 Distance Based on Topics

We will use the terms nodes and documents interchangeably. We use similarity between underlying topic distributions $\theta^{(d_i)}$ and $\theta^{(d_j)}$ for documents d_i and d_j to calculate the distance or edge cost between documents d_i and d_j. While a number of probabilistic measures of distance have been proposed (see Steyvers and Griffiths [2007]), we propose a geometrical measure of distance or edge cost, the Manhattan distance metric. We connect two documents if they share any terms. Being a cost function itself, the heuristic distance for a node m is given by the straight line distance to document t . Manhattan distance follows the triangle inequality and hence $hScore$ is an admissible heuristic for the $A^*Search$ algorithm to be employed later. $fScore(l)$ can be evaluated as the sum of $gScore(l)$ and $hScore(l)$. For details about our notation please refer to Table 10.1.

Table 10.1: Overview of notation and formulae used for edge costs, path costs and $scores$ in $A^*Search$.

Define	Notation/Formula	Explanation
Document d_i	$\theta^{(d_i)} = (\theta_1^{(d_i)}, \ldots, \theta_T^{(d_i)})$	T dimensional vector of normalized topic weights, representing a document in a simplex space.
Edge e_{ij}	$< d_i, d_j >$	Edge between documents d_i and d_j.
Cost of edge, $c(e_{ij})$	$c(e_{ij}) = c_{ij} = \sum_{t=1}^{T} \Delta_{(ij)t}$ where $\Delta_{(ij)t} = \lvert\theta_t^{(d_i)} - \theta_t^{(d_j)}\rvert$.	Manhattan distance between documents d_i and d_j.
Path P, from s to t , with L edges.	$P =< s\ , d_{P(1)}, d_{P(2)}, \ldots, d_{L-1}, t\ >$	$d_{P(i)}$ is ith document in P after s , Length of path P is L.
Shortest path from s to t.	P^*	
Cost of path, $c(P)$	$c(P) = \sum_{e_{ij} \in P} c(e_{ij}) = \sum_{t=1}^{T} \Delta_t^{(P)}$ where $\Delta_t^{(P)} = \sum_{e_{ij} \in P} \Delta_{(ij)t}$.	$\Delta_t^{(P)}$ is the contribution by the tth topic towards the total cost of P.
Cost of shortest path from s to t, P^*	$c(P^*) = \sum_{t=1}^{T} \Delta_t^*$ where $\Delta_t^* = \sum_{e_{ij} \in P^*} \Delta_{(ij)t}$. Δ_t^* is the same as $\Delta_t^{(P^*)}$.	Δ_t^* is the contribution by the tth topic towards the cost of the shortest path from s to t.
$gScore(l)$ for a node l, cost of shortest path from s to l using $A^*Search$	$gScore(l) = \sum_{t=1}^{T} \Delta_t^{gScore(l)}$	$\Delta_t^{gScore(l)}$ is the contribution by the tth topic towards the cost of shortest path from s to l.
$hScore(m)$ for heuristic distance between node m and goal node g.	$hScore(m) = \sum_{t=1}^{T} \Delta_{(mg)t}$ where $\Delta_{(mg)t} = \lvert\theta_t^{(m)} - \theta_t^{(g)}\rvert$.	Heuristic distance between nodes m and g is the Manhattan distance between the nodes.

10.3 Formulating a Supervised Topic Model

After the initial path is provided to the user, the user injects his feedback into the system depending on his mental model of the data. He selects a sequence of documents $\mathcal{C} =< C_1, \ldots, C_K >$ which should *perhaps* be included in the shortest path between documents s and t . The order of the documents in feedback is important since the investigator expects to see the documents in the order specified in $\mathcal{C}$; this order is a reflection of his understanding of how the story should progress and subsequently how evidence ought to be marshaled. The path P^* is obtained by *stitching* sequentially the shortest paths between s and C_1, C_1 and C_2 and so on, including C_K and t, all in the original topic space. Although the user's feedback is only in terms of of the sequence $\mathcal{C} =< C_1, \ldots, C_K >$, we simplify our assumption as P^* being the sequence of documents favored by the user.

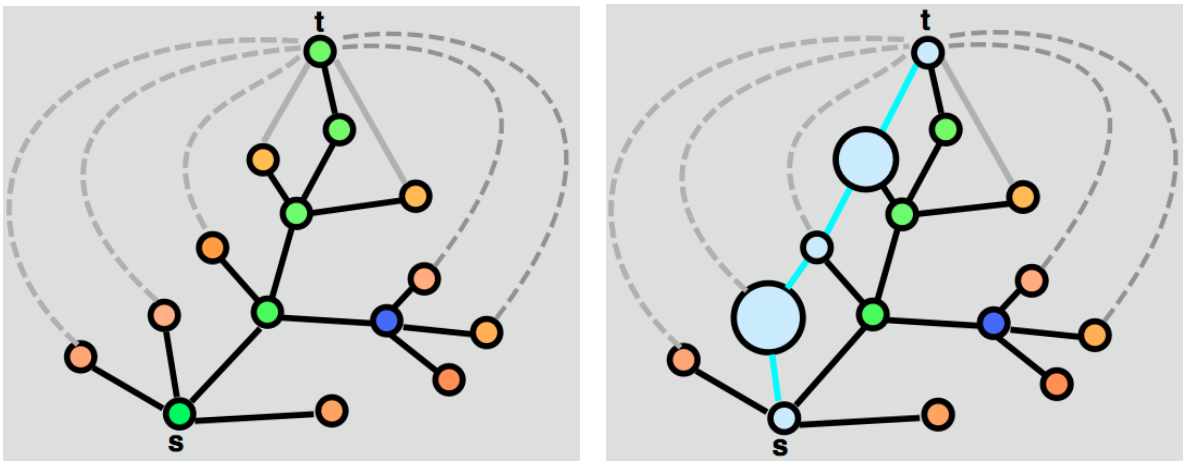
Such a feedback by the user might be motivated after reading documents in C, or after searching for certain terms that seem contextual to the story connecting documents s and t (and which are present in C). There could be various scenarios under which the investigator introduces a specific document or a sequence of documents as being pertinent to the context of the story; we however ignore such considerations that precede the investigator's feedback. We assume that there is a visual analytic platform that allows the requisite interactions for a *sensemaking* in document corpus.

The likelihood in probabilistic topic model does not explicitly take in to account any concept of similarity between two documents. By the generative model, the term distribution in a document is modeled by the mixture distribution $p(\eta_i) = \sum_{j=1}^{T} p(\eta_i|z_i = j)p(z_i = j)$. We supplement the term-document data $\mathfrak{D}$, with feedback data $\mathfrak{R}$, that specifies the user's preference of P^* over alternate paths from s to t . We use $\mathfrak{R}$ to modify the problem to a supervised learning problem. As a supervised problem, the idea is to search for parameters that *most likely* represent a document as a mixture over T independent distributions over the vocabulary, consistent with feedback represented in $\mathfrak{R}$. We now elaborate in the next few sections, what we mean by alternate *stories*, explicitly show the form of these inequalities in terms of P^* and the alternate *stories*, and derive the inequalities that represent $\mathfrak{R}$ using ideas from the *inverse shortest path problem*.

10.4 Alternate/Candidate *Stories*

Remember, the documents explored by $A^*Search$ in the original topic space i.e. the set of open and closed nodes, induce an acyclic graph denoted by $G(V, E)$. The orange nodes in Figure 10.1 are open nodes in one such graph G. Denote the open set by $\mathcal{O}$. Any path from s to t via $o \in \mathcal{O}$ is a candidate *story* based on $A^*Search$. We denote these paths by $P^{(o)}$, indexed by the open node $o \in \mathcal{O}$. For some of these paths, the heuristic cost from an orange node o to t will equal the actual path cost as there exists an edge between

Figure 10.1: **Left**: The path with green nodes is the initial *story* and hence the shortest path from s to t before incorporating feedback. The gray paths (dashed and solid) are alternate *stories* which were abandoned by the $A^*Search$. **Right**: Post feedback, the user defined *story* P^*, in blue. It is not the shortest path in *this* topic space. The documents that the user wants to be in the *story* are large circles. We intend to find the topic space where the blue path is shorter than all the other alternate paths from s to t.



o and t (solid gray edge in Figure 10.1). These paths are called *complete* paths. Paths corresponding to other open nodes which do not have direct edges to t (dashed gray path in Figure 10.1) will be referred to as *incomplete* paths.

10.5 Representing Relationships as a System of Inequalities

Let $\mathcal{O} = \{o_1, \ldots, o_O\}$ be the O open nodes in $\mathcal{O}$. To enforce the user feedback that P^* is the shortest path over all paths from s to t, our system of inequalities is:

$$\begin{aligned} c(P^*) &\leq c(P^{(o_1)}) \\ &\vdots \\ c(P^*) &\leq c(P^{(o_O)}). \end{aligned} \tag{10.1}$$

Replacing the costs with notation in Table 10.1 and simplifying our notation by indexing inequalities only by $o \in \mathcal{O}$, we obtain:

$$\begin{aligned} \sum_{t=1}^{T}(\Delta_t^* - \Delta^{(o_1)}) &\leq 0 \\ &\vdots \\ \sum_{t=1}^{T}(\Delta_t^* - \Delta^{(o_O)}) &\leq 0. \end{aligned} \tag{10.2}$$

To this set of relationships we also add another set of inequalities that constrain that the cost of an edge in the new topic space, $c(e)$, is at least as much as the cost of the edge in the initial topic space, $c_0(e)$:

$$c(e) \geq c_0(e), e \in E. \tag{10.3}$$

This is to ensure that the proximity structure of the documents does not change drastically so as to completely fluster the user. The visual analytic framework accompanying our algorithm should not result in very drastic changes in visualizations in subsequent steps such that *sensemaking* by the user in hindered. Our intention is to find the unknown parameters of the generative process, that satisfy the relationship data denoted by $\mathfrak{R}$.

10.6 Deriving System of Inequalities from Shortest Path Tolerances

Finding the shortest path in a graph between two specified nodes is a combinatorial optimization problem. The $A^*Search$ algorithm is a heuristic version of this problem. Given the shortest path in a network, the problem of finding edge costs or upper and lower limits thereof, is known as the *inverse shortest path problem*. Our goal is to find normalized topic weight vector $\theta^{(d_i)}$, for document d_i such that P^* is indeed the shortest path in the new topic space.

Table 10.2: Definitions and notations for inverse shortest path based problems.

Define	Notation/Formula	Explanation
Edge in user defined path P^*	$e^* \in P^*$	
Edge in graph G but not in user defined path P^*	$e \in E - P^*$	
Upper shortest path tolerance	β_{e^*}	Maximum cost any $e^* \in P^*$ is bounded by (all other edge costs remaining fixed), such that P^* is indeed the shortest path from s and t.
Lower shortest path tolerance	α_{e^*}	Minimum cost any $e^* \in E - P^*$ is bounded by (all other edge costs remaining fixed), such that P^* is indeed the shortest path from s and t.
	$d^{e,k}(s\ ,t\ )$	Cost of the shortest path from s to t with $c(e) = k$, all other edge costs fixed.
	$c^{e,k}(P)$	Cost of an arbitrary path P with $c(e) = k$, all other edge costs fixed.
	$d(s\ ,e,t\ )$	Cost of the shortest path from s to t via an edge $e \in E$

In our approach, we obtain the inequalities in Equations 10.2 by starting with the following observations: if the cost of an edge e^* in P^* crosses the upper threshold β_{e^*}, or if the cost of an edge e not currently in P^* falls below the lower threshold α_e, all other edge costs remaining fixed, P^* will no longer be the shortest path from s to t . Hence for P^* to be the shortest path in G,

$$c(e^*) \leq \beta_{e^*}, \text{for all } e^* \in P^*$$
$$c(e) \geq \alpha_e, \text{for all } e \in E - P^*.$$

The upper and lower shortest path tolerances were studied by Ramaswamy et al. [2005] to address questions related to sensitivity analysis for shortest paths in an undirected graph. The tolerances provided in Ramaswamy et al. [2005] are:

$$\begin{aligned} \beta_{e^*} &= d^{e^*,\infty}(s\ ,t\ ) - c(P^*) + c(e^*) \\ \alpha_e &= c(P^*) - d^{e,0}(s\ ,t\ ) \end{aligned} \tag{10.4}$$

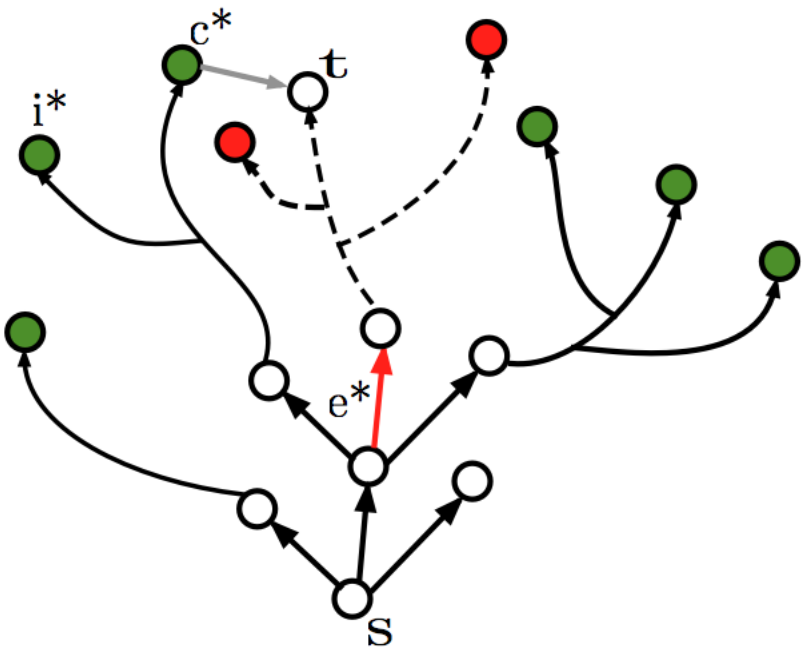
so that the inequalities for the respective edges in E become,

$$c(P^*) \leq d^{e^*,\infty}(s\ ,t\ ), \text{for all } e^* \in P^* \tag{10.5}$$

$$c(e) \geq c(P^*) - d^{e,0}(s\ ,t\ ), \text{for all } e \in E - P^*. \tag{10.6}$$

Since $c(e^*) > \beta_{e^*}$ implies $d^{e^*,\infty}(s\ ,t\ ) < c(P^*)$, which is exactly what we want to avoid, the first equality follows. To have an intuitive understanding of Equations 10.4, we note that β_{e^*} is given by the difference of two path costs – the cost of the shortest path from s to t that *bypasses* e^* (enforcing an infinite cost for e^*) i.e. $d^{e^*,\infty}(s\ ,t\ )$, and the minimum cost of P^* with e^* in the path (enforcing a zero cost for e^*), $c(P^*) - c(e^*) = c^{e^*,0}(P^*)$. Note that if $e = (l, m)$, then $d^{e,0}(s\ ,t\ ) = min(c(P^*), d(s\ ,l) + d(m, t\ ))$. For the second equality, if the shortest path from s to t is unchanged even with $c(e) = 0$ i.e. $d^{e,0}(s\ ,t\ ) = c(P^*)$, then the lower tolerance for $c(e)$ is zero. If however, the constraint $c(e) = 0$ favors an alternate path through e (and hence not P^*), the lower tolerance for e is given by the drop in path cost that this alternate path allows for over P^*.

Figure 10.2: Dashed lines represent the tree $\tau(e^*)$ and solid lines represent the tree $\tau^C(e^*)$. Green nodes are candidate open nodes in $\tau^C(e^*)$ for Equation 10.5. Red nodes are open nodes in $\tau(e^*)$ and do not contribute to Equation 10.5. The shortest path from s to t bypassing e^* is the shortest path from s to t via any of the green nodes.



To simplify our formulation of the inequalities with output from our path search algo-

rithm, we use the fact that our choice of $hScore$ is an admissible heuristic in the $A^*Search$ algorithm i.e. it does not overestimate the distance it is approximating. Hence $hScore(m) \leq d(m,t\,)$ and consequently $gScore(l) + hScore(m) \leq d(s\,,l) + d(m,t\,)$. Replacing $d^{e,0}(s\,,t\,)$ with the smaller heuristic estimate of $gScore(l) + hScore(m)$ in Equation 10.6 , we obtain a stricter inequality given by:

$$\left.\begin{array}{c} c(e) \geq c(P^*) - gScore(l) - hScore(m), \\ c(e) \geq 0, \end{array}\right\} \qquad \forall e \in E - P^*.$$

The cost of the shortest path bypassing an edge $e^* \in P^*$ is given by $d^{e^*,\infty}\,(s\,,t\,) = min_{e \in E-P^*}(d(s\,,e,t\,) \mid e^* \notin d(s\,,e,t\,))$. In Figure 10.2, let the red edge be one such $e^* \in P^*$. Denote the subtree induced by $A^*Search$, following the edge e^* by $\tau(e^*)$; this tree is denoted by dashed lines in Figure 10.2. Let the remainder of the tree corresponding to solid lines be denoted by $\tau^C(e^*)$. Based on the search process that induced the graph G, we would expect the shortest path from s to t via any edge in $\tau(e^*)$ to have e^* in it. Hence, $d^{e^*,\infty}(s\,,t\,)$ should be based on paths via edges in $\tau^C(e^*)$. Since we have output from a $A^*Search$ search, we will use path costs that are estimated heuristically by the $fScores$ for open nodes in $\tau^C(e^*)$. These open nodes are denoted by the green nodes in Figure 10.2. Hence for this specific edge e^*, the inequality $c(P^*) \leq min_{e \in E-P^*}(d(s\,,e,t\,) \mid e^* \notin d(s\,,e,t\,))$ is replaced by the following set of inequalities:

$$c(P^*) \leq fScore(o), \forall\, o \text{ in the set of open nodes in } \tau^C(e^*). \tag{10.7}$$

By the earlier logic of admissibility of $hScore$, since any $fScore$ underestimates the true distance, we are using a stricter inequality in Equation 10.7.

If this exercise is repeated for all $e^* \in P^*$, our final set of inequalities consist of the user defined path P^* being compared against all set of paths defined by the open nodes in the $A^*Search$ search algorithm, as given in Equation 10.2.

10.7 Modeling Relationships Using Auxiliary Variables

In Section 10.6 we formulated the user feedback as a set of relationships. Each relationship is an inequality in terms of path lengths, with the user defined path P^* being favored by the user over other alternate paths $P^{(o)}$ from $A^*Search$. Since the distance measure in Table 10.1 is a function of the normalized topic weights for documents, we explicitly show the dependence of an individual relationship on $\boldsymbol{\theta}$. For an inequality $r_o \equiv c(P^*) \leq c(P^{(o)})$ in Equation 10.2, we define a slack random variable λ_o (i.e. $\lambda_o < \epsilon$ for some $\epsilon < 0$), as an auxiliary variable, with expectation given by $\mathbf{E}(\lambda_\mathbf{o}) = \mu_o(\boldsymbol{\theta}) = c(P^*) - c(P^{(o)})$. Correspondingly for a relationship $r_e \equiv c(e) \geq c_0(e)$ in Equation 10.3, indexed by $e \in E$, we define a surplus random variable (i.e. λ_e is positive) with expectation given by

$\mathbf{E}(\lambda_e) = \mu_o(\boldsymbol{\theta}) = c(e) - c_0(e)$. Using Manhattan distance in the topic space (Table 10.1), $\mu_o(\boldsymbol{\theta}) = \sum_{t=1}^{T}(\Delta_t^*(\boldsymbol{\theta}) - \Delta_t^{(o)}(\boldsymbol{\theta}))$. Let the distribution of the auxiliary variable be given by $\lambda_o \sim f(.|\boldsymbol{\theta})$. For interpretation of the auxiliary variables, we focus on the slack random variables. The random variable λ_o measures the difference in path lengths between the user defined path P^* and an alternate $P^{(o)}$. If $\mu_o(\boldsymbol{\theta})$ is zero, it goes only so far as enforcing the relationship that P^* is as costly as an alternate path $P^{(o)}$. The more negative the value of its mean $\mu_o(\boldsymbol{\theta})$, the larger we would expect $P^{(o)}$ to be compared to P^*. This ensures that *on the average*, the topic space satisfies the relationship $c(P^*) \leq c(P^{(o)})$.

Conditional on a known $\boldsymbol{\theta}$, and hence a topic space to embed the documents, the joint distribution of the auxiliary variables (slack and surplus) and observed relationship data $\mathfrak{R}$ is given by:

$$f(\mathfrak{R}, \boldsymbol{\lambda}|\boldsymbol{\theta}) \propto \prod_{o \in \mathcal{O}} \{\mathbb{1}_{c(P^*) \leq c(P^{(o)})} \mathbb{1}_{\lambda_o \leq \epsilon} + \mathbb{1}_{c(e) \geq c_0(e)} \mathbb{1}_{\lambda_o \geq 0}\} f(\lambda_o|\boldsymbol{\theta}). \tag{10.8}$$

Here $\mathbb{1}_x$ is an indicator variable that is one if condition x holds and zero otherwise. Our goal is to find the set of slack and surplus variables $\boldsymbol{\lambda}$ that maximizes the probability in Equation 10.8. Let $f(\lambda_o|\boldsymbol{\theta})$ be $N(\lambda_o|\mu_o(\boldsymbol{\theta}), 1)$. By marginalizing over the auxiliary variables λ_o, our formulation is equivalent to modeling the probability of satisfying a relationship using the cumulative standard normal distribution, i.e.,

$$\begin{aligned} P(c(P^*) \leq c(P^{(o)})|\boldsymbol{\theta}) &= 1 - \Phi(\mu_o(\boldsymbol{\theta}) - \epsilon), \text{for a slack relationship in Equation 10.2,} \\ P(c(e) \geq c_0(e)|\boldsymbol{\theta}) &= \Phi(\mu_o(\boldsymbol{\theta})), \text{for a surplus relationship in Equation 10.3.} \end{aligned} \tag{10.9}$$

Here, for a standard normal variable Z, $\Phi(z) = P(Z \leq z)$. Our approach is very similar to the usage of auxiliary variables in probit regression by Albert and Chib [1993]. In probit regression, the mean of the auxiliary variable is modeled via a linear predictor with the goal of maximizing the discrimination between the *successes* and *failures* in the data. In our formulation, satisfiability of a user defined relationship is a *success*, and the probability of satisfying the relationship is modeled via the mean of the auxiliary variable. This allows us to compare and rank user defined relationships by posterior estimates of probabilities of satisfiability. Highly probable (or improbable) alternate stories can be hence be identified in the new topic space.

The mean of the auxiliary variable is a function of the topic space $\boldsymbol{\theta}$ on which the distances are defined. Our goal is to search for the parameters $\boldsymbol{\theta}$ of the topic space which explains the term distribution of a specific document as a discrete mixture, and satisfies as many of the relationships in $\mathfrak{R}$ as possible. Truncating a slack variable λ_o, to a negative region specified by ϵ allows us to search for $\boldsymbol{\theta}$ that shrinks the mean $\mu_o(\boldsymbol{\theta})$ to a negative value. The truncation ϵ was set to -20 for our examples, although in reality it could be the largest

value such that P^* is the shortest path between s and t.

The complete hierarchical model using the term-document frequency data $\boldsymbol{\eta}$ and relationship data $\mathfrak{R}$ is then given by:

$$\begin{aligned} f(\mathfrak{R}, \boldsymbol{\lambda}|\boldsymbol{\theta}) &\propto \prod_{o\in\mathcal{O}} \{\mathbb{1}_{c(P^*)\leq c(P^{(o)})}\mathbb{1}_{\lambda_o\leq\epsilon} + \mathbb{1}_{c(e)\geq c_0(e)}\mathbb{1}_{\lambda_o\geq 0}\} N(\lambda_o|\mu_o(\boldsymbol{\theta}), 1) \\ \eta_i|z_i, \phi^{(z_i)} &\sim Discrete(\phi^{(z_i)}) \\ \phi &\sim Dirichlet(\beta) \\ z_i|\theta^{(d_i)} &\sim Discrete(\theta^{(d_i)}) \\ \theta &\sim Dirichlet(\alpha), \end{aligned} \tag{10.10}$$

In Section 10.8, we provide a Gibb's sampling based approach that allows joint inference on $\boldsymbol{\theta}$ and λ.

10.8 Sampling Strategy

A Gibbs sampling strategy requires conditional posterior distributions for $\boldsymbol{z}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$. The conditional posterior distribution for z_i is given by:

$$p(z_i = j|\boldsymbol{z}_{(-i)}, \boldsymbol{\eta}) \propto p(\eta_i|z_i = j, \boldsymbol{z}_{(-i)}, \boldsymbol{\eta}_{(-i)}) p(z_i = j|\boldsymbol{z}_{(-i)}, \boldsymbol{\eta}_{(-i)}). \tag{10.11}$$

The first term in Equation 10.11 can be obtained by marginalizing over $\phi^{(j)}$ as,

$$p(\eta_i|z_i = j, \boldsymbol{z}_{(-i)}, \boldsymbol{\eta}_{(-i)}) = \int p(\eta_i|z_i = j, \phi^{(j)}) p(\phi^{(j)}|\boldsymbol{z}_{(-i)}, \boldsymbol{\eta}_{(-i)}) d\phi^{(j)}, \tag{10.12}$$

where $\phi^{(j)}$ is the distribution over terms in the dictionary for topic j, and the integral is defined over all such distributions over the M dimensional simplex . The second term in Equation 10.12 is the posterior predictive distribution for $\phi^{(j)}$ based on all the remaining data, and can be obtained as,

$$p(\phi^{(j)}|\boldsymbol{z}_{(-i)}, \boldsymbol{\eta}_{(-i)}) \propto p(\boldsymbol{\eta}_{(-i)}|\phi^{(j)}, \boldsymbol{z}_{(-i)}) p(\phi^{(j)}|\beta). \tag{10.13}$$

Since the prior on $\phi^{(j)}$ is $Dirichlet(\beta)$ and conjugate to the multinomial likelihood for $p(\boldsymbol{\eta}_{(-i)}|\phi^{(j)}, \boldsymbol{z}_{(-i)})$, the posterior distribution $p(\phi^{(j)}|\boldsymbol{z}_{(-i)}, \boldsymbol{\eta}_{(-i)})$ is given by $Dirichlet(\beta + n^{(\eta)}_{(-i,j)})$, where $n^{(\eta)}_{(-i,j)}$ is number of occurrences of term η assigned to the jth topic, except for the current term. Since Equation 10.12 is the expectation of $\phi^{(j)}_{\eta_i}$ with respect to

$Dirichlet(\beta + n^{(\eta)}_{(-i,j)})$, integrating over $\phi^{(j)}$ gives,

$$p(\eta_i | z_i = j, \boldsymbol{z}_{(-i)}, \boldsymbol{\eta}_{(-i)}) = \frac{\beta + n^{(\eta_i)}_{(-i,j)}}{M\beta + n^{(.)}_{(-i,j)}}, \tag{10.14}$$

where $n^{(.)}_{(-i,j)}$ is the number of assignments of terms to topic j, excluding the current term. Similarly,

$$p(z_i = j | \boldsymbol{z}_{(-i)}) = \int p(z_i = j | \theta^{d_i}) p(\theta^{(d_i)} | \boldsymbol{z}_{(-i)}) d\theta^{(d_i)} = \frac{\alpha + n^{(d_i)}_{(-i,j)}}{T\alpha + n^{(d_i)}_{(-i,.)}}, \tag{10.15}$$

where $n^{(d_i)}_{(-i,j)}$ is the number of tokens assigned to topic j in document i, excluding the current term. Combining Equations 10.11 and 10.15, we obtain,

$$p(z_i = j | \boldsymbol{z}_{(-i)}, \boldsymbol{\eta}) \propto \frac{\beta + n^{(\eta_i)}_{(-i,j)}}{M\beta + n^{(.)}_{(-i,j)}} \frac{\alpha + n^{(d_i)}_{(-i,j)}}{T\alpha + n^{(d_i)}_{(-i,.)}}. \tag{10.16}$$

The full conditional distribution for λ_o (corresponding to relationship r_o) is given by,

$$p(\lambda_o | \boldsymbol{\theta}, \Re) = \begin{cases} N(.|\mu_o(\boldsymbol{\theta}), 1), \lambda_o \leq \epsilon & \text{if relationship } r_o \text{ is} \leq \text{ type,} \\ N(.|\mu_o(\boldsymbol{\theta}), 1), \lambda_o > 0 & \text{if relationship } r_o \text{ is} > \text{ type.} \end{cases}$$

The full conditional distribution for the topic distribution of document d_j is,

$$\begin{aligned} p(\theta^{(d_j)} | \boldsymbol{\theta}^{(-d_j)}, \boldsymbol{\lambda}, \boldsymbol{z}) &\propto \prod_{z_i \in d_j} p(z_i | \theta^{(d_j)}) \, p(\theta^{(d_j)} | \alpha). \prod_{o \in \mathcal{O}} N(\lambda_o | \mu_o(\boldsymbol{\theta}), 1) \\ &\propto p(\theta^{(d_j)} | \boldsymbol{z}, , \alpha) \prod_{o \in \mathcal{O}} N(\lambda_o | \mu_o(\boldsymbol{\theta}), 1) \\ &\propto \prod_{t=1}^{T} \left(\theta_t^{(d_j)} \right)^{(n_t^{(d_j)} + \alpha) - 1} \prod_{o \in \mathcal{O}} N(\lambda_o | \mu_o(\boldsymbol{\theta}), 1), \end{aligned} \tag{10.17}$$

since $p(\theta^{(d_j)} | \boldsymbol{z}, , \alpha) = Dirichlet(n_t^{(d_j)} + \alpha)$. Here $n_t^{(d_j)}$ is the number of terms from document d_j assigned to topic t based on $\boldsymbol{z}$. If document d_j is not part of any relationship in $\Re$, $\theta^{(d_j)}$ is sampled from $Dirichlet(n_t^{(d_j)} + \alpha)$. Otherwise, we sample from $p(\theta^{(d_j)} | \theta^{(-d_j)}, \boldsymbol{\lambda}, \boldsymbol{z})$ using a Metropolis-Hastings step. To allow better mixing, we use a proposal strategy inspired by the stick-breaking process. The stick-breaking process automatically bounds the individual $\theta_t^{(d_j)}$ between zero and one, and their sum to one. We first propose random variables $u_1, \ldots, u_{T-1}$ truncated between zero and one, and centered by scaled $\boldsymbol{\theta}^{(d_j)}$, using

a proposal distribution $q(.)$:

$$\begin{aligned}
u_1 &\sim q\left(.|\theta_1^{(d_j)}\right), 0 < u_1 < 1 \\
u_2 &\sim q\left(.|\frac{\theta_2^{(d_j)}}{1-u_1}\right), 0 < u_2 < 1 \\
u_3 &\sim q\left(.|\frac{\theta_3^{(d_j)}}{(1-u_1)(1-u_2)}\right), 0 < u_3 < 1 \\
&\vdots \\
u_{T-1} &\sim q\left(.|\frac{\theta_{T-1}}{(1-u_1)(1-u_2)\ldots(1-u_{T-2})}\right), 0 < u_{T-1} < 1,
\end{aligned} \tag{10.18}$$

followed by the mapping, $S : \boldsymbol{u} \rightarrow \boldsymbol{\theta}_{1:T-1}^{*(d_j)}$,

$$\begin{aligned}
\theta_1^{*(d_j)} &= u_1 \\
\theta_2^{*(d_j)} &= u_2(1-u_1) \\
\theta_3^{*(d_j)} &= u_3(1-u_2)(1-u_1) \\
&\vdots \\
\theta_{T-1}^{*(d_j)} &= (1-u_{T-1})(1-u_{T-2})\ldots(1-u_2)(1-u_1).
\end{aligned} \tag{10.19}$$

The inverse mapping, $S^{-1} : \boldsymbol{\theta}_{1:T-1}^{*(d_j)} \rightarrow \boldsymbol{u}$, is given by,

$$\begin{aligned}
u_1 &= \theta_1^{*(d_j)} \\
u_t &= \frac{\theta_t^{*(d_j)}}{1-\sum_{i<j}\theta_i^{*(d_j)}}, t = 2, \ldots, T-1.
\end{aligned} \tag{10.20}$$

The Metropolis-Hastings acceptance probability for such a proposed move is given by:

$$p_{MH} = min\left(1, \frac{\left(p(\theta^{*(d_j)}|\boldsymbol{z})\prod_{o\in\mathcal{O}}N(\lambda_o|\mu_o(\boldsymbol{\theta}^*),1)\right)}{\left(p(\theta^{(d_j)}|\boldsymbol{z})\prod_{o\in\mathcal{O}}N(\lambda_o|\mu_o(\boldsymbol{\theta}),1)\right)} \times \frac{q(\boldsymbol{\theta^{*(d_j)}}_{1:T-1})}{q(\boldsymbol{u})}\left|\frac{\partial(\boldsymbol{\theta^{*(d_j)}}_{1:T-1})}{\partial(\boldsymbol{u})}\right|\right),$$

where,

$$\left|\frac{\partial(\boldsymbol{\theta^{*(d_j)}}_{1:T-1})}{\partial(\boldsymbol{u})}\right| = \left|\frac{\partial(\boldsymbol{u})}{\partial(\boldsymbol{\theta^{*(d_j)}}_{1:T-1})}\right|^{-1} = \left(\frac{1}{\prod_{t=2}^{T-1}\left(1-\sum_{i<t}\theta_i^{*(d_j)}\right)}\right)^{-1}.$$

The samples from $\boldsymbol{z}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are iteratively sampled using the full conditional distribu-

tions, to generate the joint posterior distribution of all the unknown parameters by Gibbs sampling. We apply a thinning of every 20 samples to adjust for autocorrelation from the MCMC samples. Multiple random starting points are used for the sampler to make sure that the sampler does not get stuck in a local mode.

Chapter 11

Examples

11.1 Simulated Data Example

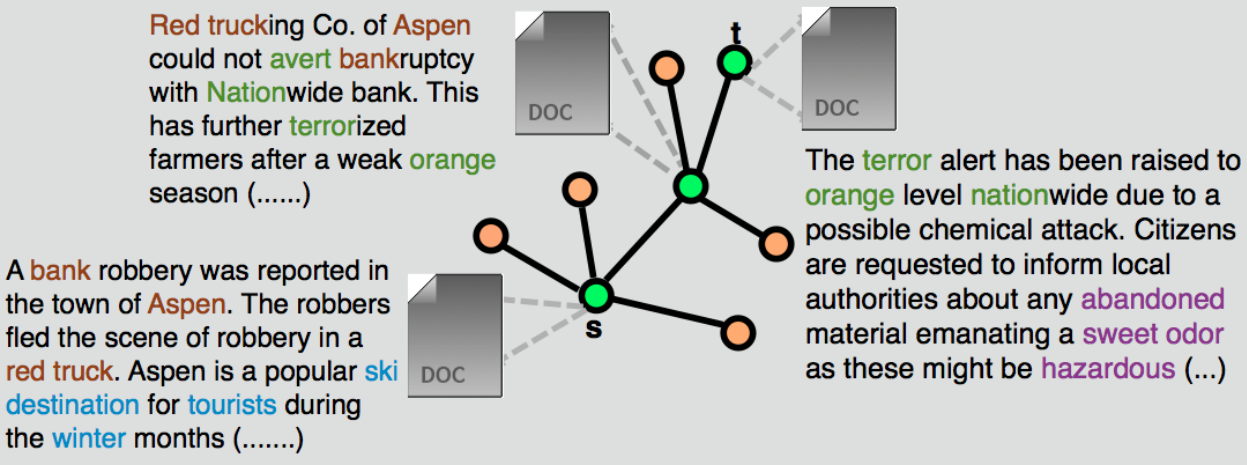
We simulate data to create a corpus of fifty documents, with terms that represent nine predefined *themes* and some random *noise* terms. Each *theme* is represented by a bag of four terms. Some example $themes$ are as follows: *Theme 1* – ***nation, terror, avert, orange***; *Theme 3* – ***hazardous, abandoned, sweet, smell***; *Theme 5* - ***ski, tourist, destination, winter***; *Theme 7* - ***bank, red, truck, aspen***; *Theme 8* - ***chemical, factory, recently, hiring***. Each document is generated using a single theme or a mixture of two themes. Apart from terms sampled from their respective theme assignments, each document also has a two *noise* terms which show up only in a specific document i.e. any two documents do not share any noise terms. The size of the dictionary is 136 terms in total – four terms for each of the nine themes and two noise terms for each of the fifty documents. Hence size of dictionary is $M = 136$, and size of corpus is $Q = 50$ documents. A pair of documents have an edge between them if they have at least one term in common. Note that since a noise term shows up in only one document in the corpus, none of the noise terms are responsible for any edge formation between documents, and hence are assumed to be of no significance to the user. Notationally, $d_1(5 \dots 6)$ represents $Document\ 1$ in the corpus with its bag of terms being generated by *Themes* 5 and 6. The indices 1 to 50 for the fifty documents in the corpus are used as identifiers.

In the following subsections, we first describe the steps involved with the INTERACTIVE STORYTELLING process applied to our data and compare *stories* before and after incorporating the user's feedback. We also show how our algorithm suggests alternate *stories* ranked by a measure of divergence from the user specified *story*. As we will see, the estimated parameters $\boldsymbol{\theta}$ of the topic space from supervised LDA often differ from the empirical composition of documents from simple LDA in the initial step. We explain this on two fronts – 1) we define a measure of association between terms which actually occur in

the document and terms which do not occur in the document, and 2), we show this measure of association is consistent at least with the proximity structure imposed amongst the documents based on the INTERACTIVE STORYTELLING algorithm.

11.1.1 INTERACTIVE STORYTELLING Applied to Simulated Data

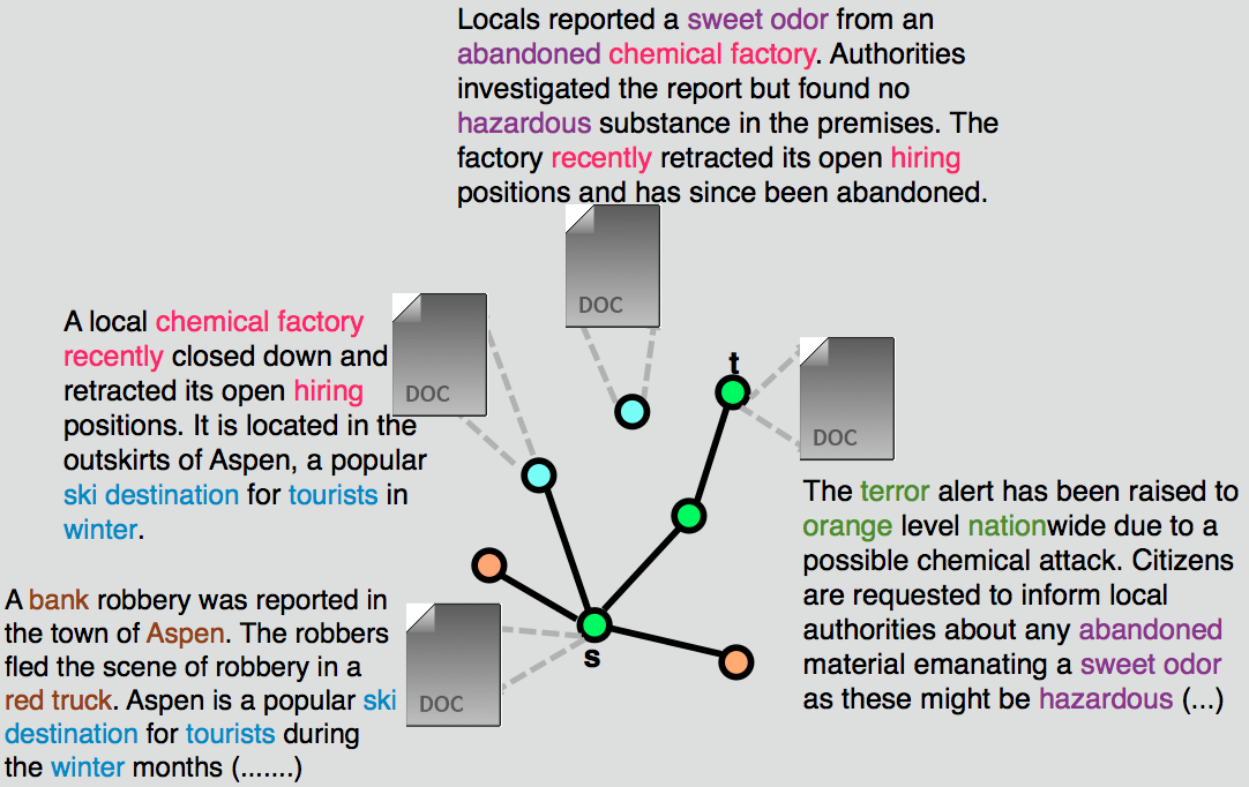
Figure 11.1: A user specifies a starting document s, describing a bank robbery, and an ending document t that alerts of a possible chemical attack. The *Storytelling* algorithm generates a *story* which connects the two documents via a document that talks about bankruptcies due to fall in orange production. The user is *not satisfied* with this *story*.



Once a predefined number of topics (in our example ten) are generated by LDA, the vector of normalized topic weights for each document induces a distance between every pair of documents. This distance information, coupled with the edge information, is used to generate a an MDS based visualization of the corpus. The user interacts with the documents and intends to understand how the documents $d_{43}(5\ldots7)$ and $d_{23}(1\ldots3)$ are linked to each other as a *story*. The document d_{43} describes a bank robbery and d_{23} mentions of a possible chemical threat. The *Storytelling* algorithm generates the story as: $d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$. The *story* is generated by searching for the shortest path from $d_{43}(5\ldots7)$ and $d_{23}(1\ldots3)$ using $A^*Search$ algorithm (Figure 11.1). The first two documents are connected using terms ***bank, red, truck, aspen*** (*theme 7*) which refer to the bank robbery using a red truck in Aspen; however the same terms in document $d_{27}(1\ldots7)$ refer to the bankruptcy of the Red trucking company in Aspen due to drop in orange production. Similarly, while the connection between documents $d_{27}(1\ldots7)$ and $d_{23}(1\ldots3)$ use the same bag of terms, their contexts are vastly different.

It is important to note that the user does not see the underlying *themes*. That overlapping terms have connected two documents and hence created a story is not of primary interest to the user; his primary goal is to read the corresponding documents (d_{43}, d_{27} and

Figure 11.2: User injects feedback by specifying two documents (blue circles) which *should* be in the *story* based on his opinion. The first document refers to the closure of a chemical factory, and the second document refers to a sweet odor characteristic of chemical weapons, emanating from a closed chemical factory.

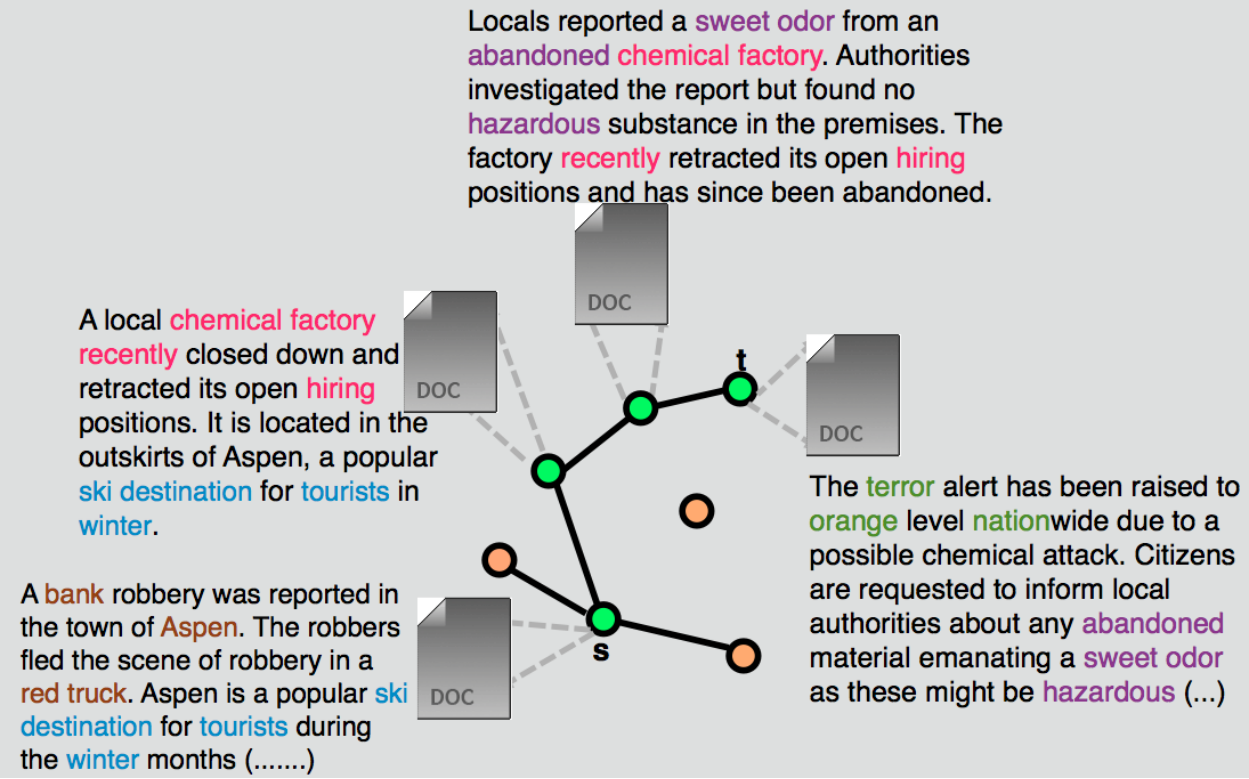


d_{23}) and validate the veracity of hypotheses that have been offered in the story. In this case the user realizes that such a story does not make much sense from his viewpoint and he tries to incorporate his expert knowledge in driving the story. He insists that documents $\mathcal{C} =< d_4(5\ldots8), d_{22}(1\ldots8) >$ should be in the story in the aforementioned order (Figure 11.2). The first document describes the closure of a chemical factory, and the second mentions sweet odor (very characteristic of chemical weapons like lewisite) emanating from a chemical factory. The user believes that these two documents might play a role in the final *story*. Based on this feedback, the user specified path from $d_{43}(5\ldots7)$ to $d_{23}(1\ldots3)$ is $P^* \equiv d_{43}(5\ldots7) \rightarrow d_4(5\ldots8) \rightarrow d_{22}(1\ldots8) \rightarrow d_{23}(1\ldots3)$ – this is the shortest path from $d_{43}(5\ldots7)$ to $d_{23}(1\ldots3)$ via the documents $d_4(5\ldots8)$ and $d_{22}(1\ldots8)$. P^* compared with a set of complete and incomplete alternate paths obtained from the initial $A^*Search$ run, to define relationships as discussed in Section 10.5 (alternate paths obtained for our simulated data from $A^*Search$ are listed in Table 11.3).

Subsequent to such a feedback, the INTERACTIVE STORYTELLING algorithm gives new topic definitions over the dictionary of terms (i.e. new topics as bag of terms), and a

new vector of normalized topic weights for each document. Under the new distance information in the graph based on the new topic weights for each document, a new story is generated: $d_{43}(5 \ldots 7) \rightarrow d_4(5 \ldots 8) \rightarrow d_{22}(1 \ldots 8) \rightarrow d_{23}(1 \ldots 3)$. In this case, the first and second documents get connected by terms ***ski, tourist, destination, winter*** (*Theme 5*), the second and third documents get connected by terms ***chemical, factory, recently, hiring*** (*Theme 8*), and the last two are connected by terms ***nation, terror, avert, orange*** (*Theme 1*). The user again might not recognize all the overlapping terms that have indeed caused the story to form, but might be satisfied with the overall *flow* of the story and hence the hypotheses the story offers.

Figure 11.3: *Story* after incorporating user's feedback based on INTERACTIVE STORYTELLING. The first two documents are linked based on the Aspen connection, the next two documents based on the abandonment of chemical factories, and the last two based on a typical odor from chemical weapons.



11.1.2 Comparing *Stories* Before and After Feedback

The final story from INTERACTIVE STORYTELLING is consistent with the user's feedback in that it includes the two documents in the order specified by the user. In general, the specified documents might not be in the final *story*, although the *themes* defining them

are expected to be in the *story*. The top ten *stories* i.e. shortest paths from $d_{43}(5\ldots7)$ to $d_{23}(1\ldots3)$, (using Yen's algorithm in Martins and Pascoal [2003]), using the topic space prior to incorporating feedback are listed in Table 11.1. We compare the content of these stories to those obtained using the topic space from supervised LDA in INTERACTIVE STORYTELLING. As shown in Table 11.1, the stories prior to the user's feedback are dominated by transitive terms ***ski, tourist, destination, winter, bank, red, truck, aspen*** (*Themes* 5 and 7). Post feedback, using the INTERACTIVE STORYTELLING algorithm, the stories in Table 11.2 are predominantly dominated by transitive terms ***chemical, factory, recently, hiring*** (*Theme* 8) and occasionally by *Themes* 9 and 6 with corresponding bags of terms which are not pertinent to our understanding the algorithm. Hence, our constraints have successfully induced a proximity structure amongst the documents in the graph such that the generated stories now prefer the terms ***chemical, factory, recently, hiring*** in their paths.

Table 11.1: Top 10 *Stories* (shortest paths) from $d_{43}(5\ldots7)$ to $d_{23}(1\ldots3)$ due to the STORYTELLING algorithm, based on the graph induced amongst the documents, by weighted topic vectors from the unsupervised LDA model.

Path from $d_{43}(5\ldots7)$ to $d_{23}(1\ldots3)$	Length of path
$d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2353
$d_{43}(5\ldots7) \rightarrow d_{20}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2353
$d_{43}(5\ldots7) \rightarrow d_{45}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2353
$d_{43}(5\ldots7) \rightarrow d_{20}(5\ldots7) \rightarrow d_{45}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2353
$d_{43}(5\ldots7) \rightarrow d_{34}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2396
$d_{43}(5\ldots7) \rightarrow d_{20}(5\ldots7) \rightarrow d_{34}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2396
$d_{43}(5\ldots7) \rightarrow d_{45}(5\ldots7) \rightarrow d_{34}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2396
$d_{43}(5\ldots7) \rightarrow d_{20}(5\ldots7) \rightarrow d_{45}(5\ldots7) \rightarrow d_{34}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2396
$d_{43}(5\ldots7) \rightarrow d_{45}(5\ldots7) \rightarrow d_{20}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2575
$d_{43}(5\ldots7) \rightarrow d_{34}(5\ldots7) \rightarrow d_{45}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	0.2617

As explained in Section 10.7, our estimate $\hat{\mu}_o(\boldsymbol{\theta})$ gives us a measure of divergence of any path $P^{(o)}$ from the user specified path P^*. This measure of divergence can be used to rank alternate paths by their expect costs – the smaller the value of $\hat{\mu}_o(\boldsymbol{\theta})$, the higher the cost of $P^{(o)}$ relative to the user specified path P^*, and hence perhaps of lesser interest to the user. The benefit of our fully Bayesian approach is that the estimate $\hat{\mu}_o(\boldsymbol{\theta})$ and hence the ranking is obtained as an output of the algorithm. The top section of Table 11.3 lists the complete paths or *stories*, and the bottom section lists the incomplete *stories* (a heuristic link in an incomplete story is denoted by $\rightsquigarrow$), both sections sorted by increasing value of $\hat{\mu}_o(\boldsymbol{\theta})$. For complete paths, the corresponding stories are clearly ranked by their costs (i.e. path lengths), since a story which shows a higher level of divergence from the user

Table 11.2: Top 10 *Stories* (shortest paths) from $d_{43}(5 \dots 7)$ to $d_{23}(1 \dots 3)$ due to the INTERACTIVE STORYTELLING algorithm, based on the graph induced amongst the documents, by weighted topic vectors from the supervised LDA model after incorporating user's feedback.

Path from $d_{43}(5 \dots 7)$ to $d_{23}(1 \dots 3)$	Length of path
$d_{43}(5 \dots 7) \rightarrow d_4(5 \dots 8) \rightarrow d_{22}(1 \dots 8) \rightarrow d_{23}(1 \dots 3)$	0.2347
$d_{43}(5 \dots 7) \rightarrow d_{12}(5 \dots 8) \rightarrow d_{22}(1 \dots 8) \rightarrow d_{23}(1 \dots 3)$	0.2591
$d_{43}(5 \dots 7) \rightarrow d_{27}(1 \dots 7) \rightarrow d_{23}(1 \dots 3)$	0.2673
$d_{43}(5 \dots 7) \rightarrow d_{45}(5 \dots 7) \rightarrow d_4(5 \dots 8) \rightarrow d_{22}(1 \dots 8) \rightarrow d_{23}(1 \dots 3)$	0.2772
$d_{43}(5 \dots 7) \rightarrow d_{27}(1 \dots 7) \rightarrow d_{22}(1 \dots 8) \rightarrow d_{23}(1 \dots 3)$	0.2843
$d_{43}(5 \dots 7) \rightarrow d_{27}(1 \dots 7) \rightarrow d_{50}(1 \dots 9) \rightarrow d_{23}(1 \dots 3)$	0.2903
$d_{43}(5 \dots 7) \rightarrow d_{34}(5 \dots 7) \rightarrow d_4(5 \dots 8) \rightarrow d_{22}(1 \dots 8) \rightarrow d_{23}(1 \dots 3)$	0.2960
$d_{43}(5 \dots 7) \rightarrow d_{12}(5 \dots 8) \rightarrow d_4(5 \dots 8) \rightarrow d_{22}(1 \dots 8) \rightarrow d_{23}(1 \dots 3)$	0.2986
$d_{43}(5 \dots 7) \rightarrow d_1(5 \dots 6) \rightarrow d_{12}(5 \dots 8) \rightarrow d_{22}(1 \dots 8) \rightarrow d_{23}(1 \dots 3)$	0.2993
$d_{43}(5 \dots 7) \rightarrow d_4(5 \dots 8) \rightarrow d_{12}(5 \dots 8) \rightarrow d_{22}(1 \dots 8) \rightarrow d_{23}(1 \dots 3)$	0.3000

specified story P^* has a higher cost associated with it. More importantly, it shows the user that incorporating $d_{25}(1 \dots 6)$ or $d_{16}(1 \dots 2)$ in the current *story* would result in bringing documents which are further away in the existing topic space. This might suggest that *Themes* 6 and 2 might not be compatible with his existing set of hypotheses connecting document $d_{43}(5 \dots 7)$ to $d_{23}(1 \dots 3)$ via transitive terms ***chemical, factory, recently, hiring*** (*Theme* 8). For incomplete *stories*, the correlation between $\hat{\mu}_o(\boldsymbol{\theta})$ and the true cost of a *story* is not so clear, although *Themes* 4 and 2 represent longer *stories*. We believe that a better heuristic distance measure in $A^*Search$ will show a high level of correlation between $\hat{\mu}_o(\boldsymbol{\theta})$ and the true cost of incomplete *stories*.

Figure 11.4: Each data point corresponds to a relationship $c(P^*) \leq c(P^{(o)})$. **X Axis**: Estimated value of $\mu(\boldsymbol{\theta})$ for a relationship. **Y Axis**: True length of *story*, $P^{(o)}$. Refer to Table 11.3 for data. The more negative the estimate of $\mu(\boldsymbol{\theta})$, the longer is the length of the *story* compared to the user defined *story*, and hence perhaps less consistent with user feedback. The circles and squares correspond to complete and incomplete *stories* respectively.

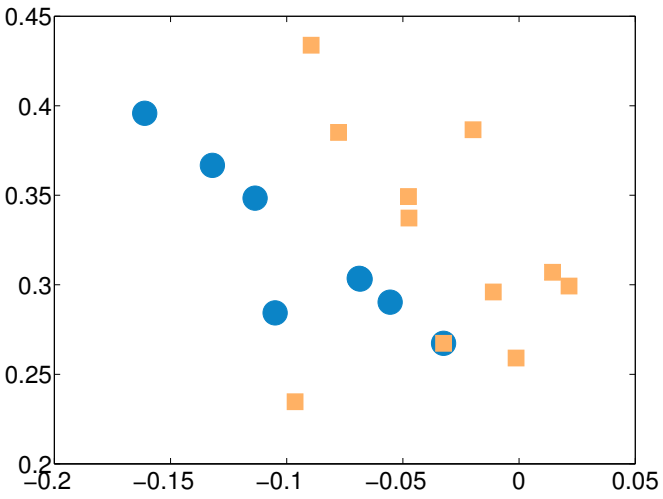
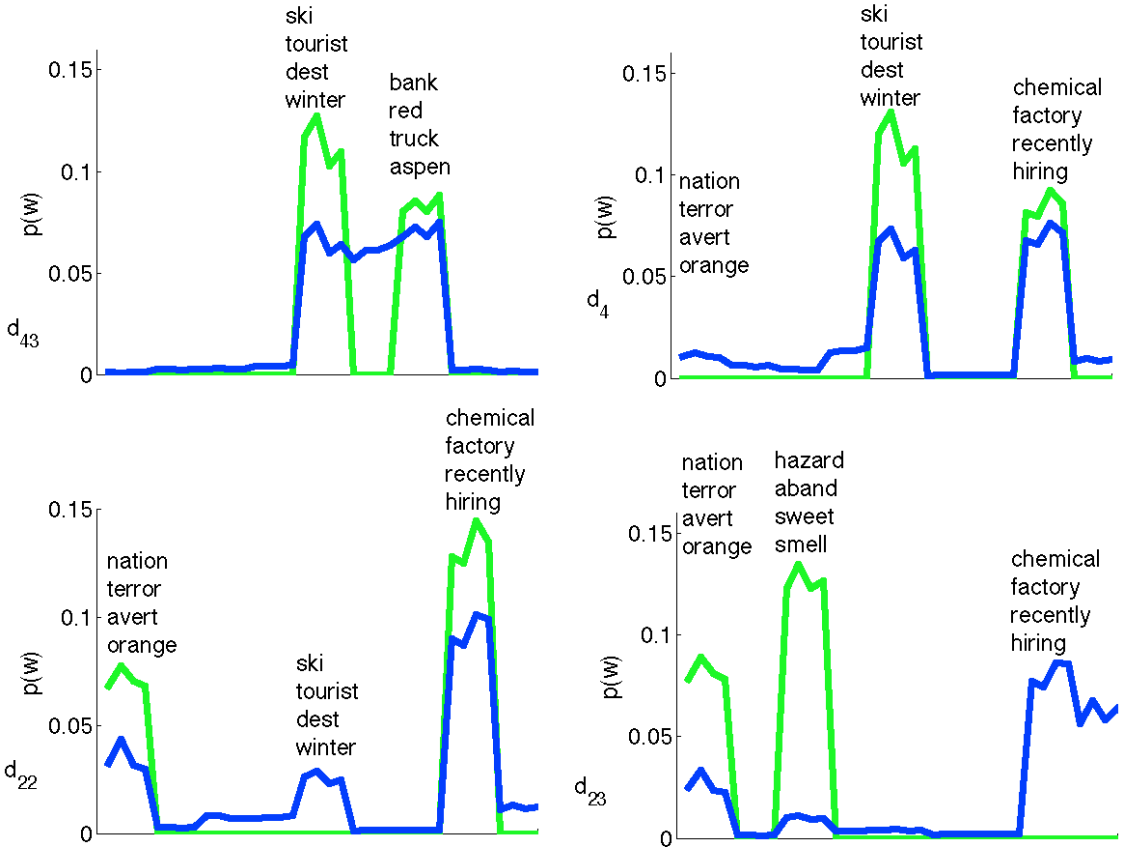


Table 11.3: Comparing measure of divergence with respect to user specified *story*, $\hat{\mu}_o(\boldsymbol{\theta})$, of complete (top section)and incomplete (bottom section) *stories* from $A^*Search$. True cost of *story* is path length of the *story*. The heuristic links in incomplete *stories* denoted by $\rightsquigarrow$. For incomplete *stories* the shortest path with the available information was obtained.

Complete and incomplete *stories* connecting $d_{43}(5\ldots7)$ to $d_{23}(1\ldots3)$ from A*Search	$\hat{\mu}_o(\boldsymbol{\theta})$	True cost of *story*
$d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{25}(1\ldots6) \rightarrow d_{23}(1\ldots3)$	-0.1611	0.3958
$d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{16}(1\ldots2) \rightarrow d_{23}(1\ldots3)$	-0.1320	0.3667
$d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{13}(1\ldots8) \rightarrow d_{23}(1\ldots3)$	-0.1136	0.3484
$d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{22}(1\ldots8) \rightarrow d_{23}(1\ldots3)$	-0.1050	0.2843
$d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{21}(1\ldots3) \rightarrow d_{23}(1\ldots3)$	-0.0688	0.3036
$d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{24}(1) \rightarrow d_{23}(1\ldots3)$	-0.0684	0.3032
$d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{50}(1\ldots9) \rightarrow d_{23}(1\ldots3)$	-0.0555	0.2903
$d_{43}(5\ldots7) \rightarrow d_{27}(1\ldots7) \rightarrow d_{23}(1\ldots3)$	-0.0325	0.2673
$d_{43}(5\ldots7) \rightarrow d_{4}(5\ldots8) \rightsquigarrow d_{23}(1\ldots3)$	-0.0964	0.2347
$d_{43}(5\ldots7) \rightarrow d_{26}(7) \rightsquigarrow d_{23}(1\ldots3)$	-0.0895	0.4338
$d_{43}(5\ldots7) \rightarrow d_{8}(2\ldots5) \rightsquigarrow d_{23}(1\ldots3)$	-0.0777	0.3851
$d_{43}(5\ldots7) \rightarrow d_{41}(4\ldots7) \rightsquigarrow d_{23}(1\ldots3)$	-0.0476	0.3493
$d_{43}(5\ldots7) \rightarrow d_{7}(4\ldots5) \rightsquigarrow d_{23}(1\ldots3)$	-0.0474	0.3373
$d_{43}(5\ldots7) \rightarrow d_{47}(7\ldots9) \rightsquigarrow d_{23}(1\ldots3)$	-0.0198	0.3866
$d_{43}(5\ldots7) \rightarrow d_{34}(5\ldots7) \rightsquigarrow d_{23}(1\ldots3)$	-0.0111	0.2960
$d_{43}(5\ldots7) \rightarrow d_{12}(5\ldots8) \rightsquigarrow d_{23}(1\ldots3)$	-0.0013	0.2591
$d_{43}(5\ldots7) \rightarrow d_{1}(5\ldots6) \rightsquigarrow d_{23}(1\ldots3)$	0.0215	0.2993
$d_{43}(5\ldots7) \rightarrow d_{11}(7\ldots9) \rightsquigarrow d_{23}(1\ldots3)$	0.0144	0.3070

11.1.3 Understanding Term-Document Distributions Before and After Feedback

Figure 11.5: Plot of probability weights for terms before (green) and after (blue) feedback for documents in *story* after feedback. Terms not occurring in the document have non-negligible weights to induce proximity that is consistent with user feedback.



The essence of the INTERACTIVE STORYTELLING algorithm is supervised LDA under path based relationships imposed by the user. Such imposed relationships might result in parameters for the generative process that do not agree with the actual contents of the document i.e. while the user feedback is satisfied, it comes with a cost. The output of the algorithm might suggest high probabilities for terms that do not even occur in the document. For our example here with a simulated document corpus, such high probabilities for non-occurring terms in a document, need to be explained with reference to the terms occurring in the document. For example, Figure 11.5 compares the probabilities associated with each term in documents of P^*, before (green) and after (blue) the user feedback was incorporated in to LDA. Prior to incorporating feedback, the probabilities estimated for terms using LDA clearly agree with the actual presence (or absence) of terms in the documents. After incorporating feedback using our supervised LDA algorithm, we esti-

mate that ***ski, tourist, destination, winter*** has some mass for document $d_{22}(1 \ldots 8)$ so that it is presumably closer to document $d_4(5 \ldots 8)$. Similarly, the algorithm estimates positive probabilities for the terms ***chemical, factory, recently, hiring*** in document $d_{23}(1 \ldots 3)$ to bring it closer to document $d_{22}(1 \ldots 8)$, which indeed has the terms. In both cases, the documents actually do not have the terms, but are brought it to account for the feedback imposed by the user. To reconcile our aposteriori *term-document distribution* from INTERACTIVE STORYTELLING with the *empirical distribution of the terms* in the document, we define an overall measure of association of any occurring term with the non-occurring terms in the document and establish its relationship to proximity of documents.

The posterior joint distribution over a term w in dictionary $\boldsymbol{w}$ for topic z, $p(w|z = j, \mathfrak{R}, \mathfrak{D})$, affords us to obtain predictive conditional distributions of the form $p(w^*|w, \mathfrak{R}, \mathfrak{D})$ for the corpus (ignoring the predictive notation),

$$p(w^*|w) = \sum_{z=1}^{T} p(w^*|z = j)p(z = j|w), \text{ for some } w^*, w \in \boldsymbol{w} \tag{11.1}$$

or even document-specific predictive conditional distributions of the form $p(w^*|w \in d_i, \mathfrak{R}, \mathfrak{D})$ as,

$$p(w^*|w \in d_i) = \sum_{z=1}^{T} p(w^*|z = j)p(z = j|w \in d_i), \text{ for some } w^*, w \in d_i \tag{11.2}$$

These are defined as corpus word associations and document specific word associations respectively. We use the document specific word association to construct a metric that measures the relative importance of a term occurring in the document, with respect to the terms not occurring in the document. For any document d_i, let $\{w_1, \ldots, w_c\}$ be the terms occurring in the document, and $\{w_1^*, \ldots, w_m^*\}$ be the terms not occurring in the document. Again, ignoring the posterior predictive notation, for a specific document d_i,

$$p(w_1^*) = \sum_{k=1}^{c} p(w_1^*, w_k) = \sum_{k=1}^{c} p(w_1^*|w_k)p(w_k).$$

Hence, the total probability mass distributed over the terms of the dictionary for docu-

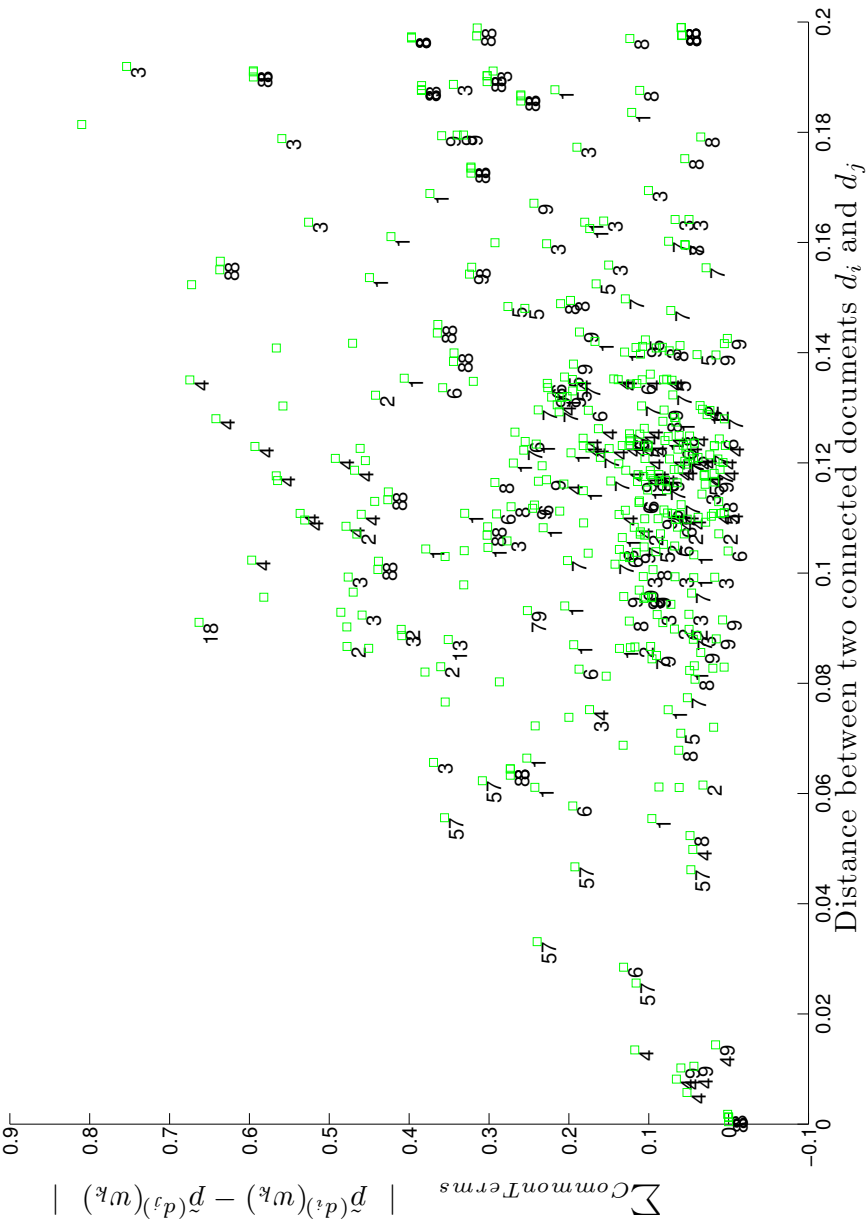
ment d_i, can be represented as,

$$\begin{aligned} 1 &= p(w_1) + \cdots + p(w_c) + p(w_1^*) + \cdots + p(w_m) \\ &= \sum_{k=1}^{c} (1 + p(w_1^*|w_k) + \cdots + p(w_m^*|w_k))\, p(w_k) \\ &= \sum_{k=1}^{c} (1 + \kappa_k) p(w_k) \\ &= \sum_{k=1}^{c} (p(w_k) + \tilde{p}(w_k)). \end{aligned} \tag{11.3}$$

We interpret κ_k as the *inflation factor* by which w_k is to be sampled from the dictionary, over its own sampling rate of $p(w_k)$, in order to replace the terms which do not occur in document d_i. A high value of κ_k implies that the terms not occurring in d_i have a high level of *association* with the term w_k. The term $\tilde{p}(w_k)$ measures the marginal probability of w_k being sampled, if pairs of terms of the form (w^*, w) are sampled for document d_i i.e. it is the probability of co-occurrence of w_k with all non-occurring terms in document d_i. A high value of $\tilde{p}(w_k)$ suggests that w_k occurs relatively often as a pair with a non-occurring term in document d_i. Hence $\tilde{p}(w_k)$ is also an *overall measure of association* of w_k with the non-occurring terms in the document. We normalize $\tilde{p}(w_k)$ by $\sum_{k=1}^{m} \tilde{p}(w_k)$ for the discussion that follows, and notate it by $\tilde{p}^{(d_i)}(w_k)$ for document d_i. For instance, in document $d_4(5 \ldots 8)$, the *overall measure of association* for terms corresponding to *Theme* 8 have increased after incorporating user feedback, while the *overall measure of association* for terms corresponding to *Theme* 5 have decreased. To the user, this suggests that terms corresponding to *Theme* 8 have a higher probability of co-occurrence with non-occurring terms, than terms in *Theme* 5 in document d_4.

It is important to establish the connection between the differences in *overall measures of association* for overlapping terms between two documents, with their distance in the supervised topic space. Consider an edge $e_{ij} =< d_i, d_j >$ in the graph induced by the similarity between the documents in the supervised topic space, and the cost of such an edge is c_{ij}. Let $w \in \boldsymbol{w}$ be the terms that overlap between d_i and d_j. In Figure 11.6, for all possible edges e_{ij} in the document graph, we plot the differences in *overall measures of association* between d_i and d_j given by $\sum_{w_k \in \boldsymbol{w}} \left| \tilde{p}^{(i)}(w_k) - \tilde{p}^{(j)}(w_k) \right|$, with c_{ij}. The plot ensures, that for document pairs which are close to each other, the *overall measure of association* for overlapping terms is a good indicator for understanding their dissimilarities. For documents which are distant from each other, the *overall measures of association* is not a reliable measure to understand their dissimilaries. This suggests, the proximity structure from our supervised latent dirichlet allocation can be explained with more certainty using the *overall measures of association* for shorter distances.

Figure 11.6: Plot of differing *overall measures of association* between documents d_i and d_j, $\sum_{w_k \in \boldsymbol{w}} \left| \tilde{p}^{(i)}(w_k) - \tilde{p}^{(j)}(w_k) \right|$, with the cost of the edge between d_i and d_j, c_{ij}, after incorporating user feedback using the INTERACTIVE STORYTELLING algorithm.



11.2 Atlantic Storm Dataset

11.2.1 Term Filtering

Our next example is based on the Atlantic Storm dataset. The dataset consists of $Q = 111$ documents and $M = 707$ unique terms. The terms were obtained after removing the top 10% terms based on *Gini index* [Dixon et al., 1987]. *Gini index* is based on the distribution of the term within the documents and measures the inequality of term occurrence in the documents. Consider an term w_i that occurs f_{ij} times in document d_j. The *Gini index* for w_i is given by,

$$GI_i = \frac{\sum_{j=1}^{Q} \sum_{k=1}^{Q} |f_{ij} - f_{ik}|}{2Q^2 \mu_i},$$

where μ_i is the average frequency for term w_i and given by,

$$\mu_i = \frac{\sum_{j=1}^{Q} f_{ij}}{Q}.$$

The *Gini index* ranges from zero, when w_i occurs equally frequently in all the documents $d_1, \ldots, d_Q$, to a theoretical maximum of one when none but one of the documents in a corpus of infinite size ($Q \to \infty$) has the term w_i (with any non-zero frequency). It is noteworthy that the frequency of term w_i in the lone document does not affect *Gini index*. The *Gini index* can be used to rank (and hence filter) terms. In the *Atlantic Storm* dataset, we remove 10% terms with the highest *Gini indices* i.e. terms that occur in all or almost all documents are removed. In typical intelligence analysis, its usually a very rare word occurring in a very few documents, which is of primary interest to an user or an intelligence analyst.

11.2.2 Understanding Topic Spaces Before and After Feedback

We fix the number of topics to be $T = 20$, and hyperparameters $\alpha = 0.05/T$ and $\beta = 0.01$. The hyperparameter specifications are as recommended by Steyvers and Griffiths [2007]. The number of topics were chosen by looking at clusters of documents after LDA and visualization using MDS. The initial view of the documents is based on topics obtained from LDA. The topics are visualized in Figure 11.10 and they show two distinct clusters with respect to topic similarity. We use the Manhattan distance between normalized term weight vectors of the two topics in our discussion here, to be consistent with the distance metric used for the remainder of the algorithm. Alternative measures of distance between topics including Kullback-Leibler divergence can also be used. An accompanying visualization perhaps also visualizes the documents in a screen as in Fig-

ure 11.7. The user specifies documents *CIA06* and *NSA16* as the starting and ending documents respectively, subsequent to which the Storytelling algorithm provides the *story* $CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA16$. The user is unsatisfied with the initial story and specifies that documents *CIA08* and *NSA09* should be in the *story*, with *NSA09* coming after *CIA08*. The INTERACTIVE STORYTELLING algorithm provides with new topic definitions after constraining the user defined path to be smaller than paths from an assortment of complete and incomplete alternate *stories* connecting documents $CIA06$ and $NSA16$ (listed in Table 11.5). The alternate paths are obtained from the $A^*Search$ with $CIA06$ and $NSA16$ as the *start* and *goal* nodes respectively in the document network. In the new topic space, the shortest path and hence the *story*, is given by $P^* \equiv CIA06 \rightarrow CIA08 \rightarrow DIA01 \rightarrow NSA09 \rightarrow NSA16$ (Figure 11.7). Indeed in includes documents $CIA08$ and $NSA09$ in the order specified by the user and brings in document $DIA01$ as another document which might be pertinent to the user's hypotheses.

Figure 11.7: Spatial visualization using Multidimensional Scaling of 111 documents before incorporating user feedback. The Green documents are the terminal documents (*start* and *goal* documents). The Blue documents with solid arrows is the initial *story*. The Orange documents are the documents that the user insists in the *story*. The shaded line shows the user defined alternative *story*.

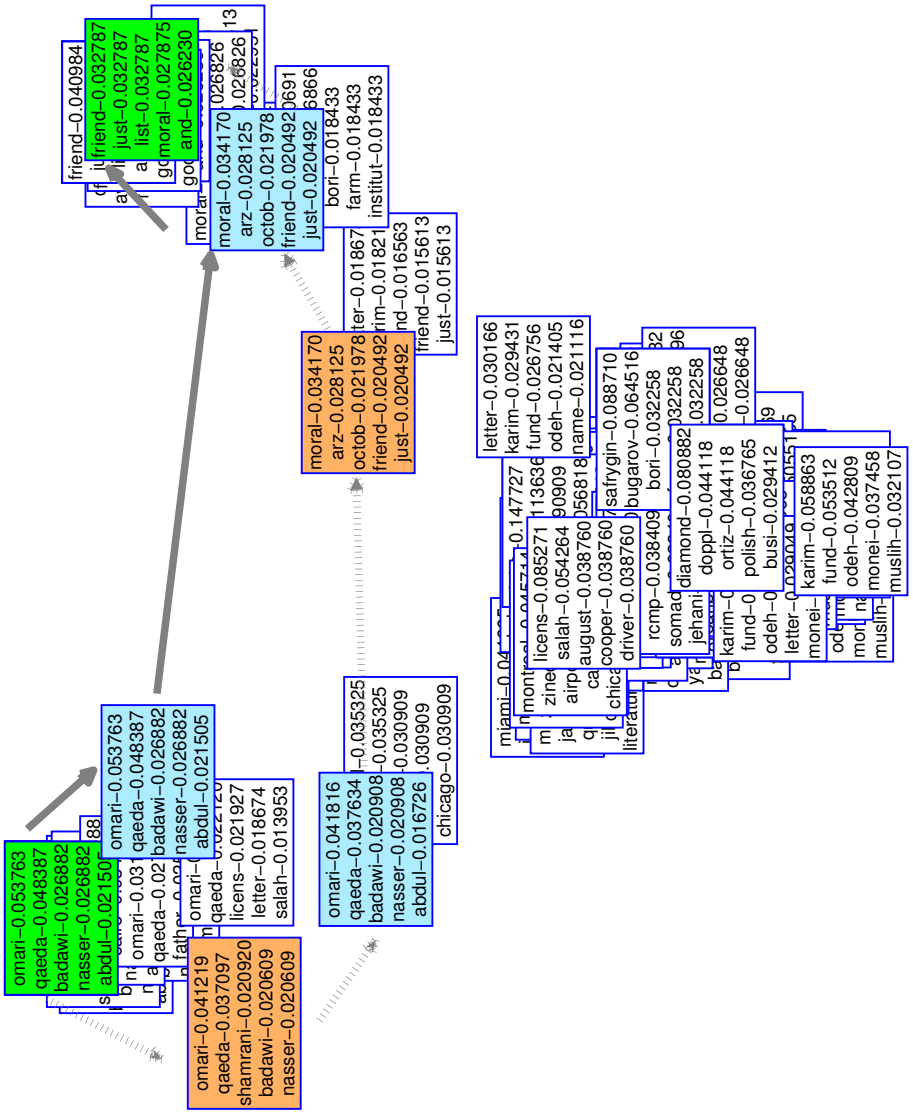
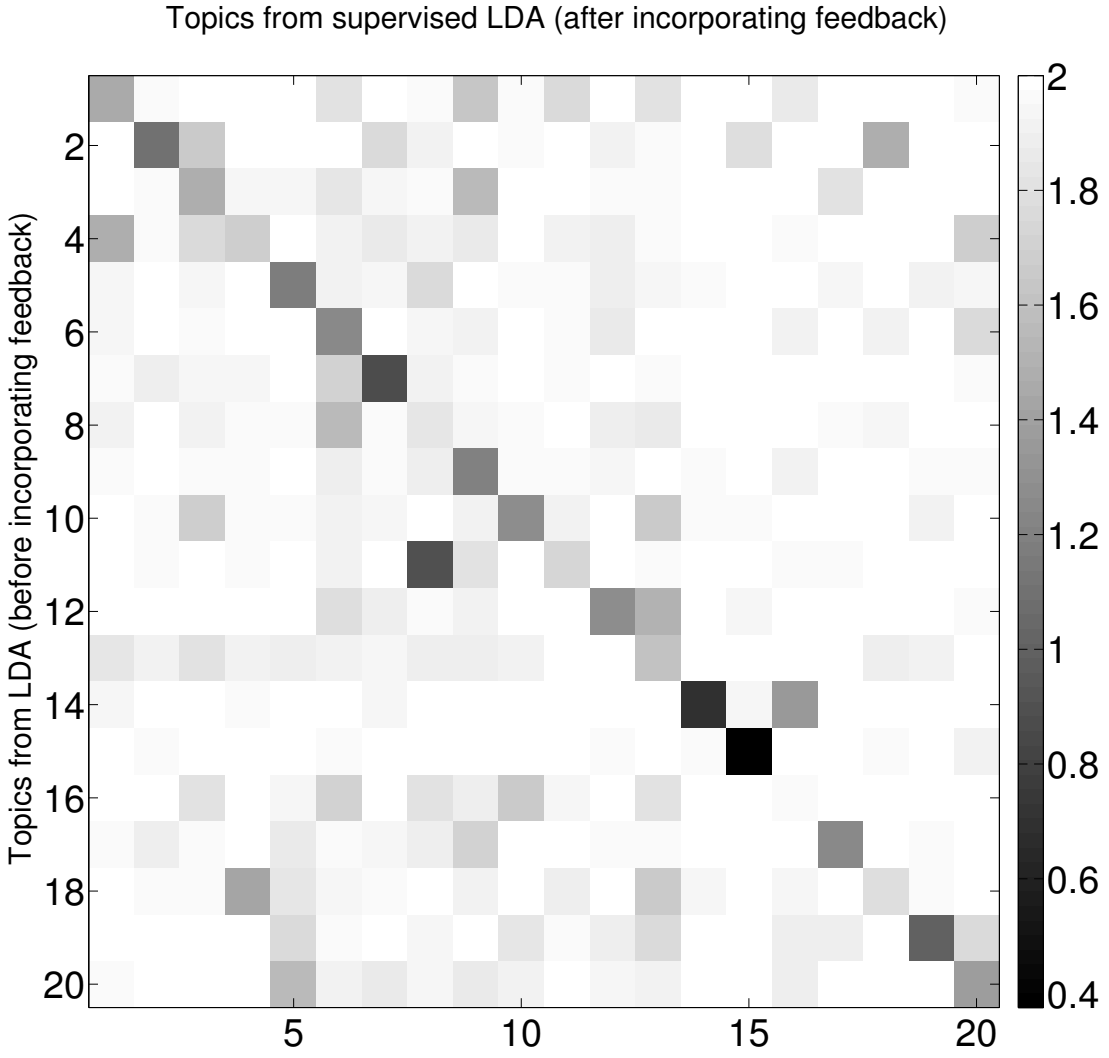


Figure 11.8: Visualization Manhattan distance between a topic from LDA prior to incorporating feedback (row) and a topic from supervised LDA after incorporating feedback (col). The darker the cell color, the closer are the topics. $T = 20$ in both cases.



The topic space under the supervised LDA in Interactive storytelling is given in Figure 11.9. Table 11.4 lists the top ten bag of terms from each topic based on probability weights, before(left) and after(right) incorporating the feedback; the LDA based topic on the left has been aligned with the closest supervised LDA based topic on the right, based on Manhattan distance. We visualize the Manhattan distance between all pairwise topics before and after incorporating feedback in Figure 11.8. Each LDA based topic has been aligned with the closest supervised LDA based topic along the diagonal. The graphic shows that the topic redefinitions after incorporating feedback are quite different from

the initial topics from LDA. We note that the farthest a topic can be from another topic is 2 distance units; it is a possibility when terms between two topics have no overlap.

11.2.3 Inference On Alternate *Stories* in Supervised Topic Space

The estimate of the difference in path lengths, $\hat{\mu}_o(\boldsymbol{\theta})$, between the user defined *story* P^*, and the alternate paths from $A^*Search$ provides us with a metric to rank competing *stories* in the current topic space. The competing *stories* and the estimates $\hat{\mu}_o(\boldsymbol{\theta})$, are tabulated in Table 11.5, and the bag of transitive terms in each alternate *story* is listed in Table 11.6. The list of competing *stories* include both complete and incomplete *stories* from the $A^*Search$. As in Example 11.1, we also calculate the actual length of the complete shortest path from $CIA06$ to $NSA16$ (using Yen's algorithm in Martins and Pascoal [2003]) via the documents specified in the alternate *stories* listed in Table 11.5. The most competitive *stories*, i.e. *stories* with path lengths slightly longer than P^*, have a value of $\hat{\mu}_o(\boldsymbol{\theta})$ very close to zero but negative. The *stories* with paths lengths much larger compared to P^*, have highly negative values of $\hat{\mu}_o(\boldsymbol{\theta})$. This is what we expect, and hence is a vindication of our methodology. For *stories* of medium path length, the relationship between its actual path length and $\hat{\mu}_o(\boldsymbol{\theta})$ is not completely monotonic. Figure 11.11 shows this relationship. For $\hat{\mu}_o(\boldsymbol{\theta})$ which are in the middle, inference should be done with caution. Hence the key point from this graphic is that extreme values (high or low) of $\hat{\mu}_o(\boldsymbol{\theta})$ are reasonably good for comparison of competing *stories* in the current topic space.

A second motivation of this graphic (Figure 11.11) is to understand the role of transitive terms in the progression of these alternate *stories*. For stories grouped by $\hat{\mu}_o(\boldsymbol{\theta})$, we superimposed the terms cloud of the transitive terms in the stories. The most competitive *stories* share ***insurg, hasam, badawi, farooq*** as the most frequently occurring transitive terms. The *stories* which are most inconsistent with the user feedback (based on highly negative values of $\hat{\mu}_o(\boldsymbol{\theta})$) share ***badawi, treat, octob*** as most frequently occurring transitive terms. *Stories* which are in between the extreme values for $\hat{\mu}_o(\boldsymbol{\theta})$ share ***khost, badawi, octob, treat, februari*** as most frequently occurring transitive terms. Hence our output of $\hat{\mu}_o(\boldsymbol{\theta})$ provides the user with a metric to rank and compare his preferred *story* P^* and allied hypotheses, with other competing *stories* and hypotheses.

11.2.4 Interpreting Similarity Between Documents in Supervised Topic Space

Figure 11.12 and 11.13 visualize the documents in a two-dimensional screen using Multidimensional Scaling before and after incorporating feedback respectively. A pair of documents close to each other on the screen can be assumed to have very similar estimated topic weights. However, the normalized vector of topic weights for a document after in-

corporating feedback might show disagreements with respect to terms actually occurring in the document, we will use the *overall measure of association* of terms in the document to understand similarities between documents. We show here that pairs of documents which are close to each other, share terms with high *overall measures of association*. This gives the user a method to understand why pairs of documents are closer to each other after incorporating the feedback. As shown also in Example 11.1, comparing terms with high *overall measures of association* for document pairs which are farther apart from each other, is not a dependable way to make interpretations about their dissimilarities after incorporating feedback.

Table 11.4: Topic definitions before and after feedback

Topic definition from LDA (Before feedback)	Topic definition from supervised LDA (After feedback)
friend,just,list,and,good, mirada,prepaid,acuna,don,fine	letter,ojinaga,calamar,and,ciudad mirada,send,strain,acuna,ago
licens,salah,august,cooper,driver, motel,bean,beandali,car,park	motel,regist,august,car,licens negra,owner,park,piedra,present
shamrani,militia,houston,attend,group, kansa,unit,aryan,casino,explos	odeh,literatur,muslih,convers,embassi kansa,arabia,bean,beandali,casino
letter,name,ojinaga,calamar,eln, went,arlington,chetum,guan,member	diseas,eln,foot,guan,mail, mouth,cartagena,control,farc,type
apart,mosqu,othman,yasser,baker, british,dahdah,heathrow,period,riyad	sufaat,cairo,mosqu,year,british, othman,queri,stai,appl,attend
omari,qaeda,badawi,nasser,abdul farooq,insurg,ahm,alias,chemic	nami,scada,omari,system,qaeda chicago,jamal,went,amsterdam,group
montreal,zinedin,airport,car,chicago, rafiki,french,mehdi,abu,haf	name,zinedin,car,rafiki,concern french,haf,mehdi,ticket,abu
nami,scada,system,usa,util, fenkel,access,contact,control,electr	diamond,apart,check,doppl,ortiz polish,rent,custom,english,gave
safrygin,bugarov,bori,farm,institut mark,come,frequent,recogn,sold	bugarov,shamrani,safrygin,moral,sizov, houston,militia,america,attend,good
arz,moral,air,arabia,bueno, embassi,argentina,offici,ambassador,appli	miami,area,bueno,argentina,arz dandani,avail,school,speak,air
diamond,doppl,ortiz,polish,busi, custom,tanzanit,transfer,form,islam	air,busi,best,chetum,friend, boat,especi,expert,iran,narcot
derwish,bafaba,rcmp,somad,jehani, abdul,amsterdam,bomb,angel,canada	bafaba,somad,usa,canadian,june, angel,culver,deposit,sungkar,written
karim,fund,odeh,monei,muslih, scholarship,donat,area,person,british	derwish,fund,karim,atmani,orang, bomb,rcmp,monei,scholarship,util
doha,ayyash,insur,central,claim, compani,island,farmer,plum,websit	doha,ayyash,insur,compani,file, central,claim,farmer,auto,mid

Continued on next page

Table 11.4 – continued from previous page

Topic definition from LDA (Before feedback)	Topic definition from supervised LDA (After feedback)
blake,charlott,qasim,rifai,bail, bond,carolina,mustafa,bailout,clark	blake,charlott,qasim,cooper,rifai, bail,bond,licens,carolina,compani
miami,jamal,jihad,literatur,quso, raid,attempt,flight,kill,shibh	island,plum,secur,websit,egyptian long,post,refuge,central,circl
octob,bugarov,sizov,regist,check, activ,atlant,employ,involv,moroccan	octob,unit,rental,atlant,sahara, storag,clipper,hyderabad,majest,sand
atmani,diseas,bomb,file,foot, initi,mouth,mzoudi,baltimor,icna	salah,fenkel,initi,mzoudi,person, icna,laurel,licens,possess,virginia
tour,book,sufaat,left,caribbean, baltimor,miami,accommod,cargo,nyc	tour,baltimor,book,left,accommod nyc,queen,caribbean,class,cruis
shakur,cairo,father,stai,univers kolokov,abdullah,attend,egyptian,loan	shakur,father,arlington,cargo,caribbean, nasser,abdul,abdullah,caraca,interview

Table 11.5: Complete and incomplete *stories* ranked by increasing value of estimated $\hat{\mu}_o(\boldsymbol{\theta})$. The closer (and negative) the value of $\hat{\mu}_o(\boldsymbol{\theta})$ to zero, the more consistent the *story* is to the user defined *story*.

Complete and incomplete *stories* connecting $CIA06$ to $NSA16$ from A*Search	$\hat{\mu}_o(\boldsymbol{\theta})$
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA40 \rightarrow NSA16$	-0.2766
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA28 \rightarrow NSA16$	-0.2751
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow FBI37 \rightarrow NSA16$	-0.2738
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow NSA06 \rightarrow NSA16$	-0.2641
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow NSA04 \rightarrow NSA16$	-0.2521
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow CIA03 \rightarrow NSA16$	-0.2458
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow FBI23 \rightarrow NSA16$	-0.2428
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA17 \rightarrow NSA16$	-0.2244
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA15 \rightarrow NSA16$	-0.2168
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow NSA21 \rightarrow NSA16$	-0.2155
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA08 \rightarrow NSA16$	-0.2122

Continued on next page

Table 11.5 – continued from previous page

Complete and incomplete *stories* connecting $CIA06$ to $NSA16$ from A*Search	$\hat{\mu}_o(\boldsymbol{\theta})$
$CIA06 \rightarrow CIA20 \rightarrow FBI02 \rightarrow NSA16$	-0.2014
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow CIA39 \rightarrow NSA16$	-0.2003
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA19 \rightarrow NSA16$	-0.1998
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA20 \rightarrow NSA16$	-0.1897
$CIA06 \rightarrow CIA20 \rightarrow FBI13 \rightarrow NSA16$	-0.1896
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA07 \rightarrow NSA16$	-0.1865
$CIA06 \rightarrow CIA20 \rightarrow FBI04 \rightarrow NSA16$	-0.1815
$CIA06 \rightarrow CIA33 \rightarrow FBI18 \rightarrow NSA16$	-0.1775
$CIA06 \rightarrow CIA02 \rightarrow FBI01 \rightarrow NSA16$	-0.1772
$CIA06 \rightarrow CIA20 \rightarrow CIA34 \rightarrow NSA16$	-0.1762
$CIA06 \rightarrow CIA33 \rightarrow CIA32 \rightarrow NSA16$	-0.1753
$CIA06 \rightarrow CIA20 \rightarrow FBI31 \rightarrow NSA16$	-0.1749
$CIA06 \rightarrow CIA33 \rightarrow CIA26 \rightarrow NSA16$	-0.1749
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA11 \rightarrow NSA16$	-0.173
$CIA06 \rightarrow CIA33 \rightarrow CIA31 \rightarrow NSA16$	-0.1726
$CIA06 \rightarrow CIA33 \rightarrow CIA27 \rightarrow NSA16$	-0.1719
$CIA06 \rightarrow CIA20 \rightarrow CIA21 \rightarrow NSA16$	-0.1676
$CIA06 \rightarrow CIA33 \rightarrow FBI35 \rightarrow NSA16$	-0.1649
$CIA06 \rightarrow CIA33 \rightarrow CIA42 \rightarrow NSA16$	-0.1633
$CIA06 \rightarrow CIA33 \rightarrow FBI39 \rightarrow NSA16$	-0.1622
$CIA06 \rightarrow CIA02 \rightarrow FBI22 \rightarrow NSA16$	-0.161
$CIA06 \rightarrow CIA20 \rightarrow NSA05 \rightarrow NSA16$	-0.1578
$CIA06 \rightarrow CIA33 \rightarrow FBI17 \rightarrow NSA16$	-0.1571
$CIA06 \rightarrow CIA02 \rightarrow FBI07 \rightarrow NSA16$	-0.1515
$CIA06 \rightarrow CIA33 \rightarrow FBI20 \rightarrow NSA16$	-0.1473
$CIA06 \rightarrow CIA20 \rightarrow NSA22 \rightarrow NSA16$	-0.1472
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA16$	-0.1461
$CIA06 \rightarrow CIA08 \rightarrow FBI40 \rightarrow NSA16$	-0.1299
$CIA06 \rightarrow CIA02 \rightarrow CIA18 \rightarrow NSA16$	-0.124
$CIA06 \rightarrow CIA08 \rightarrow CIA04 \rightarrow NSA16$	-0.123

Continued on next page

Table 11.5 – continued from previous page

Complete and incomplete *stories* connecting $CIA06$ to $NSA16$ from A*Search	$\hat{\mu}_o(\boldsymbol{\theta})$
$CIA06 \rightarrow CIA02 \rightarrow FBI32 \rightarrow NSA16$	-0.1219
$CIA06 \rightarrow CIA22 \rightarrow NSA16$	-0.1187
$CIA06 \rightarrow CIA02 \rightarrow FBI36 \rightarrow NSA16$	-0.1187
$CIA06 \rightarrow CIA02 \rightarrow CIA01 \rightarrow NSA16$	-0.1166
$CIA06 \rightarrow FBI15 \rightarrow NSA16$	-0.1139
$CIA06 \rightarrow CIA08 \rightarrow FBI33 \rightarrow NSA16$	-0.1115
$CIA06 \rightarrow FBI25 \rightarrow NSA16$	-0.1112
$CIA06 \rightarrow CIA24 \rightarrow NSA16$	-0.1103
$CIA06 \rightarrow FBI26 \rightarrow NSA16$	-0.1102
$CIA06 \rightarrow CIA02 \rightarrow CIA16 \rightarrow NSA16$	-0.1056
$CIA06 \rightarrow CIA02 \rightarrow NSA02 \rightarrow NSA16$	-0.104
$CIA06 \rightarrow FBI11 \rightarrow NSA16$	-0.0931
$CIA06 \rightarrow FBI19 \rightarrow NSA16$	-0.0916
$CIA06 \rightarrow CIA08 \rightarrow FBI21 \rightarrow NSA16$	-0.0891
$CIA06 \rightarrow CIA08 \rightarrow CIA23 \rightarrow NSA16$	-0.0884
$CIA06 \rightarrow CIA13 \rightarrow NSA16$	-0.0859
$CIA06 \rightarrow CIA12 \rightarrow NSA16$	-0.085
$CIA06 \rightarrow FBI28 \rightarrow NSA16$	-0.0848
$CIA06 \rightarrow FBI16 \rightarrow NSA16$	-0.0794
$CIA06 \rightarrow CIA08 \rightarrow FBI09 \rightarrow NSA16$	-0.0787
$CIA06 \rightarrow CIA08 \rightarrow FBI34 \rightarrow NSA16$	-0.0694
$CIA06 \rightarrow DIA02 \rightarrow NSA16$	-0.0575
$CIA06 \rightarrow FBI29 \rightarrow NSA16$	-0.054
$CIA06 \rightarrow DIA03 \rightarrow NSA16$	-0.0347
$CIA06 \rightarrow CIA07 \rightarrow NSA16$	-0.0304
$CIA06 \rightarrow CIA11 \rightarrow NSA16$	-0.0282
$CIA06 \rightarrow CIA08 \rightarrow FBI06 \rightarrow NSA16$	-0.0246
$CIA06 \rightarrow CIA37 \rightarrow NSA16$	-0.0181

Table 11.6: Complete and incomplete *stories* ranked by increasing value of estimated $\hat{\mu}_o(\boldsymbol{\theta})$, with corresponding transitive terms connecting the documents in the *story*. From top to bottom, transitive terms causing connections between documents in a *story* changes. *Stories* which are ranked higher and hence are least consistent with the user defined *story* are dominated by ***octob, badawi, treat*** as transitive words. *Stories* towards the bottom of the table, and hence more consistent with the user defined *story* are dominated by ***insurg, hasham, badawi, farooq*** as transitive words.

Complete and incomplete *stories* connecting $CIA06$ to $NSA16$ from A*Search	Transitive terms connecting documents in the *story*
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA40 \rightarrow NSA16$	badawi,fight,octob,treat
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA28 \rightarrow NSA16$	badawi,destin,octob,treat,usa
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow FBI37 \rightarrow NSA16$	badawi,member,octob,treat
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow NSA06 \rightarrow NSA16$	badawi,octob,treat,went
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow NSA04 \rightarrow NSA16$	abdul,badawi,octob,treat
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow CIA03 \rightarrow NSA16$	arz,avenu,badawi,octob,treat
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow FBI23 \rightarrow NSA16$	badawi,octob,treat,went
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA17 \rightarrow NSA16$	badawi,octob,shortli,treat
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA15 \rightarrow NSA16$	badawi,octob,qaeda,treat,went
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow NSA21 \rightarrow NSA16$	badawi,don,octob,suppos,talk,treat
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA08 \rightarrow NSA16$	badawi,call,friend,just,list, morocco,octob,orang,treat
$CIA06 \rightarrow CIA20 \rightarrow FBI02 \rightarrow NSA16$	badawi,outsid,treat
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow CIA39 \rightarrow NSA16$	badawi,octob,orang,treat
$CIA06 \rightarrow CIA20 \rightarrow DIA01 \rightarrow CIA19 \rightarrow NSA16$	badawi,octob,qaeda,treat
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA20 \rightarrow NSA16$	badawi,list,octob,treat
$CIA06 \rightarrow CIA20 \rightarrow FBI13 \rightarrow NSA16$	arm,badawi,treat
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA07 \rightarrow NSA16$	badawi,octob,shipment,treat
$CIA06 \rightarrow CIA20 \rightarrow FBI04 \rightarrow NSA16$	badawi,paid,treat

Continued on next page

Table 11.6 – continued from previous page

Complete and incomplete *stories* connecting $CIA06$ to $NSA16$ from A*Search	Transitive terms connecting documents in the *story*
$CIA06 \rightarrow CIA33 \rightarrow FBI18 \rightarrow NSA16$	area,dia,khost
$CIA06 \rightarrow CIA02 \rightarrow FBI01 \rightarrow NSA16$	fahd,montreal
$CIA06 \rightarrow CIA20 \rightarrow CIA34 \rightarrow NSA16$	badawi,told,treat
$CIA06 \rightarrow CIA33 \rightarrow CIA32 \rightarrow NSA16$	khost,regard
$CIA06 \rightarrow CIA20 \rightarrow FBI31 \rightarrow NSA16$	badawi,octob,treat
$CIA06 \rightarrow CIA33 \rightarrow CIA26 \rightarrow NSA16$	februari,khost
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA11 \rightarrow NSA16$	badawi,moral,octob,shipment,treat
$CIA06 \rightarrow CIA33 \rightarrow CIA31 \rightarrow NSA16$	khost,program
$CIA06 \rightarrow CIA33 \rightarrow CIA27 \rightarrow NSA16$	februari,khost,sufaat
$CIA06 \rightarrow CIA20 \rightarrow CIA21 \rightarrow NSA16$	badawi,octob,san,treat
$CIA06 \rightarrow CIA33 \rightarrow FBI35 \rightarrow NSA16$	khost,student
$CIA06 \rightarrow CIA33 \rightarrow CIA42 \rightarrow NSA16$	call,khost,translat
$CIA06 \rightarrow CIA33 \rightarrow FBI39 \rightarrow NSA16$	hard,khost
$CIA06 \rightarrow CIA02 \rightarrow FBI22 \rightarrow NSA16$	beandali,fahd,raeed
$CIA06 \rightarrow CIA20 \rightarrow NSA05 \rightarrow NSA16$	badawi,octob,treat
$CIA06 \rightarrow CIA33 \rightarrow FBI17 \rightarrow NSA16$	khost,regard
$CIA06 \rightarrow CIA02 \rightarrow FBI07 \rightarrow NSA16$	destin,fahd,toronto
$CIA06 \rightarrow CIA33 \rightarrow FBI20 \rightarrow NSA16$	khost,student
$CIA06 \rightarrow CIA20 \rightarrow NSA22 \rightarrow NSA16$	badawi,guess,hello,just,stuff,suppos told,treat
$CIA06 \rightarrow CIA20 \rightarrow NSA09 \rightarrow NSA16$	badawi,list,martin,octob,san,treat
$CIA06 \rightarrow CIA08 \rightarrow FBI40 \rightarrow NSA16$	badawi,farooq,hasham,insurg,second
$CIA06 \rightarrow CIA02 \rightarrow CIA18 \rightarrow NSA16$	citizen,fahd
$CIA06 \rightarrow CIA08 \rightarrow CIA04 \rightarrow NSA16$	badawi,border,farooq,hasham,insurg

Continued on next page

Table 11.6 – continued from previous page

Complete and incomplete *stories* connecting $CIA06$ to $NSA16$ from A*Search	Transitive terms connecting documents in the *story*
$CIA06 \rightarrow CIA02 \rightarrow FBI32 \rightarrow NSA16$	fahd,home
$CIA06 \rightarrow CIA22 \rightarrow NSA16$	salman,yasir
$CIA06 \rightarrow CIA02 \rightarrow FBI36 \rightarrow NSA16$	fahd,heard,home
$CIA06 \rightarrow CIA02 \rightarrow CIA01 \rightarrow NSA16$	fahd,near
$CIA06 \rightarrow FBI15 \rightarrow NSA16$	action
$CIA06 \rightarrow CIA08 \rightarrow FBI33 \rightarrow NSA16$	badawi,facil,farooq,hasham,insurg,rememb
$CIA06 \rightarrow FBI25 \rightarrow NSA16$	action
$CIA06 \rightarrow CIA24 \rightarrow NSA16$	egypt,hasham
$CIA06 \rightarrow FBI26 \rightarrow NSA16$	document
$CIA06 \rightarrow CIA02 \rightarrow CIA16 \rightarrow NSA16$	canadian,fahd,get
$CIA06 \rightarrow CIA02 \rightarrow NSA02 \rightarrow NSA16$	fahd,montreal
$CIA06 \rightarrow FBI11 \rightarrow NSA16$	degre
$CIA06 \rightarrow FBI19 \rightarrow NSA16$	milit
$CIA06 \rightarrow CIA08 \rightarrow FBI21 \rightarrow NSA16$	badawi,facil,farooq,hasham,insurg
$CIA06 \rightarrow CIA08 \rightarrow CIA23 \rightarrow NSA16$	badawi,farooq,hasham,insurg, martin moral,san,show
$CIA06 \rightarrow CIA13 \rightarrow NSA16$	plan,special
$CIA06 \rightarrow CIA12 \rightarrow NSA16$	action,success
$CIA06 \rightarrow FBI28 \rightarrow NSA16$	milit
$CIA06 \rightarrow FBI16 \rightarrow NSA16$	egypt
$CIA06 \rightarrow CIA08 \rightarrow FBI09 \rightarrow NSA16$	badawi,farooq,group,hasham,insurg
$CIA06 \rightarrow CIA08 \rightarrow FBI34 \rightarrow NSA16$	badawi,farooq,hasham,insurg,shown
$CIA06 \rightarrow DIA02 \rightarrow NSA16$	hasham,insurg,khost,plan,staff

Continued on next page

Table 11.6 – continued from previous page

Complete and incomplete *stories* connecting $CIA06$ to $NSA16$ from A*Search	Transitive terms connecting documents in the *story*
$CIA06 \rightarrow FBI29 \rightarrow NSA16$	special
$CIA06 \rightarrow DIA03 \rightarrow NSA16$	degre
$CIA06 \rightarrow CIA07 \rightarrow NSA16$	pakhtia,provinc
$CIA06 \rightarrow CIA11 \rightarrow NSA16$	badawi,salman,yasir
$CIA06 \rightarrow CIA08 \rightarrow FBI06 \rightarrow NSA16$	badawi,farooq,hasham,insurg,presenc
$CIA06 \rightarrow CIA37 \rightarrow NSA16$	fahd,insurg

Figure 11.9: Spatial visualization using Multidimensional Scaling, of 20 topics from supervised Latent Dirichlet Allocation after incorporating user feedback. Distance between two topics is the Manhattan distance between the two weighted topic vectors. For each topic, the first five terms with the highest weights are given.

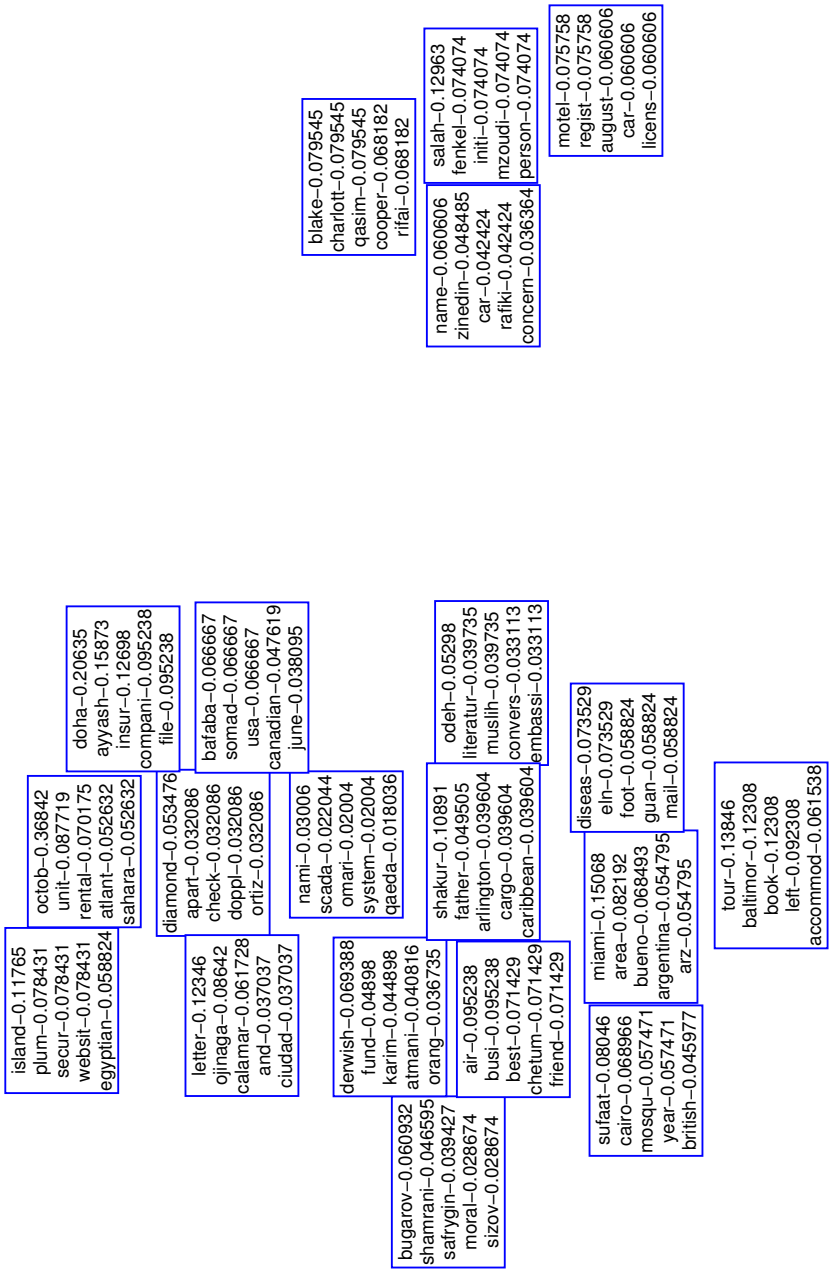


Figure 11.10: Spatial visualization using Multidimensional Scaling, of 20 topics from Latent Dirichlet Allocation before user feedback. Distance between two topics is the Manhattan distance between the two weighted topic vectors. For each topic, the first five terms with the highest weights are given.

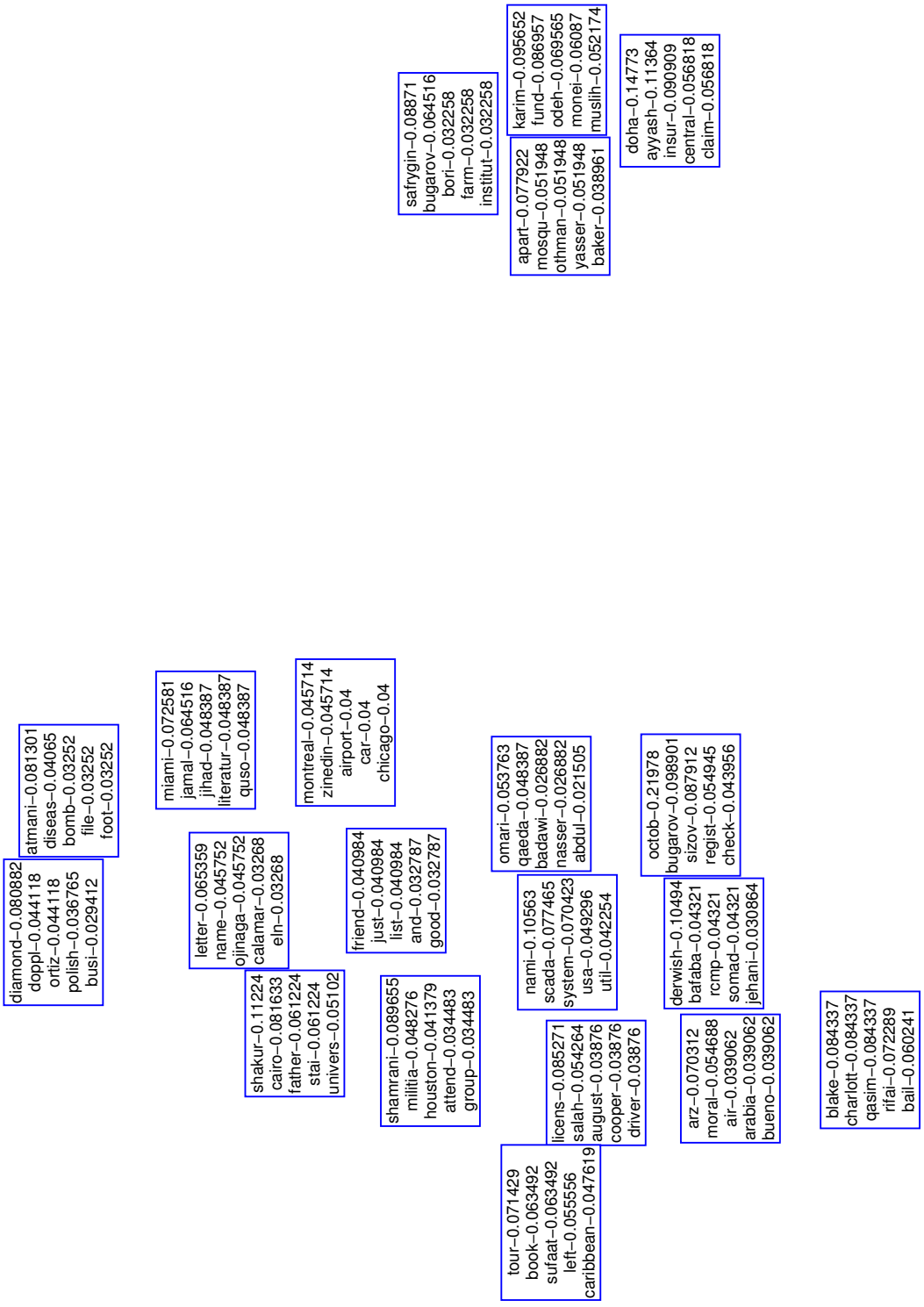


Figure 11.11: Each data point corresponds to a relationship $c(P^*) \leq c(P^{(o)})$. X Axis: Estimated value of $\mu_o(\boldsymbol{\theta})$ for a relationship. Y Axis: True length of *story*, $P^{(o)}$. The more negative the estimate of $\mu_o(\boldsymbol{\theta})$, the longer is the length of the *story* compared to the user defined *story*, and hence perhaps less consistent with user feedback. For clusters of relationships denoted by arrow, the word cloud of transitive terms causing document connections is provided. Top-left word cloud corresponds to transitive terms in *stories* which are least consistent with user defined *story*. Bottom-right word cloud corresponds to transitive terms in *stories* which are closest with user defined *story*.

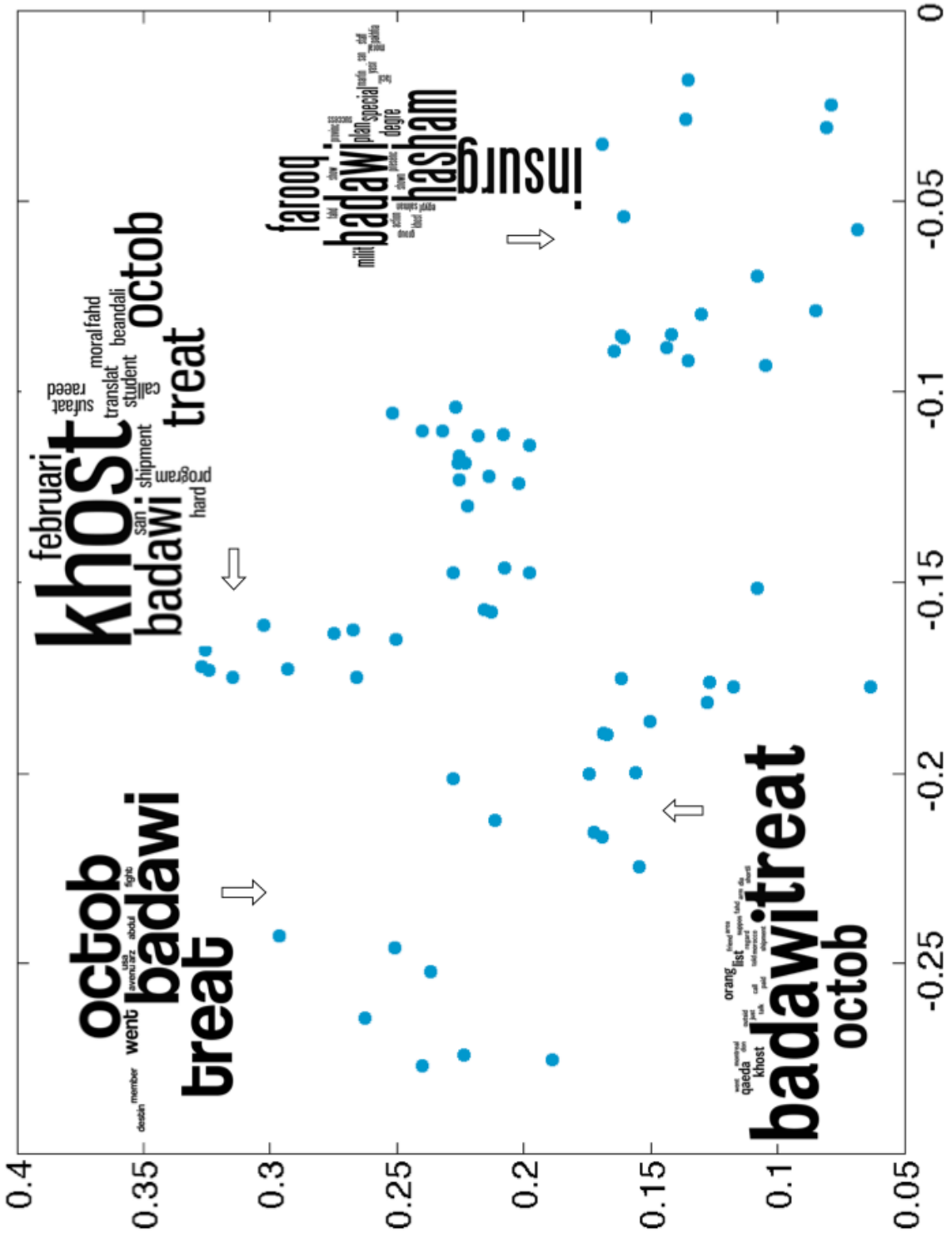


Figure 11.12: Spatial visualization using Multidimensional Scaling of 111 documents before incorporating user feedback. Distance between two documents is the Manhattan distance between the normalized topic weight vectors. Each document is represented by the top five terms with largest *overall measure of association*. Pairs of documents close to each other (colored pairs in the graphic) share terms with high *overall measure of association*.

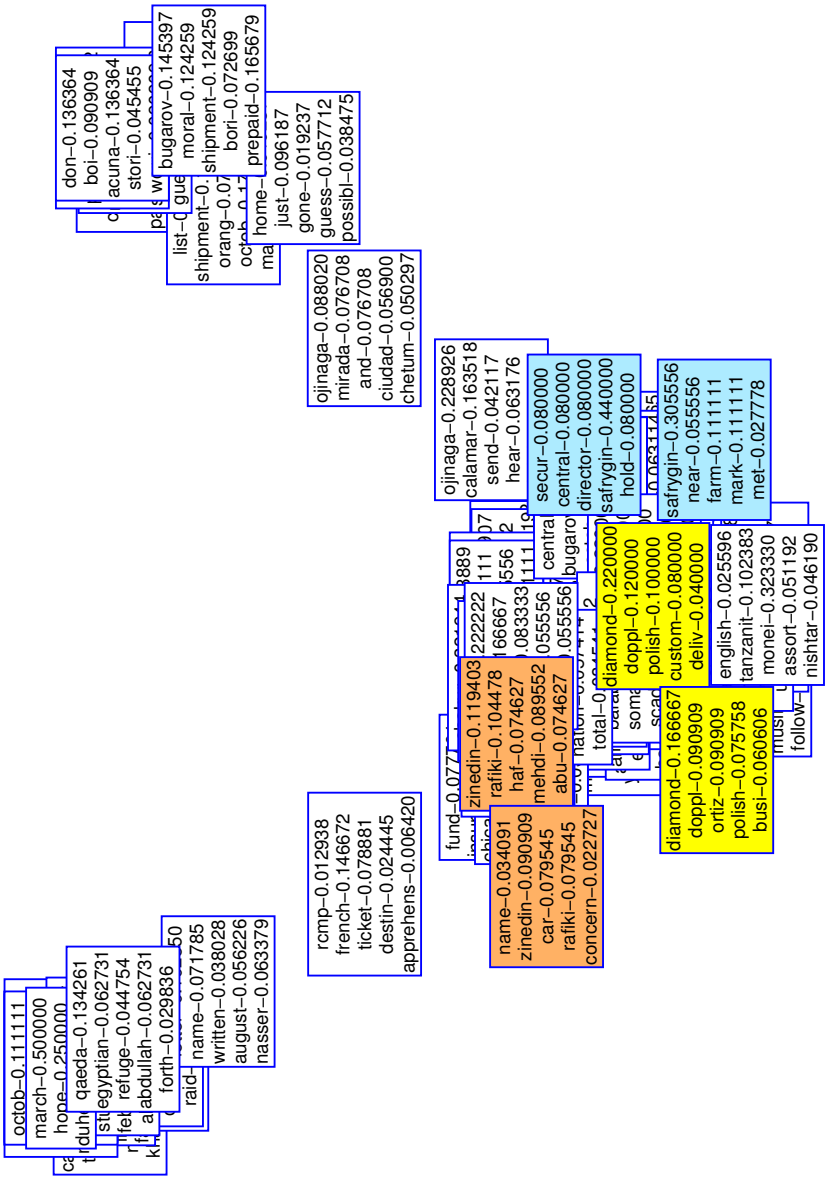
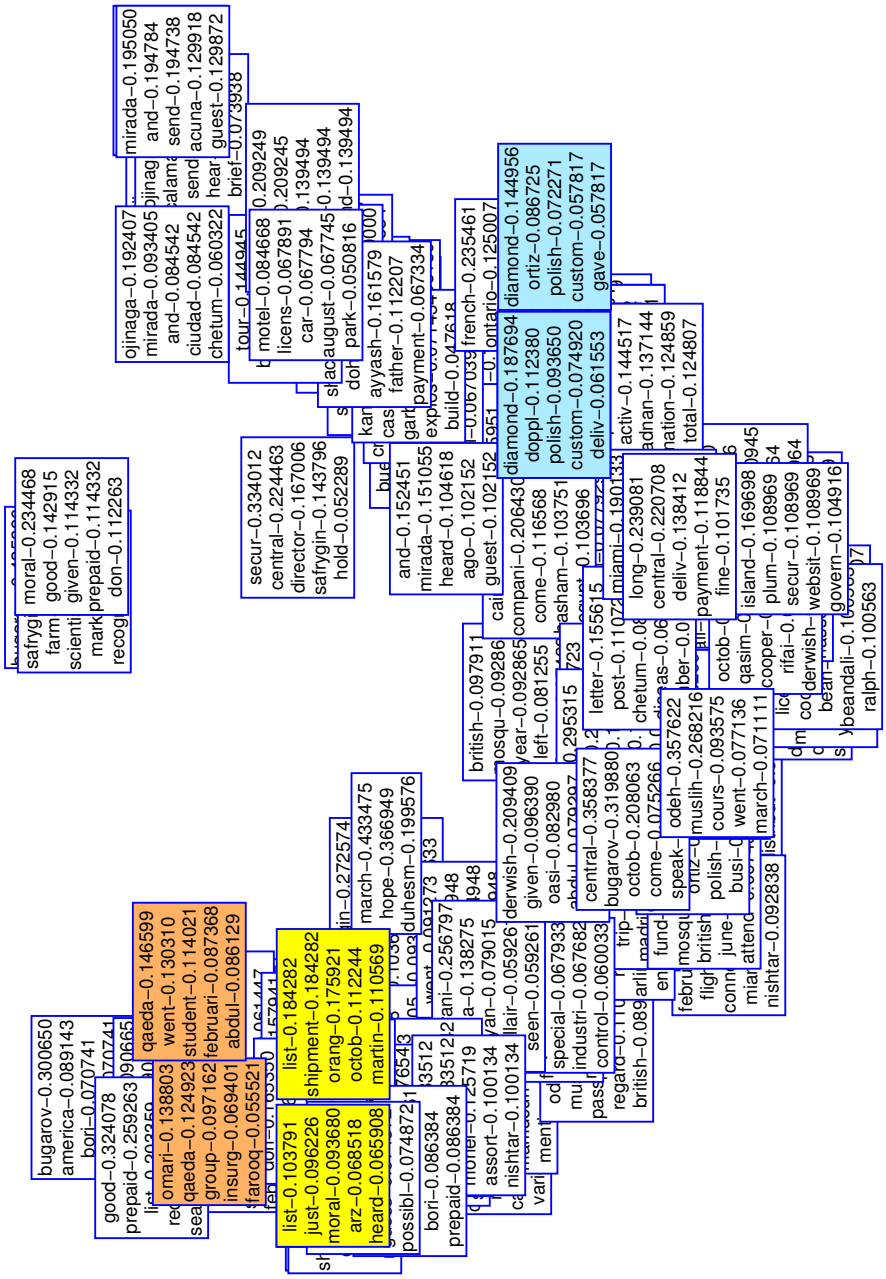


Figure 11.13: Spatial visualization using Multidimensional Scaling of 111 documents after incorporating user feedback. Distance between two documents is the Manhattan distance between the normalized topic weight vectors. Each document is represented by the top five terms with largest *overall measure of association*. Pairs of documents close to each other (colored pairs in the graphic) share terms with high *overall measure of association*.



Chapter 12

Discussion and Future Work

Documents are represented in the simplex topic space as normalized vector of topic weights. A Manhattan distance metric induces a proximity structure between the documents, while links between documents are defined by overlapping terms. The resultant document network or graph is used to connect a *start* and *goal* document via a series of connected documents as a *story*. Connections are made such that the path is the shortest path between *start* and *goal*. In the event that the user disagrees with the initial *story*, INTERACTIVE STORYTELLING incorporates feedback from users in a *V2PI* framework thus shielding the user away from the technical details of the underlying supervised topic model. The user provides feedback in terms of a sequence of documents that he wants in the *story* and appropriate documents are brought in to the *story* via a redefinition of the topic space.

A set of relationships in the form of path based inequalities are imputed to the user feedback. These inequalities are modeled probabilistically by truncated latent slack or surplus variables. The new topic space is obtained by satisfying the inequalities via a regularization on the difference between the cost of two competing *stories*. While the existence of a solution is not guaranteed via the framework, it does provide a probabilistic measure of comparing alternate possible paths as candidate *stories*. The algorithm also allows interpretation of the new topic weights with respect to terms that are in the document, using an *overall measure of association* of terms occurring in the document with terms not occurring in the document.

A key aspect of INTERACTIVE STORYTELLING is that other path or flow based constraints can be incorporated in the model if the constraints can be expressed as inequalities on path lengths or tolerances on edge costs from inverse combinatorial problems. However, any new path or flow based constraint has to be mapped to an appropriate user feedback. In a framework such as ours which depend on user interaction in a visual analytic platform, the performance has to be judged from the standpoint of user studies. Possible user interactions include highlighting a portion of text, or searching a specific term. Spatial

interactions include moving two or more documents closer to or farther from each other to signify their apparent similarity or dissimilarity based on underlying topics, specifying document(s) to be altogether ignored for path discoveries etc. A second aspect is that a more complicated hierarchical generative topic model would only involve defining a new distance metric on the documents based on the richer set of parameters in the hierarchy. In so far as our framework is concerned, the new definition of distance and the set of relationships will provide a formulation of the problem. Applications of our methodology in document networks limits its possibilities. Our algorithm provides the user with feature associations before and after the feedback in situations where predictions have to be made e.g. predictive word associations in topic modeling. Lastly, for large corpus or networks, our fully Bayesian approach has to be replaced with a variational Bayes approximation.

Bibliography

Data, data everywhere. *The Economist*, Feb, 2010.

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Society*, 88(422):669–679, 1993.

James H. Albert. Bayesian selection of log-linear models. *Canadian Journal of Statistics*, 24: 327–347, 1995.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 25–32, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553378. URL `http://doi.acm.org/10.1145/1553374.1553378`.

Mikhail Bilenko, Sugato Basu, and Raymond J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 11–, New York, NY, USA, 2004. ACM. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015360. URL `http://doi.acm.org/10.1145/1015330.1015360`.

David M. Blei and John D. Lafferty. Correlated topic models. In *NIPS*, 2005.

David M. Blei and Jon D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

S. P. Brooks, N. Friel, and R. King. Classical model selection via simulated annealing. *Journal Of The Royal Statistical Society Series B*, 65(2):503–520, 2003a.

Stephen P. Brooks, Paolo Giudici, and Gareth O. Roberts. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society. Series B*, 65(1):3–55, 2003b.

Andreas Buja, Deborah F Swayne, Michael L Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, 2008. doi: 10.1198/106186008X318440. URL http://amstat.tandfonline.com/doi/abs/10.1198/106186008X318440.

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In *Advances in Neural Information Processing Systems*, 2006. URL http://books.nips.cc/papers/files/nips19/NIPS2006_0305.pdf.

Chaitanya Chemudugunta, America Holloway, Padhraic Smyth, and Mark Steyvers. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *Proceedings of the 7th International Conference on The Semantic Web*, ISWC '08, pages 229–244, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-88563-4. doi: 10.1007/978-3-540-88564-1_15. URL http://dx.doi.org/10.1007/978-3-540-88564-1_15.

Siddhartha Chib. Marginal likelihood from the Gbibs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

Patricia Crossno, Andrew T. Wilson, Timothy M. Shead, Daniel M. Dunlavy, and Daniel M. Dunlavy. Topicview: Visually comparing topic models of text collections. In *ICTAI*, pages 936–943, 2011.

Corinne Dahinden, Giovanni Parmigiani, Mark C. Emerick, and Peter Bühlmann. Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics*, 8, 2007.

P. M. Dixon, J. Weiner, T. Mitchell-olds, and R. Woodley. Bootstrapping the Gini coefficient of inequality. *Ecology*, 68:1548–1551, 1987.

Alex Endert, Chao Han, Dipayan Maiti, Leanna House, Scotland D. Leman, and Chris North. Observation-level interaction with statistical models for visual analytics. Technical report, Blacksburg, VA, USA, 2011.

Edward I. George and Robert E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57:1317–1340, 1989.

John Geweke. Variable selection and model comparison in regression. In *Bayesian Statistics 5*, pages 609–620. University Press, 1996.

Simon J Godsill. On the relationship between Markov chain Mnote Caril methods for model uncertainty. *Journal Of Computational And Graphical Statistics*, 10(2):1–19, 2001.

Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

Peter J. Green. Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*. Oxford University Press, 2003.

Peter J. Green and Antonietta Mira. Delayed rejection in reversible jump metropolis-hastings. *Biometrika*, 88:1035–1053, 1999.

Peter Hart, Nils Nilsson, and Bertram Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, July 1968. ISSN 0536-1567. doi: 10.1109/TSSC.1968.300136. URL `http://dx.doi.org/10.1109/TSSC.1968.300136`.

David Hastie. Towards automatic reversible jump Markov chain Mnote Caril. unpublished doctoral thesis. *University of Bristol*, 26, 2005.

H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60, 1981.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial (with comments by Merlise Clyde, David Draper and Edward I. George, and a rejoinder by the authors. *Statistical Science*, 14(4):382–417, 1999.

S.C.H. Hoi, Wei Liu, M.R. Lyu, and Wei-Ying Ma. Learning distance metrics with contextual constraints for image retrieval. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2072 – 2078, 2006. doi: 10.1109/CVPR.2006.167.

Chris C. Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.

M. Shahriar Hossain, Michael Narayan, and Naren Ramakrishnan. Efficiently discovering hammock paths from induced similarity networks. *CoRR*, abs/1002.3195, 2010.

M. Shahriar Hossain, Christopher Andrews, Naren Ramakrishnan, and Chris North. Helping intelligence analysts make connections. 2011.

Yuening Hu, Jordan B. Graber, and Brianna Satinoff. Interactive topic modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 248–257, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL `http://portal.acm.org/citation.cfm?id=2002505`.

F. J. Huges. Discovery, proof, choice: The art and science of the process of intelligence analysis, case study 6. Technical report, 2005.

Ajay Jasra, David A. Stephens, and Chris C. Holmes. Population-based reversible jump Markov chain Mnote Caril. *Biometrika*, 94(4):787–807, 2007.

Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. ipca: An interactive system for pca-based visual analytics. volume 28, pages 767–774. Blackwell Publishing Ltd, 2009. doi: 10.1111/j.1467-8659.2009.01475.x. URL http://dx.doi.org/10.1111/j.1467-8659.2009.01475.x.

Robert E. Kass and Adrian E. Raftery. Bayes factors, 1995.

Daniel A. Keim, Florian Mansmann, Daniela Oelke, and Hartmut Ziegler. Visual analytics: Combining automated discovery with interactive visualizations. In *Proceedings of the 11th International Conference on Discovery Science*, DS '08, pages 2–14, 2008.

Deept Kumar, Naren Ramakrishnan, Richard F. Helm, and Malcolm Potts. Algorithms for storytelling. In *Proc. KDD'06*, pages 604–610, 2006.

Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhya: The Indian Journal of Statistics*, 60B(1):65–81, 1998.

Kyeong Eun Lee and et al. Gene selection: a Bayesian variable selection approach. *Bioinformatics*, 19:90–97, 2003.

Scotland C. Leman, Yuguo Chen, and Michael Lavine. The multiset sampler. *Journal of the American Statistical Association*, 104(487):1029–1041, 2009.

Scotland C. Leman, Leanna House, Dipayan Maiti, Alex Endert, and Chris North. A bidirectional visualization pipeline that enables visual to parametric interaction (v2pi). Technical report, Blacksburg, VA, USA, 2010.

Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 121–130, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367514. URL http://doi.acm.org/10.1145/1367497.1367514.

David Madigan and Adrian E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. Technical report, 1993.

J Manyika, M Chui, B Brown, J Bughin, R Dobbs, C Roxburgh, and A Hung Byers. Big data: The next frontier for innovation, competition, and productivity. *McKinsey Quarterly*, May, 2011.

Ernesto Q. V. Martins and Marta M. B. Pascoal. A new implementation of Yen's ranking loopless paths algorithm. *4OR: A Quarterly Journal of Operations Research*, 1:121–133, 2003. ISSN 1619-4500. URL http://dx.doi.org/10.1007/s10288-002-0010-2. 10.1007/s10288-002-0010-2.

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 101–110, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367512. URL `http://doi.acm.org/10.1145/1367497.1367512`.

George A. Miller. The magical number seven, plus or minus two some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.

M. A. Newton and Adrian E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56:3–48, 1994.

Terry Noreault, Michael McGill, and Matthew B. Koll. A performance evaluation of similarity measures, document term weighting schemes and representations in a boolean environment. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, SIGIR '80, pages 57–76, 1981.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://dl.acm.org/citation.cfm?id=1699510.1699543`.

Ramkumar Ramaswamy, James B. Orlin, and Nilopal Chakravarti. Sensitivity analysis for shortest path problems and maximum capacity path problems in undirected graphs. *Math. Program., Ser. A*, 102:355–369, 2005.

Sylvia Richardson and Peter J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B*, 59(4):731–792, 1997.

Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *UAI'04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*, pages 487–494, 2004.

Mark Steyvers and Tom Griffiths. *Probabilistic Topic Models*. Lawrence Erlbaum Associates, 2007. ISBN 1410615340. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/1410615340`.

L. Tierney, R. E. Kass, and J. B. Kadane. Fully exponential Laplace approximations to expectations and variances of non-positive functions. *Journal of the American Statistical Society*, 84:710–716, 1989.

Jean-Pierre Vila, Véréne Wagner, and Pascal Neveu. Bayesian nonlinear model selection and neural networks: a conjugate prior approach. *Neural Networks, IEEE Transactions on*, 11(2):265 –278, Mar 2000. ISSN 1045-9227. doi: 10.1109/72.838999.

Hanna M. Wallach. Topic modeling: beyond bag-of-words. In *NIPS 2005 Workshop on Bayesian Methods for Natural Language Processing*, 2005.

Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 153–162, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0055-1. doi: 10.1145/1835804.1835827. URL `http://doi.acm.org/10.1145/1835804.1835827`.

Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2002.