Temporal Sequencing of Documents

Michael Gervers[1] and Gelila Tilahun[1]

[1]University of Toronto, Canada

Corresponding author: Gelila Tilahun , `gelila.tilahun@utoronto.ca`

Abstract

We outline an unsupervised method for temporal rank ordering of sets of historical documents, namely American State of the Union Addresses and DEEDS, a corpus of medieval English property transfer documents. Our method relies upon effectively capturing the gradual change in word usage via a bandwidth estimate for the non-parametric Generalized Linear Models [Fan et al., 1995]. The number of possible rank orders needed to search through for cost functions related to the bandwidth can be quite large, even for a small set of documents. We tackle this problem of combinatorial optimization using the Simulated Annealing algorithm, which allows us to obtain the optimal document temporal orders. Our rank ordering method significantly improved the temporal sequencing of both corpora compared to a randomly sequenced baseline. This unsupervised approach should enable the temporal ordering of undated document sets.

Keywords

temporal ordering of documents, document dating, sotu, deeds, pattern of word usage, non-parametric regression.

I INTRODUCTION

The accurate dating of historical and heritage texts is of paramount importance to historians. On the basis of such correctly sequenced texts, historians can examine, judge, and analyze events within the context of a specific time period. Often, only the undated textual contents of the historical documents are available to historians, on the basis of which they must infer the dates of composition [Gervers, 2002]. English property-transfer documents (charters or deeds) were selected as one component of this study because of their particular nature. That is, while the earliest surviving examples from the Anglo-Saxon period (c. A.D. 670 to 1066) are invariably dated, only 300 out of a total of approximately 1,600 can be considered originals. Experts have noted that many supposed Anglo-Saxon documents are actually later forgeries, but nonetheless are difficult to distinguish from genuine charters; "Thus in some cases, the date given is either demonstrably fictional or suspect" (Robert Getz, personal communications, April 25, 2019) or, in many cases, the charters survived only in later copies made centuries after the date of issue, resulting in genuine errors arising from misreading or miscopying by scribes. Common instances include miscopying of Roman numerals, confusion about a witness's identity, or the names of reigning monarchs [Whitelock, 1979, Cubitt, 1999]. For examples of documents allegedly forged or fabricated to justify social realities and for political motivations, see Hiatt [2004].

When the Anglo-Saxon political and judicial system was largely replaced by that of the Normans following their conquest of England in 1066, an entirely new phenomenon was intro-

1

arXiv:2311.02578v3 [cs.CL] 21 Sep 2024

duced: the undated charter. From 1066 to circa 1307 (the start of the reign of King Edward II) only about 3% of the million or more known charters issued bear internal dates. Dating was reintroduced to the royal chancery in 1189 under King Richard the Lionhearted, but the example was not followed by the nobility and commoners for five score years more. Compared to Continental charters, which with few exceptions were regularly dated internally for the duration, the first 600 years of the English charter record has always floated on a sea of incertitude.

In historical research, the one essential principle is to identify the correct order of events. As is evident from the foregoing, that is one of the most difficult tasks in the profession. It is of far greater concern, however, than for historians alone; in fact, it is common to virtually all avenues of scholarship, not to mention of the many institutions upon which literate society depends. Undated documents are everywhere, leaving lawyers, police and spy agencies, forensic linguists, code breakers, artists and art historians, businesses, real estate agents, medical practitioners, military analysts, philosophers (the list is endless), with the responsibility of determining what event preceded or succeeded another. This study sets the stage for anyone with a series of undated, digitized texts, or even lists, to determine a chronological order thereof without having to undertake the arduous task of examining each document for contextual clues and references to specific events, and of identifying periodization through content analysis, handwriting and/or watermarks. All of these aspects can be accomplished automatically through the temporal sequencing methodology outlined below.

By way of examples in which correct temporal ordering was essential, we note that in the financial fraud investigations of Enron Corporation, forensic linguistics was used to analyze emails, memos and internal communications to re-construct the timeline of fraudulent activities even when the timestamps of these evidential materials were not always available [McLean and Elkind, 2003].

The Library of Congress contains many written documents from former presidents of the United States. For example, The Papers of Abraham Lincoln[1] is a digitized corpus of over 40K documents consisting of the correspondence, notes, memos, personal letters, drafts of speeches of Abraham Lincoln from his time as a lawyer, congressman and then as the 16th president of the United States. Chronological gaps in The Papers remain, as not all of the original letters and documents were meticulously dated or preserved. A proper chronological order would give insight into the President's evolving thoughts and ideas through a tumultuous period of American History. Another similar example is The Papers of Thomas Jefferson[2], a digitized corpus of 25K items consisting of the correspondence of Thomas Jefferson, who was a diplomat, architect, lawyer and the third president of the United States. Besides correspondence, the collection also includes his drafts of The Declaration of Independence, drafts of laws, fragments of his autobiography, his personal notes, including his records of spending and even recipes! Establishing an accurate chronological order of The Papers is crucial in understanding the personal worldview and the evolving visions about the early Republic by one of the prominent Founding Fathers.

The medieval Exeter Book[3] is another example. It is an anthology of Old English poetry and riddles from the 10th century, but the chronological order of none of the texts is known. Establishing a chronological order of the texts would give us a deeper understanding of the evolution

---

[1] https://www.loc.gov/collections/abraham-lincoln-papers/
[2] https://www.loc.gov/collections/thomas-jefferson-papers/
[3] https://www.exeter-cathedral.org.uk/learning-collections/explore-the-collections/the-exeter-book/

of Old English and the literary culture of the Anglo-Saxon people.

Previous efforts in document sorting have been directed towards the development of historical language models [Feuerverger et al., 2005, 2008, Tilahun et al., 2016, Gervers et al., 2018]. Within the broader field of information retrieval, investigators employ statistical models incorporating temporal aspects of term usage [Swan and Jensen, 2000], study the relationship between time and the retrieval of relevant documents [Li and Croft, 2003], and classify document dates according to time partitions of predefined granularity [De Jong et al., 2005]. In preparing web pages that present document lists the accuracy of time-stamping is of paramount importance. In this regard, Kanhabua and Nørvåg [2008] and Kanhabua and Nørvåg [2009] extended De Jong et al. [2005]'s work by integrating semantic-based techniques into the document pre-processing pipeline, the aim being to improve the temporal precision of web pages and web document searches where the trustworthiness of document time-stamps is often questionable. Lebanon and Zhao [2008] modelled temporal text streams, and Mani and Wilson [2000] extracted temporal expressions such as *now*, *today*, and *tomorrow* from print and broadcast news documents to resolve uncertain temporal expressions. Chambers [2012] presented models that incorporate rich linguistic features about time, whereas Vashishth et al. [2019] employed deep neural network-based methods that also exploit linguistic features about time in order to date the documents. However, as pointed out by Kotsakos et al. [2014], relying upon temporal terms for dating suffers from the drawback that terms "can be very sparse as well as ambiguous, referring to irrelevant timeframes". These authors proposed using statistical approaches based on lexical similarity and burstiness — the sudden increase in the frequency of terms within a time-frame.

In the current work, we propose TempSeq[4], an unsupervised method for the temporal *sequencing* or *ranking* of documents. This approach is designed to be applicable when the only available data are the undated documents that are to be temporally sequenced. TempSeq relies on a 'bag-of-words' approach, and does not make use of linguistic features about time, nor does it use a training set data with time-tag. In addition, our approach does not make use of specific language rules, word representations, or any other metadata information, thus presenting a potentially significant advantage in the task of document temporal ordering. TempSeq relies on measuring word usage drift under the assumption that word usage changes gradually over time, which means that the temporal variability in word usage is low. We model word usage drift via the non-parametric Generalized Linear Models regression [Fan et al., 1995], and we estimate the correct temporal sequencing of the documents to be the one that minimizes, on average, the associated kernel bandwidths (a direct measure of the temporal variability in word usage). To our knowledge, using the variability of word usage drift to ascribe a temporal sequencing for a set of documents is entirely new.

The necessity for temporal sequencing of documents arises not only in the field of information retrieval, but also in studies of heritage texts, which frequently lack timestamps, are intentionally ambiguous with respect to time of issue, or can even be outright forgeries. Often, only the textual contents of heritage documents are available, from which one must infer the dates of issue [Gervers, 2002]. Furthermore, heritage texts that are available as a training data set can be limited in number, as proportionately few documents have survived across the centuries to the present time, thus necessitating an unsupervised method for inferring document dates. The task at hand is not only to infer the temporal ranking of a collection of documents or corpus, but also to identify the terms that contribute most towards the task of identifying the correct temporal

---

[4]The TempSeq pseudo and source codes used in this paper are available at https://github.com/gitgelila/TempSeq

ranking/ordering. We believe that such terms are most likely the temporal signatures for identifying the characteristics of intentionally or inadvertently temporally mislabelled documents, or documents with missing or corrupted timestamps.

Past and present research about problems arising in document sequencing involve criteria for document classification, for example, by topic, Blei et al. [2003], Blei and Lafferty [2006], McAuliffe and Blei [2008], Taddy [2013, 2015], document indexing Roberts et al. [2016], and document ranking using latent semantic analysis Deerwester et al. [1990], Hofmann [1999]. Cohen et al. [1998] consider the problem of machine learning to order instances, as opposed to classifying them, when an algorithm's output receives feedback in the form of *preference judgments*, i.e., "statements indicating which instances should be ranked ahead of the others" [Cohen et al., 1998]. However, such lines of research have not directly addressed the problem of document ordering *per se*. Perhaps the closest approach to dealing with this problem is that of Thinniyam in her dissertation [Thinniyam, 2014]. Her approach is based on the notion of similarity (or distance) between two documents, namely the supposition that similar documents are more likely to discuss similar topics, and therefore should have closer underlying temporal footprints. She proposed a linear regression-based approach, with regression of observed document distance measures of spacings between consecutive documents. In a separate approach, Thinnayam also framed the document ordering problem in a Bayesian framework, where the distribution of pairwise distances between documents was modelled conditionally on a time-line vector, where the ith coordinate value of the vector represents the time interval between the ith document and a reference document. Herein, a Markov Chain Monte Carlo (MCMC) method was employed to sample from the posterior distribution of the timeline. Thinniyam's ordering methods fundamentally require pairwise distance measures of documents (i.e., a quantifiable measure of dissimilarity between two documents) to estimate the temporal order of a set of documents within a corpus. Other studies have shown that such measures are prone to yield spuriously high values of similarity due to an abundance of uninformative terms within the documents, including, but not limited only to stop words [Tilahun, 2011]; it is not always a straightforward matter to identify these uninformative terms, and the requisite degree of filtration. Moreover, using measures of document distance does not allow identification of the particular words or terms that are essential for determining the predicted temporal rank orders. In addition, the degree to which two documents are similar/dissimilar is highly dependent on the type of distance measures that are used Broder [1997].

In contrast to Thinnyam's previous approach, the present TempSeq method temporally ranks a set of documents even when a reliably dated training dataset is not available, and/or when there is a very limited number of documents in the set. The TempSeq method relies fundamentally on modelling the probability of occurrence of words in a given date range, thereby avoiding the need for a document distance measure. By design, TempSeq also filters out words according to their degree of uninformativeness, thereby allowing us to gain insights into the history underlying the documents by identifying words that are most putatively useful for determining the correct temporal ordering of documents. We test the TempSeq method on two corpora of heritage texts; one written in American English and the other in Latin. When a set of training data is available, TempSeq can optimize the smoothing parameter for temporal sequencing. More importantly, we show for both corpora that the TempSeq temporal sequencing method performed significantly better as compared to random sequencing, in the absence of training data.

II CORPORA

We evaluated our temporal sequencing methods on two different sets of corpora with time-tags. The first corpus consisted of 240 transcripts of the American State of the Union Address (SOTU), from the years 1790 to 2020. Each transcript had a median average length of 6400 words. This corpus is available from the R package, Arnold [2022]. The second corpus is from the Documents of Early England Data Set (DEEDS)[5]. From within this corpus, we focused on a set of 11,463 English property conveyance records issued in the years 1120 to 1300. All the records are written in Latin, and have been inspected for content by subject expert historians to accurately verify the date of issue. The Latin documents have a median length of 175 words. We chose to evaluate the TempSeq temporal document sequencing method on this corpus as it consists of documents similar to corresponding DEEDS documents from the Anglo-Saxon period, which, as mentioned in section I, have generally unreliable dates. In this project, we considered the DEEDS corpus in two different forms. In the first form, we conflate all the documents written in a given year into single texts, thus yielding 181 conflated DEEDS records, of mean average length of approximately 11,000 words. We denote the conflated collection as "DEEDS-conflated". In the second form, we denote the entire set of 11,463 unconflated records as "DEEDS-single".

III OUTLINE

When a set of training documents with known dates is available, Tilahun et al. [2012] have proposed the "Maximum Prevalence" method for their dating. This approach is based on modelling (on the basis of the training data) a curve that describes the temporal pattern of the probability of occurrence of each word from the undated documents. For example, in the DEEDS corpus, the proposed dating method achieves very reliable date estimates, giving a test set median dating error of ± 5 years within the 230 year span [Gervers et al., 2018]. High accuracy validates an underlying feature of the model, namely useful words for dating a document are those with a non-uniform probability of occurrence across a date range, and showing a gradual change in the variability of their usage changes. Words such as *et, de, huic* (in Latin), *the, to, that, on* (in English), and stop words, which appear in consistent proportion at all times, that is to say *uninformative* words, do not contribute to the date estimation of an undated document.

In section IV, we discuss our modelling approach to estimate the curves best describing the temporal pattern of the probability of occurrence of a given word/phrase, and present examples of such curves. We also examine the properties of the curve estimates (in particular, that of a smoothing parameter) in relation to the bias-variance trade-off, using the form of the bias-variance trade-off to select the optimal curve. This trade-off is at the heart of any statistical learning process. We could perfectly fit training data (zero bias) to a model by including excessive amounts of parameters (excessive under-smoothing in the case of our model). This situation, referred to as over-parametrization, risks overfitting the data because the model learns not only the pattern in the data but also the random noise and fluctuations that are present. When an overfitted model is applied to a test data, the performance is often poor. On the other extreme, when a model is under-fitted (over-smoothed in the case of our model) the effect from random noise is eliminated but at the expense of failing to learn the pattern from the data. The right amount of parameterization is one that balances bias and variance (large bias and low variance). We seek an optimization process which can balance this trade-off. This optimization seeks on the one hand to minimize bias, thereby increasing curve fluctuation to accurately track

---

[5] https://deeds.library.utoronto.ca

the empirical values of the probability of word occurrences. At the same time, the optimization minimizes variance, thereby decreasing the amount of curve fluctuation to obtain a smooth curve. The optimally smoothed curve for balancing the trade-off between these demands is a quantifiable parameter value that can be estimated using a "rule-of-thumb" smoothing parameter estimate [Fan et al., 1995]. In section V, we address the problem of temporally ordering a set of documents in the absence of a set of dated training data. To this end, we compute the average value of the optimal smoothing parameters for estimating the probability of occurrence of each word in the documents. Here, we find a close estimate of the correct temporal order of well-spaced subsets of ($m = 10$) documents by searching among all possible temporal orderings to identify the highest average optimal smoothing parameter. We carried out this search using combinatorial optimization via the Simulated Annealing algorithm. In section VI, we evaluate the TempSeq method and present its results for the two distinct corpora — the DEEDS corpus and the SOTU corpus. In addition, we identify the informative words that enabled TempSeq to establish the correct temporal order for the selected subset of documents. In section VII, we provide error analysis, and present our general conclusions in section VIII. Theoretical background and operational equations are presented in Annex A and Annex B.

IV   MODELLING THE TEMPORAL PATTERN OF WORD USAGE

Our fundamental assumption is that word usage changes gradually. We model the probability of word usage as a function of time using the local polynomial kernel regression for the generalized linear model, Fan et al. [1995]. For further details of this model, see Annex A, section 1.1.

Suppose $(D_i, t_{D_i}), i = 1, \ldots, n$ represents a sequence of data pairs, where t_{D_i} represents the date of the D_i^{th} document and n denotes the size of the data set, that is to say, the total number of documents in the collection. Let $n_w(D_i)$ denote the number of occurrences of the word (or term) w in document D_i. Finally, let $N(D_i)$ denote the total number of words (or terms) in document D_i. We are interested in estimating the probability of occurrence of the term w at time t, which is given by:

$$\hat{\pi}_{w,h}(t) = \frac{\sum_{i=1}^{n} n_w(D_i) K_h(t_{D_i} - t)}{\sum_{i=1}^{n} N(D_i) K_h(t_{D_i} - t)}. \tag{1}$$

We define the weight term K_h to be $K_h(u) \equiv h^{-1} K(u/h)$, where K is called a kernel function. Typically, K is a bell-shaped, non-negative function, with an area under the graph equalling unity. The function decays fast enough to eliminate the contributions of remote data points. (See equation 9, in Annex A, section 1.1). For the present study, we used the t-distribution function with a low degree of freedom value (equal to 5) to adequately weigh distant points.

As a weight term, K_h fits a polynomial regression curve around the data in the neighbourhood of t, where h, called the bandwidth parameter, is the size of the local neighbourhood, such that data points distal from the neighbourhood are down-weighed (for this reason, h is also referred to as the curve smoothing parameter). In simple terms, if h is very large (highly smoothed), then $\hat{\pi}_{w,h}(t) \approx \sum_{i=1}^{n} n_w(D_i) / \sum_{i=1}^{n} N(D_i)$, thus representing an overall proportional outcome of word w which does not change with t. On the other extreme, if h is very small, then, $\hat{\pi}_{w,h}(t)$ evaluated at, say, t_{D_j}, has the value $\hat{\pi}_{w,h}(t_{D_j}) \approx n_w(D_j)/N(D_j)$, which is the proportional outcome of word w in the document written at time t_{D_j}. In this case, information on the frequency of occurrence of word w in documents written at dates near to t_{D_j} has been completely ignored in determining the value of $\hat{\pi}_{w,h}(t_{D_j})$. When h is very small, the curve $\hat{\pi}_{w,h}(t)$ overfits, thus fluctuating rapidly to attain the values $n_w(D_i)/N(D_i)$ for each time point $t_{D_i}, i = 1, \ldots, n$.

Although it is possible to draw a curve that perfectly describes the empirical probability of occurrence of a word across a date range (i.e., a bias with a value of zero), the consequent high variance of the curve means that, when applied to a test data set, the curve would overfit thus depicting a very inaccurate description of the probability of occurrence of the word of interest. In the field of kernel regression, there has been extensive research on how to select the appropriate bandwidth parameter h. However, in implementing the TempSeq method, we have relied on using a rule-of-thumb selection approach (Fan et al. [1995]), which we describe in Annex B, section 2.1. The theoretical details for the derivation of equation 1, which falls under the *locally constant case* ($p = 0$), can be found in Annex A, section 1.1.

Relying on our assumption that change in word usage is gradual, we now focus on the role played by the bandwidth parameter h in setting the bias-variance trade-off of the estimator $\hat{\pi}_{w,h}(t)$. Figure 1 illustrates the probability of occurrence of the words *Drug(s)* in the SOTU corpus at different dates and for variable bandwidth settings. The x-axis illustrates the calendar year (the time t), ranging from 1790 to 2020. The y-axis illustrates the values of $\hat{\pi}_{w,h}(t)$. The asterisks show the proportion of occurrences of the words "Drug(s)" in the years for which dated SOTU documents are available. The proportion is greater than zero in a few of the years, but is zero for a few years preceding or following that time. In that circumstance, when the bandwidth value h is small, the resulting estimate, $\hat{\pi}_{w,h}(t)$, is highly variable (the closer h approaches zero, the closer its values match the recorded proportion of occurrence of the term in the training data). This behaviour is illustrated by the dashed-line curve of $\hat{\pi}_{w,h}(t)$ in the figure. When the value of h is larger (smoother), then the resultant probability curve $\hat{\pi}_{w,h}(t)$, overlaid in solid, has less variability. Conversely, when the value of h is very large, the closer the values of $\hat{\pi}_{w,h}(t)$, for all t, approach the proportion of the real occurrence of the term w across all the dates in the document set (illustrated in dotted lines). The bandwidth, which controls both the bias and the variance, is therefore a crucial parameter of the estimator $\hat{\pi}_{w,h}(t)$.

The optimal amount of smoothing of the data in figure 1 results in the solid curve, which illuminates a clear pattern in the data. The first peak in the figure (coinciding with the presidency of Richard M. Nixon (1970-1974)) refers to the emergence of the so-called *War on Drugs*, a US-led global policy aimed at the production, distribution, and use of psychoactive drugs, which was presented in Nixon's State of the Union address of 1972. That address contains the phrases '... *strong new drug treatment programs ...*', '... *by continuing our intensified war on drug abuse ...*', '... *Special Action Office for Drug Abuse Prevention ...*', '... *collective effort by nations throughout the world to eliminate drugs at their source ...*', '... *to drive drug traffickers and pushers off the streets of America ...*', '... *to curb illicit drug traffic at our borders and within our country ...*'. His 1974 address contains the phrases '... *the spiraling rise in drug addiction ...* ', '... *The Psychotropic Convention ... treaty regulating manufactured drugs worldwide ...*' and '... *the drug battle is far from over ...*'. The first peak extends to President Gerald Ford's 1976 address, which contains phrases such as, '*The sale of hard drugs is tragically on the increase again ...*' and '... *shipment of hard drugs ...*'. The second peak in figure 1 occurs decades later, around the year 2000. Then President William Clinton's 1998 address contains phrases such as, '... *to crack down on gangs and guns and drugs ...*' and '... *the largest anti-drug budget in history ...*'; his 1999 address has phrases such as '... *if you stay on drugs, you have to stay behind bars ...*' and '... *to strengthen the Safe and Drug-Free School Act ...*'. In his 2000 address, Clinton announced '... *new legislation to go after what these drug barons value the most, their money.*'. Clinton also invokes the words *drug(s)*, in the positive sense of insurance coverage for affordable prescription drugs. In that context, he used phrases such as '... *seniors now lack dependable drug coverage ...*' and '*Lifesaving drugs are an indispensable*

part of modern medicine ...'. In the following years, under President George W. Bush, policy regarding affordable drug coverage becomes a major issue domestically and globally; '... *some form of prescription drug coverage ...*' (in the 2001 address), '... *new drugs that are transforming health care in America ...*' (in the 2003 address) and '*More than 4 million require immediate drug treatment*' (in the 2003 address regarding the lack of antiretroviral drugs in Africa). Thus, the first peak in figure 1 is exclusively related to illicit drug issues, whereas the second peak, some 25 years later, is primarily related to the affordability of prescription drugs.

Figure 2 illustrates the probability of occurrence of the words *Anglicis* and *Anglis* in the DEEDS corpus. These words are often found within the form of address '*Franc[igen]is quam(et) Angl[ic]is*' (French and English), such as '... *tam presentibus quam futuris tam Francigen[is] quam Anglicis salutem Sciatis me intuitu dei assensu ...*', (... both present and future, both French and English, greeting. Know that with God's consent I have [granted] ...) and '... *omnibus ministris et fidelibus suis Francis et Anglis de Oxenfordscira ...*', (... to all his French and English ministers and servants of Oxfordshire ...). The above form of address was commonly used by French and the English barons of the time to address their subjects. However, after the province of Normandy was conquered by the French in 1204, this form of address virtually fell out of use.

Figure 3 illustrates the probability of occurrence of a common stop word '*de*' (of) in the DEEDS corpus. The asterisks in the figure show the proportion of occurrences of the word across time (1120 to 1300), and the line curve is found by smoothing the proportion of occurrences of the word across those same years. When comparing the smoothed black curves in figure 1 to figure 3, we see that the latter curve is more uniform across time (except for the years prior to 1125, when fewer documents were available). This behaviour, which we call the principle of temporal uniformity (non-uniformity) of uninformative (informative) words, is shown in the figure where the probability of occurrence of *de* has no defining temporal feature, and is uniform across the date range. For the purpose of temporally ordering a set of documents, no matter what combination of temporal ordering is evaluated in the TempSeq process, the contribution of uninformative words is immaterial.

In all of these figures 1 to 3, we see that the solid black curves have relatively the most optimal smoothing, as opposed, for example, to the highly variable dashed-line curve in figure 1. The optimal smoothing for a given curve represents the trade-off between small bias and small variance for the curve estimator. If the data analyst then randomizes the true temporal ordering of word usage, applying the optimal smoothing parameter will now result in a curve estimate that rapidly oscillates (high variance) due to seeking a minimum bias.
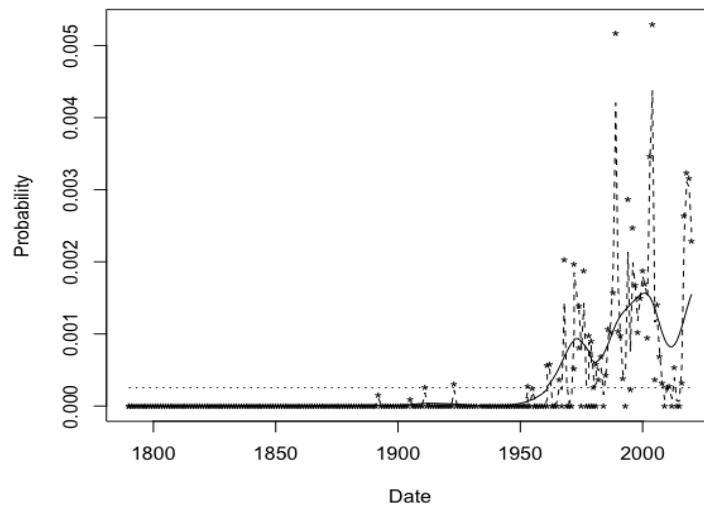
Figure 1: Asterisks show the proportion of occurrences of the words *Drug(s)* in the SOTU corpus. The solid curve is based on a larger bandwidth value than that of the dashed-lined curve. The dotted curve (the horizontal dotted line) is based on a very large bandwidth value. Date (time) is the x-axis and $\hat{\pi}_{w,h}(t)$ is the y-axis.
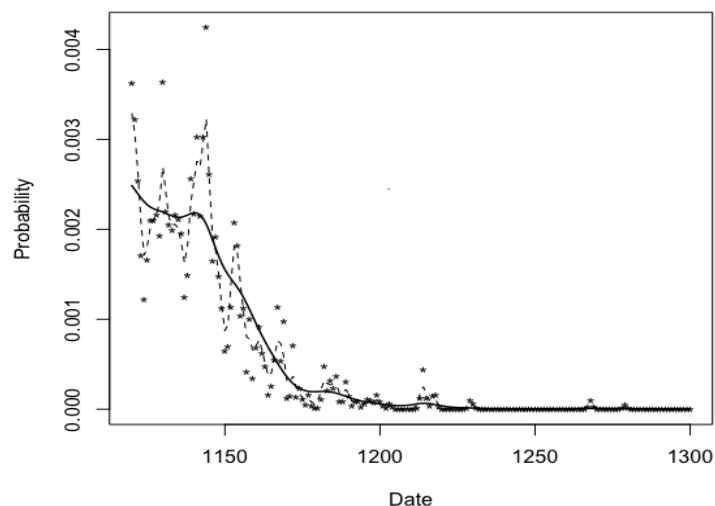

Figure 2: Asterisks show the proportion of occurrences of the phrase *Angl(ic)is* in the DEEDS corpus. The solid curve is based on a larger bandwidth value than that of the dashed-lined curve. Date (time) is the x-axis and $\hat{\pi}_{w,h}(t)$ is the y-axis.
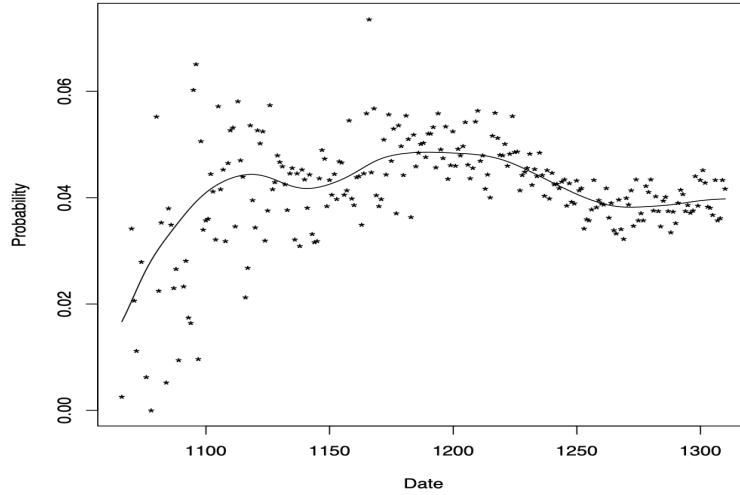
Figure 3: Asterisks show the proportion of occurrences of the word *de* (of). The smoothed solid probability curve is uniform across the date range. Date (time) is the x-axis and $\hat{\pi}_{w,h}(t)$ is the y-axis.

V THE TEMPSEQ METHOD FOR TEMPORAL SEQUENCING

5.1 Determining the Optimal Bandwidth

Let $\{D_1, D_2, \ldots, D_m\}$ be a set of m number of documents that we wish to sequence in temporal order. We assume that all the documents have a unique timestamp, and as such, there are $m!/2$ possible orders when two orderings that are the reverse of each other are equated. Without loss of generality, assume that the sequence $l = (1, 2, \ldots, m)$ represents the true temporal rank order of the m documents. Let $\sigma(l)$ represent a permutation of the ranks and let σ_0 represent permutation identity, that is to say $\sigma_0(l) = l$. For each word w, $w \in \cup_{i=1}^{m} D_i$, and temporal rank ordering of the documents under $\sigma(l))$, we compute the asymptotically optimal bandwidth value for $\hat{\pi}_{w,h}(t)$, which we denote as $h_{amise,w,\sigma(l)}$. This bandwidth value is estimated via a 'rule-of-thumb' estimate, the detail of which can be found in Annex B, sections 2.1 and 2.2. The formulation of $\hat{\pi}_{w,h}(t)$ here is subject to the condition $p = 1$ (the case of the locally linear regression, equation 10), which is more accurate than the formulation of $\hat{\pi}_{w,h}(t)$ in equation 1 (the case of locally constant regression, $p = 0$). For theoretical details, refer to Annex A, section 1.1.

Following the principle of temporal non-uniformity of informative words, the optimal smoothing parameter $h_{amise,w,\sigma(l)}$ will be larger under the correct temporal ordering of the documents, since the curve would not entail such extensive oscillation to obtain a small bias. Therefore, we would generally expect

$$h_{amise,w,\sigma_0(l)} \geq h_{amise,w,\sigma(l)} \tag{2}$$

to hold for each word w. Put another way, the rule-of-thumb bandwidth estimate of a word associated with the correct temporal ordering of documents will be larger than those bandwidth estimates based on incorrect temporal orderings. For a set of documents $\{D_1, D_2, \ldots, D_m\}$, we estimate the temporal rank order for the set of documents by first computing

$$H_{\sigma(l)} = \underset{w}{\mathrm{Median}} \left\{ h_{amise,w,\sigma(l)} : w \in \cup_{i=1}^{m} D_i \right\} \tag{3}$$

where $H_{\sigma(l)}$ is the uniform median value of the optimal bandwidths associated with each word[6] present in the m number of documents, and $\sigma(l)$ is a proposed temporal ordering. The temporal rank order estimate, $\hat{\sigma}(l)$, is the rank order which maximizes the term $H_{\sigma(l)}$ over all possible permutations; stated more succinctly,

$$\hat{\sigma}(l) = \arg\max_{\sigma} H_{\sigma(l)}. \tag{4}$$

The estimated temporal rank order is one that results, on average, in the smoothest rule-of-thumb bandwidth estimate of $h_{amise,w,\sigma(l)}$ over all the words in the m number of documents and over all possible temporal rank orders.

To verify our expectation that equation (2) in fact holds in general, which would imply that $H_{\sigma_0(l)} \geq H_{\sigma(l)}$ also holds, we conducted the following experiment separately on the DEEDS and the SOTU corpora. In each case, we randomly selected a set of ten documents with date gaps of about 20 years, i.e., one tenth of the document history, thus obtaining a trade-off between excessive computational time and fitness of the method for correct ordering. In this computation, only those words that occurred at least once in two separate documents were considered. Based on random permutations of the underlying true temporal rank order, $\sigma(l)$, we computed $H_{\sigma(l)}$. For the same set of ten documents, we also computed $H_{\sigma_0(l)}$, when the true temporal order of the documents was not permuted. We ran 100 replications of the above experiment. Figures 4(a), 4(b) and 4(c) are the box plots of $H_{\sigma(l)}$ (bandwidths) for the SOTU, DEEDS-conflated and DEEDS-single corpora, respectively. In each figure, the first box plot is that of $H_{\sigma(l)}$ where $\sigma(l)$ is a random permutation of the true temporal order of the given set of ten documents, and the second box plot depicts the case when the true temporal order of the same ten documents is maintained. As shown by these box plots, optimal bandwidths associated with the true temporal orderings are generally larger (smoother) than those associated with random orderings. In comparing all the box plots in the figures, we see that those associated with DEEDS-single more closely resemble one another. This result is not surprising, since the computation of equation 3 on sets of documents relies upon fewer words than those from the SOTU and DEEDS-conflated corpora.
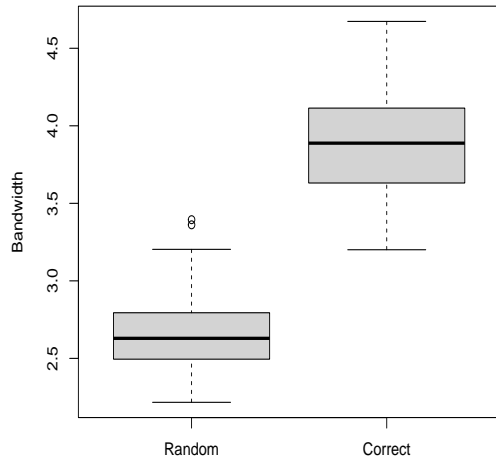
We note that no matter what the permutation of the underlying documents' sequence, the bandwidth h_{amise} associated with uninformative words remains unchanged. Therefore, the contribution of uninformative words has negligible influence on the estimation of the temporal rank order of the set of documents.

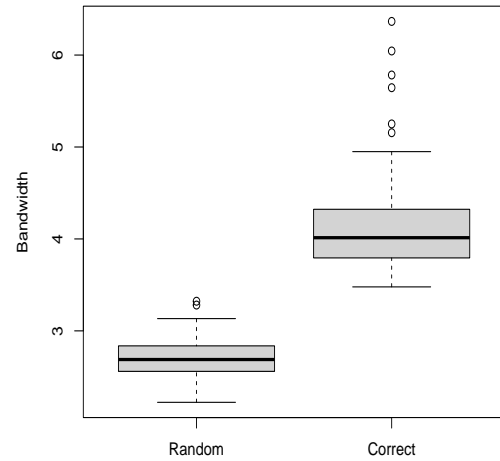## 5.2 Estimation via Simulated Annealing

With an increasing number of documents that we wish to place in order, there is a corresponding increase in the number of permutations required to search exhaustively in order to obtain $\hat{\sigma}(l)$ (equation 4). For example, when $m = 10$, the requisite number of permutations equals $10!/2$ (circa 1.8 million), where a permutation and its reverse are equated. Scaling up to a large number of possible permutations to optimize in an objective function (such as equation 4) calls for combinatorial optimization. We propose to solve this problem using the well-known Simulated Annealing algorithm [Kirkpatrick et al., 1983].

In the current task, we are attempting to find permutations of the temporal rank order of the document sets that maximize $H_{\sigma(l)}$. The optimization problem involves a search over the neighbours of a permutation element $\sigma(l)$, and the generating scheme of the candidate solution
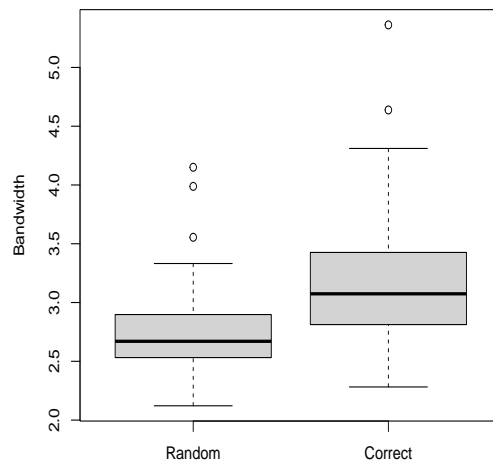
---

[6]For a word w to be included in the estimations, we required that it occurs in at least two documents as measuring the pattern of word usage fluctuations is a key element.

(a) The State of the Union Address

(b) DEEDS-conflated

(c) DEEDS-single

Figure 4: Box plots of $H_{\sigma(l)}$ (Bandwidths) versus temporal orders for 100 randomly selected sets of ten documents. In all figures, the first two box plots are that of $H_{\sigma(l)}$ for temporally randomly permuted and unpermuted sets of documents. Figure 4(a) shows the results for the SOTU corpus. Figures 4(b) and 4(c) show the results for the DEEDS corpus (DEEDS-conflated and DEEDS-single, respectively).

(neighbourhood) along with its set size are important factors in the performance of Simulated Annealing [Tian et al., 1999]. We employ a neighbourhood generating scheme proposed by those authors for the well known Travelling Salesman Problem. The proposed scheme generates a random permutation solution from the current one by reversing and/or moving a subsequence of terms. For example, the sequence $\{1, 2, 3, 4, \underline{5, 6, 7, 8}, 9, 10\}$, could generate the $\{1, \underline{8, 7, 6, 5}, 2, 3, 4, 9, 10\}$. In fact, this perturbation scheme (where a random set of subsequences with four terms that are randomly reversed and/or moved), was employed in this paper to generate the candidate neighbours for the Simulated Annealing algorithm. The authors prove that under such a random perturbation scheme to generate random permutation solutions, the Simulated Annealing algorithm converges asymptotically to the set of global optimal solutions.

VI   EVALUATION AND RESULTS

For a task of ordering temporally a set of m number of documents $\{D_1, D_2, \ldots, D_m\}$, as in section V, let the true temporal rank order of the documents be $l = (1, 2, \ldots, m)$. Let $\sigma(l)$ represent a permutation of the ranks and let σ_0 represent the permutation identity ($\sigma_0(l) = l$). We measure the extent to which two permutations are in close proximity to one another using the Spearman's (ρ) rank correlation. If $\hat{\sigma}(l)$ is the predicted temporal rank order for the set of documents $\{D_1, D_2, \ldots, D_m\}$, then the closer $|\rho(\hat{\sigma}(l), \sigma_0(l))|$ is to unity, the more accurately the predicted order matches the true order (we only consider the absolute value of the correlation since forward and reverse orders are equated, as noted above).

We randomly selected sets of 10 documents, dated approximately 24 years apart for the SOTU, and 18 years apart for the DEEDS-conflate corpora. For the random selection, we used systematic sampling, as follows: First, we randomly selected the start date document, and then every 24th year in succession for the annual SOTU, and then for ever 18th year for the annual conflation of DEEDS documents (DEEDS-conflated), or the corresponding randomly selected single DEEDS documents (DEEDS-single). If the subsequently selected year exceeded the range of dates, we cycled through from the start date. Then, we labelled the temporal rank of the resulting documents from 1 to 10. Starting from an initial random temporal order of the ten documents, we estimated their true temporal rank order as described above, and calculated the Spearman correlation for the true and shuffled ordering. In all cases, our analysis was based on words that occurred at least once in two separate documents (otherwise, no information regarding temporal ordering can be inferred). For 100 replications of the above procedure, the median of the absolute value of the correlations between the estimated and the true rank orders was 0.66 for the SOTU corpus and 0.78 for the annually conflated DEEDS corpus. As a baseline comparison, we note that the median of the absolute value of the correlations between the true temporal rank orders and their 100 random permutations was 0.24. In the computation, we only considered words that occurred more than once in the set of 10 documents. As shown in the box plots of figure 5, the TempSeq method performed significantly better than the baseline correlation. The Wilcox rank sum test and the t-test showed a statistically significant difference between the baseline correlation and the correlations associated with each of the SOTU and the DEEDS-conflated corpora. In both cases, p-value $< 2.2 \times 10^{-16}$.

Regarding the DEEDS-single corpus, the TempSeq did not perform as well as for the temporally DEEDS-conflated collection. Although statistically significantly better than the baseline correlations (for tests based on Wilcox rank sum and the t-test, p-value $< 5.5 \times 10^{-7}$), the median correlation coefficient between the estimated and the true rank orders was only 0.45 (see figure 6). This raises the question of why TempSeq under-performed on the DEEDS-single collection

as compared to the DEEDS-conflated collection. These two collections differ in that the latter was created by merging all DEEDS documents for a given year, such that the DEEDS-conflated documents had a mean of 11,000 words as compared to only 175 words for the DEEDS-single collection. As such, the temporal sequencing for the DEEDS-single collection according to equation 4 is based on a very small sample of words. Given the requirement that each word should occur at least once in two separate documents within a set, there are correspondingly fewer words informing the estimation of temporal order.

As stated in section I, the TempSeq approach for document sequencing allows the identification of words that are most informative for determining the correct temporal order. The process that we used to identify and analyze the informative words is as follows: As an example, we first considered a set of 10 documents from the DEEDS-conflated collection for which the TempSeq prediction of ordering, $\hat{\sigma}(l)$, resulted in a Spearman correlation coefficient of 0.92. Although the total number of unique words across these 10 documents was 15,757, when we considered the number of words that occurred at least once in two documents, the number of the relevant unique words declined to 5,988. Of these 5,988 words, we selected those with frequency values in the top 50th percentile. For each of these selected words, we computed the rule-of-thumb optimal bandwidth values under the predicted TempSeq order, $h_{amise,w,\hat{\sigma}(l)}$. On the basis of these optimal bandwidths, we extracted words that were in the top 88th percentile for their maximum probability of occurrence score over the temporal rank domain in equation 1. The first condition ensured that the words under consideration were those that occurred with sufficient frequency (exceeding the median). The second condition ensured that when a set of documents was listed in its correct temporal order, the informative words (i.e., the words that were most useful for attaining the correct temporal ordering using TempSeq method) were those whose occurrence clustered in near-by time periods. In turn, these clusterings resulted in higher word probability occurrence around those same time periods (refer to equation 1) as compared to the other time periods. Using the above word filtering procedure, we obtained a final list of around 360 relevant words.

To interpret meaningfully the 360 relevant words in the list, we compared them with the representative topic terms from an LDA (Latent Dirichlet Allocation, Blei et al. [2003]) topic modelling run on eight topics on the entire DEEDS-single documents [Gervers and Tilahun, 2023]. In the pre-processing stage (prior to running the LDA algorithm), each document was split into tri-gram words (sequences of three consecutive words). The topic proportions for each of the documents were aggregated in accordance with their date of issue (see Figure 6.1 in Gervers and Tilahun [2023]). When we examined the dominant topics corresponding to the dates from the set of the 10 documents, we found that the vast majority of the informative words extracted using the procedure outlined above matched the words in the tri-gram topic terms. Some examples of the informative words and their contexts include *finalis facta concordia* (indicating that the document is a 'final concordance made'), *anno regni regis* ('in the year of the king's reign'), *scripto sigillum meum* ('marking the document (with) my seal', indicating the sealing of a grant), *pro omni seruitio* ('for all service', indicating that a transfer was not a simple donation), and *perpetuam elemosinam* ('in perpetual alms', indicating that the transfer was a donation).

For the SOTU corpus, we similarly selected a set of 10 documents, each one separated by 23 years during the interval from 1810 to 2019. The Spearman correlation coefficient obtained between the TempSeq ordering method and the true orders for this set of documents was 0.95. The informative words were extracted and examined in the same way as described for the DEEDS-

conflated corpus. Of a total of 29,537 words, there were 3,957 words that occurred at least once in at least two documents, of which the top 20th percentile of the maximum probability of occurrence score over the temporal rank domain in equation 1 yielded 435 words. For the purposes of illustration, we present three words from the short list of relevant words: *Britain*, *Families*, and *Court*. A bar graph of the frequencies of these words, counted from the selected set of 10 presidential speeches, is illustrated in figure 7 .

We then ran an LDA topic modelling with five topics, where in the pre-processing stages, documents were split into bi-grams (sequences of two consecutive words), over the entire SOTU corpus. To enable an interpretation of the words that drove the temporal ordering, we examined the high ranking (top 20) bi-gram words of a topic associated with fiscal and commercial interests of the US. Within this topic, *Great Britain* was one of the high ranking words. One of the uni-gram forms of the term, *Britain*, turned up in the list of the top relevant words. From the 1810 SOTU address by James Madison, the term *Britain* was invoked in the context of the naval blockade suffered by the US when the Napoleonic Wars had spilled into the Atlantic. Twenty-three years later, *Britain* was discussed in the context of final settlement on the US North-East boundary and navigational safety concerns; another twenty-three years later, in 1856, Franklin Pierce's address discussed Britain in the contexts of her desire to dominate the Panama routes (and US refusal thereof), rights to fisheries, increasing trade between the US and British Provinces in North America, and maritime rights regarding immunity from seizure: '... *the private property of subjects and citizens of a belligerent on the high seas ... by the public armed vessels of the other belligerent, except it be contraband.*' Britain was a subject in Ruthford Hayes' 1879 address regarding the settlement of a dispute over rights to fisheries in Canadian waters. In two of the subsequent presidential addresses from the 10 documents, Britain was mentioned once in each, and not mentioned thereafter. Thus, the correct temporal order (and the TempSeq ordering method) of the 10 selected documents optimized a gradual pattern of change in the usage of the word *Britain*.

Another informative word to TempSeq analysis is *Families*, although it was not ranked highly from LDA, either as a uni-gram or as a portion of a word in a bi-gram. In the set of 10 SOTU documents, the word *Families* was barely mentioned prior to the address by Harry Truman in 1948. The frequent usage of that word appears in the later time periods. For example, in the 1948 Address, Harry Truman mentioned *Families* in the contexts of a social safety net and policies aiming to raise the standard of living for ordinary Americans. For example, we note '*public housing for low-income families*', the provision of price support for farm commodities to enable '*farm families ... to catch up with the standards of living enjoyed in the cities*' and anti-inflation measures to fight the '*undermining* [of] *the living standards of millions of families*'. In Lyndon B. Johnson's 1968 address, the word *Families* was invoked to boast about the increase in the wealth accumulation of '*most American families*' as '*more and more families own their own homes ... television sets*'. He urged congress to authorize more money to allow '*new housing units for low and middle-income families*' to be built in order for '... *thousands of families to become homeowners, not rent-payers*'. In William Clinton's 1996 Address, the word *Families* is invoked in the context of sheltering '*working families*' from the effects of government cuts. In spite of cuts and a shrinking government, Clinton states his belief in the possibilities of cultivating '*stronger families*'. Further, he speaks of the challenge to '*strengthen America's families*' and thanks his wife for having taught him '*the importance of families and children*'. Clinton also challenges '*America's families to work harder to stay together*' because '*families who stay together not only do better economically, their children do better as well*'. Furthermore, the word *Families* is invoked in the context of health insurance policies –'... *over*

one million Americans in working families have lost their health insurance'. The 2019 Address by Donald Trump invokes the word *Families* in his statement '*We passed a massive tax cut for working families*'. The word *Families* is also mentioned in the context of victims of criminal violence whom Trump had met – '*I have gotten to know many wonderful angel moms, dads and families*'.

Among the 10 selected presidential speeches, *Court* was found to be an informative word, despite not being ranked highly by LDA. Examining the bar graph in figure 7, there is an abundant usage of the word (34 times) in the 1925 address by Calvin Coolidge. Although occurring a few times in the other prior presidential speeches (except for that of James Madison), the word *Court* did not occur after this date. Emerging as a great power after the end of World War I, the United States sought a foreign policy with global influence. Calvin Coolidge invoked the word *Court* primarily in discussing his administration's support in joining the Permanent Court of International Justice, which had been set-up in 1922. In his speech, Coolidge encouraged the senate to support adherence to the Court by arguing that the United States' interests would not be negatively affected, for example in stating '*... by supporting the court we do not assume any obligations under the league ...*'; '*... the statute creating the court shall not be amended without consent ...*'; '*No provision of the statute ... give[s] [the] court any authority to be a political rather than a judicial court*', and '*If we support the court, we can never be obliged to submit any case which involves our interests for its decision*'.

The word *Court* was also invoked in the 1879 address by Rutherford B. Hayes, although to a lesser extent than compared to that of Calvin Coolidge. Here, the primary contexts of usage involved criminal offenses and court administration. In the context of criminal offences, there is the example of the practice of polygamy in Utah, which would no longer be defended under the constitutional guarantee of religious freedom, '*The Supreme Court of the United States has decided the law to be within the legislative power of Congress*'. Another instance concerns the urgency to introduce a justice system to prosecute criminals in the newly acquired territory of Alaska: '*bill authorizing ... detention of persons charged with criminal offenses, and providing for an appeal to United States courts ...*'. In the context of court administration, we find the phrases, '*The business of the Supreme Court is at present largely in arrears*'; '*... magistrates who compose the court can accomplish more than is now done*', and '*... additional circuit judges and the creation of an intermediate court of errors and appeals, which shall relieve the Supreme Court of a part of its jurisdiction ...* '.
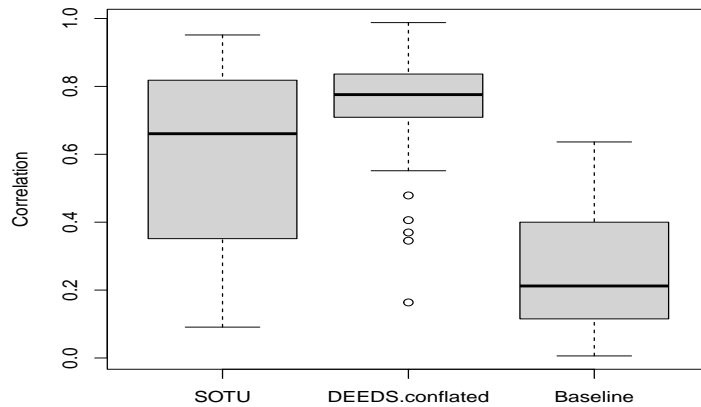
Figure 5: Box plots of the correlation coefficients (in absolute terms) of the estimated rank orders of sets of 10 documents and their true rank orders, replicated 100 times. The first plot corresponds to the State of the Union Address corpus (SOTU), the second to the DEEDS-conflated corpus, and the final plot is the baseline (random).
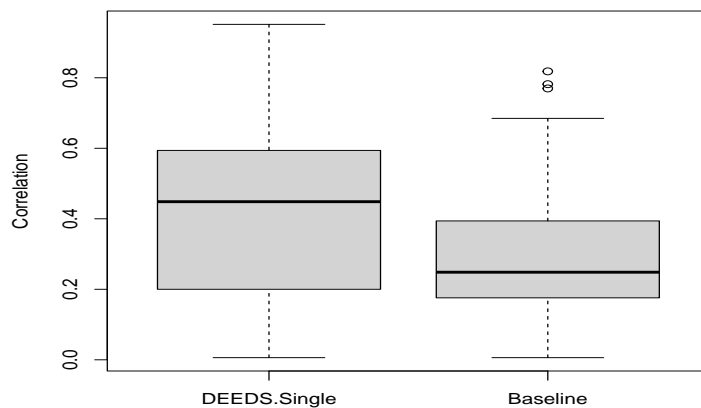

Figure 6: Box plots of the correlation coefficients (in absolute terms) of the estimated rank orders of sets of 10 documents and their true rank orders, replicated 100 times. The first plot corresponds to the DEEDS-single corpus, and the second to the baseline (random).
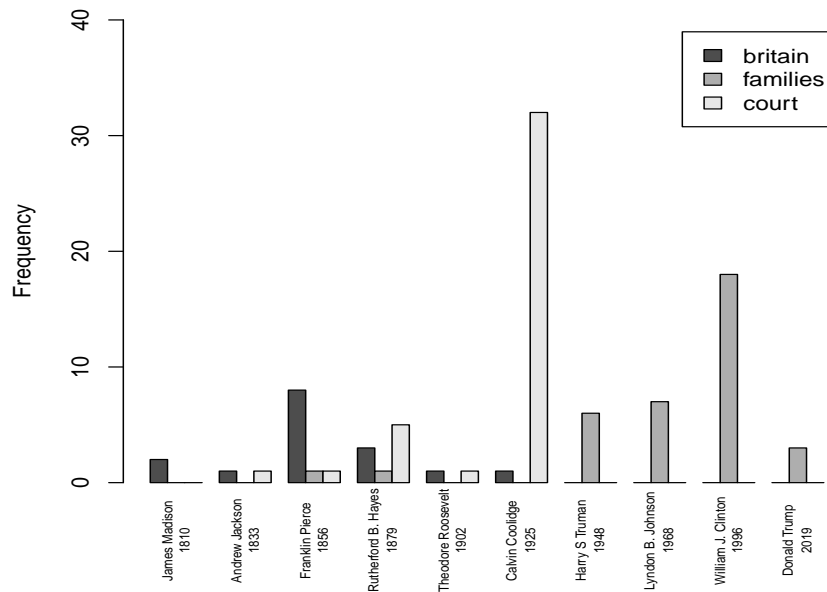
Figure 7: A bar graph of the frequencies of the usage of the words *Britain*, *Families* and *Court* in each presidential speech, indicated by year.

VII   ERROR ANALYSIS

We conducted an error analysis on the subset of the 100 replication sets of 10 randomly selected documents for which TempSeq under-performed, with a cut-off of correlation coefficients falling below the 10th percentile. For the SOTU corpus, these were the sets of documents for which correlation coefficient between estimated temporal ordering via the TempSeq method and their true temporal ordering was less than 0.27. For the richer DEEDS-conflated corpus, the corresponding threshold correlation was 0.62. When comparing the average bandwidth values, equation (4), of the estimated temporal orderings for such sets of documents to that of the average bandwidth value under their correct temporal orderings, i.e., $H_{\sigma_0(l)}$, the values of $H_{\sigma_0(l)}$ were generally larger for the latter case, as shown in figures 8 and 9. This reflects the lesser variability of word usage, and the gradual change in word frequency with time. The under-performance of TempSeq on the sets of documents under discussion is therefore explicable by the inadequate search runs of the Simulating Annealing algorithm that searches for the optimal temporal ordering in equation (4).

VIII   CONCLUSION

A natural question that arises at this point is whether employing the models used in Large Language Models (LLM), which have shown unprecedented capabilities in understanding and generating natural language, can be used to solve the temporal sequential ordering problem described above. The fundamental problem with the possible application of LLM for the research question posed in this paper is that the amount of textual data required to train an LLM is typically massive (for example, the first BERT model was trained on over 3.3B words, Devlin [2018]). By comparison, the SOTU corpus has only a total of around 1.5M words, and the DEEDS corpus has only a total of around 2M words, with the dates of documents in each corpus spread across two and half centuries. Even to employ static word embeddings, which
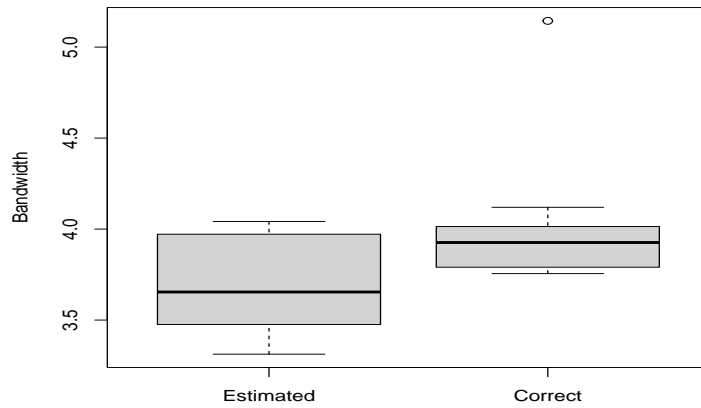
Figure 8: On the left side is a box plot illustrating the bandwidth values obtained for the State of the Union (SOTU) documents with the lowest 10th percentile correlations. On the right side is a corresponding box plot illustrating the bandwidth values obtained under the correct temporal ordering for the documents with the lowest 10th percentile correlations.
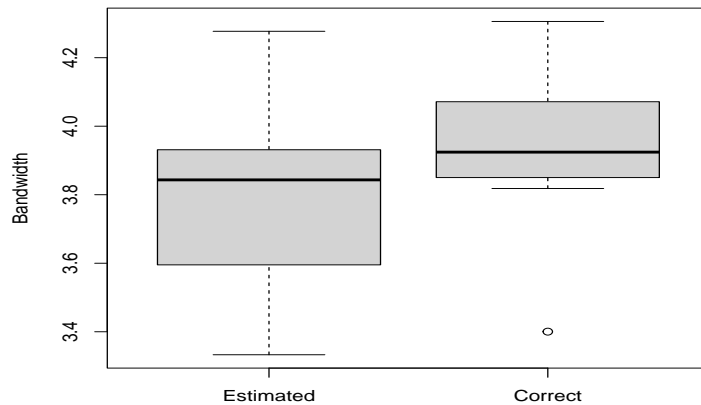

Figure 9: On the left side is a box plot illustrating the bandwidth values obtained for the DEEDS-conflated documents with the lowest 10th percentile correlations. On the right side is a box plot illustrating the bandwidth values obtained under the correct temporal ordering for the documents with the lowest 10th percentile correlations.

are ways to represent words as vectors in a multidimensional vector space, the smallest training sets required for reliable word representation using Word2Vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014] were 24M words (a subset of Google News corpus) and 1B tokens (2010 Wikipedia), respectively. If pre-trained LLM models were to be leveraged for the task at hand, those models would need to have been trained on the right kinds of corpora, in the sense of topics, genre, time periods and language. Moreover, assuming we can obtain effective word representations using LLM, it is not clear how we could use them for the task of temporal ordering of a set of documents. More pertinently, the research question posed in this paper seeks to temporally order a set of documents (for example, sets of ten documents following the TempSeq experiments in sections V and VI) when a training corpus is not necessarily available,

as is the case for many heritage texts. Far below the size of data required for LLM, a set of ten documents from SOTU has a total average count of 64K words (an average of 6400 words per Presidential Address), and a set of ten documents from DEEDS has a total average count of 1750 words (an average of 175 words per document).

Motivated by problems arising in the dating of historical and heritage texts, we set about in this paper to develop a method for assigning temporal rank orders to a sequence of documents when the date of issue is either missing or uncertain. In the historical context, the limited number and length of surviving manuscript texts presents a particular challenge in ordering. Our unsupervised method for document rank ordering relies on the principle that word usage changes gradually, on the scale of decades. Our method effectively captures changing word usage in the DEEDS and SOTU corpora by the bandwidth estimates. As shown (in section VI), the median of the correlation values for both corpora were significantly higher than the baseline from random ordering of sets of 10 documents. However, when the sizes of each of the documents are composed of fewer words, as in the case of the DEEDS-single, the TempSeq method doesn't perform as well due to the lack of adequate number of words on which to base the necessary estimates.

In practice, a reliable document rank ordering method should furnish the opportunity to identify particularly informative words for estimating the correct document orderings (for example, words with relatively high values of $h_{amise,w,\sigma(l)}$ for the optimal ordering). Equally, such a method provides us with the potential to identify anachronistic words, which may have been inserted into documents for nefarious reasons, such as the case of forged documents mentioned in section I.

In our current experimental design, we selected orderings of documents that were approximately twenty years apart and extending over 200 years; our procedures for ordering performed significantly better than the randomization. In the future, we shall examine the performance of our method when selected documents are separated at variable time intervals. We also plan to examine the temporal rank orderings of documents when the number of words in the documents is extremely limited, namely the particular case of Anglo-Saxon texts.

References

Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.

Taylor B. Arnold. United States presidential state of the union addresses, 2022. URL https://github.com/statsmaths/sotu/. R package version $>=$ 3.5.0.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Andrei Z Broder. On the resemblance and containment of documents. In *International Conference on Compression and Complexity of Sequences*, pages 21–29. IEEE Computer Society, Los Alamitos, California, 1997.

Nathanael Chambers. Labeling documents with timestamps: Learning from their time expressions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 98–106. Association for Computational Linguistics, 2012.

William W Cohen, Robert E Schapire, and Yoram Singer. Learning to order things. In *Advances in Neural Information Processing Systems*, pages 451–457, 1998.

Catherine Cubitt. Finding the forger: An alleged decree of the 679 council of Hatfield. *The English Historical Review*, 114(459):1217–1248, 1999.

Franciska De Jong, Henning Rode, and Djoerd Hiemstra. Temporal language models for the disclosure of historical text. In *Humanities, Computers and Cultural Heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, pages 161–168. Koninklijke Nederlandse Academie van Wetenschappen, 2005.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jianqing Fan, Nancy E Heckman, and Matt P Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429):141–150, 1995.

Andrey Feuerverger, Peter Hall, Gelila Tilahun, and Michael Gervers. Distance measures and smoothing methodology for imputing features of documents. *Journal of Computational and Graphical Statistics*, 14(2):255–262, 2005.

Andrey Feuerverger, Peter Hall, Gelila Tilahun, and Michael Gervers. Using statistical smoothing to date medieval manuscripts. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honour of Professor Pranab K. Sen*, pages 321–331. Institute of Mathematical Statistics, 2008.

Michael Gervers. *Dating Undated Medieval Charters*. Boydell & Brewer Ltd, 2002.

Michael Gervers and Gelila Tilahun. Topic modeling and the resolution of a medieval English diplomatic enigma. In Žarko Vujošević and Nebojša Porčić, editors, *Archives and Archival Research in the Digital Environment. A Thematic Volume.*, pages 151–170. Belgrade: University of Belgrade-Faculty of Philosophy, 2023.

Michael Gervers, Gelila Tilahun, Shima Khoshraftar, Roderick Mitchell, and Ariella Elema. The dating of undated medieval charters. *Journal of the British Records Association*, 53(136):1–33, 2018.

Alfred Hiatt. *The making of medieval forgeries: false documents in fifteenth-century England*. University of Toronto Press, 2004.

Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

Nattiya Kanhabua and Kjetil Nørvåg. Improving temporal language models for determining time of non-timestamped documents. In *International Conference on Theory and Practice of Digital Libraries*, pages 358–370. Springer, 2008.

Nattiya Kanhabua and Kjetil Nørvåg. Using temporal language models for document dating. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 738–741. Springer, 2009.

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598): 671–680, 1983.

Dimitrios Kotsakos, Theodoros Lappas, Dimitrios Kotzias, Dimitrios Gunopulos, Nattiya Kanhabua, and Kjetil Nørvåg. A burstiness-aware approach for document dating. In *Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1003–1006. ACM, 2014.

Guy Lebanon and Yang Zhao. Local likelihood modeling of temporal text streams. In *Proceedings of the 25th International Conference on Machine Learning*, pages 552–559. ACM, 2008.

Xiaoyan Li and W Bruce Croft. Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 469–475, 2003.

Inderjeet Mani and George Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 69–76, 2000.

Jon D McAuliffe and David M Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128, 2008.

Bethany McLean and Peter Elkind. The smartest guys in the room. *The Amazing Rise*, 2003.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Margaret E Roberts, Brandon M Stewart, and Edoardo M Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003, 2016.

Russell Swan and David Jensen. Timelines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining*, pages 73–80, 2000.

Matt Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108

(503):755–770, 2013.

Matt Taddy. Document classification by inversion of distributed language representations. *arXiv preprint arXiv:1504.07295*, 2015.

Ramya Thinniyam. *On Statistical Sequencing of Document Collections*. PhD thesis, University of Toronto, 2014.

Peng Tian, Jian Ma, and Dong-Mo Zhang. Application of the simulated annealing algorithm to the combinatorial optimisation problem with permutation property: An investigation of generation mechanism. *European Journal of Operational Research*, 118(1):81–94, 1999.

Gelila Tilahun. *Statistical Methods for Dating Collections of Historical Documents*. PhD thesis, University of Toronto, 2011.

Gelila Tilahun, Andrey Feuerverger, and Michael Gervers. Dating medieval English charters. *The Annals of Applied Statistics*, 6(4):1615–1640, 2012.

Gelila Tilahun, Michael Gervers, and Roderick Alexander Mitchell. Statistical approaches to the diplomatics of institutional topography. *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde*, 62(1):351–364, 2016.

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. Dating documents using graph convolution networks. *arXiv preprint arXiv:1902.00175*, 2019.

Dorothy Whitelock. *English Historical Documents I, 2nd edn*. Cambridge University Press, 1979.

A ANNEX 1

1.1 Non-parametric Kernel Regression for the Binomial Model

Suppose $(n_w(D_i), N(D_i), t_{D_i})$, $i = 1, \ldots, n$ represents a sequence of data pairs, where t_{D_i} represents the date of the $D_i{}^{th}$ document and n denotes the size of the data; that is, the total number of documents in the collection. Let $n_w(D_i)$ denote the number of occurrences of the word (or term) w in document D_i. Let $N(D_i)$ denote the total number of words (or terms) in document D_i. We are interested in modelling the probability of occurrence of the term *w* at time t. The Generalized Linear Models (GLM) for the binomial family is therefore a natural point of departure.

The GLM assume that the conditional likelihood of the response Y, given the explanatory variables X, has an exponential family form

$$f(y \mid x) = \exp\left\{ \frac{\theta(x) - b(\theta(x))}{a(\phi)} + c(y, \phi) \right\} \tag{5}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$ are known functions, but the value of the *dispersion* parameter, ϕ is not necessarily known. The parameter $\theta(\cdot)$ is called the *canonical* parameter. The conditional mean and variance of the above model can be shown to be

$$m(x) = \mathrm{E}(Y \mid X = x) = b'(\theta(x)),$$

and

$$\mathrm{Var}(Y \mid X = x) = a(\phi)b''(\theta(x))$$

In the parametric form of GLM, a function g of the conditional expectation is regressed on the variable X as

$$g(m(x)) = x^t \boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is a vector of the regression coefficients. If $g = (b')^{-1}$, then g is designated as the *link* function because it links the conditional expectation to the canonical parameter θ, such that we model $\theta(x) = x^t\beta$. Given independently observed data $\{(X_i, Y_i), i = 1, \ldots, n\}$, the β values are estimated by maximizing the joint conditional likelihood (in the form of equation (5) or equivalently, by maximizing the joint conditional log-likelihood) over the β's:

$$\max_{\beta} L(\vec{y} \mid \vec{x}) = \max_{\beta} \sum_{i=1}^{n} \log(f(y_i \mid x_i)) \tag{6}$$

In the binomial family setting, let r be the number of trials and Y the number of successes in the r trials. Let X be the predictor variable such that $Y \sim \mathrm{Bin}(r, \pi(X))$, where $\pi(X)$ is the probability of success. Our interest is

in estimating the mean of the sample proportion rather than the mean number of successes. Letting $Y^* = Y/r$, we wish to estimate $\pi(x) = \mathrm{E}(Y^* \mid X = x, r)$. Suppose $\{(X_i, r_i, Y_i), i = 1, \ldots, n\}$ are samples drawn from (X, r, Y) where $Y_i = r_i Y_i^* \sim \mathrm{Bin}(r_i, \pi(X_i))$. The form of the joint conditional log-likelihood can be written as

$$L(Y_1^*, \ldots, Y_n^* \mid (X_i, r_i) : i = 1, \ldots, n) = \sum_{i=1}^{n} \left\{ \frac{Y_i^* \theta(X_i) - \log(1 + \exp(\theta(X_i)))}{1/r_i} + \log \binom{r_i}{r_i Y_i^*} \right\}$$

where $\theta(X_i) = \log\left(\frac{\pi(X_i)}{(1 - \pi(X_i))}\right)$, $b(\theta(X_i)) = \log(1 + \exp(\theta(X_i)))$, $\phi_i = r_i$, $a(\phi_i) = 1/r_i$, and $c(Y_i^*, \phi_i) = \log\binom{r_i}{r_i Y_i^*}$ Agresti [2002] .

For the above binomial example, we model the canonical link function $\theta(x) = g(x, \boldsymbol{\beta})$ as a polynomial of degree at most p $(p \ll n)$ in the predictor variable, and $\boldsymbol{\beta}$ is the vector of coefficients of this polynomial. In viewing the above as a reformulation of equation (6), we wish to maximize (with respect to the $\boldsymbol{\beta}$'s)

$$\sum_{i=1}^{n} \log\{f(Y_i | X_i, g(X_i; \boldsymbol{\beta}))\}. \tag{7}$$

One of the deficiencies from which this model suffers is its lack of flexibility, namely that the optimal values of the $\boldsymbol{\beta}$'s are *global* — a set of parameter values over the entire domain. In the context of the problem addressed in this paper, we do not have a pre-defined idea as to the number of parameters that are necessary to model the probability of occurrence of tokens via the canonical link function (the *logit*, $\theta(x)$) as it varies over a given range of time. Our aim is to model the probability *locally* — that is, to relax the global polynomial assumption and to allow the $\boldsymbol{\beta}$'s to adjust locally within a small neighbourhood of the domain space [Fan et al., 1995].

The local modelling approach thus leads to the following new *local log-likelihood* objective function:

$$L(\boldsymbol{\beta}(x)) = \sum_{i=1}^{n} \log\{f(Y_i | X_i, g(X_i; \boldsymbol{\beta}))\} K_h(X_i - x)$$

where

$$\hat{\boldsymbol{\beta}}(x) = \arg \max_{\beta} L(\boldsymbol{\beta}(x)).$$

We define $K_h(u) \equiv h^{-1} K(u/h)$ where K is a kernel function and the scaled factor h is the associated bandwidth. The kernel function K is typically a continuous, unimodal, symmetric, and non-negative. It satisfies the condition $\int_{-\infty}^{\infty} K(x) dx = 1$ and decays fast enough to eliminate the contributions of remote data points. The kernel function could be a Gaussian distribution among many other possibilities, although in this paper, we used the t-distribution function with a low degree of freedom value (equal to 5), so as not excessively to discount distant data points. As a weight term, K_h fits a polynomial regression curve around the data in the neighbourhood of x, where h is the size of the local neighbourhood.

We build locally flexible $\boldsymbol{\beta}$'s at u in the neighbourhood of the point x for the canonical link function using the following expansion:

$$\theta(u) = g(u, \boldsymbol{\beta}) \approx \beta_0(x) + \beta_1(x)(u - x) + \cdots + \beta_p(x)(u - x)^p \tag{8}$$

where $\beta_j(x) = \frac{\theta^{(j)}(x)}{j!}$. Maximizing with respect to the $\boldsymbol{\beta}$'s, when the polynomial order is $p = 0$, which is to say the *locally-constant* regression case, we obtain

$$L(\boldsymbol{\beta}(x)) = \sum_{i=1}^{n} \left[Y_i \beta_0(x) - r_i b(\beta_0(x)) \right.$$
$$\left. + \log \binom{r_i}{Y_i} \right] K_h(X_i - x).$$

Allowing $\hat{\beta}_0(x) = \hat{\theta}_h(x)$ to maximize the above expression, we find that

$$\hat{\pi}(x) \equiv \hat{\pi}_h(x) = \exp(\hat{\beta}_0(x))/(1 + \exp(\hat{\beta}_0(x)))$$
$$= \sum_{i=1}^{n} \frac{Y_i K_h(X_i - x)}{r_i K_h(X_i - x)} \tag{9}$$

where $\hat{\pi}(x)$ is an estimate of the probability of success at x. When $p = 1$, it follows

$$
\begin{aligned}
L(\boldsymbol{\beta}(x)) \quad = \quad & \sum_{i=1}^{n} \Big[Y_i(\beta_0(x) + \beta_1(x)(X_i - x)) \\
& -r_i b(\beta_0(x) + \beta_1(x)(X_i - x)) \\
& + \log \binom{r_i}{Y_i} \Big] K_h(X_i - x)
\end{aligned}
$$

where $\hat{\beta}_0(x)$ and $\hat{\beta}_1(x)$ maximize the above equation. The maximizers can be found using numerical methods, such as that of Newton-Raphson, where the initial value of $\beta_0(x)$ is set to be the solution for the local polynomial estimator $p = 0$ together with $\beta_1(x) = 0$. The estimator $\hat{\pi}(x)$ (which doesn't have a closed form solution) is given by

$$
\hat{\pi}(x) \equiv \hat{\pi}_h(x) = \exp(\hat{\beta}_0(x))/(1 + \exp(\hat{\beta}_0(x))) \tag{10}
$$

B ANNEX 2

2.1 Bandwidth Estimation

Regarding the notations used in this section, refer to Annex A.

Fan et al. [1995] derive a rule-of-thumb bandwidth parameter h estimate for the curve $\hat{\pi}_h(x)$. Recalling that $\theta(X_i) = \log\left(\frac{\pi(X_i)}{(1 - \pi(X_i))}\right)$ (Annex A, section 1.1), the curve estimate $\hat{\pi}_h(x)$ was determined having first estimated the canonical *logit* function $\hat{\theta}_h(x)$ where the local polynomial fitting is for $p = 1$, in equation 8 of Annex A, section 1.1. The error incurred when estimating $\theta(x)$ with $\hat{\theta}_h(x)$ is measured using the asymptotic mean squared error (AMSE) criterion:

When $h \to 0$ and $nh \to \infty$, as $n \to \infty$,

$$
\begin{aligned}
\mathrm{AMSE}\{\hat{\theta}_h(x)\} \quad = \quad & \left\{ \int z^2 K(z)dz \; \times \; \frac{\theta^{(2)}(x)}{2!} h^2 \right\}^2 \\
& + \sigma^2(x; K) n^{-1} h^{-1}
\end{aligned}
$$

where

$$
\sigma^2(x; K) = \mathrm{var}(Y|X = x)^{-1} f(x)^{-1} \times \int K^2(z)dz. \tag{11}
$$

The expansion of $\mathrm{AMSE}\{\hat{\theta}_h(x)\}$ above is split into the squared bias and variance terms of $\hat{\theta}_h(x)$, reflecting the bias and variance trade-offs. As the asymptotic expansion shows, low values of the bandwidth parameter h decrease the bias at the cost of a high variance (insufficient smoothing). We also note that sparser regions of the design density $f(x)$ result in larger variance of the estimator. The unknown terms, such as $\theta^{(2)}(x)$ and $\sigma^2(x; K)$ would still need to be estimated. A convenient approach is rather to approximate the error for $\hat{\theta}_h(x)$ via the asymptotic mean integrated squared error (AMISE) defined to be

$$
\mathrm{AMISE}\{\hat{\theta}_h(x)\} = \int \mathrm{AMSE}\{\hat{\theta}_h(x)\} f(x) w(x) dx
$$

where the design density f and the weight function w are included for stability reasons. According to this error criterion, the optimal bandwidth is given by:

$$
h_{amise} = (A/B)^{1/5}
$$

where $A = \int \sigma^2(x; K) f(x) w(x) dx$ and

$$
B = \left\{ \int z^2 K(z)dz \right\}^2 \times \left\{ \int \theta^{(2)}(x)^2 f(x) w(x) dx \right\} n \tag{12}
$$

The unknown quantities, $\sigma^2(x; K)$ and $\theta(x)$ can be estimated by fitting a qth-degree polynomial parametric fit where $q \geq p + 1 = 2$. The estimated bandwidth h_{amise} provides us with a rough and quick approach to calculating a bandwidth value to use in practice.

For theoretical results related to the estimator of $\hat{\pi}_h(x)$, such as the asymptotic distribution when the bandwidth $h \to 0$ and $nh \to \infty$ (thus allowing us to create a confidence band around the estimator), and for the form of the bias and variance of the estimator when x is an interior and a boundary point, refer to Fan et al. [1995].

2.2 Bandwidth Computation

All the computations were performed using the R language and environment for statistical computing, R Core Team [2023]. For codes, refer to https://github.com/gitgelila/TempSeq.

The bandwidth value $h_{amise,w,\sigma(l)}$ from section V is estimated following the above rule-of-thumb procedure to produce h_{amise}. To compute $h_{amise,w,\sigma(l)}$, we first need to estimate the unknown quantities $\sigma^2(x; K)$ and $\theta(x)$. For a particular permutation, $\sigma(l)$, of the true temporal sequence $l = (1, 2, \ldots, 10)$ of a set of 10 documents, and on which the TempSeq method is to be run (see section V), the data has the form

$$\{(n_w(D_{\sigma(l)_i}), N(D_{\sigma(l)_i}), t(D_{\sigma(l)_i})), i = 1, \ldots, 10\}.$$

The notation $\sigma(l)_i$ identifies the ith document after permutation. $n_w(D_{\sigma(l)_i})$ counts the number of occurrences of the word w in document $D_{\sigma(l)_i}$ which has a total of $N(D_{\sigma(l)_i})$ number of words. $t(D_{\sigma(l)_i})$ is the temporal rank of the date of issue of document $D_{\sigma(l)_i}$.

Using the *glm* (generalized linear model) function from the R statistical package, a parametric second degree polynomial logistic regression was fit to $\theta(x)$. This fit was used to estimate $\theta^{(2)}(x)$ and also $\sigma^2(x; K)$. The kernel function $K(x)$ is the Student's t-density function with degree of freedom equal to 5. The weight term $w(x) \geq 0$ was set to equal $1/10$ at each of the 10 temporal positions of the independent variable, and zero elsewhere. The term $\int K^2(z)dz$ (in equation 11) was numerically computed by randomly drawing samples from the Student's t-distribution with 5 degrees of freedom. The second moment of the Student's t-distribution with 5 degrees of freedom, $\int z^2 K(z)dz = 5/3$.