The Annals of Applied Statistics
2020, Vol. 14, No. 4, 1699–1723
https://doi.org/10.1214/20-AOAS1344
© Institute of Mathematical Statistics, 2020

MINING EVENTS WITH DECLASSIFIED DIPLOMATIC DOCUMENTS

BY YUANJUN GAO[1], JACK GOETZ[2], MATTHEW CONNELLY[3] AND RAHUL MAZUMDER[4]

[1]*Department of Statistics, Columbia University, gaoyuanjun0430@gmail.com*
[2]*Department of Statistics, University of Michigan, jackrgoetz@gmail.com*
[3]*MIT Sloan School of Management, Operations Research Center and Center for Statistics, MIT, rahulmaz@mit.edu*
[4]*Department of History, Columbia University, mjc96@columbia.edu*

Since 1973, the U.S. State Department has been using electronic record systems to preserve classified communications. Recently, approximately 1.9 million of these records from 1973–77 have been made available by the U.S. National Archives. While some of these communication streams have periods witnessing an acceleration in the rate of transmission, others do not show any notable patterns in communication intensity. Given the sheer volume of these communications, far greater than what had been available until now, scholars need automated statistical techniques to identify the communications that warrant closer study. We develop a statistical framework that can identify from a large corpus of documents a handful that historians would consider more interesting. Our approach brings together techniques from nonparametric signal estimation, statistical hypothesis testing and modern optimization methods—leading to a set of tools that help us identify and analyze various geometrical aspects of the communication streams. Dominant periods of heightened activities, as identified through these methods, correspond well with historical events recognized by standard reference works on the 1970s.

1. Introduction. For more than 40 years, social scientists have been developing datasets for the quantitative analysis of world politics. The last decade has witnessed a dramatic increase in activity in this area, much of it focused on automatic event detection for purposes of explaining and predicting political crises (Beieler et al. (2016)). All of these efforts, however, have used public information, such as newspaper or wire service reporting. Rather than directly measuring political activity, these systems can only count what reporters write about, which can vary over time and geography, depending on many extraneous factors (Jenkins and Maher (2016)). Together with the intrinsic challenges in automatic extraction, this has produced datasets that purport to track the same kind of events, such as political protests, but that are completely uncorrelated (Hanna (2014)). Moreover, some of the most important political activity is not immediately reported and may not become publicly known until decades later, when formerly secret records are declassified. The sheer volume of these records can make it difficult, even for the diligent researcher, to identify individual events and assess their relative importance.

In this paper we study a new dataset of declassified documents and use statistical methods to identify political events and explore how these heterogeneous events manifest themselves in the form of different geometric characteristics of these diplomatic communication streams. Since 1973, the State Department has been using electronic records systems to preserve classified communications. The National Archives[1] now makes available over 1.4 million declassified cables from 1973–77 as well as the metadata of more than 0.4 million other communications originally delivered by diplomatic pouch. They are all machine readable—creating many opportunities for statistical analyses.

Received May 2019; revised February 2020.
Key words and phrases. U. S. History, National Archives, signal estimation, fused lasso, optimization.
[1]Website: https://aad.archives.gov/aad/series-list.jsp?cat=WR43.

Our goal is to explore methods that can automatically identify statistically interesting events in an important corpus of historical documents which will continue to grow year-by-year as millions of additional communications are declassified. We contend that these "interesting" statistical patterns correspond to heightened diplomatic activity and validate our findings with standard reference works on the 1970s. A statistically interesting pattern can mean several things; we explore how they relate to heterogeneous political events. To provide some intuition, this can correspond to sudden localized changes or abrupt "jumps" in communication traffic, regardless of the overall series-specific baseline activity (a communication stream may be very active or have very low traffic intensity overall). There can also be continuous periods in a communication stream, where the data lies consistently above a series-specific baseline that corresponds to a representative global activity level of that stream. These are "bursts" of activity in the temporal structure of the document streams that probably correspond with heightened diplomatic activity, such as the start or end of a war. An interesting event can also correspond to heightened traffic intensity that plays out over longer periods, such as an increase over time.

When these communications were first entered in the State Department system, they were assigned one or more TAGS (Traffic Analysis by Geography and Subject) which indicate to what countries or subjects each cable is related. For example, "VS" signifies South Vietnam, and "SHUM" concerns human rights. By using these content-based TAGS as the feature, we avoid the complication of language processing and focus on identifying statistically relevant activity patterns based on the traffic of communication streams. Unfortunately, reliable text data is unavailable for many thousands of the cables in this corpus, either because State Department storage systems failed to preserve it or because only the metadata has been declassified for cables that have been deemed to contain sensitive information. In this context, the TAGS-specific features appear to be quite useful and effective in identifying events of importance to a social scientist.

1.1. *A brief exploratory description of the data.* A glimpse of processed data in the form of communication streams is shown in Figure 1. The data shows that there is less traffic on weekends and holidays (including the end of the year). In addition, the number of communications sent in 1973 seems to be smaller compared to later years, due to fewer records. We study the time series at a granular level by restricting to different types of TAGS. In Figure 1, panels (a)–(d) represent the communication streams when restricted by TAGS type. Panels (a)–(c) show noticeable forms of increased activities in portions of the series; these are indicative of "interesting" historical events. For example, in panel (a) the increased activity in July 1974 corresponds to the Cyprus coup; in panel (b) the increase in number of diplomatic communications in April 1975 corresponds to the Fall of Saigon; in panel (c) multiple bursts correspond to the annual United Nations General Assembly meetings. In addition to these visible bursts there seem to be some shorter periods of heightened activities, such as the smaller peaks for VS (South Vietnam) a year after the fall of Saigon corresponding to the ensuing refugee crisis.

In contrast to panels (a)–(c) in Figure 1, panel (d), for cables related to Finland (FI), does not seem to show any period of heightened activity during the time period under consideration. These prototypes are representative of the different TAGS-specific series: Exploratory analyses of the database of TAGS-specific communication streams suggest that there are several series with some "interesting event" (as in panels (a)–(c)), while others seem to be less active (as in panel (d)). Changes in the proportion of a particular TAGS appearing in a communication stream seem to be better representatives of identifying whether a period is active or not, as compared to tracking the corresponding counts.

Due to the noticeable difference in the number of cables that were communicated over the weekdays and holidays—as a preprocessing step, we filtered out the days where the total number of cables being communicated were very small.

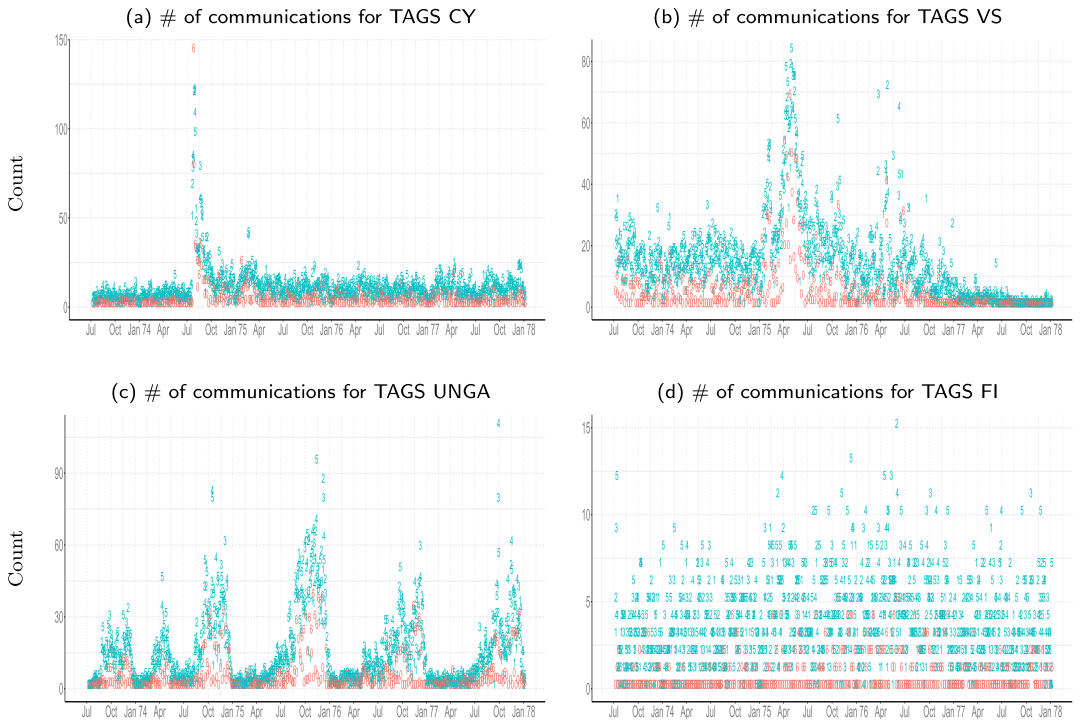

FIG. 1. *Figures showing counts of communications sent on each day, in the period* 1973–1977. *The numbers in the plot represent day-of-week* (0-*Sunday*, 1-*Monday*, 2-*Tuesday*, ..., 6-*Saturday*), *with weekdays colored in blue and weekends in red. Figures* (*a*)–(*d*) *show the communications restricted to different TAGS. The apparent heightened activities in the communication streams correspond to*: (*a*) *Cyprus coup*, (*b*) *Fall of Saigon* (*the most intense one*), (*c*) *the yearly United Nation General Assembly meetings. There does not seem to be any interesting activity for panel* (*d*), *showing cables related to Finland. A goal of this paper is to create statistical methods to automatically identify series with heightened diplomatic communications and further describe their structural patterns.* .

1.2. *Scope of this work.* A first goal of our work is to quantitatively define traits that separate communications like panel (d) from panels (a)–(c). We develop statistical methods that can *mine* these (TAGS-specific) time series and identify communication streams that exhibit statistically interesting activities. Once they are identified, we develop algorithms (Sections 2 and 3) that perform a deeper investigation of each series and identify time intervals where the signal undergoes abrupt localized changes in communication traffic. We also present methods to quantify and contrast these various geometric patterns.

Our general methodological approach is inspired by principles in statistical signal segmentation and change-point modeling with origins in 1950s (Page (1954)); see Brodsky and Darkhovsky (1993), Truong, Oudre and Vayatis (2018) for excellent overview(s) of the topic. On the algorithmic front we employ ideas from first-order methods in continuous optimization (Nesterov (2004), Nesterov (2013)) that complement state-of-the-art approaches in change-point detection (Johnson (2013), Killick, Fearnhead and Eckley (2012)). Statistical models for change-point detection have enjoyed a great deal of success across several application domains spanning speech processing, financial analysis, bioinformatics, climatology, network traffic, gait analysis, text processing, among others. Similar models are also employed in the context of burstiness analysis (Kleinberg (2003)). Such ideas are enhanced in important ways for identifying events in Twitter communication streams; see, for example, Atefeh and Khreich (2015) for a nice survey. Our main goal in this paper is to build upon classical and modern statistical signal estimation/inference tools and enhance them in suitable ways so that they can provide insights to a historian on a new dataset available from the

National Archives. In this work we establish a synergy between statistics and social science perspectives. As we discuss in Section 4, analysis of diplomatic documents using TAGS-based features relevant to a historical scientist presents a unique set of challenges. The scope of our work is quite different from the analysis of text-based features to mine events in Twitter communication streams.

2. Statistical methodology. We first present a brief outline of the main statistical approaches pursued in this paper. Section 2.1 addresses how we can use a global testing approach to determine whether a TAGS-specific communication stream, among several hundreds, is interesting or not. To further explore the geometry of the underlying signal, we use a regularized negative log-likelihood criterion based on the fused lasso penalty (Tibshirani et al. (2005)) and also its ℓ_0-counterpart (Killick, Fearnhead and Eckley (2012), Boysen et al. (2009), Johnson (2013)). For efficient computation we propose a unified framework for these optimization problems which seem to be promising alternatives to prior approaches (Johnson (2013), Killick, Fearnhead and Eckley (2012)). We use hypothesis testing ideas based on sample splitting (Wasserman and Roeder (2009)) to associate p-values to the detected jumps; see Section 2.3. Inspired by Kleinberg (2003), the jumps in an individual series are aggregated to obtain "bursts," leading to a rank-ordering of political events across the corpus. Finally, in Section 3 we discuss how to estimate the underlying proportions with models that are more flexible than piecewise constant segments.

2.1. *Identifying interesting communication streams.* Consider a TAGS-specific series $(y_t, n_t), t = 1, \ldots, N$, where, y_t denotes the number of documents containing the specific TAGS among n_t cables, with proportion p_t. We will assume that the conditional distributions of $(y_t|n_t, p_t)$'s are independent across t. We are interested in the following question:

Is there any evidence of (localized) heightened intensity of the proportions, compared to a baseline model, where all proportions are the same?

To measure a localized change (increase) in intensity, we fix a window of size 2Δ and consider all the points in the Δ neighborhood of a time point i, given by $\mathrm{N}(\Delta; i) = \{j : 1 \le j \le N, |j - i| \le \Delta\}$. The average proportion in this neighborhood,

$$p_i^{\mathrm{ave}} := \sum_{j \in \mathrm{N}(\Delta;i)} n_j p_j \Big/ \sum_{j \in \mathrm{N}(\Delta;i)} n_j,$$

is a measure of communication traffic around the reference time point i. We say that a large value of p_i^{ave}, compared to a baseline value p, indicates the presence of an intense localized activity of some form.[2] We hypothesize such an activity to be associated with an event of historical interest and, subsequently, validate this belief by factoring in the insights of a historian or social scientist.

Formally, we consider a global testing approach with H_0: $p_t = p\ \forall t$ vs. H_1: there exists an i such that p_i^{ave} is larger than the (global) average proportion. Inspired by popularly used scan statistics (Glaz, Naus and Wallenstein (2001)), we propose the following test statistic:

$$\mathcal{T} = \max_t T_t \quad \text{where, } T_t := \big(\widehat{p_t^{\mathrm{ave}}} - \hat{p}_{H_0}\big)/\hat{\sigma}_t, \tag{2.1}$$

where, $\hat{p}_{H_0}$ is the (estimated) global proportion of the signal estimated under the null hypothesis; $\widehat{p_t^{\mathrm{ave}}}$ is an estimate of p_t^{ave}, that is, $\widehat{p_t^{\mathrm{ave}}} = \sum_{j \in \mathrm{N}(\Delta;i)} y_j / \sum_{j \in \mathrm{N}(\Delta;i)} n_j$. Furthermore, $\hat{\sigma}_t$ is the estimated standard deviation of $\widehat{p_t^{\mathrm{ave}}}$ evaluated under the null (H_0): If $\hat{p}_{H_0}$ denotes

[2] At this point we do not offer an explanation of the exact geometric reason behind such an heightened activity—we address this at a later stage in the paper.

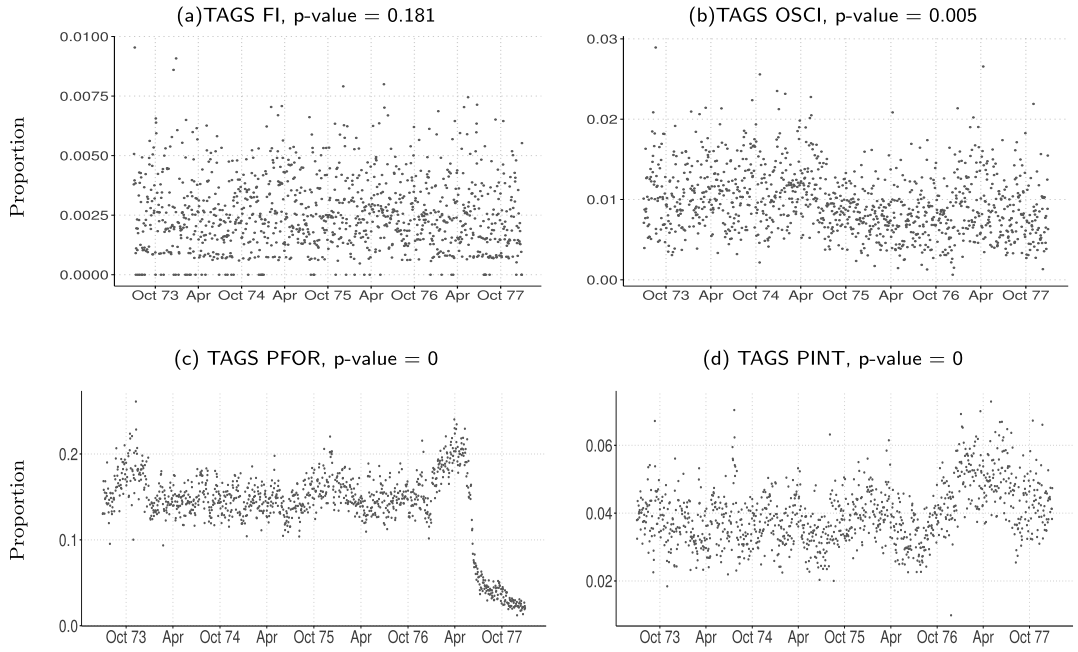

FIG. 2. *Communication streams with different significance scores based on the framework of Section* 2.1: (*a*): (*TAGS FI*) *relating to Finland, corresponds to a null model*; (*b*): (*TAGS OSCI*) *relating to scientific grants, shows a weak deviation from the null model—perhaps due to a slow decreasing trend in the series*; (*c*): (*TAGS PFOR*) *relating to foreign policy—this generally shows significant deviation from the null model which is due to sudden changes/jumps after Oct.* '76; (*d*): (*TAGS PINT*) *relating to internal political affairs, shows deviation from the null model but not due to a jump as prominent as* (*c*)*—this series seems to exhibit some systematic pattern of heightened activity after Oct.* '76, *leading to a small p-value. The small p-values suggest the presence of a statistically interesting event in each series and can be used to identify interesting communication streams. The p-value, however, does not provide additional insights into the finer structural patterns of the streams. Additional examples can be found in Figure* 2 *in the Supplementary Material* (Gao et al. (2020)).

the estimate of p under the null, then $\hat{\sigma}_t^2 = (\hat{p}_{H_0}(1-\hat{p}_{H_0}))/(\sum_{j\in N(\Delta;i)} n_j)$; see the Supplementary Material (Gao et al. (2020)) for a derivation. T_t measures the strength of a locally contiguous period of heightened activity; we take the supremum over all time points t to get $\mathcal{T}$. The larger the value of $\mathcal{T}$, the more pronounced is the localized traffic compared to the baseline value $\hat{p}_{H_0}$. We use a permutation based approach to compute the null distribution of $\mathcal{T}$. Figure 2 (see also Figure 2 in the Supplementary Material (Gao et al. (2020)) for additional examples) shows different communication streams with their associated p-values. A large[3] p-value for panel (a), Figure 2 (corresponding to TAGS FI) signifies a lack of interesting activity in this series; this aligns with an expert's understanding that, during this period, there was limited diplomatic activity at the international scale related to TAGS FI.

To understand the sensitivity of results to Δ, a summary of how many cables survive different p-value thresholds for different choices of Δ are provided in the Supplementary Material (Gao et al. (2020)). While the TAGS-specific p-values are found to change with Δ, the overall results remain quite stable.[4] Note that we use this step to simply remove a small fraction of the communication streams from further downstream analysis. The results

[3]We note the choice of a p-value threshold (i.e., whether it is deemed to be large or small) may depend upon the subjective intuition of a practitioner.

[4]Usually, the p-values smaller than 0.001 become smaller with increasing values of Δ; larger p-values remain large.

in Figure 2 suggest an important limitation of statistic (2.1)—this framework in itself does not offer much insight into the geometry of the signal. This motivates the methods presented in Section 2.2.

2.2. *Identifying jumps in communication streams.* We propose methods to explore finer structural properties of the series that are not informed by the methods in Section 2.1. Inspired by popularly used signal segmentation/estimation methods (Killick, Fearnhead and Eckley (2012), Tibshirani et al. (2005), Mammen and van de Geer (1997)), we seek to identify breaks or jumps in a piecewise constant approximation of the signal $t \mapsto p_t$.

Regularized maximum likelihood. Using the notation of Section 2.1, we assume $(y_t|n_t, p_t) \overset{\text{ind}}{\sim} \text{Bin}(n_t, p_t)$ for $t = 1, \ldots, N$ where, p_t denotes the probability of success and N denotes the total number of time points. This leads to a joint likelihood (conditional on $\{(n_t, p_t)\}_{t\geq 1}$) given by

$$P\big(\{y_t\}_1^N | \{(n_t, p_t)\}_1^N\big) = \prod_{t=1}^{N} \binom{n_t}{y_t} p_t^{y_t} (1-p_t)^{(n_t - y_t)}. \tag{2.2}$$

Note that an unconstrained maximum likelihood estimator will lead to $\hat{p}_t = y_t/n_t$ for all t which overfits the data. Therefore, additional structural constraints on p_t are needed for interpretable models. Using the standard logistic parametrization, $p_t = \exp(\theta_t)/(1+\exp(\theta_t))$, the negative log-likelihood (2.2) in terms of the variables $\{\theta_t\}_1^N$ is

$$\sum_{t=1}^{N} \{-y_t\theta_t + n_t \log(1+\exp(\theta_t))\} - \sum_{t=1}^{N} \log\left(\binom{n_t}{y_t}\right), \tag{2.3}$$

where the second term above does not depend upon $\theta_t, t \geq 1$. The expression in (2.3) is convex in $\{\theta_t\}_1^N$—we consider $\theta_t, t \geq 1$ to be our natural parameters. We first discuss a method to approximate $t \mapsto \theta_t$ by a piecewise constant signal (generalizations are discussed in Section 3). A location where the latent signal $t \mapsto \theta_t$ exhibits a discontinuity will be called a "jump" in the communication stream. We consider the following regularized criterion:

$$\underset{\theta_t, 1\leq t\leq N}{\text{minimize}} \sum_{t=1}^{N} \big(-y_t\theta_t + n_t \log(1+\exp(\theta_t))\big) + \lambda H(\boldsymbol{\theta}), \tag{2.4}$$

where $\mathcal{L}(\boldsymbol{\theta}) := \sum_{t=1}^{N}(-y_t\theta_t + n_t \log(1+\exp(\theta_t)))$ is the part of (2.3), depending upon $\{\theta_t\}_1^N$ (i.e., the data-fidelity term) and $H(\boldsymbol{\theta})$ is the regularizer. $H(\boldsymbol{\theta})$ encourages the estimated θ_t's (and hence the proportion p_t's) to be piecewise constant, and the regularization parameter $\lambda > 0$ controls the amount of shrinkage. Two examples of $H(\boldsymbol{\theta})$ we study appear in earlier works in signal estimation (Tibshirani et al. (2005), Johnson (2013), Killick, Fearnhead and Eckley (2012))—we present a simultaneous analysis for both these choices:

- ℓ_1-segmentation (fused lasso): Here, $H(\boldsymbol{\theta}) = H_{\ell_1}(\boldsymbol{\theta}) = \sum_{t=1}^{N-1} |\theta_{t+1} - \theta_t|$—this penalizes the total variation of a signal and may also be thought as a soft version of the number of jumps in $\theta_t, t \geq 1$.
- ℓ_0-segmentation: Here, we take $H(\boldsymbol{\theta}) = H_{\ell_0}(\boldsymbol{\theta}) = \sum_{t=1}^{N-1} \mathbf{1}(\theta_{t+1} \neq \theta_t)$ which penalizes the number of jumps in the signal $\theta_t, t \geq 1$.

We assume above and in the discussion below that the time points are equally spaced. If they are not equispaced, the penalty function needs to be adjusted in a straightforward fashion, as discussed in the Supplementary Material (Gao et al. (2020)).

Choice of regularizer. For the ℓ_1 penalty $H_{\ell_1}(\boldsymbol{\theta})$, Problem (2.4) is a convex optimization problem. This is commonly referred to as the fused lasso or total-variation penalty (Tibshirani et al. (2005), Mammen and van de Geer (1997)) and used in the context of signal estimation wherein the underlying signal is assumed to have a small total variation norm. $H_{\ell_1}(\boldsymbol{\theta})$ shrinks the successive coefficient differences $\{\theta_{t+1} - \theta_t\}_t$ to zero and, due to the presence of the ℓ_1-norm, encourages sparsity in $\theta_{t+1} - \theta_t$'s leading to a piecewise constant signal $t \mapsto \theta_t$. The shrinkage effect of the ℓ_1-penalty severely penalizes large values of the jumps $\theta_{t+1} - \theta_t$. Hence, this penalty leads to a model with many jumps, especially when the tuning parameter is chosen so as to obtain a model with good data-fidelity (e.g., if the tuning parameter is chosen based on validation set tuning). To obtain a model with fewer jumps, the regularization parameter needs to be made larger—in the process, important jumps may be missed. These observations are well known in the context of the usual lasso estimator in regression; see, for example, Bertsimas, King and Mazumder (2016), Mazumder, Friedman and Hastie (2011). An alternative is to use an ℓ_0-based penalty (Killick, Fearnhead and Eckley (2012), Boysen et al. (2009)) which directly penalizes the number of jumps and is agnostic to the precise value of the jump. Both these penalty functions are popularly used in change-point detection in statistics; see for example, Truong, Oudre and Vayatis (2018) for a recent review. The rich literature on ℓ_0 and ℓ_1-based approaches seems to have grown somewhat independently of one another, with curious links and differences between the two approaches. Indeed the ℓ_0 and ℓ_1-based estimators have different operating characteristics—this is also seen in our numerical experiments. For example, if the amount of shrinkage for the ℓ_0-penalty is not very high, one can obtain a signal with short segments. Furthermore, being agnostic to the magnitude of the jumps, the ℓ_0-based estimator may lead to a signal estimate that is "spiky"; this is often ameliorated with ℓ_1-based estimators which shrinks the magnitude of a jump. For additional discussion on the delicate differences between ℓ_1 and ℓ_0-based estimators in the regression context,[5] we refer the reader to the recent works of Hazimeh and Mazumder (2018), Mazumder, Radchenko and Dedieu (2017). The signal estimation problem we study here differs from the regression problem and poses a unique set of challenges. Here, we present algorithms that can compute solutions for both the ℓ_1 and ℓ_0-based estimators. This allows us to gather useful insights about the estimators in the context of the problems studied here. More importantly, this will allow practitioners to make an informed decision regarding what might be appropriate in their context.

Other regularizers beyond the ℓ_0 and ℓ_1 penalties, alluded to above, are also used in the context of change-point models; see, for example, Killick, Fearnhead and Eckley (2012), Truong, Oudre and Vayatis (2018). In Section 3 we discuss another regularizer that encourages a piecewise linear description of $t \mapsto \theta_t$.

2.2.1. *Model fitting*: *Optimization algorithms.* Developing efficient specialized solvers for Problem (2.4) is a challenging task. Johnson (2013), Killick, Fearnhead and Eckley (2012) propose appealing dynamic programming based algorithms for Problem (2.4). However, as far as we can tell, the software packages made available by Johnson, Killick, Fearnhead and Eckley do not provide implementations for the general form of Problem (2.4)—they present easy-to-use interfaces for the least squares loss function. We present an alternative method to obtain good quality solutions to Problem (2.4) by using first-order optimization methods (Nesterov (2004)). To this end, we rely on efficient dynamic programming solvers proposed by Johnson and Killick, Fearnhead and Eckley for the *least squares* loss with the $H_{\ell_1}(\boldsymbol{\theta})$ or $H_{\ell_0}(\boldsymbol{\theta})$ penalty.

[5]The differences of these estimators depend upon a multitude of factors, such as signal to noise ratios, feature correlations, model sparsity, sample size, number of features, etc.

We note that the framework presented below applies to problems more general than Problem (2.4). In particular, they apply to a more general class of problems than can be handled via dynamic programming methods (Johnson (2013), Killick, Fearnhead and Eckley (2012))—they are complementary to the suite of algorithms used in change point models[6] and may be of independent interest. We present proximal gradient descent methods (Beck and Teboulle (2009)) for problems of the composite form (Nesterov (2013)),

$$\min_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}) := \mathcal{L}(\boldsymbol{\theta}) + \lambda H(\boldsymbol{\theta}), \tag{2.5}$$

where $\mathcal{L}(\boldsymbol{\theta})$ is a function with Lipschitz continuous gradient,

$$\big\|\nabla\mathcal{L}(\mathbf{u}) - \nabla\mathcal{L}(\mathbf{v})\big\| \le \ell\|\mathbf{u}-\mathbf{v}\|, \quad \forall \mathbf{u}, \mathbf{v} \in \Re^N. \tag{2.6}$$

In the case of Problem (2.4), we have $\ell = \frac{1}{4}\max_{i=1}^N n_i$. This follows by noting that the ith coordinate of $\nabla\mathcal{L}(\mathbf{u})$ is: $\{\nabla\mathcal{L}(\mathbf{u})\}_i = -y_i + n_i\exp(u_i)/(1+\exp(u_i))$ and $\nabla^2\mathcal{L}(\mathbf{u})$ is a diagonal matrix with the ith diagonal entry satisfying

$$\big\{\nabla^2\mathcal{L}(\mathbf{u})\big\}_{ii} = n_i\exp(u_i)/\big(1+\exp(u_i)\big)^2 \le \frac{1}{4}n_i, \quad i = 1,\dots,N. \tag{2.7}$$

Hence, the largest eigenvalue of $\nabla^2\mathcal{L}(\mathbf{u})$, that is, $\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{u})) \le \frac{1}{4}\max_{i=1}^N n_i$, which justifies the choice of ℓ, as above. For a fixed $L \ge \ell$, the proximal gradient algorithm performs the following updates (for all $k \ge 0$):

$$\boldsymbol{\theta}_{k+1} \in \arg\min_{\boldsymbol{\theta}} \frac{L}{2}\left\|\boldsymbol{\theta} - \left(\boldsymbol{\theta}_k - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right)\right\|_2^2 + \lambda H(\boldsymbol{\theta}). \tag{2.8}$$

This leads to a decreasing sequence of objective values $\phi(\boldsymbol{\theta}_{k+1}) \le \phi(\boldsymbol{\theta}_k)$ for $k \ge 0$. If $\phi(\boldsymbol{\theta})$ is bounded below (which is true for Problem (2.4) as soon as $n_i > 0$ for all i), then $\phi(\boldsymbol{\theta}_k)$ converges to a finite value. We now study the fate of the sequence $\phi(\boldsymbol{\theta}_k)$ (and $\boldsymbol{\theta}_k$), depending upon the choice of $H(\boldsymbol{\theta})$.

2.2.1.1. *The fused lasso penalty* ($H_{\ell_1}(\boldsymbol{\theta})$). Due to the convexity of $H(\boldsymbol{\theta})$, the function $\phi(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$. Using standard results in proximal gradient methods, $\boldsymbol{\theta}_k$ converges to a minimum of Problem (2.5) with the penalty function $H_{\ell_1}(\boldsymbol{\theta})$. In terms of convergence rates of objective values, it follows from Beck and Teboulle (2009) that

$$\phi(\boldsymbol{\theta}_k) - \phi(\boldsymbol{\theta}^*) \le \frac{L}{2k}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2^2, \tag{2.9}$$

where $\boldsymbol{\theta}^*$ is an optimal solution to Problem (2.5)—hence, the sequence $\phi(\boldsymbol{\theta}_k)$ converges to the minimum of Problem (2.5) at a worst case rate of $O(1/k)$. If $\mathcal{L}(\boldsymbol{\theta})$ is strongly convex in $\boldsymbol{\theta}$, the sequence $\phi(\boldsymbol{\theta}_k)$ exhibits a linear rate (Nesterov (2013)) of convergence. Let $\mu_k = \min_{i=1,\dots,N}\{\nabla^2\mathcal{L}(\boldsymbol{\theta}_k)\}_{ii}$ denote the smallest diagonal entry of the Hessian of $\mathcal{L}(\boldsymbol{\theta}_k)$. Since, $\boldsymbol{\theta}_k$'s are uniformly bounded[7] and $\min_i n_i > 0$, then $\mu := \inf_k \mu_k > 0$. The convergence rate is given by (Nesterov (2013))

$$\phi(\boldsymbol{\theta}_k) - \phi(\boldsymbol{\theta}^*) \le \left(1 - \frac{\mu}{4L}\right)^k \big(\phi(\boldsymbol{\theta}_0) - \phi(\boldsymbol{\theta}^*)\big). \tag{2.10}$$

The rates in (2.9) and (2.10) are interesting to interpret. If $\mu/(4L)$ is small, the sublinear rate (2.9) explains the convergence speed of $\phi(\boldsymbol{\theta}_k)$ in the initial stages of the algorithm, after

[6]We note that, due to the generality of our methods, they may not lead to optimal solutions to Problem (2.4) if $H(\boldsymbol{\theta})$ is nonconvex.

[7]Note that, for Problem (2.5), $\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ has a finite minimizer (as $n_i > 0$ for all i); hence, an optimal solution to Problem (2.5) is finite, and the sequence $\{\boldsymbol{\theta}_k\}$ is uniformly bounded.

which the linear rate (2.10) will kick in. If $\mu/(4L)$ is large, the linear rate of convergence dominates, and the algorithm converges very fast. Interestingly, the convergence speed of the algorithm adapts to the better of the linear or sublinear rates without explicitly changing the algorithm.

We note that accelerated variants (Nesterov (2013)) of proximal gradient descent can also be used; they improve the convergence rate in (2.9) to $O(1/k^2)$.

Note that subproblem (2.8) is a problem of the form (for some $\lambda' > 0$)

$$\underset{\mathbf{u}\in\Re^N}{\text{minimize}}\ \frac{1}{2}\|\mathbf{u}-\bar{\mathbf{u}}\|_2^2+\lambda'\sum_{i=1}^{N-1}|u_{i+1}-u_i|. \tag{2.11}$$

This can be solved very efficiently via dynamic programming (Johnson (2013)) with a worst case cost of $O(N)$—in fact, for $N\approx 10^6$ the solver of Johnson (2013) takes usually around 0.03 seconds on a modest desktop computer. Hence, obtaining a solution to Problem (2.4) for a similar size usually takes a second or so.

2.2.1.2. *The ℓ_0-segmentation penalty ($H_{\ell_0}(\boldsymbol{\theta})$).* The algorithm above can also be applied for the penalty $H_{\ell_0}(\boldsymbol{\theta})$. In update (2.8) we set $H(\boldsymbol{\theta})$ to $H_{\ell_0}(\boldsymbol{\theta})$. Instead of Problem (2.11), we solve the following jump penalized least squares problem:

$$\underset{\mathbf{u}\in\Re^N}{\text{minimize}}\ \frac{1}{2}\|\mathbf{u}-\bar{\mathbf{u}}\|_2^2+\lambda'\sum_{i=1}^{N-1}\mathbf{1}(u_{i+1}\neq u_i) \tag{2.12}$$

which can be computed efficiently using the dynamic programming algorithm(s) of Johnson, Killick, Fearnhead and Eckley. Intuitively, our proposed proximal method approximates the smooth part of the loss function (by a quadratic objective as in a Newton method). A key component of the nonconvexity in the optimization problem (2.4) lies in solving (2.12) which can be solved to *optimality* via dynamic programming. The proximal gradient algorithm, outlined above, is different from the dynamic programming algorithms of Johnson and Killick, Fearnhead and Eckley that can solve Problem (2.4) to optimality. When direct comparisons are possible, our proposed algorithm seems to be faster than Johnson (2013) (See Section 2.2.2 for details). While our algorithm may lead to a local solution of Problem (2.4), in our numerical experiments, solutions were often found to be near optimal.

Describing the properties of the sequence $\boldsymbol{\theta}_k, k\geq 1$ is subtle for Problem (2.4) (with penalty function $H_{\ell_0}(\boldsymbol{\theta})$) due to nonconvexity of the optimization problem. Following Bertsimas, King and Mazumder (2016), it can be shown that the sequence $\phi(\boldsymbol{\theta}_k)$ is decreasing, bounded below,[8] and it converges to ϕ^* (this may depend upon the initialization). We say that $\tilde{\boldsymbol{\theta}}$ is a first-order stationary point for Problem (2.4) if setting $\boldsymbol{\theta}_k=\tilde{\boldsymbol{\theta}}$ leads to $\boldsymbol{\theta}_{k+1}=\tilde{\boldsymbol{\theta}}$. We say that $\boldsymbol{\theta}_k$ is an ϵ-accurate first order stationary point for Problem (2.5) if $\|\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k\|_2^2\leq\epsilon$. Following the convergence analysis in Bertsimas, King and Mazumder (2016), Theorem 3.1, we obtain the following finite-time convergence rate of $\boldsymbol{\theta}_k$ to a first order stationary point:

$$\min_{0\leq k\leq K}\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}\|_2^2\leq\frac{2(\phi(\boldsymbol{\theta}_0)-\phi^*)}{K(L-\ell)}. \tag{2.13}$$

The above convergence rate is conservative; in practice, $\phi(\boldsymbol{\theta}_k)$ is found to converge much faster (usually, within 10 iterations or so).

[8]This is satisfied under minor conditions, as discussed earlier.

2.2.1.3. *A constrained variant of Problem* (2.5). The above framework for the penalized problem can be extended to a constrained version of the form

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \quad \text{s.t. } H(\boldsymbol{\theta}) \leq \kappa, \tag{2.14}$$

where, $\kappa \geq 0$ is the regularization parameter (in constrained form). An important instance of the above corresponds to the choice $H(\boldsymbol{\theta}) = H_{\ell_0}(\boldsymbol{\theta})$—in which case we constrain the total number of jumps to a pre-set value κ. To obtain solutions to Problem (2.14), we will need to modify the proximal operator (2.8) to

$$\boldsymbol{\theta}_{k+1} \in \arg\min_{\boldsymbol{\theta}} \left\| \boldsymbol{\theta} - \left(\boldsymbol{\theta}_k - \frac{1}{L} \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right) \right\|_2^2 \quad \text{s.t. } H(\boldsymbol{\theta}) \leq \kappa.$$

For $H(\boldsymbol{\theta}) = H_{\ell_0}(\boldsymbol{\theta})$ this proximal operator can (once again) be solved using the dynamic programming framework of Johnson (2013) (at a slightly higher cost than the jump penalized least squares problem). When $H(\boldsymbol{\theta}) = H_{\ell_1}(\boldsymbol{\theta})$, the solutions obtained from (2.14) are in one-to-one correspondence to solutions from (2.5). This is no longer true when $H(\boldsymbol{\theta}) = H_{\ell_0}(\boldsymbol{\theta})$—it may be possible that, for certain choices of κ, Problem (2.5) does not lead to a solution with $H(\boldsymbol{\theta}) = \kappa$ for any value of λ. In this sense, formulation (2.14) with $H(\boldsymbol{\theta}) = H_{\ell_0}(\boldsymbol{\theta})$ may be more favorable than the penalized version.

2.2.2. *Related work* (*Algorithms*). The origins of change-point modeling in statistics date back to 1950s (Page (1954)); see Brodsky and Darkhovsky (1993), Truong, Oudre and Vayatis (2018) for excellent overview(s) of the topic. Dynamic programming based segmentation for curve-fitting appeared in 1960s (Bellman and Roth (1969)). Approximate segmentation methods, for example, based on binary segmentation appear in Olshen et al. (2004), Scott and Knott (1974)—these are popular heuristics that may not lead to an optimal solution to the nonconvex ℓ_0-penalized problem, as pointed out by Killick, Fearnhead and Eckley (2012) and others. Some more recent exact segmentation algorithms appear in Auger and Lawrence (1989), Jackson et al. (2005)—for a sequence of length N, these methods have a computational cost of $O(N^2)$. Killick, Fearnhead and Eckley (2012) use a pruning step to improve the algorithm of Jackson et al. (2005). Related algorithms, based on improvements of dynamic programming, also appear in Johnson (2013).

The two works most related to the segmentation problem, discussed in Section 2.2.1, are Johnson (2013), Killick, Fearnhead and Eckley (2012). The `R` package `changepoint` provides a user-friendly interface for mean and/or variance changes for Gaussian data and some other distributions (though not the Binomial distribution considered in this paper). In the Supplementary Material (Gao et al. (2020)) we present a comparison of the change points obtained by the algorithm in `changepoint` (for mean changes in Gaussian data) vs. the method presented here for the ℓ_0 jump penalized problem.

Under some assumptions the method of Killick, Fearnhead and Eckley has an expected computation cost of $O(N)$, though the worst-case cost is $O(N^2)$. Johnson describe dynamic programming algorithms for the ℓ_1 and ℓ_0-penalties with a variety of loss functions (separable across time points). The `R` package of Johnson provides a user-friendly interface for the squared error loss for both these penalty functions—their cost is $O(N)$. In our experience the algorithm of Johnson was found to be faster than Killick, Fearnhead and Eckley (2012) for large instances of Problem (2.12) (Gaussian ℓ_0 segmentation): For problems with $N = 10^5$ and $N = 10^6$, the method of Johnson exhibits a 300x-fold and 2000x-fold improvement (respectively) in runtimes over the algorithm of Killick, Fearnhead and Eckley. The signals we consider are much smaller (usually, $N \approx 1500$), hence the time difference between Johnson and Killick, Fearnhead and Eckley is less pronounced (Johnson (2013) shows a 5x speed improvement) for computing a solution to Problem (2.12). However, for every TAGS-specific

time-series, we use multiple calls to this function via (2.8) and consider multiple values of λ. When aggregated across all the TAGS-specific time series, there is an overall time-benefit in using the framework of Johnson.

For the ℓ_1-based segmentation problem with the Binomial likelihood, we compared our algorithm with the code of Johnson. For signals of length $N = 10^4$, $N = 10^5$ and $N = 10^6$, the implementation[9] of Johnson takes around 1.5, 18 and 197 seconds (respectively)—our method exhibits 187, 267 and 540x-fold speedups (respectively).

Since Killick, Fearnhead and Eckley, Johnson rely on dynamic programming, the class of loss functions considered is rather limited. Our proposed framework can address a larger family of loss functions. Since we rely on existing efficient solvers for Problem (2.8), our algorithms are also efficient, with a cost of $O(N)$ per iteration. Thus, Section 2.2.1 presents an easy-to-implement and useful tool for signal segmentation tasks.

Signal segmentation problems also appear in the computer science/data-mining literature. In a seminal paper, Kleinberg formalizes models for detecting bursts in events in the context of a continuous steam of events (e.g., email messages over time) and discrete time events (events arriving in batches as in conference papers)—they have close ties to change-point models in statistics. For the first case, Kleinberg models email messages arriving over time with an exponential interarrival rate. These rates can take values in a finite set (a.k.a. states) and are of the form $q_i = q_0 s^i$ for some scale factor s, and q_0 is a prespecified base rate. The rates can change over time, but there is a penalty for an increase in the current rate to a newer one. He uses dynamic programming algorithms for hidden Markov models to perform the estimation. The computational cost increases with the number of states—the `R` package `bursts` (that implements the algorithms of Kleinberg (2003)) is usually found to be much slower than the implementations of Johnson (2013), Killick, Fearnhead and Eckley (2012) alluded to above. For the case of events occurring in batches over discrete periods of time, Kleinberg uses a binomial model for every time point—there are two models corresponding to success probabilities q_1, q_0, with q_0 denoting the known baseline and $q_1 = q_0 s$ for some prespecified scale factor s. Our framework is different: While Kleinberg allows p_t to take two prespecified values, we allow p_t to take a continuum of values in $[0, 1]$. We allow for a flexible family of penalty functions H_{ℓ_0} and H_{ℓ_1} that penalize any increase/decrease in p_t, while Kleinberg penalizes only an increase. It is also known that Kleinberg's model may lead to under-smoothing—to this end, localized window averaging is often recommended as a smooth preprocessing step. The regularized likelihood framework considered in this paper systematically addresses the smoothing/under-smoothing tradeoff by adjusting the regularization parameter λ (e.g., based on cross-validation).

2.2.3. *Estimated signal.* To gather some intuition about the behavior of the estimators described above, we consider a synthetic example in Figure 3 and some real datasets in Figures 4 and 5.

2.2.3.1. *Illustration with a synthetic dataset.* In this synthetic example (Figure 3) the underlying (true) signal is piecewise constant with three levels up to time point t_0; there is a right discontinuity at t_0 after which it becomes linear.[10] Note that the underlying signal is not piecewise constant—there is model misspecification due to the linearity on the right part of the signal. This example is chosen to shed light into the behaviors of the ℓ_0 and ℓ_1-based

[9] We note that Johnson (2013) presents an implementation in `R` for this problem, unlike the least squares problem which is written in `C` with a `R` wrapper. Our code is written in `R`.

[10] More specifically, data is generated by $y_t \overset{\text{ind}}{\sim} \text{Bin}(n_t = 200, p_t)$, $t = 1, \ldots, 1203$, where $p_t = 0.5$ for $1 \le t \le 200$; $p_t = 0.6$ for $201 \le t \le 500$; $p_t = 0.8$ for $501 \le t \le 550$; $p_t = 0.55 + (t - 550)/3000$ for $551 \le t \le 1203$. (Note that a constant value of n_t is a simplification—in the real dataset n_t depends upon t.)

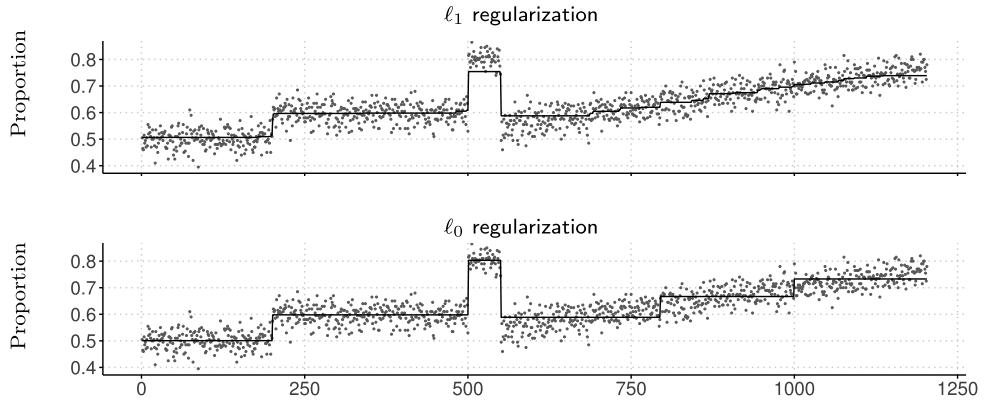

FIG. 3. *Estimators obtained from Problem* (2.4) *with* ℓ_1 (*upper panel*) *and* ℓ_0 (*lower panel*) *regularization. The data is synthetic and the underlying signal contains two sharp jumps* (*on the left*) *and a gradual increasing trend* (*on the right*). *We use cross-validation to select a value of* λ. *The* ℓ_1 *penalty shrinks the estimated probability during a big burst* ($501 \leq t \leq 550$) *and leads to more jumps during the gradual increase period* ($551 \leq t \leq 1203$). *The* ℓ_0*-based estimator leads to a better estimate of the signal burst and leads to fewer jumps during the gradual increase period*. .

estimators for the real datasets studied herein, where there is obvious model misspecification. Figure 3 presents the signal estimates (for both the ℓ_0 and ℓ_1 penalties) at the cross-validated choices of the tuning parameter; we use k-fold (with $k = 10$) cross validation (Hastie, Tibshirani and Friedman (2009)) which is also used in the `R` package `genlasso` (Since we want to ensure each fold is representative of the time series, instead of randomly assigning points to a fold, we systematically assign points by placing every kth point into the same fold). For both schemes the estimated signals $\{\hat{p}_t\}$ serve as good (overall) approximations of $\{p_t\}$; however, there are some important differences. First of all, the ℓ_1-segmentation scheme leads to biased estimates, and the bias becomes quite prominent in estimating the jump at the centre of the signal. This behavior is not present for the ℓ_0-scheme. In addition, the estimates for the linear component (at the right) also differ across the ℓ_0 and ℓ_1 schemes. The ℓ_0 regularizer leads to a fewer number of segments (here, three), compared to the ℓ_1-penalty which has several smaller jumps.

2.2.3.2. *Illustration on TAGS series*. Figures 4 (TAGS UNGA) and 5 (TAGS VS) show the estimated signal proportions obtained via estimator (2.4). Both of the penalty functions do a good job in estimating a piecewise constant version of the underlying signal. The ℓ_0 scheme leads to fewer jumps than its ℓ_1 counterpart for a comparable data-fidelity. The figures also show fitted signals for a few other values of λ around the cross-validated choice at the center[11] (λ increases as one moves down the rows); we include the tuning parameter selected by the one-standard error rule (Hastie, Tibshirani and Friedman (2009)) (see also the `R` package `genlasso`). We can see that, as λ decreases, the algorithm captures a more granular structure of the data and estimates more jumps.

Figure 4 shows communication traffic corresponding to TAGS specific to the U.N. General Assembly. The cyclical jumps correspond to the regular fall meetings of the General Assembly. In addition, our signal estimate suggests additional peaked activities, for example, a jump in April–May 1974. It seems that this jump is of smaller intensity compared to the other regular peaks corresponding to the annual meetings. Further investigation revealed that

[11]For Figure 4 the λ values were: (a) for the ℓ_0-penalty: 2.2, 4.8 and 13.7 (top to bottom) and (b) for the ℓ_1-penalty: 39.2, 66.3 and 189 (top to bottom). For Figure 5, the λ values were: (a) for the ℓ_0-penalty: 2.9, 8.1 and 23.2 (top to bottom) and (b) for the ℓ_1-penalty: 86.1, 245.7 and 701 (top to bottom).

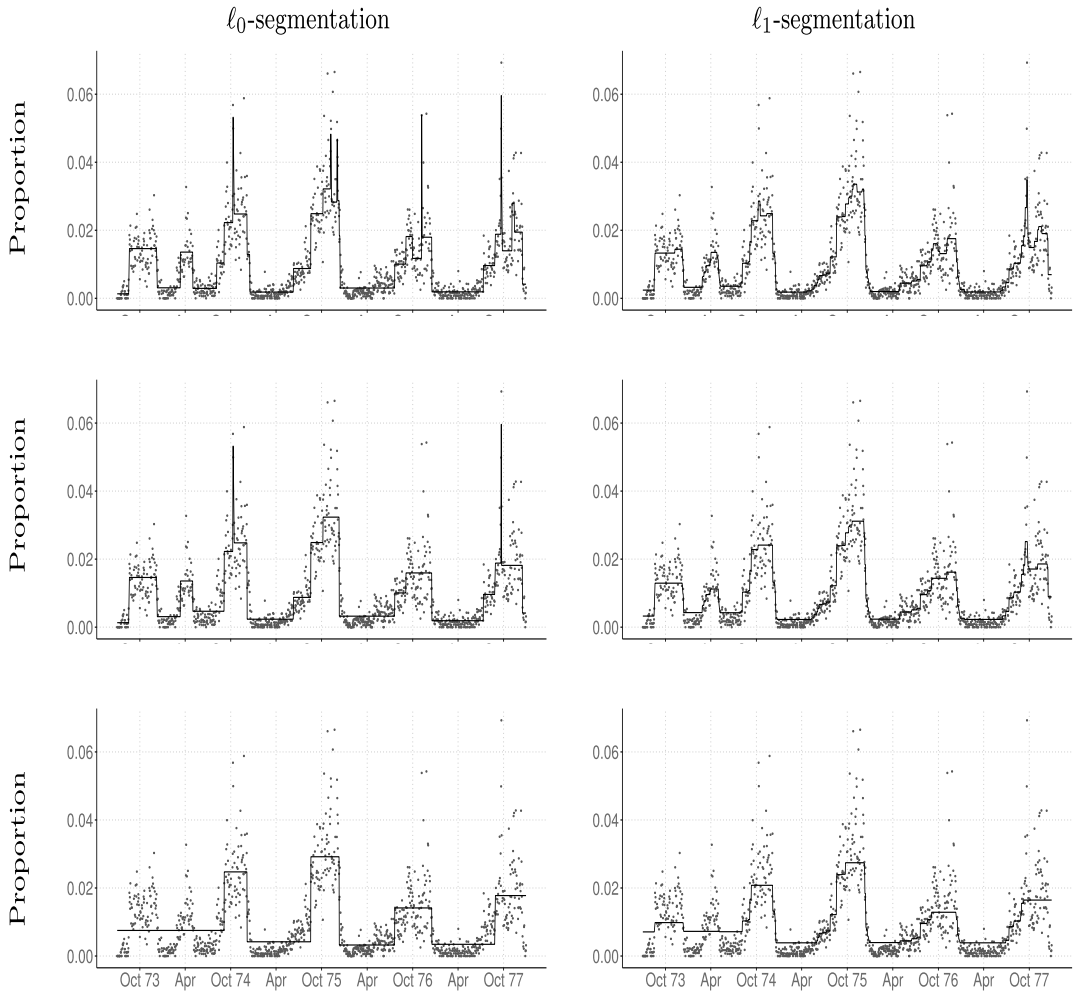

FIG. 4. *Figure showing the raw proportions* (*in blue dots*) *for TAGS UNGA* (*U.N. General Assembly*) *and the estimated proportions* $\hat{p}_t, t \geq 1$, *as obtained from the regularization framework in Problem* (2.4). *The left panel shows the estimates obtained with the* ℓ_0*-segmentation penalty, and the right panel shows the estimates with the* ℓ_1*-segmentation penalty. The middle rows correspond to the optimal* λ (*as discussed in the text*). *It shows how, in between the cyclical jumps in U.N.-related communications relating to the regular fall meetings of the General Assembly, there was also a jump in April–May* 1974. *This occurred when Algeria called a special session to demand U.N. support for a "New International Economic Order." We show a few additional choices of the regularization parameter for each example*.

this jump occurred when Algeria called a special session to demand U.N. support for a "New International Economic Order." In Figure 5 (TAGS related to South Vietnam) we observe that the most intense jump corresponds to the fall of Saigon and the end of the Vietnam War. The signal estimate suggests that this event is accompanied by a peak in communication traffic in April 1975—this happened with the collapse of the South Vietnamese regime and the rush to evacuate American personnel. A social scientist might also be interested to identify and explore smaller jumps; here, they correspond to the refugee crisis that continued into the following year.

2.3. *A deeper investigation of jumps*. The framework of Section 2.2 can be used to obtain a simple piecewise constant approximation of the underlying signal. Upon further investigation they offer insights into the behavior of communication streams. Section 2.3.1 discusses

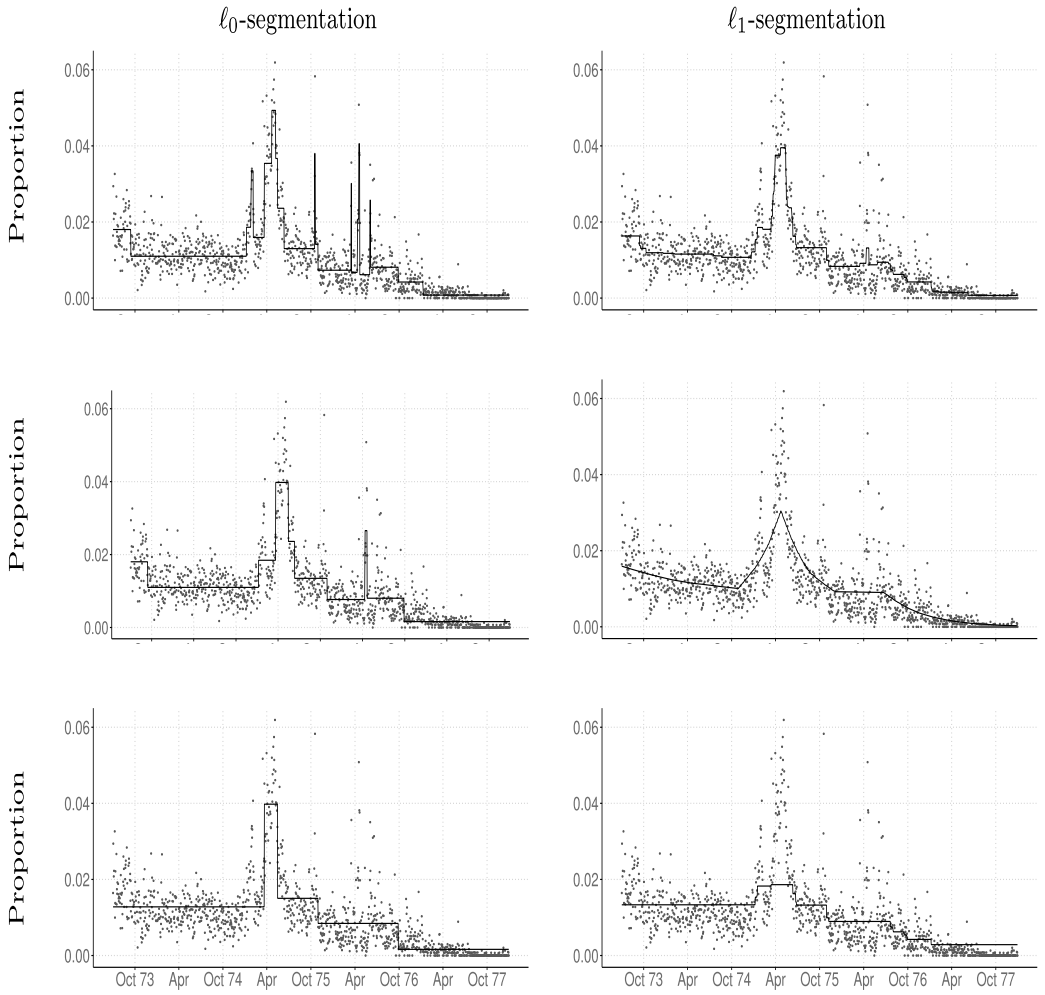

FIG. 5. *Figure showing the raw proportions* (*in blue dots*) *for TAGS VS* (*South Vietnam*) *and the estimated proportions* $\hat{p}_t, t \geq 1$, *as obtained from the regularization framework in Problem* (2.4). *The left panel shows the results for the* ℓ_0*-segmentation penalty and the right panel the* ℓ_1*-penalty. The middle rows correspond to the optimal* λ (*as discussed in the text*), *and we show a few additional choices of the regularization parameter for each example. The biggest burst corresponds to the fall of Saigon and the end of the Vietnam War. A social scientist might select one or another depending on whether they would want to identify smaller jumps that correspond, in this case, to the refugee crisis that followed the defeat of South Vietnam.*

how to quantify the intensity of a jump using sample splitting ideas (Wasserman and Roeder (2009)). In Section 2.4 we show how these jumps can be aggregated to obtain the notion of a "burst" (Kleinberg (2003)) in a communication stream. We also present related social science perspectives.

2.3.1. *How intense is a jump?*. A jump estimated by the ℓ_0 or ℓ_1-segmentation procedure may reflect: (a) a discontinuity in the signal, as we saw in the first half of Figure 3 (in this case the signal is well approximated by locally constant segments with pieces adapting to the data) and/or (b) a localized trend in the signal, as we saw in the second half of Figure 3. A jump in (b) is a consequence of the slope of $t \mapsto p_t$ and not a discontinuity. A piecewise constant signal can be considered as an approximation of the underlying (linear) trend in $t \mapsto p_t$. Given an estimate of $\{\hat{p}_t\}$, a scholar accustomed to analyzing events through a close reading of historical documents may ask:

- Which of these jumps might be important or are indicative of a historical event of interest?
- Can one obtain a rank ordering of the jumps based on their intensities?

We formalize this as follows: given an estimate $\{\hat{p}_t\}$ and a set of candidate jumps, can we obtain a scoring for their strengths and sizes? This would lead to a smaller set of jumps that merit closer scrutiny. Toward this end, we use a sample splitting[12] procedure: a subsample of size 50% of the data is used for estimating the location of the jumps, and the remaining held out part of the data is used to associate a p-value score (the method is described below) to each jump that is identified in the first stage. In other words, the training set is created by randomly choosing half the cables for each day, with the remaining half set aside for testing purposes.

Suppose $\hat{t}$ is a candidate change point based on the first part of the sample (used for estimating the signal). We denote the time points on the left of $\hat{t}$ as $L(\hat{t})$ and those on the right of $\hat{t}$ as $R(\hat{t})$—these segments $L(\hat{t})$ and $R(\hat{t})$ do not contain any jumps. We assume that p_t for $t \in L(\hat{t})$ are all equal to $p(L, \hat{t})$, and p_t for $t \in R(\hat{t})$ are all equal to $p(R, \hat{t})$. We test the null hypothesis (H_0) that the proportions on the left and right parts of $\hat{t}$ are equal, $p(L, \hat{t}) = p(R, \hat{t})$, vs. the alternative ($H_1$) that $p(L, \hat{t}) \neq p(R, \hat{t})$. We use the likelihood ratio test statistic for this purpose, where the null distribution is computed based on a permutation test.

Note that a candidate jump obtained at the cross-validated choice of the tuning parameter need not have a low p-value.[13] The p-values, thus obtained, can be used to: (a) devise a scoring mechanism to rank-order multiple jumps observed in a series and/or (b) prune out redundant jumps and identify ones that exhibit a significant difference in proportions between the left and right intervals. Scheme (b) is useful for estimators obtained by the ℓ_1-segmentation scheme, as this is known (Boysen et al. (2009), Killick, Fearnhead and Eckley (2012)) to make false discoveries of change-point locations (even if the underlying signal is piecewise constant). In our application the underlying signal is not piecewise constant—it simply serves as an useful approximation. In this case the p-value scores appear to measure the strength of a jump.

Figure 6 shows the communication stream for TAGS CVIS (Consular Affairs-Visas) and the estimated signal obtained via ℓ_0-segmentation. We also computed the p-value scores for

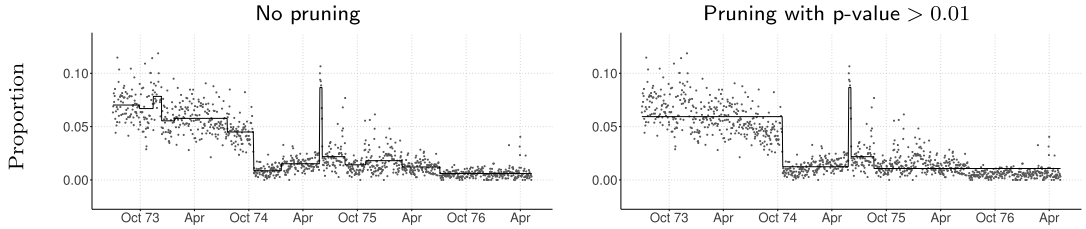

FIG. 6. *Figure showing the communication stream for TAGS CVIS—the estimated signal is obtained from the ℓ_0-segmentation scheme (at the cross-validated choice of λ). We compute the p-values (based on sample splitting, as described in Section 2.3.1) for every candidate jump location and prune the jump locations (and refit the signal with the new jump locations) based on the mentioned thresholds. The refitted signal is shown on the right panel.*

[12] Since we have a large number of samples (or cables), the size of the training set after sample splitting is still quite large.

[13] A jump obtained from the estimated $\{\hat{p}_t\}$ may be due to a linear rise in the signal which need not correspond to a significant change in local proportions. Our experiments indicate that jumps in $\{\hat{p}_t\}$, that correspond to gradual linear rises in the signal, have higher p-values associated with them when compared to sudden or abrupt changes in $\{\hat{p}_t\}$.

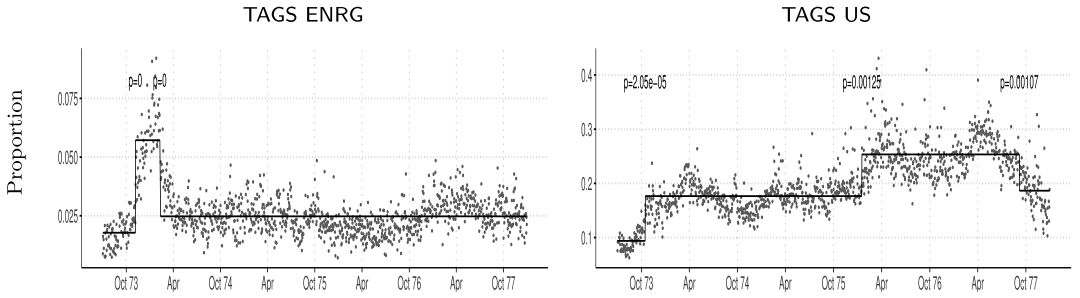

FIG. 7. *TAGS ENRG* (*energy*) *and TAGS US* (*for cables relating to the U.S.*) *with p-values associated with the estimated jumps* (*using the framework in Section* 2.3.1). *The jumps for ENRG are much sharper and indicate rapid* (*though not instantaneous*) *changes in mean*, *starting with the* 1973 *OPEC oil embargo*. *This gives them low p-values*. *In contrast*, *the two jumps to the right of the signal for US are during less rapid changes in mean and thus have slightly larger p-values* ($\sim 10^{-3}$). *Figure* 8 *presents more flexible fits of the underlying signal*. (*The notation* $p = x$ *is a shorthand for p-value being equal to* x.)

each potential jump location, as suggested via the ℓ_0-segmentation fit. The sharp jump in TAGS CVIS that persists across the panels, starts around mid September 1975, and ends around early October 1975. This spike is rather curious from a social science perspective. Prior to this period the number of communications related to visa applications sharply decreased not because this kind of activity diminished, but because archivists decided to stop retaining these records. The exception is a two-week period in September 1975, when the number of CVIS records is comparable or even higher than before. Some concern Indochinese refugees, but many others involve the FBI, especially visas for people coming from the U.S.S.R. and other communist states. Few, if any, of these FBI communications have message text. This is intriguing given the (limited) information available to us from the declassified cable documents and suggests the need for further investigation. This is the kind of anomaly that a social scientist might miss without using this kind of statistical framework.

Figure 7 presents two examples for TAGS ENRG (energy) and TAGS US (for cables relating to the United States). The jumps in TAGS ENRG are much sharper and indicate rapid (though not instantaneous) changes in mean, starting with the 1973 OPEC oil embargo. For the TAGS US figure, the p-values are indicative of whether a jump is due to a shift in the piecewise constant level or a linear trend—the p-values are larger when there is a linear trend rather than a sharp jump (as in a piecewise constant signal). We illustrate the intuition conveyed above via a synthetic example; see Figure 1 in the Supplementary Material (Gao et al. (2020)).

2.4. *From jumps to bursts*. We discuss how to summarize a single communication stream (corresponding to a specific TAGS) with a score that aggregates different jumps into a "burst;" this terminology was introduced by Kleinberg in the context of event detection. Informally speaking, a burst corresponds to a stretch of time where a communication stream depicts traffic larger than a baseline value. The approach we present here is a bit different from the work of Kleinberg (see also discussion in Section 2.2.2), who uses a Binomial model with two states at every time point. Kleinberg defines the weight of a burst to be the aggregated sum of the differences in the loss function (data fidelity term) across these two states. In our approach, we do not restrict p_t to two a priori specified states—instead, we allow for a continuum of values that are obtained by solving a regularized signal segmentation problem.

2.4.1. *Computation of the strength of a burst*. Suppose we are given an estimate of a baseline proportion p_0 (we discuss how to compute this below) for a communication stream.

A "burstiness period" or, simply, "burst" corresponds to a time interval where the estimated signal lies above the baseline value p_0 and is given by $\mathsf{T} = [t_{\text{start}}, t_{\text{end}}]$, where $\hat{p}_t > p_0, \forall t \in \mathsf{T}$. Inspired by Kleinberg (2003), we define the strength $S(\mathsf{T})$ of the burst as the logarithm of the likelihood ratio (here, the numerator is the likelihood of the signal and the denominator is that evaluated at the baseline) given by $S(\mathsf{T}) = \sum_{t\in\mathsf{T}}(\log L(\hat{p}_t|n_t, y_t) - \log L(p_0|n_t, y_t))$, where $L(\hat{p}_t|n_t, y_t)$ denotes the likelihood at time t. As the baseline p_0 is specific to a communication stream, the score $S(\mathsf{T})$ represents a deviation from this global baseline. $S(\mathsf{T})$ is different than the magnitude of a jump given by $\hat{p}_{t+1} - \hat{p}_t$; it takes into account the deviation of $\hat{p}_t$ from the baseline p_0 as well as the duration of the burst given by the length of T. A large value of $S(\mathsf{T})$ means that a large part of the likelihood is explained by deviations from the baseline and, therefore, corresponds to a strong burst. Note that each TAGS-specific communication stream can have multiple bursts leading to multiple intervals T, each with an assigned strength $S(\mathsf{T})$.

2.4.1.1. *Choice of baseline.* The baseline value p_0 should be representative of the behavior of the TAGS-specific communication stream. The global proportion of a communication stream is a reasonable choice. We set p_0 to be one standard deviation larger than the global proportion

$$p_0 = \bar{p} + \sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}, \quad \text{where, } \bar{p} = \frac{\sum_{t=1}^{N} y_t}{\sum_{t=1}^{N} n_t}, \bar{n} = \frac{1}{N}\sum_{t=1}^{N} n_t.$$

A robust estimate like the median can also be used instead of the average. In our experiments we found that the top-ranked slots (cf Table 1) were relatively agnostic to the choice of the baseline p_0.

2.4.2. *Interpretation of bursts.* Table 1 presents the top thirty bursts with the start and end dates as well as the date with the highest burst strength score. A close study of the content of the cables shows that not all of these bursts correspond with what scholars would recognize as an event of historical importance. After all, the cable TAGS that diplomats used do not necessarily correspond with diplomatic activity. For instance, the second biggest burst is made up of cables related to transportation (ETRN)—a TAGS that was commonly used, and overused, from when we begin to have records continuing until 1974, when diplomats' use of this TAGS was largely discontinued. The biggest burst, for CVIS (visas), has a similar pattern (as shown in Figure 6). But in this case, it appears to reflect a decision by archivists to stop preserving records related to visas (Langbart, Fischer and Roberson (2007)). To the model, both of these look like bursts, but they simply reflect administrative procedures rather than historical events.

The bursts that follow, on the other hand, appear to correspond well with historical events. The next 10 include the Carter administration's prioritization of human rights (SHUM), Anwar Sadat's surprise visit to Israel (PGOV), the Southeast Asian Boat People crisis (SREF), the U.S. withdrawal from the International Labor Organization (PORG), the conclusion of the Panama Canal Treaty (PDIP), the 1973 Yom Kippur War (XF, for Middle East), Portugal's withdrawal from Angola (AO) and the 1974 crisis over Cyprus (CY).

To validate these results, we consulted four standard reference works on U.S. foreign relations (Brune and Burns (2003), Flanders and Flanders (1993), Bruce et al. (1997), De Conde et al. (2002)). The editors are all domain experts, but the varying content of each one reflects the different ways social scientists evaluate historical significance. Nevertheless, all of the aforementioned events appear in every one of these reference works, suggesting our framework succeeds in identifying events that are broadly recognized as historically important.

TABLE 1

Top 30 *bursts identified using* ℓ_0 *segmentation algorithm, using the method in Section* 2.4 *to compute burst strengths. For interpretations regarding the bursts, please see the discussion in Section* 2.4.2

	TAGS	Meaning	start	end	peak	Burst Strength
1	ETRN	Economic Affairs-Transportation	1973-07-02	1974-08-09	1973-09-28	5146.05
2	CVIS	Consular Affairs-Visas	1973-07-02	1975-01-02	1974-06-28	4839.35
3	SHUM	Social Affairs-Human Rights	1977-01-19	1977-12-30	1977-11-18	2872.02
4	US	United States	1976-01-28	1977-09-16	1976-04-15	2516.03
5	PGOV	Political Affairs-Government	1977-06-03	1977-12-30	1977-11-18	2484.57
6	SREF	Social Affairs-Refugees	1975-04-22	1976-07-20	1976-06-02	1662.58
7	SOPN	Social Affairs-Public Opinion and Information	1976-11-26	1977-12-30	1977-08-26	1597.14
8	PORG	Political Affairs-Policy Relations With International Organizations	1977-06-15	1977-12-30	1977-11-11	1547.35
9	PDIP	Political Affairs-Diplomatic and Consular Representation	1977-05-24	1977-12-30	1977-09-02	1462.93
10	XF	Middle East	1973-10-09	1973-12-19	1973-10-16	1453.76
11	AO	Angola	1975-11-08	1976-02-23	1975-11-10	1439.58
12	CY	Cyprus	1974-07-15	1974-07-29	1974-07-20	1378.79
13	VM	Vietnam	1977-10-11	1977-12-30	1977-10-12	1365.45
14	PDEV	Political Affairs-National Development	1977-06-13	1977-12-30	1977-08-31	1344.70
15	VS	Vietnam (South)	1973-07-02	1975-06-06	1975-04-25	1150.46
16	UNGA	UN General Assembly	1975-08-19	1975-12-13	1975-11-07	1044.29
17	CARR	Consular Affairs-Americans Arrested Abroad	1977-06-01	1977-12-30	1977-06-28	951.98
18	MCAP	Political Affairs-Military Capabilities	1973-07-02	1974-08-15	1974-07-03	903.26
19	ENRG	Economic Affairs-Energy	1973-11-08	1974-02-21	1974-01-25	760.64
20	PBOR	Political Affairs-Boundary and Sovereignity Claims	1977-07-01	1977-12-30	1977-11-09	685.75
21	OVIP	Operations-VIP Travel Arrangements	1974-10-09	1974-11-09	1974-10-31	607.17
22	RH	Rhodesia	1976-09-01	1977-12-30	1977-08-31	569.89
23	AEMR	Administration-Emergency and Evacuation	1975-03-28	1975-05-12	1975-04-28	524.59
24	MPLA	Popular Movement for the Liberation of Angola	1975-11-07	1976-02-24	1976-02-18	507.98
25	MSG	Marine Security Guards	1976-09-02	1977-12-30	1977-11-28	507.27
26	OREP	Operations-Congressional Travel	1976-10-27	1976-11-18	1976-11-02	481.08
27	PRG	Provisional Revolutionary Government of South Vietnam	1975-01-16	1975-02-06	1975-02-03	470.51
28	MNUC	Military and Defense Affairs-Military Nuclear Applications	1977-03-11	1977-12-30	1977-08-22	421.53
29	UNGA	UN General Assembly	1974-09-05	1974-12-05	1974-10-10	417.20
30	CB	Cambodia (Khmer Republic)	1973-07-02	1975-05-21	1975-04-16	370.14

A systematic evaluation of hundreds of bursts for historical significance lies outside the scope of this paper. But the relative proportion of recognized historical events appears to diminish as one examines smaller bursts, like the ones ranked in the range 13–22. They include the denouncement of the Vietnamese War (VM and VS), the OPEC oil embargo (ENRG), the Vladivostok summit (OVIP) and negotiations to end white rule in Rhodesia (RH). This

suggests that the ranking of the bursts, while not necessarily corresponding to historical importance, does reflect the likelihood that each one will correspond to events historians have already recognized as significant. But among the unrecognized events, like a 1975 U.N. General Assembly debate over the command of foreign military forces in South Korea, there are some that appear to merit closer scrutiny. The identification of such unstudied episodes, no less than rank-ordering well-known events, is valuable for historical scholarship.

3. A generalization beyond piecewise constant segments. A major focus of Section 2 was on approximating a communication stream with a piecewise constant signal. This framework does help us answer some key data-driven questions of interest to a political scientist, based on a first order (i.e., piecewise constant) approximation of the communication streams. However, as we discussed before, many of the signals are not piecewise linear. We now investigate more flexible signal approximations that provide us insights into the finer behavior of the signals. A natural extension of a piecewise constant estimate $\{\hat{p}_t\}$ is a piecewise linear estimate. However, we need to address certain technical issues to incorporate this structure into our likelihood framework, as we discuss below.

Let us consider the usual signal denoising problem with data: $\tilde{y}_i = \mu_i + \epsilon_i$, for $i = 1, \ldots, N$ where, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Suppose we would like to estimate $\boldsymbol{\mu}$ such that it is piecewise linear. A method to achieve this is by using the ℓ_1 trend-filtering approach (Kim et al. (2009), Tibshirani (2014)). Here, one uses a convex regularizer $H^{\text{tf}}_{\ell_1}(\boldsymbol{\mu}) = \sum_t |\mu_{t+2} - 2\mu_{t+1} + \mu_t|$ to obtain a signal with piecewise linear segments,

$$\underset{\boldsymbol{\mu}}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{N}(\tilde{y}_i - \mu_i)^2 + \lambda H^{\text{tf}}_{\ell_1}(\boldsymbol{\mu}). \tag{3.1}$$

The penalty function $H^{\text{tf}}_{\ell_1}(\boldsymbol{\mu})$ encodes the ℓ_1-norm on the discrete second order derivative of the signal $\{\mu_t\}$ assuming that the time points are all equally spaced. $H^{\text{tf}}_{\ell_1}(\boldsymbol{\mu})$ can be interpreted as a convexification of its ℓ_0 version: $H^{\text{tf}}_{\ell_0}(\boldsymbol{\mu}) = \sum_t \mathbf{1}(\mu_{t+2} - 2\mu_{t+1} + \mu_t \neq 0)$ that counts the number of different piecewise linear segments.

Our situation is different from the denoising example (with least squares loss), as outlined above. Since we are working under the modeling assumption: $(y_t|n_t, p_t) \sim \text{Bin}(n_t, p_t)$ with $p_t = \exp(\theta_t)/(1 + \exp(\theta_t))$, imposing a trend filtering penalty on p_t will lead to a difficult nonconvex optimization problem due to the nonlinear dependence of p_t on θ_t. Instead, we let the latent parameter $t \mapsto \theta_t$ be piecewise linear; this leads to a computationally tractable estimation framework based on convex optimization. Toward this end, we propose an adaption of Problem (2.4) by using the regularizer $H(\boldsymbol{\theta}) = H^{\text{tf}}_{\ell_1}(\boldsymbol{\theta})$:

$$\underset{\theta_t, 1\le t\le N}{\text{minimize}} \quad \sum_{t=1}^{N}\big(-y_t\theta_t + n_t \log\big(1 + \exp(\theta_t)\big)\big) + \lambda H^{\text{tf}}_{\ell_1}(\boldsymbol{\theta}). \tag{3.2}$$

Figure 8 shows the results of estimates obtained from some communication streams using the ℓ_1-trend filtering penalty. If the time points are not equally spaced, then this penalty needs to be modified; see, for example, Kim et al. (2009) and also Section 5 in the Supplementary Material (Gao et al. (2020)).

Computation. The proximal gradient-stylized update (2.8) can be adapted to the setting described above with $H(\boldsymbol{\theta}) = H^{\text{tf}}_{\ell_1}(\boldsymbol{\theta})$. To solve the proximal operator, we use existing specialized solvers for the ℓ_1-trend filtering problem for the least squares loss function—in particular, we found the interior point solver[14] of Kim et al. (2009) to work quite nicely for

[14]We use the R package wrapper available from https://github.com/hadley/l1tf.

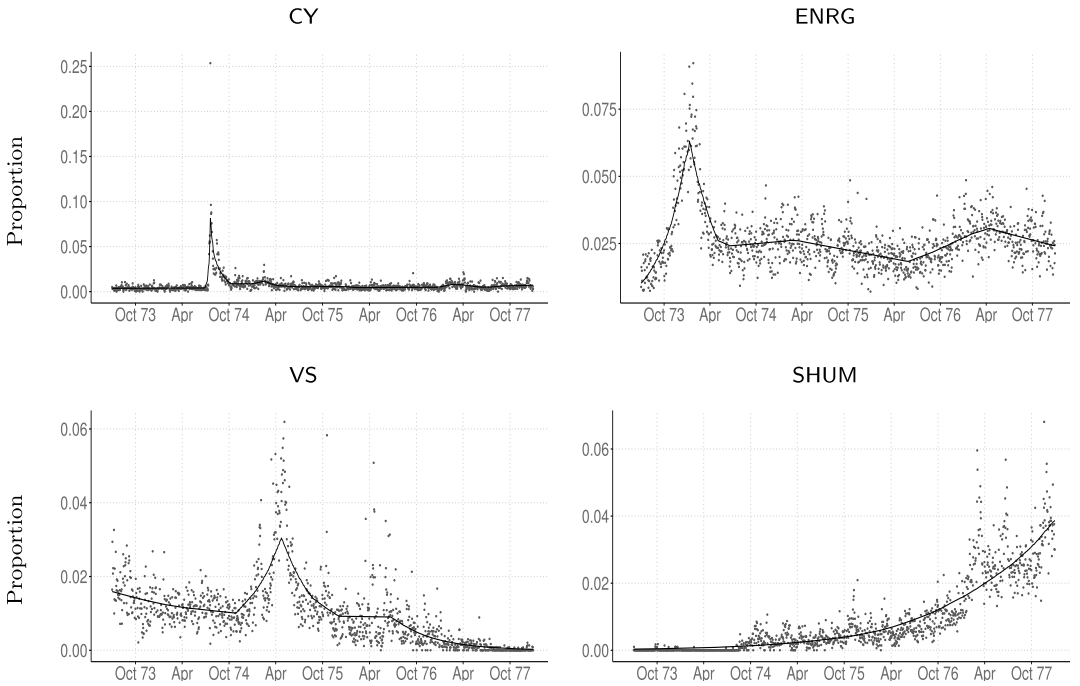

FIG. 8. *Figure showing the estimates obtained from Problem* (2.4) *with the* ℓ_1*-trend filtering regularizer* (*see Section* 3). *The sharp spike in the CY* (*Cyprus*) *communication stream corresponds to an unanticipated event*, *when Greek forces launched a coup with the goal of annexing Cyprus*. *The first peak for the second stream* (*ENRG*) *corresponds to the* 1973 *energy crisis*, *after the OPEC oil ministers announced an embargo during the Yom Kippur War. The peak for VS*, *for South Vietnam*, *corresponds to the Fall of Saigon in* 1975, *which marked the end of the Vietnam War. SHUM*, *for communications related to human rights*, *shows the increasing attention the State Department gave to this subject*, *especially after the election of President Jimmy Carter.* .

the problem sizes encountered in this paper. We note, however, that solving Problem (3.2) is computationally more demanding than the piecewise constant segmentation approach. The main difference is due to the proximal operator: Problem (2.11) can be solved much more efficiently than Problem (3.1). As Problem (2.4) is convex, the sequence (2.8) converges to a minimum of the optimization problem (note that the minimum exists under minor assumptions). The convergence rates outlined in (2.9) and (2.10) will also apply to this problem.

If we set $H(\boldsymbol{\theta}) = H^{\mathrm{tf}}_{\ell_0}(\boldsymbol{\theta})$, the resulting Problem (2.4) becomes a challenging nonconvex optimization problem—in this case, the proximal operator is harder to solve compared to the fused ℓ_0 problem (2.12). Hence, in this paper we limit our attention to the convex ℓ_1-trend filtering regularizer.

Social science interpretation. The shapes of the estimated signal (see Figure 8) capture key differences between different kinds of events in the history of U.S. foreign relations—these would have been less obvious using the piecewise constant signal approximation framework of Section 2.2. Some crises, like the Cyprus coup, occur with no warning, but also make little difference in the longer-term level of attention and activity. Others, like the OPEC oil embargo, are similarly unexpected, but signal the beginning of a period marked by moments of heightened activity well over the previous baseline. Still others, like the Fall of Saigon, build to a climax, and then gradually subside. Finally, the rise of human rights as a concern for policymakers is gradual but seemingly inexorable. The taxonomy of these different patterns provide useful insights to a social scientist in identifying and classifying different kinds of events. There are a few broad categories of patterns. From the sudden and unexpected, to the

gradual rise and fall, to longer-term trends, these patterns can help social scientists develop a taxonomy of historical events to gain a deeper understanding of heterogenous data.

4. Related work (event detection literature). We present a brief contextualization of our work in regard to the event detection literature within the computer science/data mining community. Event detection in text communication streams, such as news or social media platforms (e.g., Twitter), is a fairly rich area of research; see for example, the nice review by Atefeh and Khreich (2015) for an overview of event detection in Twitter communication streams. Starting with the seminal work of Kleinberg (2003), important contributions have been made in the field of feature-pivot techniques, where one attempts to identify context-specific words that depict a sharp rise in frequency as an event emerges. An important line of research in this area—see, for example, He, Chang and Lim (2007), Pui Cheong Fung et al. (2005)—focuses on how to understand contributions from different words toward describing an event. A key focus in this line of work is an elaborate design of specialized features to describe the content of messages (keywords, hashtags, advanced text/context based features, etc), interactions among users, etc. For many of these approaches, text-processing methods play an important role. The dataset we analyze in this paper and its application context is different from event detection in Twitter streams; the corpus we study has limited textual data. Nevertheless, we present a preliminary analysis of text-based data for our problem in Section 5.1.

The communication streams we study are naturally characterized by TAGS, defined by the State Department when they were entered into the system. Our emphasis is on the use of statistical methods that make our models and results interpretable to a social scientist (see Section 5 for further discussion). As mentioned by Atefeh and Khreich (2015), many of the complex methods used in the context of analyzing Twitter streams are based on an ad hoc selection of thresholds—this is understandable given the complexity of the problem. In contrast, such heuristics are mostly avoided in our approach.

5. Concluding remarks and discussion. In this paper we present statistical methods to analyze diplomatic cable communications during 1973–1977, recently made available by the U.S. National Archives. The complexity and heterogeneity of different historical events present themselves in the form of various geometric patterns in the communication streams. An important challenge of this work has been in identifying and proposing a suitable suite of statistical tools useful to a social scientist to glean insights from this newly available data. Our study focuses on the TAGS-specific communication streams and understanding how geometric characteristics of these streams relate to events of historical significance.

We present a global testing framework to identify which, among potentially thousands of communication streams, exhibit interesting statistical activity that merit further downstream analysis. We propose signal segmentation methods based on ℓ_1 and ℓ_0 penalization to identify structural breaks, a.k.a. jumps in the communication streams. The proposed algorithms are complementary to the area of change-point models in statistics and appear to be faster than existing implementations. We present a sample-splitting framework to perform statistical inference on the detected jumps and discuss a simple but effective notion of combining jumps to bursts, following Kleinberg (2003). Finally, we present extensions of piecewise constant signals to model the underlying process of communication stream traffic.

5.1. *A social science perspective.* Results available from our statistical analysis in some cases correspond to well-known events, while in other cases they lead to curious findings that a social scientist might have missed in the absence of statistical tools such as the ones presented herein. The tools proposed here may be used (with suitable modifications) in other

contexts, beyond the dataset analyzed in this paper. With the ever growing volume of digital content, it will become increasingly important for social scientists to devise a range of new methods to identify patterns and anomalies. When a corpus consists of hundreds of millions of emails—as is the case with the Obama White House files—in the absence of a suitable statistical framework, it may be challenging to filter out less interesting communications without obscuring unexpected and potentially important information. Historical events are quite heterogeneous, and social scientists need methods, such as the ones presented herein, that can help them prioritize which communications they should examine most closely.

An important characteristic of our framework—of particular appeal to a social scientist—is that it permits efficient algorithmic processing of large corpora. Our methodology does not require prior knowledge of the content of the communications; it is thus agnostic to specific biases that may be imposed by a practitioner. At the same time, our proposed framework is flexible enough to be able to capture very different kinds of events. Our statistical methods also present a clear guideline of what geometric patterns these events might correspond to. Throughout our project we have observed an interesting synergy between statistics and social science perspectives during the interpretation of the findings.

Our results in Sections 2.3 and 3 suggest how a social scientist may use our framework to identify and classify different kinds of events. Some of these events can be described by jumps (piecewise constant signals), while others are better explained by piecewise linear segments. While each one is unique, there are clearly different classes, from the sudden and unexpected, to the gradual rise and fall, to longer-term trends, to cyclical patterns. Categorizing these different phenomena provides a useful heuristic in analyzing heterogenous data and could help social scientists develop a taxonomy of historical events. Developing a comprehensive statistical framework to perform shape-based grouping of these taxonomies is an interesting topic for future research.

Our framework is perhaps most important in supporting a new, more inductive approach to historical analysis, where it may not be possible to know in advance what communications should be examined more closely. Since the "archive" is increasingly digital and archivists are no longer able to create traditional finding aids to guide researchers, it is increasingly difficult to decide which events or trends were most important. Our framework makes it possible to start instead with the data and then conduct a statistical analysis of communications that deviate from established communications patterns.

Inductive and deductive approaches are not mutually exclusive. For instance, we asked a historian, who had just published a history of the 1970s, to independently identify and rank the most important events. There was substantial overlap between his list and the one generated by our methodology. But there were also interesting differences which raised what he called "crucial questions" about how we judge historical significance:

"Is that most vital quality to be assessed only in [the] perspective of hindsight, or can we use quantitative aggregation of contemporary data to achieve novel perspectives?"

Our framework can already help researchers meet the challenge of exponentially larger archives and do what machines can never do: interpret complex data, assess historical significance and determine what this history really means for the present and the future.

Using text data to understand regime changes. While our work has focused on cable metadata and communication volume, similar methods could be applied to textual data. Unfortunately, the text is unavailable for many thousands of the cables in this corpus, either because State Department storage systems failed to preserve it or because only the metadata has been declassified for cables that have been deemed to contain sensitive information. But we performed a small-scale analysis to explore what might be learned from what data is available.

We studied communications relating to several regime changes in the 1970s by using the name of the incoming leader as a guiding feature. In this analysis we looked at the weekly

proportion of cables with the country in question (CY for Cyprus, CI for Chile, AR for Argentina and PK for Pakistan) which also contain the name of the leader or leaders of the coup in the text (SAMPSON for CY, PINOCHET for CI, VIDELA or GUZZETTI for AR, and ZIA-UL-HAQ or ZIA for PK). Documents with unavailable text data were not included in the analysis. We applied our framework to detect potential change points (by restricting the number of jumps to at most two). Our results appear in Figure 3 of the Supplementary Material (Gao et al. (2020)). In three of the four cases we examined, Pakistan, Cyprus and Argentina, our framework clearly shows the change in communications patterns that corresponds with a coup. This is, as opposed to, for example, South Vietnam, where the peak corresponding to the fall of Saigon is preceded by a long period of increasing activity. Although all three do have cables mentioning the coup leader before the event, these mentions are not persistent enough to shift the location.

Chile presents an interesting exception, in which the change point appears well after the regime change. There is increased interest in Pinochet leading up to it, but U.S. diplomats did not recognize he was a key figure in military plotting against the Allende government. The Nixon administration faced accusations of backing the coup once it succeeded, leading it to be cautious initially in contacts with the new leaders. That did not start to change until the spring of 1974 (Harmer (2011)).

Even though the textual data for the cable corpus is incomplete, rich textual data is available for other corpora necessitating an in-depth analysis of text processing methods. There is already a large body of work in the context of event detection in Twitter data (Atefeh and Khreich (2015)) using text processing tools. We leave the combination of statistical analysis of communications streams and techniques from natural language processing as topics of future research.

Acknowledgments. The authors would like to thank the Editor and anonymous reviewers for helpful comments that substantially improved the paper. R. Mazumder was supported by ONR (grant # N000141512342) and NSF-IIS (grant # 1718258); M. Connelly was supported by NSF-RIDIR (grant # 1637159). The authors would like to thank David Blei, David Madigan and Shawn Simpson for helpful comments and encouragement and Raymond Hicks for his help in providing us access to the text data used in Section 5. The authors will also like to thank the workshop participants of "Famine and Feast–International Historical Research in the Digital Age" (London, U.K.; 2015); and seminar participants at the U.S. Census Bureau (2018) for comments on the work. Preliminary results from this article appeared in a media article on `buzzfeed.com`.[15]

SUPPLEMENTARY MATERIAL

Supplement to "Mining events with declassified diplomatic documents" (DOI: 10.1214/20-AOAS1344SUPPA; .pdf). In the Supplementary Material, we provide additional discussion for (i) the derivation of $\mathcal{T}$ in (2.1); (ii) sensitivity analysis of the global testing results in Section 2.1; (iii) the comparison of the ℓ_0 segmentation approach we used for Problem (2.4) versus PELT (Killick, Fearnhead and Eckley (2012)); (iv) discussion of local p-values on synthetic data; and (v) modification of Problem (2.8) to handle irregularly spaced time points.

Supplement to "Mining events with declassified diplomatic documents" (DOI: 10.1214/20-AOAS1344SUPPB; .zip). This supplement contains `R` scripts of the fused ℓ_0 and ℓ_1-penalized problems discussed in the paper.

[15] Article www.buzzfeed.com/josephbernstein/can-a-computer-algorithm-do-the-job-of-a-historian?

REFERENCES

ATEFEH, F. and KHREICH, W. (2015). A survey of techniques for event detection in Twitter. *Comput. Intell.* **31** 132–164. MR3318075 https://doi.org/10.1111/coin.12017

AUGER, I. E. and LAWRENCE, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **51** 39–54. MR0978902 https://doi.org/10.1016/S0092-8240(89)80047-3

BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527 https://doi.org/10.1137/080716542

BEIELER, J., BRANDT, P. T., HALTERMAN, A., SCHRODT, P. A. and SIMPSON, E. M. (2016). Generating political event data in near real time: Opportunities and challenges. In *Computational Social Science* (R. M. Alvarez, ed.) 98–120. Cambridge Univ. Press, Cambridge. https://doi.org/10.1017/CBO9781316257340.005

BELLMAN, R. and ROTH, R. (1969). Curve fitting by segmented straight lines. *J. Amer. Statist. Assoc.* **64** 1079–1084. MR0246456

BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. MR3476618 https://doi.org/10.1214/15-AOS1388

BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348 https://doi.org/10.1214/07-AOS558

BRODSKY, B. E. and DARKHOVSKY, B. S. (1993). *Nonparametric Methods in Change-Point Problems*. *Mathematics and Its Applications* **243**. Kluwer Academic, Dordrecht. MR1228205 https://doi.org/10.1007/978-94-015-8163-9

BRUCE, W., JENTLESON, PATERSON, T. G. and RIZOPOULOS, N. X. (1997). *Encyclopedia of US Foreign Relations* **2**. Oxford Univ. Press, New York.

BRUNE, L. H. and BURNS, R. D. (2003). Chronological History of U.S. Foreign Relations: 1933–1988. Routledge. ISBN 9780415939164. Available at https://books.google.com/books?id=nSe0gs-YaTkC.

DE CONDE, A., BURNS, R. D., LOGEVALL, F. and KETZ, L. B. (2002). *Encyclopedia of American Foreign Policy*. Scribner's, New York.

FLANDERS, S. A. and FLANDERS, C. N. (1993). Dictionary of American Foreign Affairs. Macmillan Library Reference.

GAO, Y., GOETZ, J., CONNELLY, M. and MAZUMDER, R. (2020). Supplement to "Mining events with declassified diplomatic documents." https://doi.org/10.1214/20-AOAS1344SUPPA, https://doi.org/10.1214/20-AOAS1344SUPPB

GLAZ, J., NAUS, J. and WALLENSTEIN, S. (2001). *Scan Statistics*. *Springer Series in Statistics*. Springer, New York. MR1869112 https://doi.org/10.1007/978-1-4757-3460-7

HANNA, A. (2014). Assessing gdelt with handcoded protest data. (accessed: July 29, 2016). www.badhessian.org.

HARMER, T. (2011). *Allende's Chile and the Inter-American Cold War*. Univ. of North Carolina Press, Chapel Hill, NC.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HAZIMEH, H. and MAZUMDER, R. (2018). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Preprint. Available at arXiv:1803.01454.

HE, Q., CHANG, K. and LIM, E.-P. (2007). Analyzing feature trajectories for event detection. In *Proceedings of the* 30*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 207–214. ACM, New York.

JACKSON, B., SCARGLE, J. D., BARNES, D., ARABHI, S., ALT, A., GIOUMOUSIS, P., GWIN, E., SANGTRAKULCHAROEN, P., TAN, L. et al. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.* **12** 105–108.

JENKINS, J. C. and MAHER, T. V. (2016). What should we do about source selection in event data? Challenges, progress, and possible solutions. *Int. J. Sociol.* **46** 42–57.

JOHNSON, N. A. (2013). A dynamic programming algorithm for the fused lasso and L_0-segmentation. *J. Comput. Graph. Statist.* **22** 246–260. MR3173713 https://doi.org/10.1080/10618600.2012.681238

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418 https://doi.org/10.1080/01621459.2012.737745

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 https://doi.org/10.1137/070690274

KLEINBERG, J. (2003). Bursty and hierarchical structure in streams. *Data Min. Knowl. Discov.* **7** 373–397. MR2011139 https://doi.org/10.1023/A:1024940629314

LANGBART, D., FISCHER, W. and ROBERSON, L. (2007). The national weights and measures laboratory. appraisal of records covered by N1-59-07-3-P Technical report, Tech. rep, College Park: National Archives.

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931 https://doi.org/10.1214/aos/1034276635

MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). *SparseNet*: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. MR2894769 https://doi.org/10.1198/jasa.2011.tm09738

MAZUMDER, R., RADCHENKO, P. and DEDIEU, A. (2017). Subset selection with shrinkage: Sparse linear modeling when the snr is low. Preprint. Available at arXiv:1708.03288.

NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization*: *A Basic Course*. *Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598 https://doi.org/10.1007/978-1-4419-8853-9

NESTEROV, YU. (2013). Gradient methods for minimizing composite functions. *Math. Program.* **140** 125–161. MR3071865 https://doi.org/10.1007/s10107-012-0629-5

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics* **5** 557–572.

PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100–115. MR0088850 https://doi.org/10.1093/biomet/41.1-2.100

PUI CHEONG FUNG, G., XU YU, J., YU, P. S. and LU, H. (2005). Parameter free bursty events detection in text streams. In *Proceedings of the* 31*st International Conference on Very Large Data Bases* 181–192. VLDB Endowment.

SCOTT, A. J. and KNOTT, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* 507–512.

TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487 https://doi.org/10.1214/13-AOS1189

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641 https://doi.org/10.1111/j.1467-9868.2005.00490.x

TRUONG, C., OUDRE, L. and VAYATIS, N. (2018). A review of change point detection methods. CoRR. Available at abs/1801.00718.

WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689 https://doi.org/10.1214/08-AOS646