Local Likelihood Modeling of the Concept Drift Phenomenon

by

G. Lebanon and Y. Zhao
Purdue University

Technical Report #07-10

Department of Statistics
Purdue University

October 2007

Local Likelihood Modeling of the Concept Drift Phenomenon

Guy Lebanon and Yang Zhao
Department of Statistics
Purdue University – West Lafayette

October 1, 2007

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 temporal document sequences. We examine the asymptotic bias and variance and present an experimental study using the extensive RCV1 dataset of Reuters news stories.

1 Introduction

Time stamped documents such as news stories 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 is generated by a time dependent process $z^{(t)} \sim p_t(z), t \in I \subset \mathbb{R}$ whose approximation $\{\hat{p}_t : t \in I\}$ becomes the main objective of the inference task. We assume that the time t is a continuous quantity, even in cases where the realized time points form a discrete sample. For example, assuming that the time stamps represent the days of the year when the documents were authored, we assume that the set $\{1, \ldots, 365\}$ is a discrete sample from a underlying continuous interval $[1, 365]$. We further assume that the data samples $z^{(t)}$, sampled from p_t, correspond to pairs $z^{(t)} = (x, y)$ constituting a document x and a categorial-valued label y. Such pairs (x, y) appear often in practice, for example with y corresponding to the document topic [14], sentiment [11], author [9] or Email spam/no-spam [10].

Assuming that our data is a set of time stamped documents and labels $(t, (x, y))$, 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 for language modeling which is an important component in the speech recognition and machine translation tasks.

We demonstrate the presence of concept drift in practice by considering the Reuters RCV1 dataset. The Reuters RCV1 dataset, described in [6], contains over 800,000 news stories gathered in a period spanning 365 consecutive days and categorized according to topic. Figure 1 displays the temporal change in the relative frequency (number of appearance in a document divided by document length) of three words: `million`, `common`, and `Handelsgesellschaft` (German trade unions) for documents in the most popular RCV1 category titled `earn`. 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 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 $\{p_t(x, y) : t \in I\}$ based on the the local likelihood principle. Local likelihood is a locally weighted version of the loglikelihood with the weights corresponding to the difference between a reference time point and the

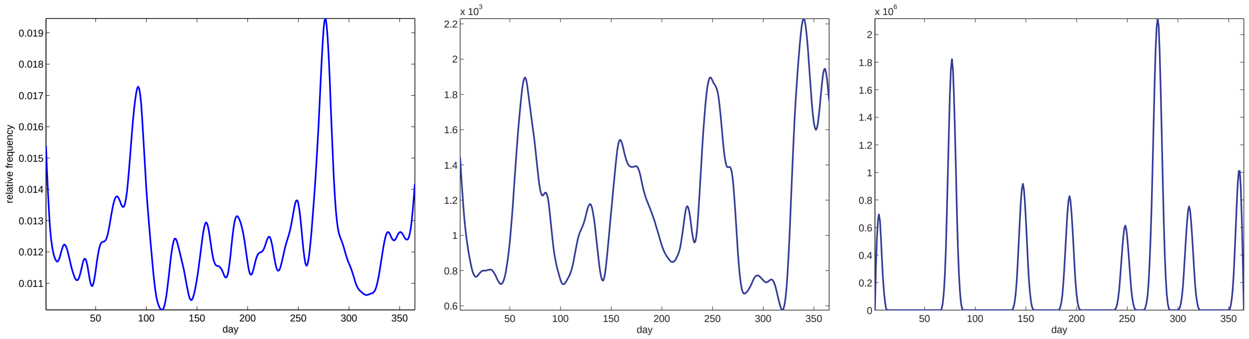

Figure 1: Estimated relative frequency (number of appearance in a document divided by document length) of three words for the most popular category in RCV1 titled `earn` as a function of time. The three panels correspond (left to right) to the words `million`, `common`, and `Handelsgesellschaft` (German trade unions). The displayed curves were smoothed by a triangular kernel of support 11 to remove sampling noise.

time associated with the sampled data. It enjoys nice theoretical properties, in particular convexity if the underlying likelihood is convex and consistency assuming a smooth drift.

After presenting a more formal discussion of concept drift in Section 3 and the definition of local likelihood in Section 4 we turn to examine in detail the case of modeling $p_t(x|y)$ with local naive Bayes and $p_t(y|x)$ with local logistic regression. In the case of the naive Bayes model, we provide a precise as well as asymptotic description of the bias and variance quantities which illuminates certain facts concerning the selection of weights and the difference between online and offline scenarios. Experiments conducted on the RCV1 dataset demonstrates the local likelihood estimation in practice and contrasts it with more standard non-local alternatives.

2 Related Work

Concept drift or similar problems 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 data streams. Early machine learning literature on the concept drift problem involved mostly computational learning theory tools [3, 5]. Hulten et al. [4] studied the problem in the context of datamining large scale streams whose distribution change in time. More recently the concept drift phenomenon has been demonstrated and studied in the context of information retrieval in large textual databases [2]. Overall, the prevailing techniques have been to train standard methods on examples obtained by filtering the data through a sliding window.

In the statistics community, local likelihood was developed by Tibshirani and Hastie [12] and is studied in the monograph [7]. Its origin is in non-parametric smoothing and regression which are now a considerable research area in statistics. Furthermore, as we show in Section 4.1 local likelihood for the naive Bayes model is very similar to the Nadaraya-Watson estimator for non-parametric regression.

3 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 $I \subset \mathbb{R}$ to a family of joint distributions*

$$t \mapsto p_t(x,y) \stackrel{\text{def}}{=} p_t(X=x, Y=y).$$

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

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

we have that Definition 1 is equivalent to a smooth parameterized curve in the simplex $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$. The curve can be restricted to lie in a submanifold $M \subset \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ if parametric assumptions are made or otherwise remain unrestricted.

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 (in most cases involving text documents $|\mathcal{Y}| \ll |\mathcal{X}|$). 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 probability of obtaining two samples from the same time point is 0. In practice, however, time is always discretized at some granularity and we often obtain multiple data points corresponding to the same time. We therefore allow the number of samples from time t, denoted by N_t, to be possibly larger than 1. Summarizing the data generation process, we have

$$D = \{(x_{tj}, y_{tj}) : t \in I' \subset I \subset \mathbb{R}, j = 1, \ldots, N_t\} \qquad t \sim g(t), \quad (x_{tj}, y_{tj}) \sim_{iid} p_t(x,y). \tag{2}$$

To illustrate these concepts in the context of the RCV1 dataset, we display in Figure 2 the total number of words per day (left) and the total number of documents per day (right) corresponding to the `earn` category. As is evident from the right panel, $g(t)$ is a highly non-uniform density corresponding to varying amount of news content in different dates.

It is easy to come up with two simple solutions to the problem of concept drift modeling. The first solution, called 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 $\{p_t, t \in I\}$ effectively modeling the concept drift as a degenerate curve in the simplex. The second simple alternative, called the extreme local model, is to model p_t using only data sampled from time t: $\{(x_{tj}, y_{tj}) : j = 1, \ldots, N_t\}$. 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 points 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

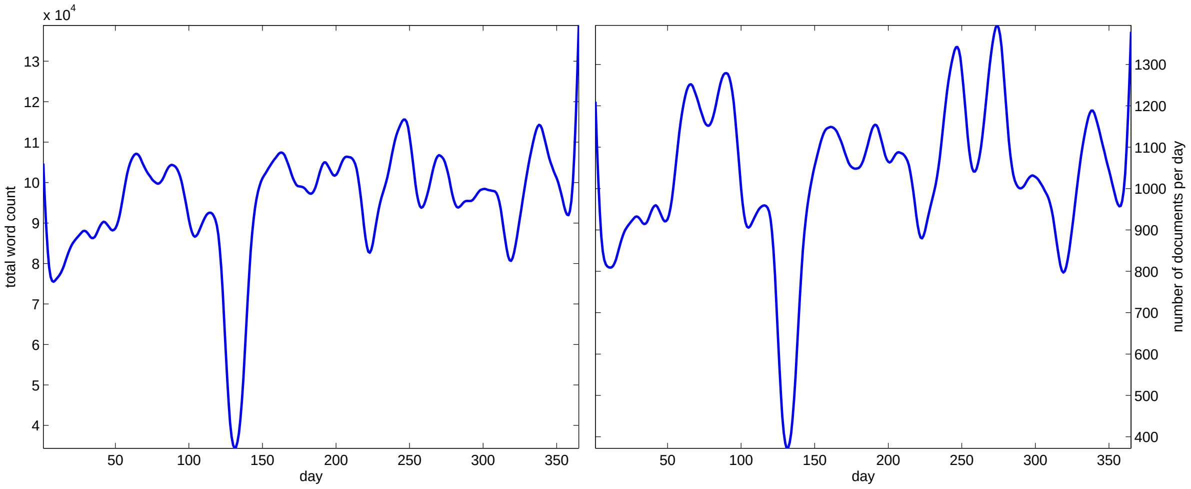

Figure 2: Total number of words per day (left) and documents per day (right) for the most popular category in RCV1 titled `earn`. The displayed curves were smoothed by a triangular kernel of support 11 to remove sampling noise.

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 ϵ increases.

4 Local Likelihood and Concept Drift Estimation

The local likelihood principle extends the ideas of non-parametric regression smoothing and density estimation to likelihood-based inference. We concentrate on using the local likelihood principle for estimating $p_t(x|y)$ and $p_t(y|x)$ which are described in the two following subsections.

4.1 Local Likelihood for Naive Bayes and n-Grams

We apply local likelihood to the problem of estimating $p_t(x|y)$ by assuming the naive Bayes assumption i.e. that $x|y$ is generated by a multinomial distribution or its n-gram extensions. Assuming documents contain words belonging to a finite dictionary of size V, the naive Bayes assumption may be stated as

$$p_t(x|y) \propto \prod_{w \in V} \theta_w^{c(w,x)}, \qquad \theta \in \mathbb{P}_V \tag{3}$$

where $c(w,x)$ represents the number of times word w appear in document x. Similarly, the n-gram model extends naive Bayes (3) by considering n-order Markov dependency. For example, in the case of $n=3$ we have

$$p_t(x|y) \propto \prod_{(w,u,v)\in V^3} \theta_{(w,u,v)}^{c((w,u,v),x)}, \qquad \theta \in \mathbb{P}_{V^3} \tag{4}$$

where $c((w,u,v),x)$ denotes the number of times the words w,u,v appear as an ordered sequence in the document x. The naive Bayes and n-gram are a mainstay of statistical text processing [8] and usually lead to accurate language modeling, especially when appropriate smoothing is used [1]. For notational simplicity we consider the problem of estimating $p_t(x)$ rather than the equivalent $p_t(x|y)$ and we concentrate on the naive Bayes model. Extending the discussion in this section to n-grams is relatively straightforward.

Applied to the concept drift problem, the local log-likelihood at time t is a smoothed version of the loglikelihood of (2) with the amount of smoothing determined by a non-negative kernel $K_h : \mathbb{R} \to \mathbb{R}$

$$\ell_t(\eta|D) \overset{\text{def}}{=} \sum_{\tau\in I'} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log p(x_{\tau j}\,;\eta) \qquad\qquad \eta\in\Theta, t\in\mathbb{R}. \tag{5}$$

We assume that the kernel function is a normalized density concentrated around 0 and parameterized by a scale parameter $h > 0$ that satisfies $K_h(r) = h^{-1}K(r/h)$ for some base function K. We further assume that K has bounded support and $\int u^r K(u)\,du < \infty$ for $r \leq 2$.

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 $\lim_{h\to 0} K_h$ approaches Dirac's delta function and $\lim_{h\to\infty} K_h$ 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 3 as h ranges from 0 to ∞.

Solving the maximum local likelihood problem for each t provides an estimation of the entire drift $\{\hat\theta_t : t\in\mathbb{R}\}$ with $\hat\theta_t = \arg\max_{\eta\in\Theta} \ell_t(\eta|D)$. 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 many other cases, however, there is no closed form maximizer in general and iterative optimization algorithms are needed in order to obtain $\hat\theta_t = \arg\max_{\eta\in\Theta} \ell_t(\eta|D)$ for all t.

We denote the length of a document in (2) by $|x_{tj}| \overset{\text{def}}{=} \sum_{v\in V} c(x_{tj}, v)$ and the total number of words in day t in (2) by $|x_t| \overset{\text{def}}{=} \sum_{j=1}^{N_t} |x_{tj}| = \sum_{v\in V}\sum_{j=1}^{N_t} c(v, x_{tj})$. We assume that the length of documents x_{tj} is independent of t and is drawn from a distribution with expectation λ. Under the above assumptions, the local likelihood (5) of the naive Bayes model becomes

$$\ell_t(\eta|D) = \sum_{\tau\in I'} K_h(t-\tau) \sum_{j=1}^{N_\tau}\sum_{w\in V} c(w, x_{\tau j}) \log \eta_w + c, \quad \eta\in\mathbb{P}_V, \tag{6}$$

and has a single global maximum whose closed form is obtained by setting to 0 the gradient of the Lagrangian

$$0 = \frac{1}{[\hat\theta_t]_w} \sum_{\tau\in I} K_h(t-\tau) \sum_{j=1}^{N_\tau} c(w, x_{\tau j}) + \lambda_w \qquad \Rightarrow \tag{7}$$

$$[\hat\theta_t]_w = \frac{\sum_{\tau\in I} K_h(t-\tau)\sum_{j=1}^{N_\tau} c(w, x_{\tau j})}{\sum_w \sum_{\tau\in I} K_h(t-\tau)\sum_{j=1}^{N_\tau} c(w, x_{\tau j})} = \frac{\sum_{\tau\in I} K_h(t-\tau)\sum_{j=1}^{N_\tau} c(w, x_{\tau j})}{\sum_{\tau\in I} K_h(t-\tau)|x_\tau|}. \tag{8}$$

The estimator $\hat\theta_t$ is a normalized linear combination of word counts where the combination coefficients are determined by the kernel function. Note that $\hat\theta_t$ is different from a Parzen window or kernel smoothing [13] of the relative frequencies $c(w, x_{\tau j})/\sum_{w'} c(w', x_{\tau j})$. Interestingly, identifying $c(w, x_{\tau j})$ in (8) as a sum $\sum Y_j$ of random variables reveals a striking similarity between $\hat\theta_t$ and the Nadaraya-Watson or locally constant regression estimator.

We distinguish between two fundamental scenarios for predicting the drift θ_t.

Offline or batch scenario: The goal is to estimate the drift $\{\theta_t : t\in\mathbb{R}\}$ offline 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 $\hat\theta_t \xrightarrow{p} \theta_t$ as indicated by Proposition 2.

Online or data-stream scenario: The goal is estimate a model for the present distribution θ_t using training data from the present or the past i.e. a dataset whose pairs are sampled from times smaller than t. This corresponds to situations where the data arrives sequentially as a temporal stream and at each

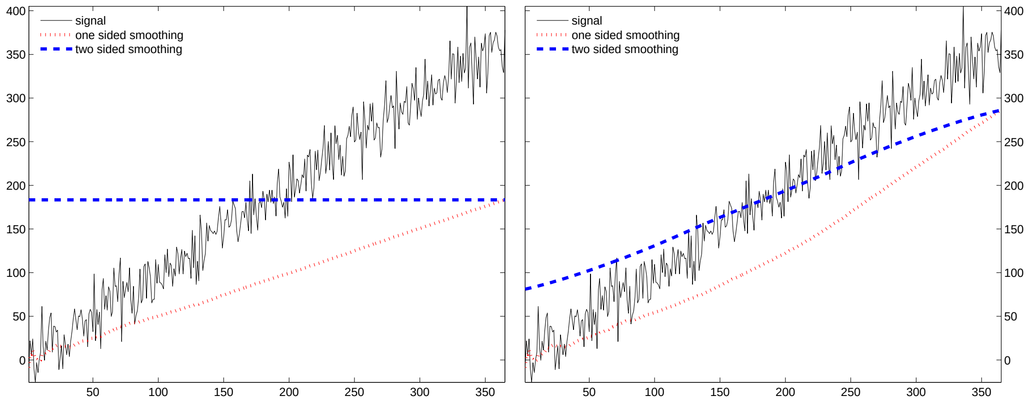

Figure 3: One-sided (online setting) and two-sided (off-line setting) smoothing of a signal $y(x) = x + \epsilon, \epsilon \sim N(0, \sigma^2), \sigma = 20$. Smoothing was performed using a triangular kernel with slope 0 (left) and 1.5 (right). Note how in both plots the online smoother consistently underestimated the linear trend while the offline smoother overestimated it at the beginning and underestimated it at the end.

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 $K(r) = 0, r \leq 0$.

The online task of predicting the present given the past incurs a substantially higher bias than the offline scenario. Figure 3 demonstrates this by displaying offline and online smoothing of a noisy linear signal. The online smoother consistently under-performs the offline smoother as it underestimates the signal in a systematic manner. The added difficulty of estimating the drift using only past data is quantified in Proposition 2.

Proposition 1. *The bias vector and variance matrix of $\hat{\theta}_t$ in (8) are*

$$\mathsf{bias}\,(\hat{\theta}_t) = \mathsf{E}\,\hat{\theta}_t - \theta_t = \frac{\sum_{\tau \in I} K_h(t-\tau)|x_\tau|\,(\theta_\tau - \theta_t)}{\sum_{\tau \in I} K_h(t-\tau)|x_\tau|} \tag{9}$$

$$\mathsf{Var}\,(\hat{\theta}_t) = \frac{\sum_{\tau \in I} K_h^2(t-\tau)|x_\tau|\,(diag(\theta_\tau) - \theta_\tau \theta_\tau^\top)}{\left(\sum_{\tau \in I} K_h(t-\tau)|x_\tau|\right)^2} \tag{10}$$

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

Proof. The random variable $c(w, x_{\tau j})$ 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 $Y \sim \mathrm{Mult}(1, \theta)$, $\mathsf{E}\,Y = \theta$, $\mathsf{Var}\,(\theta) = \mathrm{diag}(\theta) - \theta\theta^\top$. □

Examining Equations (9)-(10) reveals the expected dependency of the bias on h and θ_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 $h \to 0$. Similarly, as for slower drifts, $\|\theta_\tau - \theta_t\|, t \approx \tau$ will decrease and reduce the bias. Despite the relative simplicity of Equations (9)-(10), it is difficult to quantitatively capture the relationship between the bias and variance, the sample size, h, λ, and the smoothness of θ_t, g. Towards this goal we derive below asymptotic expansions of these quantities.

Proposition 2. *Assuming (i) θ, g are smooth in t, (ii) $h \to 0, hn \to \infty$, (iii) $g > 0$ in a neighborhood of t,* *and (iv) document lengths do not depend on t and have expectation λ, we have*

$$\textit{Offline setting}: \quad \mathsf{bias}\,(\hat{\theta}_t|I) = h^2\mu_{21}(K)\left(\dot{\theta}_t\frac{g'(t)}{g(t)} + \frac{1}{2}\ddot{\theta}_t\right) + o_P(h^2) \tag{11}$$

$$\mathsf{Var}\,(\hat{\theta}_t|I) = \frac{\mu_{02}(K)}{(nh)g(t)\lambda}(diag(\theta_t) - \theta_t\theta_t^\top) + o_P((nh)^{-1}) \tag{12}$$

$$\textit{Online setting}: \quad \mathsf{bias}\,(\hat{\theta}_t|I) = h\mu_{11}(K)\dot{\theta}_t + o_P(h) \tag{13}$$

$$\begin{aligned}\mathsf{Var}\,(\hat{\theta}_t|I) = &\left(\frac{\mu_{02}(K)}{nhg(t)\lambda} + \frac{\mu_{12}(K)g'(t)}{ng^2(t)\lambda}\right)(diag(\theta_t) - \theta_t\theta_t^\top)\\ &+ \frac{\mu_{12}(K)}{n\lambda g(t)}(diag(\dot{\theta}_t) - \dot{\theta}_t\theta_t^\top - \theta_t\dot{\theta}^\top) + o_P((nh)^{-1})\end{aligned} \tag{14}$$

where $\mu_{kl}(K) \overset{\text{def}}{=} \int u^k K^l(u)\,du$ is assumed to be finite and $\dot{\theta}_t$ is the vector $[\dot{\theta}_t]_i = \frac{d}{dt}[\theta_t]_i$.

Corollary 1. *Assuming $h \to 0, nh \to \infty$, $\hat{\theta}_t$ is pointwise consistent i.e. $\hat{\theta}_t \xrightarrow{p} \theta_t$ in both the offline and online settings.*

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 (11)-(14). We use below the notation x_{τ_i} to represent the i-training example which is associated with time τ_i.

We expand the denominator and numerator of the bias (9) 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 (11).

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

$$\left(\frac{1}{n}\sum_{i=1}^n K_h(\tau_i)|x_{\tau_i}|\right)^2 \xrightarrow{p} \left(\lambda\int K(u)(g(u) + o(1))\,du\right)^2 = \lambda^2g^2(0) + o(1))^2$$

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

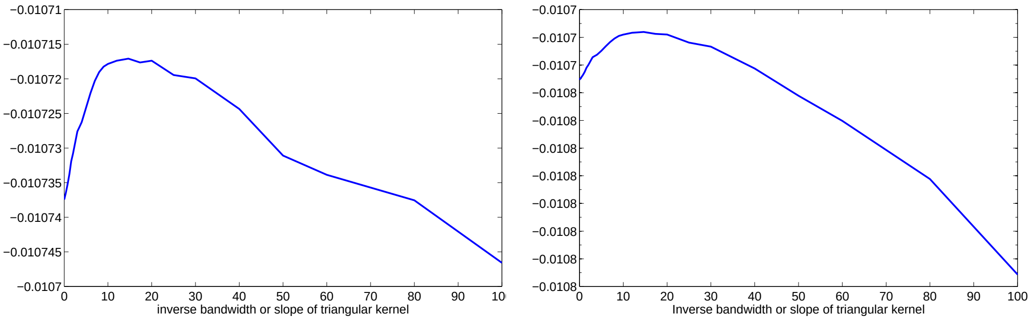

Figure 4: Log-likelihood (per word, computed for most popular 300 words) of held out test set as a function of the triangular kernel slope (or inverse bandwidth) in the offline (left) and online (right) scenarios. In both cases the extreme global model performs better than the extreme local model but both models are outperformed by an intermediate model with $0 < h < \infty$. Note that in accordance with the earlier discussion concerning the online and offline scenarios, the latter results in more accurate modeling.

where again we used the kernel symmetry to deduce $\int K^2(u)u\,du = 0$. Since $\mathsf{Var}\,(\theta_t) = (\mathrm{diag}(\theta_t) - \theta_t\theta_t^\top)$, 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

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

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

Noticing that $\dot{\mathsf{Var}}\,(\theta_t) = \mathrm{diag}(\dot{\theta}_t) - \dot{\theta}_t\theta_t^\top - \theta_t\dot{\theta}^\top$ concludes the proof. □

Proposition 2 specifies the conditions for consistency as well as the rate of convergence. In particular, the bias of online kernels converges at a linear rather than quadratic rate. In either cases, the estimator is biased and inconsistent unless $h \to 0$. Expressions (11)-(14) reveal the performance gain associated with a slower moving drift θ_t and sampling density g and with more (represented by n) and longer (represented by λ) documents. It is also possible to prove consistency of $\hat{\theta}_t$ using more general-purpose expansions that do not rely on the closed form expression (8). Such alternatives, however, rely on the central limit theorem and typically lead to constants and rates that are not as accurate as the ones in Proposition 2. We also note that the similarity between $\hat{\theta}_t$ and the Nadaraya-Watson regression estimator leads to similarity between our asymptotic expressions and the asymptotic expressions corresponding to locally constant regression e.g., [13].

We display the per-word loglikelihood of the local likelihood estimator (8) over a held-out test set in Figure 4 as a function of h^{-1}. The left panel illustrates the offline scenario while the right panel illustrate the online scenario. In both cases we used a triangular kernel containing a linear slope. In this and subsequent figures we use the inverse bandwidth h^{-1} or triangular slope rather than h since the extreme global model $h \to \infty$ turns to $h^{-1} \to 0$ which is more convenient to visualize. In both cases a triangular slope of $h^{-1} = 15$

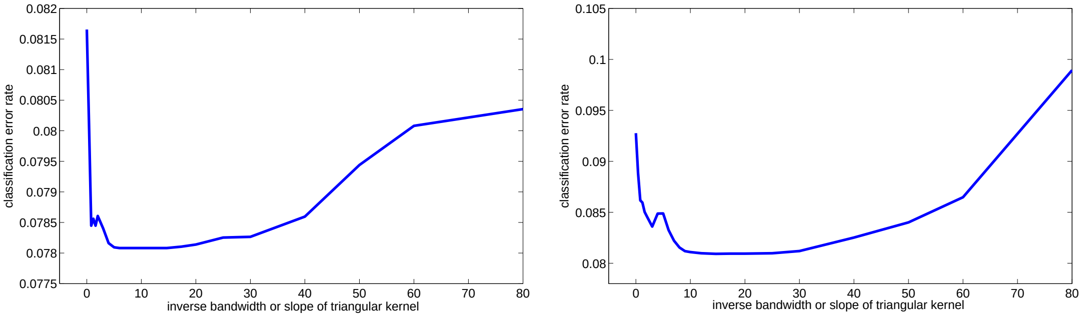

Figure 5: Classification error rate over a held-out test set of the naive Bayes classifier as a function of the inverse bandwidth parameters of a triangular kernel. The classification task was to distinguish between documents associated with the two most popular classes in RCV1. The left panel illustrates the offline scenario while the right panel illustrates the online scenario.

seems to perform best. The optimal slope of 15 corresponds to a positive support of 25 days in the online scenario and 49 days in the offline scenario i.e. $365-25$ and $365-49$ days are completely ignored. As a result, in addition to obtaining a more accurate model, the local model enjoys computational efficiency as it ignores a large portion of the training data.

We also explore the behavior of estimator (8) in the classification task. Using Bayes rule we can classify documents according to $\arg\max_{y\in\mathcal{Y}} p(x|y)p(y)$ where $p(x|y)$ is estimated separately for the different classes using Equation (8). Such classifiers are known as naive Bayes classifiers and despite their conditional independence assumption are still popular due to their simplicity and computational efficiency. Figure 5 displays the error rate over a held-out test set as a function of the inverse bandwidth for the offline (left) and online (right) scenarios.

A central issue in local likelihood modeling is selecting the appropriate bandwidth h. A practical solution is to use cross validation or some other automatic bandwidth selection mechanism. The performance of such cross validation schemes on RCV1 and their relation to the optimal bandwidth and the extreme global model is illustrated in Figure 6. From a theoretical perspective, the asymptotic bias and variance can be used to characterize the optimal bandwidth and study its properties. Minimizing the (offline) leading term of sum of component-wise MSE with respect to h we obtain the following expression

$$\hat{h}_t = \left(\frac{\mu_{02}(K)\mathrm{tr}(\mathrm{diag}(\theta_t) - \theta_t\theta_t^\top)}{4n\lambda\mu_{21}^2(K)\sum_j \left([\dot{\theta}_t]_j\, g'(t)/\sqrt{g(t)} + \sqrt{g(t)}[\ddot{\theta}_t]_j\,/2\right)^2}\right)^{1/5}. \tag{15}$$

As expected, the optimal bandwidth decreases as $n, \lambda, \|\dot{\theta}_t\|, \|\ddot{\theta}\|$ 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 7 which demonstrates the temporal change in the gradient norm (left) and the partial derivatives with respect to two of the three words in Figure 1 (right).

Perhaps more interesting than the dependency of the optimal bandwidth on $n, \lambda, \dot{\theta}_t, \ddot{\theta}_t$ is its dependency on the time sampling distribution $g(t)$. Equation (15) reveals an un-expected non-monotonic dependency of the optimal bandwidth in $g(t)$. The dependency, expressed by $\hat{h}_t \propto (\sum_{j=1}^V (c_{1j}/\sqrt{g(t)} + c_{2j}\sqrt{g(t)})^2)^{-1/5}$ is illustrated in Figure 8 (left) where we assume for simplicity that c_{1j}, c_{2j} do not change with j resulting in $(\hat{h}_t)^{-1} \propto (c_1/\sqrt{g(t)} + c_2\sqrt{g(t)})^{2/5}$. The key to understanding this relationship is the increased asymptotic bias due to the presence of the term $g'(t)/g(t)$ in Equation (11). Intuitively, the variations in $g(t)$ expressed

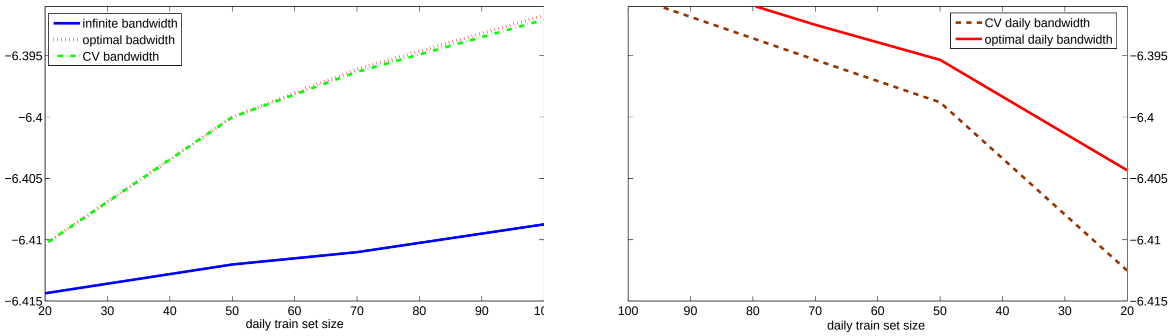

Figure 6: Per-word log-likelihood over held-out test set for various bandwidths as a function of the daily training set size. Left: The extreme global model corresponding to infinite bandwidth performs worst. Selecting the bandwidth by cross validation results in test-set loglikelihood that closely resembles that of the optimal slope. Right: Allowing the kernel scale to vary over time results in a higher modeling accuracy. Selecting a time dependent scale through cross validation presents some difficulties due to the fact that much less data is available in each of the daily cross validation problems.

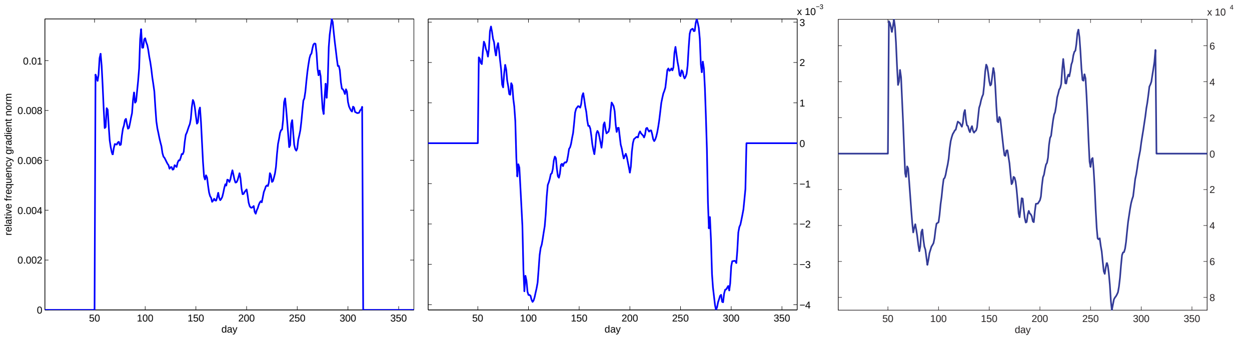

Figure 7: Estimated gradient norm (left) and partial derivatives (center, right) for the most popular category in RCV1 titled `earn` as a function of t. The center and right panels contain the partial derivatives with respect to two of the three words in Figure 1: `million` (center) `common` (right). The derivatives were estimated using local smoothing. To avoid running into boundary effects we ignore the first and last 50 days.

by $g'(t)$ introduce a bias component which alters the otherwise monotonic role of the optimal bandwidth and bias-variance tradeoff. Since $g(t)$ is highly non-uniform (as illustrated in Figure 2), this dependency of $\hat{h}_t$ on $g(t)$ is likely to play a significant role. 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 8 (right) reveals a trend similar to the theoretical dependency displayed in Figure 8 (left).

4.2 Conditional Local Likelihood for Logistic Regression

Often, the primary goal behind modeling the drift is conditional modeling i.e. predicting the value of y given x. In this case, drift modeling should focus on estimating the conditional $p_t(y|x)$ since modeling the marginal $p_t(x)$ becomes irrelevant. In contrast to the modeling of the conditional by Bayes rule $p_t(y|x) \propto p_t(x|y)p_t(y)$ described in the previous section, we explore here direct modeling of $\{p_t(y|x) : t \in I\}$ using local logistic regression.

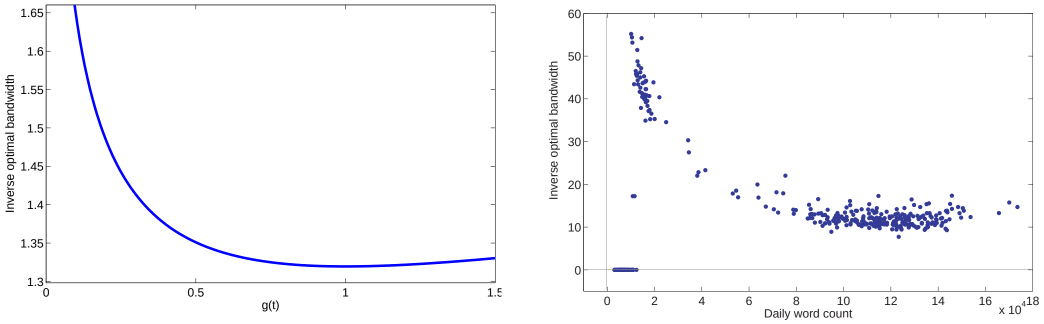

Figure 8: Left: Inverse of the optimal bandwidth derived from Equation (15) as a function of $g(t)$: $(\hat{h}_t)^{-1} \propto (c_1/\sqrt{g(t)} + c_2\sqrt{g(t)})^{2/5}$ (we take $c_1 = c_2 = 1$). Right: Inverse of the optimal bandwidth $(\hat{h}_t)^{-1}$ (averaged over the optimal bandwidths for the 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 $g(t)$.

By direct analogy to Equation (5) the conditional local likelihood estimator $p_t(y|x)$, or using parametric notation $p(y|x\,;\theta_t)$, is the maximizer of the locally weighted conditional loglikelihood

$$\ell_t(\eta|D) = \sum_{\tau \in I}^{n} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log p(y_{\tau j}|x_{\tau j};\eta) \qquad \eta \in \Theta. \tag{16}$$

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

Applying the logistic regression model $\log \frac{p(1|x\,;\theta_t)}{1-p(1|x\,;\theta_t)} = \sum_j [\theta_t]_j x_j$ or the equivalent exponential family model

$$p_t(y|x) = p(y|x\,;\theta_t) \overset{\text{def}}{=} \frac{\exp(y\langle\theta_t, x\rangle/2)}{\exp(\langle\theta_t, x\rangle/2) + \exp(-\langle\theta_t, x\rangle/2)}, \qquad y \in \{-1, +1\} \tag{17}$$

to (16) we obtain the following local conditional likelihood

$$\ell_t(\eta|D) = -\sum_{\tau \in I}^{n} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log\left(1 + \exp\left(-y_{\tau j}\langle x_{\tau j}, \eta\rangle\right)\right). \tag{18}$$

In contrast to the naive Bayes model in the previous section, expression (18) 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 (for each t) [7]. Most of the standard iterative algorithms for training logistic regression can be modified to account for the local weighting in (18). Recently popularized regularization techniques such as L_1 or L_2 penalty $c\|\eta\|^q$ may be added to Equation (18) to obtain a local regularized version equivalent to maximum posterior estimation.

Figure 9 (right) displays classification error rate over a held-out test set for local logistic regression as a function of the train set size. The documents were represented using the vector of relative frequencies making the resulting classifier (17) draw a linear decision boundary in $\mathbb{P}_V$. The plots in the figure correspond to both online and offline tricube kernels with optimal and infinite bandwidths, using 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 achieve 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.

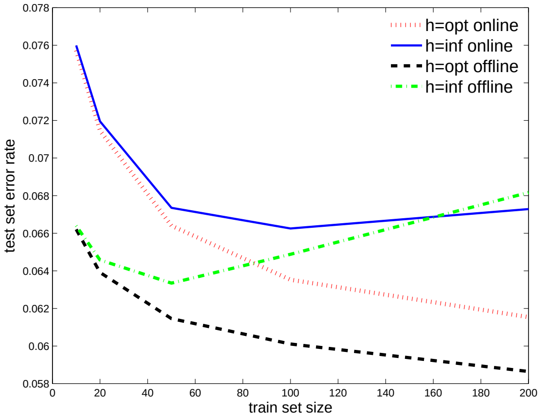

Figure 9: Classification error rate over a held-out train set for the local logistic regression model as a function of the train set size.

5 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 dataset demonstrate the validity of the local likelihood estimators in practice and contrast them with more standard non-local alternatives.

References

[1] S. Chen and J. Goodman. An empirical study of smoothing techniques for language modelling. Technical report, Harvard university, 1998.

[2] G. Forman. Tackling concept drift by temporal inductive transfer. In *Proc. of the ACM SIGIR Conference*, 2006.

[3] D. P. Helmbold and P. M. Long. Tracking drifting concepts by minimizing disagreements. *Machine Learning*, 14, 1994.

[4] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In *Proc. of the ACM SIGKDD Conference*, 2001.

[5] A. Kuh, T. Petsche, and R. L. Rivest. Learning time-varying concepts. In *Advances in Neural Information Processing Systems, 3*, 1990.

[6] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[7] C. Loader. *Local Regression and Likelihood*. Springer, 1999.

[8] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[9] F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison Wesley, 1964.

[10] G. Mulligan. *Removing the Spam: Email Processing and Filtering*. Addison Wesley, 1999.

[11] B. Pang and L. Lee. Seeing stars: Exploiting class relationship for sentiment categorization with respect to rating scales. In *Proc. of the Association of Computational Linguistics*, 2005.

[12] R. Tibshirani and T. Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82, 1987.

[13] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall/CRC, 1995.

[14] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1:69–90, 1999.