Historical Analysis of Legal Opinions with a Sparse Mixed-Effects Latent Variable Model

William Yang Wang[1] and **Elijah Mayfield**[1] and **Suresh Naidu**[2] and **Jeremiah Dittmar**[3]
[1]School of Computer Science, Carnegie Mellon University
[2]Department of Economics and SIPA, Columbia University
[3]American University and School of Social Science, Institute for Advanced Study
{ww,elijah}@cmu.edu sn2430@columbia.edu dittmar@american.edu

Abstract

We propose a latent variable model to enhance historical analysis of large corpora. This work extends prior work in topic modelling by incorporating metadata, and the interactions between the components in metadata, in a general way. To test this, we collect a corpus of slavery-related United States property law judgements sampled from the years 1730 to 1866. We study the language use in these legal cases, with a special focus on shifts in opinions on controversial topics across different regions. Because this is a longitudinal data set, we are also interested in understanding how these opinions change over the course of decades. We show that the joint learning scheme of our sparse mixed-effects model improves on other state-of-the-art generative and discriminative models on the region and time period identification tasks. Experiments show that our sparse mixed-effects model is more accurate quantitatively and qualitatively interesting, and that these improvements are robust across different parameter settings.

1 Introduction

Many scientific subjects, such as psychology, learning sciences, and biology, have adopted computational approaches to discover latent patterns in large scale datasets (Chen and Lombardi, 2010; Baker and Yacef, 2009). In contrast, the primary methods for historical research still rely on individual judgement and reading primary and secondary sources, which are time consuming and expensive. Furthermore, traditional human-based methods might have good precision when searching for relevant information, but suffer from low recall. Even when language technologies have been applied to historical problems, their focus has often been on information retrieval (Gotscharek et al., 2009), to improve accessibility of texts. Empirical methods for analysis and interpretation of these texts is therefore a burgeoning new field.

Court opinions form one of the most important parts of the legal domain, and can serve as an excellent resource to understand both legal and political history (Popkin, 2007). Historians often use court opinions as a primary source for constructing interpretations of the past. They not only report the proceedings of a court, but also express a judges' views toward the issues at hand in a case, and reflect the legal and political environment of the region and period. Since there exists many thousands of early court opinions, however, it is difficult for legal historians to manually analyze the documents case by case. Instead, historians often restrict themselves to discussing a relatively small subset of legal opinions that are considered decisive. While this approach has merit, new technologies should allow extraction of patterns from large samples of opinions.

Latent variable models, such as latent Dirichlet allocation (LDA) (Blei et al., 2003) and probabilistic latent semantic analysis (PLSA) (Hofmann, 1999), have been used in the past to facilitate social science research. However, they have numerous drawbacks, as many topics are uninterpretable, overwhelmed by uninformative words, or represent background language use that is unrelated to the dimensions of analysis that qualitative researchers are interested in.

SAGE (Eisenstein et al., 2011a), a recently proposed sparse additive generative model of language, addresses many of the drawbacks of LDA. SAGE assumes a background distribution of language use, and enforces sparsity in individual topics. Another advantage, from a social science perspective, is that SAGE can be derived from a standard logit random-utility model of judicial opinion writing, in contrast to LDA. In this work we extend SAGE to the supervised case of joint region and time period prediction. We formulate the resulting sparse mixed-effects (SME) model as being made up of mixed effects that not only contain random effects from sparse topics, but also mixed effects from available metadata. To do this we augment SAGE with two sparse latent variables that model the region and time of a document, as well as a third sparse latent

variable that captures the interactions among the region, time and topic latent variables. We also introduce a multiclass perceptron-style weight estimation method to model the contributions from different sparse latent variables to the word posterior probabilities in this predictive task. Importantly, the resulting distributions are still sparse and can therefore be qualitatively analyzed by experts with relatively little noise.

In the next two sections, we overview work related to qualitative social science analysis using latent variable models, and introduce our slavery-related early United States court opinion data. We describe our sparse mixed-effects model for joint modeling of region, time, and topic in section 4. Experiments are presented in section 5, with a robust analysis from qualitative and quantitative standpoints in section 5.2, and we discuss the conclusions of this work in section 6.

2 Related Work

Natural Language Processing (NLP) methods for automatically understanding and identifying key information in historical data have not yet been explored until recently. Related research efforts include using the LDA model for topic modeling in historical newspapers (Yang et al., 2011), a rule-based approach to extract verbs in historical Swedish texts (Pettersson and Nivre, 2011), a system for semantic tagging of historical Dutch archives (Cybulska and Vossen, 2011).

Despite our historical data domain, our approach is more relevant to text classification and topic modelling. Traditional discriminative methods, such as support vector machine (SVM) and logistic regression, have been very popular in various text categorization tasks (Joachims, 1998; Wang and McKeown, 2010) in the past decades. However, the main problem with these methods is that although they are accurate in classifying documents, they do not aim at helping us to understand the documents.

Another problem is lack of expressiveness. For example, SVM does not have latent variables to model the subtle differences and interactions of features from different domains (e.g. text, links, and date), but rather treats them as a "bag-of-features". Generative methods, by contrast, can show the causes to effects, have attracted attentions in recent years due to the rich expressiveness of the models and competitive performances in predictive tasks (Wang et al., 2011). For example, Nguyen et al. (2010) study the effect of the context of interaction in blogs using a standard LDA model. Guo and Diab (2011) show the effectiveness of using semantic information in multifaceted topic models for text categorization. Eisenstein et al. (2010) use a latent variable model to predict geolocation information of Twitter users, and investigate geographic variations of language use. Temporally, topic models have been used to show the shift in language use over time in online communities (Nguyen and Rosé, 2011) and the evolution of topics over time (Shubhankar et al., 2011).

When evaluating understandability, however, dense word distributions are a serious issue in many topic models as well as other predictive tasks. Such topic models are often dominated by function words and do not always effectively separate topics. Recent work have shown significant gains in both predictiveness and interpretatibility by enforcing sparsity, such as in the task of discovering sociolinguistic patterns of language use (Eisenstein et al., 2011b).

Our proposed sparse mixed-effects model balances the pros and cons the above methods, aiming at higher classification accuracies using the SME model for joint geographic and temporal aspects prediction, as well as richer interaction of components from metadata to enhance historical analysis in legal opinions. To the best of our knowledge, this study is the first of its kind to discover region and time specific topical patterns jointly in historical texts.

3 Data

We have collected a corpus of slavery-related United States supreme court legal opinions from Lexis Nexis. The dataset includes 5,240 slavery-related state supreme court cases from 24 states, during the period of 1730 - 1866. Optical character recognition (OCR) software was used by Lexis Nexis to digitize the original documents. In our region identification task, we wish to identify whether an opinion was written in a free state[1] (R1) or a slave state (R2)[2]. In our time identification experiment, we approximately divide the legal documents into four time quartiles (Q1, Q2, Q3, and Q4), and predict which quartile the testing document belongs to. Q1 contains cases from 1837 or earlier, where as Q2 is for 1838-1848, Q3 is for 1849-1855, and Q4 is for 1856 and later.

4 The Sparse Mixed-Effects Model

To address the over-parameterization, lack of expressiveness and robustness issues in LDA, the SAGE (Eisenstein et al., 2011a) framework draws a

[1]Including border states, this set includes CT, DE, IL, KY, MA, MD, ME, MI, NH, NJ, NY, OH, PA, and RI.

[2]These states include AR, AL, FL, GA, MS, NC, TN, TX, and VA.

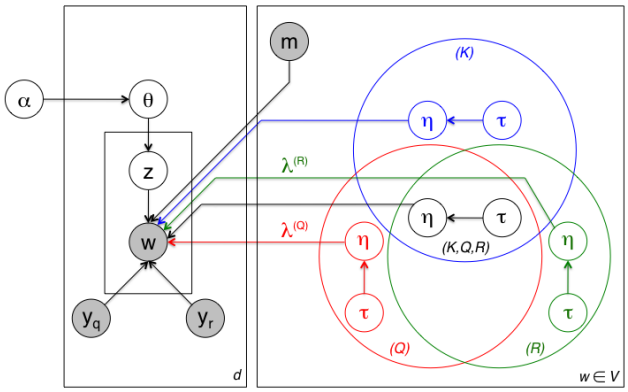

Figure 1: Plate diagram representation of the proposed Sparse Mixed-Effects model with K topics, Q time periods, and R regions.

constant background distribution m, and additively models the sparse deviation η from the background in log-frequency space. It also incorporates latent variables τ to model the variance for each sparse deviation η. By enforcing sparsity, the model might be less likely to overfit the training data, and requires estimation of fewer parameters.

This paper further extends SAGE to analyze multiple facets of a document collection, such as the regional and temporal differences. Figure 1 shows the graphical model of our proposed sparse mixed-effects (SME) model. In this SME model, we still have the same Dirichlet α, the latent topic proportion θ, and the latent topic variable z as the original LDA model. For each document d, we are able to observe two labels: the region label $y_d^{(R)}$ and the time quartile label $y_d^{(Q)}$. We also have a background distribution m that is drawn from a uninformative prior. The three major sparse deviation latent variables are $\eta_k^{(T)}$ for topics, $\eta_j^{(R)}$ for regions, and $\eta_q^{(Q)}$ for time periods. All of the three latent variables are conditioned on another three latent variables, which are their corresponding variances $\tau_k^{(T)}, \tau_j^{(R)}$ and $\tau_q^{(Q)}$. In the intersection of the plates for topics, regions, and time quartiles, we include another sparse latent variable $\eta_{qjk}^{(I)}$, which is conditioned on a variance $\tau_{qjk}^{(I)}$, to model the interactions among topic, region and time. $\eta_{qjk}^{(I)}$ is the linear combination of time period, region and topic sparse latent variables, which absorbs the residual variation that is not captured in the individual effects.

In contrast to traditional multinomial distribution of words in LDA models, we approximate the conditional word distribution in the document d as the exponentiated sum β of all latent sparse deviations $\eta_k^{(T)}, \eta_j^{(R)}, \eta_q^{(Q)}$, and $\eta_{qjk}^{(I)}$, as well as the background m:

$$
\begin{aligned}
&P(w_n^{(d)}|z_n^{(d)}, \eta, m, y_d^{(R)}, y_d^{(Q)}) \propto \beta \\
&= \exp\big(m + \eta_{z_n^{(d)}}^{(T)} + \lambda^{(R)}\eta_{y^{(r)}}^{(R)} \\
&\quad + \lambda^{(Q)}\eta_{y^{(q)}}^{(Q)} + \eta_{y^{(r)},y^{(q)},z_n^{(d)}}^{(I)}\big)
\end{aligned}
$$

Despite SME learns in a Bayesian framework, the above $\lambda^{(R)}$ and $\lambda^{(Q)}$ are dynamic parameters that weight the contributions of $\eta_{y^{(r)}}^{(R)}$ and $\eta_{y^{(q)}}^{(Q)}$ to the approximated word posterior probability. A zero-mean Laplace prior τ, which is conditioned on parameter γ, is introduced to induce sparsity, where its distribution is equivalent to the joint distribution, $\int \mathcal{N}(\eta; m, \tau)\varepsilon(\tau; \sigma)d\tau$, and $\varepsilon(\tau; \sigma)d\tau$ is the Exponential distribution (Lange and Sinsheimer, 1993). We first describe a generative story for this SME model:

- Draw a background m from corpus mean and initialize $\eta^{(T)}$, $\eta^{(R)}$, $\eta^{(Q)}$ and $\eta^{(I)}$ sparse deviations from corpus
- For each topic k
  - For each word i
    - Draw $\tau_{k,i}^{(T)} \sim \varepsilon(\gamma)$
    - Draw $\eta_{k,i}^{(T)} \sim \mathcal{N}(0, \tau_{k,i}^{(T)})$
  - Set $\beta_k \propto \exp(m+\eta_k+\lambda^{(R)}\eta^{(R)}+\lambda^{(Q)}\eta^{(Q)}+\eta^{(I)})$
- For each region j
  - For each word i
    - Draw $\tau_{j,i}^{(R)} \sim \varepsilon(\gamma)$
    - Draw $\eta_{j,i}^{(R)} \sim \mathcal{N}(0, \tau_{j,i}^{(R)})$
  - Update $\beta_j \propto \exp(m + \lambda^{(R)}\eta_j + \eta^{(T)} + \lambda^{(Q)}\eta^{(Q)} + \eta^{(I)})$
- For each time quartile q
  - For each word i
    - Draw $\tau_{q,i}^{(Q)} \sim \varepsilon(\gamma)$
    - Draw $\eta_{q,i}^{(Q)} \sim \mathcal{N}(0, \tau_{q,i}^{(Q)})$
  - Update $\beta_q \propto \exp(m + \lambda^{(Q)}\eta_q + \eta^{(T)} + \lambda^{(R)}\eta^{(R)} + \eta^{(I)})$
- For each time quartile q, for each region j, for each topic k
  - For each word i
    - Draw $\tau_{q,j,k,i}^{(I)} \sim \varepsilon(\gamma)$
    - Draw $\eta_{q,j,k,i}^{(I)} \sim \mathcal{N}(0, \tau_{q,j,k,i}^{(I)})$
  - Update $\beta_{q,j,k} \propto \exp(m + \eta_{q,j,k} + \eta^{(T)} + \lambda^{(R)}\eta^{(R)} + \lambda^{(Q)}\eta^{(Q)})$

- For each document d
  - Draw the region label $y_d^{(R)}$
  - Draw the time quartile label $y_d^{(Q)}$
  - For each word n, draw $w_n^{(d)} \sim \beta_{y_d}$

4.1 Parameter Estimation

We follow the MAP estimation method that Eisenstein et al. (2011a) used to train all sparse latent variables η, and perform Bayesian inference on other latent variables. The estimation of all variance variables τ remains as plugging the compound distribution of Normal-Jeffrey's prior, where the latter is a replacement of the Exponential prior. When performing Expectation-Maximization (EM) algorithm to infer the latent variables in SME, we derive the following likelihood function:

$$\begin{aligned}
\mathcal{L} =& \sum_d \langle \log P(\theta_d|\alpha)\rangle + \big\langle \log P(Z_n^{(d)}|\theta_d)\big\rangle \\
&+ \sum_n^{N_d} \big\langle \log P(w_n^{(d)}|z_n^{(d)}, \eta, m, y_d^{(R)}, y_d^{(Q)})\big\rangle \\
&+ \sum_k \langle \log P(\eta_k^{(T)}|0, \tau_k^{(T)})\rangle + \sum_k \langle \log P(\tau_k^{(T)}|\gamma)\rangle \\
&+ \sum_j \langle \log P(\eta_j^{(R)}|0, \tau_j^{(R)})\rangle + \sum_j \langle \log P(\tau_j^{(R)}|\gamma)\rangle \\
&+ \sum_q \langle \log P(\eta_q^{(Q)}|0, \tau_q^{Q)})\rangle + \sum_q \langle \log P(\tau_q^{(Q)}|\gamma)\rangle \\
&+ \sum_q \sum_j \sum_k \langle \log P(\eta_{q,j,k}^{(I)}|0, \tau_{q,j,k}^{(I)})\rangle \\
&+ \sum_q \sum_j \sum_k \langle \log P(\tau_{q,j,k}^{(I)}|\gamma)\rangle \\
&- \big\langle \log Q(\tau, z, \theta)\big\rangle
\end{aligned}$$

The above E step likelihood score can be intuitively interpreted as the sum of topic proportion scores, latent topic scores, the word scores, the η scores with their priors, and minus the joint variance. In the M step, when we use Newton's method to optimize the sparse deviation η_k parameter, we need to modify the original likelihood function in SAGE and its corresponding first and second order derivatives when deriving the gradient and Hessian matrix. The likelihood function for sparse topic deviation η_k is:

$$\begin{aligned}
\mathcal{L}(\eta_k) =& \langle c_k^{(T)}\rangle \mathsf{T} \eta_k \\
&- C_d \log \sum_q \sum_j \sum_i \exp(\lambda^{(Q)} \eta_{qi} + \lambda^{(R)} \eta_{ji} \\
&+ \eta_{ki} + \eta_{qjki} + m_i) - \eta_k \mathsf{T} diag(\langle (\tau_k^{(T)})^{-1}\rangle) \eta_k^{(T)}/2
\end{aligned}$$

and we can derive the gradient when taking the first order partial derivative:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \eta_k^{(T)}} =& \langle c_k^{(T)}\rangle - \sum_q \sum_j \langle C_{qjk}\rangle \beta_{qjk} \\
&- diag(\langle (\tau_k^{(T)})^{-1}\rangle) \eta_k^{(T)}
\end{aligned}$$

where $c_k^{(T)}$ is the true count, and β_{qjk} is the log word likelihood in the original likelihood function. C_{qjk} is the expected count from combinations of time, region and topic. $\sum_q \sum_j \langle C_{qjk}\rangle \beta_{qjk}$ will then be taken the second order derivative to form the Hessian matrix, instead of $\langle C_k\rangle \beta_k$ in the previous SAGE setting.

To learn the weight parameters $\lambda^{(R)}$ and $\lambda^{(Q)}$, we can approximate the weights using a multiclass perceptron-style (Collins, 2002) learning method. If we say that the notation of $\sum V^{(\bar{R})}$ is to marginalize out all other variables in β except $\eta^{(R)}$, and $P(y_d^{(R)})$ is the prior for the region prediction task, we can predict the expected region value $\hat{y}_d^{(R)}$ of a document d:

$$\begin{aligned}
\hat{y}_d^{(R)} &\propto \arg\max_{\hat{y}_d^{(R)}} \exp\Big(\sum V^{(\bar{R})} \log \beta + \log P(y_d^{(R)})\Big) \\
&= \arg\max_{\hat{y}_d^{(R)}} \Bigg(\exp\Big(\sum V^{(\bar{R})} \big(m + \eta_{z_n^{(d)}}^{(T)} + \lambda^{(R)} \eta_{y_d^{(R)}}^{(R)} \\
&\quad + \lambda^{(Q)} \eta_{y_d^{(Q)}}^{(Q)} + \eta_{y_d^{(R)}, y_d^{(Q)}, z_n^{(d)}}^{(I)}\big)\Big) P(y_d^{(R)})\Bigg)
\end{aligned}$$

If the symbol δ is the hyperprior for the learning rate and $\dot{y}_d^{(R)}$ is the true label, the update procedure for the weights becomes:

$$\lambda_d^{(R')} = \lambda_d^{(R)} + \delta(\dot{y}_d^{(R)} - \hat{y}_d^{(R)})$$

Similarly, we derive the $\lambda^{(Q)}$ parameter using the above formula. It is necessary to normalize the weights in each EM loop to preserve the sparsity property of latent variables. The weight update of $\lambda^{(R)}$ and $\lambda^{(Q)}$ is bound by the averaged accuracy of the two classification tasks in the training data, which is similar to the notion of minimizing empirical risk (Bahl et al., 1988). Our goal is to choose the two weight parameters that minimize the empirical classification error rate on training data when learning the word posterior probability.

5 Prediction Experiments

We perform three quantitative experiments to evaluate the predictive power of the sparse mixed-effects model. In these experiments, to predict the region and time period labels of a given document, we

jointly learn the two labels in the SME model, and choose the pair which maximizes the probability of the document.

In the first experiment, we compare the prediction accuracy of our SME model to a widely used discriminative learner in NLP – the linear kernel support vector machine (SVM)[3]. In the second experiment, in addition to the linear kernel SVM, we also compare our SME model to a state-of-the-art sparse generative model of text (Eisenstein et al., 2011a), and vary the size of input vocabulary W exponentially from 2^9 to the full size of our training vocabulary[4]. In the third experiment, we examine the robustness of our model by examining how the number of topics influences the prediction accuracy when varying the K from 10 to 50.

Our data consists of 4615 training documents and 625 held-out documents for testing. While individual judges wrote multiple opinions in our corpus, no judges overlapped between training and test sets. When measuring by the majority class in the testing condition, the chance baseline for the region identification task is 57.1% and the time identification task is 32.3%. We use three-fold cross-validation to infer the learning rate δ and cost C hyperpriors in the SME and SVM model respectively. We use the paired student t-test to measure the statistical significance.

5.1 Quantitative Results

5.1.1 Comparing SME to SVM

We show in this section the predictive power of our sparse mixed-effects model, comparing to a linear kernel SVM learner. To compare the two models in different settings, we first empirically set the number of topics K in our SME model to be 25, as this setting was shown to yield a promising result in a previous study (Eisenstein et al., 2011a) on sparse topic models. In terms of the size of vocabulary W for both the SME and SVM learner, we select three values to represent dense, medium or sparse feature spaces: $W_1 = 2^9$, $W_2 = 2^{12}$, and the full vocabulary size of $W_3 = 2^{13.8}$. Table 1 shows the accuracy of both models, as well as the relative improvement (gain) of SME over SVM.

When looking at the experiment results under different settings, we see that the SME model always outperforms the SVM learner. In the time quartile prediction task, the advantage of SME model is more salient. For example, with a medium density feature space of 2^{12}, SVM obtained an accuracy of 35.8%, but SME achieved an accuracy of 40.9%, which is a 14.2% relative improvement ($p < 0.001$) over SVM. When the feature space becomes sparser, the SME obtains an increased relative improvement ($p < 0.001$) of 16.1%, using full size of vocabulary. The performance of SVM in the binary region classification is stronger than in the previous task, but SME is able to outperform SVM in all three settings, with tightened advantages ($p < 0.05$ in W_2 and $p < 0.001$ in W_3). We hypothesize that it might because that SVM, as a strong large margin learner, is a more natural approach in a binary classification setting, but might not be the best choice in a four-way or multiclass classification task.

Method	Time	Gain	Region	Gain
SVM (W_1)	33.2%	–	69.7%	–
SME (W_1)	36.4%	9.6%	71.4%	2.4%
SVM (W_2)	35.8%	–	72.3%	–
SME (W_2)	40.9%	14.2%	74.0%	2.4%
SVM (W_3)	36.1%	–	73.5%	–
SME (W_3)	41.9%	16.1%	74.8%	1.8%

Table 1: Compare the accuracy of the linear kernel support vector machine to our sparse mixed-effects model in the region and time identification tasks ($K = 25$). *Gain: the relative improvement of SME over SVM.*

5.1.2 Comparing SME to SAGE

In this experiment, we compare SME with a state-of-the-art sparse generative model: SAGE (Eisenstein et al., 2011a).

Most studies on topic modelling have not been able to report results when using different sizes of vocabulary for training. Because of the importance of interpretability for social science research, the choice of vocabulary size is critical to ensure understandable topics. Thus we report our results at various vocabulary sizes W on SME and SAGE. To better validate the performance of SME, we also include the performance of SVM in this experiment, and fix the number of topics $K = 10$ for the SME and SAGE models, which is a different value for the number of topics K than the empirical K we used in the experiment of Section 5.1.1. Figure 2 and Figure 3 show the experiment results in both time and region classification task.

In Figure 2, we evaluate the impacts of W on our time quartile prediction task. The advantage of the SME model is very obvious throughout the experiments. Interestingly, when we continue to increase

[3]In our implementation, we use LibSVM (Chang and Lin, 2011).

[4]To select the vocabulary size W, we rank the vocabulary by word frequencies in a descending order, and pick the top-W words.

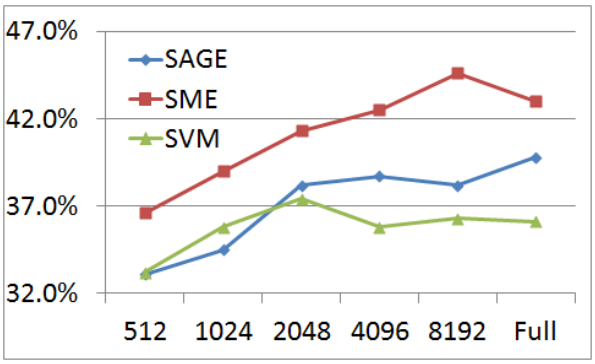

Figure 2: Accuracy on predicting the time quartile varying the vocabulary size W, while K is fixed to 10.

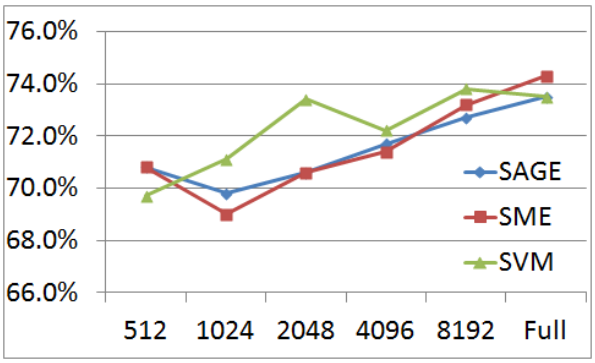

Figure 3: Accuracy on predicting the region varying the vocabulary size W, while K is fixed to 10.

the vocabulary size W exponentially and make the feature space more sparse, SME obtains its best result at $W = 2^{13}$, where the relative improvement over SAGE and SVM is 16.8% and 22.9% respectively ($p < 0.001$ under all comparisons).

Figure 3 shows the impacts of W on the accuracy of SAGE and SME in the region identification task. In this experiment, the results of SME model are in line with SAGE and SVM when the feature space is dense. However, when W reaches the full vocabulary size, we have observed significantly better results ($p < 0.001$ in the comparison to SAGE and $p < 0.05$ with SVM). We hypothesize that there might be two reasons: first, the K parameter is set to 10 in this experiment, which is much denser than the experiment setting in Section 5.1.1. Under this condition, the sparse topic advantage of SME might be less salient. Secondly, in the two tasks, it is observed that the accuracy of the binary region classification task is much higher than the four-way task, thus while the latter benefits significantly from this joint learning scheme of the SME model, but the former might not have the equivalent gain[5].

[5]We hypothesize that this problem might be eliminated if the two tasks in SME have similar difficulties and accuracies, but this needs to be verified in future work.

5.1.3 Influence of the number of topics K

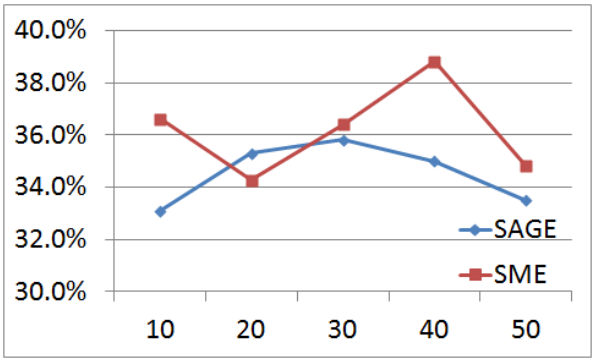

Figure 4: Accuracy on predicting the time quartile varying the number of topics K, while W is fixed to 2^9.

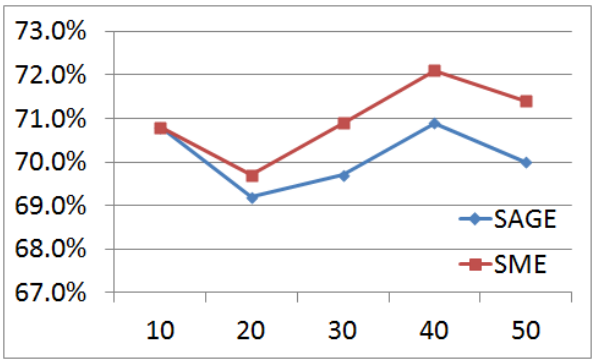

Figure 5: Accuracy on predicting the region varying the number of topics K, while W is fixed to 2^9.

Unlike hierarchical Dirichlet processes (Teh et al., 2006), in parametric Bayesian generative models, the number of topics K is often set manually, and can influence the model's accuracy significantly. In this experiment, we fix the input vocabulary W to 2^9, and compare the mixed-effect model with SAGE in both region and time identification tasks.

Figure 4 shows how the variations of K can influence the system performance in the time quartile prediction task. We can see that the sparse mixed-effects model (SME) reaches its best performance when the K is 40. After increasing the number of topics K, we can see SAGE consistently increase its accuracy, obtaining its best result when $K = 30$. When comparing these two models, SME's best performance outperforms SAGE's with an absolute improvement of 3%, which equals to a relative improvement ($p < 0.001$) of 8.4%. Figure 5 demonstrates the impacts of K on the predictive power of SME and SAGE in the region identification task.

	Keywords discovered by the SME model
Prior to 1837 (Q1)	pauperis, footprints, American Colonization Society, manumissions, 1797
1838 - 1848 (Q2)	indentured, borrowers, orphan's, 1841, vendee's, drawer's, copartners
1849 - 1855 (Q3)	Frankfort, negrotrader, 1851, Kentucky Assembly, marshaled, classed
After 1856 (Q4)	railroadco, statute, Alabama, steamboats, Waterman's, mulattoes, man-trap
Free Region (R1)	apprenticed, overseer's, Federal Army, manumitting, Illinois constitution
Slave Region (R2)	Alabama, Clay's Digest, oldest, cotton, reinstatement, sanction, plantation's
Topic 1 in Q1 R1	imported, comaker, runs, writ's, remainderman's, converters, runaway
Topic 1 in Q1 R2	comaker, imported, deceitful, huston, send, bright, remainderman's
Topic 2 in Q1 R1	descendent, younger, administrator's, documentary, agreeable, emancipated
Topic 2 in Q1 R2	younger, administrator's, grandmother's, plaintiffs, emancipated, learnedly
Topic 3 in Q2 R1	heir-at-law, reconsidered, manumissions, birthplace, mon, mother-in-law
Topic 3 in Q2 R2	heir-at-law, reconsideration, mon, confessions, birthplace, father-in-law's
Topic 4 in Q2 R1	indentured, apprenticed, deputy collector, stepfather's, traded, seizes
Topic 4 in Q2 R2	deputy collector, seizes, traded, hiring, stepfather's, indentured, teaching
Topic 5 in Q4 R1	constitutionality, constitutional, unconstitutionally, Federal Army, violated
Topic 5 in Q4 R2	petition, convictions, criminal court, murdered, constitutionality, man-trap

Table 2: A partial listing of an example for early United States state supreme court opinion keywords generated from the time quartile $\eta^{(Q)}$, region $\eta^{(R)}$ and topic-region-time $\eta^{(I)}$ interactive variables in the sparse mixed-effects model.

Except that the two models tie up when $K = 10$, SME outperforms SAGE for all subsequent variations of K. Similar to the region task, SME achieves the best result when K is sparser ($p < 0.01$ when $K = 40$ and $K = 50$).

5.2 Qualitative Analysis

In this section, we qualitatively evaluate the topics generated vis-a-vis the secondary literature on the legal and political history of slavery in the United States. The effectiveness of SME could depend not just on its predictive power, but also in its ability to generate topics that will be useful to historians of the period. Supreme court opinions on slavery are of significant interest for American political history. The conflict over slave property rights was at the heart of the "cold war" (Wright, 2006) between North and South leading up to the U.S. Civil War. The historical importance of this conflict between Northern and Southern legal institutions is one of the motivations for choosing our data domain.

We conduct qualitative analyses on the top-ranked keywords[6] that are associated with different geographical locations and different temporal frames, generated by our SME model. In our analysis, for each interaction of topic, region, and time period, a list of the most salient vocabulary words was generated. These words were then analyzed in the context of existing historical literature on the shift in attitudes and views over time and across regions. Table 2 shows an example of relevant keywords and topics.

This difference between Northern and Southern opinion can be seen in some of the topics generated by the SME. Topic 1 deals with transfers of human beings as slave property. The keyword "remainderman" designates a person who inherits or is entitled to inherit property upon the termination of an estate, typically after the death of a property owner, and appears in Northern and Southern cases. However, in Topic 1 "runaway" appears as a keyword in decisions from free states but not in decisions from slave states. The fact that "runaway" is not a top word in the same topic in the Southern legal opinions is consistent with a spatial (geolocational) division in which the property claims of slave owners over runaways were not heavily contested in Southern courts.

Topic 3 concerns bequests, as indicated by the term "heir-at-law", but again the term "manumissions", ceases to show up in the slave states after the first time quartile, perhaps reflecting the hostility to

[6]Keywords were ranked by word posterior probabilities.

manumissions that southern courts exhibited as the conflict over slavery deepened.

Topic 4 concerns indentures and apprentices. Interestingly, the terms indentures and apprenticeships are more prominent in the non-slave states, reflecting the fact that apprenticeships and indentures were used in many border states as a substitute for slavery, and these were often governed by continued usage of Master and Servant law (Orren, 1992).

Topic 5 shows the constitutional crisis in the states. In particular, the anti-slavery state courts are prone to use the term "unconstitutional" much more often than the slave states. The word "man-trap", a term used to refer to states where free blacks could be kidnapped purpose of enslaving them. The fugitive slave conflicts of the mid-19th century that led to the civil war were precisely about this aversion of the northern states to having to return runaway slaves to the Southern states.

Besides these subjective observations about the historical significance of the SME topics, we also conduct a more formal analysis comparing the SME classification to that conducted by a legal historian. Wahl (2002) analyses and classifies by hand 10989 slave cases in the US South into 6 categories: "Hires", "Sales", "Transfers", "Common Carrier", "Black Rights" and "Other". An example of "Hires" is Topic 4. Topics 1, 2, and 3 concern "Transfers" of slave property between inheritors, descendants and heirs-at-law. Topic 5 would be classified as "Other".

We take each of our 25 modelled topics and classify them along Wahl's categories, using "Other" when a classification could not be obtained. The classifications are quite transparent in virtually all cases, as certain words (such as "employer" or "bequest") clearly designate certain categories (respectively, such as "Hires" or "Transfers"). We then calculate the probability of each of Wahl's categories in Region 2. We then compare these to the relative frequencies of Wahl's categorization in the states that overlap with our Region 2 in Figure 6 and do a χ^2 test for goodness of fit, which allows us to reject difference at 0.1% confidence.

The SME model thus delivers topics that, at a first pass, are consistent with the history of the period as well as previous work by historians, showing the qualitative benefits of the model. We plan to conduct more vertical and temporal analyses using SME in the future.

6 Conclusion and Future Work

In this work, we propose a sparse mixed-effects model for historical analysis of text. This model is built on the state-of-the-art in latent variable modelling and extends that model to a setting where metadata is available for analysis. We jointly model those observed labels as well as unsupervised topic modelling. In our experiments, we have shown that the resulting model jointly predicts the region and the time of a given court document. Across vocabulary sizes and number of topics, we have achieved better system accuracy than state-of-the-art generative and discriminative models of text. Our quantitative analysis shows that early US state supreme court opinions are predictable, and contains distinct views towards slave-related topics, and the shifts among opinions depending on different periods of time. In addition, our model has been shown to be effective for qualitative analysis of historical data, revealing patterns that are consistent with the history of the period.

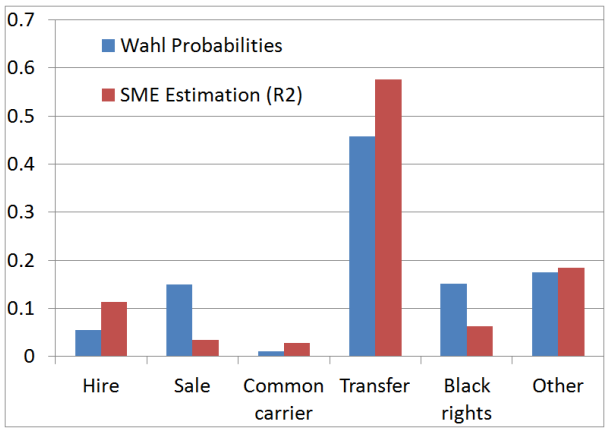

Figure 6: Comparison with Wahl (2002) classification.

This approach to modelling text is not limited to the legal domain. A key aspect of future work will be to extend the Sparse Mixed-Effects paradigm to other problems within the social sciences where metadata is available but qualitative analysis at a large scale is difficult or impossible. In addition to historical documents, this can include humanities texts, which are often sorely lacking in empirical justifications, and analysis of online communities, which are often rife with available metadata but produce content far faster than it can be analyzed by experts.

Acknowledgments

We thank Jacob Eisenstein, Noah Smith, and anonymous reviewers for valuable suggestions. William Yang Wang is supported by the R. K. Mellon Presidential Fellowship.

References

Lalit R. Bahl, Peter F. Brown., Peter V. de Souza, and Robert L. Mercer. 1988. A new algorithm for the estimation of hidden Markov model parameters. In *IEEE Inernational Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 493–496.

Ryan S.J.D. Baker and Kalina Yacef. 2009. The state of educational data mining in 2009: a review and future visions. In *Journal of Educational Data Mining*, pages 3–17.

David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, pages 993–1022.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent System Technologies*, pages 1–27.

Jake Chen and Stefano Lombardi. 2010. *Biological data mining*. Chapman and Hall/CRC.

Michael Collins. 2002. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8.

Agata Katarzyna Cybulska and Piek Vossen. 2011. Historical event extraction from text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 39–43.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1277–1287.

Jacob Eisenstein, Amr Ahmed, and Eric. Xing. 2011a. Sparse additive generative models of text. *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 1041–1048.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011b. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 1365–1374.

Annette Gotscharek, Andreas Neumann, Ulrich Reffle, Christoph Ringlstetter, and Klaus U. Schulz. 2009. Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data (AND 2009)*, pages 69–76.

Weiwei Guo and Mona Diab. 2011. Semantic topic models: combining word distributional statistics and dictionary definitions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 552–561.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence (UAI 1999)*, pages 289–296.

Thorsten Joachims. 1998. Text categorization with support vector machines: learning with many relevant features.

Kenneth Lange and Janet S. Sinsheimer. 1993. Normal/independent distributions and their applications in robust regression.

Dong Nguyen and Carolyn Penstein Rosé. 2011. Language use as a reflection of socialization in online communities. In *Workshop on Language in Social Media at ACL*.

Dong Nguyen, Elijah Mayfield, and Carolyn P. Rosé. 2010. An analysis of perspectives in interactive settings. In *Proceedings of the First Workshop on Social Media Analytics (SOMA 2010)*, pages 44–52.

Karen Orren. 1992. Belated feudalism: labor, the law, and liberal development in the united states.

Eva Pettersson and Joakim Nivre. 2011. Automatic verb extraction from historical swedish texts. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 87–95.

William D. Popkin. 2007. *Evolution of the judicial opinion: institutional and individual styles*. NYU Press.

Kumar Shubhankar, Aditya Pratap Singh, and Vikram Pudi. 2011. An efficient algorithm for topic ranking and modeling topic evolution. In *Proceedings of International Conference on Database and Expert Systems Applications*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, pages 1566–1581.

Jenny Bourne Wahl. 2002. *The Bondsman's Burden: An Economic Analysis of the Common Law of Southern Slavery*. Cambridge University Press.

William Yang Wang and Kathleen McKeown. 2010. "got you!": automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1146–1154.

William Yang Wang, Kapil Thadani, and Kathleen McKeown. 2011. Identifyinge event descriptions using co-training with online news summaries. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 281–291.

Gavin Wright. 2006. Slavery and american economic development. *Walter Lynwood Fleming Lectures in Southern History*.

Tze-I Yang, Andrew Torget, and Rada Mihalcea. 2011. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104.