

# Local Likelihood Modeling of the Concept Drift Phenomenon

Yang Zhao  
Department of Statistics  
Purdue University

Guy Lebanon  
College of Computing  
Georgia Institute of Technology

Yanjun Zhao  
College of Computing  
Georgia Institute of Technology

September 26, 2009

## Abstract

Temporal text data is often generated by a time-changing process or distribution. Such a drift in the underlying distribution cannot be captured by stationary likelihood techniques. We consider the application of local likelihood methods to generative and conditional modeling of categorical temporal data such as time-stamped document sequences. The resulting model corresponds to a local  $n$ -gram model in the generative case and local logistic regression in the conditional case. We examine the asymptotic bias and variance of the local estimator and their implications to the optimal kernel bandwidth. We discuss various regularization schemes and demonstrate the proposed estimators using an experimental study on the Reuters RCV1 news data and AOL query logs.

## 1 Introduction

Time stamped documents such as news stories and emails often cannot be accurately modeled by a single time invariant distribution. An alternative is to assume that the concepts underlying the distribution generating the data drift with time. In other words, the data  $z^{(t_1)}, \dots, z^{(t_k)}$ , is generated by a time dependent process  $\{p_t : t \in [a, b] \subset \mathbb{R}\}$ ,  $z^{(t)} \sim p_t(z)$  whose approximation  $\{\hat{p}_t : t \in [a, b]\}$  becomes the main objective of the learning task.

We assume that the time  $t$  is a continuous quantity, even in cases where the realized times are discrete. For example, if the time stamps represent the days of the year when the documents were authored, we assume that the set  $\{1, \dots, 365\}$  is a discrete sample from a underlying continuous interval  $[1, 365]$ . We further assume that the data sampled from  $p_t$  correspond to pairs  $z^{(t)} = (x, y)$  constituting a document  $x$  and a categorical-valued label  $y$ . Such pairs  $(x, y)$  appear often in practice, for example with  $y$  corresponding to the document topic (Yang, 1999; Lewis et al., 2004), sentiment (Pang and Lee, 2004, 2005; Lebanon and Mao, 2008), author (Mosteller and Wallace, 1964; Airoidi et al., 2006) or Email spam/no-spam (Mulligan, 1999).

The drift  $p_t(x, y)$  can be characterized by considering the temporal transition of the joint distribution  $p_t(x, y)$ , the conditionals  $p_t(y|x)$ ,  $p_t(x|y)$ , or the marginals  $p_t(x)$ ,  $p_t(y)$ . The choice of which of the distributions above to model depends on the application at hand. For example, modeling  $p_t(y|x)$  is usually sufficient for document classification purposes while modeling  $p_t(x|y)$  is necessary

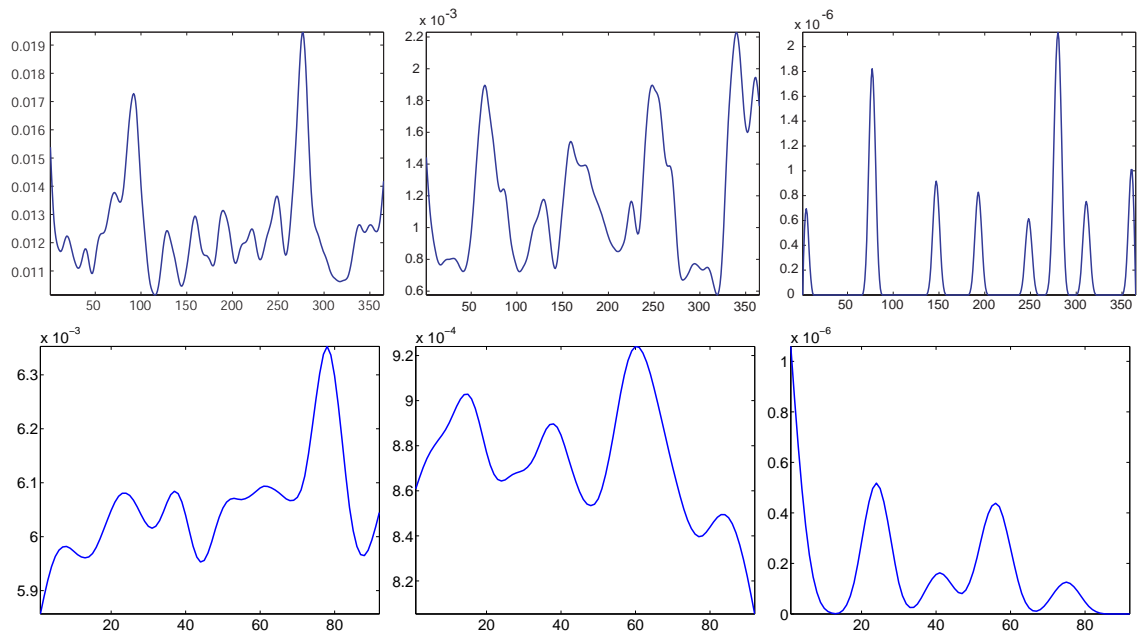


Figure 1: Estimated relative frequency (number of appearances in a document divided by document length) of words from the most popular category in RCV1 and AOL as a function of time. The upper three panels correspond to the words `million`, `common`, and `Handelsgesellschaft` (trade unions in German) in RCV1 dataset. The lower three panels correspond to the words `free`, `lottery`, and `evansville.net` in AOL dataset. The displayed curves were smoothed by convolving the counts with a smoothing kernel.

for language modeling which is an important component in speech recognition, machine translation, and information retrieval.

We demonstrate the presence of concept drift in practice by considering the Reuters RCV1 dataset (Lewis et al., 2004) and the AOL dataset (Pass et al., 2006). The Reuters RCV1 data contains news stories authored during a period of 365 consecutive days. The AOL dataset contains queries issued by AOL users during a period of three months. More details regarding these datasets are provided in the appendix.

The upper part of Figure 1 displays the temporal change in the relative frequency (number of appearance in a document divided by document length) of three words in RCV1 dataset: `million`, `common`, and `Handelsgesellschaft` (trade unions in German) for documents in the most popular RCV1 category titled `CCAT`. It is obvious from these plots that the relative frequency of these words vary substantially in time. For example, the word `Handelsgesellschaft` appear in 8 distinct time regions, representing time points in which German trade unions were featured in the Reuters news archive. The lower part of Figure 1 displays the temporal change in the relative frequency of three words: `free`, `lottery`, and `evansville.net` in AOL dataset. Similar to RCV1 dataset, the relative frequency of these words vary largely in time. However, due to the short length of web queries, even the most frequent words in the AOL data such as `free` are still sparse compared to RCV1 data.

The temporal variation in relative frequencies illustrated by Figure 1 corresponds to a drift in

the distribution generating the data. Since the drift is rather pronounced, standard estimation methods based on maximum likelihood are not likely to accurately model the data. In this paper, we consider instead estimating the entire drift model  $\{p_t(x, y) : t \in [a, b]\}$  based on the local likelihood principle. Local likelihood is a locally weighted version of the loglikelihood with the weights determined by the difference between the time points associated with the sampled data and the time at which inference takes place. It enjoys nice theoretical properties, in particular concavity if the underlying likelihood is concave and statistical consistency.

After presenting a more formal discussion of concept drift in Section 2 and the definition of local likelihood in Section 3 we turn to examine in detail the case of modeling  $p_t(x|y)$  with local likelihood for  $n$ -grams and modeling  $p_t(y|x)$  with local likelihood for logistic regression. In the case of 1-grams or the naive Bayes model, we provide a precise as well as asymptotic description of the bias and variance which generalize the asymptotic bias and variance of the Nadaraya-Watson local regression. This generalization illuminates key properties concerning the selection of weights and the difference between the online and offline scenarios. We also consider different regularization schemes for the local  $n$ -gram model that correspond to a Bayesian model. Experiments conducted on the RCV1 and AOL datasets demonstrate the local likelihood estimation in practice and contrast it with more standard non-local alternatives.

## 2 The Concept Drift Phenomenon and its Estimation

Formally, the concept drift phenomenon may be thought of as a smooth flow or transition of the joint distribution of a random vector. We will focus on the case of a joint distribution of a random vector  $X$  and a random variable  $Y$  representing predictor and response variables. We will also restrict our attention to temporal or one dimensional drifts.

**Definition 1.** *Let  $X$  and  $Y$  be two discrete random vectors taking values in  $\mathcal{X}$  and  $\mathcal{Y}$ . A smooth temporal drift of  $X, Y$  is a smooth mapping from an interval  $[a, b] \subset \mathbb{R}$  to a family of joint distributions*

$$t \mapsto p_t(X = x, Y = y). \quad (1)$$

By restricting ourselves to discrete random variables we can obtain a simple geometrical interpretation of concept drift. Denoting the simplex of all distributions over the set  $S$  by

$$\mathbb{P}_S \stackrel{\text{def}}{=} \left\{ r \in \mathbb{R}^{|S|} : \forall i r_i \geq 0, \sum_{i=1}^{|S|} r_i = 1 \right\} \quad (2)$$

we have that Definition 1 is equivalent to a smooth parameterized curve in the simplex  $\mathbb{P}_{\mathcal{X} \times \mathcal{Y}}$ .

The drift in the joint distribution can be decomposed in several ways. The first decomposition  $p_t(x, y) = p_t(x|y)p_t(y)$  is useful for generative modeling and the second decomposition  $p_t(x, y) = p_t(y|x)p_t(x)$  is useful for conditional modeling. In the generative case we will focus on modeling  $p_t(x|y)$  since modeling  $p_t(y)$  is typically an easier problem due to its lower dimensionality (it is often the case that  $|\mathcal{Y}| \ll |\mathcal{X}|$ ). In the case of conditional modeling, we focus on modeling  $p_t(y|x)$  and we ignore the drift in the marginal  $p_t(x)$  since it is irrelevant for discriminative tasks.

In both cases we assume that our data is a set of time-stamped labeled documents sampled from  $p_t(x, y)$  where the time points  $t$  are sampled from a distribution  $g(t)$ . If  $g$  is a continuous

density, the number of samples at time  $t$ , denoted by  $N_t$ , is no greater than 1 with probability 1. In practice, however, we allow  $N_t$  to be larger than 1 in order to account for the discretization of time. We thus have the following decomposition of data

$$D = D_{t_1} \cup \dots \cup D_{t_r} \quad \text{where } t_1, \dots, t_r \in [a, b] \quad \text{and} \quad (3)$$

$$D_{t_i} = \{(x^{(t_i, j)}, y^{(t_i, j)}) : j = 1, \dots, N_{t_i}\}.$$

The notation  $x^{(t_i, j)}, y^{(t_i, j)}$  refers to the  $j$ -individual document and label associated with time  $t_i$ .  $D_{t_i}$  refers to the collection of all document and labels associated with time  $t_i$ .

We display in Figure 2 the total number of words per day (left) and the total number of documents per day (right) for the RCV1 (top) and AOL (bottom) datasets. As is evident from the two right panels,  $g(t)$  is a highly non-uniform density corresponding to varying amount of news content and queries in different dates. Dividing the number of documents per day (left) by the number of words per day (right) we obtain a surprisingly little variation among the number of words per document. We thus conclude that documents tend to have similar lengths but the number of documents per day vary substantially.

It is easy to come up with two simple solutions to the problem of concept drift modeling. The first solution, which we call the extreme global model, is to simply ignore the temporal drift and use all of the samples in  $D$  regardless of their time stamp. This approach results in a single global model  $\hat{p}$  which serves as an estimate for the entire flow  $\{p_t, t \in [a, b]\}$  effectively modeling the concept drift as a degenerate curve equivalent to a stationary point in the simplex. The second simple alternative, which we call the extreme local model, is to model  $p_t$  using only data sampled from time  $t$  i.e.  $\{(x^{(t, j)}, y^{(t, j)}) : j = 1, \dots, N_t\}$ . This alternative decomposes the concept drift estimation into a sequence of disconnected estimation problems.

The extreme local model has the benefit that if the individual estimation problems are unbiased, the estimation of the concept drift is unbiased as well. The main drawback of this method is the high estimation variance resulting from the relatively small number of daily samples  $N_t$  used to estimate the individual models. Furthermore, assuming  $D$  is finite we can only estimate the drift in the finite number of time points appearing in the dataset  $D$  (since we have no training data for the remaining time points). On the other hand, the extreme global model enjoys low variance since it uses all data points to estimate  $p_t$ . Its main drawback is that it is almost always heavily biased due to the fact that samples from one distribution  $p_{t_1}$  are used to estimate a different distribution  $p_{t_2}$ .

It is a well known fact that the optimal solution in terms of minimizing the mean squared estimation error usually lies between the extreme local and extreme global models. An intermediate solution can trade-off increased bias for reduced variance and can significantly improve the estimation accuracy. Motivated by this principle, we employ local smoothing in forming a local version of the maximum likelihood principle which includes as special cases the two extreme models mentioned above. The intuition behind local smoothing in the present context is that due to the similarity between  $p_t$  and  $p_{t+\epsilon}$ , it makes sense to estimate  $p_t$  using samples from neighboring time points  $t + \epsilon$ . However, in contrast to the global model the contribution of points sampled from  $p_{t+\epsilon}$  towards estimating  $p_t$  should decrease as  $\epsilon$  increases.

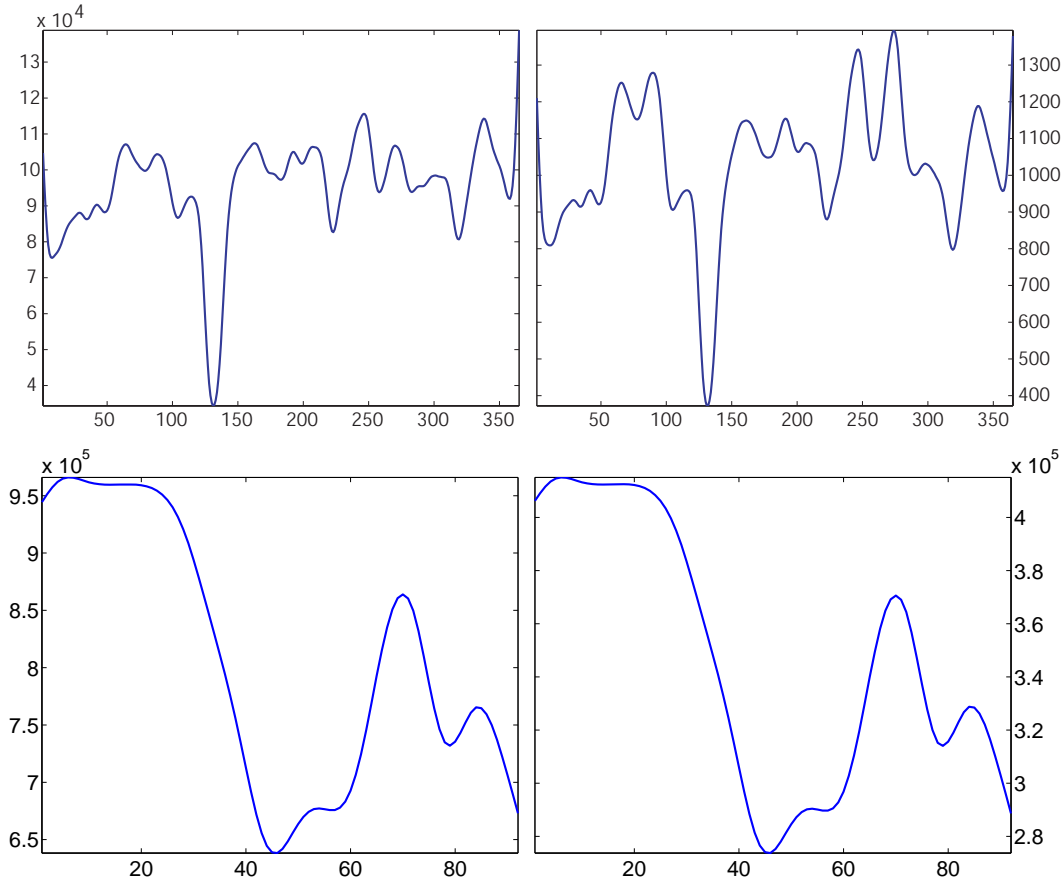


Figure 2: the total number of words per day (left) and the total number of documents per day (right) for the most popular RCV1 class (top) and AOL (bottom) datasets. As is evident from the two right panels,  $g(t)$  is a highly non-uniform density corresponding to varying amount of news content and queries in different dates. Dividing the number of documents per day (left) by the number of words per day (right) we obtain a surprisingly little variation among the number of words per document. We thus conclude that documents tend to have similar lengths but the number of documents per day vary substantially. The displayed curves were smoothed to remove sampling noise.

### 3 Local Likelihood and Concept Drift

The local likelihood principle extends the ideas of non-parametric regression smoothing and density estimation to likelihood-based inference. In Section 3.1 we concentrate on using the local likelihood principle for estimating  $p_t(x|y)$  and in Section 4 we focus on estimating  $p_t(y|x)$ .

#### 3.1 Local Likelihood for $n$ -Gram Estimation

Some of the most popular language modeling tools are  $n$ -Gram models which correspond to  $n$ -order Markov chain dependency. They are used to build language models for state-of-the-art information retrieval, speech recognition, and machine translation. An important component of such systems is the selection of appropriate mixtures and backing off policy for different Markov orders. For example, an interpolated tri-gram models could be  $p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x) + \alpha_3 p_3(x)$  where  $p_1, p_2, p_3$  correspond to  $n$ -gram models for  $n = 1, 2, 3$  respectively. We concentrate, however, on demonstrating local likelihood for a single  $n$ -gram with the understanding that appropriate mixtures or backing off policies can be used to improve the language modeling performance. For ease of notation we use the case  $n = 1$ , which is referred to as unigram or naive Bayes. Extending the discussion to  $n > 1$  is straightforward. More information on  $n$ -gram models and other language models may be found in (Chen and Goodman, 1998; Manning and Schutze, 1999).

Formally, the naive Bayes model assumes that that  $x|y$  is generated by a multinomial distribution

$$p_t(x|y) \propto \prod_{w \in V} \theta_w^{c(w,x)}, \quad \theta \in \mathbb{P}_V \quad (4)$$

where  $c(w, x)$  represents the number of times word  $w$  appears in document  $x$  and  $V$  is a finite dictionary of possible words. Applied to the concept drift problem, the local log-likelihood at time  $t$  is a smoothed or weighted version of the loglikelihood of the data  $D$  in (3) with the amount of smoothing determined by a non-negative smoothing kernel  $K_h : \mathbb{R} \rightarrow \mathbb{R}$

$$\ell_t(\theta|D) \stackrel{\text{def}}{=} \sum_{\tau \in I'} K_h(t - \tau) \sum_{j=1}^{N_\tau} \log p(x^{(\tau,j)}; \theta). \quad (5)$$

We assume that the kernel function is a normalized density concentrated around 0 and parameterized by a scale parameter  $h > 0$  reflecting its spread and satisfying the relation

$$K_h(r) = h^{-1} K_1(r/h)$$

where we denote  $K = K_1$  and refer to it as the kernel base form. We further assume that  $K$  has bounded support and  $\int u^r K(u) du < \infty$  for  $r \leq 2$ . The monograph (Wand and Jones, 1995) contains additional details on smoothing kernels and their use in non-parametric statistics.

Three popular kernel choices are the tricube, triangular and uniform kernels, defined as  $K_h(r) = h^{-1} K(r/h)$  where the  $K(\cdot)$  functions are respectively

$$K(r) = (1 - |r|^3)^3 \cdot 1_{\{|r| < 1\}} \quad (6)$$

$$K(r) = (1 - |r|) \cdot 1_{\{|r| < 1\}} \quad (7)$$

$$K(r) = 2^{-1} \cdot 1_{\{|r| < 1\}}. \quad (8)$$

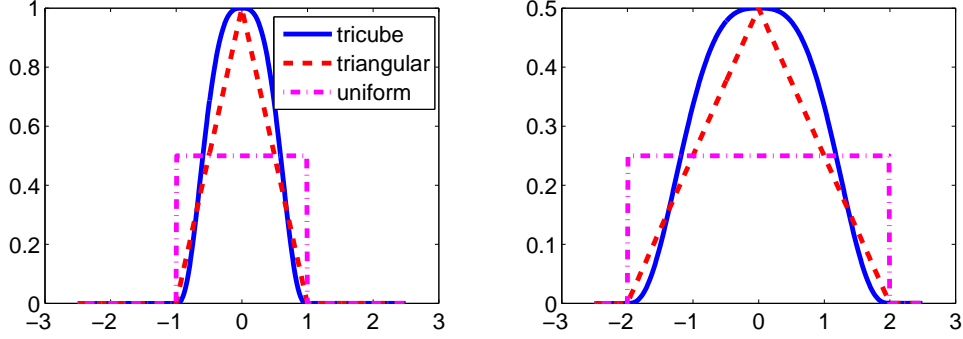


Figure 3: Tricube, triangular, and uniform kernels with scale  $h = 1$  (left) and  $h = 2$  (right).

Figure 3 displays these kernels for  $h = 1$  (left) and  $h = 2$  (right). The uniform kernel is the simplest choice and leads to a local likelihood (5) equivalent to filtering the data by a sliding window i.e.  $\hat{\theta}_t$  is computed based on data from adjacent time points with uniform weights. Unfortunately, it can be shown that the uniform kernel is suboptimal in terms of its statistical efficiency or rate of convergence to the underlying distribution (Wand and Jones, 1995). Surprisingly, the triangular kernel has a higher statistical efficiency than the Gaussian kernel and is the focus of our experiments in this subsection. We use the tricube kernel in the next subsection.

The scale parameter  $h$  is central to the bias-variance tradeoff. Large  $h$  represents more uniform kernels achieving higher bias and lower variance. Small  $h$  represents a higher degree of locality or lower bias but higher variance. Since  $\lim_{h \rightarrow 0} K_h$  approaches Dirac’s delta function and  $\lim_{h \rightarrow \infty} K_h$  approaches a constant function the local log-likelihood (5) interpolates between the loglikelihoods of the extreme local model and the extreme global model mentioned in Section 2 as  $h$  ranges from 0 to  $+\infty$ .

Solving the maximum local likelihood problem for each  $t$  provides an estimation of the entire drift  $\{\hat{\theta}_t : t \in \mathbb{R}\}$  with  $\hat{\theta}_t = \arg \max_{\theta \in \Theta} \ell_t(\theta|D)$ . In the case of the naive Bayes or  $n$ -gram model we obtain a closed form expression for the local likelihood maximizer  $\hat{\theta}_t$  as well as convenient expressions for its bias and variance. In general, however, there is no closed form maximizer and iterative optimization algorithms are needed in order to obtain  $\hat{\theta}_t = \arg \max_{\theta \in \Theta} \ell_t(\theta|D)$  for all  $t$ .

We denote the length of a document in (3) by  $|x^{(t,j)}| \stackrel{\text{def}}{=} \sum_{v \in V} c(x^{(t,j)}, v)$  and the total number of words in day  $t$  in (3) by  $|x^{(t)}| \stackrel{\text{def}}{=} \sum_{j=1}^{N_t} |x^{(t,j)}| = \sum_{v \in V} \sum_{j=1}^{N_t} c(v, x^{(t,j)})$ . We assume that the length of documents  $x^{(t,j)}$  is independent of  $t$  and is drawn from a distribution with expectation  $\lambda$  (this assumption is motivated by the data as indicated in Figure 2).

Under the above assumptions, the local likelihood (5) of the naive Bayes model becomes

$$\ell_t(\theta|D) = \sum_{\tau \in I'} K_h(t - \tau) \sum_{j=1}^{N_\tau} \sum_{w \in V} c(w, x^{(\tau,j)}) \log \theta_w \quad (9)$$

where  $\theta \in \mathbb{P}_V$ . The local likelihood has a single global maximum whose closed form is obtained by

setting to 0 the gradient of the Lagrangian

$$0 = \frac{1}{[\hat{\theta}_t]_w} \sum_{\tau \in I} K_h(t - \tau) \sum_{j=1}^{N_\tau} c(w, x^{(\tau,j)}) + \lambda_w$$

to obtain

$$[\hat{\theta}_t]_w = \frac{\sum_{\tau \in I} K_h(t - \tau) \sum_{j=1}^{N_\tau} c(w, x^{(\tau,j)})}{\sum_{\tau \in I} K_h(t - \tau) |x^{(\tau)}|}. \quad (10)$$

The estimator  $\hat{\theta}_t$  is a normalized linear combination of word counts where the combination coefficients are determined by the kernel function and normalized by the number of words in different days. We note that  $\hat{\theta}_t$  in (10) is different from a weighted averaging of the relative frequencies  $c(w, x^{(\tau,j)}) / \sum_{w'} c(w', x^{(\tau,j)})$ .

We distinguish between two fundamental scenarios for predicting the drift  $\theta_t$ .

**Offline scenario:** The goal is to estimate the drift  $\{\theta_t : t \in \mathbb{R}\}$  given the entire dataset  $D$ . In this case we will consider symmetric kernels  $K(r) = K(-r)$  which will achieve an increased convergence rate of  $\hat{\theta}_t \rightarrow \theta_t$  as indicated by Proposition 2.

**Online scenario:** The goal is estimate a model for the present distribution  $\theta_t$  using training data from the past i.e. a dataset whose time stamps are strictly smaller than  $t$ . This corresponds to situations where the data arrives sequentially as a temporal stream and at each time point a model for the present is estimated using the available stream at that time. We realize this restriction by constraining  $K$  to satisfy  $K(r) = 0, r \leq 0$ .

### 3.2 Bias-Variance Analysis of $\hat{\theta}_t$

As with other statistical estimators, the accuracy of  $\hat{\theta}_t$  may be measured in terms of its mean squared error  $E(\hat{\theta}_t - \theta_t)^2$  which decomposes as the sum of the squared bias and variance of  $\hat{\theta}_t$ . Note that the expectation and variance in this case are taken with respect to the generation of the training data  $D$ . Examining these quantities allow us to study the convergence rate of  $\hat{\theta}_t \rightarrow \theta$  and its leading coefficient .

**Proposition 1.** *The bias vector  $bias(\hat{\theta}_t) \stackrel{\text{def}}{=} E\hat{\theta}_t - \theta_t$  and variance matrix of  $\hat{\theta}_t$  in (10) are*

$$bias(\hat{\theta}_t) = \frac{\sum_{\tau \in I} K_h(t - \tau) |x^{(\tau)}| (\theta_\tau - \theta_t)}{\sum_{\tau \in I} K_h(t - \tau) |x^{(\tau)}|} \quad (11)$$

$$Var(\hat{\theta}_t) = \frac{\sum_{\tau \in I} K_h^2(t - \tau) |x^{(\tau)}| (diag(\theta_\tau) - \theta_\tau \theta_\tau^\top)}{(\sum_{\tau \in I} K_h(t - \tau) |x^{(\tau)}|)^2} \quad (12)$$

where  $diag(z)$  is the diagonal matrix  $[diag(z)]_{ij} = \delta_{ij} z_i$ .

*Proof.* The random variable (RV)  $c(w, x^{(\tau,j)})$  is distributed as a sum of multivariate Bernoulli RVs, or single draws from multinomial distribution. The expectation and variance of the estimator are that of a linear combination of iid multinomial RVs. To conclude the proof we note that for  $Y \sim \text{Mult}(1, \theta)$ ,  $EY = \theta$ ,  $\text{Var}(\theta) = \text{diag}(\theta) - \theta\theta^\top$ .  $\square$

Examining Equations (11)-(12) reveals the expected dependency of the bias on  $h$  and  $\theta_t$ . The contribution to the bias of the terms  $(\theta_\tau - \theta_t)$ , for large  $|\tau - t|$ , will decrease as  $h$  decreases since the kernel becomes more localized and will reduce to 0 as  $h \rightarrow 0$ . Similarly, for slower drifts,  $\|\theta_\tau - \theta_t\|, t \approx \tau$  will decrease and reduce the bias.

Despite the relative simplicity of Equations (11)-(12), it is difficult to quantitatively capture the relationship between the bias and variance, the sample size,  $h, \lambda$ , and the smoothness of  $\theta_t, g$ . Towards this goal we derive the following asymptotic expansions. Below, the notation  $g_n \xrightarrow{P} f$  represents convergence in probability of  $g_n$  to  $f$  i.e.  $\forall \epsilon > 0, P(|g_n - f| > \epsilon) \rightarrow 0$ , and  $g_n = o_P(f_n)$  represents  $g_n/f_n \xrightarrow{P} 0$ .

**Proposition 2.** *Assuming (i)  $\theta, g$  are smooth in  $t$ , (ii)  $h \rightarrow 0, hn \rightarrow \infty$ , (iii)  $g > 0$  in a neighborhood of  $t$ , and (iv) document lengths do not depend on  $t$  and have expectation  $\lambda$ , the bias vector and variance matrix are in the offline case*

$$\text{bias}(\hat{\theta}_t|I) = h^2 \mu_{21}(K) \left( \dot{\theta}_t \frac{g'(t)}{g(t)} + \frac{1}{2} \ddot{\theta}_t \right) + o_P(h^2) \quad (13)$$

$$\text{Var}(\hat{\theta}_t|I) = \frac{\mu_{02}(K)}{(nh)g(t)\lambda} (\text{diag}(\theta_t) - \theta_t \theta_t^\top) + o_P((nh)^{-1})$$

and in the online case

$$\text{bias}(\hat{\theta}_t|I) = h \mu_{11}(K) \dot{\theta}_t + o_P(h) \quad (14)$$

$$\begin{aligned} \text{Var}(\hat{\theta}_t|I) &= \left( \frac{\mu_{02}(K)}{nhg(t)\lambda} + \frac{\mu_{12}(K)g'(t)}{ng^2(t)\lambda} \right) (\text{diag}(\theta_t) - \theta_t \theta_t^\top) \\ &\quad + \frac{\mu_{12}(K)}{n\lambda g(t)} (\text{diag}(\dot{\theta}_t) - \dot{\theta}_t \theta_t^\top - \theta_t \dot{\theta}_t^\top) + o_P((nh)^{-1}) \end{aligned}$$

where  $\dot{\theta}_t$  is the vector  $[\dot{\theta}_t]_i = \frac{d}{dt}[\theta_t]_i$  and

$$\mu_{kl}(K) \stackrel{\text{def}}{=} \int t^k K^l(t) dt < \infty \quad 0 \leq k, l \leq 2.$$

*Proof.* The proof follows standard expansions similar to the ones used in studying local polynomial regression but modified to our setting. We start by expanding the numerator and denominator of the bias and variance in the offline case. Our main tools are the law of large numbers, changing the integration variable, and Taylor series expansion. For notational simplicity we assume below that  $t = 0$ . The arguments below may be modified at some notational expense for  $t \neq 0$  to produce Equations (13)-(14). In the proof below we use slightly different notation with  $x_{\tau_i}$  representing the  $i$ -training example which is associated with time  $\tau_i$ .

We expand the denominator and numerator of the bias (11) multiplied by  $1/n$ :

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n K_h(\tau_i) |x_{\tau_i}| &\xrightarrow{P} \lambda \int g(t) K_h(t) dt = \lambda h^{-1} \int g(t) K(t/h) dt = \lambda \int g(uh) K(u) du \\ &= \lambda \int K(u) (g(0) + o(1)) du = \lambda g(0) + o(1). \end{aligned}$$

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n K_h(\tau_i) |x_{\tau_i}| (\theta_{\tau_i} - \theta_0) \xrightarrow{p} \lambda h^{-1} \int g(t) (\theta_t - \theta_0) K(t/h) dt = \lambda \int g(uh) (\theta_{uh} - \theta_0) K(u) du \\
& = \lambda \int (g(0) + g'(0)uh + g''(0)u^2h^2/2 + o(u^2h^2)) (\dot{\theta}_0uh + \ddot{\theta}_0u^2h^2/2 + o(u^2h^2)) K(u) du \\
& = \lambda h^2 \mu_{21}(K) \left( g'(0)\dot{\theta}_0 + \frac{1}{2}g(0)\ddot{\theta}_0 \right) + o(h^2).
\end{aligned}$$

Above, we used the offline assumption by exploiting the symmetry of the kernel to deduce  $\int K(u)u du = 0$ . Dividing the two expansions and replacing  $o(h^2)$  with  $o_P(h^2)$  due to the law of large numbers approximation establishes (13).

Similarly we expand the denominator and numerator of the variance matrix times  $1/n^2$  and  $1/n$  respectively

$$\begin{aligned}
& \left( \frac{1}{n} \sum_{i=1}^n K_h(\tau_i) |x_{\tau_i}| \right)^2 \xrightarrow{p} \left( \lambda \int K(u) (g(u) + o(1)) du \right)^2 = \lambda^2 g^2(0) + o(1)^2 \\
& \frac{1}{n} \sum_{i=1}^n K_h^2(\tau_i) |x_{\tau_i}| \text{Var}(\theta_{\tau_j}) \xrightarrow{p} \lambda h^{-2} \int K^2(t/h) g(t) \text{Var}(\theta_t) dt = \lambda h^{-1} \int K^2(u) g(uh) \text{Var}(\theta_{uh}) du \\
& = \lambda h^{-1} \int K^2(u) (g(0) + g'(0)uh + o(uh)) (\text{Var}(\theta_0) + \text{Var}(\theta_0)uh + o(uh)) du \\
& = \lambda h^{-1} g(0) \text{Var}(\theta_0) \mu_{02}(K) + o(h)
\end{aligned}$$

where again we used the kernel symmetry to deduce  $\int K^2(u)u du = 0$ . Since  $\text{Var}(\theta_t) = (\text{diag}(\theta_t) - \theta_t \theta_t^\top)$ , dividing the second expansion by the first and dividing by  $n^{-1}$  provides the desired result.

In the online setting, the kernel is no longer symmetric and  $\int K(u)u du \neq 0$  which lowers the rate of convergence. The expansions of the numerator of the bias and variance are

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n K_h(\tau_i) |x_{\tau_i}| (\theta_{\tau_i} - \theta_0) \xrightarrow{p} \lambda \int (g(0) + g'(0)uh + o(uh)) (\dot{\theta}_0uh + o(uh)) K(u) du \\
& = \lambda h \mu_{11}(K) \dot{\theta}_0 g(0) + o(h).
\end{aligned}$$

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n K_h^2(\tau_i) |x_{\tau_i}| \text{Var}(\theta_{\tau_j}) \xrightarrow{p} \frac{\lambda}{h} \int K^2(u) (g(0) + g'(0)uh + o(uh)) (\text{Var}(\theta_0) + \text{Var}(\theta_0)uh + o(uh)) du \\
& = \frac{\lambda}{h} g(0) \text{Var}(\theta_0) \mu_{02}(K) + \lambda \mu_{12}(K) (g(0) \text{Var}(\theta_0) + g'(0) \text{Var}(\theta_0)) + o(h).
\end{aligned}$$

Noticing that  $\text{Var}(\theta_t) = \text{diag}(\dot{\theta}_t) - \dot{\theta}_t \theta_t^\top - \theta_t \dot{\theta}_t^\top$  concludes the proof.  $\square$

### 3.3 MSE, MISE and their Dependence on the Drift Parameters

Since the component-wise mean squared error  $\text{mse}([\hat{\theta}_t]_i) = \mathbb{E}([\hat{\theta}_t]_i - [\theta_t]_i)^2$  decomposes as the sum of the variance and the squared bias we have the following direct corollary of Proposition 2.

**Corollary 1.** Under the assumptions in Proposition 2, the component-wise mean squared error  $mse([\hat{\theta}_t]_i) = E([\hat{\theta}_t]_i - [\theta_t]_i)^2$ ,  $i = 1, \dots, V$  are for the offline case

$$mse([\hat{\theta}_t]_i) = h^4 \mu_{21}^2(K) \left( [\hat{\theta}_t]_i \frac{g'(t)}{g(t)} + \frac{1}{2} [\ddot{\theta}_t]_i \right)^2 + \frac{\mu_{02}(K)}{nhg(t)\lambda} [\theta_t]_i (1 - [\theta_t]_i) + o_P(h^4 + (nh)^{-1}).$$

and for the online case

$$mse([\hat{\theta}_t]_i) = h^2 \mu_{11}^2(K) [\dot{\theta}_t]_i^2 + \left( \frac{\mu_{02}(K)}{nhg(t)\lambda} + \frac{\mu_{12}(K)g'(t)}{ng^2(t)\lambda} \right) [\theta_t]_i (1 - [\theta_t]_i) + \frac{\mu_{12}(K)}{n\lambda g(t)} [\dot{\theta}_t]_i (1 - 2[\theta_t]_i) + o_P(h^2 + (nh)^{-1}).$$

**Corollary 2.** Under the assumptions in Proposition 2, and in particular  $h \rightarrow 0, nh \rightarrow \infty$ , the estimator  $\hat{\theta}_t$  is consistent i.e.  $\hat{\theta}_t \xrightarrow{P} \theta_t$  in both the offline and online settings.

*Proof.* The proof follows from the fact that under these conditions we have convergence in the second moment of the components of  $\hat{\theta}_t - \theta_t$  to 0.  $\square$

Proposition 2 is important as it specifies the conditions for consistency as well as the rate of convergence. The conditions specified in Proposition 2 for consistency of the estimator (in particular  $h \rightarrow 0, nh \rightarrow \infty$ ) are standard conditions in non-parametric kernel smoothing and are similar to those of other related estimators such as the kernel density estimator and the Nadaraya-Watson local regression estimator.

The rates of convergence indicated by the argument of  $o_P(\cdot)$  in Proposition 2 are important as they quantify the rates at which the estimators converges to the underlying drift. In particular, it is interesting to note the fact that the bias of online kernels converges at a linear rather than the quadratic rate of the offline kernels. This is a quantification of the fact that looking at the past and future helps predict the present more than looking only at the past.

Additional important insights we obtain from corollary 1 and (13)-(14) are the dependency of the bias, variance, and mse on the drift parameters: the drift speed indicated by  $\dot{\theta}_i$ , the rate of change of the log sampling density  $d \log g(t) = g'(t)/g(t)$ , the number of documents  $n$ , and the expected length of the documents  $\lambda$ . Intuitively, the estimation task is easier if the drift is slower ( $\dot{\theta}_t$  is smaller), the time sampling variation ( $d \log g(t)/dt$ ) is smaller, and there are more and longer documents. Using expressions (13)-(14) we confirm these intuitive observations and quantify them: the bias is reduced as the drift speed and time sampling variation are lower while the variance is reduced as we have more (denoted by  $n$ ) and longer (denoted by  $\lambda$ ) documents.

Corollary 1 and expressions (13)-(14) also reveal somewhat less intuitive insights. First, the variance grows linearly with  $[\theta_t]_i(1 - [\theta_t]_i)$  (in the offline case; the online variance is slightly more complicated). In the case of documents, the word probabilities  $\theta_i$  are typically very small and thus the larger they are the higher the  $[\theta_t]_i(1 - [\theta_t]_i)$  factor is in the asymptotic variance (see Figure 4, left). Note also that  $[\theta_t]_i(1 - [\theta_t]_i)$  is the inverse Fisher information of the binomial and therefore bounds the variance of the optimal estimator.

Another interesting observation is that in the case of large time intervals the factor  $g'(t)/g(t)$  tends to be very small (see Figure 4 (right)). In such cases the  $g'(t)/g(t)$  factor associated with the first offline bias term is likely to be negligible compared to the second term making the bias increase linearly with  $\ddot{\theta}_t$  and independent of  $\theta_t$ . This indicates zero (or nearly zero) offline bias for linear drift as its second derivative is zero.

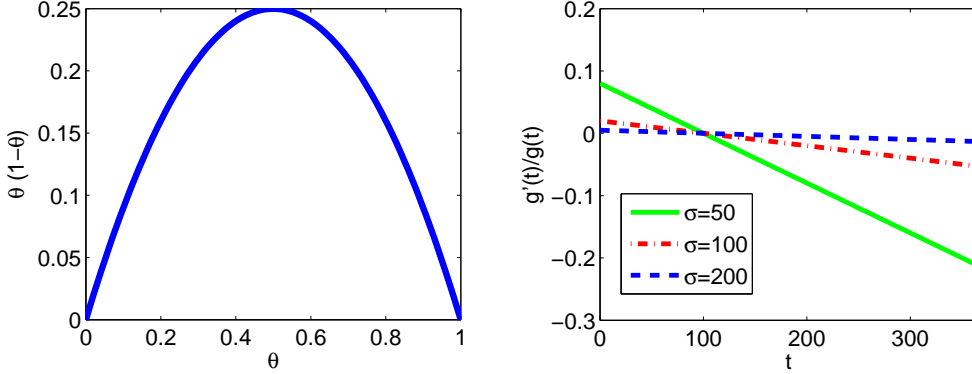


Figure 4: Left: The variance increases linearly with  $[\theta_t]_i(1 - [\theta_t]_i)$  which is monotonically increasing in  $[\theta_t]_i$  for small values such as word probabilities. Right: The  $g'(t)/g(t)$  factor associated with the first bias term is likely to be negligible compared to the second term making the bias increase linearly with  $\ddot{\theta}_t$  and independent of  $\dot{\theta}_t$  (indicating zero bias for linear drift). The figure plots  $g'(t)/g(t)$  for  $t \in [1, 365]$  and  $g(t) = N(100, \sigma)$  with  $\sigma = 50, 100, 200$ .

The above proposition and corollary are expressed in terms of the mean squared error at a particular time point  $t$ . This is suitable in cases where we are interested in estimation accuracy at a specific time point such as the present time. In other cases, more insightful criteria are the integrated squared error (ise) and the mean integrated square error (mise) which average the estimation error over  $t$

$$\text{ise}([\hat{\theta}]_i) = \int ([\hat{\theta}_t]_i - [\theta_t]_i)^2 dt \quad (15)$$

$$\text{mise}([\hat{\theta}]_i) = \mathbb{E} \int ([\hat{\theta}_t]_i - [\theta_t]_i)^2 dt = \int \text{mse}(\hat{\theta}_t) dt \quad (16)$$

where the last equality follows from changing the order of the two integrals.

The integrated version of Corollaries 1 and 2 are listed below.

**Corollary 3.** *Under the assumptions in Proposition 2, we have in the offline case*

$$\text{mise}([\hat{\theta}]_i) = h^4 \mu_{21}^2(K) \mu_{02} \left( [\dot{\theta}]_i \frac{g'(t)}{g(t)} + \frac{1}{2} [\ddot{\theta}]_i \right) + \frac{\mu_{02}(K)}{nh\lambda} \mu_{01} \left( \frac{[\theta_t]_i(1 - [\theta_t]_i)}{g(t)} \right) + o_P(h^4 + (nh)^{-1}).$$

A similar expansion for the online case is straightforward. Under these assumptions and in particular  $h \rightarrow 0, nh \rightarrow \infty$  the total mise  $\sum_i \text{mise}([\hat{\theta}]_i)$  converges to 0 in both the online and offline scenarios.

### 3.4 Bandwidth Selection and Optimal Convergence Order

A central issue in local likelihood modeling and non-parametric estimation in general is selecting the appropriate bandwidth  $h$ . Such a selection is critical to effective modeling and is the subject of substantial research. Figure 5 displays the RCV1 test set loglikelihood for the online and offline scenarios as a function of the (triangular) kernel's bandwidth. As expected, offline kernels performs

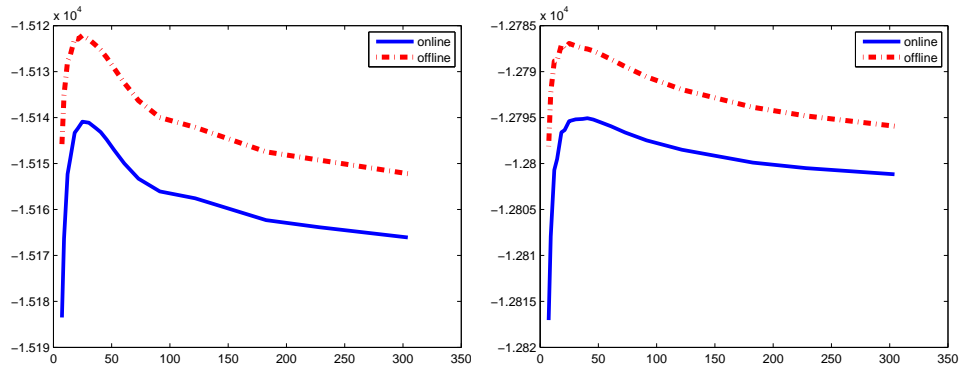


Figure 5: Log-likelihood of held out test set as a function of the triangular kernel’s bandwidth for the two largest RCV1 categories (CCAT (left) and GCAT (right)) and the most frequent 500 words. Training set size was 100 documents per day and test set performance was averaged over repeated sampling to remove noise. In all four cases, the optimal bandwidth seems to be approximately 25 which indicates a support of 25 days for the online kernels and 50 days for the offline kernels.

better than online kernels with both achieving the best performance for a bandwidth approximately 25 which corresponds to a support of 25 days in the online scenario and 50 days in the offline scenario. Similar results are displayed in Figure 6 for the AOL dataset where the optimal bandwidth is 7 indicating offline support of 14 days. Note that in addition to obtaining higher accuracy than the global model corresponding to  $h \rightarrow \infty$ , the local model enjoys computational efficiency as it ignores a large portion of the training data.

The first and perhaps most obvious technique for selecting  $h$  is maximum likelihood cross validation (MLCV). In this technique, the dataset  $D$  is randomly partitioned to two subsets  $D = D_A \cup D_B$  - one  $D_A$  used to construct a local likelihood estimator  $\hat{\theta}_t^{(D_A)}$  and one used to evaluate the loglikelihood of the estimator

$$\ell^{(D_A)}(D_B) = \sum_{i \in D_B} \log p_{\hat{\theta}_t^{(D_A)}}(x_i).$$

Ten-fold cross validation averages this process ten times where each time 90% of the data is kept for constructing  $\hat{\theta}^*$  and 10% of the data is used to evaluate the log-likelihood of  $\hat{\theta}^*$ . The MLCV estimator then proceeds to select the bandwidth  $h$  that maximizes the ten-fold cross validation function

$$h_{\text{MLCV}} = \arg \max_{h > 0} \sum_{j=1}^{10} \ell^{(D_{A_j})}(D_{B_j}).$$

On RCV1 data, the performance of such cross validation schemes is extremely good and the estimated bandwidth possesses test set loglikelihood that is almost identical to the optimal bandwidth (see Figure 7, left). Allowing the kernel scale to vary over time results in a higher modeling accuracy than using fixed bandwidth for all dates (see Figure 7, right). A time-dependent cross validation procedure may be used to approximate the time-dependent optimal bandwidth which performs slightly better than the fixed-date cross validation estimator. Note that the accuracy

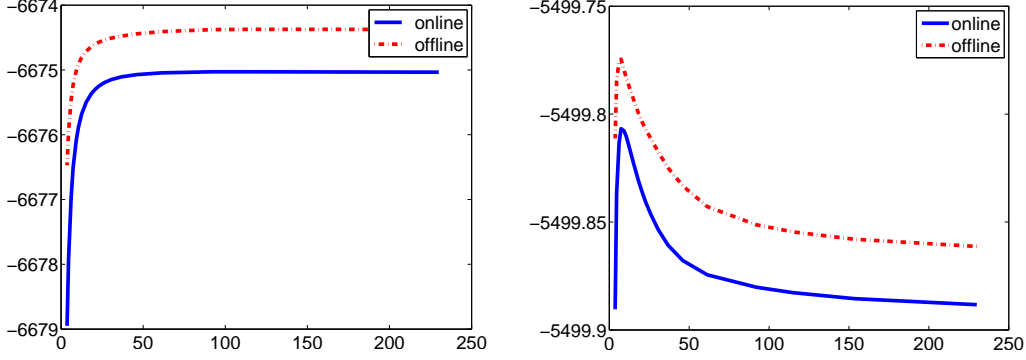


Figure 6: Log-likelihood of held out test set as a function of the triangular kernel’s bandwidth for the AOL dataset with 20000 documents per day (left) and 80000 docs per day (right)). Due to the short document length of queries, 20000 documents per day are not sufficient to motivate a non-global estimator. In the case of 80000 documents per day, the optimal bandwidth in both cases is around 7 which indicates a support of 7 days for the online kernel and 14 days for the offline kernel.

with which the cross validation estimator approximates the optimal bandwidth is lower in the time-dependent or varying bandwidth situation due the fact that much less data is available in each of the daily cross validation problems.

While performing well on the RCV1 data, in some cases the MLCV is problematic from both theoretical and practical perspectives (DasGupta, 2008). In these cases, an alternative is least squares cross validation (LSCV) which is based on the following decomposition of the mise (16)

$$\text{mise}([\hat{\theta}]_i) = \text{E} \int [\hat{\theta}_t]_i^2 dt - 2\text{E} \int [\hat{\theta}_t]_i [\theta_t]_i dt + \int [\theta_t]_i^2 dt \quad (17)$$

and noting that the third term does not depend on  $h$ . We can thus construct an unbiased estimator for  $\text{mise}([\hat{\theta}]_i) - \int [\theta_t]_i^2 dt$  as follows

$$\text{LSCV}(h, i) = \int [\hat{\theta}_t]_i^2 dt - 2n^{-1} \sum_{j=1}^n [\hat{\theta}_t^{(-j)}]_i \quad (18)$$

where  $\sum_{j=1}^n [\hat{\theta}_t^{(-j)}]_i$  is the local likelihood estimator for the drift using the dataset  $D$  but omitting the  $i$ -observation. Assuming we are interesting in minimizing the mise over all the parameters of  $\hat{\theta}$  we obtain

$$\hat{h}_{\text{LSCV}} = \arg \min_{h>0} \sum_{i=1}^V \text{LSCV}(h, i) \quad (19)$$

which is an unbiased estimator of the minimizer of the total mise  $\arg \min_{h>0} \sum_{i=1}^V \text{mise}([\hat{\theta}]_i)$ .

From a theoretical perspective, we can get additional insights by analytically minimizing the leading terms of the mse or mise as a function of  $h$ . The resulting minimizer expresses in closed

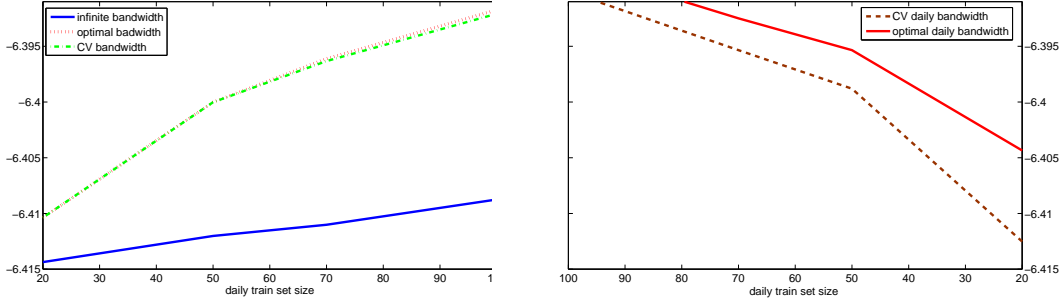


Figure 7: Per-word log-likelihood over held-out test set for various bandwidths as a function of the daily training set size. Left: The extreme global model corresponding to  $h \rightarrow \infty$  performs worst. Selecting the bandwidth by cross validation results in an accurate estimate and test-set loglikelihood almost identical to that of the optimal slope. Right: Allowing the kernel scale to vary over time results in a higher modeling accuracy than using fixed bandwidth for all dates.

form the dependency of the optimal bandwidth on the problem parameters  $n, \lambda, \dot{\theta}, \ddot{\theta}, g(t), \theta_t$ . For example minimizing the leading term of  $\sum_{j=1}^V \text{mse}([\hat{\theta}_t]_i)$  we obtain

$$\hat{h}_t^5 = \frac{\mu_{02}(K) \text{tr}(\text{diag}(\theta_t) - \theta_t \theta_t^\top)}{4n\lambda\mu_{21}^2(K) \sum_j \left( [\dot{\theta}_t]_j g'(t) / \sqrt{g(t)} + \sqrt{g(t)} [\ddot{\theta}_t]_j / 2 \right)^2}. \quad (20)$$

**Proposition 3.** *Under the assumptions in Proposition 2 in the offline case we have*

$$\inf_{h>0} \text{mse}([\hat{\theta}_t]_i) = n^{-4/5} \left\{ \mu_{21}^2(K) \left( [\dot{\theta}_t]_i \frac{g'(t)}{g(t)} + \frac{1}{2} [\ddot{\theta}_t]_i \right)^2 + \frac{\mu_{02}(K)}{g(t)\lambda} [\theta_t]_i (1 - [\theta_t]_i) + o_P(1) \right\}. \quad (21)$$

*Proof.* Equation 21 follows from minimizing the mse as a function of  $h$  and substituting the resulting minimizer back in the mse.  $\square$

Proposition 3 indicates that the mse of the local likelihood estimator converges to 0 at the rate  $n^{-4/5}$  which is only slightly lower than the parametric rate of  $n^{-1}$  (DasGupta, 2008) with the latter rate being achieved if there is no drift and the estimation problem becomes maximum likelihood for the multinomial model. The particular  $n^{-4/5}$  rate is dependent, however, on successful bandwidth selection rules such as the MLCV or LSCV.

As expected, the optimal bandwidth decreases as  $n, \lambda, \|\dot{\theta}_t\|, \|\ddot{\theta}_t\|$  increases. Intuitively this makes sense since in these cases the variance decreases and bias either increases or stays constant. In practice,  $\dot{\theta}_t, \ddot{\theta}_t$  may vary significantly with time which leads to the conclusion that a single bandwidth selection for all  $t$  may not perform adequately. These changes are illustrated in Figure 8 which demonstrates the temporal change in the gradient norm.

A more surprising result is the non-monotonic dependency of the optimal bandwidth on the time sampling distribution  $g(t)$ . The dependency, expressed by

$$\hat{h}_t \propto \left( \sum_{j=1}^V (c_{1j} / \sqrt{g(t)} + c_{2j} \sqrt{g(t)})^2 \right)^{-1/5}$$

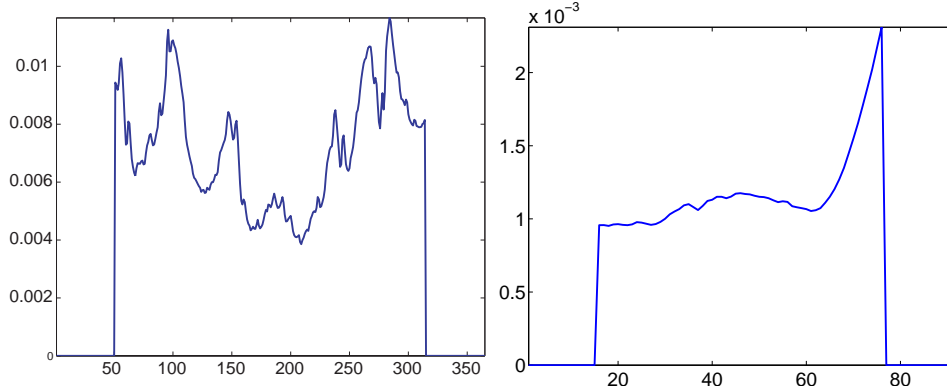


Figure 8: Estimated gradient norm for the most popular category in RCV1 (left) and AOL (right) as a function of  $t$ . The derivatives were estimated using local smoothing. To avoid running into boundary effects we ignore the first and last 50 days in RCV1 and 15 days in AOL.

is illustrated in Figure 9 (left) where we assume for simplicity that  $c_{1j}, c_{2j}$  do not change with  $j$  resulting in

$$(\hat{h}_t)^{-1} \propto (c_1/g(t) + c_2g(t) + c_3)^{1/5}.$$

The key to understanding this relationship is the increased asymptotic bias due to the presence of the term  $g'(t)/g(t)$  in Equation (13). Indeed, plotting the inverse of the optimal bandwidth (we actually average that quantity over the word-specific optimal bandwidths for different words) for the RCV1 data as a function of the daily word count (which is proportional to  $g(t)$ ) in Figure 9 (right) reveals a trend similar to the theoretical dependency displayed in Figure 9 (left). The mismatch in absolute numbers is due to the proportionality constant that is hard to determine in practice.

We finally point out that different words  $w$  have different parameters  $[\theta_t]_w$  and parameter derivatives  $[\dot{\theta}_t]_w$ . As a result, it is unlikely that a single bandwidth will work best for all words. Frequent words are likely to benefit more from narrow kernel smoothing than rare words which almost never appear. One possible solution is to use regularization as described in Section 5. A second solution is to group the words according to the corresponding probability and drift speed  $[\theta_t]_w, [\dot{\theta}_t]_w$  and use different bandwidths for each group, determined by cross validation. A lower bandwidth should be used for frequent words while a high bandwidth should be used for rare words. A systematic investigation of these topics is beyond the scope of this paper.

## 4 Local Likelihood for Logistic Regression

Often, the primary goal is predicting the value of  $y$  given  $x$  rather than density estimation as in language modeling. In this section, we focus on such cases and describe modeling the conditional drift  $\{p_t(y|x) : t \in I\}$  using local likelihood for logistic regression. By direct analogy to Equation (5) the conditional local likelihood estimator  $p_t(y|x)$  is the maximizer of the locally weighted conditional

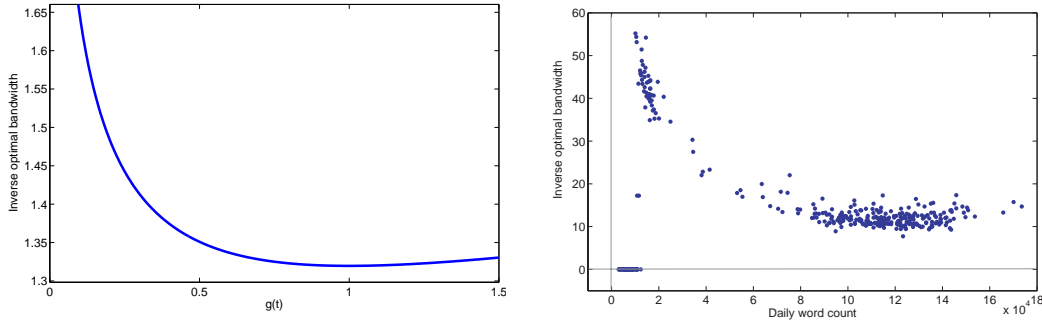


Figure 9: Left: Inverse of the optimal bandwidth derived from Equation (20) as a function of  $g(t)$ :  $(\hat{h}_t)^{-1} \propto (c_1/\sqrt{g(t)} + c_2\sqrt{g(t)})^{2/5}$  (we take  $c_1 = c_2 = 1$ ). The graph shows the non-monotonic dependency between  $\hat{h}^{\text{opt}}$  and  $g(t)$ . Right: Inverse of the optimal bandwidth  $(\hat{h}_t)^{-1}$  (averaged over the optimal bandwidths for the top 3000 words) as a function of the daily word count (which is proportional to  $g(t)$ ). The two graphs show an interesting correspondence between theory and practice and illustrate the non-monotonic dependency between the optimal bandwidth and  $g(t)$ .

loglikelihood

$$\ell_t(\eta|D) = \sum_{\tau \in I} K_h(t - \tau) \sum_{j=1}^{N_\tau} \log p(y^{(\tau,j)} | x^{(\tau,j)}; \eta) \quad \eta \in \Theta. \quad (22)$$

As in the generative case, the kernel parameter  $h$  balances the degree of the kernel's locality and controls the bias-variance tradeoff.

Denoting by  $f(x)$  the vector of relative frequencies in the document  $x$ , the logistic regression model

$$\log \frac{p(1|x; \theta_t)}{1 - p(1|x; \theta_t)} = \langle \theta_t, f(x) \rangle, \quad \theta \in \mathbb{R}^V$$

leads to the following local conditional likelihood

$$\ell_t(\eta|D) = - \sum_{\tau \in I} K_h(t - \tau) \sum_{j=1}^{N_\tau} \log \left( 1 + e^{-y^{(\tau,j)} \langle x^{(\tau,j)}, \eta \rangle} \right).$$

In contrast to the naive Bayes model in the previous section, the local likelihood does not have a close form maximizer. However, it can be shown that under mild conditions it is a concave problem exhibiting a single global maximum (Loader, 1999). Most of the standard iterative gradient-based algorithms for training logistic regression can be modified to account for the local weighting introduced by the smoothing kernel. Moreover, recently popularized regularization penalties such as  $c\|\eta\|^q$ ,  $q = 1, 2$  may be added to the local likelihood in order to obtain a local regularized version equivalent to maximum posterior estimation.

Figure 10 (right) displays classification error rate over a held-out test set for local logistic regression as a function of the train set size. The classification task was predicting the most popular class vs the second most popular class in RCV1. The plots in the figure contrast the performance of the online and offline tricube kernels with optimal and infinite bandwidths, using

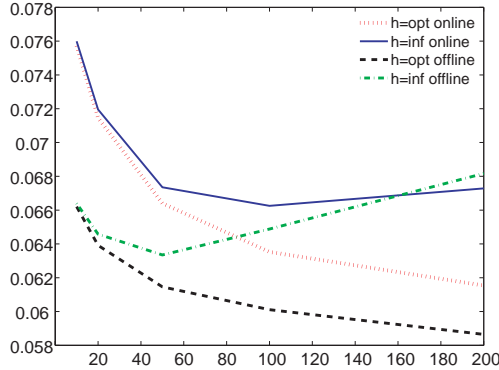


Figure 10: Classification error rate over a held-out test set for the local logistic regression model as a function of the train set size.

$L_2$  regularization. The optimization was carried out using a modification of the logistic regression BBR package. The local model achieved a relative reduction of error rate over the global model by about 8%. Note that as expected, the online kernel generally achieves worse error rates than the offline kernels. In all the experiments mentioned above we averaged over multiple random samplings of the training set to remove sampling noise.

## 5 Penalized Local Likelihood

As indicated by the analysis in the previous sections applying a bandwidth  $h \ll \infty$  for rare words is likely to result in poor estimation as there simply are not enough word appearances in small neighborhoods. For example, a word with low probability that appears twice in the corpus - once at the beginning and once at the end cannot benefit from local estimates. One potential strategy to overcome this is to use a bandwidth  $h \ll \infty$  for non-rare words and bandwidth  $\infty$  for rare words. In this section we consider an alternative technique based on regularization towards the global estimates.

We focus on the following penalized local likelihood whose maximizer is denoted by  $\hat{\theta}_t^*$

$$\ell_t^*(\theta|D) = \sum_{\tau \in I'} K_h(t - \tau) \sum_{j=1}^{N_\tau} \sum_{w \in V} c(w, x_{\tau j}) \log \theta_w - \sum_w \lambda_w |\log \theta_w - \log q_w|.$$

The regularization term shrinks the local likelihood towards the global estimate  $q_w$  through an  $L_1$  penalty on the logarithms of the probabilities. Rearranging the summation and letting  $c_w = \sum_{\tau \in I'} K_h(t - \tau) \sum_{j=1}^{N_\tau} c(w, x_{\tau j})$ , we can focus on an equivalent but simplified objective function

$$\ell_t^*(\theta) = \sum_{w \in V} c_w \log \theta_w - \lambda_w |\log \theta_w - \log q_w|. \quad (23)$$

Our penalized estimator  $\hat{\theta}_t^*$  shows some resemblance to the lasso estimator due to the  $L_1$  penalty. It also bears some similarity to the linear interpolation smoothing as described in Section 5.3. Assuming that  $\lambda_w$  does not vary with  $w$ , very large  $\lambda_w$  result in  $\hat{\theta}_t^* = q$  while very small  $\lambda_w$  result

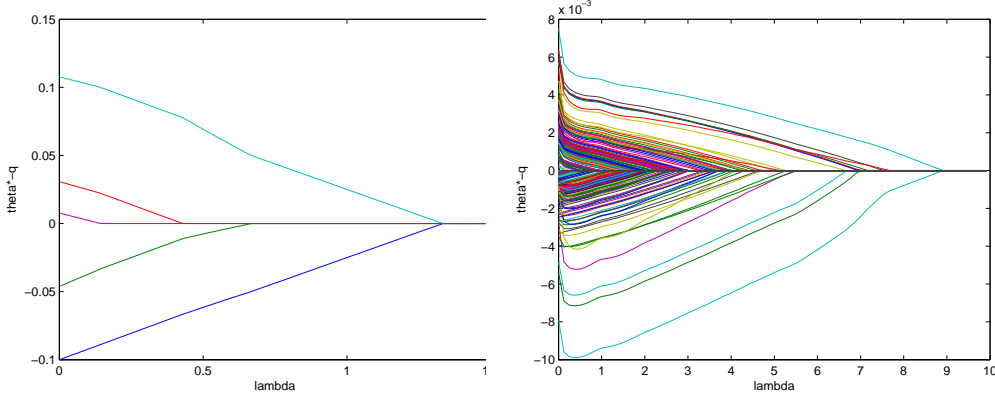


Figure 11: Deviation of penalized estimate from global model  $[\hat{\theta}_t^*]_w - q_w$  as a function of  $\lambda_w$ . Left: A toy problem with count vector  $(0, 2, 3, 4, 4)$  and  $q = (.1, .2, .2, .2, .3)$ . Right: Estimated word probabilities from documents in one day of the RCV1 data set.

in  $\hat{\theta}_t^* = \hat{\theta}_t$ . The departure of  $[\hat{\theta}_t^*]_w$  from  $q_w$  is shown in Figures 11-12 as a function of  $\lambda_w$  (assumed in this case to be the same for all  $w$ ). As  $\lambda_w$  increase we obtain a sparse difference vector  $\hat{\theta}_t^* - q$  with deviations from 0 only for a few select words.

## 5.1 Optimization problem

When  $\theta_t$  is two dimensional and  $\lambda_1 = \lambda_2$ , Equation (23) can be maximized by taking partial derivatives and setting them to 0. The solution  $\hat{\theta}_t^*$  is a soft threshold function applied to the non-penalized maximum likelihood estimate  $\hat{\theta}_t$ :

$$[\hat{\theta}_t^*]_i = \begin{cases} [\hat{\theta}_t]_i + \lambda_i/N & [\hat{\theta}_t]_i < q_i - \lambda_i/N \\ q_i & q_i - \lambda_i/N \leq [\hat{\theta}_t]_i \leq q_w + \lambda_i/N \\ [\hat{\theta}_t]_i - \lambda_i/N & q_i + \lambda_i/N < [\hat{\theta}_t]_i \end{cases}$$

In general, however, the maximization of Equation (23) is neither concave nor smooth. However, by analyzing the derivatives of (23) we are able to derive an effective algorithm that calculates  $\hat{\theta}_t^*$  and show that there is a unique local maximum.

We first introduce the notation  $e_{i,j} = (0, \dots, 1, \dots, -1, \dots, 0)$  for the vector that is 1 at position  $i$ ,  $-1$  at position  $j$  and 0 everywhere else and denote the one-sided directional derivative of  $\ell_t^*(\cdot)$  along a vector  $v$  as

$$\frac{\partial \ell_t^*(\theta)}{\partial v} \stackrel{\text{def}}{=} \lim_{h \rightarrow 0^+} \frac{\ell(\theta + hv) - \ell(\theta)}{h}$$

**Lemma 1.** *If  $c_w \leq \lambda_w$  for all  $w$ ,  $\hat{\theta}_t^* = q$ .*

*Proof.* The partial derivative of (23) is

$$\frac{\partial \ell_t^*}{\partial \theta_w} = \frac{c_w - \lambda_w \text{sgn}(\theta_w - q_w)}{\theta_w}.$$

Since  $c_w \leq \lambda_w$ , the partial derivative is positive when  $\theta_w < q_w$  and negative when  $\theta_w > q_w$ , thus the maximum is reached at  $\theta_w = q_w$ . It follows that the maximizer  $\hat{\theta}_t^*$  is equal to  $q$ .  $\square$

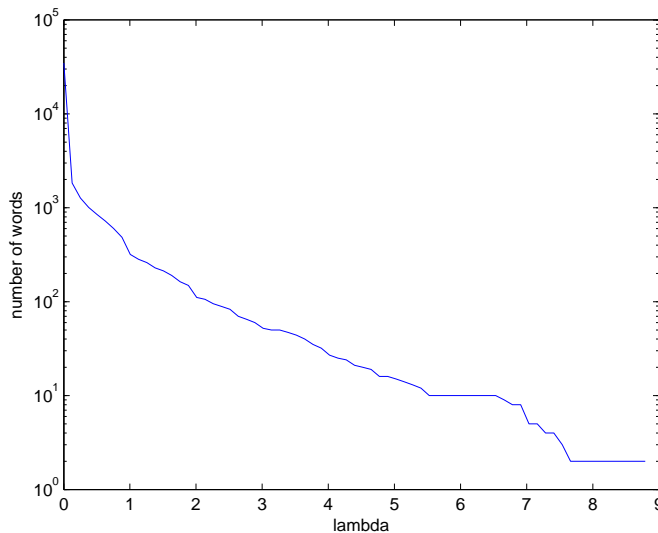


Figure 12: Number of words whose penalized estimator  $[\hat{\theta}_t^*]_w$  deviates from the global model  $q_w$  as a function of  $\lambda$ . The log-scale on the  $y$ -axis indicates very fast drop.

**Proposition 4.** *The estimator  $\hat{\theta}_t^*$  satisfies*

$$\hat{\theta}_t^* \in D = \{(\theta_1, \dots, \theta_k) : \theta_w \leq q_w \text{ for all } w \text{ that } c_w \leq \lambda_w\}$$

and  $\ell_t^*(\theta)$  is a concave function on that domain.

*Proof.* If the condition for lemma 1 is satisfied,  $D$  is reduced to a single point set  $\{q_w\}$  and the result follows directly from lemma 1.

Otherwise, there exists  $i$  such that  $c_i > \lambda_i$ . For any  $w$  that  $c_w \leq \lambda_w$ , consider the directional derivative along  $e_{w,i}$

$$\frac{\partial \ell_t^*}{\partial e_{w,i}} = \frac{c_w - \lambda_w \operatorname{sgn}(\theta_w - q_w)}{\theta_w} - \frac{c_i - \lambda_i \operatorname{sgn}(\theta_i - q_i)}{\theta_i}. \quad (24)$$

When  $\theta_w > q_w$ , the first term is

$$\frac{c_w - \lambda_w \operatorname{sgn}(\theta_w - q_w)}{\theta_w} = \frac{c_w - \lambda_w}{\theta_w} \leq 0.$$

The second term is always positive regardless of the sign of  $\theta_i - q_i$ . So  $\partial \ell_t^* / \partial e_{w,i} < 0$  for  $\theta_w > q_w$ , which means that  $\ell_t^*$  is monotonically decreasing along  $e_{w,i}$  which contradicts a possibility that for the maximizer  $\theta_w^* > q_w$ .

To prove the concavity of  $\ell(\theta)$  on  $D$ , notice  $\ell_t^*$  is a sum of the following terms

$$\ell_w(\theta_w) = c_w \log \theta_w - \lambda_w |\log \theta_w - \log q_w|. \quad (25)$$

By examining the derivatives of the terms above it can be shown that they are concave on  $D$  for all  $w$  resulting in the concavity of  $\ell_t^*(\theta)$ .  $\square$

As a result, we can maximize Equation (23) by analyzing the derivatives in a similar fashion to the Karush-Kuhn-Tucker condition. A special treatment is needed since  $\ell$  is only piece-wise smooth.

**Lemma 2.** *A vector  $v = (v_1, \dots, v_k)$  satisfying  $\sum_i v_i = 0$  can be decomposed as*

$$v = \sum_{(i,j) \in C} d_{i,j} e_{i,j} \quad (26)$$

for a subset  $C$  of all  $(i, j)$  pairs. In addition the decomposition satisfies  $d_{i,j} > 0$  and  $v_i > 0, v_j < 0$  for all  $(i, j) \in C$ .

*Proof.* We prove by induction in size of support of  $v$ ,  $\text{supp}(v) = \{i : v_i \neq 0\}$ . The lemma is trivial when  $\text{supp}(v)$  is empty. For a general  $v$ , let  $v_m$  be the component with smallest non-zero absolute value. Since  $\sum_i v_i = 0$ , there exists a  $v_n$  with the opposite sign. Let

$$u = \begin{cases} v - v_m e_{m,n} & \text{if } v_m > 0, \\ v + v_m e_{n,m} & \text{if } v_m < 0 \end{cases}. \quad (27)$$

Since  $|v_n| \geq |v_m|$  in both cases above  $u$  is a vector not disagreeing with  $v$  in sign.  $u$  has a smaller support than  $v$  and also satisfies  $\sum u_i = 0$ , thus by induction we have the decomposition  $u = \sum_{(i,j) \in C'} d_{i,j} e_{i,j}$ . Plugging it back into (27),  $v$  can be decomposed with one more term as

$$v = \begin{cases} v_m e_{m,n} + \sum_{(i,j) \in C'} d_{i,j} e_{i,j} & \text{if } v_m > 0, \\ -v_m e_{m,n} + \sum_{(i,j) \in C'} d_{i,j} e_{i,j} & \text{if } v_m < 0 \end{cases}$$

satisfying the decomposition requirements and proving the lemma.  $\square$

**Lemma 3.** *For vectors  $u$  and  $v$  for which  $\text{sgn}(u_i) \text{sgn}(v_i) \geq 0$  for all  $i$ , we have*

$$\frac{\partial \ell}{\partial(u+v)} = \frac{\partial \ell}{\partial u} + \frac{\partial \ell}{\partial v}$$

*Proof.* The directional derivative of  $\ell_w(\theta)$  as defined in (25) is

$$\frac{\partial \ell_w(\theta)}{\partial v} = v_w \frac{c_w - \lambda_w s_w(v_w)}{\theta_w}$$

where

$$s_w(v_w) = \begin{cases} \text{sgn}(\theta_w - q_w) & \theta_w \neq q_w \\ \text{sgn}(v_w) & \theta_w = q_w \end{cases}.$$

By inspection, it can now be shown that

$$\frac{\partial \ell_w}{\partial(u+v)} = \frac{\partial \ell_w}{\partial u} + \frac{\partial \ell_w}{\partial v} \quad (28)$$

and since  $\partial \ell_i^* / \partial v = \sum_w \partial \ell_w / \partial v$  for any vector  $v$ , we can sum (28) over all  $w$  to get the result.  $\square$

**Proposition 5.**  $\hat{\theta}^*$  is the maximizer of (23) if and only if there exists  $\beta > 0$  such that for all  $w$  such that  $\hat{\theta}_w^* \neq q_w$ ,

$$\forall w : \hat{\theta}_w^* \neq q_w \quad \beta = \frac{c_w - \lambda_w \operatorname{sgn}(\hat{\theta}_w^* - q_w)}{\hat{\theta}_w^*} \quad (29)$$

$$\forall w : \hat{\theta}_w^* = q_w \quad \frac{c_w - \lambda_w}{\hat{\theta}_w^*} \leq \beta \leq \frac{c_w + \lambda_w}{\hat{\theta}_w^*}. \quad (30)$$

*Proof.* “ $\Rightarrow$ ”: The one sided directional derivative

$$\frac{\partial \ell}{\partial e_{i,j}} = \begin{cases} \frac{c_i + \lambda_i}{\theta_i} & \theta_i < q_i \\ \frac{c_i - \lambda_i}{\theta_i} & \theta_i \geq q_i \end{cases} - \begin{cases} \frac{c_j + \lambda_j}{\theta_j} & \theta_j \leq q_j \\ \frac{c_j - \lambda_j}{\theta_j} & \theta_j > q_j \end{cases} \quad (31)$$

should be non-positive for all  $i, j$ . We prove (29) and (30) by examining two cases.

If  $\hat{\theta}^* \neq q$ , Consider all  $(i, j)$  pairs such that  $\hat{\theta}_i^* \neq q_i$  and  $\hat{\theta}_j^* \neq q_j$ . Note such  $(i, j)$  must exist since  $\hat{\theta}^*$  is different from  $q$  in at least two indexes, otherwise they can not both sum up to 1. We can verify that  $\partial \ell_t^*(\hat{\theta}^*) / \partial e_{i,j} = -\partial \ell_t^*(\hat{\theta}^*) / \partial e_{j,i}$ , so both of them should be 0 leading to

$$\frac{c_i - \lambda_i \operatorname{sgn}(\hat{\theta}_i^* - q_i)}{\hat{\theta}_i^*} = \frac{c_j - \lambda_j \operatorname{sgn}(\hat{\theta}_j^* - q_j)}{\hat{\theta}_j^*} = \beta.$$

and to (29). We have  $\hat{\theta}_i^* < q_i$  for at least some  $i$ , thus  $\beta = (c_i + \lambda_i) / \hat{\theta}_i^* > 0$ . For  $w$  such that  $\hat{\theta}_w^* = q_w$  and  $j$  such that  $\hat{\theta}_j^* \neq q_j$  we can solve the inequality  $\partial \ell / \partial e_{w,j} \leq 0$  to obtain (30).

In case  $\hat{\theta}^* = q$ , solving  $\partial \ell / \partial e_{i,j} \leq 0$  gives us

$$\frac{c_i - \lambda_i}{\theta_i} \leq \frac{c_j + \lambda_j}{\theta_j}$$

for all  $i, j$  pairs. So such  $\beta$  still exists but is not necessarily unique.

“ $\Leftarrow$ ”: We only need to prove  $\hat{\theta}^*$  is a local maximum in  $D$ . Then by Proposition 4,  $\hat{\theta}^*$  is the global maximizer of (23).

To prove  $\hat{\theta}^*$  is a local maximum, it is sufficient to show that for a vector  $v$  pointing to a direction in the simplex  $\mathbb{P}_V$ ,  $\partial \ell(\hat{\theta}^*) / \partial v \leq 0$ . We can check it is true for  $v = e_{i,j}$  by (29) and (30). In a general case, we decompose  $v = \sum_{(i,j) \in C} d_{i,j} e_{i,j}$  according to lemma 2. It also affirms each  $e_{i,j}$  in the decomposition does not disagree with  $v$  in sign, so we can apply lemma 3 to get the result. To show that  $\hat{\theta}^* \in D$ , from (29), for any  $c_w < \lambda_w$ ,

$$c_w - \lambda_w \operatorname{sgn}(\hat{\theta}_w^* - q_w) = \beta \hat{\theta}_w^* > 0$$

so  $\operatorname{sgn}(\hat{\theta}_w^* - q_w) \leq 0$  and  $\hat{\theta}_w^* \leq q_w$ . □

The following corollary rephrases proposition 5 in a convenient form that leads to the algorithm in figure 13, which efficiently obtains  $\hat{\theta}^*$ .

**Corollary 4.** Denote

$$\begin{aligned} A^+ &= \{w : \theta_w^* > q_w\} \\ A^- &= \{w : \theta_w^* < q_w\} \\ A^0 &= \{w : \theta_w^* = q_w\} \\ n_w^- &= (c_w - \lambda_w)/q_w \\ n_w^+ &= (c_w + \lambda_w)/q_w. \end{aligned}$$

$\hat{\theta}^*$  is the maximizer of (23) if and only if there exists  $\beta > 0$  such that:

- $\beta < n_w^-$  for  $w \in A^+$ .
- $n_w^- \leq \beta \leq n_w^+$  for  $w \in A^0$ .
- $n_w^+ < \beta$  for  $w \in A^-$ .

and when  $A^+ \cup A^- \neq \emptyset$

$$\beta = \frac{\sum_{w \in A^+} (c_w - \lambda_w) + \sum_{w \in A^-} (c_w + \lambda_w)}{\sum_{w \in A^+ \cup A^-} q_w}. \quad (32)$$

Moreover,  $\hat{\theta}^*$  can be calculated by

$$\hat{\theta}_w^* = \begin{cases} (c_w - \lambda_w)/\beta & w \in A^+ \\ q_w & w \in A^0 \\ (c_w + \lambda_w)/\beta & w \in A^- \end{cases}. \quad (33)$$

*Proof.* The inequalities on  $\beta$  and (33) come directly from equation (29) and (30). To show (32), from (29) we have

$$c_w - \lambda_w \operatorname{sgn}(\hat{\theta}_w^* - q_w) = \beta \theta_w^*$$

sum up this equation for all  $w \in A^+ \cup A^-$ , and note  $\sum_{w \in A^+ \cup A^-} \hat{\theta}_w^* = \sum_{w \in A^+ \cup A^-} q_w$ .  $\square$

Note that the algorithm computes  $\beta$  in addition to  $\hat{\theta}^*$ . In each iteration, a maximum and minimum is examined from the set  $A_i^0$  which decreases gradually, thus can be implemented effectively as a priority queue. The time complexity of the algorithm is  $O(k + (k - |A^0|) \log k)$ . The following proposition proves the correctness of this algorithm.

**Proposition 6.** The algorithm in figure 13 calculates  $\hat{\theta}^*$  that maximizes the penalized likelihood function (23).

*Proof.* Step 1 in the algorithm checks the condition if  $\hat{\theta}^* = q$  and stops if so. Otherwise, from corollary 4,  $\beta < n_w^-$  for all  $w \in A^+$ . Since  $A^+$  is not empty,  $\beta < \max n_w^-$ , hence  $\arg \max n_w^- \in A^+$ . Similarly  $\arg \min n_w^+ \in A^-$ . This get us started with  $A_1^-$  and  $A_1^+$  in step 2.

Knowing  $A_i^- \subset A^-$  and  $A_i^+ \subset A^+$ , the algorithm expands them in step 5 and 6, until both hit the target. Let  $v = \arg \max_{w \in A_i^0} n_w^-$ , we want to show the following to justify step 6: if  $n_v^- > \beta_i$ , then  $n_v^- > \beta$ , thus  $v \in A^+ \setminus A_i^+$ . Step 5 follows a similar argument.

1. If  $\min n_w^+ \geq \max n_w^-$ , stop with  $\hat{\theta}^* = q$ .
2. Let  $A_1^- = \{\arg \min n_w^+\}$  and  $A_1^+ = \{\arg \max n_w^-\}$ .
3. Begin loop at step  $i = 1$ .
4. Calculate  $\beta_i$  from (32) using the current  $A_i^+$  and  $A_i^-$ .
5. If  $\min_{w \in A_i^0} n_w^+ < \beta_i$ , then let  $A_{i+1}^- = A_i^- \cup \{\arg \min_{w \in A_i^0} n_w^+\}$ , otherwise  $A_{i+1}^- = A_i^-$ .
6. If  $\max_{w \in A_i^0} n_w^- > \beta_i$ , then let  $A_{i+1}^+ = A_i^+ \cup \{\arg \max_{w \in A_i^0} n_w^-\}$ , otherwise  $A_{i+1}^+ = A_i^+$ .
7. If either  $A_{i+1}^+ \neq A_i^+$  or  $A_{i+1}^- \neq A_i^-$ , let  $i \leftarrow i+1$  and loop back to step 4. Otherwise terminate the algorithm with  $\hat{\theta}^*$  calculated from (33).

Figure 13: Algorithm for solving  $\hat{\theta}^*$ .

$\beta$  can be decomposed as

$$\beta = \frac{\sum_{w \in A^+} (c_w - \lambda_w) + \sum_{w \in A^-} (c_w + \lambda_w)}{\sum_{w \in A^+ \cup A^-} q_w} \quad (34)$$

$$= \xi_0 \beta_i + \sum_{w \in A^+ \setminus A_i^+} \xi_w n_w^- + \sum_{w \in A^- \setminus A_i^-} \xi_w n_w^+ \quad (35)$$

and  $\xi_0$  and  $\xi_w$  are positive coefficients that sum up to 1:

$$\xi_0 = \frac{\sum_{w \in A_i^+ \cup A_i^-} q_w}{\sum_{w \in A^+ \cup A^-} q_w}$$

$$\xi_w = q_w / \sum_{w \in A^+ \cup A^-} q_w.$$

We use (35) in the following way. To prove  $x \geq \beta$ , it is sufficient to show  $x \geq \beta_i$ ,  $x \geq n_w^-$  for  $w \in A^+ \setminus A_i^+$  and  $x \geq n_w^+$  for  $w \in A^- \setminus A_i^-$ . The inequality is strict if one of the conditions is strict. Let  $u = \arg \max_{w \in A^- \setminus A_i^-} n_w^+$ . If  $n_u^+ > n_v^-$ , then  $n_u^+ > \beta_i$ ,  $n_u^+ > n_v^- \geq n_w^-$  for all  $w \in A_i^0$  which includes  $A^+ \setminus A_i^+$ , and  $n_u^+ \geq n_w^+$  for all  $w \in A^- \setminus A_i^-$  by the definition of  $u$ . Thus from the decomposition of  $\beta$  in (35),  $n_u^+ > \beta$ . This contradicts corollary 4 since  $u \in A^-$ . Hence  $n_v^- \geq n_u^+ \geq n_w^+$  for all  $w \in A^- \setminus A_i^-$ . Combining this with  $n_v^- > \beta_i$  and  $n_v^- \geq n_w^-$  for all  $w \in A^+ \setminus A_i^+$ , we can use (35) again to show  $n_v^- > \beta$ . Thus from corollary 4 we conclude that  $v \in A^+$ .

When the algorithm ends,  $n_w^- \leq \beta_i \leq n_w^+$  for all  $w \in A_i^0$ , thus all the conditions for corollary 4 are met for  $\beta = \beta_i$ .  $\square$

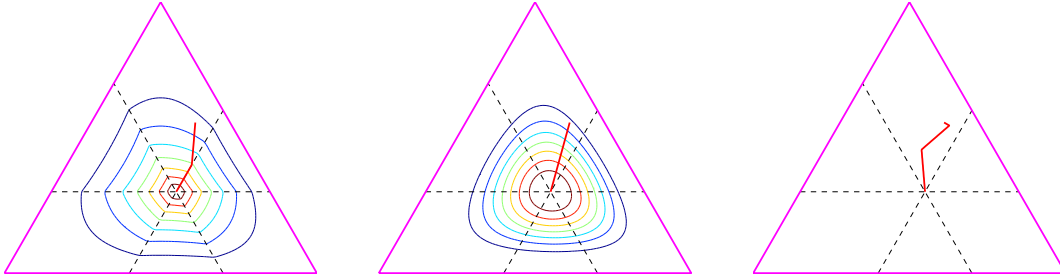


Figure 14: Different regularization methods in the low dimensional case  $|V| = 3$ . The parameter space is a 2-simplex, with the lower left corner corresponds to word 1, lower right corner corresponds to word 2, top corner corresponds to word 3. The dashed lines intersect at the crossing point  $q = (.3 .4 .3)$ . A solution path for a count vector  $c = (1 3 5)$  is shown with varying amount of regularization. **Left:** The solution path  $\hat{\theta}^*$  for different values of  $\lambda = \lambda_w$  displayed on a contour plot of the prior  $p(\theta)$  in (36) with  $\lambda_w = 1$ . **Middle:** Linear interpolation estimate on a contour plot of Dirichlet distribution, with mode at  $q$ . **Right:** Solution path of absolute discounting with changing  $d$ .

## 5.2 Bayesian Interpretation

Like other penalized likelihood methods, there is a Bayesian interpretation for  $\hat{\theta}^*$  as the maximum posterior estimate with the prior

$$p(\theta; \lambda, q) = \prod_{i=1}^k \left( \frac{\theta_w}{q_w} \right)^{-\lambda_w \text{sgn}(\theta_w - q_w)}. \quad (36)$$

The mode of this distribution is at  $q$ , and the density decays at the boundary of the simplex  $\mathbb{P}_V$ . It has several non-smooth “ridges” when  $\theta_w = q_w$  for some  $w$ . An illustration of a low-dimensional case is shown in Figure 14 (left).

## 5.3 Comparison with other regularization methods

There are other regularization methods that use a prior or “back-off” distribution. One class of methods is linear interpolation, which includes additive smoothing, Jelinek-Mercer smoothing and Witten-Bell smoothing. It is also a Bayesian estimate with Dirichlet prior distribution. Another class is absolute discounting which includes Kneser-Ney smoothing as a special case. The differences among these methods are mainly in the choice of the smoothing parameters. In this section we compare our penalized local likelihood with some alternative methods from these classes.

Linear interpolation methods estimate the probability as

$$\theta_w^{LI} = (1 - \eta) \frac{c_w}{N} + \eta q_w \quad (37)$$

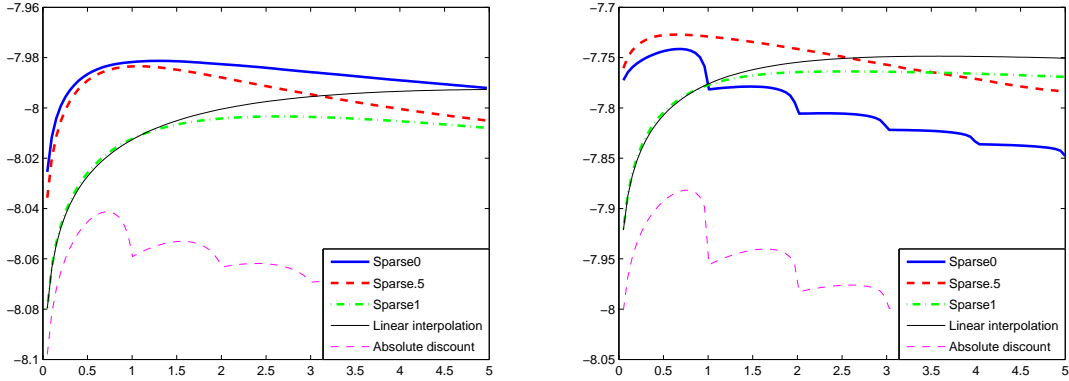


Figure 15: Left: Cross entropy of true probabilities and estimated values for different regularization methods on simulated data.  $q$  is equal to  $\theta$  in 65% of words. Right: Log likelihood for different methods on a held-out test set. Data are news stories from one day in the RCV1 corpus. Experiment is repeated multiple times to remove noise.

where  $0 \leq \eta \leq 1$  is an interpolation hyper-parameter. For  $\eta = \lambda/(N + \lambda)$ ,  $\theta_w^{LI}$  is the maximizer of

$$\ell(\theta) = \sum_w (c_w \log \theta_w + \lambda q_w \log \theta_w) \quad (38)$$

$$= \sum_w c_w \log \theta_w - \lambda q_w (\log q_w - \log \theta_w) \quad (39)$$

The latter term is the Kullback-Leibler divergence between  $\theta$  and  $q$ . Although the sum is always non-negative, the individual items are not. That means deviation of  $\theta_w$  from  $q_w$  will cancel each other for some  $w$ . It is interesting to note that replacing the parenthesis with absolute value, (39) becomes precisely our penalized likelihood function (23) with  $\lambda_w = \lambda q_w$ .

Absolute discounting methods estimate the probability as

$$\theta_w^{AD} = \frac{\max\{c_w - d, 0\}}{N} + r q_w \quad (40)$$

where  $r$  is chosen to make the distribution sum to 1. The hyper-parameter  $d$  is usually chosen between 0 and 1. By taking  $\lambda_w = d$  in our method, every word in the set  $A^+$  has an effective count  $c_w - d$ .

We compared 3 variants of our estimator with linear interpolation and absolute discounting. Specifically, we examined  $\lambda_w = \lambda q_w$  (**Sparse1**),  $\lambda_w = d$  (**Sparse0**) and  $\lambda_w = \lambda \sqrt{q_w}$  (**Sparse.5**).

In figure 15, the performance of the 5 methods are shown for simulated and real world data. The horizontal axis is regularization strength, which is  $d$  for absolute discounting and **Sparse0**. For other 2 variants, they are aligned so that  $\sum_w \lambda_w$  are the same. Linear interpolation is aligned with **Sparse1** for the similarity of penalty term. When  $q$  is equal to the true  $\theta$  for many words, the penalized local likelihood estimators **Sparse0** and **Sparse.5** outperform the other methods. When  $q$  is not such a good guess, the penalized local likelihood estimators perform as competitive as linear interpolation. Absolute discounting generally performed poorly in these experiments.

## 5.4 Choice of $\lambda_w$

We concentrate in this subsection on automatic selection of the regularization parameter in the case  $\lambda_w = \lambda\sqrt{q_w}$  which performed very well in our experiments.

An often used scheme is to choose  $\lambda$  with the highest cross-validation score. However, it is often computationally expensive to calculate it for all  $t$ . We propose an approximate leave-one-out cross-validation score that is simpler to compute and has a nearly identical performance. Under the assumption that all the words are drawn independently from a multinomial distribution, the leave-one-out cross-validation score is

$$s(\lambda; c) = \sum_{w \in V} c_w \log \theta_w^*(c_w^{-1}; \lambda) \quad (41)$$

where  $c_w^{-1} = (c_1, \dots, c_w - 1, \dots, c_k)$  is the count vector with word  $w$  removed. Since it may not be practical to calculate  $\theta_w^*(c_w^{-1}; \lambda)$  for all  $w \in V$  we propose to approximate  $\hat{\theta}^*(c_w^{-1}; \lambda)$  base on (33). Instead of using  $\beta(c_w^{-1}; \lambda)$ , we use  $\beta(c, \lambda)$  which is available from the calculation of  $\hat{\theta}^*$ :

$$\tilde{\theta}_w(c_w^{-1}; \lambda) = \begin{cases} (c_w - 1 - \lambda_w) / \beta(c, \lambda) & c_w - 1 - \lambda_w > \beta(c, \lambda) q_w \\ q_w & c_w - 1 - \lambda_w < \beta(c, \lambda) q_w < c_w - 1 + \lambda_w \\ (c_w - 1 + \lambda_w) / \beta(c, \lambda) & c_w - 1 + \lambda_w < \beta(c, \lambda) q_w \end{cases} \quad (42)$$

The difference caused by this approximation is negligible in practice (figure 16 left).

In a simulation study, we generated multinomial counts from a probability vector  $\theta$  on 34803 words.  $\theta$  is chosen to represent word distribution of a typical day. The difference between the log-likelihoods of the cross validation estimator and of the estimator with the best possible  $\lambda$  is shown in figure 16 (right). Most of the time the cross validation estimator will lose less than 0.001 in the log-likelihood and with more words, the difference diminishes. Combining the two graphs above we conclude that the approximate cross validation estimator provides a computationally efficient and statistically accurate automatic selection of the regularization parameter  $\lambda$ .

## 6 Related Work

Concept drift or similar phenomena under different names have been studied in a number of communities. It has recently gained interest primarily due to an increase in the need to model large scale temporal data streams.

Early machine learning literature on the concept drift problem involved mostly computational learning theory tools (Helmbold and Long, 1994; Kuh et al., 1991). Hulten et al. (2001) studied the problem in the context of datamining large scale streams whose distribution change in time. More recently, Forman (2006) studied the concept drift phenomenon in the context of information retrieval in large textual databases. Sharan and Neville (2007) consider the modeling of temporal changes in relational databases and its application to text classification. Overall, the prevailing techniques have been to train standard methods on examples obtained by filtering the data through a sliding window. Our approach generalizes this perspective as a sliding window is equivalent to a constant kernel. On the other hand, using non-constant kernels allow higher statistical accuracy. The local likelihood framework also provides a more formal and systematic analysis of the estimation accuracy.

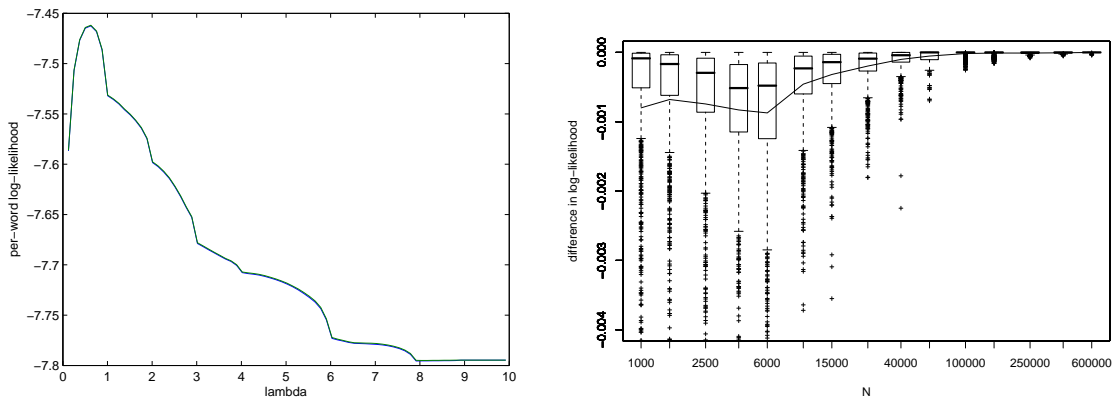


Figure 16: **Left:** Using documents from one day in the RCV1 corpus, the approximated leave-one-out cross validation score approximates the usual leave-one-out score with such a high accuracy that the two curves overlap each other. **Right:** In a simulation study, the difference of per-word log-likelihood of leave-one-out  $\lambda$  and best possible  $\lambda$  is very small. The line shows mean difference, which is typically on the scale of  $10^{-3}$ .

In statistics, the ideas of local likelihood were developed by Tibshirani and Hastie (1987) within the context of non-parametric smoothing and regression. More details on local likelihood can be found in the monographs (Loader, 1999; Wand and Jones, 1995).

## 7 Discussion

A large number of textual datasets such as emails, webpages, news stories, etc. contain time stamped documents. For such datasets, considering a drifting rather than a stationary distribution is often appropriate. The local likelihood framework provides a natural extension for many standard likelihood models to the concept drift scenario. As the drift becomes more noticeable and the data size increases the potential benefits of local likelihood methods over their extreme global or local counterparts increase.

In this paper we illustrate the drift phenomenon and examine the properties of the local likelihood estimator including the asymptotic bias and variance tradeoff and optimal bandwidth. Experiments conducted on the RCV1 and AOL datasets demonstrate the framework in practice and contrast the results with the developed theory.

## References

- Airoldi, E. M., Anderson, A., Fienberg, S. E., and Skinner, K. K. (2006). Who wrote Ronald Reagan’s radio addresses? *Bayesian Analysis*, 1:288–320.
- Chen, S. and Goodman, J. (1998). An empirical study of smoothing techniques for language modelling. Technical report, Harvard university.
- DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer.

- Forman, G. (2006). Tackling concept drift by temporal inductive transfer. In *Proc. of the ACM SIGIR Conference*.
- Helmbold, D. P. and Long, P. M. (1994). Tracking drifting concepts by minimizing disagreements. *Machine Learning*, 14.
- Hulten, G., Spencer, L., and Domingos, P. (2001). Mining time-changing data streams. In *Proc. of the ACM SIGKDD Conference*.
- Kuh, A., Petsche, T., and Rivest, R. L. (1991). Learning time-varying concepts. In *Advances in Neural Information Processing Systems*, 3.
- Lebanon, G. and Mao, Y. (2008). Non-parametric modeling of partially ranked data. In *Advances in Neural Information Processing Systems 20*. The MIT Press.
- Lewis, D., Yang, Y., Rose, T., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.
- Loader, C. (1999). *Local Regression and Likelihood*. Springer.
- Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Mosteller, F. and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*. Addison Wesley.
- Mulligan, G. (1999). *Removing the Spam: Email Processing and Filtering*. Addison Wesley.
- Pang, B. and Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of the Association of Computational Linguistics*.
- Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationship for sentiment categorization with respect to rating scales. In *Proc. of the Association of Computational Linguistics*.
- Pass, G., Chowdhury, A., and Torgeson, C. (2006). A picture of search. In *The First International Conference on Scalable Information Systems*.
- Sharan, U. and Neville, J. (2007). Exploiting time varying relationships in statistical relational models. In *Proc. of the Joint 9th WebKDD and 1st SNA-KDD Workshops*.
- Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82.
- Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall/CRC.
- Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1:69–90.

## A Description of Datasets

In this paper we conducted experiments on the Reuters RCV1 dataset and the AOL query-log dataset. A description of these datasets is found below. For more information see (Lewis et al., 2004) and (Pass et al., 2006).

### A.1 RCV1 Dataset

Reuters Corpus Volume I (RCV1) contains over 800,000 news stories which are provided by Reuters, Ltd. for research purposes. Over 11,000 stories are produced a day by Reuters' editorial division and made available via online databases and other archival products. RCV1 is drawn from these online databases. It consists of all English language stories produced by Reuters journalists spanning 365 days between August 20, 1996, and August 19, 1997. The stories have been formatted in XML and vary from a few hundred to several thousand words in length. The dataset is categorized across three dimensions: topics, industries, and regions. Special topic codes were assigned to describe the major subjects of a story.

In our experiments, the RCV1 dataset is pre-processed as follows. First the xml/html tags are removed and non-alphabetic characters (including numbers) are removed. Then all words are lowercased and stemmed while stopwords are discarded. The single character words are removed since generally they are meaningless. Lastly the vocabulary was built ignoring terms appearing less than  $k$  ( $k = 5$  in the experiments) times.

### A.2 AOL Dataset

The AOL search query log dataset, which was provided by AOL for non-commercial research use, contains about 20 million web queries from 650,000 users. It records all English language web queries from users over 3 months between March 01, 2006, and May 31, 2006. Each query record contains an anonymous user ID number, the query issued by the user, and the time at which the query was submitted for search. If the user clicked on a search result, the rank of the item on which they clicked and the domain portion of the URL in the clicked result are also recorded. The whole dataset contains about 36 million lines of data and near 21 million instances of new queries, where each query averages 3.5 words.

We pre-processed the AOL dataset in a similar manner to the RCV1 dataset: removing non-alphabetic characters, lowercasing and stemming the words, cleaning stopwords, removing single character words, constructing the vocabulary, and removing extremely rare words. The web queries contain a huge amount of web url such as `apple.com` which were parsed as one word `apple.com` rather than two words `apple` and `com`.