Score-based Change Detection for Gradient-based Learning Machines

Lang Liu, Joseph Salmon*, Zaid Harchaoui

Technical Report no. 652
Department of Statistics
University of Washington

June 17, 2019

Abstract

The widespread use of machine learning algorithms calls for automatic change detection algorithms to monitor their behavior over time. As a machine learning algorithm learns from a continuous, possibly evolving, stream of data, it is desirable and often critical to supplement it with a companion change detection algorithm to facilitate its monitoring and control. We present a versatile score-based change detection method that can detect a change in any number of (hidden) components of a machine learning model. The proposed statistical hypothesis test can be readily implemented for any machine learning model implemented in a differentiable programming framework. We establish the consistency of the hypothesis test and show how to calibrate it based on the theoretical results. We illustrate the versatility of the approach on linear regression models, time series models, text topic models, and latent-variable models on synthetic and real data.

1 Introduction

Statistical machine learning models are now fostering progress in numerous technological applications, *e.g.,* visual object recognition, game playing, speech and language processing, as well as in many scientific domains, *e.g.,* astrophysics, genomics, neuroscience, and sociology. This progress has been fueled most recently by the flexible statistical machine learning modeling software libraries designed within a differentiable programming framework (*e.g.,* Pytorch [Paszke et al., 2017a] and TensorFlow [Abadi et al., 2015]).

First-order optimization algorithms such as accelerated batch gradient methods or averaged stochastic gradient methods are then perfectly adapted to this framework, opening up the possibility of gradient-based training of machine learning models from a continuous stream of data. As a machine learning algorithm learns from a continuous, possibly evolving, stream of data, it is desirable to supplement it with tools to facilitate its monitoring and control.

Recent remarkable failures of intelligent learning systems such as [Metz, 2018] or [Knight, 2018] as they were unleashed in the wild showed the need for such tools. As N. Wiener had foreseen

*CNRS, University Montpellier, Montpellier, France.

already in the 1960s "the very speed of operation of modern digital machines stands in the way of our ability to perceive and think through the indications of the danger". To pursue his argument, the speed of monitoring intelligent learning systems should be comparable with the one of training such systems.

Therefore, in order to be relevant, the same way the training of machine learning models is now automatic and effortless, the monitoring of machine learning models should be automatic and effortless. Humans monitoring machines should have at hand automatic monitoring tools to scrutinize the learned model as it evolves over time. Statistical decision theory and hypothesis testing to detect changes is the natural framework to design such tools.

We make the following contributions in this paper. We first introduce a generic change monitoring method called *autograd-test*. Then we show how to calibrate it to test for changes for a prescribed false alarm level. We establish the level consistency and the power consistency of the test. Finally we illustrate the versatility of the approach for time series models and text topic models.

Related work. Early works in statistical change detection considered so-called control charts such as the Shewart control chart [Shewhart, 1931] to continuously inspect a stream of datapoints under parametric assumptions of the distribution of the data. The popular CuSum method [Page, 1954, 1957] automatically adjusts the detection threshold then served as a basis for change detection methods in the parameters derived on a case by case basis for various statistical models; see [Hinkley, 1970, Deshayes and Picard, 1986, Basseville and Nikiforov, 1993] and references therein. These classical approaches are either based on (possibly generalized) likelihood ratios [Lorden, 1971] or on residuals and therefore not directly suited for statistical models implemented in a differentiable programming framework. Furthermore they are limited to low-dimensional statistical regimes where the sample size is much larger than the number of parameters to be estimated.

Change detection approaches based on the Fisher's efficient score statistic are more adapted to a differentiable programming framework, where the score statistic, *i.e.,* the gradient of the log-likelihood, can be readily computed. Such test statistics have received much less attention, limited to testing change in mean or other basic types of change, only in low-dimensional statistical regimes; see [Box and Ramírez, 1992, Horváth and Parzen, 1994, Luceño, 1999, Apley and Chin, 2007]. On the other hand, in [Enikeeva and Harchaoui, 2019] the authors consider the change in mean problem in multivariate Gaussian random vectors in a high-dimensional statistical regime, where changes may only occur on an unknown subset of the components. In this paper we show how to expand the application of score-based change detection test statistics in the context of differentiable programming and equip them with scanning procedures that allow them to tackle high-dimensional settings.

2 Score-based change detection

We first formalize the task of detecting changes in model parameters as a statistical hypothesis testing problem. We then construct three test statistics based on the score function that serve as building blocks for the proposed *autograd-test*. Finally, we discuss an adaptation to online settings.

Change detection and model components. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space, and $X_1, \ldots, X_n$ in $\mathcal{X}$ be a sequence of observations associated with a family of joint probability distributions $\{\mathbb{P}_\theta(X_1, \ldots, X_n) : \theta \in \Theta \subset \mathbb{R}^d\}$, where we assume the parameter θ is independent of n. A time point

$\tau_0 \in [n-1] := \{1, \dots, n-1\}$ is called a *changepoint* if $X_k \sim \mathbb{P}_{k,\theta} := \mathbb{P}_\theta(\cdot|X_1, \dots, X_{k-1})$ (conditional probability of X_k) for $k \in [\tau_0]$ and $X_k \sim \mathbb{P}_{k,\theta+\Delta}$ for $k \in \{\tau_0+1, \dots, n\}$, where $\theta, \theta+\Delta \in \Theta$ and $\Delta \neq 0$. We aim at determining whether there exists a changepoint in this sequence, which we formalize as hypothesis testing:

(P0) Testing the existence of a change within $[n-1]$,

$$H_0 : X_k \sim \mathbb{P}_{k,\theta}, k \in [n]$$
$$H_1 : \exists \tau \in [n-1] \text{ and } \Delta \neq 0, s.t.\ \ X_k \sim \mathbb{P}_{k,\theta}, k \in [\tau] \text{ and } X_k \sim \mathbb{P}_{k,\theta+\Delta}, k = \tau+1, \dots, n.$$

There are scenarios where the change may not happen simultaneously in all d components of the parameter, or only a subset of these components is of interest. In such cases, testing for simultaneous change can result in low sensitivity (or true positive rate); see *e.g.* [Enikeeva and Harchaoui, 2019]. Instead, we test for *sparse alternatives*, that is, only a subset of these components changes, and we call them *changed components*. For this purpose, we consider the following alternatives: the existence of one change point among p predefined components, then we generalize to the case of unknown p components, and finally, we adapt to the case when the number of changed components is unknown:

(P1) Testing the existence of a change in given p components, indexed by T (*i.e.,* $T \subset [d]$ and $|T| = p$), of the parameter within $[n-1]$:

$$H_1 : \exists \tau \in [n-1], \Delta \in \text{span}^\star(T), s.t.\ \ X_k \sim \mathbb{P}_{k,\theta}, k \in [\tau]; X_k \sim \mathbb{P}_{k,\theta+\Delta}, k = \tau+1, \dots, n,$$

where $\text{span}^\star(T) := \text{span}\{e_i : i \in T\} \backslash \{0\}$ and $\{e_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$.

(P2) Testing the existence of a change in unknown p components within $[n-1]$:

$$H_1 : \exists \tau \in [n-1], \Delta \in \bigcup_{T \in \mathcal{T}_p} \text{span}^\star(T), s.t.\ X_k \sim \mathbb{P}_{k,\theta}, k \in [\tau]; X_k \sim \mathbb{P}_{k,\theta+\Delta}, k = \tau+1, \dots, n,$$

where $\mathcal{T}_p$ is the collection of all subsets of size p of $[d]$.

(P3) Testing the existence of a change in an unknown number of components within $[n-1]$:

$$H_1 : \exists \tau \in [n-1] \text{ and } \Delta \in \bigcup_{p \in \mathcal{P}} \bigcup_{T \in \mathcal{T}_p} \text{span}^\star(T), s.t.\ \begin{cases} X_k \sim \mathbb{P}_{k,\theta}, & k = 1, \dots, \tau \\ X_k \sim \mathbb{P}_{k,\theta+\Delta}, & k = \tau+1, \dots, n, \end{cases}$$

where $\mathcal{P} \subset [d]$ and we refer to it as the *change cardinality set*.

Likelihood score and score-based testing. For simplicity, we assume from now on that $\mathbb{P}_\theta$ is absolutely continuous *w.r.t.* the Lebesgue measure with density p_θ. We define $\ell_{m:n}(\theta, \Delta) = \sum_{k=m}^n \log p_{\theta+\Delta \mathbb{1}\{k>\tau\}}(X_k|X_1, \dots, X_{k-1})$, to be the conditional log-likelihood of $X_m, \dots, X_n$ for every $m \in [n]$, and $\mathbb{1}\{\cdot\}$ is the indicator function. We write $\ell_{m:n}(\theta) = \ell_{m:n}(\theta, 0)$ for short. In particular, $\ell_n(\theta)$ is the log-likelihood of $(X_i)_{i=1}^n$ under H_0. When the true value $\theta = \theta_0$ is known and the change location τ is fixed, the testing Problem (P0) reduces to

$$H_0 : X_k \sim \mathbb{P}_{k,\theta_0}, k \in [n]$$
$$H_1 : X_k \sim \mathbb{P}_{k,\theta_0}, k \in [\tau] \text{ and } X_k \sim \mathbb{P}_{k,\theta_0+\Delta}, k = \tau+1, \dots, n, \Delta \neq 0.$$

The *score function w.r.t.* Δ is defined to be the gradient of the log-likelihood function, $S_{(\tau+1):n}(\theta) := \nabla_\Delta \ell_n(\theta) = \nabla_\theta \ell_{(\tau+1):n}(\theta)$. Under the null hypothesis (and under certain conditions detailed in

[Wellner, 2010]), the score is asymptotically normal with mean 0 and covariance $\mathcal{I}_0$ such that the observed Fisher information *w.r.t.* Δ, $\hat{\mathcal{I}}_{(\tau+1):n}(\theta)$, satisfies [1]

$$\frac{1}{n}\hat{\mathcal{I}}_{(\tau+1):n}(\theta_0) := -\frac{1}{n}\nabla^2_{\Delta}\ell_n(\theta_0) = -\frac{1}{n}\nabla^2_{\theta}\ell_{(\tau+1):n}(\theta_0) \to_p \mathcal{I}_0, \text{ as } n \to \infty \ .$$

Algorithm 1 Autograd-test

Input: data $(X_i)_{i\in[n]}$, log-likelihood function ℓ, MLE $\hat{\theta}_n$, thresholds α_l and α_s.
Compute H_p for each $p \in \mathcal{P}$.
Compute $(\nabla^2_\theta \ell_n(\hat{\theta}_n))^{-1}$.
for $\tau = 1$ **to** $n-1$ **do**
  Compute $\nabla_\theta \ell_{(\tau+1):n}(\hat{\theta}_n)$
  Compute $\nabla^2_\theta \ell_{(\tau+1):n}(\hat{\theta}_n)$.
  Compute $\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau)$ by (3).
  Compute $R_n(\tau)$ by (2).
  $v_s \leftarrow \mathrm{diag}(\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau))^{-1} S^2_{(\tau+1):n}(\hat{\theta}_n)$.
  for $p \in \mathcal{P}$ **do**
    Let T_p be the index set of the largest p components of v_s.
    Compute $R_n(\tau, T_p)$.
  end for
end for
Compute $\psi_{\mathrm{lin}}(\alpha_l)$ by (5).
Compute $\psi_{\mathrm{scan}}(\alpha_s)$ by (6).
Output: $\psi(\alpha) = \psi_{\mathrm{lin}}(\alpha_l) \vee \psi_{\mathrm{scan}}(\alpha_s)$.

Algorithm 2 Autograd-test-CuSum

Input: data stream $X_1, X_2, \ldots, X_n$, log-likelihood function ℓ, initial MLE $\hat{\theta}$, threshold α.
Sample from M to estimate $q_M(\alpha)$.
Initialization: $t \leftarrow m$, $S_{\mathrm{full}} \leftarrow S_{\mathrm{par}} \leftarrow S_{1:m}(\hat{\theta})$, $\hat{\mathcal{I}}_{\mathrm{full}} \leftarrow \hat{\mathcal{I}}_{\mathrm{par}} \leftarrow \hat{\mathcal{I}}_m(\hat{\theta})$, $R_{\min} \leftarrow S^\top_{\mathrm{full}}(\hat{\mathcal{I}}_{\mathrm{full}})^{-1} S_{\mathrm{full}}$.
while $t \le n$ and $R_{\mathrm{par}} \le N q_M(\alpha)/t$ **do**
  $t \leftarrow t+1$.
  $\hat{\theta} \leftarrow \hat{\theta} + \eta \nabla_\theta \ell_t(\hat{\theta})$.
  $S_j \leftarrow S_j + \nabla_\theta \ell_t(\hat{\theta})$ for $j \in \{\mathrm{full}, \mathrm{par}\}$.
  $\hat{\mathcal{I}}_j \leftarrow \hat{\mathcal{I}}_j + \nabla_\theta \ell_t(\hat{\theta}) \nabla_\theta \ell_t(\hat{\theta})^\top$ for $j \in \{\mathrm{full}, \mathrm{par}\}$.
  $R_{\mathrm{full}} \leftarrow S^\top_{\mathrm{full}} \left(\hat{\mathcal{I}}_{\mathrm{full}}\right)^{-1} S_{\mathrm{full}}$.
  if $R_{\mathrm{full}} \le R_{\min}$ **then**
    $S_{\mathrm{par}} \leftarrow 0$, $\hat{\mathcal{I}}_{\mathrm{par}} \leftarrow 0$, $R_{\mathrm{par}} \leftarrow 0$
    $R_{\min} \leftarrow R_{\mathrm{full}}$.
  else
    $R_{\mathrm{par}} \leftarrow S^\top_{\mathrm{par}} \left(\hat{\mathcal{I}}_{\mathrm{par}}\right)^{-1} S_{\mathrm{par}}$.
  end if
end while
Output: t.

The score statistic for this problem is given by

$$S^\top_{(\tau+1):n}(\theta_0) \left[\hat{\mathcal{I}}_{(\tau+1):n}(\theta_0)\right]^{-1} S_{(\tau+1):n}(\theta_0) \ , \tag{1}$$

which is asymptotically χ^2_d-distributed [2] and thus can be calibrated using quantiles of χ^2_d. When θ is unknown, we consider it as a nuisance parameter, the score statistic now becomes

$$R_n(\tau) := S^\top_{(\tau+1):n}(\hat{\theta}_n) \left[\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau)\right]^{-1} S_{(\tau+1):n}(\hat{\theta}_n) \ , \tag{2}$$

where, compared to (1), we replace θ_0 with $\hat{\theta}_n$, the maximum likelihood estimator (MLE) of θ under H_0, and replace $\hat{\mathcal{I}}_{(\tau+1):n}(\theta_0)$ by the Schur complement of the block $\nabla^2_\theta \ell_n(\hat{\theta}_n)$ of the observed Fisher information *w.r.t.* (θ, Δ) (see, for instance [Wellner, 2010] for details):

$$\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau) = -\nabla^2_\theta \ell_{(\tau+1):n}(\hat{\theta}_n) + \left(\nabla^2_\theta \ell_{(\tau+1):n}(\hat{\theta}_n)\right)^\top \left(\nabla^2_\theta \ell_n(\hat{\theta}_n)\right)^{-1} \nabla^2_\theta \ell_{(\tau+1):n}(\hat{\theta}_n) \ , \tag{3}$$

A natural statistic for testing Problem (P0) is the *linear statistic* in the terminology of [Enikeeva and Harchaoui, 2019],

$$R_n = \max_{\tau\in[n-1]} R_n(\tau) \ . \tag{4}$$

[1] Here $\to_p$ represents convergence in probability, and we will omit "as $n \to \infty$" if there is no confusion.
[2] The χ^2 distribution with d degrees of freedom.

Under sparse alternatives (P1), the changed components are fixed so that the score statistic becomes $\max_{\tau\in[n-1]} R_n(\tau, T) = \left[S^\top_{(\tau+1):n}(\hat{\theta}_n)\right]_T \left\{\left[\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau)\right]_{T,T}\right\}^{-1} \left[S_{(\tau+1):n}(\hat{\theta}_n)\right]_T$, where v_T is the subvector of $v \in \mathbb{R}^d$ indexed by T and $M_{T,T}$ is the sub-matrix of $M \in \mathbb{R}^{d\times d}$ indexed by (T, T). For Problem (P2), a natural choice is $\max_{\tau\in[n-1]} \max_{T\in\mathcal{T}_p} R_n(\tau, T)$ while to adapt to unknown p in (P3), we use the *scan statistic* $R_n(\mathcal{P}) = \max_{p\in\mathcal{P}} \max_{\tau\in[n-1]} \max_{T\in\mathcal{T}_p} \frac{1}{H_p} R_n(\tau, T)$, where H_p is a pre-determined threshold discussed later. The algorithm to compute the autograd-test is presented in Alg. 3.

Score statistics, differentiable programming, and component screening. An attractive feature of score statistics in the age of differentiable programming software libraries such as PyTorch [Paszke et al., 2017a] and TensorFlow [Abadi et al., 2015] is their straightforward computation using automatic differentiation. Another attractive feature of score statistics *w.r.t.* generalized likelihood ratio (GLR) statistics is their flexibility to screen individual components or groups of components. To be more specific, if only the parameters indexed by the set T^* are of interest, the parameters indexed by $[n]\backslash T^*$ can be seen as constants (so associated derivatives *w.r.t.* can be discarded) and one can proceed as before. Doing so, we neglect irrelevant information in the data and focus on catching specific signals, while for the GLR statistics, maximization under the new null ($\Delta_{T^*} = 0$) is required, imposing additional computational burden.

Online extension. Often, we might not have access to the full dataset (or it is too large to be stored in memory), so we also develop an online version of the autograd-test. For computational consideration, we assume from now on that the data are *i.i.d.*. Consequently, we use $\sum_{t=1}^n \nabla_\theta \ell_t(\hat{\theta}) \nabla_\theta \ell_t(\hat{\theta})^\top / n$ to estimate the expected Fisher information, since it does not require the calculation of second derivatives.

A naive modification to the score statistic in the online setting is $R_t = S^\top_{1:t}(\hat{\theta}_t) \left[\hat{\mathcal{I}}_t(\hat{\theta}_t)\right]^{-1} S_{1:t}(\hat{\theta}_t)$, with stopping time $t_a = \min\{t : R_t > h_t\}$ in which h_t is some fixed threshold. To control the false discovery rate (FDR) of this approach, we take the repeated significance tests [Ghosh and Sen, 1991] viewpoint. We first train the model using a sample of size m from the null distribution, and compute $S_{1:m}(\hat{\theta}_m)$ and $\hat{\mathcal{I}}_m(\hat{\theta}_m)$ as an initialization. Then for each new arrival observation, we update the score and information to compute new test statistic, and compare it with the threshold until rejection. Now the overall FDR, $\mathbb{P}_{\theta_0}[R_m > h_m(\alpha)] + \sum_{t=m+1}^n \mathbb{P}_{\theta_0}\left[R_j \le h_j(\alpha), \forall j \in [1, t-1]; R_t > h_t(\alpha)\right]$, can be controlled within α by approximating the threshold as $h_t(\alpha) = n q_M(\alpha)/t$, where n is the sample size of the whole dataset, M is defined by $M = \sup\{W(t)^\top W(t) : t \in [0, 1]\}$ and $W(t)$ here is the standard multi-dimensional Wiener process. To compute $q_M(\alpha)$, we follow the discretization methods presented in [Glasserman, 2013] to directly sample from the distribution of M, then estimate $q_M(\alpha)$ by the sample quantile. For the scan counterpart, we leave it for future work.

One drawback of this statistic is that $\hat{\theta}_t$ needs to be updated whenever a new observation arrives, and hence $S_{1:t}$ and $\hat{\mathcal{I}}_t$ must be evaluated at a different value. As a result, the update of R_t requires an operation of $\mathcal{O}(t)$ complexity, which leads to an algorithm of computational complexity at least $\mathcal{O}(t^2)$. Another challenge, as mentioned in [Apley and Chin, 2007], lies in the ineffectiveness of the cumulative score statistic when the change appears at some later time. Intuitively, for observations $(X_1, \ldots, X_t)$ coming from the null distribution, the score function $S_{1:t}(\hat{\theta}_t)$ is close to zero since $\mathbb{E}_0[S_1(\theta_0)] = 0$, while the observed Fisher information accumulates as $\hat{\mathcal{I}}_t(\hat{\theta}_t) \approx t\mathcal{I}_0$. If the number of observations after change is relatively small, then $t\mathcal{I}_0$ may have a dominating effect so that

the statistic will also be small. To address these problems, we propose the *autograd-test-CuSum* algorithm as illustrated in Alg. 4 (for more discussion see Appendix A).

3 Level and power

In this section we summarize asymptotic behavior of the proposed score-based tests under null and alternatives (for precise statements and proofs see the supplementary material). We then demonstrate the consistency of these tests in both level and power.

Given a significance level α, it is desirable that the tests are at least asymptotically of level α, *i.e.,* the type I error (or false discovery rate) is asymptotically upper bounded by α. To this end, we derive the asymptotic distribution under the null hypothesis in the following theorem.

Theorem 1 (Asymptotic normality under the null)**.** *Let $X_1, \ldots, X_n$ be a sequence of random variables with a family of joint probability density functions $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$, where the parameter θ is assumed to be independent of n. Suppose that the true parameter $\theta_0 \in \operatorname{int}(\Theta)$ (the interior of Θ), the log-likelihood is twice continuously differentiable, the observed information matrix converges in probability to $\mathcal{I}_0$, and $\hat{\theta}_n$, the maximum likelihood estimator (MLE), exists and weakly converges to a normal distribution. Then, for any $\tau_n \in \mathbb{Z}_+$ such that $\tau_n/n \to \lambda \in (0,1)$, we have*

$$\frac{n}{\tau_n(n-\tau_n)} \hat{\mathcal{I}}_n(\hat{\theta}_n; \tau) \to_p \mathcal{I}_0 \ .$$

If further the normalized score can be written as the sum of a martingale difference sequence up to an $o_p(1)$ term, and this sequence satisfies the Lindeberg conditions[3], then we also have:

$$\sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{(\tau_n+1):n}(\hat{\theta}_n) \to_d \mathcal{N}(0, \mathcal{I}_0) \ ,$$

In particular, if the martingale difference sequence is stationary and ergodic w.r.t. its natural filtration, the conclusion holds.

Most conditions in Th. 1 are standard for the MLE to be asymptotically normal. The conditions on the normalized score are required by the Lindeberg theorem for martingales [Van der Vaart, 2013] to prove the asymptotic normality of the score for dependent models. In fact, under suitable regularity conditions, they are satisfied in independent and identically distributed models, hidden Markov models [Bickel et al., 1998], and stationary autoregressive–moving-average models [Douc et al., 2014]. If the model of the data satisfies all the conditions in Th. 1, then $R_n(\tau) \to_d \chi^2_d$ under the null. Note that the linear statistic is the maximum of $R_n(\tau)$ over $\tau \in [n-1]$, so we use the Bonferroni correction to compensate for multiple comparisons, which gives the critical value $q_{\chi^2_d}(\alpha/n)$—the upper (α/n)%-quantile of χ^2_d. Consequently, the decision rule for the linear statistic is

$$\psi_{\text{lin}}(\alpha) := \mathbb{1}\left\{ R_n > q_{\chi^2_d}(\tfrac{\alpha}{n}) \right\} \ . \tag{5}$$

Similarly, since the asymptotic distribution of $R_n(\tau, T)$ is χ^2_p, the critical value for testing Problem (P1) is $q_{\chi^2_p}(\alpha/n)$. For Problem (P2), we maximize $R_n(\tau, T)$ over $T \in \mathcal{T}_p$, so a further Bonferroni

[3] See [Van der Vaart, 2013, Chapter 4.5] for details.

correction leads to the critical value $q_{\chi^2_p}\big(\alpha/n\,|\mathcal{T}_p|\big) = q_{\chi^2_p}\big(\alpha/n\binom{d}{p}\big)$. For Problem (P3), we derive the *scan statistic*, a statistic adaptive to an unknown number of changed components defined by

$$\psi_{\text{scan}}(\alpha) := \mathbb{1}\left\{ \max_{p\in\mathcal{P}} \frac{\max_{\tau\in[n-1]} \max_{T\in\mathcal{T}_p} R_n(\tau, T)}{q_{\chi^2_p}\big(\alpha/\big(\binom{d}{p} n(p+1)^2\big)\big)} > 1 \right\} \ , \tag{6}$$

where $(p+1)^2$ is required to guarantee an asymptotic α level [4]; see supplementary material for technical details. Other corrections are possible, but the former provides small thresholds when the signal of change is weak (p is small).

To incorporate the strengths of these two tests, we propose *autograd-test*, a combination of them:

$$\psi(\alpha) := \max\{\psi_{\text{lin}}(\alpha_l), \psi_{\text{scan}}(\alpha_s)\} \ , \tag{7}$$

where $\alpha_l + \alpha_s = \alpha$. The choice of α_l and α_s should be based on prior knowledge regarding how likely the change is sparse (*i.e.,* they should be equal without any prior information).

The next proposition verifies the consistency in power of the score test for changepoint detection under fixed alternatives by assuming the independence of the data.

Proposition 1 (Fixed alternatives)**.** *Let $X_1, \ldots, X_n$ be independent observations and $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be a family of marginal density functions such that there exists $\tau_n \in [n-1]$ with $X_1, \ldots, X_{\tau_n} \sim p_{\theta_0}$, $X_{\tau_n+1}, \ldots, X_n \sim p_{\theta_1} (\theta_1 \neq \theta_0)$, and $\tau_n/n \to \lambda \in (0,1)$. Assume that[5] $\lambda D_{\text{KL}}(p_{\theta_0} \| p_\theta) + \bar{\lambda} D_{\text{KL}}(p_{\theta_1} \| p_\theta)$ (with $\bar{\lambda} := 1 - \lambda$) has a unique minimizer $\theta^* \in \text{int}(\Theta)$, and $\mathbb{E}_{\theta_l}[-\nabla^2_\theta \ell(\theta^*)]$ is positive definite for $l \in \{0, 1\}$. Then there exists a sequence of MLE, provided smoothness and regularity conditions on the log-likelihood, such that $\hat{\theta}_n \to_p \theta^*$ and*

$$\mathbb{P}\big(R_n(\tau_n) > q_{\chi^2_d}(\alpha)\big) \to 1 \ .$$

The conditions in Prop. 1 are counterparts of standard assumptions for analyzing the asymptotic behavior of classic score statistic under fixed alternatives. The condition regarding the KL divergences is needed for the MLE to have a valid limit under fixed alternatives. In addition, **??** implies $H_p(\alpha)/n \to 0$, so it follows immediately that the score-based tests have power converging to one.

In hypothesis testing, it is also of great interest to show the asymptotic behavior of tests under local alternatives, that is, the signal of change weaken as the sample size increases.

Proposition 2 (Local alternatives)**.** *Let $X_1, \ldots, X_n$ be independent observations and $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be a family of marginal density functions such that there exists $\tau_n \in [n-1]$ with $X_1, \ldots, X_{\tau_n} \sim p_{\theta_0}$, $X_{\tau_n+1}, \ldots, X_n \sim p_{\theta_n}$, where $\theta_n = \theta_0 + hn^{-1/2}$ with $h \neq 0$, and $\tau_n/n \to \lambda \in (0,1)$. Assume that θ_0 is the unique maximizer of $\mathbb{E}_{\theta_0}[\ell(\theta)]$ and $\mathbb{E}_{\theta_0}[-\nabla^2_\theta \ell(\theta_0)] = \mathcal{I}_0$ is positive definite. Then there exists a sequence of MLE $\hat{\theta}_n$, provided smoothness and regularity conditions on the log-likelihood, such that*

$$\frac{n}{\tau_n(n-\tau_n)} \hat{\mathcal{I}}_n(\hat{\theta}_n; \tau) \to_p \mathcal{I}_0 \quad , \quad \sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}_d\left(\bar{\lambda} h, \mathcal{I}_0^{-1}\right), \tag{8}$$

$$\textit{and} \quad \sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{(\tau_n+1):n}(\hat{\theta}_n) \to_d \mathcal{N}_d(\sqrt{\lambda\bar{\lambda}}\, \mathcal{I}_0 h, \mathcal{I}_0) \ . \tag{9}$$

Hence, $R_n(\tau_n)$ weakly converges to a non-central chi-square distribution with degrees of freedom d and parameter $\lambda\bar{\lambda} h^\top \mathcal{I}_0 h$.

[4] We only need $\sum_{p\in\mathcal{P}} 1/(p+1)^2 < 1$ for controlling the level of the test.

[5] We use the notation D_{KL} for the Kullback-Leibler divergence here.

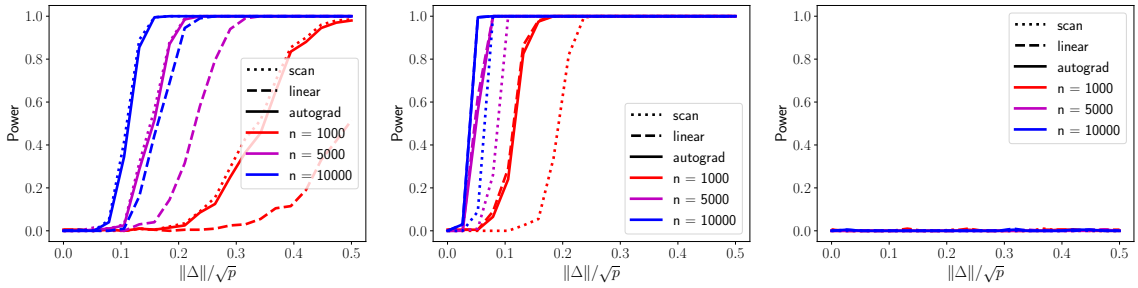

Figure 1: Power versus magnitude of change for linear regression (left: $p = 1$; middle: $p = 20$; right: $p = 1$ with restriction excluding the changed component).

4 Experiments

In this section, we perform simulations to evaluate empirical behavior of our methods on synthetic data generated from a linear regression model, a time series model, a hidden Markov model (HMM), and a text topic model described in [Stratos et al., 2015], which is essentially an HMM with structured discrete emission distribution. We apply our approach to detect changes in this topic model on subtitles of TV shows. Due to the shortage of space, we summarize experimental settings and summarize our findings here. More details and additional results (including the ones for HMM) can be found in the supplementary material.

Experimental settings. For each sample size n, we generate the first half of the sample from a model with parameter θ_0. Then we obtain θ_1 by adding δ to the first p components of θ_0 (so that $\delta = \|\Delta\| / \sqrt{p}$ where $\Delta = \theta_1 - \theta_0$ quantifies the magnitude of change). Next, we generate the second half from the same model with parameter θ_1 and run the linear test, the scan test, and the autograd-test on this dataset, where the significance levels are set to be $\alpha = 2\alpha_l = 2\alpha_s = 0.05$. We repeat this procedure 200 times and approximate the power of these tests by rejection frequency. Finally, we plot the power curve by varying values of δ, where we use three different types of lines to represent three tests, and different colors to indicate different sample sizes.

Linear regression model. We consider a linear regression model with 100 slope coefficients and intercept (*i.e.,* 101 parameters) and investigate two sparsity levels, $p = 1$ and $p = 20$. As shown in the left and middle plots of Fig. 1, when the change is sparse, the scan test and the autograd-test share similar power curves and both outperform the linear test significantly. When the change is less sparse, all tests' performance improve to a large extent since the change signal becomes stronger, with the scan test tending to perform poorer than the other two. This empirically illustrates that 1) the scan test works better in detecting sparse changes, 2) the linear test is more powerful for non-sparse changes and 3) the autograd-test achieves comparable performance in both situations.

Additionally, we consider the same linear regression model except that the detection is limited to 50 components of parameters (*i.e.,* regard the rest 51 components as constants). When the restricted components contain the changed one, all tests have improved performance; when the abnormal component is outside the scope of the detection, the power is below 0.01 no matter how strong the signal is (see the right plot in Fig. 1). Hence, these tests can filtrate irrelevant information by simply confining the detection to parameters of interest.

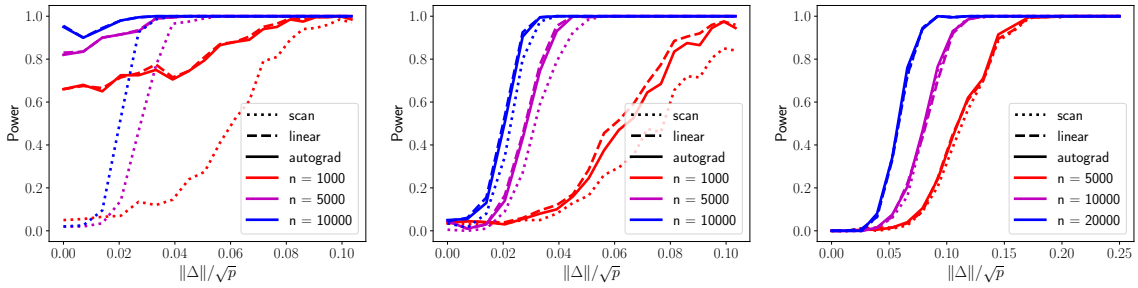

Figure 2: Power versus magnitude of change for dependent models (left: ARMA$(3, 2)$ without restriction; middle: ARMA$(3, 2)$ with restriction; right: text topic model with 3 hidden states).

Time series model. We then investigate an autoregressive–moving-average (ARMA) model. As shown in the left plot of Fig. 2, the scan test works fairly well for this model. However, the other two have extremely high false discovery rate (FDR). This problem gets more severe as n increases, and hence is not due to lack of accuracy of the MLE. It turns out that it is caused by the non-homogeneity of model parameters—the derivatives *w.r.t.* AR coefficients tend to be of different magnitude compared to the ones *w.r.t.* MA coefficients. This results in ill-conditioned Schur complement (3) and subsequent unstable computation of the linear statistic. On the contrary, the scan statistic only inverts submatrices of sizes approximately $\sqrt{d} \times \sqrt{d}$ with reduced condition number (the parameters selected by the scan statistic are all AR coefficients in our experiments). We remark that in such situations we can select a small (or even zero) significance level for the linear part of the autograd-test (so the scan part has a dominating effect) to obtain reasonable results. If we restrict the detection of these tests in the AR coefficients, all tests are now consistent in level (see the middle plot of Fig. 2).

Text topic model. The right figure of Fig. 2 contains results for a topic model with $N = 3$ hidden states. All three tests show consistent behavior and the scan test are less competitive than the rest two because of low sparsity. As N increases, the FDRs start to be out of control, but, different from the case of ARMA models, this problem is alleviated as n increases.

We now consider a real data application. We collect subtitles of the first two seasons of four different TV shows—Friends, Modern Family, the Sopranos, and Deadwood—where the former two are assumed to be "polite" and the latter two are "rude". The task is to detect changes in rudeness level. For every pair of different seasons, we train the aforementioned topic model on them, and then apply (only) the scan test for detection (for FDR consideration discussed above).

As demonstrated in Table 1, the scan test does a perfect job in reporting shifts in rudeness level. However, it tends to be liberal in this task, that is, the false discovery rate is relatively high. For ("polite", "polite") pairs, there are two false positives and one NA because the MLE of the transition matrix contains zero which makes the statistic undefined; while for ("rude", "rude") pairs, 9 out of 16 are false positives. Additionally, to eliminate sequential effect of episodes and obtain more robust results, we randomly shuffle the episodes (as a whole) in each season, then detect changes using these new data. Results suggest a similar phenomenon.

We remark that rudeness is definitely not the only factor that contributes to the difference between two shows, and there is no reason to believe it is the only factor that the scan test utilizes to detect changes (it might not be one). But the results are promising in the sense that the scan

Table 1: Rejection status for pair-wise experiments.

	F1	F2	M1	M2	S1	S2	D1	D2
F1	NR	NR	NR	NR	R	R	R	R
F2	NR	NR	R	NR	R	R	R	R
M1	NR	R	NR	NR	R	R	R	R
M2	NR	NR	NR	NA	R	R	R	R
S1	R	R	R	R	NR	NR	R	R
S2	R	R	R	R	NR	NR	R	R
D1	R	R	R	R	R	R	NR	R
D2	R	R	R	R	R	R	NR	NR

statistic can neglect some low level discrepancies and focus on "global information" in language level. As we already discussed, we can confine the detection to some specific parameters, if it is possible to determine which ones are related to the rudeness, and obtain a more appropriate test for this specific task.

Online extension. For the autograd-test-CuSum algorithm, we use a linear regression model with 11 parameters. We first train and initialize the model with 1000 observations. Then we run the algorithm on samples of different sizes with a change at time 1000. Besides the power, its performance is also assessed by the delay of detections, measuring how fast the algorithm can detect the anomaly when a change happens. With relatively small n, the autograd-test-CuSum is consistent in level and achieves power 1. As n increases, the power presents an increment first then a downward trend, and we speculate this is due to the approximation error of online updating; while the FDR also rises then oscillates around 0.2, which calls for a correction to the threshold, accounting for the reinitialization employed in autograd-test-CuSum. For the delay of detections, data with larger n tend to bring about longer delay and sharper decline (with the change magnitude), since the growth of change magnitude has more impact on data with bigger post-change sample. The reader is referred to the supplementary material for more experiments.

Conclusion We introduced a versatile change monitoring method called *autograd-test* and its online variant *autograd-test-CuSum*. The experimental results showed that the calibration of these test statistics based on our theoretical arguments brings about change detection algorithms that are able to capture subtle changes in parameters for various machine learning models, ranging from time series models to text topic models, in a wide range of statistical regimes. The extension of this approach to machine learning models trained with implicit regularization techniques would be an interesting venue to explore in future work.

Acknowledgements

This work was supported by NSF CCF-1740551, NSF DMS-1810975, the program "Learning in Machines and Brains" of CIFAR, and faculty research awards.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

Daniel W Apley and Chang-Ho Chin. An optimal filter design approach to statistical process control. *Journal of Quality Technology*, 39(2):93–117, 2007.

Michèle Basseville and Igor V Nikiforov. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

Peter J Bickel, Ya'acov Ritov, and Tobias Ryden. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):1614–1635, 1998.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

George Box and José Ramírez. Cumulative score charts. *Quality and Reliability Engineering International*, 8(1):17–27, 1992.

O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer-Verlag New York, 1st edition, 2005.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Jean Deshayes and Dominique Picard. Off-line statistical analysis of change-point models using non parametric and likelihood methods. In Albert Basseville, Michèleand Benveniste, editor, *Detection of Abrupt Changes in Signals and Dynamical Systems*, pages 103–168, Berlin, Heidelberg, 1986. Springer Berlin Heidelberg. ISBN 978-3-540-39726-7.

Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear time series: Theory, methods and applications with R examples*. Chapman and Hall/CRC, 2014.

Farida Enikeeva and Zaid Harchaoui. High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, 47(4):2051–2079, 2019.

Bhaskar Kumar Ghosh and Pranab Kumar Sen. *Handbook of sequential analysis*. CRC Press, 1991.

Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer Science & Business Media, 2013.

David V Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17, 1970.

Lajos Horváth and Emanuel Parzen. Limit theorems for Fisher-score change processes. *Lecture Notes-Monograph Series*, pages 157–169, 1994.

Will Knight. A self-driving Uber has killed a pedestrian in Arizona. *Ethical Tech*, March 2018.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

Thomas A Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982.

Alberto Luceño. Average run lengths and run length probability distributions for cuscore charts to control normal mean. *Computational Statistics & Data Analysis*, 32(2):177–195, 1999.

Rachel Metz. Microsoft's neo-Nazi sexbot was a great lesson for makers of AI assistants. *Artificial Intelligence*, March 2018.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Ewan S Page. Estimating the point of change in a continuous process. *Biometrika*, 44(2):248–252, 1957.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017a.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017b.

Walter Andrew Shewhart. *Economic control of quality of manufactured product*. ASQ Quality Press, 1931.

Karl Stratos, Michael Collins, and Daniel Hsu. Model-based word embeddings from decompositions of count matrices. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1282–1291, 2015.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Aad W Van der Vaart. Time series. Universiteit Leiden, 2013. URL `http://pub.math.leidenuniv.nl/~vaartawvander/timeseries/dictaat.pdf`.

Jon A Wellner. Efficient likelihood estimation and related tests. University of Washington, 2010. URL `https://www.stat.washington.edu/jaw/COURSES/580s/581/LECTNOTES/ch4-rev1.pdf`.

A Algorithmic implementation

Algorithm 3 Autograd-test

Input: data $(X_i)_{i\in[n]}$, log-likelihood function ℓ, MLE $\hat{\theta}_n$, thresholds α_l and α_s.
Compute H_p for each $p \in \mathcal{P}$.
Compute $(\nabla^2_\theta \ell_n(\hat{\theta}_n))^{-1}$.
for $\tau = 1$ **to** $n-1$ **do**
    Compute $\nabla_\theta \ell_{(\tau+1):n}(\hat{\theta}_n)$
    Compute $\nabla^2_\theta \ell_{(\tau+1):n}(\hat{\theta}_n)$.
    Compute $\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau)$ by (3).
    Compute $R_n(\tau)$ by (2).
    $v_s \leftarrow \mathrm{diag}(\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau))^{-1} S^2_{(\tau+1):n}(\hat{\theta}_n)$.
    for $p \in \mathcal{P}$ **do**
        Let T_p be the index set of the largest p components of v_s.
        Compute $R_n(\tau, T_p)$.
    end for
end for
Compute $\psi_{\mathrm{lin}}(\alpha_l)$ by (5).
Compute $\psi_{\mathrm{scan}}(\alpha_s)$ by (6).
Output: $\psi(\alpha) = \psi_{\mathrm{lin}}(\alpha_l) \vee \psi_{\mathrm{scan}}(\alpha_s)$.

Algorithm 4 Autograd-test-CuSum

Input: data stream $X_1, X_2, \ldots, X_n$, log-likelihood function ℓ, initial MLE $\hat{\theta}$, threshold α.
Sample from M to estimate $q_M(\alpha)$.
Initialization: $t \leftarrow m$, $S_{\mathrm{full}} \leftarrow S_{\mathrm{par}} \leftarrow S_{1:m}(\hat{\theta})$, $\hat{\mathcal{I}}_{\mathrm{full}} \leftarrow \hat{\mathcal{I}}_{\mathrm{par}} \leftarrow \hat{\mathcal{I}}_m(\hat{\theta})$, $R_{\mathrm{min}} \leftarrow S^\top_{\mathrm{full}} (\hat{\mathcal{I}}_{\mathrm{full}})^{-1} S_{\mathrm{full}}$.
while $t \le n$ and $R_{\mathrm{par}} \le N q_M(\alpha)/t$ **do**
    $t \leftarrow t+1$.
    $\hat{\theta} \leftarrow \hat{\theta} + \eta \nabla_\theta \ell_t(\hat{\theta})$.
    $S_j \leftarrow S_j + \nabla_\theta \ell_t(\hat{\theta})$ for $j \in \{\mathrm{full}, \mathrm{par}\}$.
    $\hat{\mathcal{I}}_j \leftarrow \hat{\mathcal{I}}_j + \nabla_\theta \ell_t(\hat{\theta}) \nabla_\theta \ell_t(\hat{\theta})^\top$ for $j \in \{\mathrm{full}, \mathrm{par}\}$.
    $R_{\mathrm{full}} \leftarrow S^\top_{\mathrm{full}} \left(\hat{\mathcal{I}}_{\mathrm{full}}\right)^{-1} S_{\mathrm{full}}$.
    if $R_{\mathrm{full}} \le R_{\mathrm{min}}$ **then**
        $S_{\mathrm{par}} \leftarrow 0$, $\hat{\mathcal{I}}_{\mathrm{par}} \leftarrow 0$, $R_{\mathrm{par}} \leftarrow 0$
        $R_{\mathrm{min}} \leftarrow R_{\mathrm{full}}$.
    else
        $R_{\mathrm{par}} \leftarrow S^\top_{\mathrm{par}} \left(\hat{\mathcal{I}}_{\mathrm{par}}\right)^{-1} S_{\mathrm{par}}$.
    end if
end while
Output: t.

Algorithmic details for autograd-test are summarized in Alg. 3. We first compute the threshold H_p for all $p \in \mathcal{P}$, and invert the Hessian of the log-likelihood ℓ_n at $\hat{\theta}_n$. Then, for all $\tau \in [n-1]$, we get the derivative of conditional log-likelihood $\nabla_\theta \ell_{(\tau+1):n}(\hat{\theta}_n)$, the Hessian of the conditional log-likelihood $\nabla^2_\theta \ell_{(\tau+1):n}(\hat{\theta}_n)$ and the Schur complement $\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau)$ thanks to (3). Next, the score statistic $R_n(\tau)$ is evaluated with Eq. (2). For $p \in \mathcal{P}$, maximizing $R_n(\tau, T)$ over all possible $T \in \mathcal{T}_p$ can be costly. So, we greedily approximate the maximizer by selecting the largest components of $R_n(\tau, T)$ for $|T| = 1$,*i.e.,* we choose T_p as the indices of the largest p components in $\mathrm{diag}(\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau))^{-1} S^2_{(\tau+1):n}(\hat{\theta}_n)$ (we write S^2 to represents the element-wise square), and substitute $R_n(\tau, T_p)$ to $R_n(\tau, \mathcal{T}_p)$.

Alg. 3 relies on the *Autograd* package in *PyTorch* [Paszke et al., 2017b][6]. Hence, we have designed a universal testing framework for changepoint detection, applying to any probabilistic model whose likelihood can be handled by *PyTorch*.

To analyze the computational complexity of our algorithm, we assume that the cost of evaluating $\nabla_\theta \ell_n$ is linear in the sample size n, that is, $\mathcal{O}(n)$. This implies computing the Hessian matrix $\nabla^2_\theta \ell_n$ is of complexity $\mathcal{O}(nd)$. Calculating the inverse of the Hessian costs at most $\mathcal{O}(d^3)$, so that the overall cost for step 3 is $\mathcal{O}(nd + d^3)$. The computation of thresholds have time complexity $\mathcal{O}(d\,|\mathcal{P}|)$, which is dominated by the former operation since $|\mathcal{P}| \le d$.

Next, let us analyze the computational cost for each $\tau \in [n-1]$. Steps 5-8 have complexity $\mathcal{O}((n-\tau)d)$, $\mathcal{O}(d^2)$, $\mathcal{O}(d^3)$, and $\mathcal{O}(d \log d)$, respectively. The loop over $p \in \mathcal{P}$ costs at most $\mathcal{O}(|\mathcal{P}|^4)$ because evaluating $R_n(\tau, T)$ has complexity $\mathcal{O}(|T|^3)$. To summarize, the overall computational complexity of Alg. 3 is $\mathcal{O}(n^2 d + n d^3 + n\,|\mathcal{P}|^4)$.

In practice, the change cardinality set $\mathcal{P}$ is usually set to be $\{1, \ldots, \lceil \sqrt{d} \rceil\}$, and the the overall cost becomes $\mathcal{O}(n^2 d + n d^3)$. If the observations are independent and identically distributed, or the log-likelihood function admits a simple recursive computation, we can reduce the complexity to $\mathcal{O}(nd + nd^3)$ which is linear in n. However, there are some scenarios, especially for latent variable

[6] *PyTorch* calculates the gradient of any combination of standard functions

models, where even evaluating the likelihood function is too expensive. In such cases our algorithm would also be intractable, unless there is an efficient way to compute the gradients. For instance, to make the calculation of the derivative and Hessian for hidden Markov models feasible, we invoke the Fisher's identity [Dempster et al., 1977] and Louis' identity [Louis, 1982] to interchange derivative and expectation, and then implement the fixed-interval smoothing [Cappé et al., 2005] to obtain the derivative and Hessian recursively.

Online extension. As illustrated in Alg. 4, at each time t, a new data point X_t arrives, and the MLE is updated by a first-order optimization method in the direction $\nabla_\theta \ell_t(\hat{\theta}_{t-1})$. Next, the new score function S is obtained by adding the score of the t-th observation at the updated $\hat{\theta}_t$ to the previous score. The update rule for the observed Fisher information $\hat{\mathcal{I}}$ is similar, leading to a constant time (*w.r.t.* the sample size) procedure.

With regard to the second drawback (the ineffectiveness of the cumulative score statistic when the change appears at some later time), we use a correction technique as in CUSUM [Page, 1954], *i.e.,*

$$R_t^\star = R_t - \min_{j\in[t]} R_j.$$

In Alg. 4, $R_{\min}$ is used to store the minimum of all *full statistics* R_{full} computed so far. If at time t the current R_{full} is not larger than $R_{\min}$, we think there is no changepoint before time t and reinitialize the *partial score* S_{par} and the *partial information* $\hat{\mathcal{I}}_{\text{par}}$. Importantly, the stopping criteria is based on the *partial statistic* R_{par} rather than R_{full}. Now at time $t \le \tau$, $S_{\text{full}} \approx t\mathbb{E}[S_0] = 0$, $\hat{\mathcal{I}}_{\text{full}} \approx t\mathcal{I}_0$, and $R_{\text{full}} \approx 0$; at time $t > \tau$, $S_{\text{full}} \approx (t-\tau)S_\tau$, $\hat{\mathcal{I}}_{\text{full}} \approx \tau\mathcal{I}_0 + (t-\tau)\mathcal{I}_\tau \approx t\mathcal{I}_*$, and $R_{\text{full}} \approx (t-\tau)^2 S_\tau^\top (\mathcal{I}_*)^{-1} S_\tau / t > 0$. Consequently, the reinitialization is likely to happen before τ, which prevents the accumulation of Fisher information when there is no change.

B Proofs

Theorem 1. *Let $X_1, \ldots, X_n$ be a sequence of random variables with a family of joint probability density functions $\{p_\theta(X_1, \ldots, X_n) : \theta \in \Theta \subset \mathbb{R}^d\}$, where the parameter θ is assumed to be independent of n. Assume that the true parameter $\theta_0 \in \text{int}(\Theta)$ (the interior of Θ), and that the following conditions hold:*

C1 : $\ell_n(\theta) := \log p_\theta(X_1, \ldots, X_n)$ *is twice continuously differentiable within* Θ.

C2 : $-\nabla_\theta^2 \ell_n(\theta_0)/n \to_p \mathcal{I}_0$ *where* $\mathcal{I}_0 \in \mathbb{R}^{d\times d}$ *is a positive definite matrix.*

C3 : The MLE $\hat{\theta}_n$ *exists and* $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}(0, \mathcal{I}_0^{-1})$ *(convergence in distribution).*

Then for any $\tau_n \in \mathbb{Z}_+$ *such that* $\tau_n/n \to \lambda \in (0,1)$, *we have* $\dfrac{n}{\tau_n(n-\tau_n)} \hat{\mathcal{I}}_n(\hat{\theta}_n; \tau) \to_p \mathcal{I}_0$.

If further the following conditions hold

C4 : The normalized score can be written as the sum of a martingale difference sequence, up to an $o_p(1)$ *term, w.r.t. to some filtration* $\{\mathcal{F}_t\}_{t\in\mathbb{Z}}$, *that is,*

$$Z_n(\theta_0) := \frac{1}{\sqrt{n}} S_{1:n}(\theta_0) = \frac{1}{\sqrt{n}} \nabla_\theta \ell_{1:n}(\theta_0) = \sum_{k=1}^{n} \frac{M_k}{\sqrt{n}} + o_p(1) \ ,$$

where $\mathbb{E}[M_k | \mathcal{F}_{k-1}] = 0, \forall k \in [n]$.

In addition, this martingale sequence satisfies the Lindeberg conditions:

C4-(a) : $\sum_{k=1}^{n} \mathbb{E}[M_k M_k^\top | \mathcal{F}_{k-1}]/n \to_p \mathcal{I}_0$ *and*
C4-(b) : $\forall \varepsilon > 0$ *and* $\alpha \in \mathbb{R}^d$, $\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[(\alpha^\top M_k)^2 \mathbb{1}\{|\alpha^\top M_k| > \sqrt{n}\varepsilon\} | \mathcal{F}_{k-1}\right] \to_p 0$.

Then, we can also obtain asymptotic normality of the score statistics:

$$\sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{(\tau_n+1):n}(\hat{\theta}_n) \to_d \mathcal{N}(0, \mathcal{I}_0) \ ,$$

In particular, if $\{M_k\}_{k\in\mathbb{Z}}$ *is a stationary and ergodic martingale difference series w.r.t. its natural filtration, the conclusion holds.*

Proof Condition C3 implies $\hat{\theta}_n \to_p \theta_0$, then by Conditions C1 and C2 we can derive

$$-\frac{1}{n-\tau_n}\nabla_\theta^2 \ell_{(\tau_n+1):n}(\hat{\theta}_n) = -\frac{1}{n-\tau_n}[\nabla_\theta^2 \ell_{1:n}(\hat{\theta}_n) - \nabla_\theta^2 \ell_{1:\tau_n}(\hat{\theta}_n)] \to_p \frac{1}{1-\lambda}\mathcal{I}_0 - \frac{\lambda}{1-\lambda}\mathcal{I}_0 = \mathcal{I}_0 \ .$$

Thus,

$$\frac{n}{\tau_n(n-\tau_n)}\hat{\mathcal{I}}_n(\hat{\theta}_n; \tau) \to_p \frac{1}{\lambda}\mathcal{I}_0 - (\frac{1}{\lambda} - 1)\mathcal{I}_0 = \mathcal{I}_0 \ .$$

Furthermore, according to Condition C3 and Taylor expansion,

$$Z_n(\theta_0) = Z_n(\hat{\theta}_n) - \nabla_\theta Z_n(\theta_n^\star)^\top(\hat{\theta}_n - \theta_0) = -\frac{1}{\sqrt{n}}\nabla_\theta Z_n(\theta_n^\star)^\top \sqrt{n}(\hat{\theta}_n - \theta_0) \ ,$$

where $\theta_n^\star$ is between θ_0 and $\hat{\theta}_n$ so that $\theta_n^\star \to_p \theta_0$, then we also have $-\nabla_\theta Z_n(\theta_n^\star)/\sqrt{n} \to_p \mathcal{I}_0$. Note that $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$, we can obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1) \ .$$

Moreover, by Lindeberg theorem for martingales [Van der Vaart, 2013] and Cramér-Wold device [Billingsley, 2008], Condition C4 implies $Z_n(\theta_0) \to_d \mathcal{N}(0, \mathcal{I}_0)$, and thus $Z_n(\theta_0) = O_p(1)$ as $n \to \infty$. It follows that

$$\begin{aligned}
\frac{S_{(\tau_n+1):n}(\hat{\theta}_n)}{\sqrt{n-\tau_n}} &= \frac{S_{(\tau_n+1):n}(\theta_0)}{\sqrt{n-\tau_n}} + \frac{1}{\sqrt{n-\tau_n}}\nabla_\theta S_{(\tau_n+1):n}^\top(\theta_n^\star)(\hat{\theta}_n - \theta_0) \\
&= \frac{S_{(\tau_n+1):n}(\theta_0)}{\sqrt{n-\tau_n}} + \frac{(\nabla_\theta S_n(\theta_n^\star) - \nabla_\theta S_{\tau_n}(\theta_n^\star))^\top}{\sqrt{n(n-\tau_n)}}\sqrt{n}(\hat{\theta}_n - \theta_0) \\
&= \frac{S_{(\tau_n+1):n}(\theta_0)}{\sqrt{n-\tau_n}} + \left(\frac{\lambda}{\sqrt{1-\lambda}} - \sqrt{\frac{1}{1-\lambda}}\right)\mathcal{I}_0\mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1) \\
&= -\sqrt{\frac{\tau_n}{n-\tau_n}} Z_{\tau_n}(\theta_0) + \sqrt{\frac{n}{n-\tau_n}} Z_n(\theta_0) + \frac{\lambda - 1}{\sqrt{1-\lambda}} Z_n(\theta_0) + o_p(1) \\
&= -\frac{\sqrt{\lambda}}{\sqrt{1-\lambda}} Z_{\tau_n}(\theta_0) + \frac{\lambda}{\sqrt{1-\lambda}} Z_n(\theta_0) + o_p(1) \ ,
\end{aligned}$$

Now by applying Lemma 2, we have

$$\sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{(\tau_n+1):n}(\hat{\theta}_n) \to_d \mathcal{N}\left(0, \frac{1}{\lambda^2}\frac{\lambda^2}{1-\lambda}(1 - 2\lambda + \lambda)\mathcal{I}_0\right) =_d \mathcal{N}(0, \mathcal{I}_0) \ .$$

In particular, if the sequence $\{M_k\}_{k\in\mathbb{Z}}$ is stationary and ergodic, then by stationarity there exists a fixed measurable function $f : \mathbb{R}^\infty \to \mathbb{R}^\infty$ such that $\forall k \in \mathbb{Z}$

$$\mathbb{E}[M_k M_k^\top | M_{k-1}, M_{k-2}, \dots] = f(M_{k-1}, M_{k-2}, \dots) \ .$$

almost surely. Due to the ergodicity of M_k, the series $N_k = f(M_{k-1}, M_{k-2}, \dots)$ is also ergodic so that $\bar{N}_n \to_{a.s.} \mathbb{E}[N_1]$. Similarly, given $c > 0$,

$$G_n(c) := \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\alpha^\top M_k^2 | \mathbb{1}\left\{\left|\alpha^\top M_k\right| > c\right\} | \mathcal{F}_{k-1}\right] \ ,$$

converges almost surely to their expectation for any $\alpha \in \mathbb{R}^d$. This expectation can be arbitrarily small by setting c to be large. For every $\epsilon > 0$, the sequence

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\alpha^\top M_k^2 | \mathbb{1}\left\{\left|\alpha^\top M_k\right| > \epsilon\sqrt{n}\right\} | \mathcal{F}_{k-1}\right] \le G_n(c) \ ,$$

for sufficiently large n given an arbitrary c, and thus converges almost surely to zero. ■

Remark. For *i.i.d.* models with finite second moment, the normalized score reads $Z_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla_\theta \ell_k(\theta_0)$ and under regularity conditions, $\mathbb{E}[\nabla_\theta \ell_k(\theta_0)] = 0$, so $(\nabla_\theta \ell_k(\theta_0))_{k\in[n]}$ is a martingale difference sequence. Moreover,

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[\nabla_\theta \ell_k(\theta_0)(\nabla_\theta \ell_k(\theta_0))^\top] = \mathbb{E}[\nabla_\theta \ell_1(\theta_0)(\nabla_\theta \ell_1(\theta_0))^\top] = \mathcal{I}_0 \ ,$$

and for any $\varepsilon > 0$ and any $\alpha \in \mathbb{R}^d$,

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\alpha^\top \nabla_\theta \ell_k(\theta_0) \mathbb{1}\Big(\Big|\alpha^\top \nabla_\theta \ell_k(\theta_0)\Big| > \sqrt{n}\varepsilon\Big)\right] \\
&= \mathbb{E}\left[\alpha^\top \nabla_\theta \ell_1(\theta_0) \mathbb{1}\Big(\Big|\alpha^\top \nabla_\theta \ell_1(\theta_0)\Big| > \sqrt{n}\varepsilon\Big)\right] \to 0 \ ,
\end{aligned}$$

since $\alpha^\top \nabla_\theta \ell_1(\theta_0) = O_p(1)$. Therefore, Condition C4 holds.

Lemma 1. *Under the assumptions in Th. 1, the scan test is asymptotically of level α.*

Proof With $\mathbb{E}_0$ and $\mathbb{P}_0$ the expectation and probability distribution under the null hypothesis

$$\begin{aligned}
\mathbb{E}_0[\psi_{\text{scan}}(\alpha)] &\le \sum_{p\in\mathcal{P}} \mathbb{P}_0\left(\frac{\max_{\tau\in[n-1]} \max_{T\in\mathcal{T}_p} R_n(\tau, T)}{q_{\chi^2_p}\big(\alpha/\big(\binom{d}{p} n(p+1)^2\big)\big)} > 1\right) \\
&\le \sum_{p\in\mathcal{P}} \sum_{\tau=1}^{n-1} \sum_{T\in\mathcal{T}_p} \mathbb{P}_0\left(\frac{R_n(\tau, T)}{q_{\chi^2_p}\big(\alpha/\big(\binom{d}{p} n(p+1)^2\big)\big)} > 1\right) \\
&= \sum_{p\in\mathcal{P}} \sum_{\tau=1}^{n-1} \sum_{T\in\mathcal{T}_p} \frac{\alpha}{\binom{d}{p} n(p+1)^2} + o(1) < \sum_{p=1}^{\infty} \frac{\alpha}{(p+1)^2} + o(1) < \alpha + o(1) \ ,
\end{aligned}$$

which implies this test is asymptotically of level α. ■

Lemma 2. *Let $\{M_k, \mathcal{F}_k\}_{k\in\mathbb{Z}_+}$ be a martingale difference sequence satisfying Conditions C4-(a) and C4-(b) in Theorem 1, and $Z_n = \sum_{k=1}^n M_k/\sqrt{n}$, then for every sequence $\tau_n \in \mathbb{Z}_+$ such that $\tau_n/n \to \lambda \in (0,1)$, we have*

$$\begin{pmatrix} Z_{\tau_n}\sqrt{\tau_n/n} \\ Z_n \end{pmatrix} \to_d \mathcal{N}\left(0, \begin{pmatrix} \lambda\mathcal{I}_0 & \lambda\mathcal{I}_0 \\ \lambda\mathcal{I}_0 & \mathcal{I}_0 \end{pmatrix}\right) \ .$$

Moreover,

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_{\tau_n} - \theta_0 \\ \hat{\theta}_n - \theta_0 \end{pmatrix} \to_d \mathcal{N}\left(0, \begin{pmatrix} \lambda^{-1}\mathcal{I}_0^{-1} & \mathcal{I}_0^{-1} \\ \mathcal{I}_0^{-1} & \mathcal{I}_0^{-1} \end{pmatrix}\right) \ .$$

Proof According to Cramér-Wold device, it is sufficient to show that for any $(a^\top, b^\top) \in \mathbb{R}^{2d}$,

$$a^\top\sqrt{\frac{\tau_n}{n}}Z_{\tau_n} + b^\top Z_n \to_d \mathcal{N}\left(0, \lambda(a+b)^\top\mathcal{I}_0(a+b) + (1-\lambda)b^\top\mathcal{I}_0 b\right), \text{as } n\to\infty \ .$$

In fact,

$$a^\top\sqrt{\frac{\tau_n}{n}}Z_{\tau_n} + b^\top Z_n = \sum_{k=1}^{\tau_n}(a+b)^\top\frac{M_k}{\sqrt{n}} + \sum_{k=\tau_n+1}^{n} b^\top\frac{M_k}{\sqrt{n}}.$$

Let $X_{n,k} = (a+b)^\top M_k$, if $k \in [\tau_n]$; and $X_{n,k} = b^\top M_k$, if $k \in \{\tau_n+1, \ldots, n\}$. Then $\{X_{n,k}, \mathcal{F}_k\}_{k\in\mathbb{Z}}$ is also a martingale difference sequence. Additionally,

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^n \mathbb{E}[X_{n,k}^2|\mathcal{F}_{k-1}] &= \frac{1}{n}\sum_{k=1}^{\tau_n}(a+b)^\top\mathbb{E}[M_kM_k^\top|\mathcal{F}_{k-1}](a+b) + \frac{1}{n}\sum_{k=\tau_n+1}^{n} b^\top\mathbb{E}[M_kM_k^\top|\mathcal{F}_{k-1}]b \\
&= \frac{\tau_n}{n}\frac{1}{\tau_n}\sum_{k=1}^{\tau_n} a^\top\,\mathbb{E}[M_kM_k^\top|\mathcal{F}_{k-1}](a+2b) + \frac{1}{n}\sum_{k=1}^{n} b^\top\,\mathbb{E}[M_kM_k^\top|\mathcal{F}_{k-1}]b \\
&\to_p \lambda a^\top\mathcal{I}_0(a+2b) + b^\top\mathcal{I}_0 b = \lambda(a+b)^\top\mathcal{I}_0(a+b) + (1-\lambda)b^\top\mathcal{I}_0 b \ ,
\end{aligned}$$

and, for any $\varepsilon > 0$,

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^n \mathbb{E}[X_{n,k}^2\mathbb{1}(|X_{n,k}| > \varepsilon\sqrt{n})|\mathcal{F}_{k-1}] \\
&= \frac{1}{n}\sum_{k=1}^{\tau_n}\mathbb{E}\left[\left((a+b)^\top M_k\right)^2\mathbb{1}(|(a+b)^\top M_k| > \varepsilon\sqrt{n})\middle|\mathcal{F}_{k-1}\right] \\
&+ \frac{1}{n}\sum_{k=\tau_n+1}^{n}\mathbb{E}\left[\left(b^\top M_k\right)^2\mathbb{1}(|b^\top M_k| > \varepsilon\sqrt{n})\middle|\mathcal{F}_{k-1}\right] \to_p 0 \ ,
\end{aligned}$$

by Condition C4-(b). Therefore, the statement holds by invoking the Lindeberg theorem for martingales, and it follows that

$$\begin{aligned}
\sqrt{n}\begin{pmatrix} \hat{\theta}_{\tau_n} - \theta_0 \\ \hat{\theta}_n - \theta_0 \end{pmatrix} &= \begin{pmatrix} \mathcal{I}_0^{-1}\sqrt{\frac{n}{\tau_n}}Z_{\tau_n} + o_p(1) \\ \mathcal{I}_0^{-1}Z_n + o_p(1) \end{pmatrix} = \begin{pmatrix} \mathcal{I}_0^{-1}/\lambda & 0 \\ 0 & \mathcal{I}_0^{-1} \end{pmatrix}\begin{pmatrix} \sqrt{\tau_n/n}Z_{\tau_n} \\ Z_n \end{pmatrix} + o_p(1) \\
&\to_d \mathcal{N}\left(0, \begin{pmatrix} \lambda^{-1}\mathcal{I}_0^{-1} & \mathcal{I}_0^{-1} \\ \mathcal{I}_0^{-1} & \mathcal{I}_0^{-1} \end{pmatrix}\right) \ .
\end{aligned}$$

■

Proposition 1. *Given an independent sample $X_1, \ldots, X_n$ and a family of marginal density functions $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ (for each individual X_i) satisfying that there exists $\tau_n \in [n-1]$ such that $X_1, \ldots, X_{\tau_n} \sim p_{\theta_0}$, $X_{\tau_n+1}, \ldots, X_n \sim p_{\theta_1} (\theta_1 \neq \theta_0)$, and $\tau_n/n \to \lambda \in (0,1)$. Assume the following conditions hold:*

C'1 : $F(\theta) := \lambda D_{KL}(p_{\theta_0} \| p_\theta) + \bar{\lambda} D_{KL}(p_{\theta_1} \| p_\theta)$ *has a unique minimizer* $\theta^* \in int(\Theta)$, *where* $\bar{\lambda} = 1 - \lambda$ *and* D_{KL} *is the KL-divergence.*

C'2 : Θ *contains an open neighborhood* Θ^* *of* θ^* *for which*

C'2-(a) : $\ell(\theta) := \ell(\theta|x) := \log p_\theta(x)$ *is twice continuously differentiable in* θ *almost surely.*

C'2-(b) : $\nabla^3_{ijk} \ell(\theta|x)$ *exists and satisfies* $\left|\nabla^3_{ijk} \ell(\theta|x)\right| \leq M_{ijk}(x)$ *for* $\theta \in \Theta^*$ *and* $i, j, k \in [d]$ *almost surely with* $\mathbb{E}_{\theta_l} M_{ijk}(X) < \infty$ *for* $l \in \{0, 1\}$.

C'3 : $\mathbb{E}_{\theta_l}[\nabla_\theta \ell(\theta^*)] = \nabla_\theta \mathbb{E}_{\theta_l}[\ell(\theta)]|_{\theta=\theta^*} = S_l^*$ *for* $l \in \{0, 1\}$.

C'4 : $\mathbb{E}_{\theta_l}[-\nabla^2_\theta \ell(\theta^*)] = \mathcal{I}_l^*$ *is positive definite for* $l \in \{0, 1\}$.

Then there exists a sequence of MLE such that $\hat{\theta}_n \to_p \theta^*$ *and*

$$\tfrac{1}{n} R_n(\tau_n) \to_p (\bar{\lambda} S_1^*)^\top (\mathcal{I}^*)^{-1} (\bar{\lambda} S_1^*), \text{ where } \mathcal{I}^* = \bar{\lambda} \mathcal{I}_1^* - \bar{\lambda} \mathcal{I}_1^* \left(\lambda \mathcal{I}_0^* + \bar{\lambda} \mathcal{I}_1^*\right)^{-1} \bar{\lambda} \mathcal{I}_1^* \ , \tag{10}$$

where $\mathcal{I}^*$ *is a positive definite matrix. If in addition* $S_1^* \neq 0$, *then*

$$\mathbb{P}\big(R_n(\tau_n) > q_{\chi^2_d}(\alpha)\big) \to 1 \ .$$

Proof For simplicity we assume that the likelihood equation $\nabla_\theta \ell_n(\theta) = 0$ has a unique solution $\hat{\theta}_n$, otherwise we choose $\hat{\theta}_n$ to be the solution closest to θ^* (this is possible since we are proving the existence). We firstly prove that $\hat{\theta}_n \to_p \theta^*$. For $\varepsilon > 0$ sufficiently small, let $B_\varepsilon = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\| \leq \varepsilon\} \subset \Theta^*$ and $\mathrm{bd}(B_\varepsilon)$ be the boundary of B_ε. We will show that, for sufficiently small ε,

$$\mathbb{P}\left(\ell_n(\theta) < \ell_n(\theta^*), \forall \theta \in \mathrm{bd}(B_\varepsilon)\right) \to 1 \ . \tag{11}$$

This implies, with probability converging to one, $\ell_n(\theta)$ has a local maximum (also a solution to the likelihood equation) in B_ε so $\hat{\theta}_n \in B_\varepsilon$. Consequently, $\mathbb{P}(\|\hat{\theta}_n - \theta^*\| > \varepsilon) \to 0$.

To prove (11), we write for any $\theta \in \mathrm{bd}(B_\varepsilon)$

$$\begin{aligned}
\frac{1}{n}[\ell_n(\theta) - \ell_n(\theta^*)] =& \frac{1}{n}(\theta - \theta^*)^\top \nabla_\theta \ell_n(\theta^*) - \frac{1}{2}(\theta - \theta^*)^\top \left(-\frac{1}{n} \nabla^2_\theta \ell_n(\theta^*)\right)(\theta - \theta^*) \\
&+ \frac{1}{6n} \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d (\theta_i - \theta_i^*)(\theta_j - \theta_j^*)(\theta_k - \theta_k^*) \nabla_{ijk} \ell_n(\bar{\theta}_n) \\
:=& D_1 + D_2 + D_3 \ ,
\end{aligned}$$

where $\bar{\theta}_n \in B_\varepsilon$ satisfies $\left\|\bar{\theta}_n - \theta^*\right\| \leq \|\theta - \theta^*\|$. Note that, by Condition C'3:

$$\begin{aligned}
D_1 \to_p & (\theta - \theta^*)^\top \left[\lambda \mathbb{E}_{\theta_0}[\nabla_\theta \ell(\theta^*)] + \bar{\lambda} \mathbb{E}_{\theta_1}[\nabla_\theta \ell(\theta^*)]\right] \\
&= (\theta - \theta^*)^\top \nabla_\theta \left[\lambda \mathbb{E}_{\theta_0}[\ell(\theta)] + \bar{\lambda} \mathbb{E}_{\theta_1}[\ell(\theta)]\right]\big|_{\theta=\theta^*} \\
&= -(\theta - \theta^*)^\top \nabla_\theta \left[\lambda D_{KL}(p_{\theta_0} \| p_\theta) + \bar{\lambda} D_{KL}(p_{\theta_1} \| p_\theta)\right]\big|_{\theta=\theta^*} \\
&= 0 \ ,
\end{aligned}$$

where the last equality follows from Condition C'1. Moreover, by Condition C'4,

$$D_2 \to_p -\frac{1}{2}(\theta-\theta^*)^\top \left(\lambda\mathcal{I}_0^* + \bar{\lambda}\mathcal{I}_1^*\right)(\theta-\theta^*) \leq -\frac{1}{2}\lambda_{\min}\varepsilon^2 \ ,$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\lambda\mathcal{I}_0^* + \bar{\lambda}\mathcal{I}_1^*$. If we let ε small enough such that $\mathrm{bd}(B_\varepsilon) \subset \Theta^*$, then according to Condition C'2, for all $\theta \in \mathrm{bd}(B_\varepsilon)$,

$$\begin{aligned}
|D_3| &\leq \frac{1}{6n}\sum_{ijk} \|\theta_i - \theta_i^*\| \left\|\theta_j - \theta_j^*\right\| \|\theta_k - \theta_k^*\| \sum_{l=1}^n \left|\nabla_{ijk}\ell(\bar{\theta}_n|X_l)\right| \\
&\leq \frac{1}{6}\varepsilon^3 \sum_{ijk} \frac{1}{n}\sum_{l=1}^n M_{ijk}(X_l) \\
&\to_p \frac{\varepsilon^3}{6}\sum_{ijk}\left(\lambda\mathbb{E}_{\theta_0}[M_{ijk}(X)] + \bar{\lambda}\mathbb{E}_{\theta_1}[M_{ijk}(X)]\right) \ .
\end{aligned}$$

Hence, for any given $\delta > 0$, any $\varepsilon > 0$ sufficiently small, any n sufficiently large, with probability larger than $1-\delta$, we have, for all $\theta \in \mathrm{bd}(B_\varepsilon)$,

$$\begin{aligned}
|D_1| &< \varepsilon^3 \\
D_2 &< -\lambda_{\min}\varepsilon^2/4 \\
|D_3| &\leq A\varepsilon^3 \ ,
\end{aligned}$$

where $A > 0$ is a constant. It follows that,

$$D_1 + D_2 + D_3 < \varepsilon^3 + A\varepsilon^3 - \frac{\lambda_{\min}}{4}\epsilon^2 = \left((A+1)\varepsilon - \frac{\lambda_{\min}}{4}\right)\varepsilon^2 < 0, \text{if } \varepsilon < \frac{\lambda_{\min}}{4(A+1)} \ ,$$

and thus (11) holds.

Now according to continuous mapping theorem and Slutsky's theorem (see, for instance [Billingsley, 2008]) and to Eq. (3)

$$\begin{aligned}
\frac{1}{n}S_{(\tau_n+1):n}(\hat{\theta}_n) &\to_p \bar{\lambda}S_1^* \\
\frac{1}{n}\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau_n) &\to_p \bar{\lambda}\mathcal{I}_1^* - \bar{\lambda}\mathcal{I}_1^*\left(\lambda\mathcal{I}_0^* + \bar{\lambda}\mathcal{I}_1^*\right)^{-1}\bar{\lambda}\mathcal{I}_1^* \equiv \mathcal{I}^* \ ,
\end{aligned}$$

where $\mathcal{I}^*$ is positive definite since both $\mathcal{I}_0^*$ and $\mathcal{I}_1^*$ are positive definite. This implies

$$\begin{aligned}
\frac{1}{n}R_n(\tau_n) &= \left(\frac{1}{n}S_{(\tau_n+1):n}(\hat{\theta}_n)\right)^\top \left(\frac{1}{n}\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau_n)\right)\left(\frac{1}{n}S_{(\tau_n+1):n}(\hat{\theta}_n)\right) \\
&\to_p (\bar{\lambda}S_1^*)^\top(\mathcal{I}^*)^{-1}(\bar{\lambda}S_1^*) \ .
\end{aligned}$$

If in addition one has $S_1^* \neq 0$, then it follows from the positive definiteness of $\mathcal{I}^*$ that $\mathbb{P}\left(R_n(\tau_n) > q_{\chi_d^2}(\alpha)\right) \to 1$. ■

Proposition 2. *Given an independent sample $X_1, \ldots, X_n$ and a family of marginal density functions $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ satisfying that there exists $\tau_n \in [n-1]$ such that $X_1, \ldots, X_{\tau_n} \sim p_{\theta_0}$, $X_{\tau_n+1}, \ldots, X_n \sim p_{\theta_n}$ in which $\theta_n = \theta_0 + hn^{-1/2}$ with $h \neq 0$, and $\tau_n/n \to \lambda \in (0,1)$. We denote the joint probability measure of $X_1, \ldots, X_n$ as $\mathbb{P}_{\theta_0,\theta_n}(\tau_n)$. Assume the following conditions hold:*

C"1 : θ_0 *is the unique maximizer of* $\mathbb{E}_0[\ell(\theta)]$.

C"2 : Θ *contains an open neighborhood* Θ_0 *of* θ_0 *for which*

C"2-(a) : $\ell(\theta) := \ell(\theta|x) := \log p_\theta(x)$ *is twice continuously differentiable in* θ *almost surely.*

C"2-(b) : $\nabla^3_{ijk}\ell(\theta|x)$ *exists and satisfied* $\left|\nabla^3_{ijk}\ell(\theta|x)\right| \leq M_{ijk}(x)$ *for* $\theta \in \Theta_0$ *and* $i,j,k \in [d]$ *almost surely with* $\mathbb{E}_{\theta_0} M_{ijk}(X) < \infty$.

C"3 : $\mathbb{E}_{\theta_0}[\nabla_\theta \ell(\theta_0)] = \nabla_\theta \mathbb{E}_{\theta_0}[\ell(\theta)]|_{\theta=\theta_0} = S_0$.

C"4 : $\mathbb{E}_{\theta_0}[-\nabla^2_\theta \ell(\theta_0)] = \mathcal{I}_0$ *is positive definite.*

Then there exists a sequence of MLE $\hat{\theta}_n$ such that

$$\frac{n}{\tau_n(n-\tau_n)}\hat{\mathcal{I}}_n(\hat{\theta}_n;\tau) \to_p \mathcal{I}_0 \quad , \quad \sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}_d\left(\bar{\lambda}h, \mathcal{I}_0^{-1}\right), \tag{12}$$

$$\text{and} \quad \sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{(\tau_n+1):n}(\hat{\theta}_n) \to_d \mathcal{N}_d(\sqrt{\lambda\bar{\lambda}}\, \mathcal{I}_0 h, \mathcal{I}_0) \ . \tag{13}$$

Hence, $R_n(\tau_n)$ converges to a non-central chi-square distribution with degrees of freedom d and parameter $\lambda\bar{\lambda}h^\top \mathcal{I}_0 h$.

Proof In this proof we firstly analyze the behavior of the score statistic under the null hypothesis, then we use Le Cam's third lemma (see [Van der Vaart, 2000]) to attain the asymptotic distribution of the test statistic under local alternatives.

Under $\mathbb{P}_0 = \mathbb{P}_{\theta_0}$, an argument similar to the one in Proposition 1 implies that there exists a sequence of MLE such that $\hat{\theta}_n \to_p \theta_0$, then (12) directly follows from the proof in Theorem 1. Furthermore, by Condition C"2-(a) and the mean value theorem, there exists $\bar{\theta}_n$ such that $\|\bar{\theta}_n - \theta_0\| \leq \|\hat{\theta}_n - \theta_0\|$, and

$$0 = \frac{1}{\sqrt{n}} S_{1:n}(\hat{\theta}_n) = \frac{1}{\sqrt{n}} S_{1:n}(\theta_0) - \frac{1}{n}\nabla_\theta S_{1:n}(\bar{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta_0) \ ,$$

Therefore,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathcal{I}_0^{-1}\frac{1}{\sqrt{n}} S_{1:n}(\theta_0) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\ell}_i(\theta_0) + o_p(1) \ ,$$

where $\tilde{\ell}_i(\theta_0) = \mathcal{I}_0^{-1}\nabla_\theta \ell_i(\theta_0)$ is the normalized score for X_i. Additionally, the log-likelihood ratio is asymptoticly linear:

$$\log \frac{d\mathbb{P}_{\theta_0,\theta_n}(\tau_n)}{d\mathbb{P}^n_{\theta_0}} = \ell_{(\tau_n+1):n}(\theta_n) - \ell_{(\tau_n+1):n}(\theta_0) = h^\top \frac{1}{\sqrt{n}} S_{(\tau_n+1):n}(\theta_0) - \frac{\bar{\lambda}}{2} h^\top \mathcal{I}_0 h + o_p(1).$$

For any $a \in \mathbb{R}^d$, it follows from the multivariate Central Limit Theorem [Billingsley, 2008] that

$$\begin{aligned}\begin{pmatrix} a^\top\sqrt{n}(\hat{\theta}_n - \theta_0) \\ \log \frac{d\mathbb{P}_{\theta_0,\theta_n}(\tau_n)}{d\mathbb{P}^n_{\theta_0}} \end{pmatrix} &= \frac{1}{\sqrt{n}}\left[\sum_{i=1}^{\tau_n}\begin{pmatrix} a^\top\tilde{\ell}_i(\theta_0) \\ 0 \end{pmatrix} + \sum_{i=\tau_n+1}^{n}\begin{pmatrix} a^\top\tilde{\ell}_i(\theta_0) \\ h^\top\nabla_\theta\ell_i(\theta_0) \end{pmatrix}\right] - \begin{pmatrix} 0 \\ \frac{\sigma^2}{2} \end{pmatrix} + o_p(1) \\ &\to_d \mathcal{N}_2\left(\begin{pmatrix} 0 \\ -\sigma^2/2 \end{pmatrix}, \begin{pmatrix} a^\top\mathcal{I}_0^{-1}a & \bar{\lambda}a^\top h \\ \bar{\lambda}a^\top h & \sigma^2 \end{pmatrix}\right) \ ,\end{aligned}$$

where $\sigma^2 := \bar{\lambda} h^\top \mathcal{I}_0 h$. Hence the assumptions of Le Cam's third lemma are fulfilled, we conclude that, under $\mathbb{P}_{\theta_0,\theta_n}(\tau_n)$,

$$a^\top \sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}\left(\bar{\lambda} a^\top h, a^\top \mathcal{I}_0^{-1} a\right) \ ,$$

and by the Cramer-Wold device, the second part of Eq. (12) holds.

Notice that, under $\mathbb{P}_{\theta_0}$,

$$\begin{aligned}\frac{1}{\sqrt{n}} S_{(\tau_n+1):n}(\hat{\theta}_n) &= \frac{1}{\sqrt{n}} S_{(\tau_n+1):n}(\theta_0) - \bar{\lambda}\mathcal{I}_0 \sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1) \\ &= \frac{1}{\sqrt{n}} \left[\sum_{i=1}^{\tau_n} -\bar{\lambda} \nabla_\theta \ell_i(\theta_0) + \sum_{i=\tau_n+1}^{n} \lambda \nabla_\theta \ell_i(\theta_0) \right] + o_p(1) \ .\end{aligned}$$

Similarly we get

$$\frac{1}{\sqrt{n}} S_{(\tau_n+1):n}(\hat{\theta}_n) \to_d \mathcal{N}_d(\lambda\bar{\lambda}\mathcal{I}_0 h, \lambda\bar{\lambda}\mathcal{I}_0) \ ,$$

and (13) follows. ■

C Experimental details

In this section, we perform simulations to evaluate empirical behavior of our methods on synthetic data generated from a linear regression model, a time series model, a hidden Markov model (HMM), and a text topic model described in [Stratos et al., 2015], which is essentially an HMM with discrete emission distribution such that only one is positive of the emission probabilities (conditioning on different states) for each category. We apply our approach to detect changes in this topic model on subtitles of TV shows.

Experimental settings. Let $\mathbb{P}_{k,\theta}$ be the model (conditional probability), n be the sample size, $\theta_0 \in \mathbb{R}^d$ be the parameter before change, and $\delta \geq 0$ be the magnitude of change. We generate the first half of the sample from $\mathbb{P}_{1,\theta_0}, \ldots, \mathbb{P}_{[n/2],\theta_0}$. Then we obtain θ_1 by adding δ to the first p components of θ_0 (so that $\delta = \|\Delta\| / \sqrt{p}$ with $\Delta = \theta_1 - \theta_0$). Next, we generate the second half from $\mathbb{P}_{[n/2]+1,\theta_1}, \ldots, \mathbb{P}_{n,\theta_1}$, and run the linear test given in (5), the scan test given in (6), and the autograd-test given in (7) on this dataset, where the significance levels are set to be $\alpha = 2\alpha_l = 2\alpha_s = 0.05$. And these statistics are computed only for $\tau \in [n/10, 9n/10]$ to prevent encountering ill-conditioned Fisher information matrix. We repeat this procedure 200 times and approximate power by the frequency of rejections. We plot the power curve by varying the values of δ, where we use three different types of lines to represent three tests, and different colors to indicate different sample sizes.

Linear regression model. We consider a linear regression model with 100 slope coefficients and intercept (*i.e.,* $d = 101$), and investigate two sparsity levels, $p = 1$ and $p = 20$. The coefficients and intercept are fixed to be zero. All the entries of the design matrix and error terms are generated independently from a standard normal distribution. As shown in Fig. 3, when the change is sparse ($p = 1$), the scan test and the autograd-test share similar power curves and both outperform the

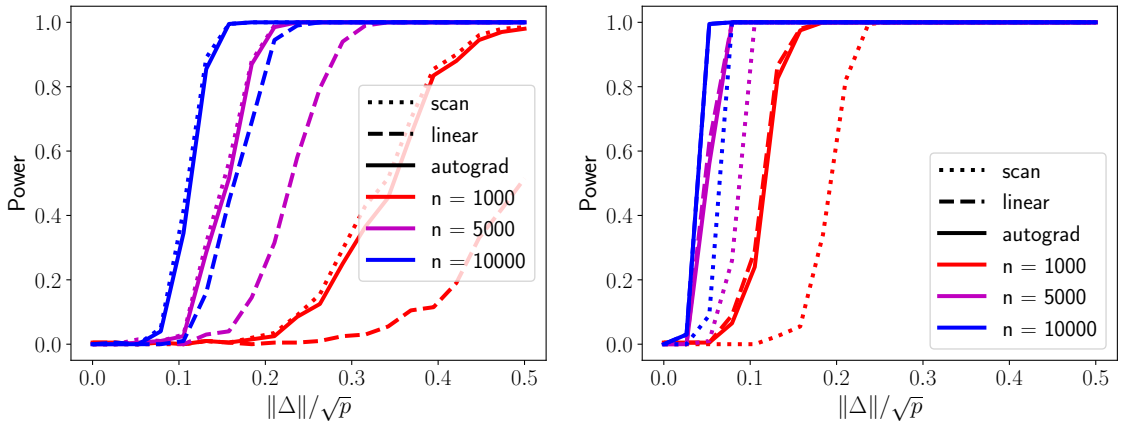

Figure 3: Power versus magnitude of change for linear models (left: $p = 1$; right: $p = 20$). Three types of lines indicates three tests, and three colors indicates three sample sizes.

linear test significantly. When the change is less sparse ($p = 20$), all tests' performance improve to a large extend since the change signal becomes strong, with the scan test tending to perform poorer than the other two. This empirically illustrates that 1) the scan test works better in detecting sparse changes, 2) the linear test is more powerful for non sparse changes and 3) the autograd-test achieves comparable performance in both situations.

Additionally, we consider the same linear regression model with $p = 1$, except that the detection is limited to 50 components of the slope coefficients (*i.e.,* regard the rest 51 components as constants). Results in Fig. 4 show that when the restricted components contain the changed one, all tests have improved performance, while the linear test improves to a larger extend. When the abnormal component is outside the scope of the detection, the true positive rate is below 0.01 no matter how strong the signal is. Hence, these tests can filtrate irrelevant information by simply confining the detection to parameters of interest.

Time series model. We then investigate two different autoregressive–moving-average models—ARMA$(3, 2)$ and ARMA$(6, 5)$. To acquire stationary time series, we have a slightly different procedure. We firstly sample $p_0 \in \{3, 6\}$ values that are larger than 1, say $\lambda_1, \ldots, \lambda_{p_0}$, then use the coefficients of the polynomial $f_0(x) = \prod_{i=1}^{p_0}(x - \lambda_i^{-1})$ as AR coefficients; MA coefficients are obtained similarly. Furthermore, the post-change AR coefficients are created by adding δ to those p_0 values and extracting the coefficients from $f_1(x) = \prod_{i=1}^{p_0}(x - (\lambda_i + \delta)^{-1})$. The error terms follow a normal distribution with mean 0 and standard deviation 0.1. We remark that for ARMA models we do not have exact control of $\|\Delta\| / \sqrt{p}$. Readers need to be careful about the range of x-axis in Fig. 5.

As we can see, the scan test works fairly well for these two ARMA models. However, the linear test and the autograd-test have extremely high type I error (*i.e.,* false discovery rate or FDR). This problem gets even severer as the sample size increases, and hence is not due to lack of accuracy of the MLE. It turns out that it is caused by the non-homogeneity of model parameters—the derivatives *w.r.t.* AR coefficients tend to be of different magnitude compared to the ones *w.r.t.* MA coefficients. This results in ill-conditioned Schur complement (3) and subsequent unstable computation of the linear statistic. On the contrary, the scan statistic only inverts the submatrix of size $p \times p$. Since $p \lessapprox \sqrt{d}$, the submatrix has much smaller condition number (the parameters selected by the scan

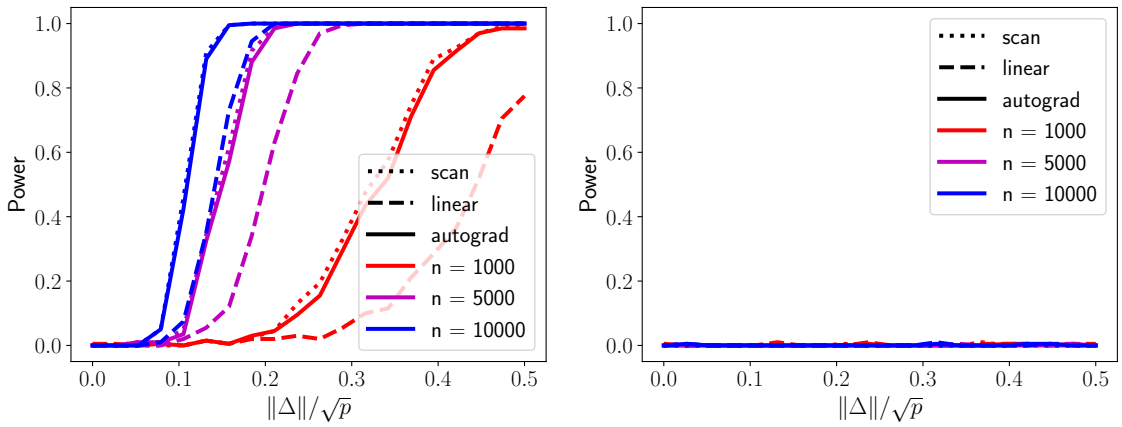

Figure 4: Power versus magnitude of change for linear regression with restriction (left: contains the changed component; right: does not contain the changed component). Three types of lines indicates three tests, and three colors indicates three sample sizes.

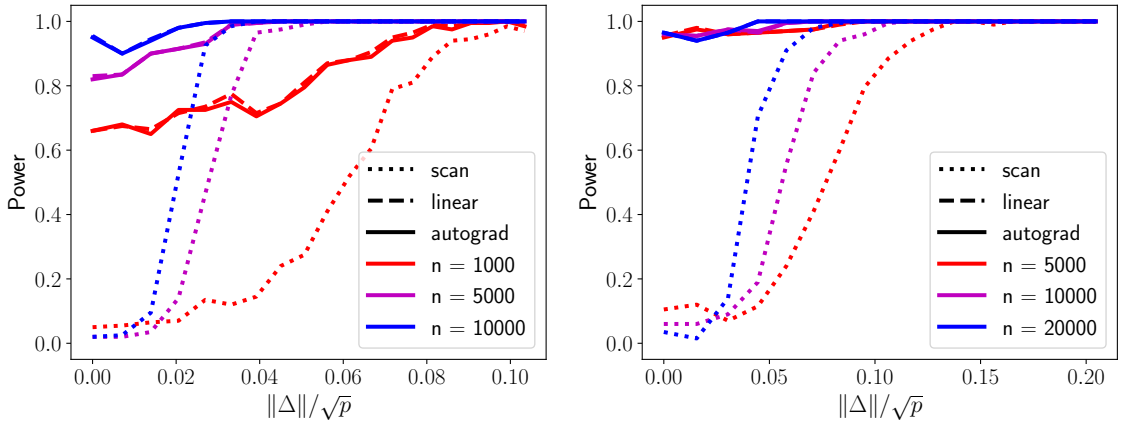

Figure 5: Power versus magnitude of change for ARMA$(3, 2)$ (left) and ARMA$(6, 5)$ (right). Three types of lines indicates three tests, and three colors indicates three sample sizes.

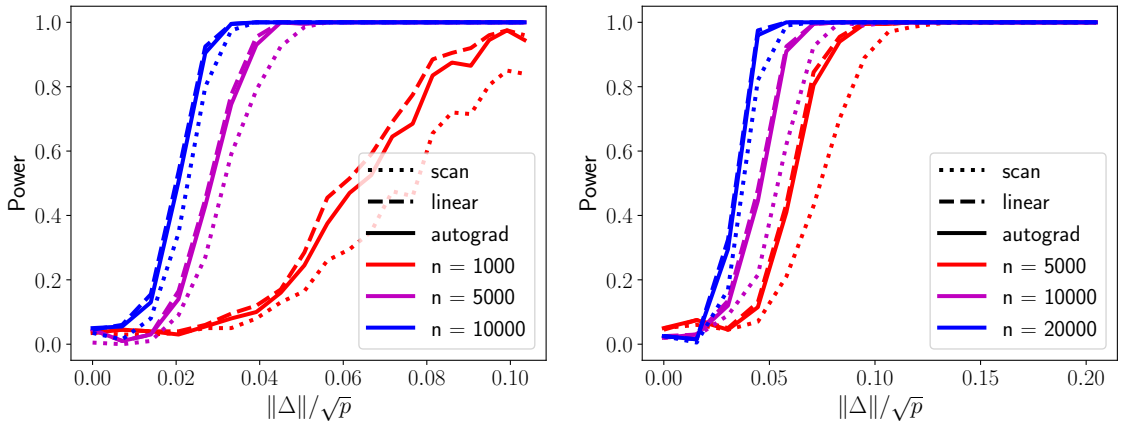

Figure 6: Power versus magnitude of change for ARMA models with restricted components (left: ARMA$(3, 2)$; right: ARMA$(6, 5)$). Three types of lines indicates three tests, and three colors indicates three sample sizes.

statistic are all AR coefficients in our experiments). Therefore, the scan statistic can produce reasonable results even though the parameters are heterogeneous. We remark that in such situations we can select a small (or even zero) significance level for the linear part of the autograd-test (so the scan part has a dominating effect) to obtain comparable results. If we restrict the detection of these tests in the AR coefficients, as presented in Fig. 6, all three tests are now consistent in level, and the linear test and the autograd-test are slightly more powerful than the scan test.

Hidden Markove model. We then investigate HMMs with a number of hidden states $N \in \{3, 7, 15\}$ and normal emission distribution. Its transition matrix is sampled in the following way: each row (the distribution of next state conditioning on current state) is the sum of vector $(1/(2N), \ldots, 1/(2N))$ and a Dirichlet sample with concentration parameters $(0.5, \ldots, 0.5)$ (these two vectors are of length N). Conditioning on the state being $k \in \{0, \ldots, N-1\}$, the emission distribution has mean k and standard deviation $0.01 + 0.09k/(N-1)$ so that they are evenly distributed within $[0.01, 0.1]$. Note that the transition matrix subjects to the constraint that each row must sum to one, so we only view entries in the first $N-1$ columns as transition parameters. Results are shown in Fig. 7.

When $N = 3$, three tests have almost identical performance. When $N = 7$, the change becomes sparser, and subsequently, the scan test and the autograd-test demonstrate outperform the linear test. When $N = 15$, the linear test and autograd-test become inconsistent, but, different from the situation for ARMA models, the inconsistency are alleviated as the sample size increases. Note that for $N = 15$ some states pair might be in low frequency, so the estimate of the associated transition probability can be of poor accuracy. This sometimes results in non-invertible full Fisher information. We view this situation as lack of evidence and do not reject the null hypothesis, which accounts for the power oscillating around 0.8.

Text topic model. Finally, we exam text topic model with different parameter schemes: $(N, M) \in \{(3, 6), (7, 20), (15, 150)\}$, where N is the number of hidden states, and M is the number of categories for emission distribution. We use the same way as HMMs to generate its transition matrix. The

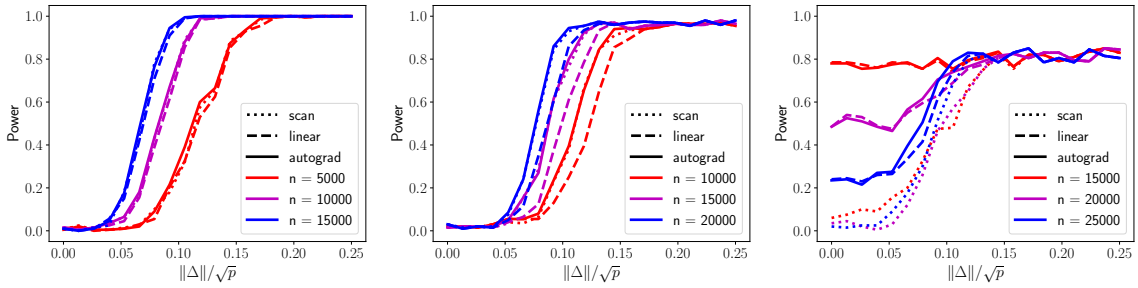

Figure 7: Power versus magnitude of change for HMMs with N hidden states (left: $N = 3$; middle: $N = 7$; right: $N = 15$). Three types of lines indicates three tests, and three colors indicates three sample sizes.

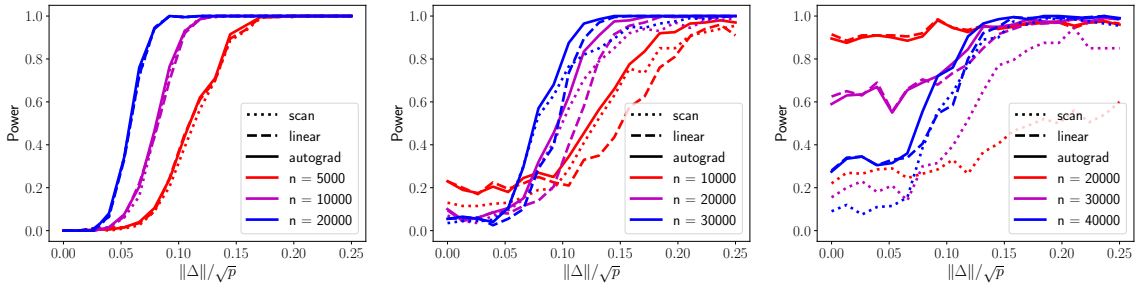

Figure 8: Power versus magnitude of change for text topic models (left: $(N, M) = (3, 6)$; middle: $(N, M) = (7, 20)$; right: $(N, M) = (15, 150)$). Three types of lines indicates three tests, and three colors indicates three sample sizes.

emission matrix is sampled analogically except that each row can only have one nonzero entry, which is required by this model. As shown in Fig. 8, the results are similar to the ones for HMMs. Since the topic model has more parameters than the HMM with the same N, the inconsistency in level is exacerbated.

We now consider a real data application. We collect subtitles of the first two seasons of four different TV shows—Friends, Modern Family, the Sopranos, and Deadwood [7]—where the former two are assumed to be "polite" and the latter two are assumed to be "rude". The task is to detect changes in rudeness level.

After preprocessing steps (such as remove punctuation and stop words, tokenization, and lemmatization, see attached code for complete steps), we arrange all the text in each season into a long series of words. For every pair of series, we train the aforementioned topic model on them, where the number of hidden states is chosen to be $[\sqrt{n/100}]$ so that each entry in the transition matrix are estimated using about 100 observations on average, and the number of categories is the size of vocabulary built from the training corpus. In order to avoid dealing with ill-conditioned information matrices, we only apply the scan test to detect changes within the middle half sample (*i.e.,* from $[n/4]$ to $[3n/4]$) on this dataset. We also restrict the detection in transition parameters because latent variables tend to capture global information while emission parameters are much easier to alter due to the shift of high frequency words.

[7] Downloaded from *http://www.tvsubtitles.net.*

Table 2: Rejection status for pair-wise experiments.

	F1	F2	M1	M2	S1	S2	D1	D2
F1	NR	NR	NR	NR	R	R	R	R
F2	NR	NR	R	NR	R	R	R	R
M1	NR	R	NR	NR	R	R	R	R
M2	NR	NR	NR	NA	R	R	R	R
S1	R	R	R	R	NR	NR	R	R
S2	R	R	R	R	NR	NR	R	R
D1	R	R	R	R	R	R	NR	R
D2	R	R	R	R	R	R	NR	NR

As demonstrated in Table 2, the scan test does a perfect job in reporting shifts in rudeness level. However, it tends to be liberal in this task, that is, the false discovery rate is relatively high. For ("polite", "polite") pairs, there are two false positives and one NA because the MLE of the transition matrix contains zero which makes the statistic undefined; while for ("rude", "rude") pairs, 9 out of 16 are false positives, suggesting the existence of discrepancy in other aspects. These results are only based on one experiment, and the order of episodes in each season remain unchanged during the experiment, so the results may be subject to some unknown effects of the order.

To eliminate this possibility and obtain more robust results, we randomly shuffle the episodes (as a whole) in each season, then detect changes using these new series. We repeat this process 200 times, and count the frequency of rejections. Table 3 shows a similar phenomenon, and, as expected, the scan test is particularly good at distinguishing two shows from being the same one. For (F, M) and (M, F) pairs, four of them have a high rejection rate, which can be viewed as FDR in this task. When it comes to (S, D) and (D, S) pairs, most rejection rates are extremely high except for pairs coming from the same show.

We remark that rudeness is definitely not the only factor that contributes to the difference between two shows, and there is no reason to believe it is the only factor that the scan test utilizes to detect changes (it might not be one). But the results are promising in the sense that the scan statistic is able to neglect some low level discrepancies and focus on "global information" in language level. As we already discussed, we can confine the detection to some specific parameters, if it is possible to determine which ones are related to the rudeness, and obtain a more appropriate test for this specific task.

Online extension. For the autograd-test-CuSum algorithm, We consider a linear regression model with 10 slope coefficients. We first train and initialize the model with 1000 observations. Then we generate τ observations before change and $n - \tau$ observations after change (with $p = 1$), and run the autograd-test-CuSum algorithm until fulfilling the stopping criterion or reaching the end of the sequence. We denote the stopping time by t_a, with $t_a = n + 1$ suggesting stop with no rejection. If $\tau \leq t_a \leq n$, we call it a detection; if $t_a < \tau$, we view it as an abnormal stop. To assess the performance of the autograd-test-CuSum method, we repeat the procedure 200 times and compute the proportion of detections among normal stops. Another important metric is the delay of detections, measuring how fast the algorithm can detect the anomaly when a change actually happens. This is usually quantified by the conditional mean delay, $\mathbb{E}[t_a - \tau | \tau \leq t_a \leq n]$, and we

Table 3: Rejection rate for pair-wise experiments.

	F1	F2	M1	M2	S1	S2	D1	D2
F1	0.060	0.230	0.235	0.305	0.995	0.975	0.995	1.000
F2	0.195	0.115	0.525	0.425	0.955	0.975	1.000	1.000
M1	0.235	0.460	0.020	0.180	0.980	0.975	1.000	1.000
M2	0.300	0.405	0.155	0.000	0.985	0.960	1.000	1.000
S1	1.000	0.985	0.975	0.975	0.135	0.200	1.000	0.995
S2	0.995	0.995	0.985	0.925	0.190	0.220	1.000	0.985
D1	1.000	1.000	1.000	0.990	0.970	0.980	0.175	0.305
D2	1.000	1.000	0.985	1.000	0.995	0.990	0.305	0.195

estimate it by the average of $t_a - \tau$ among all detections.

In the first scenario, we fix $\tau = 1000$ and vary $n - \tau \in \{1000, 5000, 10000, 20000, 30000\}$. When the sample size is relatively small ($n - \tau = 1000$), the autograd-test-CuSum is of level 0.05 and achieves power 1 as the magnitude of change amplifies. When $n - \tau = 5000$, the power, as well as the FDR, presents a significant increment. As the sample size increases, the FDR slightly rises but the power declines greatly. We speculate that this phenomenon is due to the approximation error of the MLE, the score, and the information matrix since we are computing them in an online fashion. For the conditional mean delay, as expected, data with larger sample size tend to bring about longer delay and sharper decline (with the magnitude of change), since the growth of change magnitude has more impact on data with bigger post-change sample.

In the second scenario, we fix $n - \tau = 5000$ and choose $\tau \in \{1000, 5000, 10000, 20000, 30000\}$. The the true discovery rate shows a downward trend as τ increase, which coincides with the intuition that the larger τ is, the less evidence the data possess in the existence of a change. For strong signal the conditional mean delay exhibit a similar behavior as the first scenario while it is exactly the opposite case for weak signal. This is due to the fact that virtually the conditional mean delay is computed using only a small portion of the experiments where there is a detection, and this kind of detection behaves like random noise (as $\delta = 0$) rather than a reaction of signals.

As last, we remark that even though the type I errors are out of control, they are close to 0.2, which calls for a correction to the threshold, accounting for the reinitialization employed in autograd-test-CuSum.

We also run the autograd-test-CuSum algorithm on linear regression with 100 slope coefficients, where the model is initially trained using 5000 observations. As exhibited in Fig. 10, contrary to linear regression with 10 slope coefficients, the power increases as the sample size gets larger in both situations, and the type I error becomes uncontrolled; while for the conditional mean delay, it shows similar trends.

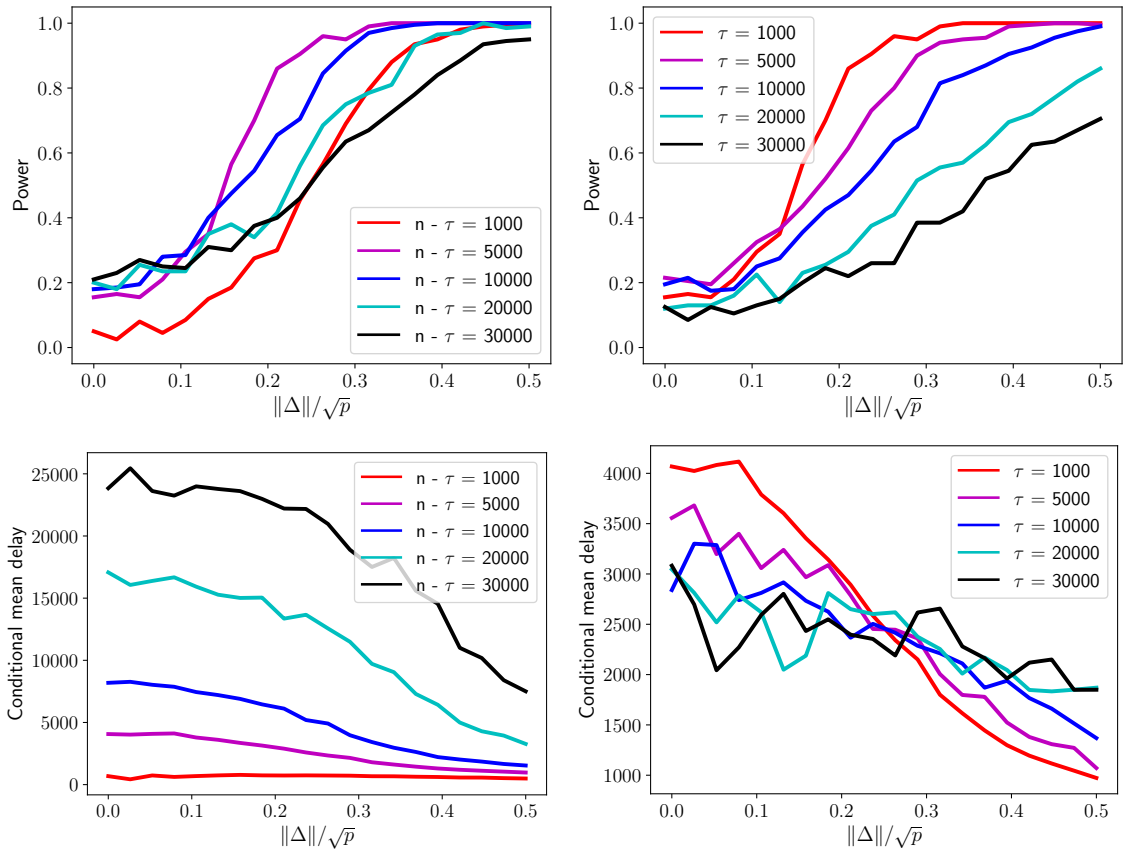

Figure 9: Power versus magnitude of change (up) and conditional average run length versus magnitude of change (down) for autograd-test-CuSum (left: $\tau = 1000$; right: $n - \tau = 5000$).

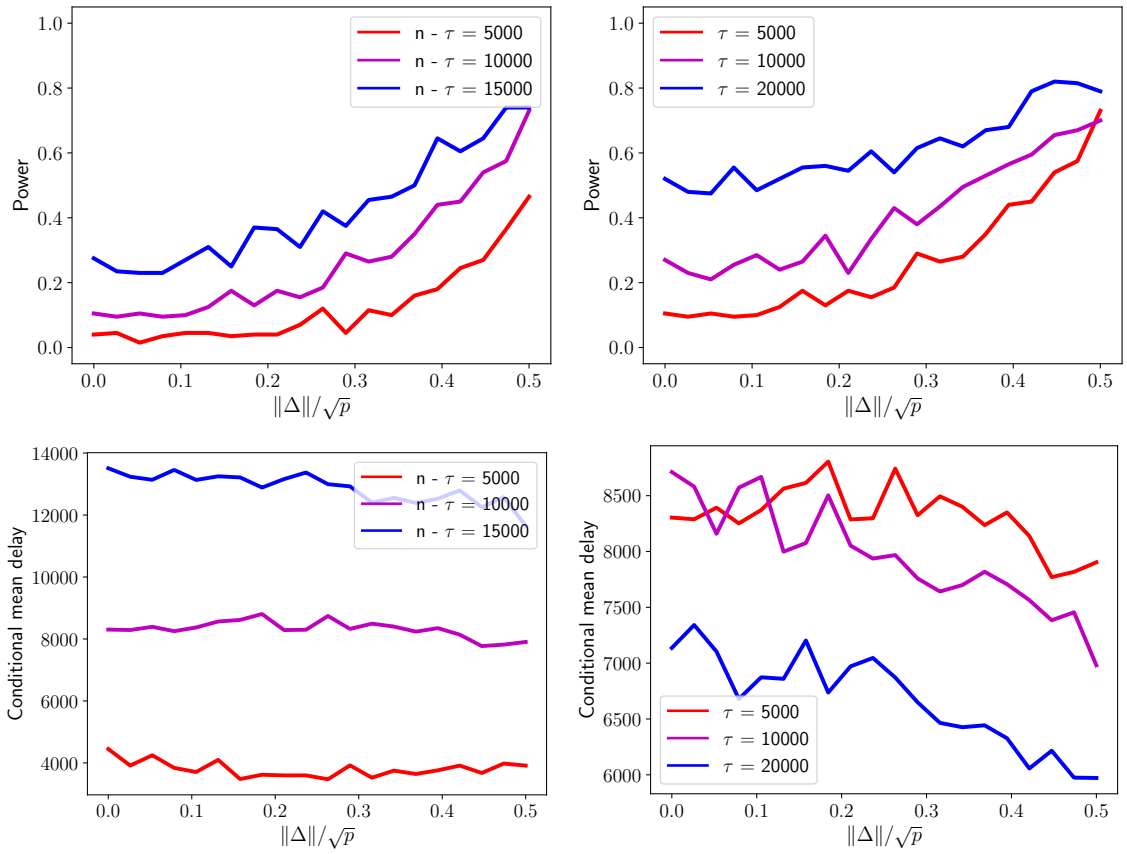

Figure 10: Power versus magnitude of change (up) and conditional average run length versus magnitude of change (down) for autograd-test-CuSum (left: $\tau = 5000$; right: $n - \tau = 10000$).