Statistical Methods for Dating Collections of Historical Documents

by

Gelila Tilahun

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Statistics
University of Toronto

Abstract

Statistical Methods for Dating Collections of Historical Documents

Gelila Tilahun

Doctor of Philosophy

Graduate Department of Statistics

University of Toronto

2011

The problem in this thesis was originally motivated by problems presented with documents of Early England Data Set (DEEDS). The central problem with these medieval documents is the lack of methods to assign accurate dates to those documents which bear no date.

With the problems of the DEEDS documents in mind, we present two methods to impute missing features of texts. In the first method, we suggest a new class of metrics for measuring distances between texts. We then show how to combine the distances between the texts using statistical smoothing. This method can be adapted to settings where the features of the texts are ordered or unordered categoricals (as in the case of, for example, authorship assignment problems).

In the second method, we estimate the probability of occurrences of words in texts using nonparametric regression techniques of local polynomial fitting with kernel weight to generalized linear models. We combine the estimated probability of occurrences of words of a text to estimate the probability of occurrence of a text as a function of its feature – the feature in this case being the date in which the text is written. The application and results of our methods to the DEEDS documents are presented.

Acknowledgements

It was the vision of Professor Michael Gervers, director of the DEEDS project at the University of Toronto, to create a centralized database of digitized medieval English manuscripts so that historians could upload via the internet their undated documents in order to have them dated. The idea was to have an automated document dating system based on sound statistical methodologies. Building a high quality training dataset on which to test the statistical methodologies was not an easy task – it involved many experts of the Latin language and of medieval history, including numerous graduate students from the Medieval Studies Program, University of Toronto, working under the direction of Professor Gervers. Currently, the DEEDS database consists of over 11,000 documents. I would like to acknowledge the enormous contributions of the members of the DEEDS project – without their efforts and insight gained into the nature of these documents, the development of practical statistical methodologies would not have been possible.

I am deeply indebted to my thesis advisor, Professor Andrey Feuerverger, for constant advice and encouragement over the years, for freely sharing many valuable ideas, and for pushing me always towards the next higher level. He made me question and analyze, and re-question and re-analyze my ideas so that my thoughts ultimately could be as clear and sharp as they could be. He is also a phenomenal editor and without the input of his skills, my dissertation would not have been possible.

I am deeply indebted to Professor Peter Hall for the interest that he showed in this work, and for his contributions to the distance-based methods of Chapter 5. He showed me that an originally crude distance-based dating methodology that I had based on document distance measures could be framed as a kernel regression problem, thereby generalizing the method and obtaining better results. The consistency result

of the date estimate from this chapter is due to Professor Hall, although its exposition as supplied in this thesis, and any of its shortcomings, are my own.

I am greatly indebted to Professors David Andrews and Radford Neal for having served as a committee members prior to my taking a leave of absence during my research work, and for their very considerable encouragement over the years. In particular, both Professors Andrews and Neal were extremely generous with their time and freely offered many valuable ideas and suggestions. Many of the ideas they offered could be useful for further refining the document dating methodologies developed in this thesis and could have wider applicability in text analysis more generally.

I would like to thank Professors Nancy Reid and Keith Knight for serving as committee members subsequent to my leave of absence, and to Professor Mike Evans for participating in my final oral defense. Their comments and suggestions during the committee meetings and at the SGS oral defense were very valuable.

I would also like to thank Professor Benjamin Kedem of the University of Maryland for coming to Toronto to serve as the external examiner. His comments and questions prior to and during the oral defense were both insightful and stimulating.

I am grateful to acknowledge the financial support out of grants by the Natural Sciences and Engineering Research Council of Canada, from an Ontario Graduate Scholarship for Science and Technology, and out of a grant by Google Inc.

I am also indebted to my friend and former colleague Rodolfo Fiallos, previous member of the DEEDS project. His attempts in dating the DEEDS documents (Chapter 2) are full of insight in the ways humans use language. He is also an excellent teacher and was a devoted member of the DEEDS project.

I would also like to extend my gratitude to my good friends Jemila Hamid and Eshetu Atenafu not only for their moral support, but also for sacrificing their lunch

hours on countless occasions to discuss with me statistical issues as related (and not related) to my thesis. I wish I could be half as useful to them as they have been for me.

A heart felt thank you to Dermot Whelan who never said "no" or, "that doesn't fit my job description" or, "I'm too busy" to help me with matters related to my computer, printings etc....

A heartfelt thank you to the staff at the Statistics Department, and in particular Andrea Carter, for her encouragement to push forward, and for taking care of the endless amount of SGS paper work, emails, and what not.

To the very close friends of my youth, Monica Whalley and Jonathan "Tani" Sheps. Without your friendship, I don't know what my life would have been like, but it certainly wouldn't be half as rich. Thank you for your unfailing support and encouragement. Thank you Tani for the many interesting discussions on the similarity between problems in text analysis and problems in Biology.

Many thanks to my friend Linn Clark for looking after my daughter on so many occasions. Her efforts have allowed me to free up time to work on my research.

I am indebted to my in-laws, Carol and Ken Drudge of Komoka, Ontario. Not only are they pleasant and cheerful, encouraging and supportive; and not only were they always available to look after my children in order that I could work my thesis; and not only do they take our family on vacations, but they arrive at my home with floor-to-ceiling bookshelves, refinished oak chairs, tables, coat racks etc... all freshly built in Ken's wood working shop.

A heart felt appreciations to my mother Yemtu-Negus (Malka) Abraham for her help, especially during the last crunch of my thesis completion. On short notice, she would travel for seven hours on a Greyhound to arrive at my home in order to take over, for weeks on end, the domestic duties of cooking and cleaning and child minding.

Many thanks to my sister Bethel Tilahun for her encouragement and support, and the surprise visit on the day of my defense.

To my young daughters Tseday and Maraki Drudge. Thank you for understanding in your own ways that I had to spend so much time with my computer and books. Now you have a free Mama to spend lots of time with you.

Finally, to my partner Keldon Drudge. I am not enough of a writer to convey my gratitude to him for all the things he had to put up with in order to see me complete my thesis. He not only taught me how to program in C computer language, a necessary part of my thesis work, but his quick and razor-sharp mind was always at my disposal to help me plough through technical mathematical details, and to discuss with me many of the ideas that arose in my research. He encourages me and believes in me, supports our family financially, and spends a great deal of his time with our children. And he cooks excellent chicken cacciatore. This one is for you.

For Keldon

Contents

Chapter 1

Introduction

1.1 Background of the DEEDS Project

By studying records of property transfers, medieval historians can teach us about the social, political and economic dynamics of a particular period in history. In order to carry out such work, it is critical that original legal documents, such as deeds, or records of property holdings and transfers are accurately dated. Unlike other parts of Europe, it was not until the accession of Richard I (1189) that it became customary in England to date royal charters, and more than a century would pass until the reign of Edward II, 1307-27, before private charters routinely began to bear dates. This administrative style has left over one million undated property transfer documents (deeds) in British archives from the 12th and 13th centuries alone.

The DEEDS[1] project group founded in 1975 under the direction of Michael Gervers[2] at the University of Toronto, has been transcribing into machine readable form English charters dating from around the 11th to the 14th century from across England and

[1]DEEDS stands for Documents of Early England Data Set
[2]Professor of History, University of Toronto

Wales. These documents are written in the language of their time, Latin. The goal of this research group is the development of methodologies for dating these property conveyance documents. The DEEDS corpus, as at the time of our work, consisted of over 6000 documents from England and Wales, although only 3353 of them are used in the data analysis of this thesis[3].

The DEEDS corpus consists of documents registering property conveyances. Each document describes rights and obligations pertaining to a property. The documents used in the analysis of this thesis span almost 400 years, the earliest being dated 1089, and the latest 1466. Traditionally, there have been two types of methods applied to the dating of DEEDS manuscripts, commonly referred to as "external" and "internal" methods. Documents bearing dates within the text itself (about 5% of the extant corpus) are considered to be internally dated. Even then, one cannot be sure whether such a date represents the actual date when the document was drafted, or rather, for example, the date of the property transfer, or the date in which this transfer was recorded, or in the case of copies, the date on which the copied document was registered in a cartulary or deed book. (See Gervers, 2000, p. 4). The method of external dating refers to the use of external evidence, for example, person or place names in a dated document, compared to those in an undated one can assist in assigning a date to the undated document. However, unless there happens to be a reference to a particular datable event in the document, this method of document dating has some significant drawbacks. For instance, there are cases in which lists contain names that are closely replicated in preceding or in succeeding generations. As noted by Gervers (Gervers, 2000, p. 14), many of the names do not have identifiable counterparts elsewhere, and even if such counterparts are found, there is no sure way

[3]There were only 3353 documents available to us at the start of this study.

of knowing if one is dealing with the same person or only with a namesake. Gervers (Gervers, 2000, p. 16) gives an example in which he used name association to date a document only to discover many years later that his estimate was off by at least 40 years – the problem being that the names used in that document were largely similar to those in another witnessed by the grandchildren of the individuals appearing in the earlier document. It is a well known custom of the English high middle ages that Christian names succeeded from generation to generation, as for example in the names "William son of William son of William son of William", or "William son of Richard son of William son of Richard" (Gervers, 2000, p. 18). Such a tradition was "so widespread that some village rosters at least would appear to have been remarkably similar from one generation to the next for upwards of a hundred years and probably more" (Gervers, 2000, p. 18). Using name association in the dating of documents can therefore frequently be very misleading.

It has been suggested that studying the forms of the scripts (palaeography) in which the DEEDS documents are written could be of significant help in the dating process. However, this is not really an option since the vast majority of those documents have survived only as copies. For similar reasons, sigillography (the study of seals) is useful only in those rare instances where documents have survived in the original along with their seals (Gervers, 2000, p. 18).

When scholars attach imprecise circa dates to such documents, the circa dates tend to take on a historical value of their own. Subsequent historians are prone to use them as precise dates, which leads to even greater inaccuracies when they calculate dates for documents which they may themselves be editing. The result is a compounding of errors. One of the ways to remedy the problems that arise from dating by internal and external methods, is to study the similarities of phrase or word patterns between undated charters and dated ones, for it is well known to medieval

historians that the usage, form and content of the language of medieval documents were constantly changing over time. Gervers (Gervers, 2000, p. 19) cites the example of the phrase *amicorum meorum vivorum et mortuorum* ("of my friends living and dead") which had a life span of some 90 years (from 1150 to 1240) in the DEEDS database. Such forms of language in grant offers are in 'vogue' for a certain period of time and then eventually die out, so that the occurrence of such a phrase in an undated document can be very valuable in the dating process (Gervers, 2000, p. 19).

In the case of DEEDS, the dating of the documents was done very carefully, in that only those documents that could be dated to within a year of the time of issue form part of the data set. Thus, only accurately dated documents form the DEEDS corpus. The purpose of the present thesis is to develop statistical methodologies to estimate the dates of undated documents by comparing phrases and word patterns from the undated ones to those occurring in the dated ones.

As already mentioned, 3353 DEEDS documents were used for the data analysis in this thesis. These documents, written in Latin, come from already published sources. Those transcriptions have not been checked against the originals. This means that if editors mistakenly left out certain words from certain documents, it is the revised documents that form part of the corpus. Moreover, a large number of the documents in cartularies were themselves copied from earlier charters, and – as would modern day editors – medieval scribes may have abbreviated, left out, or replaced some phrases with those more appropriate to their time.

There is also reason to suspect that some variation can be attributed to the type of document, such as the place of issue, as well as the issuer (for instance, a member of the clergy), and the institution responsible for its production. For these reasons, every document in the DEEDS data set is accompanied with descriptive data under the following headings:

a) Document

b) Person

c) Property/Compensation

d) Lease

e) Relation

f) Linkage

g) Role

Under the header 'Document', data on the document type is noted, such as whether the document is a transfer or an agreement, the place where the transaction took place, and type of seal used. Under the header 'Person', information, such as the name to whom the document was issued, the modern equivalent of the first and last names, occupation, nationality, etc., is recorded. Under the header 'Property/Compensation' information, such as the type and value of the property, location of the property, and the number and quantity of the property are listed. Under the header 'Lease', temporal information related to the property, such as duration of an agreement, etc., is listed. While information for Document, Person, Property/Compensation, and Lease were extracted from the original documents, the last three headers, 'Relation', 'Linkage', and 'Role', express connections between individuals, between properties, or between people and properties, respectively, as determined by historians. For details, see Gervers *et. al.* (1990).

The data analysis in this thesis uses only the documents as they occur in the database and does not make use of the above mentioned additional information.

We now give an example of what one of the documents in the DEEDS database looks like. Numbers are placed between two semi-colons, and punctuation and paragraphing have been omitted. The first line consists of a file identification number for the document followed by the date (year) of the document.

"00640214","1237",

Haec est finalis concordia facta in curia domini regis apud Westmonasterium a die S Johannis Baptistae in !xv! dies anno regni regis Henrici filii regis Johannis !xxi! coram Roberto de Lexinton Willelmo de Eboraco Ada filio Willelmi Willelmo de Culewurth justitiariis et aliis domini regis fidelibus tunc ibi praesentibus inter Johannem Baioc quaerentem et Robertum Sarum episcopum et capitulum deforciantes per Radulfum de Haghe positum loco ipsorum ad lucrandum vel perdendum de advocatione ecclesiae de Waye Bayouse unde assisa ultimae praesentationis summonita fuit inter eos in eadem curia scilicet quod praedictus T recognovit advocationem praedictae ecclesiae cum pertinentiis esse jus ipsorum episcopi et capituli et ecclesiae suae Sarum ut illam quam idem episcopus et capitulum Sarum habent de dono Alani de Baiocis patris praedicti Johannis cujus haeres ipse est et idem episcopus et capitulum praedictum concesserunt pro se ob successoribus suis eidem Johanni ut eidem ecclesiae quotiescunque tota vita ipsius eam vacare contigerit possit idoneam personam praesentare ita quod quicunque pro tempore fuerit persona ejusdem ecclesiae ad praesentationem ipsius Johannis reddet singulis annis praedictis episcopo et capitulo sex marcas argenti de praedicta ecclesia apud Sarum nomine pensionis scilicet ad festum S Michaelis !xx! solidos ad Natale Domini !xx! solidos ad Pascha !xx! solidos ad nativitatem beati Johannis Baptistae !xx! solidos et post decessum ipsius Johannis advocatio praedictae ecclesiae cum pertinentiis remanebit praedictis episcopo et capitulo Sarum et eorum successoribus quieta de haeredibus ipsius Johannis in perpetuum Et praeterea idem episcopus et capitulum praedictum concesserunt pro se et successoribus suis quod ipsi de caetero invenient unum capellanum divina celebrantem singulis diebus anni in capella beati Johannis sita infra parochiam de Waye pro anima praedicti Johannis et pro animabus haeredum suorum et antecessorum suorum et pro cunctis fidelibus in perpetuum et idem episcopus et capitulum praedictum et successores sui invenient ornamenta libros et luminaria sufficientia in eadem capella in perpetuum

1.2 Description of the data

In this section, we provide descriptive statistics of the DEEDS documents. Figure 1.1 is a histogram for the dates of the 3353 DEEDS documents available to us. It shows that most of the documents date from the mid 12th century to the 14th century, after which the dating of private charters became commonplace. The mean date of the DEEDS charters is 1247, and the standard deviation is 46 years.

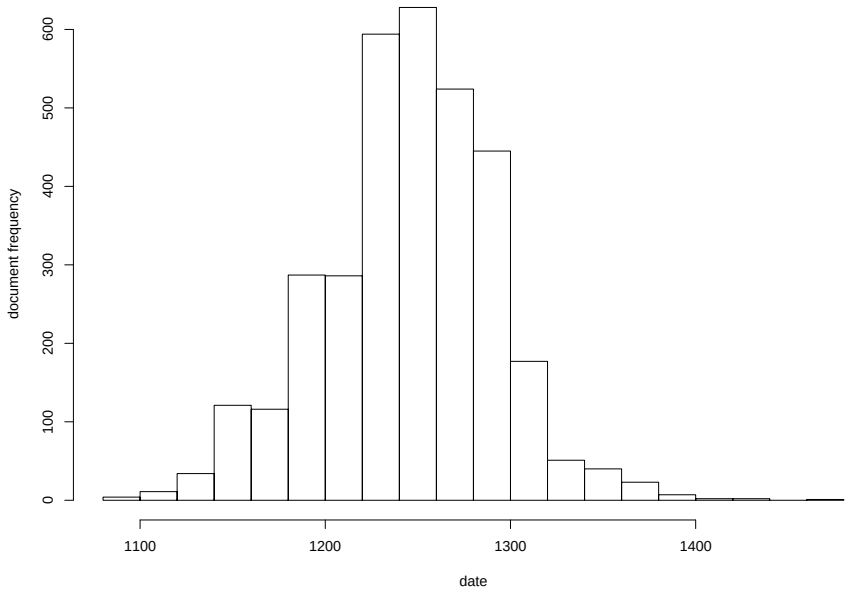


Figure 1.1: DEEDS documents distribution of dates presented as a histogram. The total number of documents is 3353.

Among these 3353 documents, there are a total of 50,006 distinct words. Of these, 28,282 (approximately 56%) occur only once, and are therefore not useful in the dating methodologies based on distance and maximum prevalence (Chapter 5 and Chapter 6, respectively). Table 1.1 lists the percentage of times that words occur relative to words that have occurred at least twice. We see, for example, that among words

that have occurred more than once, words that have occurred only two or three times constitute about half (33% + 15%) of them.

Table 1.1: Frequency of occurrences of the distinct words

Word frequency	*As a percentage of the words that have occurred more than once*
28,282 occur only once	
7223 occur twice	33%
3265 occur three times	15%
4952 occur more than 10 times	23%
3094 occur more than 20 times	14%
1644 occur more than 50 times	7.5%
1004 occur more than 100 times	4.6%
264 occur more than 500 times	1.2%
109 occur more than 1000 times	0.5%

From Figure 1.2, which plots the document lengths of the DEEDS manuscripts against their dates, we can see that there is no clear relationship between the length of a document and its date. Some quantiles for the document lengths are:

minimum length = 10 words

1st quartile length = 150 words

median length = 201 words

mean length = 232.4 words

3rd quartile length = 273 words

maximum length = 984 words.

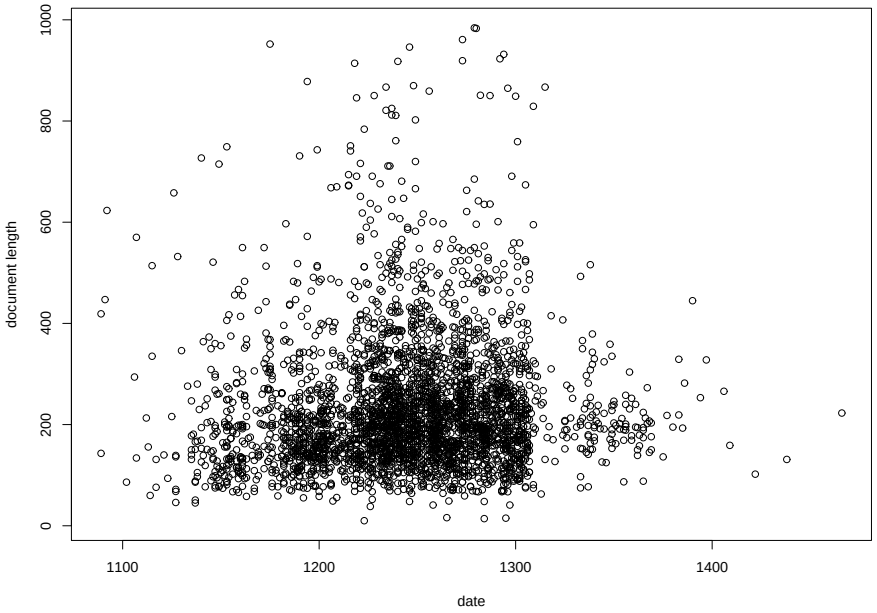


Figure 1.2: Plot of date versus length of the DEEDS documents.

1.3 Outline of thesis

This thesis develops statistical methodologies to estimate the dates of the DEEDS documents which may, in turn, be used to estimate the dates of undated similar documents. This Chapter provided the historical background for the DEEDS documents data set, as well as some descriptive statistics concerning the properties of the documents themselves. In Chapter 2, we describe some earlier attempts to estimate the dates of the DEEDS documents, and in particular the work of the DEEDS project group. In Chapter 3, we discuss basic concepts and methods from the branch of computer science known as *information retrieval*, which focuses on methodologies used in the retrieval of documents relevant to a particular query or search term. The notion of distance metrics between two documents is introduced there and a new class of metrics for measuring distances between documents, generalizing previous distance

measures between documents in the literature, is presented. (These new classes of distance measures will be used in Chapter 5). Chapter 4 discusses some relevant statistical tools which are used in Chapter 6 (and to some extent in Chapter 5.) In particular, the concepts of kernel density estimation and local polynomial regression as it applies in the framework of generalized linear models (local GLM), are discussed.

Chapters 5 and 6 contain the main original work of the thesis. In Chapter 5, we adapt ideas from information retrieval to measure the "distances" between the documents in a training set and a particular document that we wish to date, and we use kernel smoothing on the weighted distance information of the elements of the training set to construct a date estimator. A fundamentally different dating method is described in Chapter 6. It involves using local kernel smoothing in the GLM to estimate the probabilities of occurrences of word sequences as a function of time, and then uses these probabilities to estimate the most likely date in which a document was written. We call this the "method of maximum prevalence". We are later able to measure and compare the accuracy of both of our dating methods since we do know a priori the "true" dates of the DEEDS documents. We are also able to compare the accuracy of our two methods to the accuracy that has previously been achieved by the DEEDS project team.

In the concluding Chapter 7, we broaden the scope to describe work that involves the use of statistical methods for text categorization, such as computational linguistics, and authorship assignment, as studied by Mosteller and Wallace (1963) in the case of the federalist papers. In the spirit of that work, we also examine other methods of extracting informative or useful words, such as the use of regression trees. These ideas can be viewed as ways to extend the methods of Chapters 5 and 6.

Chapter 2

Previous Analyses of the DEEDS Data

Prior to our involvement in the DEEDS project, algorithms for dating undated documents were developed by personnel in the DEEDS group based on carefully considered ad hoc procedures. Although these methods were developed by persons not trained in statistics, these nevertheless were methods having considerable inherent interest, and in certain aspects actually went beyond what persons with statistical training may have been able to come up with. In this chapter, we will describe one of these methods, developed by Rodolfo Fiallos, and for which sufficient documentation has survived to allow us to give a relatively accurate description. Our reference for this chapter is Fiallos (2000) and personal communications.

2.1 The method of Rodolfo Fiallos

The algorithm for the dating procedure that had been developed by Rodolfo Fiallos of the DEEDS group employs the matching of word and phrase patterns (i.e. particular

sequences of consecutive words, also referred to as 'matching patterns') between a validation data set and those of a training data set. The underlying assumption of this dating procedure, as indicated in the previous chapter, is based on the belief that there should be a relatively high concentration of matching patterns near the true date of a document in question. The characteristics of matching patterns believed to be most important for the dating process according to Fiallos are:

- Length: the number of words in a matching pattern.
- Lifetime: the difference, in years, between the last and the first occurrence of the matching pattern. If a matching pattern occurs in one given year only, then its 'lifetime' is assigned the value 0.
- Currency: the ratio of lifetime of the pattern to the number of distinct years in which the matching pattern has occurred. Currency may be thought of as measure of the average number of lifetime-years per occurrence of a matching pattern.

A function based on these three variables was constructed to provide a numerical value of the 'importance' of a matching pattern. This function will be described in the following subsection

We note here that prior to implementation of the algorithm, a process of modification and standardization was applied to all of the documents in the training and validation sets in order to avoid break-up of word patterns due solely to differing numeric expressions, differing person or place names, and/or minor differences in spelling. For instance, all 'v's were replaced with 'u's, and all numeric expressions were replaced with '#'s. Also, all punctuation was eliminated, and names were replaced with 'P's.

2.2 Description of the algorithm

The method developed by Rodolfo Fiallos of the DEEDS group was as follows. Denote by $\mathcal{D}$ the document that we wish to date, and suppose that it consists of an ordered sequence of n consecutive words $\{W_1, \cdots, W_n\}$. It is understood that we have available a training set of T documents whose dates are known to us for use in the procedure we shall describe. A 'matching pattern' of length k is defined to be a sequence of k consecutive words, $\{W_i, W_{i+1}, \cdots, W_{i+k-1}\}$ for $i = 1, 2, \cdots, n-k+1$. For every 'matching pattern' in the document $\mathcal{D}$, ranging in length from 1 to n, a number 'MT' (which stands for 'Multiplicador Total' in Spanish, or 'Total Multiplier' in English) is computed. MT is a function of what was assumed (as in the previous subsection) to be the three most important characteristics of a matching pattern – namely length, lifetime and currency. It is defined as

$$MT = M_1(\text{length}) \times M_2(\text{lifetime}) \times M_3(\text{currency})$$

where M_1, M_2 and M_3, respectively, are particular functions of the pattern's length, lifetime and currency. Moreover, these functions are defined in such a manner that the larger the MT value of a pattern is, the more informative it is considered to be for the dating process. The definitions of M_1, M_2 and M_3 used by Rodolfo Fiallos are as follows:

$$M_1 = 1 \qquad \text{if length} \leq 3,$$

and

$$M_1 = 1 + C_0 \times (\text{length} - 3) \qquad \text{if length} > 3$$

where C_0 is a number allowed to take on values in the set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Both the cut-off point for length in the definition for M_1, and the number C_0 are

determined by trial and error. The purpose of the function M_1 is to give more weight to the longer matching patterns.

Next,

$$\begin{aligned}
M_2 &= C_1, && \text{if lifetime} = 0 \text{ (i.e. the pattern does not occur elsewhere)} \\
M_2 &= 1, && \text{if } 0 < \text{lifetime} \leq C_2 \\
M_2 &= (-0.9 \times \text{lifetime} + C_3 - 0.1 \times C_2)/(C_3 - C_2), && \text{if } C_2 < \text{lifetime} \leq C_3 \\
M_2 &= 0.1, && \text{if lifetime} > C_3
\end{aligned}$$

where C_1 takes on values in $\{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, C_2 takes values in $\{10, 20, 30, 40, 50\}$, and C_3 takes on values in $\{40, 60, 80, 100, 120\}$. Here again, the various cut-off values C_1, C_2 and C_3 are determined by trial and error. Note that as lifetime of the pattern increases, the smaller the value of the function M_2 becomes. This is because it is supposed that matching patterns having longer lifetimes are less informative since they are more spread out and not clustered around a particular date.

Finally,

$$\begin{aligned}
M_3 &= (1/32) \times \text{length}, && \text{if currency} = 0 \\
M_3 &= 1.0, && \text{if } 0 < \text{ currency } \leq C_4 \\
M_3 &= (-0.5/C_4) \times \text{currency} + 1.5, && \text{if } C_4 < \text{currency} \leq 2 \times C_4 \\
M_3 &= (-0.4/C_4) \times \text{currency} + 1.3, && \text{if } 2 \times C_4 < \text{currency} \leq 3 \times C_4 \\
M_3 &= 0.1, && \text{if currency} > 3 \times C_4
\end{aligned}$$

where

$$\text{currency} = \frac{\text{lifetime}}{\text{number of distinct years in which pattern occurs}},$$

and C_4 takes on values in $\{2, 3, 4, 5, 6\}$. Here again, the cut-off value C_4 is determined through trial and error. The larger the value of currency (i.e. the more sparse the

occurrence of the matching pattern is over the years of its lifetime) the smaller is the corresponding M_3 value. When no matching word pattern exists in the training set over which this algorithm is applied, the lifetime = 0, and therefore the value of currency = 0. Note that since the absence of a matching word pattern in the training set is presumed to most likely be due to its length (i.e. longer 'matching patterns' are less likely to be found in the training set), M_3 is still set to take on a small fraction of the value of the length in cases where the pattern does not occur in the training set (i.e. when the currency = 0).

Once the MT values have been computed for all of the 'matching patterns' in the document $\mathcal{D}$, those MT values are summed within each of the years for which training data are available. Specifically, for each year for which training data are available, the MT values of all the 'matching patterns' in $\mathcal{D}$ which also occur in that year are summed, thereby resulting in an overall function of MT over time. As an example, we refer to Gervers (1998) in which the 'matching pattern' *dictum redditum cum omnibus pertinentiis suis* is discussed. For a certain choice of C_0, C_1, C_2, C_3 and C_4, this pattern has an MT value of 1.60. Now this pattern occurs in the training set in year 1275. The MT value at year 1275 is therefore incremented by the amount 1.60, and the MT value at year 1275 will keep increasing if more matching word pattern are found in that year corresponding to patterns occurring in the document $\mathcal{D}$. In this way, the values of MT, for the document $\mathcal{D}$ is built up as a function over time.

Actually, not all the matching patterns were used in the Fiallos' process; only matching patterns with MT values higher than some threshold value, say 30, were used. The rationale for this truncation was to reduce noise arising from the MT values of patterns considered to be relatively uninformative.

Now, one also needs to account for the fact that the number of available training documents varies over time. Therefore, prior to assigning the year corresponding to

the highest value of MT as the most likely date for the document $\mathcal{D}$, the values of MT are each first divided by the number of documents in the training set for each of the years in which training data are found, so as to standardize the MT values; these new MT values are referred to as 'Global MT', or GMT values for short. Furthermore, the dates can be collapsed (i.e. grouped) into given date ranges – for example, the year axis can be divided into 20 year periods, say, and the GMT value for each such period would then be taken as the sum of the GMT values over the years in the period. The advantage of this process of collapsing of dates is that it provides a less variable global picture of the overall relationship between date and GMT. Once the date range that corresponds to the highest value of GMT has been identified, we may expand that range by 10 years to either side of the optimal date range to create a region of dates now called a "search zone" which is thought to be the range of dates most likely to contain the true date of the document $\mathcal{D}$. Then, for such a given search zone (centred at the optimal date range), the date range is then reduced from the initial 20 year period to, say, 10 years, and new GMTs are computed over the shorter intervals. A new optimal date range is thus identified and a new search zone is then created by expanding by 5 years on either side of this new date range. Finally, by reducing the size of the search zone to a single date range (in this instance, the target was to date the document to within a five year period), we finally obtain the estimated date range (which is a 5 year period) for the age of the document $\mathcal{D}$. Figure 2.1 which plots actual versus estimated dates illustrates the performance of Fiallos' method on 1484 documents based on a training set of approximately 3500 documents. The mean absolute error for the method was found to be 16 years.

Finally, we should mention that all of the constants indicated above were chosen via a leave-one-out cross-validation type of procedure applied to 1484 of the 3500 documents. Specifically, if T is the total number of documents in the available col-

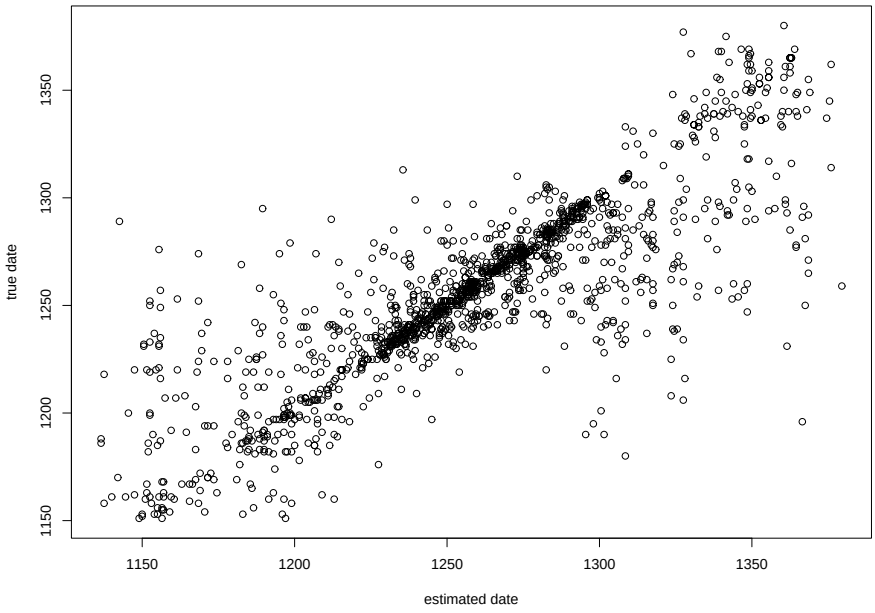


Figure 2.1: Estimated dates by Fiallos' method versus the true dates for 1484 documents selected from within a data set of approximately 3500 documents. The mean of the absolute error is 16 years.

lection, the dates of 1484 of the documents were estimated based on the remaining $T - 1$ training documents. Furthermore, these same 1484 documents were used in producing Figure 2.1. Thus, since no fresh "test set" was used to produce it, the MAE estimate of 16 years is likely to be optimistic.

Chapter 3

Some Previous Work on Text Analysis

3.1 Computational linguistics

There is a substantial technical literature pertaining to the analysis of textual documents. We review a number of these methods in this chapter, including computational linguistics, authorship assignation, and information retrieval. In particular, in our discussion of information retrieval we review set theoretic measures of document similarity which will be used in one of the methods that we will propose.

3.1.1 n-gram models

Statistical language models are used to estimate the distribution of natural language phenomena for the purposes of speech recognition, handwriting recognition, machine translation and spelling correction. Essentially, statistical language models estimate the probability distributions of various linguistic units, such as words, sentences, and whole documents. Since there are a large number of parameters to estimate, the

availability of a large volume of training data is crucial.

One of the fundamental questions that arises in language modelling is the extent to which one needs to use knowledge about the language which is being analyzed. Remarkably, it turns out that some of the most successful techniques in language modelling are devoid of any knowledge of linguistic structure or language theory, and what is essentially being analyzed, are just sequences of words which might as well be sequences of arbitrary symbols (Rosenfeld, 2000). Below we present one of the basic concepts of statistical language modelling, the so called n-gram model. We used Chen and Goodman (1998) as our main reference here.

We first illustrate the n-gram model for the case $n = 2$. (Note that 2-grams are also called *bigrams*). If a sentence s is composed of an ordered sequence of l words $\omega_1, \omega_2, \ldots, \omega_l$, then (if viewed in the appropriate context) the probability of the occurrence of this sentence $P(s)$ can be expressed as

$$\begin{aligned} P(s) &= P(\omega_1)P(\omega_2|\omega_1)\cdots P(\omega_l|\omega_1,\ldots,\omega_{l-1}) \\ &= \prod_{i=1}^{l} P(\omega_i|\omega_1,\ldots,\omega_{i-1}) \end{aligned}$$

where we define $P(\omega_1) \equiv P(\omega_1|\omega_0)$. In a bigram, it is assumed that the probability of a word only depends on the word immediately preceding it. Thus, under the bigram model we have (essentially)

$$P(s) = \prod_{i=1}^{l} P(\omega_i|\omega_1,\ldots,\omega_{i-1}) = \prod_{i=1}^{l} P(\omega_i|\omega_{i-1}) \ . \tag{3.1}$$

To estimate $P(\omega_i|\omega_{i-1})$, we let $c(\omega_{i-1}\omega_i)$ denote the number of times the bigram $\omega_{i-1}\omega_i$ occurred in the available training text, and we take, as our estimate

$$\hat{P}(\omega_i|\omega_{i-1}) = \frac{c(\omega_{i-1}\omega_i)}{\sum_\omega c(\omega_{i-1}\omega)}. \tag{3.2}$$

More generally, in the n-gram model, the conditional probability of a word depends only on the $n-1$ words preceding it. For an n-gram model (where $n > 2$), the

analogous form of equations (3.1) and (3.2) respectively are

$$P(s) = \prod_{i=1}^{l} P(\omega_i|\omega_1, \ldots, \omega_{i-1}) = \prod_{i=1}^{l} P(\omega_i|\omega_{i-n+1}^{i-1})$$

and

$$P_{ML}(\omega_i|\omega_{i-n+1}^{i-1}) \equiv \hat{P}(\omega_i|\omega_{i-n+1}^{i-1}) = \frac{c(\omega_{i-n+1}^{i})}{\sum_{\omega} c(\omega_{i-n+1}^{i-1}\omega)} \tag{3.3}$$

where ω_i^j is defined to be the string $\omega_i \ldots \omega_j$, and where we take ω_0 to be $<$ bos $>$ (i.e. beginning of sentence) and ω_{l+1} to be $<$ eos $>$ (i.e. end of sentence). In the linguistics literature, the estimate in equation (3.3) is referred to as the *maximum likelihood estimate*, or *ML* for short.

3.1.2 Smoothing the n-gram estimates

We may see from equation (3.1) that if any of the terms in the right hand-side, $P(\omega_i|\omega_{i-1})$, were estimated to equal zero, then $P(s)$ would be estimated to equal zero. For example, in speech recognition, the goal is to find a sentence s that maximizes $P(s|A) \propto P(A|s)P(s)$ for a given signal A. If the prior $P(s)$ is estimated as zero, then $P(s|A)$ will equal zero no matter how unambiguous the acoustic signal is (Chen and Goodman (1998)). In order to address such difficulties, *smoothing* techniques for the ML estimates have been developed.

Here smoothing refers to a technique designed to smooth out the peaks and troughs of the ML estimates of equation (3.3) so that high probabilities are adjusted downwards and low probabilities are adjusted upwards. It also prevents word strings from being assigned a zero probability. Furthermore, when probabilities are estimated from small counts, smoothing has the potential for improving estimation significantly (Chen and Goodman, (1998)).

As an example, consider one of the simplest smoothing methods called *Additive Smoothing*. In this method, under an n-gram model, we add a small δ, typically $0 < \delta \leq 1$, to the observed count of each word, yielding

$$\hat{P}_{add}(\omega_i|\omega_{i-n+1}^{i-1}) = \frac{\delta + c(w_{i-n+1}^{i})}{\delta|V| + \sum_{\omega} c(\omega_{i-n+1}^{i-1}\omega)}.$$

Here V, the vocabulary, is the set of all words in the corpus. In practice, $|V|$ is often limited to 10,000 words or less, and words not found in the vocabulary are mapped to a single distinguished word called '*unknown word*'. Note that if $\delta = 0$ we recover the ML estimate. In practice, Additive Smoothing performs poorly. There are however many other smoothing methods in the literature with superior performance which involve interpolation among the higher- and lower-order n-gram models, such as methods proposed by Jelinek and Mercer (1980), Witten and Bell (1991), and Kneser and Ney (1995). Many other popular ones are also discussed and their performance evaluated in Chen and Goodman(1998). Further references in the areas of language modelling can be found in Charniak (1993) and Manning and Schutze (1999).

3.2 Authorship assignation: The Federalist Papers

The Federalist Papers are a series of essays that were published anonymously (under the name Publius) from 1787 to 1788 and were written by Alexander Hamilton, James Madison and John Jay. They were intended to persuade the citizens of New York to ratify the Constitution. Of the 77 essays that were published, it is known that 5, and only those 5, were written by Jay, that 43 were written by Hamilton, and 14 by Madison. Three of the essays are known to have been jointly written by Hamilton and Madison. The authorship of the remaining 12 essays, however, which were written by either Hamilton or Madison, is in dispute.

To decide these authorships, Mosteller and Wallace (1963) carried out a statistical study of word usage using the known writings of Hamilton and Madison. The words they studied were non-contextual and therefore their rates of use was expected to be nearly invariant under changes of topic. Examples of such words, called function words, are articles, pronouns and prepositions, such as *the, an, while, whilst, also, upon*, etc. Mosteller and Wallace then used the rates of word usage in the known works of Hamilton and Madison to infer the authorship of the disputed papers. We will briefly describe their methodology.

Mosteller and Wallace used Bayesian methods to analyze word usage. They assumed word usage follows a Poisson distribution, and moreover assumed that word occurrences are independent of each other. For example, consider the usage of the word *also*. They assume the use of this word in an essay is adequately represented by a Poisson distribution, with parameters $w\mu_H$ and $w\mu_M$, where w is essay length in thousands of words, and the μ's are the rates per thousand for each of Hamilton and Madison respectively. To illustrate, assume the rates are known to be $\mu_H = 0.25$ and $\mu_M = 0.5$. For a paper of length $w = 2$ (i.e. 2000 words), Table 3.1 below gives the probabilities for the usage of the word *also* by each author.

Suppose now that the word *also* occurs four times in a paper of 2000 words known to have been written by either Hamilton or Madison. To evaluate the odds that the paper was written by Hamilton, let p_1 and $p_2 = 1 - p_1$ be prior probabilities for the hypothesis 1 and hypothesis 2, where hypothesis 1 is that Hamilton wrote the paper, and hypothesis 2 is that Madison wrote it. Let $f_i(x), i = 1, 2$ be the conditional probability of observing that the number of time the word *also* occurs is x, given that hypothesis i is true. From Bayes' theorem,

$$\text{Odds}(1, 2|x) \equiv \frac{P(\text{hypothesis } 1|x)}{P(\text{hypothesis } 2|x)} = \left(\frac{p_1}{p_2}\right)\left(\frac{f_1(x)}{f_2(x)}\right) \tag{3.4}$$

Table 3.1: Table for Poisson probabilities with $w\mu_H = 0.5$ and $w\mu_M = 1.0$.

Frequency	Hamilton	Madison
0	.607	.368
1	.303	.368
2	.0758	.184
3	.0126	.0613
4	.00158	.0153
5	.000158	.00307
6	.0000132	.000511

Adopted from Mosteller and Wallace (1963), p. 289.

which maybe interpreted as

$$\text{Final odds} \;=\; \text{Initial odds} \times \text{Likelihood ratio}.$$

Next, we can extend equation (3.4) to more than one word count, say to words having counts $x_1, \ldots, x_n$. Assuming independence among word occurrences, and taking the logarithms leads to

$$\log(\text{final odds}) \;=\; \log(\text{initial odds}) + \sum_{i=1}^{n} \log\left(\frac{f_1(x_i)}{f_2(x_i)}\right) \;. \tag{3.5}$$

Under the Poisson model, this log-likelihood ratio equals

$$\sum_i x_i \log(\mu_{iH}/\mu_{iM}) - w(\mu_{iH} - \mu_{iM})$$

where μ_{iH} and μ_{iM} have the obvious meanings and w is document length in thousands of words. Note that in (3.4) and (3.5) we are somewhat abusing notation by not indicating the dependence of f_1 and f_2 on the specific words.

The initial odds p_1/p_2 represent any prior beliefs about the authorship. For example, if it seems very certain that Hamilton wrote the paper (say, $p_1 = 0.99$ and

$p_2 = 0.01$), then the initial odds for Hamilton would be $p_1/p_2 = 99$. The likelihood ratio would be $f_1(4)/f_2(4) \approx 0.1$, if it were based only on the word *also*, and so the final odds are:

$$\text{Odds(Hamilton, Madison}|x = 4) = 99 \times 0.1 = 9.9,$$

or about 10 to 1 in favour of Hamilton. If our prior belief concerning the authorship is just a toss-up, then the authorship of the paper will just be the likelihood ratio, i.e. 10 to 1 in favour of Madison. Mosteller and Wallace argue that although the initial odds do vary from person to person, for example, the prior beliefs held by different historians, the strength of the data, as evidenced through the likelihood ratio factor, would typically overwhelm most of the subjectivity in the initial odds. They state that a likelihood ratio of 10^{-6}, for example, would convert an initial odds of 10^3 for Hamilton to final odds of 10^{-3} (i.e. 1000 to 1 for Madison). For this reason, Mosteller and Wallace's study was focused on methods for the evaluation of the likelihoods.

In computing the likelihood ratios, Mosteller and Wallace used the parameters μ_H and μ_M for the different words, but how those parameters are determined is an important issue. Although at first there seem to be vast numbers of words on which to base estimates – 94,000 written by Hamilton, and 114,000 by Madison, large sample theory for estimating μ_H and μ_M does not work well since the occurrences of the individual words are each relatively small. Moreover, we cannot be confident that the rate of word usage remains constant from one paper to another. For that reason μ_H and μ_M are estimated using Bayesian method once again.

For each word, let

$$\sigma = \mu_H + \mu_M \quad \text{and} \quad \tau = \frac{\mu_H}{\mu_H + \mu_M}. \tag{3.6}$$

Note that σ essentially measures the average frequency of a word's usage per 1000 words, and τ measures the word's ability to discriminate. Mosteller and Wallace

assumed that τ follows the Beta distribution $\tau^{\gamma-1}(1-\tau)^{\gamma-1}/\mathrm{B}(\gamma,\gamma)$ with equal arguments γ, and that $\gamma = \beta_1 + \beta_2\sigma$ with the β's typically both being positive. Note that an increase in σ implies a decrease in the variability of τ. The values of β_1 and β_2 were estimated from a data set of 90 function words chosen on the basis of having either high or low frequency in the English language. For any particular word, the distribution of its τ, based on the estimated parameters β_1 and β_2, is then used as the prior for that word.

Let now x_H and x_M be the counts observed for a given word in each of the combined known writings of Hamilton and Madison respectively. Following Bayes' theorem, the posterior density of (σ,τ) given x_H and x_M is,

$$\pi(\sigma,\tau|x_H,x_M) = C(x_H,x_M)\pi(\sigma,\tau)p(x_H,x_M|\sigma,\tau)$$

where C is a function of x_H and x_M, $\pi(\sigma,\tau)$ is the prior density of (σ,τ), and $p(x_H,x_M|\sigma,\tau)$ is the probability of observing the counts x_H and x_M given that the parameters of the model are σ and τ. The goal is to evaluate the posterior density $\pi(\sigma,\tau|x_H,x_M)$ for each word and then to use the mode of this density in equation (3.6) so as to determine estimates of the rates μ_H and μ_M for that word.

Returning to the *also* example, there were 26 counts out of 94,000 words in Hamilton's text, and 80 counts out of 114,000 words in Madison's text of the word *also* with corresponding true rates $\mu_H = \sigma\tau$ and $\mu_M = \sigma(1-\tau)$ per thousand words. Assuming independence of word usage between Hamilton and Madison, the log-likelihood of observing these counts under the Poisson distribution is:

$$\begin{aligned}
\log(p(x_H,x_M|\sigma,\tau)) &= \log(p(x_H|\sigma,\tau)) + \log(p(x_M|\sigma,\tau)) \\
&= -94\sigma\tau + 26\log(94\sigma\tau) - \log(26!) \\
&\quad -114\sigma(1-\tau) + 80\log(114\sigma(1-\tau)) - \log(80!)\ .
\end{aligned}$$

The logarithm of the prior density $\pi(\sigma, \tau)$, where τ follows the Beta distribution with equal parameters γ, and where the values $\beta_1 = 10$ and $\beta_2 = 0$ are determined from the data, is given by

$$\begin{aligned} \log(\pi(\sigma,\tau)) &= \log(\pi(\tau|\sigma)) + \log(\pi(\sigma)) \\ &= \text{constant} + (10-1)\log(\tau(1-\tau)) \ . \end{aligned}$$

The constant includes $\log(\text{Beta}(10, 10))$ and a flat prior assigned to σ. Finally, the posterior density for (σ, τ) is evaluated to be:

$$\begin{aligned} \log(\pi(\sigma,\tau|x_H,x_M)) &= \text{constant} - \frac{94+114}{2}\sigma + (80+26)\log(\sigma) \\ &\quad +(114-94)\sigma(\tau-1/2) + (26+10-1)\log(\tau) \\ &\quad +(80+10-1)\log(1-\tau). \end{aligned}$$

The mode of the above posterior gives $\hat{\sigma} = 0.99$ and $\hat{\tau} = 0.316$, which implies $\hat{\mu}_H = 0.31$ and $\hat{\mu}_M = 0.67$. Since $\hat{\mu}_H$ and $\hat{\mu}_M$ are rates per thousand words, if we wish to evaluate the likelihood ratio of Hamilton versus Madison in a paper of 2000 words (the typical length of an essay in the disputed papers) for the word *also*, we multiply both ratios by 2 and compute $f(4|0.62)/f(4|1.34)$ from the Poisson table.

Applying the log-odds (3.5) to the counts of the so-called function words from each of the 12 disputed papers where $\hat{\mu}_H$ and $\hat{\mu}_M$ for each word are evaluated from the known texts of Hamilton and Madison, Mosteller and Wallace computed the log-odds for each of the disputed 12 papers.

Mosteller and Wallace also carried out their statistical study of the 12 disputed papers assuming the counts of function words in the text follow negative binomial distributions, although here we have only described the analysis when the word counts were assumed to follow Poisson distributions. On the basis of both studies, Mosteller and Wallace inferred that all 12 of the disputed papers were written by Madison.

3.3 Some techniques from information retrieval

Information retrieval is the study of methods for finding or retrieving documents "relevant" to a particular query or information request. Information retrieval can be divided into three main areas of research[1]: content analysis, information structure, and evaluation. Briefly, content analysis is concerned with describing the content of documents in a form suitable for computer processing. Information structure is concerned with devising a retrieving strategy or methodology (by examining the relationships between documents) so as to be able to effectively retrieve relevant documents. Evaluation is concerned with analyzing the effectiveness of a retrieving strategy. In this thesis we will be mainly concerned with retrieval methods, and in particular with the classification of documents by date, using (in one of our methods) algorithms based on notions of similarity between documents (formal definitions will follow). Popular measures of similarity that we will focus on are those based on vector space models. It is worth pointing out that there are many measures of similarity between documents (more than 60 according to McGill *et al.* (1975)), but the differences in retrieval performance achieved between the different similarity measures tend to be relatively insignificant (van Rijsbergen, C.J., (1979), p. 24).

Before we describe retrieval strategy models, we need to define some basic terms. The reference used for this section and the next subsection is from Grossman and Frieder (2004).

A *word* is defined as a string of nonblank characters which appear in the full text of the document. A *stop word* is defined to be a word which occurs often in a document but has no purpose for information retrieval purposes – for example, *the, a, for,* and so on . A *term* is a sequence of words in a document. In the information

[1]See van Rijsbergen, (1979) p. 5.

retrieval literature however, term has a stricter definition. In addition to stop words, suffixes and/or prefixes are also first removed from the words. Since in this thesis we are developing dating algorithms which do not use knowledge about the language in which the documents are written, we have chosen to stick to a loosely defined usage of the word "term". A *query* $\mathcal{Q}$ is a set of terms representing the user's information needs. We also assume that there is a finite set $D_1, D_2, D_3, \ldots, D_n$ of documents on which the query search is performed. For our purposes, each query is just a document from the DEEDS data set which we are trying to date, and the dates of the retrieved documents will be used in one of our dating procedures.

3.3.1 Vector space models for documents

The *vector space model*, developed by Salton *et al.* (1975), involves representing documents as vectors. The vector is constructed by first assigning to each term – which might be a single word or a longer phrase – a *term weight*, i.e. a quantitative measure of the importance of that particular term (we will discuss shortly how term weight is determined). The length of the vector is the number of distinct terms occurring in the corpus of documents, and the coordinate points of the vector (for the document) correspond to term weights of the distinct terms.

Once documents have vector representations, say of length k, similarity between two documents is measured as follow. Let document $\mathcal{D}$ be represented by the vector $\{d_1, d_2, \ldots, d_k\}$, and let the query document $\mathcal{Q}$ be represented by the query vector $\{q_1, q_2, \ldots, q_k\}$ where d_i (respectively q_i) is the weight of the (same) ith term in $\mathcal{D}$ (respectively $\mathcal{Q}$). The next step is to define a function that measures the closeness of these two vectors. Generally, this function is the inner product (or dot product), essentially measuring the angle between the vectors, with the underlying premise

being that if the query and the document are similar, then their vectors should be pointing in the same general direction.

Term weights can be determined in a number ways. One could use:

1) the frequency of the ith term f_i in a document (commonly known as term frequency or *tf* for short)

2) the logarithm of the frequency of the term, say $\log(f_i + 1)$

3) the value $f_i \log(N/n_i)$ where N is the total number of documents in the collection, and n_i is the number of documents that contain the term i. The expression $\log(N/n_i)$ is known as *inverse document frequency (idf)*

The quantity *tf* $\times$ *idf*, i.e., $f_i \log(N/n_i)$, is perhaps the most widely used weight in information retrieval. As we see in the *tf* $\times$ *idf* expression, when the ith term occurs in most documents, that is when n_i is close to N, then the value of *tf* $\times$ *idf* is close to zero, which is appropriate since the i^{th} term is then not really discriminating among documents.

Example

In this example, we will use the term frequency (tf) as term weight. Let

$\mathcal{D}$ = "We like to hop on top of Pop"

$\mathcal{Q}$ = "You must not hop on Pop"

We base the similarity measure between $\mathcal{D}$ and $\mathcal{Q}$ on the angle-based cosine measure, where the cosine of the angle between two documents (query $\mathcal{Q}$ and document $\mathcal{D}$) is given by:

$$\text{sim}_C(D, Q) = \frac{\sum_{i=1}^{m} d_i q_i}{\sqrt{\sum_{i=1}^{m} d_i^2}\sqrt{\sum_{i=1}^{m} q_i^2}} \tag{3.7}$$

To measure the cosine angle between $\mathcal{D}$ and $\mathcal{Q}$ based on term frequency, *tf* (where term, in this illustration, will mean individual words), we begin with the set of the union of words between $\mathcal{D}$ and $\mathcal{Q}$: {we, like, to, hop, on, top, of, pop, you, must, not}.

The *tf* vector representing $\mathcal{D}$, denoted as $\mathcal{D}_{tf}$, is the frequency of each of the words of the above set in document $\mathcal{D}$ itself. Therefore, $\mathcal{D}_{tf} = (1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$. Similarly, $\mathcal{Q}_{tf} = (0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1)$. It follows that

$$\text{sim}_C(\mathcal{D}_{tf}, \mathcal{Q}_{tf}) = \frac{3}{\sqrt{8}\sqrt{6}} = 0.43$$

The advantage of the cosine measure is that the lengths of the document and query vectors are normalized. Had we used the inner product alone, longer documents would likely be found to be more similar to the query simply because there is more chance for the matching of terms to occur – and not because the longer document is actually more relevant to the query. On the other hand, the normalization term creates its own deficiencies. For example, if

$\mathcal{Q}'$ = "We like to hop on top of Pop we like to hop on top of Pop"

then, using the frequency of the distinct words as the term weights,

$$\text{sim}_C(\mathcal{D}_{tf}, \mathcal{Q}'_{tf}) = 1.$$

Documents $\mathcal{D}$ and $\mathcal{Q}'$ would now both be pointing in exactly the same direction in document space, even though the document lengths differ. Zhang and Korfhage (1999) have proposed methods of integrating angle similarity measures with distance similarity measures in order to discriminate between documents that point in the same directions but are however different due to their lengths.

3.3.2 Similarity measures between texts

We begin with set theoretic similarity measure initially used for query search purposes in the context of the World Wide Web. It was originally suggested by Broder (1998).

Let $\mathcal{D}$ be a document consisting of words in the order $(w_1, w_2, \ldots, w_n)$, noting that the words are not necessarily distinct. We define a *shingle of size* k (or k-

shingle) to mean a sequence s_k of k consecutive words from the document $\mathcal{D}$, that is $s_k = (w_{t+1}, w_{t+2}, \ldots, w_{t+k})$ where $0 \leq t \leq n-k$. Define $s_k(\mathcal{D})$ to be the set of all not necessarily distinct k-shingles of $\mathcal{D}$, and define $\mathcal{S}_k(\mathcal{D})$ to be the set of distinct k-shingles of $\mathcal{D}$. The following is an example taken from Broder (1998) where the shingle size is chosen to be $k = 2$. If

$$\mathcal{D} = (\text{a rose is a rose is a rose})$$

then the set of all its 2-shingles is given by

$$s_2(\mathcal{D}) = \{\{\text{a rose}\}, \{\text{rose is}\}, \{\text{is a}\}, \{\text{a rose}\}, \{\text{rose is}\}, \{\text{is a}\}, \{\text{a rose}\}\}$$

and the set of distinct 2-shingles is given by

$$\mathcal{S}_2(\mathcal{D}) = \{\{\text{a rose}\}, \{\text{rose is}\}, \{\text{is a}\}\} \ .$$

Once the shingle order k has been fixed, Broder defines the resemblance of two documents $\mathcal{D}_1$ and $\mathcal{D}_2$ as

$$res_k(\mathcal{D}_1, \mathcal{D}_2) = \frac{\|\mathcal{S}_k(\mathcal{D}_1) \cap \mathcal{S}_k(\mathcal{D}_2)\|}{\|\mathcal{S}_k(\mathcal{D}_1) \cup \mathcal{S}_k(\mathcal{D}_2)\|} \ , \tag{3.8}$$

where $\| \ \|$ denotes the order of a finite set. Using this definition of resemblance, he defines a distance between two documents to be

$$dis_k(\mathcal{D}_1, \mathcal{D}_2) = 1 - res_k(\mathcal{D}_1, \mathcal{D}_2) \ ;$$

as noted by Broder (1998), this distance measure is actually a metric, i.e. it satisfies the triangle inequality.

The angle-based and the set theoretic similarity measures discussed above can be viewed as being special cases of a class of correspondences or measures of document resemblance, the formal definition of which we will discuss next. This material is based on Feuerverger, Hall, Tilahun and Gervers (2005,2008).

Let $n(i)$ and $n(j)$ denote the number of words of $\mathcal{D}_i$ and $\mathcal{D}_j$, respectively. In this notation, $\mathcal{D}_i = (\omega_1, \omega_2, \ldots, \omega_{n(i)})$. Let $\mathrm{N}(i,j)$ denote the order of the set $\mathcal{S}_k(\mathcal{D}_i) \cup \mathcal{S}_k(\mathcal{D}_j)$. For each $l,\ 1 \leq l \leq \mathrm{N}(i,j)$, $\nu_l(i,j)$ denotes the number of times the l^{th} element of $\mathcal{S}_k(\mathcal{D}_i) \cup \mathcal{S}_k(\mathcal{D}_j)$ occurs in $s_k(\mathcal{D}_i)$. Denote the numbers of (not necessarily distinct) shingles of order k of $\mathcal{D}_i$ and $\mathcal{D}_j$ by $N(i)$ and $N(j)$, respectively. It follows that $N(i) = \sum_l \nu_l(i,j) = ||s_k(\mathcal{D}_i)|| = n(i) - k + 1$ and $N(j) = \sum_l \nu_l(j,i) = ||s_k(\mathcal{D}_j)|| = n(j) - k + 1$. (Here, the dependence of $N(i)$ and $N(j)$ on k has been suppressed).

Let $f(u,v)$ be a bivariate, non-negative function. Define the kth order *correspondence* between the documents $\mathcal{D}_i$ and $\mathcal{D}_j$ to be

$$\mathrm{corr}_k(i,j) = \frac{\sum_{l=1}^{N(i,j)} f\left(\nu_l(i,j), \nu_l(j,i)\right)}{F\left\{\nu_1(i,j), \cdots, \nu_{N(i,j)}(i,j); \nu_1(j,i), \cdots, \nu_{N(j,i)}(j,i)\right\}}$$

where we choose a normalizing function F (a function of two vector arguments) which has the property that for all documents $\mathcal{D}_i$ and $\mathcal{D}_j$ the following two conditions are satisfied:

(a) $0 \leq \mathrm{corr}_k(i,j) \leq 1$

and

(b) $\mathrm{corr}_k(i,j) = 1$ whenever $\mathcal{D}_i = \mathcal{D}_j$.

The *correspondence distance* associated with the correspondence $\mathrm{corr}_k(i,j)$ is then defined to be:

$$d_k(i,j) = 1 - \mathrm{corr}_k(i,j).$$

As we shall see below, $d_k(\cdot,\cdot)$ does not always turn out to be a metric.

Examples of correspondences

Type (I):

$f(u,v) = u^\alpha v^\alpha$ and

$$F(\vec{u},\vec{v}) = F(u_1, \cdots, u_{N(i,j)}, v_1, \cdots, v_{N(i,j)}) = \left(\sum_{l=1}^{N(i,j)} u_l^{2\alpha}\right)^{1/2} \left(\sum_{l=1}^{N(i,j)} v_l^{2\alpha}\right)^{1/2}$$

for an α such that $0 < \alpha < \infty$.

Type (II):

$f(u, v) = u^\alpha v^\alpha$ and

$$F(\vec{u}, \vec{v}) = \sum_{l=1}^{N(i,j)} (u_l^{2\alpha} + v_l^{2\alpha} - u_l^\alpha v_l^\alpha)$$

for an α such that $0 < \alpha < \infty$.

Type (III):

$f(u, v) = I(u > 0, v > 0)$ and

$$F(\vec{u}, \vec{v}) = \sum_{l=1}^{N(i,j)} \left(I_{(u_l>0)} + I_{(v_l>0)} - I_{(u_l>0)} I_{(v_l>0)} \right)$$

Type (IV):

$f(u, v) = \min(u, v)$ and

$$F(\vec{u}, \vec{v}) = \min \left(\sum_{l=1}^{N(i,j)} u_l, \sum_{l=1}^{N(i,j)} v_l \right) = \min(N(i), N(j)).$$

All of the correspondences of types (I) to (IV) satisfy the conditions (a) and (b). The angle based document similarity which we defined at (3.7), for example, is a type (I) correspondence with $\alpha = 1$. If we take $\alpha = 1/2$ in (I), we obtain a correspondence which Quang *et al.* (1999) term "Hellinger similarity". Specifically, letting

$$\pi_l = \frac{u_l}{\sum_{1 \le l \le N(i,j)} u_l} \quad \text{and} \quad \xi_l = \frac{v_l}{\sum_{1 \le l \le N(i,j)} v_l},$$

the resulting correspondence distance may be written as

$$d_k(i, j) \;=\; 1 - \sum_{1 \le l \le N(i,j)} \sqrt{\pi_l \xi_l}.$$

It can be easily verified that

$$\sqrt{d_k(i, j)} = \left(1 - \sum_{1 \le l \le N(i,j)} \sqrt{\pi_l \xi_l} \right)^{1/2} = \left(\frac{1}{2} \sum_{1 \le l \le N(i,j)} \left[\sqrt{\pi_l} - \sqrt{\xi_l} \right]^2 \right)^{1/2}$$

and that the square root of the Hellinger correspondence distance is actually a metric. If u and v were discrete probability distributions rather than counts, then this would just be the standard Hellinger distance for multinomials.

A shortcoming of correspondences of type (I) is that in general, the associated correspondence distances do not satisfy the triangle inequality, and so do not define a metric. As a counterexample, let $\vec{u} = (1, 0)$, $\vec{w} = (1, 1)$, and $\vec{v} = (0, 1)$, and consider the distances between $\vec{u}$ and $\vec{w}$, and $\vec{w}$ and $\vec{v}$. This example shows that the angle based difference measure (which is of type (I)) is not a metric.

Comparing type (I) correspondence to type (II), we see that the only difference is in the normalization function F, so that the main geometric aspect of the relationship remains the same. Furthermore, the value of F in type (I) does not exceed that for type (II) (which can be seen by an application of the Cauchy-Schwartz inequality), and therefore the correspondence distance for type (I) is always smaller than or equal to that for type (II). An advantage of the type (II) correspondence distances is the fact that they are metrics for all values of α, $(0 < \alpha < \infty)$; in particular, the triangle inequality is preserved. For a proof, see Feuerverger, Hall, Tilahun, Gervers (2004).

The Broder resemblance measure (3.8) is an instance of a correspondence measure of type (III); in fact type (III) correspondence, and hence the resemblance measure, can be interpreted as the limit (as $\alpha \downarrow 0$) of measures of type (II). Unlike the type (III) correspondence distance, type (II) allows us to choose a continuously variable range of distance measures through the selection of α. The values of α can be chosen to, in effect, allow different weightings to apply to the shingle counts $\nu_l(i, j)$. As a limit, type (III) ignores the actual shingle counts and only considers the presence or absence of a particular shingle.

The type (IV) correspondence distance is related to the so-called variation, or L_1 distance. The variation distance between two discrete distributions $p = (p_1, \cdots, p_n)$

and $q = (q_1, \cdots, q_n)$ is given by:

$$\mathrm{dist}_V(p, q) = \sum_{i=1}^{n} |p_i - q_i| = 2 - 2\sum_{i=1}^{n} \min(p_i, q_i).$$

(The above equality can be obtained if we note that $|p_i - q_i| = \max(p_i, q_i) - \min(p_i, q_i)$, $p_i + q_i = \max(p_i, q_i) + \min(p_i, q_i)$ and $\sum q_i = \sum p_i = 1$). If we define $p_l = v_l(i, j)/\mathrm{N}(i)$ and $q_l = v_l(j, i)/\mathrm{N}(j)$, the frequency of the lth shingle occurrence in documents $\mathcal{D}_i$ and $\mathcal{D}_j$ respectively, then the type (IV) correspondence distance for shingle frequencies is given by:

$$d_k(i, j) = 1 - \frac{\sum_l \min(p_l, q_l)}{\min(\sum_l p_l, \sum_l q_l)} = 1 - \sum_{l=1}^{N(i,j)} \min(p_l, q_l) = \frac{1}{2}\mathrm{dist}_V(p, q).$$

Again, unlike the type (II) correspondences, type (IV) lacks a parameter with which we can assign different weights to shingle counts $\nu_l(i, j)$.

We remark that the geometric aspects of the relationship between the type (I) and (II) correspondence distances is preserved but with the added advantage that the type (II) correspondence distance are metrics, and therefore the meaning of "distance" with the type (II) correspondence distances is intuitively more appealing. In that sense, type (II) is regarded as being preferable to type (I).

In summary, we see that the most commonly used type (I) correspondence distance differs from type (II) only in the normalization factor used. Type (II) is also preferable to types (III) or (IV) since the later two types of correspondences lack parameters which allow them to assign differing weights to the shingle counts.

There is precedence in the use of distance measure functions in the field of information retrieval. As mentioned already, Broder (1998) has applied type (III) correspondence distance measure for query search in the context of the World Wide Web. In the field of colour-based image retrieval, conventional distance functions such as L^1, L^2 and L^∞ have found applications. See, for example, Stehling, Nascimento and

Falcão (2003). Other types of distance measures in the context of, for example, video image and audio-based retrieval techniques are further discussed in Djeraba (2003).

In the field of vector space based information retrieval techniques, methods are used to retrieve documents that more than just literal match to a query term. Methods in this direction employ the use of various matrix decomposition techniques to a sparse, term-by-document matrices. For detailed developments in this area, see Berry and Browne (2005).

3.3.3 Evaluation

Once we have chosen a document retrieval strategy, the next step is to evaluate its effectiveness. When a query is submitted to a system, a number of documents from the collection are retrieved, some of which are considered relevant. The relevance of a document must typically be determined by human experts in the discipline of the document. An underlying assumption in such evaluations is that if a given retrieving strategy fares well under experimental conditions, then it is likely to perform favourably under the real world situation. In a perfect system, the number of retrieved and relevant documents would be identical, but in reality, systems typically retrieve many non-relevant documents. To measure the effectiveness of retrieval strategies, two ratios, *recall* and *precision* may be computed. These are defined as:

$$\text{precision} = \frac{\text{number of relevant documents retrieved}}{\text{number of retrieved documents}}$$

$$\text{recall} = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents}}.$$

A 100% recall value can of course be achieved by simply retrieving all of the documents in the collection. Precision however would control for the system's inefficiency by evaluating how many of the retrieved documents are actually relevant.

The assessment of recall and precision can be complicated since most systems only rank their document collection with respect to a query (using document similarity measures, for instance). There are then two approaches taken to compute precision and recall values. The first method specifies a fixed number of documents for each query and then determines the average precision and recall scores over a set of queries. The second method specifies a set of recall values, and for each recall value, the average precision score is computed over a set of queries. The latter method gives rise to what is known in the literature as a precision-recall curve. Comparing two retrieval strategies using their precision-recall curves is difficult since, for instance, the first method can outperform the second one at the high and low values of recall, whereas the second method can outperform the first for moderate recall values.

The performances of different retrieval strategies can be tested on a U.S. government-sponsored database known as TREC (Text REtrieval Conference). This database consists of document sets for both training purposes as well as for testing purposes. These document sets are also accompanied by a set of queries (known as 'topics' in TREC) as well as by relevance judgements.

3.4 Some further recent literature

Forman (2006) provides an explorative study for text classification problems where the features of the data change over time. Sharan and Neville (2007) explore "time-varying relationships in statistical relational models" where they provide an application to text classification. They also mention, as a potential application, the case of fraud detection, whereby it would be informative to find out who is working with whom, as well as the pattern of communication over time among suspects.

Zhao *et al.* (2009) model the frequencies of occurrences of words across time in

time-stamped documents using local logistic regression. Apparently, unaware of the results of Fan *et al.* (1995), their method is actually a special case of locally constant logistic regression. The theoretical framework in the work of Fan *et al.* (1995) will provide a basis for our work in modelling the probabilities of occurrences of words in Chapter 6. The model we will propose can handle the problems of edge biases and can be generalized to higher degrees of the local polynomial of the kernel regression. We note that asymptotic results, such as bias, variance and distribution of the estimators are readily available for local polynomial fitting and also that they are superior for polynomials of odd degree.

Chapter 4

Relevant Statistical Tools

The key techniques that we discuss in this chapter are non-parametric regression and local smoothing in generalized linear models. Non-parametric regression, and in particular its GLM local smoothing versions, will be used extensively in our work on the maximum prevalence method for dating documents in Chapter 6. Choosing optimal bandwidths for such procedures will also be important for us, so for this reason we will emphasize those details in our discussions in the present chapter. Our main references for this chapter are Härdle, W. (1990, 1991), Simonoff, J. (1996), Fan and Gijbels (1996, 2000), Sarda and Vieu (2000), Schimek (2000), Silverman (1986) and Wand and Jones (1995).

The techniques and results from kernel density estimation theory form a foundation for the developments reviewed in this chapter. A discussion of the relevant theory of kernel density estimation is provided in Appendix A.

This chapter is a self-contained background material. Those familiar with the materials in this chapter can go straight to Chapter 5 and Chapter 6, the chapters containing the main contributions of this thesis.

4.1 Non-parametric regression

The goal of a regression curve is to fit a relationship between an explanatory variable X and an output variable Y. Suppose we have bivariate i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and we wish to estimate the function $m(x) = E(Y|X = x)$ assuming a model of the form

$$Y = m(X) + \sigma(X)\epsilon$$

with $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = 1$, where X and ϵ are independent. We will let f denote the density of the random variables $X_1, \ldots, X_n$.

In nonparametric regression curve estimation, we are basically interested in a weighted average of the response Y within a certain neighbourhood of x. We weight each observation Y_i depending on the distance of X_i to x. We then estimate the value of $m(x)$ as

$$\hat{m}_h(x) = n^{-1} \sum_{i=1}^{n} w_{hi}(x; X_1, \ldots, X_n) Y_i$$

where $w_{hi}(\cdot)$ is a weight function which depends on the particular kernel estimation technique being used, and on the distance between x and the explanatory variables X, and where h is a smoothing parameter. The goal is to determine which choice of weight function and smoothing parameter are most appropriate to our data.

4.1.1 The Nadaraya-Watson regression estimator

One of the earliest and most common of the nonparametric regression estimators is the Nadaraya-Watson estimator. Introduced by Nadaraya (1964) and Watson (1964), the Nadaraya-Watson estimate traces its origin to the so-called regressogram. The regressogram is for regression estimation what the histogram is for density estimation. The set of values of X are partitioned into J subintervals, denoted by B_j, and the

average value of the Y's are taken using these B_j intervals. The resulting regression estimate is a step-function, defined for any $x \in B_j$ by

$$\hat{m}_j(x) = \frac{\sum_{i=1}^n Y_i I(X_i \in B_j)}{\sum_{i=1}^n I(X_i \in B_j)} \ .$$

A natural way to extend the regressogram is to introduce a window-based estimate where the values of the Y's are locally averaged, i.e., in which the averaging that takes place is centered at the point x. More precisely, for any x, we define the estimator

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n Y_i I(X_i \in [x-h, x+h])}{\sum_{i=1}^n I(X_i \in [x-h, x+h])} \ .$$

The Nadaraya-Watson estimate uses a kernel weight function K, generalizing the second estimator. The *kernel* function K is a continuous, bounded and a symmetric real function satisfying the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1 \ .$$

Define the kernel $K_h(u) = h^{-1}K(u/h)$ with scale factor h called *bandwidth*. The *Nadaraya-Watson* estimator is defined as

$$\hat{M}_h(x) = \frac{n^{-1}\sum_{i=1}^n Y_i K_h(X_i - x)}{n^{-1}\sum_{i=1}^n K_h(X_i - x)}, \ \text{ provided } \sum_{i=1}^n K_h(X_i - x) \neq 0,$$

and equals zero otherwise. The weight function corresponding to this estimator is given by

$$w_{hi}(x) = \frac{\frac{1}{h}K(\frac{X_i - x}{h})}{\hat{f}_h(x)} \tag{4.1}$$

where

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)$$

is the kernel density estimate for $f(x)$ (for reference, see Appendix A). Hereafter $\hat{M}_h(x)$ will denote the Nadaraya-Watson estimator. As pointed out by Härdle (1991),

some properties which follow from the definition of the Nadaraya-Watson estimator are:

a) The estimator $\hat{M}_h(x)$ is continuous if we choose a continuous kernel.

b) Observations Y_i are weighted more in areas where the corresponding X_i's are sparse but effectively fewer Y_i's are used in the averaging.

c) If $h \to \infty$, the weight $w_{hi}(x) \to 1$, and the estimate $\hat{M}_h(x) \to \bar{Y}$. Thus large bandwidths lead to oversmoothing. If $h \to 0$, then the weight $w_{hi}(x) \to n$ if $x = X_i$, and becomes undefined for all other values of x. Therefore, $\hat{M}_h(X_i) \to Y_i K(0)/K(0) = Y_i$ as $h \to 0$; that is, small bandwidths result in data interpolation.

If the predictors X_i are random and have an unknown density, then the Nadaraya-Watson kernel estimator is the most natural one to use. If the density function of the predictors, $f_X(x)$, is known, then the weight

$$w_{hi}(x) = \frac{h^{-1}K(\frac{X_i - x}{h})}{f_X(x)} \tag{4.2}$$

is preferable. For example, if the random variables are observations taken at some regular intervals (assume without loss of generality that the observations lie in $[0, 1]$), then we may think of the observations as being drawn from a uniform distributions on $[0, 1]$ and we would take $f = I_{[0,1]}$.

If the predictors, on the other hand, are not random but, say, are a fixed set of nearly equally spaced ordered numbers $x_1, \ldots, x_n$, then without loss of generality, we can assume that the x observations lie in the unit interval $[0, 1]$. We then use the weight (4.2), but we take the estimate of the "density" of $f_X(x)$ to be

$$\hat{f}_X(x) = \frac{1}{n(x_i - x_{i-1})} \quad \text{for } x \in (x_{i-1}, x_i]$$

since $x_i - x_{i-1} \approx 1/n$. More generally, substituting this form into equation (4.1), we

obtain a kernel estimator due to Priestley and Chao (1972), namely,

$$\hat{m}_{PC}(x) = h^{-1} \sum_{i=1}^{n} (x_i - x_{i-1}) K\left(\frac{x_i - x}{h}\right) Y_i$$

where $x_0 = 0$.

An estimator for the case when the predictors are ordered, but not necessarily equally spaced, was given by Gasser and Müller (1979):

$$\hat{m}_{GM}(x) = \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_h(u - x) du$$

where $s_i = \frac{1}{2}(X_i + X_{i+1})$, $X_0 = -\infty$, $X_{n+1} = +\infty$. We note here that from a function approximation point of view, the Nadaraya-Watson and the Gasser and Müller estimators are both solutions to the locally constant least squares regression problem given by

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - \theta)^2 w_i \ .$$

When $w_i = K_h(X_i - x)$ and $w_i = \int_{s_{i-1}}^{s_i} K_h(u - x) du$, we obtain the Nadaraya-Watson and the Gasser-Müller estimators respectively. Asymptotically, the variances of $\hat{m}_{GM}$ and $\hat{m}_{PC}$ are 1.5 time bigger than the variance of the Nadaraya-Watson estimator, but on the other hand, their biases are smaller. Also, comparatively, the bias form of the Nadaraya-Watson estimator is more complicated. For further details, see Fan and Gijbels (1996, p. 17).

4.1.2 Properties of the Nadaraya-Watson estimator

In this section, we examine some properties of the Nadaraya-Watson estimator. Let the kernel $K(x)$ be an arbitrary density function satisfying the conditions:

$$\text{(a)} \quad \sup_{-\infty < x < \infty} K(x) < \infty.$$

$$\text{(b)} \quad \lim_{|x| \to \infty} |x| K(x) = 0.$$

(c) $K(x) = K(-x)$.

(d) $x^2K(x) \in L^1(-\infty, \infty)$.

One of the ways to measure the closeness of $\hat{M}_h(x)$ to the true regression curve $m(x)$ is to study its mean-squared error. Assume $EY_i^2 < \infty$, and that $m(x)$, $f(x)$ and $f(x)E(Y^2|X = x)$ are continuous, with $f(x) > 0$. Using similar methods as those used to calculate the mean-squared error for density estimates, we have the following result (for details, see Härdle (1991) p. 135):

If $m(x)$ is twice differentiable, then

$$\begin{aligned} MSE(\hat{M}_h(x)) &= \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}||K||_2^2 + \frac{h^4}{4}\left(m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right)^2 \mu_2^2(K) \\ &+ o((nh)^{-1}) + o(h^4), \quad \text{as } h \to 0,\ nh \to \infty \end{aligned}$$

where $\sigma^2(x) = \text{Var}(Y|X = x)$. Dropping the order terms, we define the $AMSE$ (asymptotic MSE)

$$\begin{aligned} AMSE(\hat{M}_h(x)) &= \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}||K||_2^2 \\ &+ \frac{h^4}{4}\left(m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right)^2 \mu_2^2(K)\ . \end{aligned} \tag{4.3}$$

Here $\mu_2(K) = \int_{-\infty}^{\infty} s^2K(s)ds$. As Härdle (1991) points out, there are several points to notice regarding the $AMSE(\hat{M}_h(x))$:

a) $\hat{M}_h(x) \to m(x)$ in probability as $h \to 0$ and $nh \to \infty$ – that is, $\hat{M}_h(x)$ is a consistent estimator of $m(x)$.

b) The first term on the right of equation (4.3) describes the asymptotic variance of $\hat{M}_h(x)$. It is a function of the density $f(x)$ and the conditional variance of the response Y given the particular value of x. Note that if we are in a region of sparse data, $f(x)$ decreases, resulting in higher variance of $\hat{M}_h(x)$. The opposite is true if we are in a region with a high concentration of data; thus if we are in a region with

plenty of data, the regression curve will be more stable, i.e., smoother.

c) The second term on the right in equation (4.3) is the square of the bias of $\hat{M}_h(x)$. Near a local extremum, the bias term will be dominated by $m''(x)$ and near an inflection point of $m(x)$, it will be dominated by $m'(x)$.

d) The optimal value of the bandwidth h (i.e. one that minimizes $AMSE(\hat{M}_h(x))$) is given by

$$h_{amse} = n^{-1/5}\left(\frac{c_1}{c_2}\right)^{1/5}$$

where $c_1 = \{\sigma^2(x)/f(x)\}||K||_2^2$, and $c_2 = \left(m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right)^2 \mu_2^2(K)$.

Similar to the draw-backs we encountered when examining the properties of the h_{amse} for the kernel density estimate (refer to Appendix A), we find that the h_{amse} for the Nadaraya-Watson estimate is dependent on x as well as on the unknown functions $m(\cdot)$ and $f(\cdot)$. Methods for finding the optimal bandwidths for non-parametric regression are deferred to a subsequent section.

4.1.3 Local polynomial regression

We begin by considering the Nadaraya-Watson regression estimate from a local polynomial approximation point of view. If we assume the existence of the $(p+1)^{st}$ derivative of the mean function $m(u)$, then Taylor expanding of $m(u)$ in a neighbourhood of x we obtain

$$m(u) \simeq m(x) + m'(x)(u-x) + \frac{m''(x)}{2!}(u-x)^2 + \cdots + \frac{m^p(x)}{p!}(u-x)^p.$$

Letting $\beta_j = \frac{m^{(j)}(x)}{j!}$ (and thus suppressing the dependence of β on x), we now fit this polynomial locally by weighted least square regression. This involves solving the minimization problem

$$\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j\right)^2 K_h\left(X_i - x\right) \tag{4.4}$$

where $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$ is a kernel function that assigns a weight to each data point, and h is a bandwidth that controls the size of the local neighbourhood.

Framing the above problem in matrix notation, let

$$\mathbf{X}_x = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^p \end{pmatrix},$$

and let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \text{ and } \hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{pmatrix}.$$

Also, letting $\mathbf{W}_x$ be the $n \times n$ diagonal matrix

$$\mathbf{W}_x = \operatorname{diag}\left[K_h\left(X_1 - x\right), \ldots, K_h\left(X_n - x\right)\right],$$

the least squares problem (4.4) may be written as

$$\min_{\beta}(\mathbf{Y} - \mathbf{X}_x\beta)^T\mathbf{W}_x(\mathbf{Y} - \mathbf{X}_x\beta) \tag{4.5}$$

where $\beta = \{\beta_0, \cdots, \beta_p\}^T$. When $\mathbf{X}_x = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, and $\beta = (\beta_0)$, then the above least squares problem becomes merely

$$\min_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 K_h(X_i - x)$$

for which the resulting minimizer, $\hat{\beta}_0$, is just the Nadaraya-Watson estimator. Since the Nadaraya-Watson estimator thus results when the polynomial order is $p = 0$, it is also referred to as the locally constant estimator.

In general, if $\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x$ is invertible, then

$$\hat{\beta} = (\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x\mathbf{Y}.$$

Letting $\hat{m}_p(x)$ denote the estimate of $m(x)$ based on the p^{th} polynomial order, $\hat{m}_p(x)$ is taken simply as the intercept term $\hat{\beta}_0$ in that model. In matrix notation, we have

$$\hat{m}_p(x) = \mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x\mathbf{Y} \tag{4.6}$$

where $\mathbf{e}_r^T$ is a $1\times(p+1)$ vector having the value 1 at the r^{th} entry and zero everywhere else. More generally, the estimate of the ν^{th} derivative of $m(x)$ based on the p^{th} polynomial order is given by

$$\widehat{m_p^{(\nu)}}(x) = \nu! \times \hat{\beta}_\nu = \nu!\mathbf{e}_{\nu+1}^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x\mathbf{Y}.$$

From equation (4.6) one can see that problems can arise if $\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x$ is not invertible. However, in the case that the predictors are fixed and assumed to be distinct, by taking h to be large enough invertibility can generally be guaranteed.

4.1.4 Some properties of local polynomial estimators

In this section we describe the bias and variance for the local polynomial regression estimator $\hat{m}_p(x)$ for the cases of polynomial order $p = 0$ and $p = 1$. Assume, without loss of generality, that the support of the design density $f(x)$ is bounded and coincides with $[0, 1]$. Also, without loss of generality, assume the support of the kernel K to be $[-1, 1]$. Then, the support of $K_h(x - \cdot)$ will be $\mathcal{E}_{x,h} = \{z : |z - x| \leq h\}$. We will call x an *interior point* if $\mathcal{E}_{x,h} \subset \text{supp}(f)$, and a *boundary point* otherwise. Thus, x is a boundary point if and only if $x = \alpha h$ or $x = 1 - \alpha h$ for some $0 \leq \alpha < 1$. Essentially, this means that the weight of the kernel will "spill" outside of the support of f. Furthermore, x is an interior point if and only if $h \leq x \leq 1 - h$.

Some further notation we need is as follows: let

$$\mu_j(K) = \int u^j K(u)du, \quad \text{and} \quad \nu_j(K) = \int u^j K^2(u)du \ .$$

Also, let $\mathbf{N}_p$ be an $(p+1)\times(p+1)$ matrix having the (i,j)th entry equal to $\mu_{i+j-2}(K)$ and let $\mathbf{M}_p(u)$ be the same as $\mathbf{N}_p$ but with the first column replaced by $(1, u, \cdots, u^p)$. Define $K_{(p)}$ to be the $(p+1)$th order kernel function, that is,

$$\int K(u)du = 1,\ \int u^r K(u)du = 0,\ r = 1, \ldots, p,\ \int u^{p+1}K(u)du \neq 0.$$

We can construct a higher $p+1$ order kernel from the kernel K using the formula

$$K_{(p)}(u) = \frac{|\mathbf{M}_p(u)|}{|\mathbf{N}_p|}K(u)\ ,$$

where $|\cdot|$ means determinant. We note that $K_{(0)} = K_{(1)} = K$, and in general, for p even, $K_{(p)} = K_{(p+1)}$. Following the exposition of Wand and Jones (1995), equation (4.6) implies that

$$E(\hat{m}_p(x)|X_1, \ldots, X_n) = \mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x E\left(\mathbf{Y}|X_1, \cdots, X_n\right) \tag{4.7}$$

where, using Taylor's expansion,

$$E\left(\mathbf{Y}|X_1, \ldots, X_n\right) = \begin{pmatrix} m(X_1) \\ \vdots \\ m(X_n) \end{pmatrix} = \begin{pmatrix} m(x) + m'(x)(X_1 - x) + \cdots \\ \vdots \\ m(x) + m'(x)(X_n - x) + \cdots \end{pmatrix}.$$

Therefore,

$$\begin{aligned} E\left(\hat{m}_p(x)|X_1, \ldots, X_n\right) &= \mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x \\ &\times \left\{ \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} m(x) \ + \begin{pmatrix} (X_1 - x) \\ \vdots \\ (X_n - x) \end{pmatrix} m'(x) + \cdots \right\}. \end{aligned} \tag{4.8}$$

For the case when the local polynomial is of degree $p = 0$, we have $\mathbf{X}_x = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, and $(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x) = \sum_{i=1}^n K_h(X_i - x)$. Therefore, (4.8) simplifies to

$$E(\hat{m}_0(x)|X_1, \ldots, X_n) - m(x) =$$

$$m'(x)\frac{\sum_{i=1}^{n}(X_i-x)K_h(X_i-x)}{\sum_{i=1}^{n}K_h(X_i-x)}+\cdots+\frac{m^{(i)}(x)}{i!}\frac{\sum_{i=1}^{n}(X_i-x)^iK_h(X_i-x)}{\sum_{i=1}^{n}K_h(X_i-x)}+\cdots$$

which gives us an expression for the bias when $p=0$.

For the case when the local polynomial is of degree $p=1$, $\mathbf{X}_x=\begin{pmatrix}1 & X_1-x\\ \vdots & \\ 1 & X_n-x\end{pmatrix}$ and so the expansion in equation (4.8) can be written as

$$E(\mathbf{Y}|X_1,\ldots,X_n)=\mathbf{X}_x\begin{pmatrix}m(x)\\ m'(x)\end{pmatrix}+\frac{1}{2}m''(x)\begin{pmatrix}(X_1-x)^2\\ \vdots\\ (X_n-x)^2\end{pmatrix}+\cdots$$

where, from equation (4.7),

$$\begin{aligned}&E(\hat{m}_1(x)|X_1,\ldots,X_n)-m(x)\\ &=\mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_xE\left(\mathbf{Y}|X_1,\cdots,X_n\right)-m(x)\\ &=\frac{1}{2}m''(x)\mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x\begin{pmatrix}(X_1-x)^2\\ \vdots\\ (X_n-x)^2\end{pmatrix}+\cdots\end{aligned}$$

thus giving an expression for the form of the bias when $p=1$.

Using analysis similar to section $(A.3)$ for the kernel density estimator, the asymptotic form of the conditional bias and conditional variance of $\hat{m}_p(x)$ when $p=0$ and $p=1$ are thus given as follows: when $p=0$, the asymptotic bias is

$$\begin{aligned}E(\hat{m}_0(x)-m(x)|X_1,\ldots,X_n) &= h^2\mu_2(K)\left\{m'(x)f'(x)/f(x)\right.\\ &\quad\left.+\frac{1}{2}m^{(2)}(x)\right\}+o_p(h^2)\ .\end{aligned}\tag{4.9}$$

And when $p=1$, the asymptotic bias is

$$E(\hat{m}_1(x)-m(x)|X_1,\ldots,X_n)=\frac{1}{2}h^2m''(x)\mu_2(K)+o_p(h^2)\ .\tag{4.10}$$

The variance of both $\hat{m}_0(x)$ and $\hat{m}_1(x)$ can be shown to be given by

$$\text{Var}(\hat{m}_p(x)|X_1,\ldots,X_n) = \frac{||K||_2^2\sigma^2(x)}{nhf(x)} + o_p((nh)^{-1})$$

for both $p = 0$ and $p = 1$ where $\sigma^2(x) = \text{Var}(Y|X = x)$.

More generally we have:

Theorem 1 (Ruppert and Wand (1994)[1]) *Assume* $[0,1]$ *is the support of* f, $f(x) > 0$ *and that* $f(\cdot)$, $m^{(p+1)}(\cdot)$ *and* $\sigma^2(\cdot)$ *are continuous in a neighbourhood of* x. *Further, assume that* $h \to 0$ *and* $nh \to \infty$, *and that* x *is an interior point, that is,* $h \leq x \leq 1-h$. *Then the asymptotic conditional variance of* $\hat{m}_p(x)$ *(where* p *is the order of the local polynomial) is given by*

$$Var(\hat{m}_p(x)|X_1,\ldots,X_n) = \frac{||K_{(p)}||_2^2\sigma^2(x)}{nhf(x)} + o_p((nh)^{-1}) \quad . \tag{4.11}$$

When p is odd:

The asymptotic conditional bias for p odd is given by

$$\text{Bias}(\hat{m}_p(x)|X_1,\ldots,X_n) = \frac{h^{p+1}m^{(p+1)}(x)\mu_{p+1}(K_{(p)})}{(p+1)!} + o_p(h^{p+1}) \quad . \tag{4.12}$$

When p is even:

If furthermore $f'(\cdot)$ *and* $m^{(p+2)}(\cdot)$ *are continuous in a neighbourhood of* x, *and* $nh^3 \to \infty$, *the asymptotic conditional bias for* p *even is given by*

$$\begin{aligned} \text{Bias}(\hat{m}_p(x)|X_1,\ldots,X_n) &= h^{p+2}\left[\frac{m^{(p+1)}(x)f'(x)}{f(x)(p+1)!} + \frac{m^{(p+2)}(x)}{(p+2)!}\right] \\ &\times \mu_{p+2}(K_{(p)}) + o_p(h^{p+2}) \quad . \end{aligned} \tag{4.13}$$

We can now write the conditional mean square error as, for p odd:

$$\begin{aligned} MSE(\hat{m}_p(x)|X_1,\ldots,X_n) &= \left[\frac{h^{p+1}m^{(p+1)}(x)\mu_{p+1}(K_{(p)})}{(p+1)!}\right]^2 + \frac{||K_{(p)}||_2^2\sigma^2(x)}{nhf(x)} \\ &+ o_p(h^{2p+2} + (nh)^{-1}) \quad , \end{aligned} \tag{4.14}$$

[1] As adapted from Simonoff (1998), p. 140, and Fan and Gijbels (1996), p. 62.

and for p even:

$$MSE(\hat{m}_p(x)|X_1,\ldots,X_n) = h^{2p+4}\left[\frac{m^{(p+1)}(x)f'(x)}{f(x)(p+1)!} + \frac{m^{(p+2)}(x)}{(p+2)!}\right]^2 \times(\mu_{p+2}(K_p))^2 + \frac{||K_{(p)}||_2^2\sigma^2(x)}{nhf(x)} + o_p(h^{2p+4}+(nh)^{-1})\,. \quad (4.15)$$

An important point to note is that when the local polynomial order p is even, the order of the asymptotic conditional biases of $\hat{m}_p(x)$ and $\hat{m}_{p+1}(x)$ are the same, namely $O_p(h^{p+2})$. Therefore, the locally constant estimator ($p = 0$) and the locally linear estimator ($p = 1$) both have $O_p(h^2)$ bias asymptotically. Defining $AMSE(\hat{m}_p(x))$ to be the leading asymptotics of $MSE(\hat{m}_p(x))$, from equations (4.14) and (4.15), we can determine the asymptotic rate of the bandwidth that minimizes the conditional $AMSE(\hat{m}_p(x))$. Specifically, when p is even, the optimal rate is $h = O(n^{-1/(2p+5)})$, yielding $MSE(\hat{m}_p(x)) = O_p(n^{-(2p+4)/(2p+5)})$, and when p is odd, the optimal rate is $h = O(n^{-1/(2p+3)})$, yielding $MSE(\hat{m}_p(x)) = O_p(n^{-(2p+2)/(2p+3)})$. This means that the optimal asymptotic conditional $MSE(\hat{m}_p(x))$ based on the local constant and local linear estimators both have order $O_p(n^{-4/5})$. For $p = 2$ or $p = 3$, the optimal $MSE(\hat{m}_p(x)) = O_p(n^{-8/9})$, and so on. Furthermore, we note that the form of the bias for p odd is simpler than for p even, insofar as for p odd we do not require knowledge of the design density function $f(x)$ (Simonoff (1996), p. 141).

4.1.5 Boundary bias of local polynomial estimators

In Figure 4.1, we show a scatter plot of y versus x for the regression function $y = 5x+\epsilon$, where 100 x's were generated from a standard normal distribution, and the ϵ's were also generated (independently) from a normal standard distribution. The figure also shows the $p = 0$ (Nadaraya-Watson) and the $p = 1$ (local linear) regression curve estimates using a Gaussian kernel and bandwidth $h = 0.2$. Note that on the left edge

of the x-axis, the Nadaraya-Watson kernel estimator curve lies above the data, and on the right edge of the x-axis, it lies below the data. On the other hand, no such edge bias is observed for the local linear estimator. One of the most common problems encountered when fitting kernel regression estimators, such as the Nadaraya-Watson estimator, is the edge or boundary bias that arises due to the absence of data at and beyond the boundaries. A major advantage of the local polynomial estimator (for odd degrees p), is that it automatically adjusts for boundary bias.

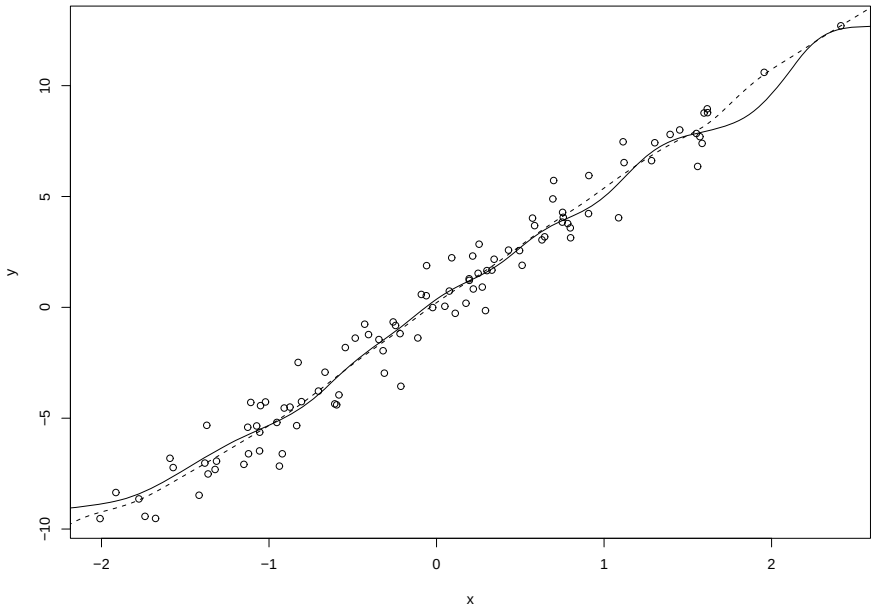


Figure 4.1: The x-axis represents 100 randomly generated standard normal random variables. The scatter plot represents a $y = 5x + \epsilon$ relationship, where the ϵ's are also generated (independently) from a standard normal distribution. The Nadaraya-Watson estimate ($p = 0$) is the solid line. The local linear estimate ($p = 1$), is the dashed line. The kernel is the standard Gaussian, and the bandwidth value is $h = 0.2$.

Analogous to the case when x is an interior point, one can derive the bias and the variance of $\hat{m}_1(x)$ and $\hat{m}_0(x)$ when x in a left boundary, that is when $x = \alpha h$,

$0 \leq \alpha < 1$, or when x is in the right boundary, that is when $x = 1 - \alpha h$.

Letting

$$\mu_{r,\alpha}(K) = \int_{-\alpha}^{1} u^r K(u) du \tag{4.16}$$

in the case when x is a left boundary point, we have

$$E(\hat{m}_1(x) - m(x)) = \frac{1}{2} h^2 m''(x) Q_\alpha(K) + o_p(h^2), \tag{4.17}$$

where

$$Q_\alpha(K) = \frac{\mu_{2,\alpha}^2(K) - \mu_{1,\alpha}(K)\mu_{3,\alpha}(K)}{\mu_{2,\alpha}(K)\mu_{0,\alpha}(K) - \mu_{1,\alpha}^2(K)} \ ,$$

and

$$E(\hat{m}_0(x) - m(x)) = m'(x) \frac{\mu_{1,\alpha}(K)}{\mu_{0,\alpha}(K)} h + O_p(h^2) \ . \tag{4.18}$$

The same holds when x is a right boundary point, except that the order of integration in (4.16) is from -1 to α. The kernel K is supported on $[-1, 1]$.

What we observe from (4.17) and (4.10) is that the bias of the local linear estimator is $O_p(h^2)$ whether x is an interior point or a boundary point. On the other hand, from (4.18) and (4.9) we see that the bias of the locally constant estimator increases from $O_p(h^2)$ to $O_p(h)$ when x goes from being an interior point to being a boundary point, unless $m'(x) = 0$. This phenomenon is evident in Figure 4.1. In particular, note that the term $\mu_{1,\alpha}(K)/\mu_{0,\alpha}(K)$ in (4.18) is positive when x is in the left boundary, and negative when x is in the right boundary. When $m'(x) > 0$, the increase to $O_p(h)$ on the left boundary means that the bias will be positive, as is indicated by our estimator $\hat{m}_0(x)$ lying above the data. Similarly, in the right boundary, the increase to $O_p(h)$ means that the bias will be negative, as is indicated by our estimator $\hat{m}_0(x)$ lying below the data there. This is due to the absence of observations at and beyond the boundary.

Boundary bias adjustment is not necessary for the local linear estimator. This is a key advantage of the local linear estimator over the locally constant (Nadaraya-Watson) estimator. This fact generalizes to the other estimators $\hat{m}_p(x)$; when p is odd, boundary bias is eliminated, but when p is even, boundary bias exists. (We are, of course, referring here to the leading order terms in the various bias expressions.) The advantage of the locally linear estimator over the locally constant estimator does come at a cost however. The asymptotic conditional variance of $\hat{m}_1(x)$ is about 3.17 times that of $\hat{m}_0(x)$ at the boundary when the Gaussian kernel and the same bandwidth h are used (Simonoff (1998), p. 143).

4.1.6 Measures of discrepancy

One of the ways in which an "optimal" bandwidth maybe defined is by introducing a global measure of discrepancy between the unknown regression curve $m(\cdot)$ and its estimate $\hat{m}_h(\cdot)$, and determining the bandwidth h which minimizes this measure of discrepancy. Some such measures, discussed in Härdle (1990 and 1991) are:

• Averaged squared error

$$ASE(h) = n^{-1}\sum_{j=1}^{n}(m(X_j) - \hat{m}_h(X_j))^2 w(X_j) \tag{4.19}$$

where $w(\cdot)$ is some assigned weight function

• Integrated squared error

$$ISE(h) = \int_{-\infty}^{\infty}(\hat{m}_h(x) - m(x))^2 w(x) f(x) dx \tag{4.20}$$

• Conditioned averaged squared error

$$MASE(h) = E(ASE(h)|X_1, \ldots X_n) \tag{4.21}$$

• Mean integrated squared error

$$MISE(h) = E(ISE(h)) \; . \tag{4.22}$$

All of the above measures are random except for $MISE(h)$ since this mean is computed over all possible samples $(X_1, Y_1), \ldots, (X_n, Y_n)$. Given such various measures of discrepancy between $m(\cdot)$ and $\hat{m}_h(\cdot)$, the question is which one is the most appropriate distance measure to use for the purposes of determining an optimal bandwidth. Härdle and Marron (1986) showed that under suitable moment conditions on Y, given the predictors, and assuming certain continuity properties on $f(x)$ and on the kernel K, bandwidth sequences that minimize $ASE(h)$, $ISE(h)$ or $MASE(h)$ also asymptotically minimize $MISE(h)$. For this reason, for purposes of bandwidth selection, it does not matter greatly which of those distance measures we choose to work with. For further discussion, refer to Härdle (1990), Chapter 5.

4.1.7 Bandwidth selection via cross-validation in local polynomial regression

We begin here by examining the average squared error $ASE(h)$, defined by equation (4.19):

$$ASE(h) = n^{-1} \sum_{i=1}^{n} (m(X_i) - \hat{m}_h(X_i))^2 w(X_i)$$

If we were to naively estimate the unknown means $m(X_i)$ with their corresponding observations Y_i, the estimate we obtain for $ASE(h)$ would become

$$r(h) = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{m}_h(X_i)]^2 w(X_i) \; . \tag{4.23}$$

The problem with this estimate is that as $h \to 0$, $\hat{m}_h(X_i) \to Y_i$, and consequently, $r(h) \to 0$. The optimal bandwidth in that case would be the smallest possible bandwidth, i.e. the one that interpolates the data.

One may remedy this problem if in the estimation of $\hat{m}_h(X_i)$ in equation (4.23), we do not use (X_i, Y_i). The *leave-one-out method* is based on omissions of one observation,

say the i th one, from the regression smoother:

$$\hat{m}_{h,-i}(X_i) = \sum_{j \neq i}^{n} w_{hj}(X_i) Y_j.$$

If we use these modified smoothers in equation (4.23), we are led to the so-called cross-validation bandwidth selection criterion

$$CV(h) = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{m}_{h,-i}(X_i)]^2 w(X_i). \tag{4.24}$$

We choose the value of h that minimizes the $CV(h)$ function.

The computation of $CV(h)$ can be computationally intensive since it requires the values of $\hat{m}_{h,-i}(X_i)$ for all $i = 1, \ldots, n$. However, if $S = S(h)$ is the $n \times n$ regression *"hat"* matrix, so that, $\hat{\mathbf{m}} = S\mathbf{Y}$ where S depends on the X variates only, then we can write

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{m}_h(X_i)}{1 - S_{ii}} \right)^2 w(X_i) \tag{4.25}$$

which is much easier to compute than (4.24). Here S_{ii} are the diagonal elements of S. This equation holds because typically

$$\hat{m}_{h,-i}(X_i) - Y_i = \frac{\hat{m}_h(X_i) - Y_i}{1 - S_{ii}} \, . \tag{4.26}$$

See, for example, Heckman [28]. The $CV(h)$ function for the Nadaraya-Watson estimator, for example, can be written in the form of equation (4.25), where $S_{ij} = K((X_j - X_i)/h) / \sum_l K((X_l - X_i)/h)$, provided that $K(0) > 0$ and $S_{ii} \neq 1$.

4.1.8 A rule of thumb bandwidth selection method for the local polynomial regression

A simple, but a rather crude bandwidth selection strategy is as follow. First consider the $MISE(h)$:

$$MISE(h) = \int \{ \text{bias}^2(\hat{m}_p(x) | X_1, \ldots, X_n) + \text{Var}(\hat{m}_p(x) | X_1, \ldots, X_n) \} f(x) w(x) dx$$

where $w \geq 0$ is some weight function. Here, the order of the polynomial p is an odd number. Then the bandwidth chosen to minimize $MISE(h)$, $\hat{h}_{opt}$, is given by

$$\hat{h}_{opt} = C_p(K) \left[\frac{\int \sigma^2(x) w(x) dx}{n \int \{m^{(p+1)}(x)\}^2 w(x) f(x) dx} \right]^{1/(2p+3)} \tag{4.27}$$

where

$$C_p(K) = \left[\frac{(p+1)!^{\,2} \int K_p^{*^2}(t) dt}{2(p+1)\{\int t^{p+1} K_p^*(t) dt\}^2} \right]^{1/(2p+3)}$$

and where

$$K_p^*(t) = \sum_{l=0}^{p} t^l K(t) \; .$$

(See Fan and Gijbels (1996), p. 67-68, 111). The unknown quantities which we need to estimate are $\sigma^2(x)$, $f(x)$ and $m^{(p+1)}(x)$. Begin by fitting a polynomial of order $p+3$ globally to $m(x)$, say $\check{m}(x)$, then take the standardized residuals sum of square as an estimate of $\sigma^2(x)$. Denote this estimate by $\check{\sigma}^2$. The denominator of (4.27) can then be estimated by

$$\frac{1}{n} \sum_{i=1}^{n} \{\check{m}^{(p+1)}(X_i)\}^2 w(X_i)$$

which leads to the simple rule of thumb bandwidth selector

$$h_{ROT} = C_p(K) \left[\frac{\check{\sigma}^2 \int w(x) dx}{\sum_{i=1}^{n} \{\check{m}^{(p+1)}(X_i)\}^2 w(X_i)} \right]^{1/(2p+3)} \; .$$

4.2 Generalized linear models and local smoothing

Generalized linear models (GLM) are particular extensions of the normal linear regression model to the case where the response variables may have distributions other than the normal – either continuous or discrete – and furthermore where the relationship between the response and the explanatory variables are not necessarily of a simple linear form. Because many of the 'nice' properties of the normal distribution family are shared by exponential families, the models considered in generalized linear

models generally assume that the conditional distribution of the response variable Y, given the associated covariate variables X, has (for a fixed, although not necessarily known values of, ϕ) an exponential family form

$$f(y|x) = \exp[\ \{\theta(x)y - b(\theta(x))\}/a(\phi) + c(y,\phi)\ ] \tag{4.28}$$

where $a(\cdot), b(\cdot)$ and $c(\cdot)$ are known functions. The parameter $\theta(\cdot)$ is called the canonical parameter, and ϕ the dispersion parameter. The conditional mean and variance in this model can be shown to be

$$m(x) = E(Y|X = x) = b'(\theta(x)) \ \ , \tag{4.29}$$

and

$$\mathrm{Var}(Y|X = x) = a(\phi)b''(\theta(x)) \ . \tag{4.30}$$

In parametric GLM, the unknown regression function $m(x)$ is typically modelled as

$$g(m(x)) = x^t\beta$$

where in view of (4.29), $g(\cdot)$ may be thought of as a function that links the conditional regression mean to a linear predictor. If $g = (b')^{-1}$, then g is referred to as the canonical link function and we then have $\theta(x) = x^t\beta$.

For our work, we will specifically require the GLM based on conditional binomial distributions where our aim is to model the probability of occurrence of words (shingles) in the DEEDS documents as a function of time. Therefore, in this model, we are interested in the mean of the sample proportion rather than in the mean number of successes. Let r be the number of trials, and Y be the number of successes in the r trials, and let X be a predictor variable such that $Y \sim B(r, \pi(X))$, where $\pi(X)$ is the probability of success. Letting $Y^* = Y/r$ be the sample proportion, we are interested

in estimating $E(Y^*|X,r)$. Let $(X_i, r_i, Y_i),\ i = 1,\ldots,n$ be independent samples drawn from the same population as (X,r,Y), and let $Y_i = r_i Y_i^* \sim B(r_i, \pi(X_i))$ where Y_i^* is the sample proportion at X_i. The joint conditional log-likelihood of the $(Y_i^*|X_i, r_i)$ may then be written in the form

$$\begin{aligned} & l(Y_1^*,\ldots,Y_n^*|(X_i,r_i) : i = 1,\ldots,n) \\ & = \sum_{i=1}^{n} \left\{ \frac{Y_i^*\theta(X_i) - \log(1+\exp(\theta(X_i)))}{1/r_i} + \log \binom{r_i}{r_i Y_i^*} \right\} \end{aligned}$$

where $\theta(X_i) = \log\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right)$. This is of the form (4.28) with $b(\theta(X_i)) = \log(1 + \exp(\theta(X_i)))$, $\phi_i = r_i, a_i(\phi_i) = 1/r_i$, and $c_i(Y_i^*, \phi_i) = \log\binom{r_i}{r_i Y_i^*}$

4.2.1 Local polynomial kernel regression for generalized linear models

Our main reference for this subsection is Fan and Gijbels (1996,2000). Suppose the function $\theta(\cdot)$ is the canonical parameter in a generalized linear model based on observed data $\{(X_i, Y_i), i = 1,\ldots,n\}$ drawn from a population (X,Y), where Y_i is the scalar response, and X_i the associated covariate vector. The function $\theta(\cdot)$ can for example be the logit transform of the conditional probabilities in a binomial model. If $g(\cdot)$ is the canonical link function, we will have $\theta(x) = g(x;\beta)$ and this is modelled as a linear combination of predictors. Specifically, $g(x;\beta)$ will be chosen to be a polynomial of degree at most p in the predictor variable, and β is the vector of coefficients of this polynomial. We then attempt to maximize over the β's the conditional log-likelihood which we write in the from

$$\sum_{i=1}^{n} l\{X_i, Y_i, g(X_i;\beta)\} \ . \tag{4.31}$$

The generalized linear model approach to modelling the link structure suffers from the same deficiencies as does parametric regression, namely insufficient flexibility. The

(generalized) local polynomial regression method generalizes the global parametric modelling approach of equation (4.31). (In that equation, the polynomial in the predictors is considered to be fixed over their entire domain). Specifically, we assume that the function g can be approximated in the neighbourhood of any point x by a polynomial of degree p,

$$\theta(u) = g(u;\theta) \approx \beta_0 + \beta_1(u-x) + \cdots + \beta_p(u-x)^p \ ,$$

where $\beta_j = \frac{\theta^{(j)}(x)}{j!}$. In the above equation, the dependence of the β's on x has been suppressed. We then use a kernel function $K_h(X_i - x)$ to weight the contribution of each data point (X_i, Y_i) according to the distance of X_i from x. We are thus led to maximize (with respect to the β's) the *local log-likelihood* which is defined as

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

$$= \sum_{i=1}^{n} l(X_i, Y_i, \beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p)K_h(X_i - x) \ . \qquad (4.32)$$

Let $\hat{\beta}_j$, $j = 0, \ldots, p$ optimize (4.32). Note that in particular, $\hat{\beta}_0$ then estimates $\theta(x)$.

We now apply the local polynomial approach to the case of the binomial distribution model discussed in the previous section. In that model, we have the conditional distribution

$$(Y_i|X_i, r_i) \sim B(r_i, \pi(X_i)) \ .$$

If we use the canonical link function (the logit), we will have

$$\theta(x) = \log\{\pi(x)/(1-\pi(x))\}$$

and

$$b(\theta) = \log\{1 + \exp(\theta)\} \ .$$

Define Y_i^* via $Y_i = r_i Y_i^*$. Then, (4.32) reduces to

$$\begin{aligned} L(\beta(x)) &= \sum_{i=1}^{n} l\{X_i, Y_i^*, g(X_i;\beta)\} K_h(X_i - x) \\ &= \sum_{i=1}^{n} \left\{ \frac{Y_i^* \theta(X_i) - \log(1 + \exp(\theta(X_i)))}{1/r_i} + \log \binom{r_i}{r_i Y_i^*} \right\} K_h(X_i - x) \\ &= \sum_{i=1}^{n} \{(\beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p) Y_i \\ &\qquad -r_i b(\beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p) \\ &\qquad + \log \binom{r_i}{Y_i} \Big\} K_h(X_i - x) \quad . \end{aligned} \tag{4.33}$$

In the case of locally constant regression ($p = 0$), the expression above becomes

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

and the maximizing value of β_0 is given by

$$\hat{\beta}_0 = \log \left(\frac{\sum_{i=1}^n Y_i K_h(X_i - x) / \sum_{i=1}^n r_i K_h(X_i - x)}{1 - \sum_{i=1}^n Y_i K_h(X_i - x) / \sum_{i=1}^n r_i K_h(X_i - x)} \right) \tag{4.34}$$

so that

$$\hat{\pi}(x) = \hat{\pi}_h(x) = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)} = \frac{\sum_{i=1}^n Y_i K_h(X_i - x)}{\sum_{i=1}^n r_i K_h(X_i - x)} \quad . \tag{4.35}$$

If the number of trials r_i is the same for all i's, then the above equation actually corresponds to the Nadaraya-Watson estimator. (In fact, it is the case that the locally constant kernel-weighted estimator ($p = 0$) results in the Nadaraya-Watson estimator for all exponential family models.)

If we maximize equation (4.33) based on expansion of $\theta(\cdot)$ up to linear order ($p = 1$), then $L(\beta(x))$ becomes

$$f(\beta_0, \beta_1) =$$

$$\sum_{i=1}^{n} \left[Y_i(\beta_0 + \beta_1(X_i - x)) - r_i b(\beta_0 + \beta_1(X_i - x)) + \log \binom{r_i}{Y_i} \right] K_h(X_i - x) \; . \tag{4.36}$$

and we need to determine

$$\arg\max_{\beta_0,\beta_1} f(\beta_0,\beta_1).$$

The partial derivatives of f with respect to β_0 and β_1, are

$$\frac{\partial f}{\partial \beta_0}(\beta_0,\beta_1) = \sum_{i=1}^{n}\{Y_i - r_i b'(\beta_0+\beta_1(X_i-x))\}K_h(X_i-x) \tag{4.37}$$

$$\frac{\partial^2 f}{\partial^2 \beta_0}(\beta_0,\beta_1) = \sum_{i=1}^{n} -r_i b''(\beta_0+\beta_1(X_i-x))K_h(X_i-x) \tag{4.38}$$

$$\frac{\partial f}{\partial \beta_1}(\beta_0,\beta_1) = \sum_{i=1}^{n}\{Y_i - r_i b'(\beta_0+\beta_1(X_i-x))\}(X_i-x)K_h(X_i-x) \tag{4.39}$$

$$\frac{\partial^2 f}{\partial^2 \beta_1}(\beta_0,\beta_1) = \sum_{i=1}^{n} -r_i (X_i-x)^2 b''(\beta_0+\beta_1(X_i-x))K_h(X_i-x) \tag{4.40}$$

$$\frac{\partial^2 f}{\partial \beta_0 \partial \beta_1}(\beta_0,\beta_1) = \sum_{i=1}^{n} -r_i (X_i-x) b''(\beta_0+\beta_1(X_i-x))K_h(X_i-x) \tag{4.41}$$

where $b'(z) = \exp(z)/1+\exp(z)$ and $b''(z) = \exp(z)/(1+\exp(z))^2$. Setting (4.37) and (4.39) to zero, the optimal values $\hat{\beta}_0$ and $\hat{\beta}_1$ will be the solutions to the equations

$$\sum_{i=1}^{n} Y_i K_h(X_i-x) = \sum_{i=1}^{n} \frac{r_i \exp(\beta_0+\beta_1(X_i-x))}{1+\exp(\beta_0+\beta_1(X_i-x))} \times K_h(X_i-x) \tag{4.42}$$

and

$$\begin{aligned} & \sum_{i=1}^{n} Y_i (X_i-x) K_h(X_i-x) \\ = & \sum_{i=1}^{n} \frac{r_i \exp(\beta_0+\beta_1(X_i-x))}{1+\exp(\beta_0+\beta_1(X_i-x))}(X_i-x) \times K_h(X_i-x)\ . \end{aligned} \tag{4.43}$$

The solutions for $\hat{\beta}_0$ and $\hat{\beta}_1$ need to be found using a numerical method such as Newton-Raphson. We may set the initial value of β_0 to be the solution for the local

polynomial estimator $p = 0$, equation (4.34), together with $\beta_1 = 0$. Similar to the case of local polynomial of degree $p = 0$, $\pi(x)$ is then estimated as

$$\hat{\pi}(x) = \hat{\pi}_h(x) = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)} \ . \tag{4.44}$$

We claim that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the global maximal values of f. To prove this, we apply the second partial derivative test (see Adams (1987), p. 135). Letting $c_i = -r_i b''(\hat{\beta}_0 + \hat{\beta}_1(X_i - x))K_h(X_i - x)$ we have that

$$\frac{\partial^2 f}{\partial^2 \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} c_i$$

$$\frac{\partial^2 f}{\partial^2 \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} c_i (X_i - x)^2$$

$$\frac{\partial^2 f}{\partial \beta_0 \partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} c_i (X_i - x) \ .$$

The determinant D of the Hessian matrix of $f(\beta_0, \beta_1)$ is

$$\begin{aligned}
D &= \{\frac{\partial^2 f}{\partial^2 \beta_0}(\hat{\beta}_0, \hat{\beta}_1)\}\{\frac{\partial^2 f}{\partial^2 \beta_1}(\hat{\beta}_0, \hat{\beta}_1)\} - \{\frac{\partial^2 f}{\partial \beta_0 \partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1)\}^2 \\
&= \left[\sum_{i=1}^{n} c_i \sum_{j=1}^{n} c_j (X_j - x)^2\right] - \left[\sum_{i=1}^{n} c_i (X_i - x)\right]^2 \\
&= \sum_{1 \le i < j \le n} \left\{c_i c_j (X_i - x)^2 + c_i c_j (X_j - x)^2 - 2 c_i c_j (X_i - x)(X_j - x)\right\} \\
&= \sum_{1 \le i < j \le n}^{n} c_i c_j \left[(X_i - x) - (X_j - x)\right]^2 \\
&\ge 0 \ .
\end{aligned}$$

If $K_h(\cdot)$ is a positive function, then $c_i < 0$ whenever $r_i > 0$ and we will have the strict inequality $D > 0$ provided the X_i are not all equal. (In the next chapter, it will be made clear why such assumptions are valid in the context of dating the DEEDS documents). Since $\frac{\partial^2 f}{\partial^2 \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} c_i < 0$ and $D > 0$ regardless of the particular values of (β_0, β_1), then $(\hat{\beta}_0, \hat{\beta}_1)$ is in fact the global maximum of (4.36).

In Chapter 6, we will use local polynomial kernel regression for the binomial model to estimate the prevalence (frequency) across time of particular shingles from the DEEDS documents. The cases will be the documents $\mathcal{D}$ of the DEEDS data set. The covariate X will represent the date in which a given document $\mathcal{D}$ is written, r will represent the total number of not necessarily distinct shingles in a set of a shingled document $\mathcal{D}$. (From Section 3.3.2, recall that this set is denoted by $s_k(\mathcal{D})$). Y/r will represent the proportion of time a given shingle appears in the document.

4.2.2 Properties of the estimator of the canonical parameter curve in the GLM model

We now provide some analysis of the asymptotic properties of the estimated binomial proportion function $\hat{\pi}_h(x)$ defined at (4.35) and (4.44). But first, we briefly describe the relevance of this estimator in our study of the DEEDS documents. (This will be expanded upon in Chapter 6). Let $\mathcal{D}$ be a document in a test set $\mathcal{A}$, and let $t_{\mathcal{D}}$ be the date in which the document $\mathcal{D}$ was written. (This date is presumed to be unknown so we are trying to estimate it). The occurrences of some given shingle s of order k in document $\mathcal{D}$ is assumed to be binomially distributed, where the number of trials is the total number shingles of order k of $\mathcal{D}$, and $\pi_h(t)$, the probability of success (i.e. of observing the shingle s at time t) is estimated using (4.35) or (4.44) (applied to the shingle counts in the training data) in the cases that the local polynomial fitting is of degree $p = 0$, or $p = 1$, respectively. To denote the dependence of $\hat{\pi}_h(x)$ on the shingle s, we write it as $\pi_{s,h}(t)$. A proposed method for obtaining a date estimate $\hat{t}_{\mathcal{D}}$ for the document $\mathcal{D}$, with some modifications which we will discuss later, is a likelihood-type procedure

$$\hat{t}_{\mathcal{D}} = \arg\max_t \prod_{s \in s_k(\mathcal{D})} \hat{\pi}_{s,h}(t) \prod_{s \notin s_k(\mathcal{D})} (1 - \hat{\pi}_{s,h}(t)).$$

The properties of the estimator $\hat{t}_{\mathcal{D}}$ will be discussed in Chapter 6 where we will also elaborate on the assumptions made regarding the set of all of the DEEDS documents.

We will examine the asymptotic distribution of $\hat{\pi}_{s,h}(x)$ as n increases to infinity, and the bandwidth h decreases to 0. Recall that $\hat{\pi}_{s,h}(t)$ is an estimate of $\pi(x) = m(x)$ (for a fixed shingle s) where

$$m(x) = E(Y^*|X = x, r) = b'(\theta(x)),$$

and where $Y^* = Y/r$, $(Y|X, r) \sim \mathrm{B}(r, \pi(x))$, $b(x) = \log(1+\exp(x))$, $\theta(x) = \log\{\pi(x)/(1-\pi(x))\}$, and

$$\mathrm{Var}(Y^*|X = x, r) = b''(\theta(x))/r \ .$$

We will also examine the forms of the bias when x is an interior point and when x is a boundary point. Issues of bias at the boundary are quite important for the analysis of the DEEDS data set due to the limited range of dates for those manuscripts, and due to the nature of the empirical design density of their dates. Except for some minor modifications, the results here are from Fan, Heckman and Wand (1995), and Fan and Gijbels (1996). We focus, in particular, on polynomials of degree $p = 0$ and $p = 1$ in the local polynomial modelling approach, although the results may be easily extended to polynomials of arbitrary degree.

In order to keep our discussions general, let $L(\beta(x))$ be the local log-likelihood as defined in equation (4.32), and let $m(x)$ and $\mathrm{Var}(Y|X = x)$ be the functions as stated in (4.29) and (4.30) respectively. The main results use the following conditions:
(1) $\mathrm{Var}(Y|X = x)$, and the design density function $f(\cdot)$, are continuous.
(2) $\mathrm{Var}(Y|X = x) > 0$
(3) $E(Y^4|X = x)$ is bounded in a neighbourhood of x.

We first discuss the case when x is an interior point of the design.

Theorem 2 (Fan *et al.* (1995)) *Suppose the conditions 1, 2 and 3 above hold, and that* $p = 0$. *If* x *is an interior point of the design density, then as* $h \to 0$ *and* $nh \to \infty$,

$$\frac{\sqrt{nh}}{\sigma(x;K)}\left[(\hat{\beta}_0 - \beta_0) - \frac{h^2\mu_2(K)}{b''(\beta_0)}\left\{m'(x)\frac{f'(x)}{f(x)} + \frac{m''(x)}{2}\right\} + o(h^2)\right] \xrightarrow{d} N(0,1)$$

where $\mu_\ell(K) = \int z^\ell K(z)dz$, and

$$\sigma^2(x;K) = \{b''(\beta_0)f(x)\}^{-1}\int K^2(z)dz.$$

Note here that $b''(\beta_0) = \mathrm{Var}(Y|X = x)$.

Theorem 3 (Fan *et al.* (1995)) *Suppose the conditions 1,2 and 3 above hold, and that* $p = 1$. *If* x *is an interior point of the design density, then as* $h \to 0$ *and* $nh \to \infty$,

$$\frac{\sqrt{nh}}{\sigma(x;K)}\left(\hat{\beta}_0 - \beta_0 - \beta_2 h^2\mu_2(K) + o(h^2)\right) \xrightarrow{d} N(0,1)$$

where $\sigma^2(x;K)$ is as stated above, and $\mu_\ell(K) = \int z^\ell K(z)dz$. Here, $\beta_2 = \frac{\theta^{(2)}(x)}{2!}$.

As we are ultimately interested in the form of the bias and variance of the estimator $\hat{m}_p(x) = b'(\hat{\beta}_0)$, some further elaboration on Theorems 2 and 3 is in order. We focus our attention on the case $p = 0$, i.e. on $\hat{m}_0(x)$, since the same argument may be used to find the asymptotic bias and variance of $\hat{m}_1(x)$.

Let

$$C_h(x) = \frac{h^2\mu_2(K)}{b''(\beta_0)}\left\{m'(x)\frac{f'(x)}{f(x)} + \frac{m''(x)}{2}\right\} + o(h^2)$$

denote the bias of $\hat{\beta}_0$ as given in Theorem 2. Taylor expanding $b'(\hat{\beta}_0)$ about the point $\beta_0 + C_h(x)$ gives,

$$b'(\hat{\beta}_0) = b'(\beta_0 + C_h(x)) + b''(\beta_h^*)(\hat{\beta}_0 - \beta_0 - C_h(x))$$

where β_h^* lies between $\hat{\beta}_0$ and $\beta_0 + C_h(x)$. Again, using a Taylor expansion of $b'(\beta_0 + C_h(x))$ about β_0, we have

$$b'(\beta_0 + C_h(x)) = b'(\beta_0) + b''(\beta_0)C_h(x) + o(h^3).$$

Therefore,

$$\sqrt{nh}\left(b'(\hat{\beta}_0) - b'(\beta_0) - b''(\beta_0)C_h(x) + o(h^3)\right) = \sqrt{nh}\, b''(\beta_h^*)(\hat{\beta}_0 - \beta_0 - C_h(x)).$$

Since β_h^* lies between $\hat{\beta}_0$ and $\beta_0 + C_h(x)$, and $C_h(x) \to 0$, it follows that $\beta_h^* \to \beta_0$. Therefore, by the continuity of the function b'', we have $b''(\beta_h^*) \to b''(\beta_0)$. Now, applying Slutsky's Theorem to Theorem 2, it follows that

$$\begin{aligned}
&\sqrt{nh}\left(b'(\hat{\beta}_0) - b'(\beta_0) - b''(\beta_0)C_h(x) + o(h^3)\right) \\
&= \sqrt{nh}\, b''(\beta_h^*)(\hat{\beta}_0 - \beta_0 - C_h(x)) \\
&\xrightarrow{d} b''(\beta_0)N\left(0, \sigma^2(x;K)\right) \stackrel{d}{=} N\left(0, \{b''(\beta_0)\}^2\sigma^2(x;K)\right)
\end{aligned}$$

as $h \to 0$ and $nh \to \infty$.

For the case $p = 1$, we let $D_h(x) = \beta_2 h^2 \mu_2(K) + o(h^2)$ as in Theorem 3. Then, following a similar argument leads to

$$\begin{aligned}
&\sqrt{nh}\left(b'(\hat{\beta}_0) - b'(\beta_0) - b''(\beta_0)D_h(x) + o(h^3)\right) \\
&\xrightarrow{d} b''(\beta_0)N\left(0, \sigma^2(x;K)\right) \stackrel{d}{=} N\left(0, \{b''(\beta_0)\}^2\sigma^2(x;K)\right) .
\end{aligned}$$

The above arguments show that the asymptotic bias and variance of $\hat{m}_0(x)$ has the same behavior as $\hat{\beta}_0$ except that the bias is multiplied by $b''(\beta_0)$ and the variance by $(b''(\beta_0))^2$. The same conclusion also holds for the estimator $\hat{m}_1(x)$.

There are several points to note regarding the leading terms in the asymptotic bias and variance for the locally constant ($\hat{m}_0(x)$) and locally linear ($\hat{m}_1(x)$) estimators. We first restate the asymptotic biases and the variance:

$$\begin{aligned}
\text{Asymptotic Bias}(\hat{m}_0(x)) &= h^2\mu_2(K)\left\{m'(x)\frac{f'(x)}{f(x)} + \frac{m''(x)}{2}\right\} , \\
\text{Asymptotic Bias}(\hat{m}_1(x)) &= b''(\beta_0)\beta_2 h^2 \mu_2(K) ,
\end{aligned}$$

and, for both $p = 0, 1$:

$$\text{Asymptotic Var}(\hat{m}_p(x)) = \{nhf(x)\}^{-1} b''(\beta_0) \int K^2(z)dz . \tag{4.45}$$

We see from the above, that for both the locally constant and locally linear fits, sparser regions of the design density $f(x)$ result in larger variance of the estimator. We also see that the asymptotic bias of the estimator based on the locally linear fit, $\hat{m}_1(x)$, does not depend on the underlying density $f(x)$ of X, which in many ways, is an attractive property. In the locally constant fit case, we see that the bias increases if the true regression $m(x)$ has a high value of curvature, $m''(x)$, at x. In both the local constant and linear fit cases, the bias depends on h^2, reflecting the fact that more smoothing results in increased bias.

For both $p = 0$ and $p = 1$, the variance of $\hat{m}_p(x)$ is the same. It is approximately proportional to the inverse of the sample size times the design density near x, with sparser design regions leading to higher variability in the estimated conditional mean. The $b''(\beta_0)$ term in the variance also confirms the intuition that higher values of the conditional variance leads to higher variability in the estimated conditional mean.

If $L(\beta(x))$ is the binomial log-likelihood given in (4.33) (rather than the general one given in (4.32)), and if we assume that the number of binomial trials $\{r_i\}_{i=1}^n$ are drawn from some distribution and are i.i.d., and furthermore, that the r_i's are independent of the X_i's, then the only modifications we would require in Theorems 2 and 3 is that the variance formulas would each have to be multiplied by $(E(r_1))^{-1}$.

We next consider the behavior of the locally constant ($p = 0$) estimator and the locally linear ($p = 1$) estimator in the boundary regions. As in the case of the locally smoothed kernel regression, we assume that the support of the kernel K lies in $[-1, 1]$ and the support of the underlying density f lies in $[0, 1]$. When x is in the left boundary, i.e. $x = \alpha h$, for $0 \leq \alpha < 1$, define

$$\mu_{\ell,\alpha}(K) = \int_{-\alpha}^{1} z^\ell K(z)dz, \quad \text{and} \quad \nu_{\ell,\alpha}(K) = \int_{-\alpha}^{1} z^\ell K^2(z)dz \ ,$$

and if x is in the right boundary, i.e. $x = 1 - \alpha h$, the integration of the above is from

-1 to α.

Theorem 4 (Fan *et al.* (1995)) *Suppose conditions 1, 2 and 3 above hold and that* $p = 0$. *If* x *is a boundary point of the design density, then as* $h \to 0$ *and* $nh \to \infty$,

$$\frac{\sqrt{nh}}{\sigma_{0,\alpha}(x;K)}\left[(\hat{\beta}_0 - \beta_0) - \frac{m'(x)h\mu_{1,\alpha}(K)}{b''(\beta_0)\mu_{0,\alpha}(K)} + o(1)\right] \xrightarrow{d} N(0,1)$$

where

$$\sigma^2_{0,\alpha}(x;K) = \frac{\nu_{0,\alpha}(K)}{b''(\beta_0)f(x)\mu^2_{0,\alpha}(K)}.$$

Theorem 5 (Fan *et al.* (1995)) *Suppose conditions 1, 2 and 3 above hold and that* $p = 1$. *If* x *is an interior point of the design density, then as* $h \to 0$ *and* $nh \to \infty$,

$$\frac{\sqrt{nh}}{\sigma_{1,\alpha}(x;K)}\left(\hat{\beta}_0 - \beta_0 - \quad \beta_2 h^2 \frac{\mu_{2,\alpha}(K)}{\mu_{0,\alpha}(K)} + o(h^2)\right) \xrightarrow{d} N(0,1)$$

where

$$\sigma^2_{1,\alpha}(x;K) = \frac{\left(\mu^2_{2,\alpha}(K)\nu_{0,\alpha}(K) - 2\mu_{1,\alpha}(K)\mu_{2,\alpha}(K)\nu_{1,\alpha}(K) + \mu^2_{1,\alpha}(K)\nu_{2,\alpha}(K)\right)}{b''(\beta_0)f(x)\left(\mu_{0,\alpha}(K)\mu_{2,\alpha}(K) - \mu^2_{1,\alpha}(K)\right)^2}\ .$$

(Here again, if $L(\beta(x))$ is the binomial log-likelihood given in (4.33), then each of the above variance formulas would have to be multiplied by $E(r_1)^{-1}$.) Using arguments similar to those following Theorems 2 and 3, we find the asymptotic biases and variances of $\hat{m}_0(x)$ and $\hat{m}_1(x)$ when x is a boundary point to be:

$$\begin{aligned}
\text{Asymptotic Bias}(\hat{m}_0(x)) &= m'(x)h\mu_{1,\alpha}(K)/\mu_{0,\alpha}(K)\ , \\
\text{Asymptotic Bias}(\hat{m}_1(x)) &= b''(\beta_0)\beta_2 h^2\mu_{2,\alpha}(K)/\mu_{0,\alpha}(K)\ ,
\end{aligned}$$

and, for $p = 0$ and $p = 1$,

$$\text{Asymptotic Var}(\hat{m}_p(x)) = (nh)^{-1}(b''(\beta_0))^2\sigma^2_{p,\alpha}(x;K)\ .$$

We see that for the case $p = 0$, the leading bias term is of order $O(h^2)$ in the interior, but $O(h)$ in the boundary. Therefore, in terms of rates of convergence for

the bias of $\hat{m}_0(x)$, the location of x makes a difference. When $p = 1$ on the other hand, the bias of $\hat{m}_1(x)$ is proportional to h^2 whether x is an interior point or a boundary point, and so no boundary adjustment for bias is necessary.

Chapter 5

Calendaring by Distance Based Methods

5.1 Calendaring via distance methods

In this chapter we present the first of our two key approaches for dating the documents in the DEEDS data set. Some of the work in this chapter has been published in Feuerverger, Hall, Tilahun and Gervers (2005, 2008).

To place our discussion here in the context of the previous chapters, we consider the following heuristic viewpoint. We wish to associate to each document in the data set an estimated date, so we are looking for a "model" of the form

$$t = m(\mathcal{D}) + error.$$

We are thus in some sense attempting to regress "date" on "document". The approach we will use here is inspired by the local polynomial models. Of course, this idea cannot be made rigorous without a formal definition of the document space, which we will not attempt here. One consequence of this will be that, for the purposes of this

approach, we will be limiting ourselves to using locally constant models. The second approach that we adopt in the next chapter will not have this limitation.

We begin by dividing up the set of DEEDS documents into three parts: a training set, a validation set, and a test set. Letting $\mathcal{T}$, $\mathcal{V}$ and $\mathcal{A}$ denote the set of document indices which identify the training, validation and test sets, respectively, we will say that document $\mathcal{D}_i$ belongs to the training set if its index $i \in \mathcal{T}$, and likewise that it belongs to the validation set or the test set if $i \in \mathcal{V}$ or $i \in \mathcal{A}$, respectively. Following the definition of *shingle* given in Chapter 3, Section 3.3.2, for a given document $\mathcal{D}$, define $\mathcal{S}_k(\mathcal{D})$ to be the set of distinct k-shingles of $\mathcal{D}$ and define $s_k(\mathcal{D})$ to be the set of all k-shingles of $\mathcal{D}$.

The method proposed in this chapter will be based on measures of similarity between documents. (Various distance measures on documents were described in Chapter 3). The underlying idea is that document distance measures capture the essential differences between documents. Suppose then that our aim is to estimate the date t_i of a given document $\mathcal{D}_i$, $i \notin \mathcal{T}$. Let k be the order of the shingle size we are using. Let $d_k(i,j)$ denote the distance between the document $\mathcal{D}_i$ and a document $\mathcal{D}_j$, for some $j \in \mathcal{T}$. Within $\mathcal{T}$, the associated dates t_j are known. We next define a kernel weight on the dates of the documents $\mathcal{D}_j$ based on their distances to the document $\mathcal{D}_i$ as follow:

$$a(i,j) = a(i,j|h_1,\ldots,h_r) = \prod_{k=1}^{r} K_{h_k}(d_k(i,j)) \quad . \tag{5.1}$$

Here K is a non-negative, non-increasing kernel function defined on the positive half-line, and the $h_1,\ldots,h_r$ denote bandwidths. (Note that if the kernel is the normal density function, only the non-increasing range is used.) Note also that in (5.1), all the shingle sizes, from $k=1$ to $k=r$ are now being used. The weight $\prod_{k=1}^{r} K_{h_k}(d_k(i,j))$ proposed here is an r-dimensional multivariate kernel which is a product of symmetric

univariate kernels, and it is therefore also symmetric. The theory for multivariate local polynomial regression is a comparatively minor extension of the univariate case; for further details, see Wand and Jones (1995), and Simonoff (1996).

Suppose the size of the training set is equal to n. Based on the locally constant kernel regression estimator (see Chapter 4, Section 4.1.3), letting the response $Y_j = t_j$ be the date associated with document $\mathcal{D}_j$, $j \in \mathcal{T}$, and letting $(t_1, \ldots, t_n)^t = (1, \ldots, 1)^t + (\epsilon_1, \ldots, \epsilon_n)$ with covariance matrix

$$W = \text{diag}\left[\frac{1}{a\,(i,1)}, \frac{1}{a\,(i,2)}, \ldots, \frac{1}{a\,(i,j)}, \ldots, \frac{1}{a\,(i,n)}\right],$$

we obtain the weighted least squares estimate $\hat{t}_i$ for the date t_i associated with document $\mathcal{D}_i$ given by

$$\begin{aligned} \hat{t}_i \equiv \hat{t} &= \arg\min_t \sum_{j \in \mathcal{T}} (t_j - t)^2 a(i,j) \\ &= \left(\sum_{j \in \mathcal{T}} t_j a(i,j)\right) \Big/ \left(\sum_{j \in \mathcal{T}} a(i,j)\right). \end{aligned} \tag{5.2}$$

The properties of this estimator will be described in Section 5.4.

Here the t_i represent the dates of the documents $\mathcal{D}_i$, but in general, we can think of the t_i's as any attributes of interest, and the $a(i,j)$'s as weights measuring the closeness between the i^{th} and the j^{th} documents. If the attribute of interest is ordered categorical, such as categorical variates measuring "experience" or "ability" of authors, then the variates can be arranged consecutively on a line with relative distances adjusted so as to reflect prior notions of closeness. The estimator (5.2) will then estimate the value of this attribute for the document $\mathcal{D}_i$. If on the other hand, the attributes are unordered categorical, as for example, the case of authorship assignment, then we could assign the m authors to the vertices of an $m-1$ dimensional simplex in Euclidean space. Prior beliefs of closeness between different authors

can be accommodated by distorting the length of the edges that connect them. The estimator (5.2) could then be used to impute authorship to document $\mathcal{D}_i$.

5.2 Bandwidth selection via cross-validation

We propose to use cross-validation to select the bandwidths $h_1, \ldots, h_r$ in (5.1) for estimating the t_i's, the dates associated with the documents $\mathcal{D}_i, i \notin \mathcal{T}$. Let $\mathcal{K}(i)$ denote the union, over $1 \leq k \leq r$, of the set of all indices $j \in \mathcal{T}$ such that $d_k(i,j)$ is among the m smallest values of that quantity, and where the integer m is some small fraction of the total number of documents in the training set. Note that $\mathcal{K}(i)$ is a collection of nearest neighbours to i. We then determine the bandwidth values, as well as the values of m, which minimize the cross-validation function

$$CV(h_1, \ldots, h_r) \;=\; \frac{1}{|\mathcal{K}(i)|} \sum_{j' \in \mathcal{K}(i)} \left(t_{j'} - \hat{t}_{-j'} \right)^2 \quad , \tag{5.3}$$

where

$$\begin{aligned} \hat{t}_{-j'} \;&\equiv\; \hat{t}_{-j'}(h_1, \ldots, h_r) \equiv \arg\min_t \sum_{j \in \mathcal{T}, j \neq j'} (t_j - t)^2 a(j', j) \\ &=\; \left(\sum_{j \in \mathcal{T}, j \neq j'} t_j a(j', j) \right) \Big/ \left(\sum_{j \in \mathcal{T}, j \neq j'} a(j', j) \right) \quad . \end{aligned} \tag{5.4}$$

Note that $\hat{t}_{-j'}$ is a date estimate based on leaving out the j^{th} document from the training set $\mathcal{T}$, and that j' only ranges over the nearby documents in $\mathcal{K}(i)$.

Note that the bandwidth selection process described above is *local* since it attempts to determine an optimal set of bandwidths (on which to base the estimate $\hat{t}_i$ of t_i) separately for each document $\mathcal{D}_i$ where the bandwidths are optimized over its nearest neighbour $\mathcal{K}(i)$. The advantage of this form of cross-validation is that the selected bandwidths are then customized for each document. The search for the optimal

bandwidths

$$(\hat{h}_1, \ldots, \hat{h}_r) = \arg \min_{(h_1,\ldots,h_r)} CV(h_1, \ldots, h_r)$$

were carried out over an r-dimensional grid. See the Appendix for the computing code. We note that if we were to choose $K(i) = \mathcal{T}$, then the optimal bandwidths in the above cross-validation will be *global*, in the sense that $(\hat{h}_1, \ldots, \hat{h}_r)$ will then be the same for all i.

Note also that the cross-validation procedure described above can also be used to estimate the mean square error $s^2(i)$ of the date estimate $\hat{t}_i$. Assuming $(\hat{h}_1, \ldots, \hat{h}_r)$ is the bandwidth vector used to estimate $\hat{t}_i$, where for each $j' \in \mathcal{K}(i)$, $\hat{t}_{j'}$ is computed using the same bandwidth, we may define an estimator of $s^2(i)$ as

$$\hat{s}^2(i) = \left\{ \sum_{j' \in \mathcal{K}(i)} (t_{j'} - \hat{t}_{-j'})^2 a(i, j' | \hat{h}_1, \ldots, \hat{h}_r) \right\} \Big/ \left\{ \sum_{j' \in \mathcal{K}(i)} a(i, j' | \hat{h}_1, \ldots, \hat{h}_r) \right\}. \quad (5.5)$$

5.3 Numerical results

As mentioned in Chapter 1, at the time of this study, the total number of dated documents in the DEEDS data set was 3353. We now describe our original decomposition of the 3353 DEEDS documents into $\mathcal{V}$, $\mathcal{T}$ and $\mathcal{A}$. We systematically took every 8^{th} document to form the validation set ($\mathcal{V}$), and every 9^{th} document to form the test set ($\mathcal{A}$). When a document satisfied both of these criteria, (i.e. every 72^{nd} document), it was assigned only to the validation set. We thus obtained 419 documents for the validation set (12.5% of the total data set), and 326 documents for the test set (9.7% of the total date set). The remaining 2608 documents (77.8% of the total data set) formed the training set ($\mathcal{T}$).

The method used for estimating the date of a document, as described in the previous sections, requires a training set and a test set only. (Following the discussion

on cross-validation (Section 5.2), a validation set would have been required had we needed to estimate prediction error for model selection, as will be the case for the dating methodology we will employ in Chapter 6). For this reason, for the purposes of the methodology described in this chapter, the validation set was combined with the test set to form a larger test set. Therefore the test set for this distance-based study contained 745 documents.

For the reasons discussed in Chapter 3, concerning the relative advantages of correspondence distance measures of type (II) over the other commonly used distance types, we chose to base our work here on the type (II) ($\alpha = 1$) correspondence distance measure. This is the correspondence distance that is geometrically the same as the cosine correspondence distance, but with the additional property of being a metric distance. We began by shingling each of the documents of the training set and of the test set into shingle orders k, where $k = 1, 2$ and 3. For each of these shingle orders, a total of 745×2608 distance computations between the documents in the test set and the documents in the training set were made. For a given single shingle order k and an $i \notin \mathcal{T}$, the set $K(i)$ of neighbours was obtained by taking the set of all indices $j \in \mathcal{T}$ such that $d_k(i, j)$ are among the m smallest values of that distance. The values of m ranged over 5, 10, 20, 100, 500, 1000. For a given pair of shingle orders, say for $k = 2$ and $k = 3$, we determined $K(i)$ by taking the union over the set of all indices $j \in \mathcal{T}$ such that $d_k(i, j)$ is among the m smallest values for either of the shingle orders $k = 2$ or $k = 3$. Here again, m ranged over the values 5, 10, 20, 100, 500, 1000. Throughout, we employed the standard normal kernel $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$, for $x \geq 0$, and the bandwidths were computed using the cross-validation method defined in (5.3). The optimal bandwidths were found by searching over a one, two or three dimensional grid depending on whether $K(i)$ was based one, two or three shingle orders, respectively. Finally, we computed the mean absolute error (MAE) and the

median absolute error (MedAE) in years, as well as the mean squared error (MSE) between the true date and the estimated date for the 745 documents in the test set. Tables 5.1- 5.7 summarize our findings.

In Tables 5.1 - 5.7, we have also included the *concordance of correlation coefficient* between the various date estimates and the true dates. The concordance of correlation coefficient measures the agreement between two variables, X and Y, along the 45 degree $X = Y$ line. It is defined as

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{5.6}$$

where σ_x^2 and σ_y^2 are the variances of the two variables, μ_x and μ_y are their corresponding means, and ρ is the coefficient of correlation between X and Y. (For further details on concordance of correlation coefficient, see Lin (1989) and Lin (2000)).

The results in Tables 5.1 - 5.3 show that shingles of order 1 gives the best results. The MAE decreases slightly as m increases. For shingle order 1, the minimum value of the MAE is 12.26 years and the maximum value is 12.90 years, with corresponding m values of 1000 and 5, respectively. For all values of m, $\sqrt{\text{MSE}}$ is around 20.5 years and the correlation coefficient ρ_c is around 0.9. The value of MedAE on the other hand drops from 7.3 to 6.3 years as m increases.

For shingle order 2, the MAE is around 14 years with an average $\sqrt{\text{MSE}}$ value of around 24 years. The MedAE decreases from around 7.0 years when m=5, 10 and 20, to around 6.4 years for m=500 and 1000. The concordance correlation coefficient ρ_C is around 0.8. For shingle order 3, the MAE is around 17 years with an average $\sqrt{\text{MSE}}$ value of around 28.5 years. The MedAE ranges from around 7.5 years (when m=500 and 1000) to 9 years when m=20. The value ρ_C is around 0.75 for all values of m.

Table 5.1: The values of $\sqrt{MSE}$, MAE, MedAE and ρ_c for shingle order 1, and for m ranging over 5, 10, 20, 100, 500 and 1000, evaluated on a test set of 745 documents.

Shingle order	m	$\sqrt{MSE}$ *(test set)*	*MAE (test set)*	*MedAE (test set)*	ρ_c *(test set)*
1	5	20.89	12.90	7.30	0.88
1	10	20.60	12.64	6.79	0.88
1	20	20.16	12.51	7.03	0.90
1	100	20.11	12.54	6.71	0.89
1	500	20.07	12.33	6.42	0.89
1	1000	20.38	12.26	6.27	0.89

Table 5.2: The values of $\sqrt{MSE}$, MAE, MedAE and ρ_c for shingle order 2, and for m ranging over 5, 10, 20, 100, 500 and 1000, evaluated on a test set of 745 documents.

Shingle order	m	$\sqrt{MSE}$ *(test set)*	*MAE (test set)*	*MedAE (test set)*	ρ_c *(test set)*
2	5	24.40	14.56	6.97	0.83
2	10	24.41	14.55	7.26	0.83
2	20	23.81	14.21	7.22	0.83
2	100	23.76	14.02	6.96	0.84
2	500	23.72	13.79	6.35	0.84
2	1000	23.96	13.90	6.37	0.84

Table 5.3: The values of $\sqrt{MSE}$, MAE, MedAE and ρ_c for shingle order 3, and for m ranging over 5, 10, 20, 100, 500 and 1000, evaluated on a test set of 745 documents.

Shingle order	m	$\sqrt{MSE}$ *(test set)*	*MAE (test set)*	*MedAE (test set)*	ρ_c *(test set)*
3	5	29.02	17.51	8.94	0.74
3	10	28.65	17.04	8.60	0.75
3	20	28.74	17.21	9.00	0.75
3	100	28.40	16.82	8.60	0.75
3	500	28.33	16.55	7.55	0.76
3	1000	28.52	16.64	7.17	0.76

Table 5.4: The values of $\sqrt{MSE}$, MAE, MedAE and ρ_c for the pair of shingle orders 1 and 2, and for m ranging over 5, 10, 20, 100, 500 and 1000, evaluated on a test set of 745 documents.

Shingle orders	m	$\sqrt{MSE}$ *(test set)*	*MAE (test set)*	*MedAE (test set)*	ρ_c *(test set)*
1 & 2	5	21.67	12.89	6.46	0.87
1 & 2	10	20.65	12.20	6.30	0.88
1 & 2	20	20.58	12.30	6.49	0.88
1 & 2	100	20.23	12.10	6.31	0.89
1 & 2	500	20.48	12.12	6.00	0.89
1 & 2	1000	20.90	12.26	6.00	0.88

Table 5.5: The values of $\sqrt{MSE}$, MAE, MedAE and ρ_c for the pair of shingle orders 1 and 3, and for m ranging over 5, 10, 20, 100, 500 and 1000, evaluated on a test set of 745 documents.

Shingle orders	m	$\sqrt{MSE}$ *(test set)*	*MAE (test set)*	*MedAE (test set)*	ρ_c *(test set)*
1 & 3	5	22.42	13.35	7.29	0.86
1 & 3	10	22.09	13.02	7.12	0.86
1 & 3	20	21.75	12.97	7.13	0.87
1 & 3	100	21.73	12.92	6.94	0.87
1 & 3	500	21.80	12.74	6.00	0.87
1 & 3	1000	22.30	12.87	5.90	0.87

Table 5.6: The values of $\sqrt{MSE}$, MAE, MedAE and ρ_c for the pair of shingle orders 2 and 3, and for m ranging over 5, 10, 20, 100, 500 and 1000, evaluated on a test set of 745 documents.

Shingle orders	m	$\sqrt{MSE}$ *(test set)*	*MAE (test set)*	*MedAE (test set)*	ρ_c *(test set)*
2 & 3	5	26.47	15.77	7.85	0.79
2 & 3	10	26.01	15.29	7.10	0.80
2 & 3	20	25.77	15.23	7.49	0.80
2 & 3	100	25.52	14.86	6.80	0.81
2 & 3	500	25.63	14.81	6.19	0.81
2 & 3	1000	25.73	14.81	6.15	0.81

Table 5.7: The values of $\sqrt{MSE}$, MAE, MedAE and ρ_c for a combination of shingle orders 1,2 and 3, and for m ranging over 5, 10, 20, 100, 500 and 1000, evaluated on a test set of 745 documents.

Shingle orders	m	$\sqrt{MSE}$ *(test set)*	*MAE (test set)*	*MedAE (test set)*	ρ_c *(test set)*
1 & 2 & 3	5	25.70	15.22	7.92	0.80
1 & 2 & 3	10	25.38	15.00	7.92	0.80
1 & 2 & 3	20	25.40	15.09	8.18	0.80
1 & 2 & 3	100	25.39	14.93	7.64	0.80
1 & 2 & 3	500	25.47	14.92	7.37	0.81
1 & 2 & 3	1000	25.52	14.92	7.26	0.81

Tables 5.4 - 5.6 show that among the combinations of two shingle orders, orders 1 and 2 give the best MAE results, with average MAE of around 12.3 years. Specifically, the minimum MAE is 12.1 years and occurs when m=100, and the maximum MAE is 12.9 years when m=5. The $\sqrt{\text{MSE}}$ is around 20.7 years and the MedAE is at 6.5 years when $m = 5$ and drops to 6 years when m=500 and 1000. The value of ρ_c also remains stable at around 0.88 for all values of m. The combination of shingle orders of 1 and 3 results in an average MAE of around 13.0 years, and $\sqrt{\text{MSE}}$ of around 22.0 years. The MedAE ranges from 5.9 to 7.1 years, with smaller values of MedAE being attained at larger values of m. The value of ρ_c is stable at 0.86. The combination of shingle orders 2 and 3 results in an average MAE of around 15.0 years. For this combination of shingle orders, the minimum MAE is 14.8 years when m=500 and 1000, and the maximum MAE is 15.8 years when m=5. The $\sqrt{\text{MSE}}$ is around 25.0 years. The MedAE ranges from 7.9 to 6.2, and here again, smaller values of MedAE

are attained at larger values of m. The value of ρ_c is at around 0.80 for all values of m.

For the combination of shingle orders 1,2 and 3, Table 5.7 shows that the value of the $\sqrt{\text{MSE}}$ is around 25.5 years for all values of m, and the MAE is around 15 years. The MedAE is around 8 years, decreasing slightly as m increases. The average value of ρ_c is about 0.80.

As can be seen from the values of $\sqrt{\text{MSE}}$, MAE and ρ_c, the date estimates based on shingle order 1 performed the best among all shingle orders, and the combination of shingle orders 1 and 2 performed the best among the combinations of shingle orders. We note that no significant improvement is gained over the results of shingle order 1 by taking the combination of shingle orders 1 and 2. This is likely related to the fact that the performance based on shingle order 2 alone is poorer than that of shingle order 1. Similarly, the shingle combination 1,2 and 3 performed poorly compared to the shingle 1 order and to the shingle order combination 1 and 2. We also note here that the mean year for the training documents is around 1246 years, and if this value was simply used as the date estimate for the documents in the test set, the MAE would be about 37 years, the $\sqrt{\text{MSE}}$ would be around 47 years and the MedAE would be around 24.5 years. Furthermore, Tables 5.1 - 5.7 also indicate that for given shingle orders, for the most part, the MAE's and MSE's (and to some extent the MedAE's) are fairly robust against the choice of m. This is useful information, since the date estimation can therefore be based on smaller values of m (and hence shorter computation times) without too much sacrifice in accuracy to the date estimation.

Figures 5.1, 5.2 and 5.3 are plots of the (presumed) true document dates versus date estimates based on shingle order 1 ($m = 500$), and on the combination of shingle orders 1 and 2 ($m = 100$), and the combination of shingle orders 1, 2 and 3 ($m = 10$), respectively. In all of the plots, and particularly in Figure 5.3, we can see evidence of

a positive edge bias on the left, and a slightly negative edge bias on the right.

The descriptions of the computer codes used in this section can be found in the appendix.

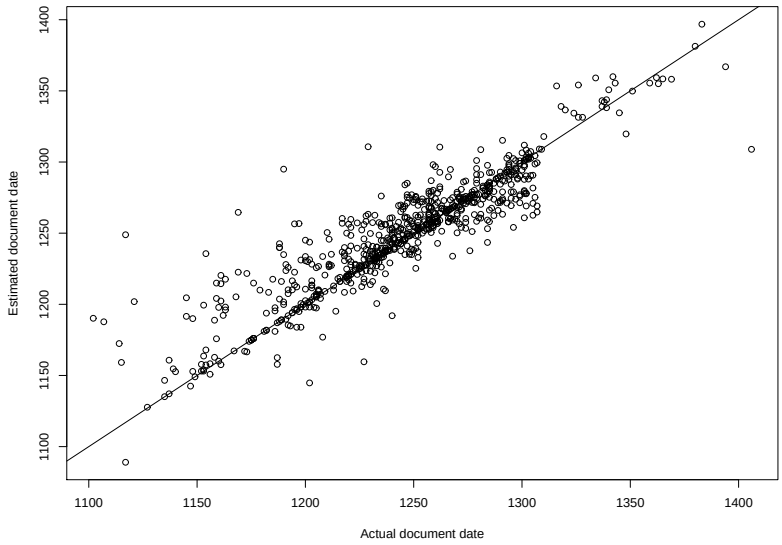


Figure 5.1: True document date versus date estimates for the 745 documents in the test set based on shingle order 1, and m=500. Solid line indicates the X=Y axis.

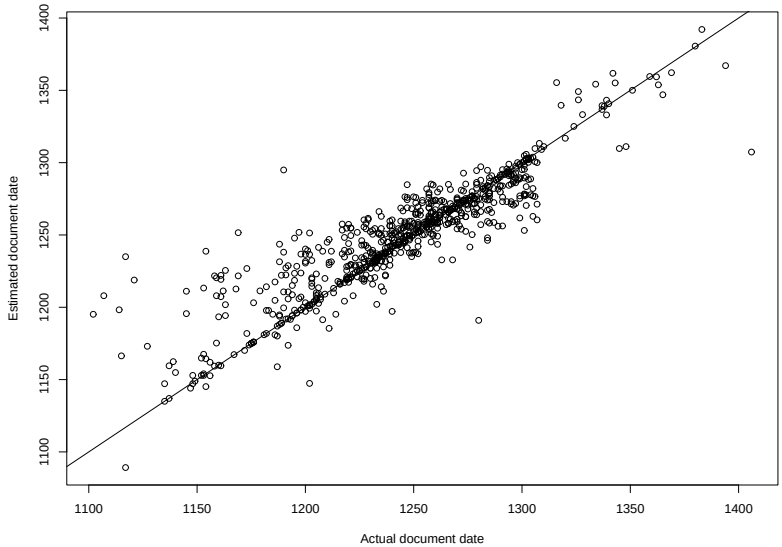


Figure 5.2: True document date versus date estimates for the 745 documents in the test set based on a combination of shingle order 1 and 2, and m=100. Solid line indicates the X=Y axis.

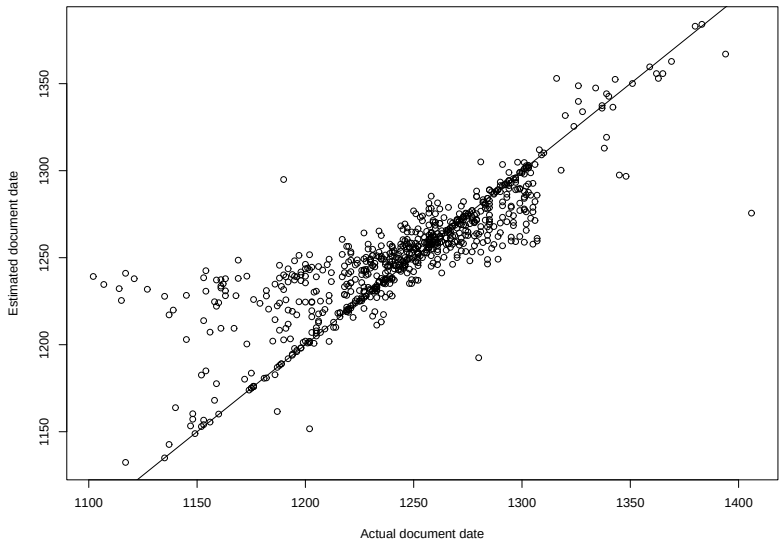


Figure 5.3: True document date versus date estimates for the 745 documents in the test set based on a combination of shingle order 1, 2 and 3, and m=10. Solid line indicates the X=Y axis.

5.4 A Consistency result for the distance-based date estimator

In this section we provide some conditions under which the date estimator (5.2) is a consistent estimator of the true date of a document (see Feuerverger, Hall, Tilahun and Gervers (2008).) Let $\mathcal{D}_0$ be an undated document written at time t_0. Let the document pair $(\mathcal{D}_0, \mathcal{D})$ denote $(\mathcal{D}_i, \mathcal{D}_j)$ where $\mathcal{D}_i = \mathcal{D}_0$ and where $\mathcal{D}_j = \mathcal{D}$ is a randomly chosen, dated document. Let $\triangle_l$ denote the distance $d_l(i, j)$ between the paired documents $(\mathcal{D}_i, \mathcal{D}_j)$. We will assume, that because the possible number of documents that can be produced using the words of the dictionary is large, $d_l(i, j)$ to be continuously distributed in $[0, 1]$. We propose to model the date T of the document

$\mathcal{D}$ and the distances $\triangle_l$ by the vector $(T, \triangle_1, \ldots, \triangle_r)$ distributed continuously in the region $(0, \infty) \times [0, 1]^r$. The training documents are assumed to be i.i.d.

Given the above setting, we will show that the kernel date estimator (5.2) will be a consistent estimator to the unknown date t_0 provided the following four assumptions hold:

(A) For a given neighbourhood of the distances, $\triangle_1, \ldots, \triangle_r$, the expectation of the mean of T converges to t_0 as the size of these neighbourhoods shrinks to zero. This condition is a statement of the "asymptotic unbiasedness" of the date of a random document $\mathcal{D}$. Formally,

$$\frac{E\left(TI(\triangle_1 \leq \delta_1, \ldots, \triangle_r \leq \delta_r)\right)}{P(\triangle_1 \leq \delta_1, \ldots, \triangle_r \leq \delta_r)} \to t_0 \tag{5.7}$$

as $\delta_1, \ldots, \delta_r \to 0$.

(B) For a given neighbourhood of distances, $\triangle_1, \ldots, \triangle_r$, the second moment of T remains bounded as the size of the neighbourhood shrinks to zero. This condition is a statement of an assumption of finite variance. Stated formally,

$$\frac{E\left(T^2I(\triangle_1 \leq \delta_1, \ldots, \triangle_r \leq \delta_r)\right)}{P(\triangle_1 \leq \delta_1, \ldots, \triangle_r \leq \delta_r)} < \infty \tag{5.8}$$

as $\delta_1, \ldots, \delta_r \to 0$.

(C) For each $c > 1$,

$$\limsup_{\delta_1, \ldots, \delta_r \to 0} \frac{P(\triangle_1 \leq c\delta_1, \ldots, \triangle_r \leq c\delta_r)}{P(\triangle_1 \leq \delta_1, \ldots, \triangle_r \leq \delta_r)} < \infty. \tag{5.9}$$

This is a technical condition on the distribution of the dates of the training documents.

(D) We also assume that the kernel K is bounded, continuous, compactly supported, and nonincreasing on the positive real line such that for some $x_0 \geq 0$, $K(x_0) > 0$; that the training documents $\mathcal{D}_j, j \in \mathcal{T}$ are independent and identically distributed as $\mathcal{D}$; and that the number of elements in the training set $N(\mathcal{T})$ increases

to infinity but in a manner such that

$$N(\mathcal{T})P(\triangle_1 \leq \delta_1, \ldots, \triangle_r \leq \delta_r) \to \infty \tag{5.10}$$

as the bandwidth $\delta_1, \ldots, \delta_r \to 0$.

Theorem 6 *If conditions (A) to (D) are satisfied, the estimator $\hat{t}_i$ defined in (5.2) is a consistent estimator of t_0, the true date of document $\mathcal{D}_0$; that is, $\hat{t}_i \xrightarrow{p} t_0$.*

Proof: It suffices to prove that the theorem holds for a kernel K expressible as a finite, positive linear combination of the form $L(x) = I(0 < x \leq c), c > 0$, since functions of this form can be used to approximate a kernel as defined in (D). Let

$$A_n = \sum_{j \in \mathcal{T}} t_j a(i,j) \text{ and } B_n = \sum_{j \in \mathcal{T}} a(i,j),$$

where

$$a(i,j) = \sum_{k=1}^{n} \alpha_k I\left(\triangle_1(i,j) \leq \delta_1^{(k)}, \ldots, \triangle_r(i,j) \leq \delta_r^{(k)}\right).$$

Let

$$\boldsymbol{\delta} = (\delta_1, \ldots, \delta_r), \quad \boldsymbol{\delta}^{(k)} = (\delta_1^{(k)}, \ldots, \delta_r^{(k)}), \quad \triangle = (\triangle_1 \ldots, \triangle_r),$$

and

$$\pi(\boldsymbol{\delta}) = P(\triangle_1 \leq \delta_1, \ldots, \triangle_r \leq \delta_r).$$

We will prove that

$$\frac{A_n}{N(\mathcal{T}) \sum_{k=1}^{n} \alpha_k \pi(\boldsymbol{\delta}^{(k)})} \xrightarrow{p} t_0$$

and

$$\frac{B_n}{N(\mathcal{T}) \sum_{k=1}^{n} \alpha_k \pi(\boldsymbol{\delta}^{(k)})} \xrightarrow{p} 1$$

when the conditions from (A) to (D) are satisfied. The result that $A_n/B_n \xrightarrow{p} t_0$ is then an immediate consequence.

From conditions (A) and (B), it follows that

$$E(A_n) = N(\mathcal{T})E\left(T\sum_{k=1}^{n}\alpha_k I\left(\triangle \le \boldsymbol{\delta}^{(k)}\right)\right) = N(\mathcal{T})\sum_{k=1}^{n}\alpha_k E\left(TI\left(\triangle \le \boldsymbol{\delta}^{(k)}\right)\right)$$

$$= N(\mathcal{T})\sum_{k=1}^{n}\alpha_k\{t_0 + o(1)\}\pi\left(\boldsymbol{\delta}^{(k)}\right),$$

and

$$\begin{aligned}
\mathrm{Var}(A_n) &= N(\mathcal{T})\mathrm{Var}\left(T\sum_{k=1}^{n}\alpha_k I\left(\triangle \le \boldsymbol{\delta}^{(k)}\right)\right) \\
&= N(\mathcal{T})\sum_{k=1}^{n}\alpha_k^2\,\mathrm{Var}\left(TI\left(\triangle \le \boldsymbol{\delta}^{(k)}\right)\right) \\
&\quad +2N(\mathcal{T})\sum_{1\le j<l\le n}\mathrm{Cov}\left(\alpha_j TI\left(\triangle \le \boldsymbol{\delta}^{(j)}\right),\ \alpha_l TI\left(\triangle \le \boldsymbol{\delta}^{(l)}\right)\right) \\
&\le \sum_{k=1}^{n}\alpha_k^2\, O\left(N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(k)}\right)\right) + 2N(\mathcal{T})\sum_{1\le j<l\le n}\alpha_j\alpha_l E\left(T^2 I\left(\triangle \le \mathbf{h}^{(l,j)}\right)\right) \\
&\le \sum_{k=1}^{n}\alpha_k^2\, O\left(N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(k)}\right)\right) + 2\sum_{1\le j<l\le n}\alpha_j\alpha_l O\left(N(\mathcal{T})\pi\left(\mathbf{h}^{(l,j)}\right)\right)
\end{aligned}$$

where $\mathbf{h}^{(l,j)} = \min\left(\boldsymbol{\delta}^{(l)}, \boldsymbol{\delta}^{(j)}\right)$. Moreover,

$$E(B_n) = N(\mathcal{T})E\left(\sum_{k=1}^{n}\alpha_k I(\triangle \le \boldsymbol{\delta}^{(k)})\right) = \sum_{k=1}^{n}\alpha_k N(\mathcal{T})\pi(\boldsymbol{\delta}^{(k)})$$

and

$$\begin{aligned}
\mathrm{Var}(B_n) &= N(\mathcal{T})\mathrm{Var}\left(\sum_{k=1}^{n}\alpha_k I(\triangle \le \boldsymbol{\delta}^{(k)})\right) = N(\mathcal{T})\sum_{k=1}^{n}\alpha_k^2\,\mathrm{Var}\left(\,I(\triangle \le \boldsymbol{\delta}^{(k)})\right) \\
&\quad +2N(\mathcal{T})\sum_{1\le j<l\le n}\alpha_j\alpha_l\mathrm{Cov}\left(\,I(\triangle \le \boldsymbol{\delta}^{(j)}),\ I(\triangle \le \boldsymbol{\delta}^{(l)})\right) \\
&\le N(\mathcal{T})\sum_{k=1}^{n}\alpha_k^2\pi(\boldsymbol{\delta}^{(k)}) + 2N(\mathcal{T})\sum_{1\le j<l\le n}\alpha_j\alpha_l\pi\left(\mathbf{h}^{(l,j)}\right)\quad .
\end{aligned}$$

Using Chebychev's inequality, for a given $\epsilon > 0$,

$$\begin{aligned}
\epsilon^2 P\left(\left|\frac{A_n}{N(\mathcal{T})\sum_{k=1}^{n}\alpha_k\pi(\boldsymbol{\delta}^{(k)})} - t_0\right| > \epsilon\right) &\le E\left(\frac{{A^2}_n}{N^2(\mathcal{T})[\sum_{k=1}^{n}\alpha_k\pi(\boldsymbol{\delta}^{(k)})]^2}\right) \\
&\quad - 2t_0 E\left(\frac{A_n}{N(\mathcal{T})\sum_{k=1}^{n}\alpha_k\pi(\boldsymbol{\delta}^{(k)})}\right) + t_0^2.
\end{aligned}$$

From the previous inequalities and condition (C), it follows that

$$
\begin{aligned}
E\ \left(\frac{A_n^2}{N^2(\mathcal{T})[\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})]^2}\right) &= \frac{\mathrm{Var}(A_n)+E^2(A_n)}{N^2(\mathcal{T})[\ \sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})\ ]^2} \\
&\leq \frac{\sum_{k=1}^n \alpha_k^2\, O\left(N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(k)}\right)\right) + 2\sum_{1\leq j<l\leq n}\alpha_j\alpha_l\, O\left(N(\mathcal{T})\pi\left(\mathbf{h}^{(l,j)}\right)\right)}{N^2(\mathcal{T})[\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})]^2} \\
&\quad + \frac{[\sum_{k=1}^n \alpha_k\{t_0+o(1)\}N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(k)}\right)]^2}{N^2(\mathcal{T})[\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})]^2} \\
&\leq \frac{\alpha_1^2\, O\left(N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(1)}\right)\right)}{N^2(\mathcal{T})\alpha_1^2\pi^2\left(\boldsymbol{\delta}^{(1)}\right)} + \frac{\alpha_2^2\, O\left(N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(2)}\right)\right)}{N^2(\mathcal{T})\alpha_2^2\pi^2\left(\boldsymbol{\delta}^{(2)}\right)} + \cdots + \frac{\alpha_n^2\, O\left(N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(n)}\right)\right)}{N^2(\mathcal{T})\alpha_n^2\pi^2\left(\boldsymbol{\delta}^{(n)}\right)} \\
&\quad + \frac{2\sum_{1<j\leq n}\alpha_j\alpha_1\, O\left(N(\mathcal{T})\pi\left(\mathbf{h}^{(1,j)}\right)\right)}{2N^2(\mathcal{T})\sum_{1<j\leq n}\alpha_j\alpha_1\pi\left(\boldsymbol{\delta}^{(1)}\right)\pi\left(\boldsymbol{\delta}^{(j)}\right)} + \frac{2\sum_{2<j\leq n}\alpha_j\alpha_2\, O\left(N(\mathcal{T})\pi\left(\mathbf{h}^{(2,j)}\right)\right)}{2N^2(\mathcal{T})\sum_{2<j\leq n}\alpha_j\alpha_2\pi\left(\boldsymbol{\delta}^{(2)}\right)\pi\left(\boldsymbol{\delta}^{(j)}\right)} \\
&\quad + \cdots + \frac{2\sum_{n-1<j\leq n}\alpha_j\alpha_{n-1}\, O\left(N(\mathcal{T})\pi\left(\mathbf{h}^{(n-1,j)}\right)\right)}{2N^2(\mathcal{T})\sum_{n-1<j\leq n}\alpha_j\alpha_{n-1}\pi\left(\boldsymbol{\delta}^{(n-1)}\right)\pi\left(\boldsymbol{\delta}^{(j)}\right)} + \{t_0+o(1)\}^2 \\
&= \frac{O\left(N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(1)}\right)\right)}{N^2(\mathcal{T})\pi^2\left(\boldsymbol{\delta}^{(1)}\right)} + \frac{O\left(N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(2)}\right)\right)}{N^2(\mathcal{T})\pi^2\left(\boldsymbol{\delta}^{(2)}\right)} + \cdots + \frac{O\left(N(\mathcal{T})\pi\left(\boldsymbol{\delta}^{(n)}\right)\right)}{N^2(\mathcal{T})\pi^2\left(\boldsymbol{\delta}^{(n)}\right)} \\
&\quad + \frac{\sum_{1<j\leq n}\alpha_j\, O\left(N(\mathcal{T})\pi\left(\mathbf{h}^{(1,j)}\right)\right)}{N^2(\mathcal{T})\sum_{1<j\leq n}\alpha_j\pi(\boldsymbol{\delta}^{(1)})\pi(\boldsymbol{\delta}^{(j)})} + \frac{\sum_{2<j\leq n}\alpha_j\, O\left(N(\mathcal{T})\pi\left(\mathbf{h}^{(2,j)}\right)\right)}{N^2(\mathcal{T})\sum_{2<j\leq n}\alpha_j\pi(\boldsymbol{\delta}^{(2)})\pi(\boldsymbol{\delta}^{(j)})} \\
&\quad + \cdots + \frac{\sum_{n-1<j\leq n}\alpha_j\, O\left(N(\mathcal{T})\pi\left(\mathbf{h}^{(n-1,j)}\right)\right)}{N^2(\mathcal{T})\sum_{n-1<j\leq n}\alpha_j\pi(\boldsymbol{\delta}^{(n-1)})\pi(\boldsymbol{\delta}^{(j)})} + \{t_0+o(1)\}^2 \\
&\quad \longrightarrow t_0^2 \quad \text{as} \quad \boldsymbol{\delta} \longrightarrow 0 \text{ and } N(\mathcal{T})\pi(\boldsymbol{\delta}) \longrightarrow \infty.
\end{aligned}
$$

Also,

$$
\begin{aligned}
\frac{-2t_0E(A_n)}{N(\mathcal{T})\sum_{k=1}^n \alpha_k\pi(\boldsymbol{\delta}^{(k)})} &= \frac{-2t_0}{N(\mathcal{T})\sum_{k=1}^n \alpha_k\pi(\boldsymbol{\delta}^{(k)})}\left\{\sum_{k=1}^n \alpha_k\{t_0+o(1)\}N(\mathcal{T})\pi(\boldsymbol{\delta}^{(k)})\right\} \\
&= -2t_0^2+o(1)\ .
\end{aligned}
$$

Therefore, as $\boldsymbol{\delta} \longrightarrow 0$ and $N(\mathcal{T})\pi(\boldsymbol{\delta}) \longrightarrow \infty$,

$$\epsilon^2 P\left(\left|\frac{A_n}{N(\mathcal{T})\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})} - t_0\right| > \epsilon\right) \longrightarrow t_0^2 - 2t_0^2 + t_0^2 = 0, \tag{5.11}$$

and we conclude that $A_n/N(\mathcal{T})\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})$ converges in probability to t_0.

Employing Chebychev's inequality once again,

$$\begin{aligned}\epsilon^2 P\left(\left|\frac{B_n}{N(\mathcal{T})\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})} - 1\right| > \epsilon\right) &\leq E\left(\frac{B_n^2}{N^2(\mathcal{T})[\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})]^2}\right) \\ &- 2E\left(\frac{B_n}{N(\mathcal{T})\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})}\right) + 1.\end{aligned}$$

From the bounds found in computing $\mathrm{Var}(B_n)$ and $E(B_n)$, and condition (C),

$$E\left(\frac{B_n^2}{N^2(\mathcal{T})[\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})]^2}\right) = \frac{\mathrm{Var}(B_n) + E^2(B_n)}{N^2(\mathcal{T})[\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})]^2}$$

$$\leq \frac{N(\mathcal{T})\sum_{k=1}^n \alpha_k^2 \pi(\boldsymbol{\delta}^{(k)}) + 2N(\mathcal{T})\sum_{1\leq j<l\leq n} \alpha_j \alpha_l \pi\left(\mathbf{h}^{(l,j)}\right)}{N^2(\mathcal{T})[\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})]^2}$$

$$+ \frac{[N(\mathcal{T})\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})]^2}{N^2(\mathcal{T})[\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})]^2}$$

$$\leq \frac{\alpha_1^2 \pi(\boldsymbol{\delta}^{(1)})}{N(\mathcal{T})\alpha_1^2 \pi^2(\boldsymbol{\delta}^{(1)})} + \frac{\alpha_2^2 \pi(\boldsymbol{\delta}^{(2)})}{N(\mathcal{T})\alpha_2^2 \pi^2(\boldsymbol{\delta}^{(2)})} + \cdots + \frac{\alpha_n^2 \pi(\boldsymbol{\delta}^{(n)})}{N(\mathcal{T})\alpha_n^2 \pi^2(\boldsymbol{\delta}^{(n)})}$$

$$+ \frac{\sum_{1<j\leq n} \alpha_j \alpha_1 \pi(\mathbf{h}^{(1,j)})}{N(\mathcal{T})\sum_{1<j\leq n} \alpha_j \alpha_1 \pi(\boldsymbol{\delta}^{(1)})\pi(\boldsymbol{\delta}^{(j)})} + \frac{\sum_{2<j\leq n} \alpha_j \alpha_2 \pi(\mathbf{h}^{(2,j)})}{N(\mathcal{T})\sum_{2<j\leq n} \alpha_j \alpha_2 \pi(\boldsymbol{\delta}^{(2)})\pi(\boldsymbol{\delta}^{(j)})}$$

$$+ \cdots + \frac{\sum_{n-1<j\leq n} \alpha_j \alpha_{n-1} \pi(\mathbf{h}^{(n-1,j)})}{N(\mathcal{T})\sum_{n-1<j\leq n} \alpha_j \alpha_{n-1} \pi(\boldsymbol{\delta}^{(n-1)})\pi(\boldsymbol{\delta}^{(j)})} + 1$$

$$\longrightarrow 1 \quad \text{as} \quad \boldsymbol{\delta} \longrightarrow 0 \text{ and } N(\mathcal{T})\pi(\boldsymbol{\delta}) \longrightarrow \infty.$$

Also,

$$\frac{-2E(B_n)}{N(\mathcal{T})\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})} = \frac{-2N(\mathcal{T})\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})}{N(\mathcal{T})\sum_{k=1}^n \alpha_k \pi(\boldsymbol{\delta}^{(k)})} = -2.$$

Therefore,

$$\epsilon P\left(\left|\frac{B_n}{N(\mathcal{T})\sum_{k=1}^{n}\alpha_k\pi(\boldsymbol{\delta}^{(k)})}-1\right|>\epsilon\right)\longrightarrow 1-2+1=0$$

and so we conclude that $B_n/N(\mathcal{T})\sum_{k=1}^{n}\alpha_k\pi(\boldsymbol{\delta}^{(k)})$ converges to 1 in probability. Finally, using Slutsky's theorem,

$$\frac{A_n}{B_n}\xrightarrow{p} t_0 \quad \text{as } \boldsymbol{\delta}\longrightarrow 0 \text{ and } N(\mathcal{T})\pi(\boldsymbol{\delta})\longrightarrow\infty,$$

which gives the required result.

Chapter 6

Calendaring by Maximum Prevalence

Following the notation established in Chapter 5, we let $\mathcal{T}$, $\mathcal{V}$ and $\mathcal{A}$ denote the training, validation, and test sets, respectively. As before, we will write $i \in \mathcal{T}$ to mean $\mathcal{D}_i \in \mathcal{T}$. Following the definition in Chapter 3, Section 3.3.2, let $s_k(\mathcal{D})$ denote the set of shingles of order k of document $\mathcal{D}$, where the elements of $s_k(\mathcal{D})$ are not necessarily distinct. For a given shingle order k, consider the fixed shingle $s \in s_k(\mathcal{D})$. Let $n_s(\mathcal{D})$ denote the number of times the shingle s occurs in $s_k(\mathcal{D})$, and let $N(\mathcal{D})$ denote the total number of elements in the set $s_k(\mathcal{D})$ (in this notation, the dependence of $N(\mathcal{D})$ on k is suppressed). Note that if the number of words in document $\mathcal{D}$ equals R , then $N(\mathcal{D}) = R - k + 1$.

In this chapter we propose a method to estimate the date of a document $\mathcal{D}$ using a local polynomial GLM model (specifically, binomial) in a logistic regression framework. We let $t_{\mathcal{D}}$ denote the actual date associated with document $\mathcal{D}$. Let n be the number of documents in the training set. For a fixed shingle s, we will think of the sequence of random variable triples $(n_s(\mathcal{D}_i), N(\mathcal{D}_i), t_{\mathcal{D}_i})$, $i = 1, 2, \ldots, n$, as being in-

dependent realizations from a population represented as $(n_s(\mathcal{D}), N(\mathcal{D}), t_{\mathcal{D}})$. We will assume that the elements of $s_k(\mathcal{D})$ occur independently of each other, in the sense that the occurrence of any one particular s in $s_k(\mathcal{D})$ does not affect the probability of occurrence of any other s in $s_k(\mathcal{D})$. (We acknowledge that this independence assumption is somewhat strong. However, studies have shown (see Domingos and Pazzani, 1996) the suprisingly good performance of certain classification problems under the assumption of independence of attributes, when clearly the attributes are highly dependent).

For each fixed s, we require an estimate $\hat{\pi}_s(t)$ of $\pi_s(t)$, namely the frequency, i.e., the probability of occurrence of the shingle s as a function of time t. (Here $\pi_s(t)$ represents the "true" proportion of the occurrence of shingle s among all shingles that have (in principle) occurred at time t.) The estimate $\hat{\pi}_s(t)$ will be based on the training data set $\mathcal{T}$ and it will be computed either as in (4.35) or as in (4.44) (with an associated bandwidth value h) depending on whether the estimate is based on locally constant or locally linear polynomial regression. Strictly speaking, for a fixed shingle s and shingle order k, we should write $\hat{\pi}_s(t) \equiv \hat{\pi}_{s,h,k}(t)$. The time scale t of the DEEDS documents ranges over the middle ages – approximately 11th century to 15th century – and we therefore assume that the probability of such documents occurring outside this time range is zero.

We next define

$$\pi_{\mathcal{D}}(t) = \prod_{s \in s_k(\mathcal{D})} \pi_s(t) \prod_{s \notin s_k(\mathcal{D})} (1 - \pi_s(t)) \ . \tag{6.1}$$

The function $\pi_{\mathcal{D}}(t)$ represents the probability of the occurrence of document $\mathcal{D}$ as function of time t. Strictly speaking, it is the probability conditioned on the length of the document. The formulation (6.1) is based on both those shingles that occurred in the document $\mathcal{D}$ as well as those shingles that did not occur in $\mathcal{D}$. The second

product on the right hand side may be thought of as being over all shingles occurring in the training set (but not in $\mathcal{D}$). Alternately, it may be thought of as being over all shingles that could, in principle, have occurred. We shall see however that this second factor is immaterial. For purposes of clarification, we note that if the shingle s was to occur, for instance, three times in $s_k(\mathcal{D})$, then we use $\pi_s^3(t)$ in (6.1).

We will estimate $\pi_{\mathcal{D}}(t)$ by $\hat{\pi}_{\mathcal{D}}(t)$ given as

$$\hat{\pi}_{\mathcal{D}}(t) = \prod_{s \in s_k(\mathcal{D})} \hat{\pi}_s(t) \prod_{s \notin s_k(\mathcal{D})} (1 - \hat{\pi}_s(t)). \tag{6.2}$$

The date of document $\mathcal{D}$, $t_{\mathcal{D}}$, will then be estimated by $\hat{t}_{\mathcal{D}}$ defined as

$$\hat{t}_{\mathcal{D}} = \arg \max_{t \in [1089,1466]} \hat{\pi}_{\mathcal{D}}(t) \ .$$

Defining $t^*_{\mathcal{D}}$ to be

$$t^*_{\mathcal{D}} = \arg \max_{t \in [1089,1466]} \pi_{\mathcal{D}}(t),$$

in Section 6.4 we will provide sufficient conditions to have $\hat{t}_{\mathcal{D}} \to t^*_{\mathcal{D}}$ as the size of the training data set increases, and the bandwidth h shrinks to zero. We use the notation $t^*_{\mathcal{D}}$, the value of t that maximizes the probability of occurrence of document $\mathcal{D}$ under our theoretical model, so as to differentiate it from the "true" date, $t_{\mathcal{D}}$, of the document $\mathcal{D}$. Such a difference, between these values occurs, for example, because $\mathcal{D}$ is of fixed finite length, and hence does not contain infinite information. We note that $\hat{\pi}_{\mathcal{D}}(t)$, and consequently $\hat{t}_{\mathcal{D}}$, are functions of both the shingle order k, as well as the bandwidth value h.

The bandwidth h will be selected as follows: Using the documents in the validation set $\mathcal{V}$ and the given shingle order k, we assign the bandwidth $h = \hat{h}_{opt}$ to be that value that minimizes the sum of the squared errors between the resulting document date estimates $\hat{t}_{\mathcal{D}} \equiv \hat{t}_{\mathcal{D},h,k}$, and the associated true document dates, $t_{\mathcal{D}}$, over the validation

set :

$$\hat{h}_{opt} = \arg\min_{h} \sum_{\mathcal{D}\in\mathcal{V}} (\hat{t}_{\mathcal{D},h,k} - t_{\mathcal{D}})^2. \tag{6.3}$$

A more adaptive form of bandwidth selection will be described in Section 6.3.

6.1 Estimating $\hat{\pi}_s(t)$

We propose to estimate $\pi_s(t)$ based on a local parametric logistic regression model as discussed in Chapter 4.

Specifically, for a given document $\mathcal{D}$, we assume the conditional distribution of $n_s(\mathcal{D})$ given $N(\mathcal{D}) = m$ and $t_{\mathcal{D}} = t$ to be binomially distributed:

$$\mathcal{L}(n_s(\mathcal{D}) \mid N(\mathcal{D}) = m, t_{\mathcal{D}} = t) \sim \text{Bin}(m, \pi_s(t)).$$

Following the results of Chapter 4, Section 4.2.1, denote the canonical transformation of $\pi_s(t)$ as $\theta_s(t) = \log\{\pi_s(t)/(1-\pi_s(t))\}$. For the locally constant (i.e. polynomial degree $p = 0$), smoothing the parameter $\theta_s(t)$ at time t using kernel K and bandwidth h, the local log-likelihood is given by equation (4.33) as

$$l(\beta_0) = \sum_{i\in\mathcal{T}} \left\{ \beta_0 n_s(\mathcal{D}_i) - N(\mathcal{D}_i) b(\beta_0) + \log \begin{pmatrix} N(\mathcal{D}_i) \\ n_s(\mathcal{D}_i) \end{pmatrix} \right\} K_h(t_{\mathcal{D}_i} - t) \tag{6.4}$$

where $b(\beta_0) = \log(1 + \exp(\beta_0))$. Note that β_0 is the local constant value of $\theta_s(t)$. Letting $\hat{\beta}_0 = \arg\max_{\beta_0} \; l(\beta_0)$, it follows that

$$\hat{\beta}_0 = \log\left(\frac{\hat{\pi}_s(t)}{1 - \hat{\pi}_s(t)}\right)$$

where

$$\hat{\pi}_s(t) = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)} = \frac{\sum_{i\in\mathcal{T}} n_s(\mathcal{D}_i) K_h(t_{\mathcal{D}_i} - t)}{\sum_{i\in\mathcal{T}} N(\mathcal{D}_i) K_h(t_{\mathcal{D}_i} - t)} \,. \tag{6.5}$$

(See equation (4.35).)

If on the other hand, we use the locally linear smoothing of the parameter $\theta_s(t)$, (i.e. smoothing the parameter $\theta_s(t)$ locally by a polynomial of degree $p = 1$), and then optimize the resulting log-likelihood with respect to β_0 and β_1, given by equations (4.42) and (4.43), we are then required to solve the pair of equations

$$\sum_{i\in\mathcal{T}} n_s(\mathcal{D}_i)K_h(t_{\mathcal{D}_i} - t) = \sum_{i\in\mathcal{T}} \frac{N(\mathcal{D}_i)\exp\{\beta_0 + \beta_1(t_{\mathcal{D}_i} - t)\}}{1 + \exp\{\beta_0 + \beta_1(t_{\mathcal{D}_i} - t)\}} K_h(t_{\mathcal{D}_i} - t)$$

and

$$\sum_{i\in\mathcal{T}} n_s(\mathcal{D}_i)(t_{\mathcal{D}_i} - t)K_h(t_{\mathcal{D}_i} - t) = \sum_{i\in\mathcal{T}} \frac{N(\mathcal{D}_i)\exp\{\beta_0 + \beta_1(t_{\mathcal{D}_i} - t)\}}{1 + \exp\{\beta_0 + \beta_1(t_{\mathcal{D}_i} - t)\}} (t_{\mathcal{D}_i} - t)K_h(t_{\mathcal{D}_i} - t)$$

for every value of t. As discussed in Chapter 4, Section 4.2.1, the solution $\hat{\beta}_o$ here is then used to estimate $\hat{\pi}_s(t)$ via the inverse of the link function:

$$\hat{\pi}_s(t) = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)} \quad .$$

Even though the estimate $\hat{\pi}_s(t)$ derived using the locally linear polynomial model has better boundary estimation properties than that of the estimator derived using the locally constant polynomial (see Chapter 4, Section 4.2.2 for discussions on this issue), we only used the local constant model in this thesis so as to avoid the numerical optimization which is necessary for the locally linear modelling (see Chapter 4, Section 4.2.1). Therefore, from this point on, $\hat{\pi}_s(t)$ will refer to the estimated probability of occurrence of the shingle s at time t based on locally constant polynomial logistic regression.

Before concluding this section, we make the following observation. Suppose we model the the number of occurrences of the shingle s at time t as a Poisson distribution with $\lambda(t)$ being the rate of occurrence of the shingle s in a reference text of length, say, r. (This is analogous to Mosteller's and Wallace's analysis of the Federalist papers.

The rate μ_H and μ_M in their case referred to the rate of a particular word usage in a text of one thousand words for Hamilton and Madison, respectively. Here, we could choose r to be 250, the average length of documents in the DEEDS data set). Then, the locally smoothed log-likelihood (4.32) is given by

$$\sum_{i\in\mathcal{T}}\{n_s(\mathcal{D}_i)\log(\lambda(t_{\mathcal{D}_i})N(\mathcal{D}_i)/r)-\lambda(t_{\mathcal{D}_i})N(\mathcal{D}_i)/r-\log(n_s(\mathcal{D}_i)!)\}K_h(t_{\mathcal{D}_i}-t)$$
$$=\sum_{i\in\mathcal{T}}\{n_s(\mathcal{D}_i)(\theta(t_{\mathcal{D}_i})+\log(N(\mathcal{D}_i)/r))-\exp(\theta(t_{\mathcal{D}_i}))N(\mathcal{D}_i)/r-\log(n_s(\mathcal{D}_i)!)\}K_h(t_{\mathcal{D}_i}-t).$$

We denote the canonical parameter by $\theta(u)=\log(\lambda(u))$ and $b_{pois}(\theta)=\exp(\theta)$. If we choose the canonical parameter to be a polynomial function of order p, for a given t,

$$\theta(t_{\mathcal{D}_i})\approx\bar{\beta}(t,t_{\mathcal{D}_i})\equiv\beta_0(t)+\beta_1(t)(t_{\mathcal{D}_i}-t)+\cdots+\beta_p(t)(t_{\mathcal{D}_i}-t)^p. \tag{6.6}$$

Then, from (4.32),

$$\begin{aligned}L_{pois}(\beta(t)) &= \sum_{i\in\mathcal{T}}\{\, n_s(\mathcal{D}_i)(\beta_0(t)+\beta_1(t)(t_{\mathcal{D}_i}-t)+\cdots+\beta_p(t)(t_{\mathcal{D}_i}-t)^p)\\ &\quad -\frac{N(\mathcal{D}_i)}{r}b_{pois}(\beta_0(t)+\beta_1(t)(t_{\mathcal{D}_i}-t)+\cdots+\beta_p(t)(t_{\mathcal{D}_i}-t)^p)\\ &\quad +n_s(\mathcal{D}_i)\log(N(\mathcal{D}_i)/r)-\log(n_s(\mathcal{D}_i)!)\,\}\,K_h(t_{\mathcal{D}_i}-t)\,.\end{aligned}$$

On the other hand, the locally smoothed log-likelihood for the binomial, $L_{bin}(\beta(t))$, as given in (4.33), is

$$\begin{aligned}L_{bin}(\beta(t)) &= \sum_{i\in\mathcal{T}}\{\, n_s(\mathcal{D}_i)(\beta_0(t)+\beta_1(t)(t_{\mathcal{D}_i}-t)+\cdots+\beta_p(t)(t_{\mathcal{D}_i}-t)^p)\\ &\quad -N(\mathcal{D}_i)b_{bin}(\beta_0(t)+\beta_1(t)(t_{\mathcal{D}_i}-t)+\cdots+\beta_p(t)(t_{\mathcal{D}_i}-t)^p)\\ &\quad +\log\binom{N(\mathcal{D}_i)}{n_s(\mathcal{D}_i)}\Big\}\,K_h(t_{\mathcal{D}_i}-t)\end{aligned}$$

where $b_{bin}(\theta)=\log\{1+\exp(\theta)\}$. Denote by $\hat{\boldsymbol{\beta}}(t)=(\hat{\beta}_0(t),\hat{\beta}_1(t),\ldots,\hat{\beta}_p(t))$ the values of the β's that maximize $L_{pois}(\beta(t))$, and denote by $\tilde{\boldsymbol{\beta}}(t)=(\tilde{\beta}_0(t),\tilde{\beta}_1(t),\ldots,\tilde{\beta}_p(t))$ the values of the β's that maximize $L_{bin}(\beta(t))$. Then, for every $j=0,1,\ldots,p$,

$$\sum_{i\in\mathcal{T}}n_s(\mathcal{D}_i)K_h(t_{\mathcal{D}_i}-t)\frac{\partial\bar{\beta}(t,t_{\mathcal{D}_i})}{\partial\beta_j}\bigg|_{\hat{\boldsymbol{\beta}}(t)}$$

$$= \sum_{i\in\mathcal{T}} \frac{N(\mathcal{D}_i)}{r} b'_{pois}(\hat{\beta}_0(t) + \hat{\beta}_1(t)(t_{\mathcal{D}_i} - t) + \cdots + \hat{\beta}_p(t)(t_{\mathcal{D}_i} - t)^p) \frac{\partial\bar{\beta}(t, t_{\mathcal{D}_i})}{\partial\beta_j}\Big|_{\hat{\boldsymbol{\beta}}(t)} K_h(t_{\mathcal{D}_i} - t)$$

and

$$\sum_{i\in\mathcal{T}} n_s(\mathcal{D}_i) K_h(t_{\mathcal{D}_i} - t) \frac{\partial\bar{\beta}(t, t_{\mathcal{D}_i})}{\partial\beta_j}\Big|_{\tilde{\boldsymbol{\beta}}(t)}$$
$$= \sum_{i\in\mathcal{T}} N(\mathcal{D}_i) b'_{bin}(\tilde{\beta}_0(t) + \tilde{\beta}_1(t)(t_{\mathcal{D}_i} - t) + \cdots + \tilde{\beta}_p(t)(t_{\mathcal{D}_i} - t)^p) \frac{\partial\bar{\beta}(t, t_{\mathcal{D}_i})}{\partial\beta_j}\Big|_{\tilde{\boldsymbol{\beta}}(t)} K_h(t_{\mathcal{D}_i} - t).$$

From the above it follows,

$$\sum_{i\in\mathcal{T}} \left(\frac{\partial\bar{\beta}(t, t_{\mathcal{D}_i})}{\partial\beta_j}\Big|_{\hat{\boldsymbol{\beta}}(t)} - \frac{\partial\bar{\beta}(t, t_{\mathcal{D}_i})}{\partial\beta_j}\Big|_{\tilde{\boldsymbol{\beta}}(t)} \right) n_s(\mathcal{D}_i) K_h(t_{\mathcal{D}_i} - t)$$
$$= \sum_{i\in\mathcal{T}} N(\mathcal{D}_i) \left\{ \frac{1}{r} b'_{pois}(\hat{\beta}_0(t) + \hat{\beta}_1(t)(t_{\mathcal{D}_i} - t) + \cdots + \hat{\beta}_p(t)(t_{\mathcal{D}_i} - t)^p) \right.$$
$$\left. - b'_{bin}(\tilde{\beta}_0(t) + \tilde{\beta}_1(t)(t_{\mathcal{D}_i} - t) + \cdots + \tilde{\beta}_p(t)(t_{\mathcal{D}_i} - t)^p) \right\} (t_{\mathcal{D}_i} - t)^j K_h(t_{\mathcal{D}_i} - t) \ .$$

Since the left hand-side of the above equation is equal to zero,

$$\sum_{i\in\mathcal{T}} \{N(\mathcal{D}_i)/r\} \left\{ b'_{pois}(\hat{\beta}_0(t) + \hat{\beta}_1(t)(t_{\mathcal{D}_i} - t) + \cdots + \hat{\beta}_p(t)(t_{\mathcal{D}_i} - t)^p) \right\} (t_{\mathcal{D}_i} - t)^j K_h(t_{\mathcal{D}_i} - t)$$
$$= \sum_{i\in\mathcal{T}} N(\mathcal{D}_i) \left\{ b'_{bin}(\tilde{\beta}_0(t) + \tilde{\beta}_1(t)(t_{\mathcal{D}_i} - t) + \cdots + \tilde{\beta}_p(t)(t_{\mathcal{D}_i} - t)^p) \right\} (t_{\mathcal{D}_i} - t)^j K_h(t_{\mathcal{D}_i} - t)$$

for every $j = 0, 1, \ldots, p$. This implies that if the locally smoothed log-likelihood functions are locally constant (i.e., the polynomial in (6.6) is of order $p = 0$), then $b'_{pois}(\hat{\beta}_0(t))/r = b'_{bin}(\tilde{\beta}_0(t))$, that is, in the case $p = 0$, the estimated probability of occurrence of the shingle s at time t, derived from the binomial model, is equal to the estimated rate of occurrence of the shingle s at time t in a text of length r derived from using the Poisson.

6.2 A note on the second factor in $\pi_{\mathcal{D}}(t)$

In this section we will argue that the second factor of

$$\pi_{\mathcal{D}}(t) = \prod_{s\in s_k(\mathcal{D})} \pi_s(t) \prod_{s\notin s_k(\mathcal{D})} (1 - \pi_s(t)) \tag{6.7}$$

may be omitted.

Consider the logarithm of the second factor in (6.10) :

$$\begin{aligned} \log \prod_{s \notin s_k(\mathcal{D})} (1 - \pi_s(t)) &= \sum_{s \notin s_k(\mathcal{D})} \log(1 - \pi_s(t)) \\ &\approx -\sum_{s \notin s_k(\mathcal{D})} \pi_s(t) \qquad (6.8) \\ &\approx -\sum_{s} \pi_s(t) = -1. \qquad (6.9) \end{aligned}$$

In (6.8) we have used the approximation $\log(1+x) \approx x$ for small x. The approximation at (6.9) follows because the number of shingles of any document $\mathcal{D}$ is small compared with the total that can occur. A typical document, for example, has on average about 200 words while the total number of distinct words in the documents of the training set is 42,978. Consequently, only the first factor of $\pi_{\mathcal{D}}(t)$ will typically be material and we shall henceforth take

$$\pi_{\mathcal{D}}(t) = \prod_{s \in s_k(\mathcal{D})} \pi_s(t) \; . \qquad (6.10)$$

Figure 6.1 is a plot of time t versus $\sum_s \log(1 - \hat{\pi}_s(t))$ where the sum over the shingles is over all the distinct words (shingle order 1) of the documents in the training set. As can be seen, the value of $\sum_s \log(1 - \hat{\pi}_s(t))$ is very close to -1 for all values of t.

6.3 Bandwidth selection

At this point, we comment on the bandwidth choice, $\hat{h}_{opt}$, defined in (6.3). Clearly, $\hat{h}_{opt}$ as a bandwidth choice is not fully optimal since we use this same bandwidth for every estimate $\hat{\pi}_s(t)$. Ideally, there should be a different "optimal" bandwidth for each shingle s. However, the additional work involved would be substantial and it seems questionable based on our experience with this data that this would result in

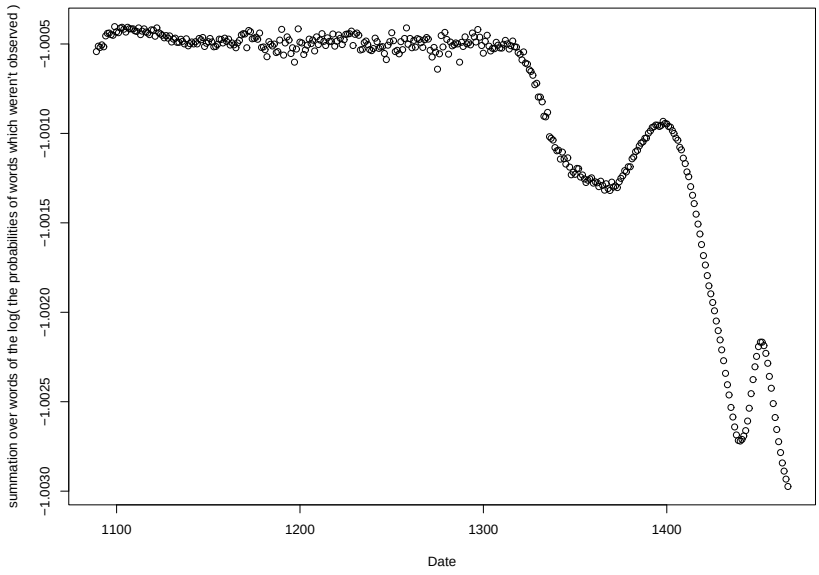


Figure 6.1: Date t versus $\sum_s \log\{(1-\hat{\pi}_s(t))\}$ where the sum is taken over all distinct words from documents in the training set. There are a total of 42,978 such words.

any great improvement of our date estimates. In the remainder of this section, we discuss how bandwidths could be chosen for each shingle.

The cross-validation technique discussed in Chapter 4, Section 4.1.7 suggests one way to select a bandwidth for $\hat{\pi}_s(t)$ specific to the shingle s. Defining $\hat{\pi}_{-i,s}(t)$ to be the same as $\hat{\pi}_s(t)$ but with document $\mathcal{D}_i$ of $\mathcal{T}$ removed from the computation, the cross-validation function for the fixed shingle s is given by

$$\mathrm{CV}_s(h) = |\mathcal{T}|^{-1} \sum_{i \in \mathcal{T}} \left(\hat{\pi}_{-i,s}(t_{\mathcal{D}_i}) - \frac{n_s(\mathcal{D}_i)}{N(\mathcal{D}_i)} \right)^2 .$$

The associated optimal bandwidth, $h_{s,cv}$, is then defined to be

$$h_{s,cv} = \arg\min_h \mathrm{CV}_s(h),$$

and this bandwidth can be used in the estimate of $\hat{\pi}_s(t)$.

For purposes of computational simplification, we can write

$$\begin{aligned} \mathrm{CV}_s(h) &= |\mathcal{T}|^{-1} \sum_{i \in \mathcal{T}} \left(\hat{\pi}_{-i,s}(t_{\mathcal{D}_i}) - \frac{n_s(\mathcal{D}_i)}{N(\mathcal{D}_i)} \right)^2 \\ &= |\mathcal{T}|^{-1} \sum_{i \in \mathcal{T}} \left(\frac{\sum_{j \in \mathcal{T}} n_s(\mathcal{D}_j) K_h(t_{\mathcal{D}_j} - t_{\mathcal{D}_i}) - n_s(\mathcal{D}_i) K(0)}{\sum_{j \in \mathcal{T}} N(\mathcal{D}_j) K_h(t_{\mathcal{D}_j} - t_{\mathcal{D}_i}) - N(\mathcal{D}_i) K(0)} - \frac{n_s(\mathcal{D}_i)}{N(\mathcal{D}_i)} \right)^2 . \end{aligned}$$

We note that the advantage of selecting a bandwidth using the above procedure is that it allows us to obtain an optimal bandwidth (in the sense finding the minimizer of $\mathrm{CV}_s(h)$) for each shingle s. Furthermore, the leave-one-out procedure would in principle allow us to work without a validation set.

6.4 Asymptotic properties of $\hat{\pi}_s(t)$ and $\hat{t}_{\mathcal{D}}$

We now discuss some asymptotic aspects concerning the estimators $\hat{\pi}_s(t)$, and the date estimator $\hat{t}_{\mathcal{D}}$. To state these asymptotic results, we require some additional notation. Assume a countably infinite set of documents $\mathcal{D}_1, \mathcal{D}_2, \ldots$ and consider the increasing sequence of training sets $\mathcal{T}_n = \{\mathcal{D}_1, \cdots, \mathcal{D}_n\}$. For each n, write h_n for the bandwidth associated with the sample size n. The conditions we impose on the kernel $K(\cdot)$ and bandwidths h are:

$$\begin{aligned} &\text{(a)} \qquad \sup_{-\infty < t < \infty} K(t) < \infty, \\ &\text{(b)} \qquad \lim_{|t| \to \infty} |t| K(t) = 0 \ . \end{aligned}$$

We also assume that $n \to \infty$, $h_n \to 0$ and $nh_n \to \infty$. Let $f(t)$ denote the density of $t_{\mathcal{D}}$, and let $Y_s(t) = E\left(n_s(\mathcal{D}) | t_{\mathcal{D}} = t\right)$ and $Y(t) = E\left(N(\mathcal{D}) | t_{\mathcal{D}} = t\right)$.

Proposition 1 *(Nadaraya 1964) Assume the characteristic function $\chi(t) = \int_{-\infty}^{\infty} e^{itx} K(x) dx$ is absolutely integrable, and let $Y_s(t)$, $Y(t)$ and $f(t)$ be continuous on a finite interval $[t_A, t_B]$. Further, assume that*

$$\min_{t \in [t_A, t_B]} f(t) > 0 \text{ and } \sum_{n=1}^{\infty} n^{-2} h_n^{-4} < \infty \ ,$$

and that the fourth moments of $n_s(\mathcal{D})$ *and* $N(\mathcal{D})$ *exist.*[1] *Then, with probability one*

$$\sup_{t\in[t_A,t_B]}\left|\frac{\sum_{i=1}^n n_s(\mathcal{D}_i)K_h(t_{\mathcal{D}_i}-t)}{\sum_{i=1}^n K_h(t_{\mathcal{D}_i}-t)}-Y_s(t)\right|\longrightarrow 0 \tag{6.11}$$

and

$$\sup_{t\in[t_A,t_B]}\left|\frac{\sum_{i=1}^n N(\mathcal{D}_i)K_h(t_{\mathcal{D}_i}-t)}{\sum_{i=1}^n K_h(t_{\mathcal{D}_i}-t)}-Y(t)\right|\longrightarrow 0 \tag{6.12}$$

as $h_n\to 0$ *and* $nh_n\to\infty$.

Note that although our data is discrete (i.e. the dates of the documents are stated in years) we are nonetheless treating the dates as being continuous. Furthermore, we base our support for the assumption that the density of the dates $f(t)$ is strictly greater than zero over the middle ages ($t_A = 1089$ to $t_B = 1466$) on empirical observation (see Figure 1.1).

Equations (6.11) and (6.12) imply that the denominator of the expression in (6.12) is non-zero and therefore we have, with probability one,

$$\sup_{t\in[t_A,t_B]}\left|\frac{\sum_{i=1}^n n_s(\mathcal{D}_i)K_h(t_{\mathcal{D}_i}-t)}{\sum_{i=1}^n N(\mathcal{D}_i)K_h(t_{\mathcal{D}_i}-t)}-\frac{Y_s(t)}{Y(t)}\right|\longrightarrow 0\ .$$

However

$$\begin{aligned} Y_s(t)=\mathrm{E}(n_s(\mathcal{D})|t_{\mathcal{D}}=t) &= \mathrm{E}(\mathrm{E}\{n_s(\mathcal{D})\mid t_{\mathcal{D}}=t,N(\mathcal{D})\}\mid t_{\mathcal{D}}=t)\\ &= \mathrm{E}(N(\mathcal{D})\pi_s(t)\mid t_{\mathcal{D}}=t)\\ &= \pi_s(t)\mathrm{E}(N(\mathcal{D})\mid t_{\mathcal{D}}=t)\\ &= \pi_s(t)Y(t). \end{aligned} \tag{6.13}$$

Therefore, with probability one

$$\sup_{t\in[t_A,t_B]}\left(\frac{\sum_{i=1}^n n_s(\mathcal{D}_i)K_h(t_{\mathcal{D}_i}-t)}{\sum_{i=1}^n N(\mathcal{D}_i)K_h(t_{\mathcal{D}_i}-t)}-\pi_s(t)\right)\longrightarrow 0\ .$$

[1]Note: since $n_s(\mathcal{D})\le N(\mathcal{D})$, we only need to require the fourth moment of $N(\mathcal{D})$ to be finite.

We thus see that under appropriate assumption, our estimators $\hat{\pi}_s(t)$ may be regarded as being consistent for $\pi_s(t)$. We next consider the estimator $\hat{t}_{\mathcal{D}}$. The method we have employed in estimating the date of a given document $\mathcal{D}$ involves estimating the probabilities over time of each of the shingles that make up the document, and then determining the point on the time axis at which the product of the estimated shingle probabilities, as in (6.2) achieves their maximum value. This value, which we denote by $\hat{t}_{\mathcal{D}}$, is our estimated date for the document $\mathcal{D}$. We will attempt to provide some informal justification for our use of $\hat{t}_{\mathcal{D}}$ as our date estimator. Firstly, however, note that one cannot really know the "true date" of a given undated document. Furthermore, the dating methodology we have presented is an approximation in the sense that it neglects idiosyncratic elements such as language usage between different type of individuals, or other such confounding elements. Moreover, we assume that the occurrences of shingles within a given document are independent of each other. In fact, all types of documents (so long as they deal with property transfer in the middle ages) are incorporated into the training set.

The question we would now like to pose for our document dating procedure is: what is the asymptotic behavior of $\hat{t}_{\mathcal{D}}$ as the size of the training set $|\mathcal{T}|$ increases to infinity, and the bandwidth h decreases to 0 in such a way that $|\mathcal{T}|h$ increases to infinity. We will attempt to justify that (under our independence assumption) as the size of the training data set $\mathcal{T}$ increases, the estimated date of the document $\hat{t}_{\mathcal{D}}$ converges in distribution to $t^*_{\mathcal{D}}$. This will hold under the following conditions (see Kim & Pollard, 1990):

(a) $\hat{\pi}_{\mathcal{D}}(t)$ converges in probability to $\pi_{\mathcal{D}}(t)$ uniformly for $t \in [1089, 1466]$.

(b) $\hat{t}_{\mathcal{D}} = \arg\max_t \hat{\pi}_{\mathcal{D}}(t)$ is $O_p(1)$.

(c) $t^*_{\mathcal{D}}$ is the unique maximizer of $\pi_{\mathcal{D}}(t)$ on $[1089, 1466]$.

We also have shown that $\hat{\pi}_{\mathcal{D}}(t)$ converges uniformly to $\pi_{\mathcal{D}}(t)$ on $[1089, 1466]$, and the continuity of $\hat{\pi}_{\mathcal{D}}(t)$ follows from the continuity of the kernel function $K(\cdot)$. Condition (b) follows from our assumption that $\hat{\pi}_{\mathcal{D}}(t)$ takes on the value zero outside of the date range $[1089, 1466]$, and takes on the values $0 \leq \hat{\pi}_{\mathcal{D}}(t) \leq 1$ on the date range $[1089, 1466]$. As for condition (c), our methodology implicitly assumes that a unique maximizer exists as the size of the training set increases and this may be justified on pragmatic grounds. Clearly, shingles which are "uninformative", i.e. those used with equal frequency across the years, or those whose frequency of use increases and decreases multiple times, or those that are only in use at an earlier and a later time, could, in rather untypical situations, cause condition (c) to be violated.

6.5 Numerical results

The DEEDS data set, consisting of a total of 3353 dated documents, was divided into three sets - the training set $\mathcal{T}$, the validation set $\mathcal{V}$, and the test set $\mathcal{A}$. Following the procedure described in Chapter 5, Section 3, we assigned 419 documents to the validation set (12.5% of the total data) and 326 documents to the test set (9.7% of the total data). The remaining 2608 documents (77.8% of the total data) form the training set.

Dates for every document $\mathcal{D}$ in the validation set $\mathcal{V}$ were estimated on the basis of the documents in the training set, using the t-distribution density function as the kernel function:

$$K(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

where ν is the number of degrees of freedom, and Γ is the Gamma function. Letting $K_h(\cdot) = (1/h)K(\cdot/h)$, we note in passing that the constant multipliers of $K_h(x)$ may be ignored in the actual numerical computation of $\pi_s(t)$ in (6.5). The bandwidth

values h used in our numerical experiments ranged from $h = 1$ to $h = 24$, and the degrees of freedom (d.f.) we used for the t-distribution ranged from 1 to 20. The combination of the bandwidth and the degrees of freedom which minimized the value of the mean squared error of the dates (MSE) on the validation set were designated to be the "optimal" parameters, and were used as the parameters of choice in computing the estimated dates for the documents in the test set $\mathcal{A}$. The above procedure was carried out for shingle sizes 1, 2, 3 and 4, resulting in four dating methodologies which we will denote by M_1, M_2, M_3 and M_4 respectively. Table 6.1 and Table 6.2 summarize these optimal bandwidths and optimal degrees of freedom, as well as the corresponding mean squared error (MSE), mean absolute error (MAE), and median absolute error (MedAE) in years, and the concordance correlation coefficient, ρ_c, (see equation (5.6) for definition) between the true date and the estimated date for documents in the validation set and the test set, respectively. For comparison purposes we have included the results for both the validation and test sets.

As we can see from the results of the dating methodologies based on documents from the test set, Table 6.2, method M_2 performed the best - it had the smallest MSE, MAE, MedAE and the highest value of ρ_c. Method M_3 seems to be the next best, followed by M_1, and finally M_4. In all cases, the MAE is consistently greater than the MedAE, indicating that there are several documents which have a relatively high error value. The histogram in Figure 6.2 is that of the value of the error of date estimates based on method M_2 for documents in the test set. The MAE and the MedAE are 9 years and 6 years, respectively.

We also attempted to combine the dating methodologies $M_1,\ M_2,\ M_3$ and M_4 in order to construct a single dating methodology, M_{total}, which would perform at least as well as any of the previous dating methodologies. We constructed M_{total} by using a weighted sum of the dating estimators based on $M_1,\ M_2,\ M_3$ and M_4. The weights

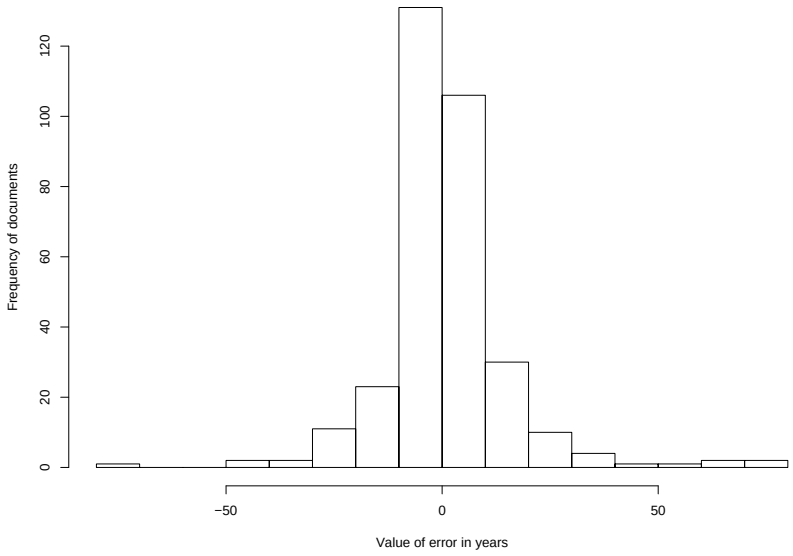


Figure 6.2: A histogram of the error in years of method M_2 based on documents in the test set. The bandwidth value is $h = 12$ and degree of freedom value is $d.f. = 3$. The 1st quartile is at 2 years, the median is at 6 years, the 3rd quartile is at 11 years and the maximum value of the absolute error is at 79 years.

were obtained by minimizing the mean square error of the estimated document dates over the validation set as follows: let $\hat{t}_{ij}$ be the date estimate of document $j \in \mathcal{V}$ based on dating methodology M_i $(i = 1, \ldots, 4)$ and let

$$(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4) = \min_{a_1,a_2,a_3,a_4} \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} (a_1\hat{t}_{1j} + a_2\hat{t}_{2j} + a_3\hat{t}_{3j} + a_4\hat{t}_{4j} - t_j)^2$$

where t_j is the true date of document j, and $\sum_i a_i = 1$. Computation results give

$$(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4) = (0.14471,\ 0.63599,\ 0.12019,\ 0.09844).$$

Using this estimation of $(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4)$ as the coefficients for combining date estimates from the test set based on dating methods $M_1,\ M_2,\ M_3$ and M_4 (we denote this method M_{total}), we find the $\sqrt{\text{MSE}}$ to be 14.5 years, MAE to be 9.2 years, MedAE

to be 6 years and $p_c = 0.95$. (Similar calculations on the validation set based on coefficients $(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4)$ results in $\sqrt{\text{MSE}}$ to be 14.3 years, MAE to be 9.3 years, MedAE to be 6 years and $p_c = 0.94$). As we can see, optimizing on MSE, M_{total} does not perform much better than using method M_2 alone, and the values of $\sqrt{\text{MSE}}$, MAE, MedAE, and p_c between the two methods are comparable.

Figures 6.3 – 6.7 provide plots of estimated document dates versus the actual document dates based on dating methodologies $M_1,\ M_2,\ M_3,\ M_4$ and M_{total}, respectively, for the 326 documents in the test set $\mathcal{A}$. In all the dating methodologies, the plots exhibit edge bias, and in particular the edge bias from M_{total} is not any smaller than that of M_2. For this reason and from the discussions in the previous paragraph, we conclude that basing our dating analysis on a single shingle order (shingle order 2 in this case) suffices.

Table 6.1: Values of $\sqrt{MSE}$, MAE, MedAE, and ρ_c for dating methodologies M1, M2, M3 and M4 evaluated on a validation set of order 419.

Dating Method	*Shingle Order*	*Optimal h*	*Optimal d.f.*	$\sqrt{MSE}$ *(val. set)*	*MAE (val. set)*	*MedAE (val. set)*	ρ_c
M1	1	8	5	18.3	11.7	7.0	0.91
M2	2	12	3	14.8	9.5	6.0	0.94
M3	3	12	5	17.0	10.1	6.0	0.92
M4	4	16	12	18.8	11.5	7.0	0.90

Table 6.2: Values of $\sqrt{MSE}$, MAE, MedAE, and ρ_c for dating methodologies M1, M2, M3 and M4 evaluated on a test set of order 326.

Dating Method	*Shingle Order*	*Optimal h*	*Optimal d.f.*	$\sqrt{MSE}$ *(test set)*	*MAE (test set)*	*MedAE (test set)*	ρ_c
M1	1	8	5	19.8	12.5	8.0	0.91
M2	2	12	3	14.7	9.0	6.0	0.94
M3	3	12	5	15.4	9.5	6.0	0.94
M4	4	16	12	22.8	12.4	7.0	0.88

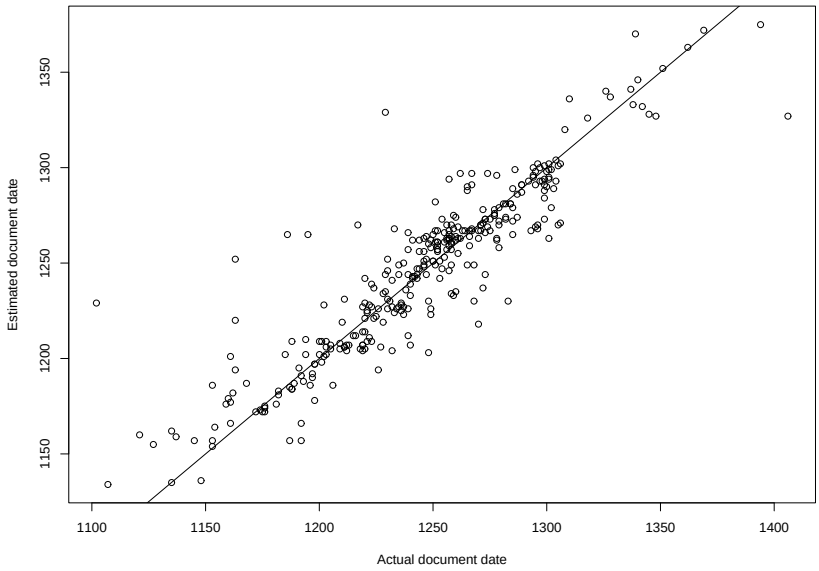


Figure 6.3: Estimated versus actual document date for the 326 documents in the test set $\mathcal{A}$ based on shingle order 1. The solid line is "X = Y" axis.

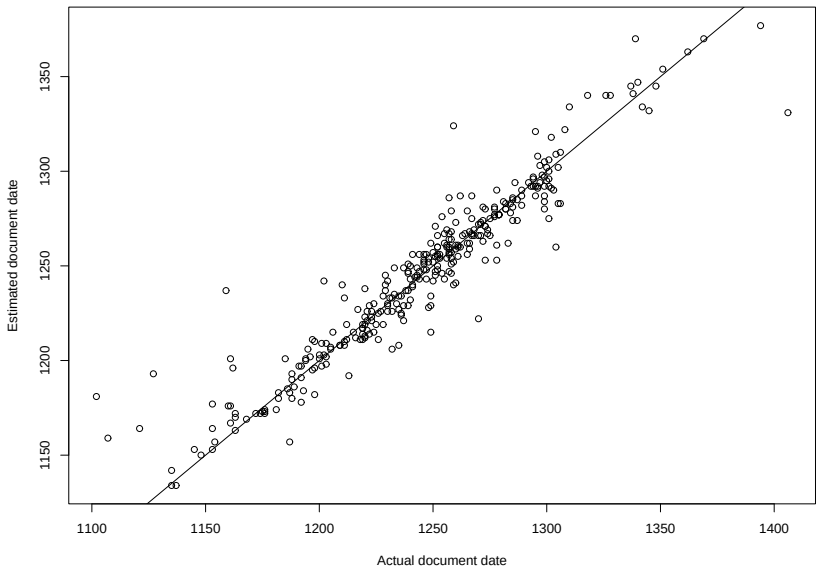


Figure 6.4: Estimated versus actual document date for the 326 documents in the test set $\mathcal{A}$ based on shingle order 2. The solid line is "X = Y" axis.

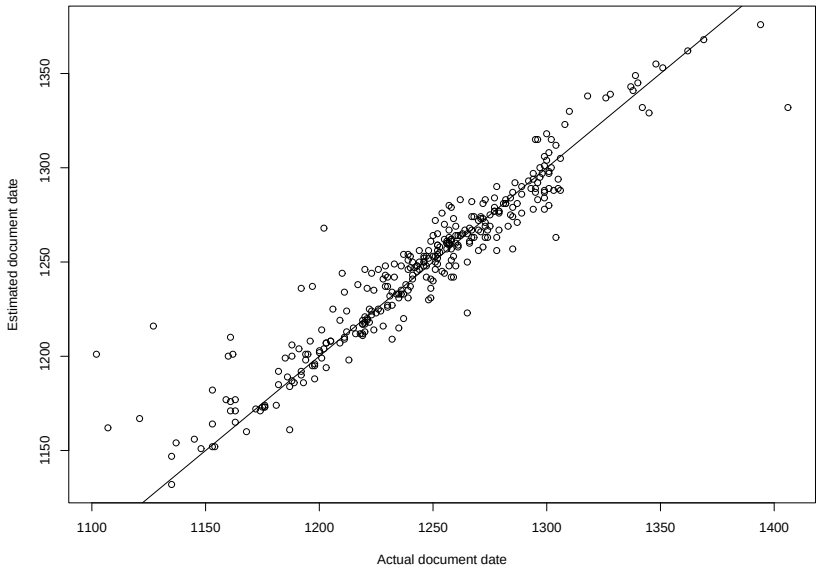


Figure 6.5: Estimated versus actual document date for the 326 documents in the test set $\mathcal{A}$ based on shingle order 3. The solid line is "X = Y" axis.

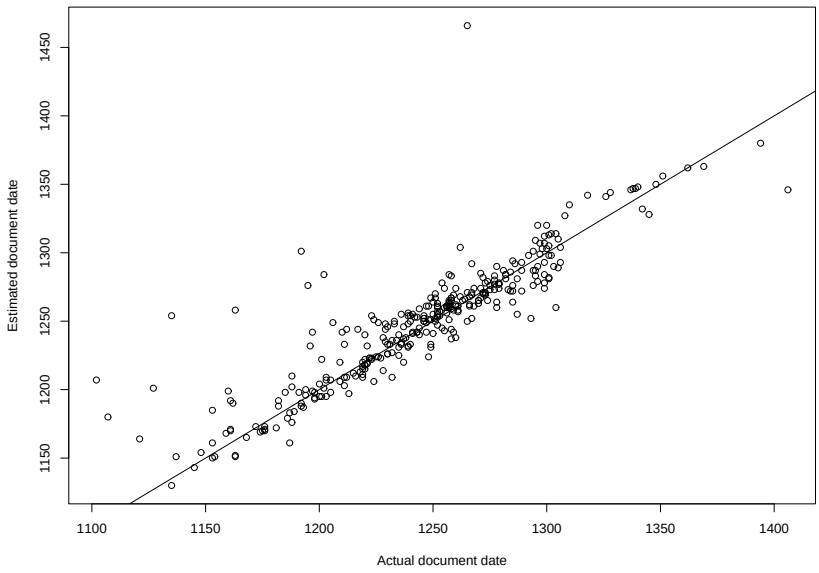


Figure 6.6: Estimated versus actual document date for the 326 documents in the test set $\mathcal{A}$ based on shingle order 4. The solid line is "X = Y" axis.

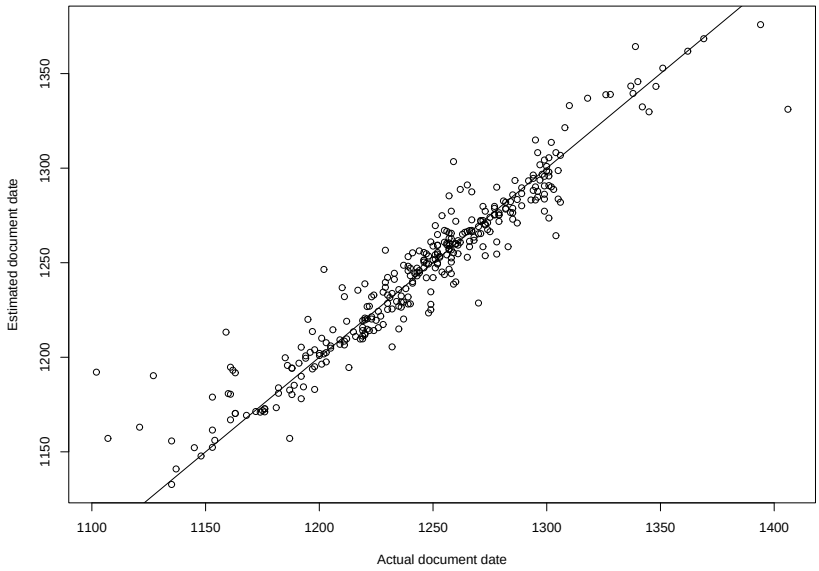


Figure 6.7: Estimated versus actual document date for the 326 documents in the test set $\mathcal{A}$ based on the combination of shingle orders 1, 2, 3 and 4 (the M_{total} method). The solid line is the "X = Y" axis.

Figures 6.8a,b – 6.11a,b show the estimate of $\text{logit}(\pi_s(t))$ and the estimate of $\pi_s(t)$, the probability of occurrence of various selected shingles across time. The data in part (a) of the figures represents the logit of the proportion of time the shingles occurred at different dates, and the data in part (b) of the figures is simply the proportions of time the shingles occurred at different dates. All the estimates are based on the local constant fit where the kernel weights are the density of the t-distribution. The bandwidth and degree of freedom associated with the kernel weight used in each of the plots are the optimal bandwidth and the optimal degree of freedom associated with the shingle order of s. For example, the optimal bandwidth and the optimal degree of freedom used for shingle order 1 are $h = 8$ and $d.f. = 5$, and for shingle order 3, $h = 12$ and $d.f. = 5$. The estimates in part (b) of the figures are obtained

by applying the respective inverse-logit transformation to figures in part (a). The data points drawn along the solid horizontal line in part (a) of the plots indicate that there is a zero occurrence of the shingle corresponding to dates on the x-axis, even though there are documents at these dates. (The value of the y-axis corresponding to the horizontal line is "artificial", in the sense that it is arbitrary and chosen only to graphically illustrate the point made in the previous sentence. The logit transform of these same date points is plotted in figures part (b), below the value 0 corresponding to the y-axis.)

Shingles were selected for plotting in Figures 6.8a,b – 6.11a,b based on the number of times they occurred across different dates. If they occurred consistently at about the same proportion across the different dates, they are regarded as being uninformative, that is, they don't provide any indication as to the date they were more likely to have been used. On the other hand, informative shingles are those that occur with highly variable proportions across different dates.

The shingles "*et*" ("and") and "*omnibus*" ("for all") are examples of uninformative shingles. As can be seen in Figure 6.8b and Figure 6.9b, the fitted curves, $\text{logit}(\hat{\pi}_s(t))$ and $\hat{\pi}_s(t)$, are both close to a horizontal line across the document dates, displaying minimal peaks or troughs. An exception can be seen around the year 1440 where the estimator suddenly peaks for the shingle *et*, and dips for the shingle *omnibus*. This phenomenon is probably due to the fact that the density of documents in the fourteen hundreds is quite low, thereby increasing the variability of the estimators in this region (see equation (4.45)). Since the variance of the estimators is a function of the bandwidth, and since the optimal bandwidth used here ($h = 8$) is obtained by minimizing the dating error based on all shingles of order 1, a bandwidth value bigger than $h = 8$ (i.e. a smoother curve) may be better suited in this region. We also note, particularly, in the plots of Figure 6.8a and Figure 6.8b, a positive bias on

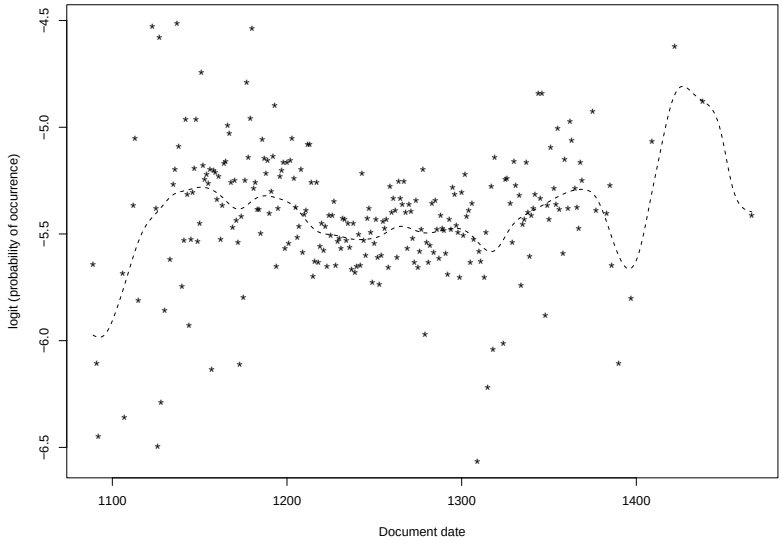
the left boundary.

The shingles "*donatorum*" ("of the donors") and "*ibidem Deo seruientibus*" or "*Deo ibidem seruientibus*" ("serving God there"), terms typically used in grants made to religious institutions, are on the other hand examples of informative shingles. For the shingle *donatorum*, we see in Figure 6.10a and Figure 6.10b that both the fitted curves, $\text{logit}(\pi_s(t))$ and $\pi_s(t)$, respectively, have prominent peaks at 1215 with another smaller nearby peak occurring at 1190. Furthermore, these curve estimates lie below the data due to the fact there are many instances where there are documents at different dates but the shingle never appears in any of these documents. This phenomenon is commonly seen in the curve estimates of various shingles.

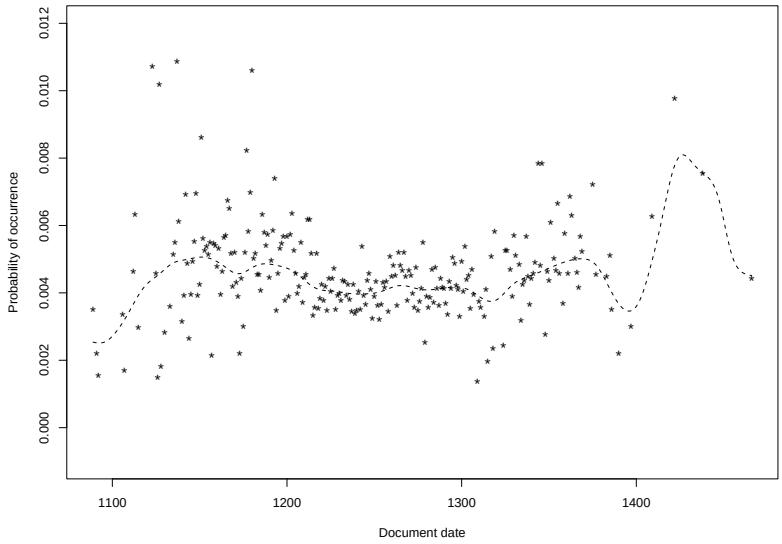
The nature of informativeness of the shingle "*ibidem Deo seruientibus*" or "*Deo ibidem seruientibus*" is different from "*donatorum*" in the sense that there are no discernable peaks that occur in the fitted curves, as can be seen in Figure 6.11a and Figure 6.11b. The fitted curves for this shingle show that its probability of occurrence is highest between the years 1115 to 1165 and drops thereafter.

Another informative shingle is "*Francis et Anglicis*" (or the version "*Franc[igen]is quam Angl[ic]is*") ("French and English"). This form of address was commonly used by French and English barons to address their men. When Normandy was lost to the French in 1204, the English renounced their tenure of lands in Normandy and the above form of address was no longer used (Gervers, 2000. p. 16). As shown in Figure 6.12a and Figure 6.12b, the estimated probability of occurrence of the above shingle declines from the early 1200's. For the purposes of illustration, we have also included additional plots overlayed on these same figures. These are plots of the functions $\text{logit}(\hat{\pi}_s(t))$ and $\hat{\pi}_s(t)$ where each function, as a function of time, is computed using a bandwidth smaller than the optimal $h = 12$. The plots of the curve estimates based on the smaller bandwidth, $h = 3$, have higher variability and are

much rougher than of those based on larger bandwidth, such as $h = 12$. This is also a theoretical fact shown by Theorem 2 and the related equation 4.45.

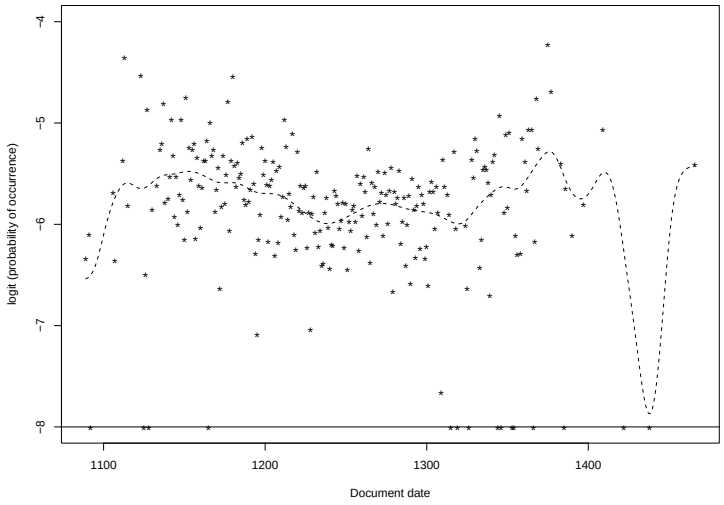


(a)

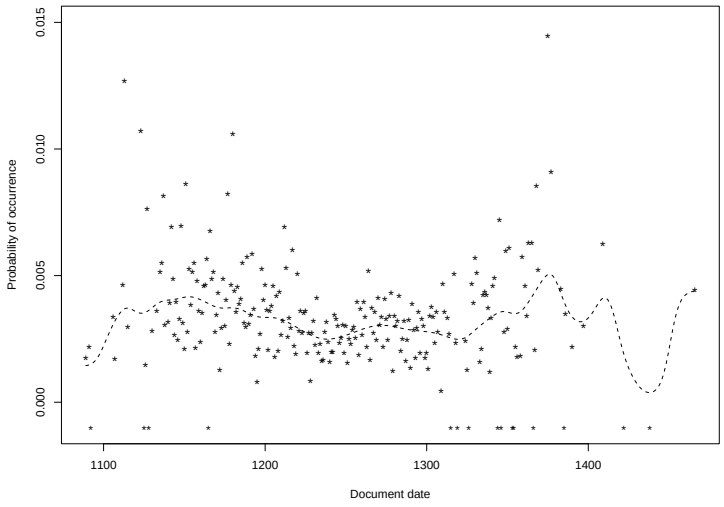


(b)

Figure 6.8: (a) Fitted logit transform of the probability of occurrence of the shingle *et* as a function of time. The curve estimates are based on degree of freedom $\nu = 5$ and bandwidth value $h = 8$. (b) Fitted probability of occurrence of the shingle as a function of time. The curve estimates are based on degree of freedom $\nu = 5$ and bandwidth value $h = 8$. The asterisk indicate the proportion of time the shingle occurs (in the case of (b), the logit of this proportion) at a date for which training documents are present.

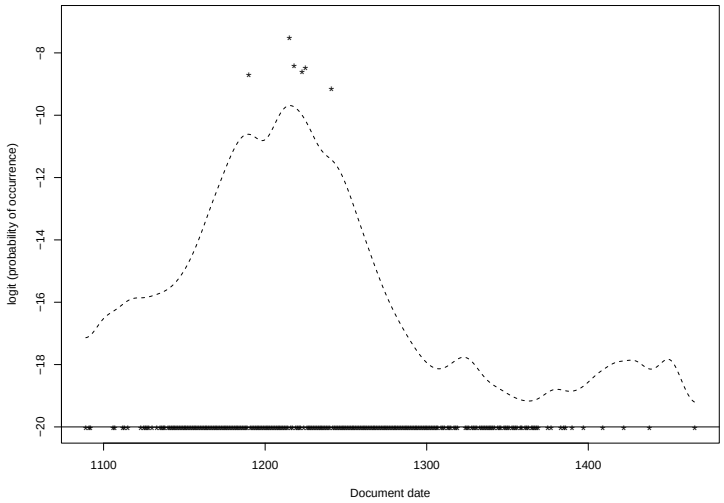


(a)

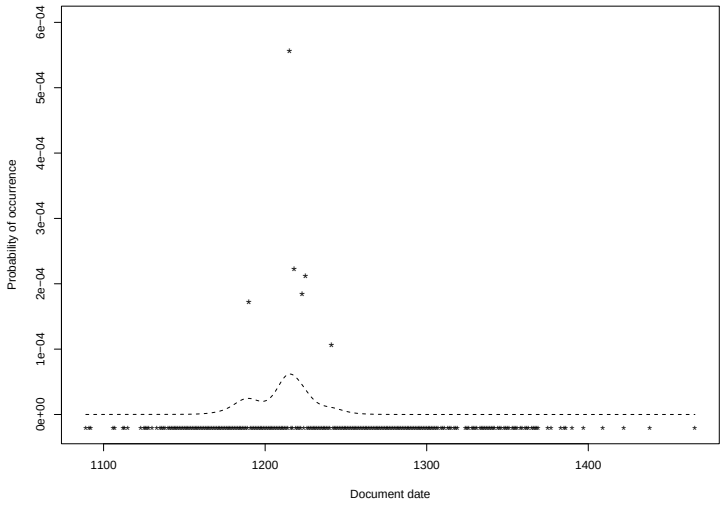


(b)

Figure 6.9: (a) Fitted logit transform of the probability of occurrence of the shingle *omnibus* as a function of time. The curve estimates are based on degree of freedom $\nu = 5$ and bandwidth value $h = 8$. (b) Fitted probability of occurrence of the shingle as a function of time. The curve estimates are based on degree of freedom $\nu = 5$ and bandwidth value $h = 8$. The asterisk indicate the proportion of time the shingle occurs (in the case of (b), the logit of this proportion) at a date for which training documents are present. The asterisk along the horizontal line in (a) indicate that the shingle was not observed even though training documents were present at the given document date. These same points are plotted below the point 0 in (b).

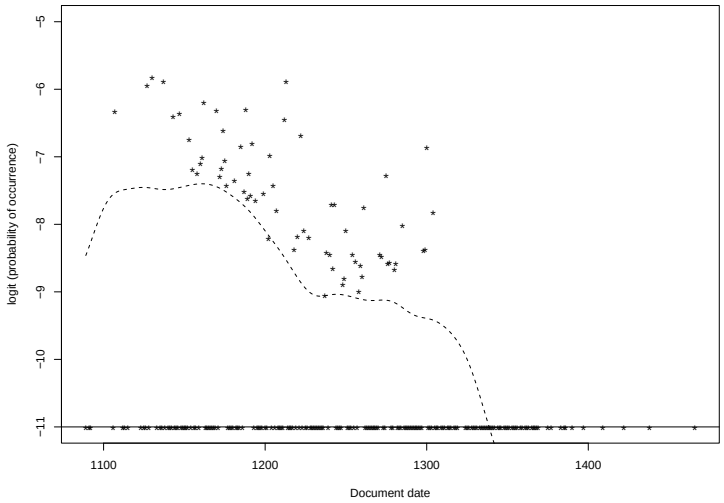


(a)

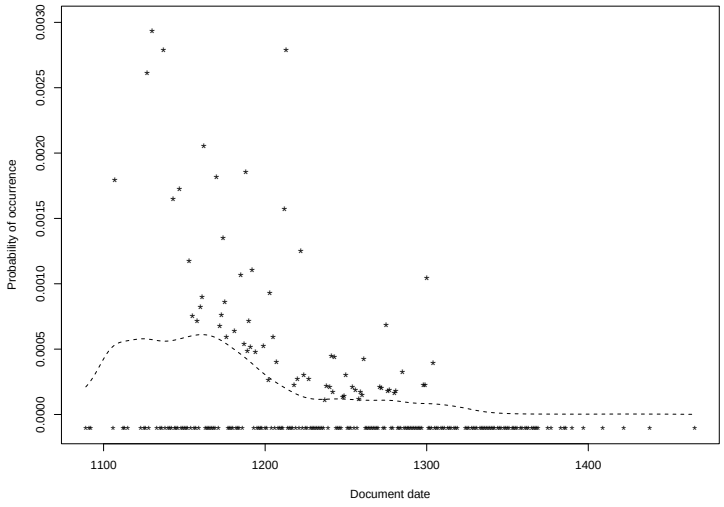


(b)

Figure 6.10: (a) Fitted logit transform of the probability of occurrence of the shingle *donatorum* as a function of time. The curve estimates are based on degree of freedom $\nu = 5$ and bandwidth value $h = 8$. (b) Fitted probability of occurrence of the shingle as a function of time. The curve estimates are based on degree of freedom $\nu = 5$ and bandwidth value $h = 8$. The asterisk indicate the proportion of time the shingle occurs (in the case of (b), the logit of this proportion) at a date for which training documents are present. The asterisk along the horizontal line in (a) indicate that the shingle was not observed even though training documents were present at the given document date. These same points are plotted below the point 0 in (b).

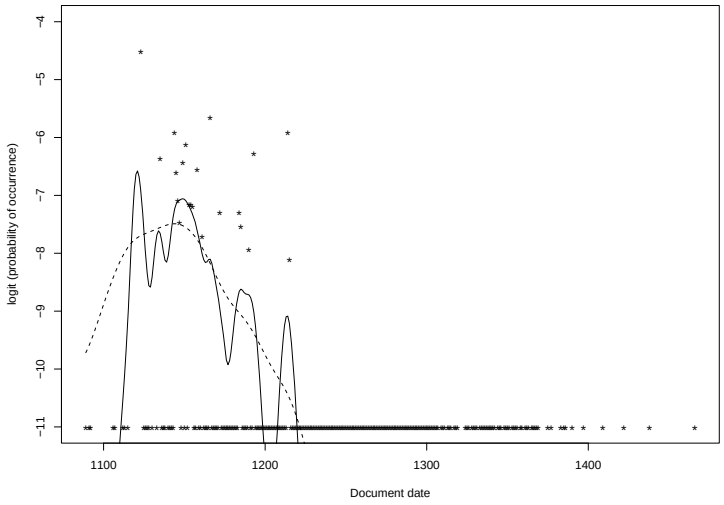


(a)

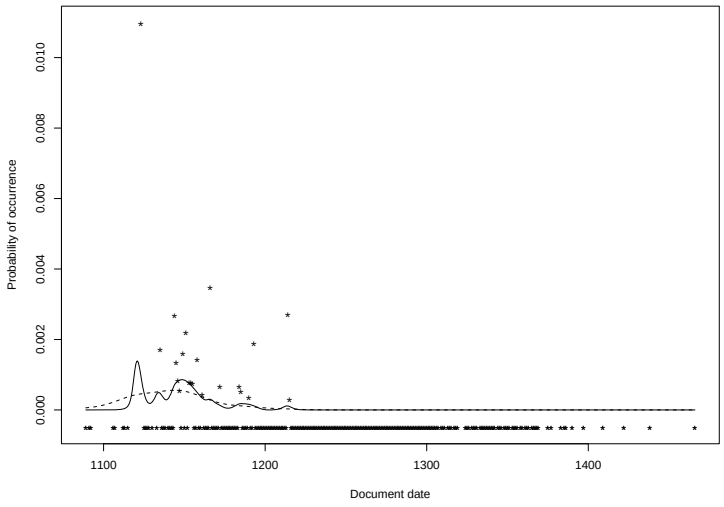


(b)

Figure 6.11: (a) Fitted logit transform of the probability of occurrence of the shingle *ibidem Deo seruientibus/ Deo ibidem seruientibus* as a function of time. The curve estimates are based on degree of freedom $\nu = 5$ and bandwidth value $h = 12$. (b) Fitted probability of occurrence of the shingle as a function of time. Curve estimates based on degree of freedom $\nu = 5$ and bandwidth values $h = 12$. The asterisk indicate proportion of time the shingle occurs (in the case of (b), the logit of this proportion) at a date for which training documents are present. The asterisk along the horizontal line in (a) indicate that the shingle was not observed even though training documents were present at the given document date. These same points are plotted below the point 0 in (b).



(a)



(b)

Figure 6.12: (a) Fitted logit transform of the probability of occurrence of the shingle *Francis et Anglicis/ Franc[igen]is quam Angl[ic]is* as a function of time. The curve estimates are based on degree of freedom $\nu = 5$ and bandwidth values $h = 12$ (dashed line), and degree of freedom $\nu = 5$ and bandwidth values $h = 3$ (solid line). (b) Fitted probability of occurrence of the shingle as a function of time. Curve estimates are based on degree of freedom $\nu = 5$ and bandwidth values $h = 12$ (dashed line), and degree of freedom $\nu = 5$ and bandwidth values $h = 3$ (solid line). The asterisk indicate the proportion of time the shingle occurs (in the case of (b), the logit of this proportion) at a date for which training documents are present. The asterisk along the horizontal line in (a) indicate that the shingle was not observed even though training documents were present at the given document date. These same points are plotted below the point 0 in (b).

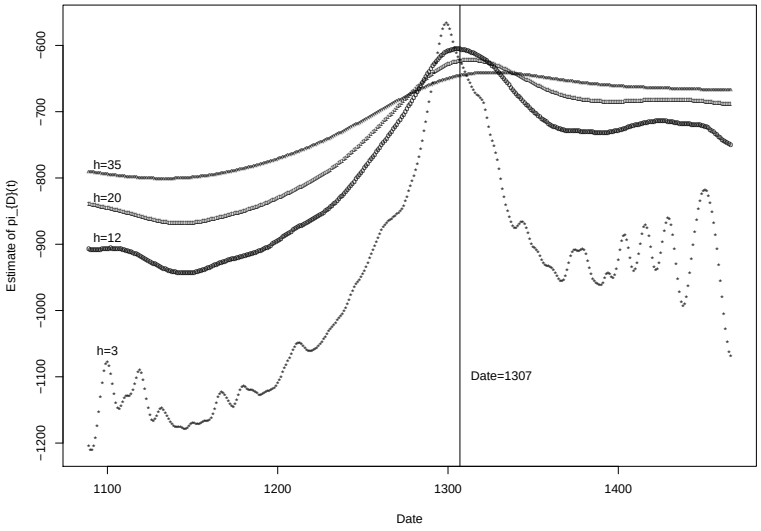
The plots in Figures 6.13a and 6.13b, are each the estimated probability of the occurrence of a document as a function of time, as given by the formula

$$\hat{\pi}_{\mathcal{D}}(t) = \prod_{s \in s_k(\mathcal{D})} \hat{\pi}_s(t).$$

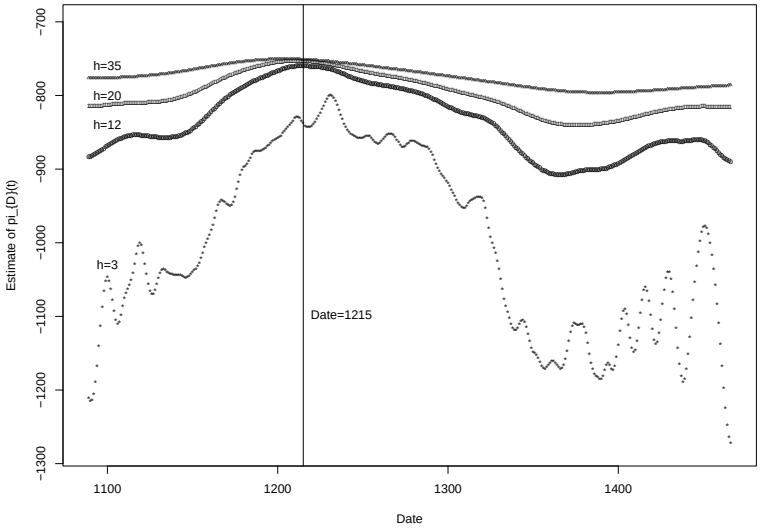
The documents used in Figure 6.13a and Figure 6.13b are both in the test set, and the function $\hat{\pi}_{\mathcal{D}}(t)$, for each document $\mathcal{D}$, is based on shingle order 2. For each document, $\hat{\pi}_{\mathcal{D}}(t)$ was computed based on four different bandwidth values; $h = 3$, 12, 20, and 35. The degree of freedom of the kernel function in all cases was set at 3. What we try to illustrate in these plots is the effect of changing bandwidth on $\hat{\pi}_{\mathcal{D}}(t)$, and ultimately, on the date estimate $\hat{t}_{\mathcal{D}} = \arg\max_t \hat{\pi}_{\mathcal{D}}(t)$. As shown in each of the plots, as the value of h increases the smoother the curve estimate of $\hat{\pi}_{\mathcal{D}}(t)$ becomes. As h decreases on the other hand, we can see the roughness (higher variance) of the curve estimator. The challenge is, of course, in deciding what level of roughness is optimal for each document. We notice, however, that no matter what the value of h is, the value of $\hat{t}_{\mathcal{D}}$ does not change much, and therefore, the method we have for determining the date estimate is relatively robust against changes in the bandwidth. The actual and the estimated dates of the document used in Figure 6.13a are 1299 and 1307, respectively, and the actual and the estimated dates of the document used in Figure 6.13b are 1222 and 1215, respectively.

Another observation we would like to make with regards to the Figure 6.13a and Figure 6.13b is that the peak of the plot in Figure 6.13a is more prominent or "pointier" than that of the plot in Figure 6.13b. Quantifying the peakedness of such plots may be a step to constructing a confidence interval for a date estimate.

The descriptions of the computer codes used in this section can be found in the appendix.



(a)



(b)

Figure 6.13: The plots (a) and (b) are that of the function $\hat{\pi}_{\mathcal{D}}(t)$ for two separate documents in the test set. The function $\hat{\pi}_{\mathcal{D}}(t)$ is plotted for varying values of the bandwidth h. In both plots, the documents were based on shingle order 2. The actual date for the document used in (a) is 1299 and the estimated date is 1307 (vertical line), for document used in (b), the actual date is 1222 and the estimated date is 1215 (vertical line). In all cases the degree of freedom $\nu = 3$.

Chapter 7

Conclusions and Future Research Directions

7.1 Combining date estimates

As discussed in Chapter 5, Section 5.3, using the mean square error as the criterion for selecting the optimal value for the parameter (there, the parameter, m, is the closest number of documents in the training set to the document which we are trying to date), we found that the distance based method based on shingle order 1 and $m = 500$ performed the best ($\sqrt{\text{MSE}} = 20.07$). However, the measurement, $\sqrt{\text{MSE}}$, MAE and MedAE based on the combined shingle orders 1 and 2 for $m = 100$, were still very close to those of shingle order 1 and $m = 500$. (See Tables 5.1 and 5.4.) In Chapter 6, Section 6.5, again using mean square error as the criterion for selecting the optimal parameters (in that case, the parameters of interest were the optimal bandwidth h, and the degree of freedom ν of the kernel) we found that for the maximum prevalence method that based on shingle order 2 with $h = 12$ and $\nu = 3$ gave the best result ($\sqrt{\text{MSE}} = 14.7$ on the test set). Considering the superior performance of shingle order

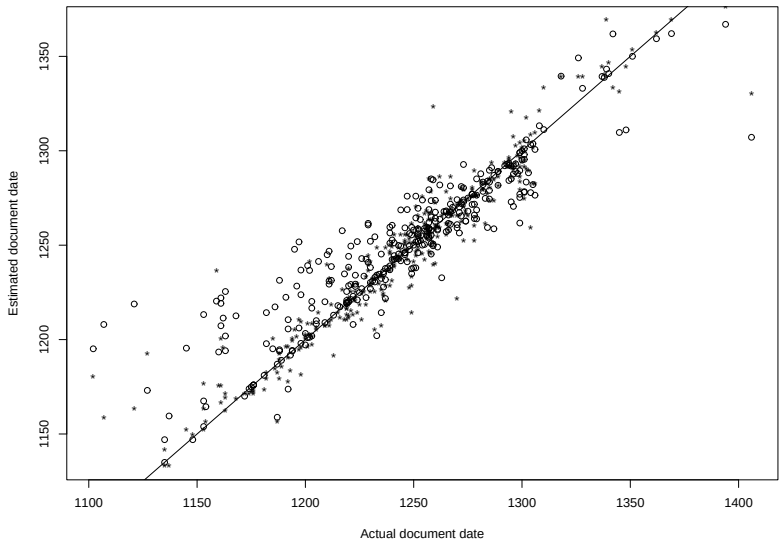
2 to all the other shingle orders in the maximum prevalence method, we decided to pick the date estimation method based on the combinations of shingle orders 1 and 2 (as oppose to shingle order 1 alone) as the best dating methodology among the distance based methods. On the test set, the combination of shingle orders 1 and 2 for $m = 100$ out performs all the other methods based on shingle combinations of orders 1 and 2 with values of m different from 100. On the 326 documents of the test set, the results for the combination of shingle orders 1 and 2 for $m = 100$ are, $\sqrt{\text{MSE}} = 20.45$, MAE = 12.32 and MedAE = 6.35.

Figure 7.1a is a plot of the best date estimate based on distance based method versus the best date estimate by the method of maximum prevalence on 326 documents from the test set. As we can see from the plot, the date estimates based on maximum prevalence method have a lower margin of error (i.e. they are closer to the $X = Y$ axis) particularly for the earlier dates (1100 to 1175) and the later dates (1325 to 1400). The plot in Figure 7.1b is the difference in date estimates (date estimates based on distance-based method minus date estimates based on maximum prevalence method) versus the actual document dates. The date estimates based on the distance-based measures are larger than those of the date estimates based on the maximum prevalence method for the years ranging from 1100 to about 1225. On the other hand, for the years ranging from about 1275 to 1400, the date estimates based on the maximum prevalence method are slightly larger than the date estimates based on the distance-based method.

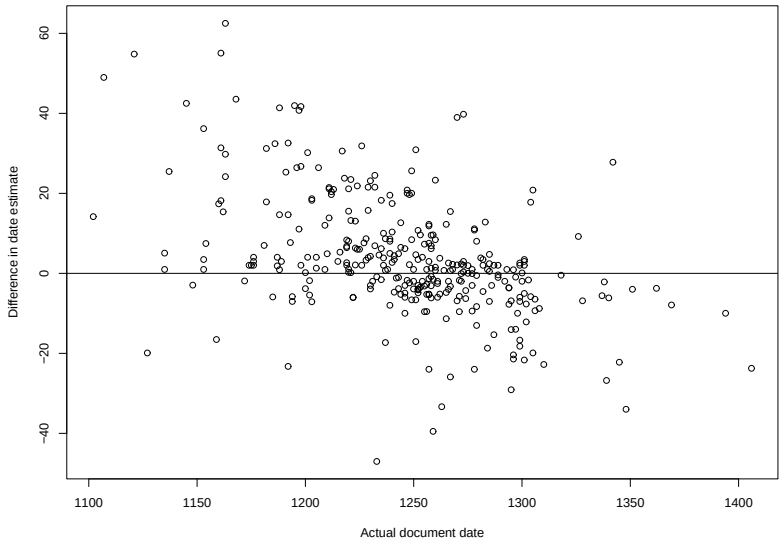
At this point we wish to address the question of whether there is a statistically significant difference between the best date estimates based on the distance-based and the maximum prevalence methods, and if there is, how to best combine these two methods for date estimation. To test the null hypothesis that the two dating methodologies are not significantly different, we employed a permutation test based

on the normalized differences in mean squared error. This test proceeds as follows.

We begin with two column vectors of length 326, containing the two sets of 326 date estimates for the documents in the test set (one set of date estimates based on the document distance-based method, and the other based on the maximum prevalence method). Also, the two vectors are matched pairs - that is, for a given row, say row i, the first and the second elements of row i are the date estimates of the ith document based on the distance-based and the maximum prevalence methods, respectively. A series of 1000 permutations of these two vectors is performed, where each permutation consists of randomly switching or not switching (with probability 1/2) the corresponding elements of the two vectors. For each such permutation, we compute the difference between the mean squared errors of the estimators corresponding to the scrambled column vectors. This produces an empirical distribution of differences of mean squared errors. Under the null hypothesis, the difference of the sample mean squared errors of the two estimators is not likely to be in the tail this distribution. Figure 7.2 is a histogram of the difference between the mean squared dating error based on 1000 permutations. We found the difference in the mean squared error between the two original dating methodologies to be 2.6, which is off scale and indicates a highly significant difference between the two dating methods.



(a)



(b)

Figure 7.1: (a) A plot of actual document date versus date estimates of documents from the test set. Date estimates based on document distance measure are indicated by "o" and date estimates based on the maximum prevalence method are indicated by "*". (b) A plot of the difference in date estimates (distance-based method minus maximum prevalence method) versus actual document date.

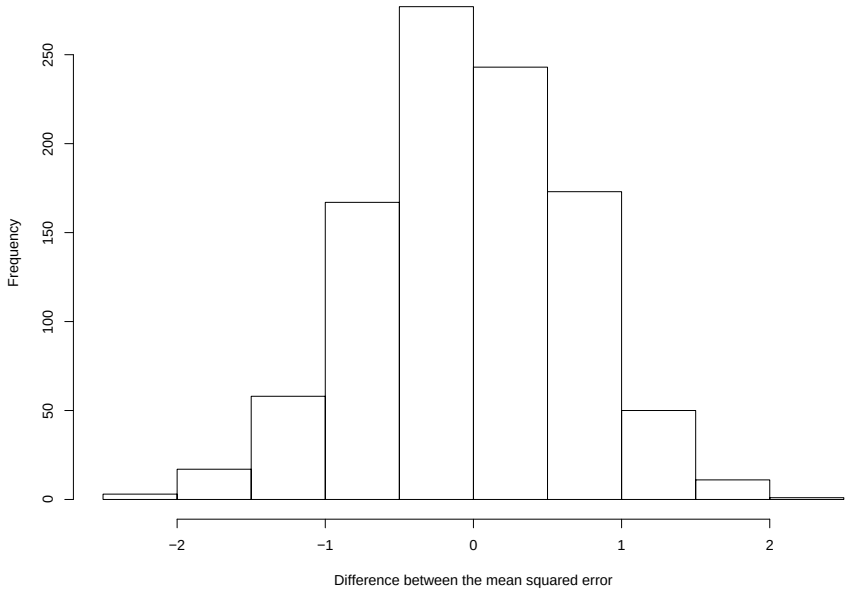


Figure 7.2: A histogram of the difference between the mean squared dating error based on 1000 permutations. The permutation was run on 326 documents from the test set. The difference in the mean squared error between the distance-based date estimates and the maximum prevalence method date estimates is 2.6.

On the question of how to combine the date estimates from the two dating methodologies, we could take the average of the date estimates, in which case we give equal weight to each of the dating methodologies in the new date estimates, or we could, for example, find the optimal coefficients for the linear combination between the two dating methodologies based on documents from the validation set. Doing this, we find the optimal coefficient for the maximum prevalence method and the distance-based method are 0.83 and 0.17, respectively. The resulting $\sqrt{MSE}$ on the test set is 13.46 years.

7.2 Comparison of the distance based method and maximum prevalence method

In this section we discuss some of the advantages and limitations of the document distance based method of Chapter 5, and the maximum prevalence method of Chapter 6 for document dating.

The document distance based method of Chapter 5 can accommodate ordered and unordered categorical variates whereas the maximum prevalence method of Chapter 6 is not designed for such purpose.

The document distance based method is simple - it is essentially weighted sums of dates from the training data where the weights are based on the distance between the training data and the document whose date we are trying to estimate. One downside of this method, however, is that unlike maximum prevalence, it does not automatically scale down the contribution of uninformative or insignificant shingles. As discussed in Chapter 6, Section 6.5, the uninformative shingles, for example, *et* and *omnibus*, have fitted curve $\hat{\pi}_s(t)$ that is close to a horizontal line. Therefore, for a document $\mathcal{D}$ containing the shingles $s =$ *"et"* and/or $s =$ *"omnibus"*, the contribution of $\hat{\pi}_s(t)$ to the evaluation of $\arg\max_t \hat{\pi}_{\mathcal{D}}(t)$ is minimal.

The document distance based method has few model assumptions whereas the maximum prevalence method assumes words (shingles) are independently and binomially distributed. As a result, the latter method allows us to model the probability of occurrence of individual shingles as a function of time. This feature of the maximum prevalence method is particulary useful for historians (see Gervers and Hamonic, 2010). Furthermore, the maximum prevalence method, at least in principle, can eliminate edge bias in the date estimates, whereas the distance-based method can not.

The document distance based method of Chapter 5 allows for the implementation

of an optimal bandwidth for each document whose date we are aiming to estimate. The maximum prevalence method of Chapter 6, however, uses the same bandwidth for all the estimates of the probability of occurrences of shingles. Perhaps customizing the bandwidths for each of these estimates would be one way to try to improve the document date estimates based on the maximum prevalence method.

7.3 Identifying informative words

In this section, we investigate the characteristics of words or shingles that are most useful for document dating. There are a total of 42,978 distinct words in the training documents, and among this large number of words one can reasonably expect there to be characteristics that render certain words more informative or significant than others. The aim of this exercise is to examine the possibility of extracting, in an automated fashion, informative words or shingles for the purposes of (in the context of this thesis) document dating. Some of the obvious characteristics that will be examined for a given word are:

(a) The standard deviation of dates when the shingle occurs. Intuitively we expect a high standard deviation would mean a lack of concentration of a shingle at any particular time frame, thus rendering it less informative.

(b) The number of documents in which the shingle occurs. This characteristic measures how rare a shingle is versus how informative it is when it occurs.

To assess the influence of the above characteristics of words for predicting dates, the following experiment was conducted. Randomly pick N documents. From these N documents, pick K words which exhibit a range of standard deviations. Let

$$x_{ij} = \begin{cases} 1 & \text{if word } i \text{ is in document } j \\ 0 & \text{otherwise} \end{cases}$$

where i ranges from 1 to K, and j ranges from 1 to N. Let $\mathbf{x}_i$ denote the vector of length N whose entries are x_{ij} where i is fixed. Also, let $d_j,\ j = 1, \ldots, N$ be the true dates of the randomly picked documents. We then ran a tree regression where we take the response variables to be the dates of the document (the d_j's) and the predictors to be the $\mathbf{x}_i$'s. Finally, let

$$y_i = \begin{cases} 1 & \text{if word } i \text{ is used in the fit} \\ 0 & \text{otherwise .} \end{cases}$$

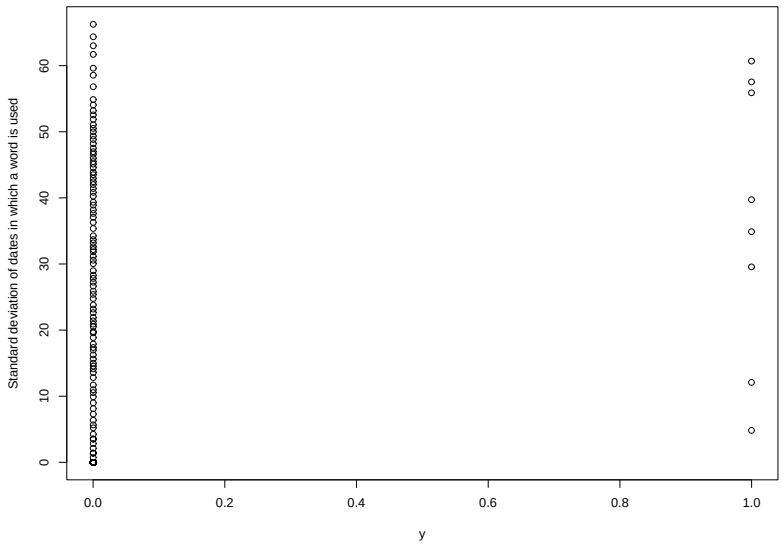
In this experiment N was taken to be 300 and K was taken to be 300.

Results: The above experiment was carried out on shingles of orders 1 and 2 separately. Of the 300 shingles of order 1, the tree regression picked 8 of them as being informative, and the remaining 292 were rendered as uninformative. For shingles of order 2, the tree regression picked 3 of the shingles as being informative and the remaining 297 as uninformative shingles. The plots in Figures 7.3a- 7.4b below display the characteristics (a) and (b) of the informative and uninformative shingles of orders 1 and 2. In Figure 7.3b, we see that the informative shingles of order 2 are those that have a higher standard deviation for the dates in which they occurred. This however is not is the case for shingles of order 1, as seen in Figure 7.3a. In the latter case, the standard deviation, as a discriminant measure, is not adequate, especially, for shingles that have arbitrary distribution (in time). Regarding characteristics (b), for both shingles of order 1 and 2, the informative shingles are those that occur with reasonably high frequency - those shingles that occur rarely or too frequently are not picked as being informative. Using the number of documents in which a shingle occurs as a measure of discrimination is supported especially for shingles of order 2.

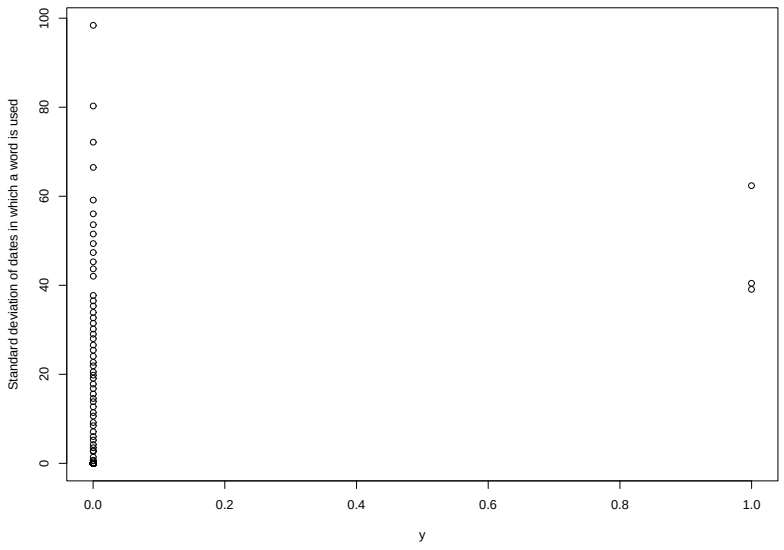
In Figure 7.5, we have plotted the probability of occurrence as a function of time of the three order 2 informative shingles - *"matris ecclesie"*, *"predictum est"* and *"quod nos"*. For purposes of comparison, in Figure 7.6, we have also selected and plotted the

probability of occurrence as a function of time of three other order 2 uninformative shingles - *"testimonium huic"*, *"beate Marie"* and *"in eadem"*. Characteristics (a) and (b) of these uninformative shingles is comparable to the three informative shingles. The data points that lie below the value 0 of the y-axis indicate that the shingles of interest do not occur in the corresponding date of the x-axis even though there are documents in those dates.

It is comforting that the findings of the tree regression, specifically the conclusion that the informative shingles are those that occur with reasonably high frequency (not too frequently or not too rarely), matches with the findings of Luhn (Luhn, 1958). While working at IBM, Luhn was trying to develop methods for creating abstracts for technical papers and magazine articles by automatic means. In this attempt, Luhn relied on Zipf's law (Zipf, 1949) which states that the product of the frequency of words, f, in a text, and the rank of their order, r, is approximately a constant, where this constant is text dependent. Luhn plotted $\log(r)$ versus $\log(f)$, and by trial and error determined an upper cut-off and a lower cut-off points on the $\log(r)$ axis. Words that exceeded the upper cut-off were considered to be too rare and those that fell below the lower cut-off were considered to be too common in discriminating content. He supposed that the ability of words to discriminate content reached its peak at the rank order exactly half way between the two cut-off points. Grossman and Frieder (1998, p.11) state that Luhn's idea is the basis of many of the techniques in information retrieval. In fact, the *idf* term weight (defined in Chapter 3, Section 3.3.1 as $\log(N/n_i)$, where N is the total number of documents in the collection and n_i is the number of documents that contain the term i,) is one such method for discriminating non-significant words.

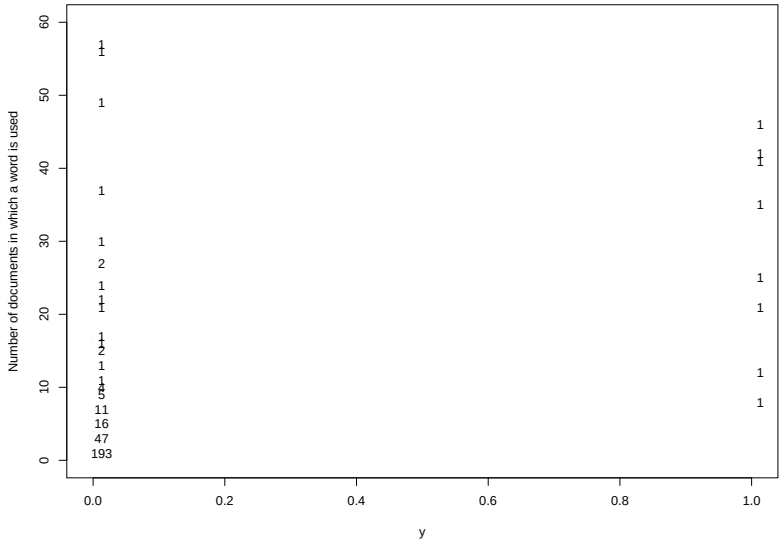


(a)

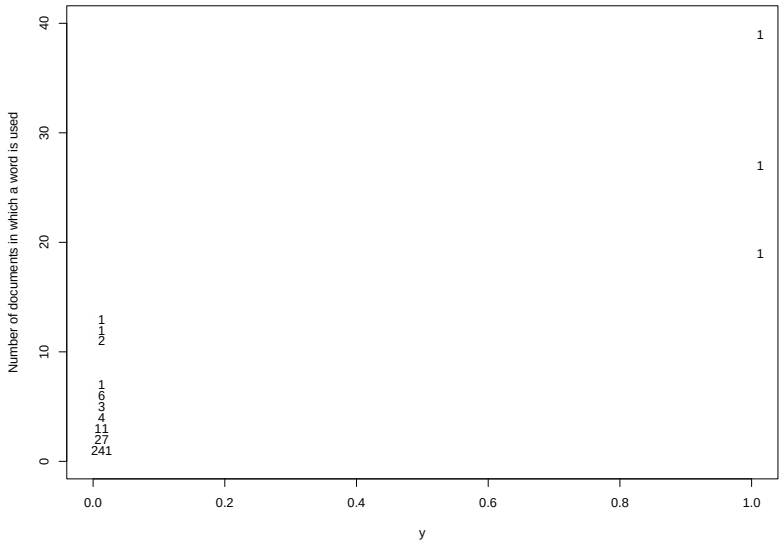


(b)

Figure 7.3: A plot of the standard deviation of dates in which a word is used versus y, where $y = 0$ indicates uninformative words and $y = 1$ informative words. Words of shingle order 1 are shown in part (a), and words of shingle order 2 are shown in part (b).

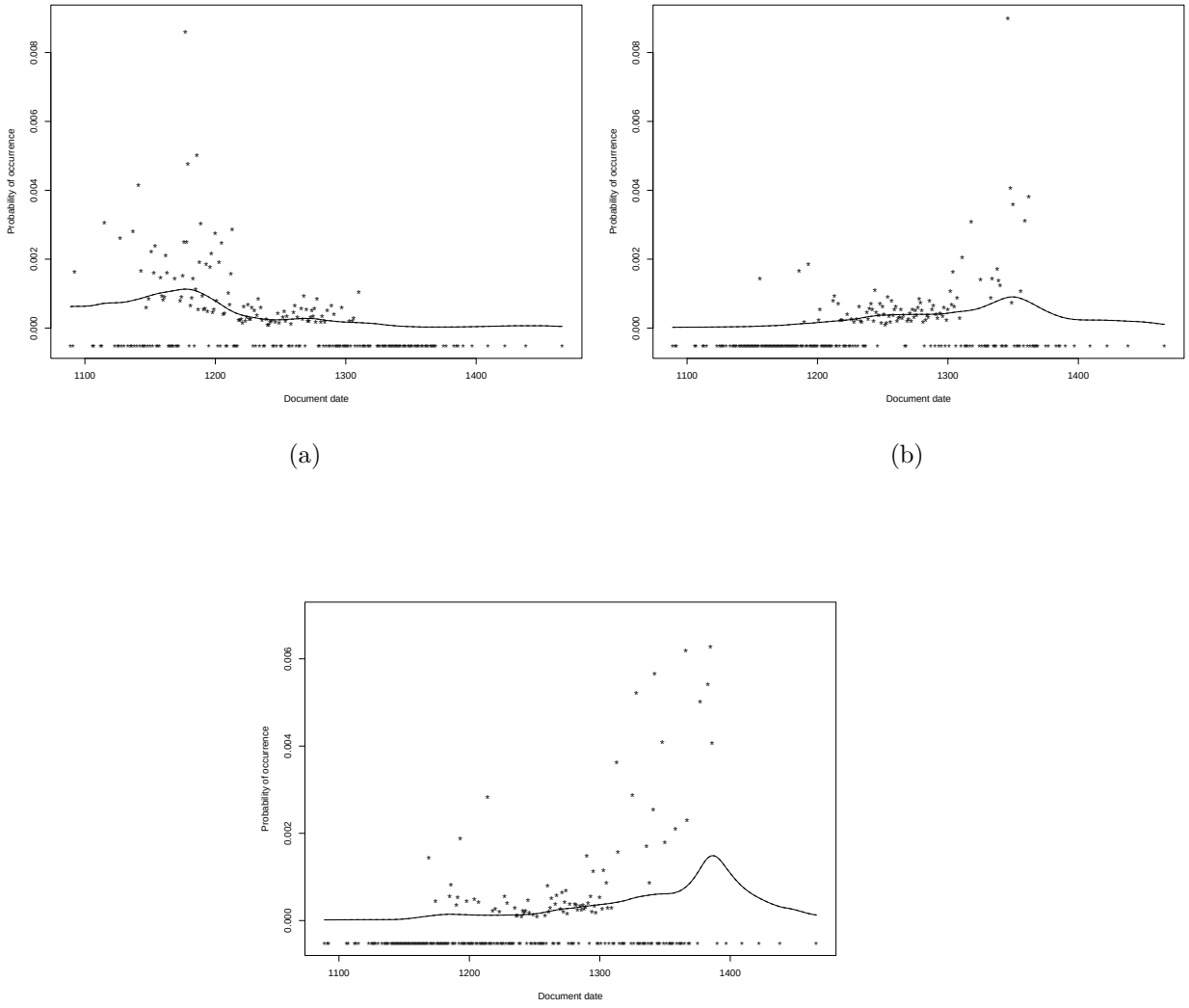


(a)



(b)

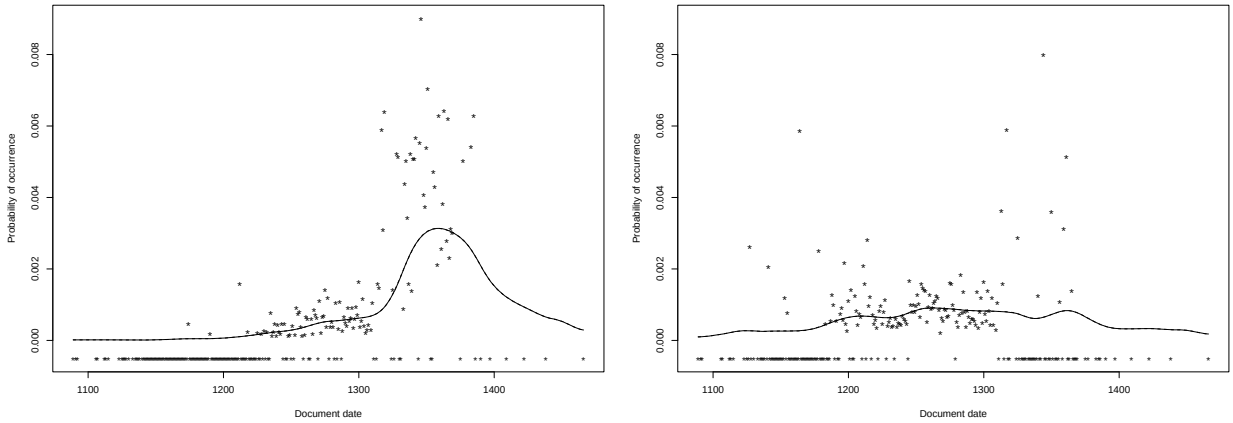
Figure 7.4: A plot of the number of documents in which a word is used versus y, where $y = 0$ indicates uninformative words and $y = 1$ informative words. Words of shingle order 1 are shown in part (a), and words of shingle order 2 are shown in part (b). The numeric labels in the plots indicate the total number of words found at '(x,y)' point.



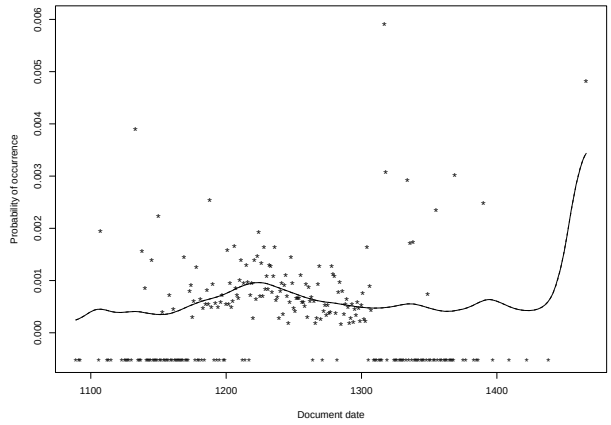
(a) (b)

(c)

Figure 7.5: Fitted probability of occurrence of the informative shingles *matris ecclesie*, *predictum est* and *quod nos* shown in plots (a), (b) and (c), respectively. Curve estimates are based on degree of freedom $\nu = 3$ and bandwidth value $h = 12$. The asterisk indicate the proportion of time the shingles occur at a date for which training documents are present. The asterisk below the 0 point indicate the shingles were not observed even though training documents were present at the given document date.



(a) (b)



(c)

Figure 7.6: Fitted probability of occurrence of the uninformative shingles *testimonium huic*, *beate Marie* and *in eadem* shown in plots (a), (b) and (c), respectively. Curve estimates are based on degree of freedom $\nu = 3$ and bandwidth value $h = 12$. The asterisk indicate the proportion of time the shingles occur at a date for which training documents are present. The asterisk below the 0 point indicate the shingles were not observed even though training documents were present at the given document date.

7.4 Future research direction

The document dating estimators we have developed in Chapter 5 (distance based method) and Chapter 6 (maximum prevalence method) are obviously functions of the training data. It would be useful from the computational and data storage aspects to examine what relationship there is between the estimation error and size of the training data. As shown in the plots of Figures 6.9b, 6.10b, 6.11b, 6.12b, and Figures 7.6a -7.6c, many words have zero occurrence at most of the document dates. An increase in training size may therefore be helpful in better estimating the probability of occurrence of these words as a function of time.

Even if the amount of training data was to increase to infinity, following the discussions in Chapter 6, Section 6.4, $\hat{t}_{\mathcal{D}} \to t^*_{\mathcal{D}}$, but $t^*_{\mathcal{D}}$ is not necessarily equal to $t_{\mathcal{D}}$, the actual document date. What accounts for the difference between $t^*_{\mathcal{D}}$ and $t_{\mathcal{D}}$, in part, we presume, is the limitation of our dating methodology. For example, we are not taking into account in the dating methodology the construction of the natural-language of Latin, nor do we weigh differently the informative and the uninformative words. Furthermore, there is inherent randomness, for example, variability of an individual's writing styles. Moreover, since a document by nature has a finite length, we are observing only a finite sample of words from what potentially would have been an infinite stream of words, even if the training date is infinite. All of this contributes to the gap between $\hat{t}_{\mathcal{D}}$ and $t^*_{\mathcal{D}}$, and $t^*_{\mathcal{D}}$ and $t_{\mathcal{D}}$.

Further, we could study the performance of date estimates based on the maximum prevalence of shingle occurrences (Chapter 6) if the associated probabilities were to be based on local polynomial kernel regression for the generalized linear models, where the order of the polynomial is odd and the generalized linear model is other than the logistic. As discussed in Chapter 4, polynomials of odd degrees have attractive

properties, such as simpler forms of bias as well as absence of edge bias. However, we note that no matter how good our model is in all aspects, our ability to analyze the DEEDS documents in this thesis is limited since the approach we have taken does not go further than analyzing the matching of words or shingles. We do not, for instance, as computational linguists would, delve into the grammatical structures of Latin (the language in which the DEEDS documents are written) nor do we take into account content meanings. For example, the content meaning of the phrase *The Big Apple* in certain cases is synonymous with *New York*. However, in the approach we have taken, the synonymity of these two phrases would not be represented in the models.

The results of our experiment from using the tree regression to classify the characteristics of informative shingles seems promising. However, we still need to incorporate the results from the tree regression in the dating methodologies discussed in Chapter 5 and Chapter 6.

Moreover, as stated in the introduction, the DEEDS documents are also accompanied with information regarding the properties of the documents, for example, the religious house which a witness belongs to, the location of the property and whether the property of a transfer or an agreement. At the moment, the dating procedures of the distance-based and the maximum prevalence methods do not make use of these informations. Finding ways to incorporate them could potentially improve the date estimates of the DEEDS documents.

Since it is useful to have knowledge of the uncertainty of any statistical estimate, it would be worth while to build a confidence interval for the date estimates of the DEEDS documents. It would be useful to investigate the accuracy of the date estimate of a document we wish to date as it relates to the "peakedness" of the curve of the probability of occurrence of the document as a function of time. This idea was briefly mentioned in the closing of Chapter 6, Section 6.5. in relation to the plots in

Figure 6.13a and Figure 6.13b. We expect that the sharper the peaks, the more accurate the date estimate, but the limitations of this intuition need to be investigated. Being able to quantify the peakedness of the curves of the probability of occurrence of documents may allow us to build confidence intervals.

It would also be interesting to examine the type of words and probability models for word frequency other than the binomial-such as the Poisson or the negative binomial as studied by Mosteller and Wallace in determining the authors of the disputed essays of the Federalist Papers (Mosteller and Wallace (1963)). In this regard, we can also draw lessons from the work of Margulis (1992) in which he investigates the validity of the multiple (nP) model of word distribution in document collection. ((nP) distribution is defined to mean a mixture of n Poisson distributions with different means). Margulis provides a test to determine whether a certain word follows an nP distribution and shows that 70% of the frequently occurring words indeed follow the nP distribution. In light of the results of Mosteller and Wallace, and Margulis, it would be useful to find ways to extract informative words on which to base the document dating algorithms. The potential benefit could be an increase in computational speed as well as accuracy in document date estimates.

In other fields of inquiry, such as in comparative linguistics, where one is interested in reconstructing a proto-language (the common ancestor of languages that form a language family), the various distance measures and their properties as discussed in Chapter 3 may be of use (see Kondrak, 2002). In the field of bioinformatics, the concept of shingles and distance measures can also be useful in the study of sequence alignment (Felsenstein, 2004; Altschul, 2003; Karlin and Altschul, 1990; Lipman *et al.*, 1989). The study of sequence alignment investigates the ways in which the sequences of DNA, RNA or proteins are aligned to another homologous sequence for the purposes of comparison in order to identify regions of similarity. This in turn

provides clues to researches to the evolutionary history and function of genes.

Appendix A

Kernel Density Estimation

The considerations which arise in kernel density estimation are central to the development of the methods of nonparametric regression and generalized linear models and local smoothing discussed in Chapter 4. In this appendix, we provide an overview of kernel density estimation, starting with the histogram estimator. The asymptotic properties of such estimators, as well as methods for bandwidth selection will be described in some detail.

A.1 The histogram

The histogram is one of the simplest non-parametric density estimators. Suppose we have a sample $X_1, \ldots, X_n$ of independent and identically distributed observations from an unknown density function f. The motivation for the construction of the histogram estimator of f is based on the definition of the density function:

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} \quad , \tag{A.1}$$

where $F(x)$ is the cumulative distribution function (c.d.f) of the random variable X. The natural estimator $\hat{F}$ of the c.d.f F is

$$\hat{F}(x) = \frac{\#\{X_i \leq x\}}{n} \quad .$$

Then, given an origin x_0 and a *bin length of* h, define the bin B_m to be

$$B_m = (\ x_0 + (m-1)h,\ x_0 + mh\ ] \ ,$$

where $h > 0$ and $m \in \mathcal{Z}$ where $\mathcal{Z}$ are the integers. From equation $(A.1)$ and the estimate $\hat{F}(x)$, we are led to estimate $f(x)$ as

$$\begin{aligned} \hat{f}(x) &= (nh)^{-1}\#\{x < X_i \leq x + h\} \\ &\approx (nh)^{-1}\#\{x_0 + (m-1)h < X_i \leq x_0 + mh\} \end{aligned}$$

for $x \in B_m$, i.e. , for $x \in [\ x_0 + (m-1)h,\ x_0 + mh\ )$. The histogram density estimate, $\hat{f}_{hist}$, is then defined as

$$\hat{f}_{hist}(x) = (nh)^{-1} \sum_{i=1}^{n} \sum_{m} I(X_i \in B_m) I(x \in B_m) \ .$$

Of course the histogram depends on the bandwidth h and the origin x_0 and its shape can be influenced significantly as x_0 varies.

A.2 The kernel density estimator

Originally suggested by Rosenblatt (1956), the kernel density estimate overcomes some of the problems of the histogram. The kernel estimator is based on a weight function K called the *kernel function*, satisfying the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1 \quad .$$

Although it is not a formal requirement, K is generally also taken to be symmetric about the origin and non-negative. Furthermore, introducing the notation

$$\mu_2(K) \equiv \int_{-\infty}^{\infty} x^2 K(x) dx \ ,$$

we typically also require that $0 < \mu_2(K) < \infty$. An example of a kernel function satisfying the above condition is the density of a normal distribution having mean zero and variance $\mu_2(K)$.

The *kernel density estimator* is then defined to be

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where h is called the *bandwidth* or the *smoothing parameter*. A more compact formula for the kernel estimator is obtained by introducing the rescaled notation $K(u) = h^{-1}K(u/h)$. This would allow us to rewrite

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \ .$$

The motivation behind the kernel density estimator can be seen if we revisit a definition of $f(x)$ equivalent to $(A.1)$:

$$f(x) = F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} \ .$$

Then, unlike the histogram estimate which divides up the line into bins, we estimate the derivative separately for each point x:

$$\hat{f}_h(x) = \frac{\text{number of } X_i\text{'s that fall in } (x-h, x+h]}{2nh} \ .$$

We can rewrite this as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \ \ ,$$

where

$$K(u) = \begin{cases} 1/2 & \text{if } -1 < u \leq 1 \\ 0 & \text{otherwise .} \end{cases}$$

The form of the resulting kernel density estimator is based on the uniform density function on (-1,1].

The density estimate inherits some of the mathematical properties of K. For example, if K is non-negative and a density, $\hat{f}$ will be too. If all derivatives of K exist (as for example when K is a normal density function), then all orders of derivatives of $\hat{f}$ will also exist.

A.3 Properties of the kernel density estimator

There are several ways to measure the "closeness" between the density estimate $\hat{f}_h$ and the underlying true density f. Some of the more commonly used measures are: the pointwise *mean squared error* (MSE) defined as

$$MSE(\hat{f}_h(x)) = E\{\hat{f}_h(x) - f(x)\}^2 ,$$

the *integrated square error (ISE)* defined as

$$\text{ISE}(\hat{f}_h) = \int \{\hat{f}_h(x) - f(x)\}^2 dx ,$$

and the *mean integrated squared error* (MISE) defined as

$$MISE(\hat{f}_h) = E \int_{-\infty}^{\infty} \{\hat{f}_h(x) - f(x)\}^2 dx .$$

Here we will examine only the $MSE(\hat{f}_h(x))$; we will discuss $ISE(\hat{f}_h)$ and $MISE(\hat{f}_h)$ under the subsection "Bandwidth Selection" where these quantities will be used for determining optimal bandwidth values.

Using basic manipulations, the $MSE(\hat{f}_h(x))$ can be rewritten as

$$\begin{aligned} MSE(\hat{f}_h(x)) &= \mathrm{Var}(\hat{f}_h(x)) + \{E\hat{f}_h(x) - f(x)\}^2 \\ &= \mathrm{Var}(\hat{f}_h(x)) + \mathrm{Bias}^2(\hat{f}_h(x)). \end{aligned}$$

The bias of $\hat{f}_n(x)$ can be computed as follows:

$$E(\hat{f}_h(x)) = \frac{1}{n}\sum_{i=1}^{n} E(K_h(x - X_i)) = \int K_h(x-u)f(u)du = \int K(s)f(x+sh)ds.$$

If we let $h \to 0$, then

$$E(\hat{f}_h(x)) \longrightarrow \int K(s)f(x)ds = f(x),$$

so that $\hat{f}_h(x)$ will become unbiased as $h \to 0$. The bias of $\hat{f}_h(x)$ can be further analyzed via Taylor expansion. Assuming $f \in C^2$, i.e. that f is twice continuously differentiable, and that the kernel K is symmetric about zero, then, as $h \to 0$, we have

$$\begin{aligned} \mathrm{Bias}(\hat{f}_h(x)) &= \int K(s)f(x+sh)ds - f(x) \\ &= \int K(s)\left\{f(x) + f'(x)sh + \frac{h^2s^2}{2}f''(x) + o(h^2)\right\}ds - f(x) \\ &= f(x) + \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2) - f(x) \\ &= \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2)\ . \end{aligned}$$

Similarly, as $nh \to \infty$, the variance of $\hat{f}(x)$ can be evaluated as follows:

$$\begin{aligned} \mathrm{Var}(\hat{f}_h(x)) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\right) = \frac{1}{n}\mathrm{Var}(K_h(x - X_1)) \\ &= \frac{1}{n}\left\{E(K_h^2(x - X_1) - E^2(K_h(x - X_1))\right\} \\ &= \frac{1}{n}\left\{\frac{1}{h}\int K^2(s)f(x+sh)ds - (f(x) + o(h))^2\right\} \\ &= \frac{1}{n}\left\{\frac{1}{h}||K||_2^2(f(x) + o(h)) - (f(x) + o(h))^2\right\} \\ &= (nh)^{-1}||K||_2^2(f(x) + o(h)) - n^{-1}(f(x) + o(h))^2 \\ &= (nh)^{-1}||K||_2^2 f(x) + o((nh)^{-1}). \end{aligned}$$

Here $||K||_2 = (\int K^2(x)dx)^{1/2}$. The asymptotics of MSE$(\hat{f}_h(x))$, as $h \to 0$ and $nh \to \infty$, is therefore determined by

$$MSE(\hat{f}_h(x)) = \frac{1}{nh}||K||_2^2 f(x) + \frac{h^4}{4}(\mu_2(K)f''(x))^2 + o((nh)^{-1}) + o(h^4)\ . \qquad (A.2)$$

Some key points which follow from the approximation (A.2), are:

(a) The asymptotically optimal bandwidth for $MSE(\hat{f}_h(x))$, i.e., the value of h that minimizes $MSE(\hat{f}_h(x))$, has the form

$$h_{mse}(x) = \arg\min_h MSE(\hat{f}_h(x)) = \left(\frac{f(x)||K||_2^2}{(f''(x))^2(\mu_2(K))^2 n}\right)^{1/5}. \qquad (A.3)$$

(b) If we increase the bandwidth parameter h, resulting in smoother density estimate, the variance term decreases but the bias term increases. Conversely, decreasing the bandwidth parameter h results in a less smooth density estimate having smaller bias but increased variance.

(c) If $h \to 0$ and $nh \to \infty$, then $MSE(\hat{f}_h(x)) \to 0$. Thus, it follows by an application of the Chebychev inequality, that $\hat{f}(x)$ will then be a consistent estimator of $f(x)$.

(d) The optimal bandwidth, $h_{mse}(x)$, depends on the unknown functions f and f'' and on the particular value of x.

The dependencies mentioned in point (d) are generally viewed as problematic since in the ordinary kernel estimator a single bandwidth value usually controls the smoothing at all parts of the density. As pointed out by Simonoff (1996, p.51), this is a significant problem with the ordinary kernel estimator. No single bandwidth value can possibly be "optimal" since the mean squared error of $\hat{f}_h(x)$ at any point x varies with both $f(x)/h$ and $h^4[f''(x)]^2$ (see equation $(A.2)$). Thus, in order to reduce MSE, h would need to increase in regions where f is large (to reduce variance) and it would need to decrease in regions where f'' is large (to reduce bias). From a practical point

of view, when using a single bandwidth as a smoothing parameter, regions with high curvature (f'' large) tend to be oversmoothed, and regions where f'' is small (such as in the tails), tend to be undersmoothed.

A.4 Bandwidth selection

For a given data set, a practical way to choose a bandwidth might be to begin with a large bandwidth (oversmooth) and then gradually decrease the bandwidth until the "structural" fluctuation of the data is reduced and "random" fluctuations start to appear (Wand and Jones, (1995), p. 58). This strategy, of course, requires knowledge about the structure of the data, for example, the locations of modes. Moreover, such a trial-and-error approach can be time consuming when there are many densities to estimate. Therefore, a solution based on some automatic procedure of bandwidth selection method is required.

Revisiting the form of $h_{mse}(x)$ at (A.3), the lack of knowledge of the functions $f(\cdot)$ and $f''(\cdot)$ is a key problem in specifying an optimal bandwidth. If we consider the $ISE(\hat{f}_h)$ measure of discrepancy between $\hat{f}_h$ and f, we note that

$$\text{ISE}(\hat{f}_h) = \int (\hat{f}_h(x) - f(x))^2 dx = \int \hat{f}_h^2(x)dx - 2\int \hat{f}_h(x)f(x)dx + \int f^2(x)dx \ .$$

Since the last term of the above equation does not depend on $\hat{f}_h$, the optimal bandwidth h corresponds to the choice that minimizes

$$\int \hat{f}_h^2(x)dx - 2\int \hat{f}_h(x)f(x)dx \ . \tag{A.4}$$

Concerning the last term of equation (A.4), we note that

$$\int \hat{f}_h(x)f(x)dx = E_X(\hat{f}_h(X)) \ .$$

We may estimate $E_X(\hat{f}_h(X))$ with the help of the *leave-one-out* estimates of $\hat{f}$ which are defined as

$$\hat{f}_{i,h}(x) = (n-1)^{-1}h^{-1}\sum_{j\neq i} K\{h^{-1}(x - X_j)\} \ ;$$

the estimate is given by

$$\widehat{E_X(\hat{f}_h(X))} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{i,h}(X_i) \ .$$

Next, the cross-validation criterion is defined to be

$$CV(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{i,h}(X_i), \tag{A.5}$$

and it has an associated optimal bandwidth value given by

$$h_{cv} = \arg\min_h CV(h) \ \ .$$

Note that the above form of cross-validation function, $CV(h)$, is an unbiased estimator of $MISE(\hat{f}_h) - ||f||_2^2$. We can see this since

$$\begin{aligned} E(CV(h)) &= E\left(\int \hat{f}_h^2(x)dx\right) - \frac{2}{n}\sum_{i=1}^{n} E(\hat{f}_{i,h}(X_i)) \\ &= MISE(\hat{f}_h) + 2E_X(\hat{f}_h(X)) - ||f||_2^2 - \frac{2}{n}E\left(\sum_{i=1}^{n} \hat{f}_{i,h}(X_i)\right) \\ &= MISE(\hat{f}_h) - ||f||_2^2 \end{aligned}$$

where we have used that $E_X(\hat{f}_h(X)) = \frac{1}{n}E\left(\sum_{i=1}^{n} \hat{f}_{i,h}(X_i)\right)$.

The cross-validation bandwidth h_{cv} is random since it depends on the values of the given sample. On the other hand, if we consider a global measure of discrepancy of $\hat{f}_h$, such as $MISE(\hat{f}_h)$, we have that

$$\begin{aligned} MISE(\hat{f}_h) &= E(ISE(\hat{f}_h)) \\ &= \int_{-\infty}^{\infty} E\{\hat{f}_h(x) - f(x)\}^2 dx \\ &= \int_{-\infty}^{\infty} \text{bias}^2(\hat{f}_h(x))dx + \int_{-\infty}^{\infty} \text{var}(\hat{f}_h(x))dx \ . \end{aligned}$$

Asymptotically, as $h \to 0$ and $nh \to \infty$ the *MISE* has the form

$$MISE(\hat{f}_h) = \frac{1}{nh}||K||_2^2 + \frac{h^4}{4}||f''||_2^2\mu_2^2(K)$$

$$+o((nh)^{-1}) + o(h^4) \ .$$

Ignoring the higher order terms, we obtain the so-called asymptotic-MISE, $(AMISE)$, defined as just

$$AMISE(\hat{f}_h) = \frac{1}{nh}||K||_2^2 + \frac{h^4}{4}||f''||_2^2\mu_2^2(K) \ . \tag{A.6}$$

The optimal bandwidth associated with $AMISE$ has the form

$$h_{amise} = \left(\frac{||K||_2^2}{||f''||_2^2\mu_2^2(K)n}\right)^{1/5} . \tag{A.7}$$

We will also let h_{mise} denote the optimal bandwidth associated with $MISE(\hat{f}_h)$.

Compared to h_{cv}, the above bandwidth is not random. Comparing it to the optimal bandwidth $h_{mse}(x)$ defined at $(A.3)$, the above optimal bandwidth does not require knowledge of the underlying density f. However, both optimal bandwidths (h_{mse} and h_{amise}), require knowledge of f''. One way to get around this problem would be to directly estimate $||f''||_2^2$ (this approach is called the *plug-in method*). However, if we were to use $||\hat{f}_h''||_2^2$ as an estimate of $||f''||_2^2$ where

$$||\hat{f}_h''||_2^2 = \frac{1}{n^2h^5}\sum_{i=1}^{n}\sum_{j=1}^{n} K'' * K''\left(\frac{X_j - X_i}{h}\right), \tag{A.8}$$

and where $*$ is convolution, then we encounter the following problem. From (A.7), the optimal bandwidth h_{amise} is proportional to $n^{-1/5}$. Scott and Terrell (1987) proved that if the kernel K and the density function f are 4-times continuously differentiable then

$$E(||\hat{f}_h''||_2^2) = ||f''||_2^2 + \frac{1}{nh^5}||K''||_2^2 + O(h^2).$$

Therefore, for a bandwidth choice of order $h \sim n^{-1/5}$, $||\hat{f}''_h||_2^2$ is asymptotically a positively biased estimate of $||f''||_2^2$. An improved estimate of $||f''||_2^2$ should therefore be

$$||\widehat{f''}||_2^2 = ||\hat{f}''_h||_2^2 - \frac{1}{nh^5}||K''||_2^2 = \frac{1}{n^2h^5}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} K'' * K''\left(\frac{X_j - X_i}{h}\right) \tag{A.9}$$

where we obtain the last equality by noting that

$$K'' * K''(0) = \int K''(0-y)K''(y)dy = \int (K''(y))^2 dy = ||K''||_2^2 \,.$$

Plugging the estimate in (A.9) into (A.6), we obtain the quantity $BCV(\hat{f}_h)$ (biased cross-validation criterion) defined as

$$BCV(h) = \frac{1}{nh}||K||_2^2 + \frac{\mu_2^2(K)}{4}\left[\frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} K'' * K''\left(\frac{X_j - X_i}{h}\right)\right] \,.$$

(The above estimate is called biased cross-validation because asymptotically, $E(BCV(h)) \neq MISE(\hat{f}_h) - ||f||_2^2$.) We denote the optimal bandwidth associated with $BCV(h)$ by h_{bcv}.

Regarding the various bandwidths choices introduced thus far, an important question is that of the most appropriate bandwidth to choose. Let h_{mise} and h_{ise} be the bandwidths which minimize $MISE(\hat{f}_h)$ and $ISE(\hat{f}_h)$ respectively. Since all of the various bandwidths are, in one way or another, based on minimizing either $MISE(\hat{f}_h)$ or $ISE(\hat{f}_h)$, if we use h_{mise} as the bandwidth of choice, we are in a situation where we are trying to minimize a quantity that is being *averaged over all possible samples* while on the other hand, h_{ise} is minimizing a quantity for the *sample at hand* (Wand and Jones, 1995, p. 80). From a conceptual point of view, using h_{ise} as the theoretical standard is thus attractive, but unfortunately, targeting h_{ise} is difficult. In particular, it is a fact that the relative rate of convergence of any data-dependent $\hat{h}$ bandwidth to h_{ise} cannot be faster than $n^{-1/10}$; that is

$$\frac{\hat{h}}{h_{ise}} = 1 + O_p(n^{-1/10}) \tag{A.10}$$

as was shown by Hall and Marron (1987). This means that the relative rate of convergence of h_{cv} and h_{bvc} to h_{ise} is quite slow. The above rate of convergence also holds if $\hat{h}$ is taken to be h_{mise}.

The relatively slow rate in (A.10) is the reason we usually take h_{mise} for theoretical comparison to any other choice of bandwidth. Unfortunately, the relative rates of convergence of both h_{cv} and h_{bcv} to h_{mise} achieve the very slow rate of $1+O_p(n^{-1/10})$.

A better bandwidth selection method that is used more widely is that of Sheather and Jones (1991). This bandwidth, which we denote by h_{sj}, achieves the rate

$$\frac{h_{sj}}{h_{mise}} = 1 + O_p(n^{-5/14}) \; .$$

Below, we explain how h_{sj} is obtained, following the exposition on bandwidth selection from Wand and Jones (1995).

The aim is to estimate h_{amise} as defined in $(A.7)$. To do so, and as noted earlier, we need to find a proper estimate of $||f''||_2^2$. Begin by noting that

$$||f^{(s)}||_2^2 = \int f^{(s)}(x)^2 dx$$

and using integration by parts,

$$\begin{aligned} ||f^{(s)}||_2^2 &= (-1)^s \int f^{(2s)}(x) f(x) dx \\ &= \int f^{(r)}(x) f(x) dx \; . \end{aligned}$$

if $r = 2s$. Define ψ_r to be

$$\psi_r = \int f^{(r)}(x) f(x) dx = E(f^{(r)}(X)) \; .$$

This motivates the estimation

$$\hat{\psi}_r(g) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_g^{(r)}(X_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_g^{(r)}(X_i - X_j)$$

where $\hat{f}_g^{(r)}(x)$ is the kernel density estimation of $f^{(r)}(x)$. Sheather and Jones show that the optimal bandwidth associated with the asymptotic MSE($\hat{\psi}_r(g)$) is

$$g_{amse,r} = \left(\frac{2K^{(r)}(0)}{-\mu_2(K)\psi_{r+2}\, n} \right)^{1/(3+r)} . \tag{A.11}$$

We can now rewrite h_{amise} defined at $(A.7)$, as

$$h_{amise} = \left(\frac{||K||_2^2}{\mu_2^2(K)\psi_4\, n} \right)^{1/5} \tag{A.12}$$

and its estimate is given by

$$\hat{h}_{amise} = \left(\frac{||K||_2^2}{\mu_2^2(K)\hat{\psi}_4(g_{amse,4})\, n} \right)^{1/5}$$

where

$$g_{amse,4} = \left(\frac{2K^{(4)}(0)}{-\mu_2(K)\psi_6\, n} \right)^{1/7} . \tag{A.13}$$

We could go on to estimate ψ_6 but as we can see, its optimal bandwidth depends on ψ_8, and as apparent in $(A.11)$, the general case is that the optimal bandwidth for estimating ψ_r depends on ψ_{r+2}.

From $(A.12)$ and $(A.13)$ we note the relationship

$$g_{amse,4} = \left[\frac{2K^{(4)}(0)\mu_2(K)^2}{||K||_2^2\mu_2(K)} \right]^{1/7} \left(-\frac{\psi_4}{\psi_6} \right)^{1/7} h_{amise}^{5/7} .$$

If we define

$$\gamma(h) = \left[\frac{2K^{(4)}(0)\mu_2(K)^2}{||K||_2^2\mu_2(K)} \right]^{1/7} \left\{ -\frac{\hat{\psi}_4(g_{amse,4})}{\hat{\psi}_6(g_{amse,6})} \right\}^{1/7} h^{5/7} ,$$

we are led to another "stage selection" problem since the estimates of $g_{amse,4}$ and $g_{amse,6}$ are themselves based on ψ_6 and ψ_8, respectively. At this stage we could choose to estimate ψ_6 and ψ_8 assuming that the design density f is the density of the

normal distribution where the mean and the variance maybe be estimated from the data. Letting $\hat{\psi}_6^{NS}$ and $\hat{\psi}_8^{NS}$ be such estimates of ψ_6 and ψ_8, respectively, we obtain

$$\hat{g}_{amse,4} = \left(\frac{2K^{(4)}(0)}{-\mu_2(K)\hat{\psi}_6^{NS}\, n} \right)^{1/7}$$

and

$$\hat{g}_{amse,6} = \left(\frac{2K^{(6)}(0)}{-\mu_2(K)\hat{\psi}_8^{NS}\, n} \right)^{1/9}.$$

Letting

$$\hat{\gamma}(h) = \left[\frac{2K^{(4)}(0)\mu_2(K)^2}{||K||_2^2 \mu_2(K)} \right]^{1/7} \left\{ -\frac{\hat{\psi}_4(\hat{g}_{amse,4})}{\hat{\psi}_6(\hat{g}_{amse,6})} \right\}^{1/7} h^{5/7},$$

the selected bandwidth is the solution to the equation

$$h = \left(\frac{||K||_2^2}{\mu_2^2(K)\hat{\psi}_4(\hat{\gamma}(h))\, n} \right)^{1/5}.$$

A.5 Further properties of the kernel density estimator

Plugging h_{amise} into equation $(A.6)$, we obtain

$$AMISE(\hat{f}_{h_{amise}}) = \frac{5}{4}(||K||_2^2)^{4/5}(||f''||_2^2 \mu_2^2(K))^{1/5} n^{-4/5}.$$

A natural question to then ask would be what the effect of the kernel function K is on the value of $AMISE(\hat{f}_{h_{amise}})$. Note that the term which measures the roughness of the underlying density, $||f''||_2^2$, is not under our control, whereas the term $(||K||_2^2)^{4/5}(\mu_2^2(K))^{1/5}$ is a function of the kernel function only, and as such, is the only term which can be controlled. If we consider a class of kernels, such that the kernels K in this class are restricted to be non-negative, with

$$\int K(u)du = 1, \quad \int uK(u)du = 0, \text{ and } \int u^2 K(u)du = a^2 < \infty,$$

then the class of kernels which minimize $C(K) = (||K||_2^2)^{4/5}(\mu_2^2(K))^{1/5}$ may be shown to be of the form

$$K^a(u) = \frac{3}{4}\frac{1}{5^{1/2}a}\left(1 - \frac{u^2}{5a^2}\right) I(|u| \leq 5^{1/2}a)$$

(Wand and Jones (1995, p. 30)). Since a is an arbitrary scale parameter, the simplest version of K^a corresponds to $a^2 = 1/5$. This leads to the kernel

$$K^*(u) = \frac{3}{4}\left(1 - u^2\right) I(|u| \leq 1)$$

which is called the *Epanechnikov kernel*. For the Epanechnikov kernel, the value of $C(K^*)^{5/4}$ is $3/(5\sqrt{5})$ (Simonoff (1994, p. 44)). To study the relative inefficiency of using other kernels, we look at the ratio of $(C(K)/C(K^*))^{5/4}$. Table A.1, adopted from Simonoff (1996, p. 44), lists this ratio for several commonly used kernels. All

Table A.1: Inefficiency of various kernels relative to the Epanechnikov kernel.

Kernel	Form	Inefficiency
Epanechnikov	$\frac{3}{4}(1-u^2)$	1
Biweight	$\frac{15}{16}(1-u^2)^2$	1.0061
Triweight	$\frac{35}{32}(1-u^2)^3$	1.0135
Gaussian	$(2\pi)^{-1/2}e^{-u^2/2}$	1.0513
Uniform	$\frac{1}{2}$	1.0758

Adopted from Simonoff (1996), p. 44

kernels except the Gaussian is zero outside of the interval $[-1, 1,]$. As is evident, the choice of kernel makes very little difference in the value of $AMISE$, and therefore in choosing a kernel, our choice may be based on other considerations, such as ease of computation or properties of $\hat{f}$ (Simonoff (1994, p. 44).

Appendix B

Computer Code Description

B.1 Description of the codes for computing document date estimates in Chapter 5

In all of the computer programs described below, the computations were carried out on a number of UNIX platforms (SGI Challenge L, running IRIX Version 6; Sun 280R, running Solaris 9; and Dell Power PC, running Solaris86 Version 8) employing the C programming language and standard UNIX commands such as 'sort', 'uniq', 'diff' 'cut' and 'wc'. Four separate programs were used to produce the date estimates. Below we describe each one separately.

(1) Description of the programming code *cleanandcut.c*

The first program, called *cleanandcut.c*, takes as an input the original file called *rawdatafile.txt* containing the DEEDS documents. It removes from the documents characters such as "=" and "=20" which indicate carriage returns left by the original editing programs. The program also sequentially assigns a number, ranging from 1 to 3353, to each of the documents. Each document is then separately written into a file

with a filename corresponding to the document number. Finally another, separate, file called *masterfile* is constructed which contains a list of all 3353 documents along with the true dates in which they were written.

(2) Description of the programming code *shingler.c*

A second program, called *shingler.c*, breaks up each of the documents (i.e. the outputs of *cleanandcut.c*) into shingles of a desired shingle order. For the purpose of this thesis, the documents were shingled into orders of 1, 2 and 3. As an example, we display outputs of the programs *cleanandcut.c* and *shingler.c* for the following document from the DEEDS data set:

"00640217","1238",

/

Memorandum quod cum inter dominum Robertum Sarum episcopum ex parte una et Henricum abbatem et conventum Scyreburne ex altera super =
amerciamentis provenientibus de pane et cerevisia venditis contra =20
assisam de hominibus praedicti abbatis infra libertates praedicti =
episcopi existentibus in hundredis de Syreburne et de Bemenistre =
discordia fuisset suscitata asserente episcopo praefato hujusmodi =20
amerciamenta ad jus suum pertinere et se et praedecessores suos in eorum =
seisina diu et in pace extitisse praedicto abbate penitus contrarium =
asserente tandem mediantibus discretis viris domino =20
Roberto decano et capitulo Sarum inter partes hujusmodi amicabilis =
compositio intercessit: videlicet quod praenominato episcopo pro se et =
successoribus suis de assensu capituli Sarum omnis =20
amerciamenta de pane et cerevisia contra assisam venditis et de =
cerevisia male braciata de omnibus hominibus in terris et feudis =
praedicti abbatis infra hundreda praedicta existentibus jus petitionem =20

et saisinam si quam habuit dicto Henrico abbati et successoribus suis =
futuris et perpetuis temporibus remittente et quieta clamante idem abbas =
promisit et se et successores suos per cartam suam et =20
conventus sui obligavit quod pro hac remissione et quieta clamatione =
praedicto episcopo et successoribus suis dimidiam marcam annuam ad =
Pascha in perpetuum persolvet apud Syreburne Et ut haec =20
compositio futuris et perpetuis temporibus rata et stabilis perseveret =
confecta sunt inde inter partes duo scripta quorum unum signatum =
sigillis praedictorum episcopi et capituli penes praedictos =20
abbatem et conventum in perpetuum remanebit et aliud signatum sigillo =
abbatis et conventus penes dictum episcopum et capitulum in perpetuum =
remanebit His testibus dominis Roberto decano Rogero =20
praecentore Ada cancellario Johanne thesaurario Sarum Egidio =
archidiacono Berksyre Stephano archidiacono Wilthesyre Hugone abbate =
Abbedesbire Henrico de Sancto Edmund Willelmo de Cambe Ricardo de =20
Cnolle Petro de Cumbe et aliis Datum apud Syreburne per manus Henrici =
Isumberd [al Ysembard] monachi !xvi! Kalend Septembris anno gratiae !m =
cc xxxviii! =20

/

The "/" indicates the separation between documents, and the identification and date of the document are located at the top left hand-side (for example, the above document is dated 1238). When the entire corpus containing all of the DEEDS documents is entered into *cleanandcut.c*, the particular document shown above, which is located at the 3rd position of the DEEDS data set, is assigned the file name *'file0003'*. The contents of *file0003* reads:

Memorandum

quod

cum

inter

dominum

Robertum

Sarum

episcopum

⋮

⋮

gratiae

!m

cc

xxxviii!

The program *shingler.c* takes as input the desired value k of the shingle order and transforms the contents of each output file of *cleanandcut.c* into a list of k-order shingles, listing these shingles in sorted order where repetitions of shingles are allowed should they occur. For $k = 3$ for example, the new file *file0003.shingled* reads as follows:

!m cc xxxviii!

!xvi! Kalend Septembris

Abbedesbire Henrico de

Ada cancellario Johanne

Bemenistre discordia fuisset

Berksyre Stephano archidiacono

Cambe Ricardo de

Cnolle Petro de

⋮

⋮

venditis contra assisam

venditis et de

videlicet quod praenominato

viris domino Roberto

(3) Description of the programming codes *corr-sh1.c*, *corr-sh2.c* and *corr-sh3.c*.

A third program – actually a set of programs – computes the type (II) ($\alpha = 1$) correspondence distance measures between all the pairs of documents in the DEEDS data set. Specifically, for each shingle order (1, 2 and 3), we have separate programs called *corr-sh1.c*, *corr-sh2.c* and *corr-sh3.c*. The results of these programs are stored in 3 directories each of which contains 3353 files containing the correspondence measures between that given document and all the other documents in the DEEDS data set.

(4) Description of the programming codes *surround-optimize-test1.c*, *surround-optimize-test12.c* and *surround-optimize-test123.c*, and *surround-optimize-val1.c*, *surround-optimize-val12.c* and *surround-optimize-val123.c*,

The fourth and the final set of programs use the distance measures produced by the third set of programs to compute the date estimates for documents in the test set based on equation (5.2). Two separate programs are used and are called *surround-optimize-test1.c*, *surround-optimize-test12.c* and *surround-optimize-test123.c*. The first of these computes date estimates based on a single shingle order, the second program computes date estimates based on a two separate shingle orders and the third program computes date estimates based on a three separate shingle orders. The optimal bandwidths associated with the distance weights in (5.1) were found

using the cross-validation procedure described in Section 5.2, equation (5.3). The search for the optimal bandwidths for each document in the test set was carried out over a k-dimensional grid where k is the shingle order. Analogously, codes *surround-optimize-val1.c*, *surround-optimize-val12.c* and *surround-optimize-val123.c* compute the date estimates for documents in the validation set.

B.2 Description of the codes for computing document date estimates in Chapter 6

We used a total of eight sets of computer programs written in the C language and various UNIX system functions to produce the results of Section 6.5. The first two programs called *cleanandcut.c* and *shingler.c* were described in Appendix B.1. We will now describe the remaining six sets of programs.

(1) Description of the programming codes *merger1.c*, *merger2.c*, *merger3.c* and *merger4.c*.

Each of the elements of a third set of computer programs, *merger1.c*, *merger2.c*, *merger3.c* and *merger4.c* merges the k-shingle order documents from the training set, for $k = 1, 2, 3$ and 4 respectively, into a file called *mergerfile*. The file contains sequences of shingles, and juxtaposed with each shingle, we indicate the date of the document in which the shingle is found. The lines of the file *mergerfile* are not necessarily distinct, so that if a shingle is repeated more than once in the same document, the shingle, along with the date of the document in which it occurred, will appear that same number of time in the file *mergerfile*.

Various UNIX commands, such as *cut, sort* and *uniq*, were also used to produce files, such as, *mergershing1sortuniq*, *mergershing2sortuniq*, *mergershing3sortuniq* and

mergershing4sortuniq. These files are sorted by shingles, and each line of the file contains the number of times that a shingle occurred, the dates in which it occurred, and the number of times it occurred within that date. The shingles are all derived from the shingled documents of the training set. As an example, below we show some of the contents of *mergershing2sortuniq*:

1 !1!@acra 1309

1 !1!@acram 1309

1 !1!@bussellam 1309

1 !1!@mesuagium 1309

1 !1!@repastum 1309

2 !1!@vomer 1309

⋮

1 zelo@iusticie 1264

1 zelum@deuocionis 1281

1 zinzeberis@ad 1294

1 zinziberis@ad 1274

1 zonam@de 1282

1 zonam@sericam 1292

In addition, we also created files called *mergerdate1sortuniq*, *mergerdate2sortuniq*, *mergerdate3sortuniq* and *mergerdate4sortuniq*, for each shingle order $k = 1, 2, 3$ and 4 respectively. These files contain the total number of non-distinct shingles for each of the training document dates. We also created files named *mergershing1onlysortuniq* , *mergershing2onlysortuniq* , *mergershing3onlysortuniq* and *mergershing4onlysortuniq*. Each of these files contain a list of distinct shingles derived from the shingled documents of the training set for shingle orders 1,2,3 and 4, respectively. As an example below is some of the content of the file *mergershing2onlysortuniq*:

!1!@acra

!1!@acram

!1!@bussellam

!1!@mesuagium

!1!@repastum

!1!@vomer

⋮

zelo@iusticie

zelum@deuocionis

zinzeberis@ad

zinziberis@ad

zonam@de

zonam@sericam

(2) Description of the programming code *dateshing.c*.

The fourth set of computer program contains *dateshing.c*. This program takes as an input, for example, the file *mergershing1sortuniq* and outputs the number of times each of the shingles in *mergershing1onlysortuniq* occurs for each consecutive date starting from the minimum date (1089) to the maximum date (1466) of the training documents. The output is stored in a file called *dateshingresultfile*. Below is an example of what this might look like for the shingle *baptiste*.

baptiste 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 2 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 0 2 0
0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 2 4 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

(3) Description of the programming code *datetotalshing.c*.

The fifth set of computer programs contain *datetotalshing.c*. This program takes as an input, for example, the file *mergerdate1sortuniq*, and outputs a file called *datetotal*. This file contains the total number of shingles for each consecutive year starting from 1089 to 1466.

(4) Description of the programming codes *dateshing1prob.c*, *dateshing2prob.c*, *dateshing3prob.c* and *dateshing4prob.c*.

The sixth set of computer programs containing *dateshing1prob.c*, *dateshing2prob.c*, *dateshing3prob.c* and *dateshing4prob.c* evaluates the Nadaraya-Watson type estimator, $\hat{\pi}_s(t)$ (see (6.5)), for each shingle s at time t ranging consecutively from 1089 to 1466, and for the shingle order $k = 1, 2, 3$ and 4, respectively. The inputs are *dateshingresultfile*, *datetotal* and a bandwidth value h. The result is stored in a directory containing files and each of these files correspond to a shingle. Each file contains $\hat{\pi}_s(t)$ for each time t for the corresponding shingle. An example of the content of such a file, *freq22156*, corresponding to the shingle *baptiste* is given below:

0.000000016137 0.000000017163 0.000000018340 0.000000019683 0.000000021208 0.000000022929

0.000000024864 0.000000027029 0.000000029439 0.000000032110 0.000000035057 0.000000038294

0.000000041836 0.000000045695 0.000000049884 0.000000054415 0.000000059300 0.000000064551

...

0.000000002795 0.000000002600 0.000000002407 0.000000002220 0.000000002039 0.000000001867

0.000000001705 0.000000001555 0.000000001418 0.000000001293 0.000000001182 0.000000001082

0.000000000995 0.000000000918 0.000000000852 0.000000000794 0.000000000745 0.000000000703

(5) Description of the programming codes *datevalshing1.c*, *datevalshing2.c*, *datevalshing3.c* and *datevalshing4.c*.

The seventh set of computer programs containing *datevalshing1.c*, *datevalshing2.c*, *datevalshing3.c* and *datevalshing4.c* takes as an input *mergershing1onlysortuniq*, *mergershing2onlysortuniq*, *mergershing3onlysortuniq* and *mergershing4onlysortuniq* respectively, and the file called *masterfile* (for a description of this file, see the description of the programming code *cleanandcut.c*). The outputs, *dateshing1valprob*, *dateshing2valprob*, *dateshing3valprob* and *dateshing4valprob* each contain a list of the estimated dates of the validation documents based on shingle order $k = 1, 2, 3$ and 4 respectively. The error in the dating of the validation documents can be computed by comparing the estimated dates to the true dates for a given bandwidth. The bandwidths (for each shingle order we have a separate bandwidth) that produce the minimum MSE between the estimated dates and the true dates (we will call these bandwidths the optimal bandwidths) are then used for computing the estimated dates of the documents in the test set.

(6) Description of the programming codes *datetestshing1.c*, *datetestshing2.c*, *datetestshing3.c* and *datetestshing4.c*.

Based on the optimal bandwidths, the eighth set of computer programs containing *datetestshing1.c*, *datetestshing2.c*, *datetestshing3.c* and *datetestshing4.c* corresponding to shingle orders 1, 2, 3 and 4 respectively, compute the date estimates of documents in the test set. The results, corresponding to each shingle order 1, 2, 3 and 4 are stored in four separate files. The computer programs described above are essentially the same as those of the seventh set of computer programs, except that the documents are now estimating those of the test set. It is these date estimates of the validation set and the test set that are used in computing the results that are presented in Table 6.1 and Table 6.2, respectively.

Bibliography

[1] Adams, R.A (1987). *Calculus of Several Variables.* Addison-Wesely Publishers Limited.

[2] Altschul, S.F (2003). BLAST algorithm. In:*Nature Encyclopedia of the Human Genome (D.N. Cooper, ed.)*. Nature Publishing Group, London, UK, **1**, pp. 328-331.

[3] Berry, M.W. and Browne, M. (2005). *Understanding search Engines – Mathematical Modeling and Text Retrieval, 2nd edition.* Society for Industrial and Applied Mathematics, Philadelphia.

[4] Broder, A.Z (1998). On the resemblance and containment of documents. In: *International Conference on Compression and Complexity of Sequences (SEQUENCES '97)*, June 11-13 1997, Positano, Italy, pp. 21-29. IEEE Computer Society, Los Alamitos, California.

[5] Charniak, E. (1993). *Statistical Language Learning*. MIT Press, Cambridge, Massachusetts.

[6] Chen, S.F. and Goodman, J. (1998). An emperical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University.

[7] Djeraba, C. (2003). *Multimedia Mining – A Highway to Intelligent Multimedia Documents*. Kluwer, Boston.

[8] Domingos, P. and Pazzani, M. (1996). Beyond independence: Conditions for optimality of the Bayes classifier. In: *Proceedings of the 13th International Conference on Machine Learning*, pp. 105-112.

[9] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman and Hall, London.

[10] Fan, J. and Gijbels, I. (2000). Local polynomial fitting. In: Schimek, M.G., *Smoothing and Regression: Approaches, Computation, and Application*. Wiley, New York, pp. 229-276.

[11] Fan, J., Heckman, N.E and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, **90**(429), pp. 141-150.

[12] Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates: Sunderland, MA.

[13] Feuerverger, A., Hall, P. , Tilahun, G. and Gervers, M. (2004). Measuring distance, and smoothing, among medieval manuscripts and pages of the World Wide Web. Available online at http://www.amstat.org/publications/jcgs/index and follow the supplemental materials link.

[14] Feuerverger, A., Hall, P. , Tilahun, G. and Gervers, M. (2005). Distance measures and smoothing methodology for imputing features of documents. *Journal of Computational and Graphical Statistics*, **14**(2), pp. 255-262.

[15] Feuerverger, A., Hall, P. , Tilahun, G. and Gervers, M. (2008). Using statistical smoothing to date medieval manuscripts. Beyond Parametrics in Interdisciplinary Research: Festschrift in Honour of P.K. Sen; N. Balakrishnan, E. Pena, M.J. Silvapulle, editors. IMS Collections, **1**, pp. 321-331.

[16] Fiallos, R. (2000). An overview of the process of dating undated medieval charters: Latest results and future developments. In: Gervers, M., *Dating Undated Medieval Charters*. Boydell Press, Woodbridge.

[17] Forman, G. (2006) Tackling Concept drift by temporal inductive transfer. In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 252-259.

[18] Gasser, T. and Müller, H.-G (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics, Springer-Verlag, New York, **757**, pp. 23-68.

[19] Gervers, M. (1998). A description of the process in which sample charter 10662 (actually dates 1274) is dated by comparing its text with the text of 1,524 dated charters in the DEEDS datebase. Supporting Document, DEEDS Project, University of Toronto.

`http://scriptor.deeds.utoronto.ca:7777/d_mech/files/sshrcc_3.pdf`

[20] Gervers, M. (2000). *Dating Undated Medieval Charters*. Boydell Press, Woodbridge.

[21] Gervers, M. and Hamonic, N. (2010). Pro Amore Dei: Diplomatic Evidence of Social Conflict During the Reign of King John. *Preprint*

[22] Grossman, D.A. and Frieder, O. (1998). *Information Retrieval: Algorithms and Heuristics*. Kluwer, Boston, Mass.

[23] Grossman, D.A. and Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics, 2nd. edition*. Kluwer, Boston, Mass.

[24] Hall, P. and Marron, J.S. (1987). On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Annals of Statistics*, **15**, pp. 163-181.

[25] Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

[26] Härdle, W. (1991). *Smoothing Techniques: With Implementation in S*. Springer-Verlag, New York.

[27] Härdle, W and Marron, J.S (1986). Random approximations to an error criterion of nonparametric statistics *Journal of Multivariate Analysis*, **20**, pp. 91-113.

[28] Heckman, Nancy. Lecture Notes, 1998-1999. *http://ugrad.stat.ubc.ca/∼nancy/526/handouts/*

[29] Jelinek, F. and Mercer, R. (1980). Interpolated estimation of Markov source parameters from sparse data. *Proceedings of the Workshop on Pattern Recognition in Practice*. Amsterdam, The Netherlands: North-Holland, May.

[30] Karlin, S. and Altschul, S.F (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, pp. 2264-68.

[31] Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, **18**, (1), pp. 191-219.

[32] Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, **1**, pp. 181-184.

[33] Kondrak, G. (2002). Algorithms for language reconstruction. Ph.D thesis, Graduate Department of Computer Science, University of Toronto.

[34] Lipman, D.J., Altschul, S.F., Kececioglu, J.D. (1989). A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, pp. 4412 - 4415.

[35] Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, pp. 255-268.

[36] Lin, L.I. (2000). A note on the concordance correlation coefficient. *Biometrics*, **56**, pp. 324-325.

[37] Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, **2**, pp. 159-165.

[38] Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

[39] Margulis, E.L. (1992). N-Poisson document modelling. *Association for Computing Machinery, SIGIR '92*, pp. 177-189.

[40] McGill, M., Koll, M., & Noreault, T. (1979). An evaluation of factors affecting document ranking by information retrieval systems. Technical Report. Syracuse, NY: School of nformation Studies, Syracuse University.

[41] Mosteller, F and Wallace, D (1963). Inference in an authorship problem. *Journal of the American Statistical Society*, **58**(302), pp. 275-302.

[42] Nadaraya, E.A. (1964) On estimating regression. *Theory of Probability and Its Applications*, **10**, pp.186-190.

[43] Parzen, E. (1962) On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, pp.1065-1076.

[44] Priestley, M.B. and Chao, M.T. (1972). Nonparametric function fitting. *Journal of the Royal Statistical Society, B*, **34**, pp. 384-392.

[45] van Rijsbergen, C.J. (1979) *Information Retrieval, 2nd. edition.* http://www.dcs.gla.ac.uk/Keith/Preface.html.

[46] Rosenblatt, M. Remarks on some non-parametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, pp.642-669.

[47] Rosenfeld, R. (2000). Two decades of statistical language modelling: where do we go from here. *Proceedings of the IEEE*, **88**(8).

[48] Salton,G., Wang, A., and Yang, C. (1975). A vector space model for information retrieval. *Journal of the American Society for Information Science*, **18**, pp.613-620.

[49] Sarda, P. and Vieu, P. (2000). Kernel Regression. In: Schimek, M.G., *Smoothing and Regression: Approaches, Computation, and Application*. Wiley, New York, pp. 229-276.

[50] Schimek, M.G. (2000). *Smoothing and Regression: Approaches, Computation, and Application*. Wiley, New York.

[51] Sharan, U. and Neville, J. (2007). Exploiting time-varying relationships in statistical relational models. In Proceedings of the Joint 9th WebKDD and 1st SNA-KDD Workshops.

[52] Scott, D.W. and Terrell, G.R. (1987) Biased and unbaised cross-validation in density estimation. *Journal of the American Statistical Association*, **82**, pp. 1131-1146.

[53] Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Ser. B*, **53**, pp. 683-690.

[54] Silverman, B. W., (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

[55] Simonoff, Jeffrey S., (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New-York.

[56] Stehling, R.O., Nascimento, M.A., Falcão, A.X. (2003). Techniques for color-based image retrieval. In: *Multimedia Mining – A Highway to Intelligent Multimedia Documents, ed. Djeraba*. Kluwer, Boston. pp. 61-82.

[57] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall/CRC, Boca Raton, London, New York, Washington, D.C.

[58] Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Ser. A*, **26**, pp. 359-372.

[59] Witten, I.H and Bell, T. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theorey*, **37**(4), July. pp.1085-1094.

[60] Zhao, Y., Lebanon, G. and Zhao, Y. (2009). Local likelihood Modeling of the concept drift phenomenon. Technical Report FODAVA-09-21, Georgia Institute of Technology.

[61] Zhang, J. and Korfhagen, R. (1999). A distance and angle similarity measure method. *Journal of the American Society for information Science*, **50**(9), pp. 772-778.

[62] Zipf, H.P., (1949) *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, Massachusetts.