Gradient-Based Monitoring of Learning Machines

Lang Liu[1] Joseph Salmon[2] Zaid Harchaoui[1]

[1] Department of Statistics, University of Washington, Seattle, USA
[2] IMAG, Univ. Montpellier, CNRS, Montpellier, France

March 08, 2020

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 generic score-based change detection method that can detect a change in any number of (hidden) components of a machine learning model trained via empirical risk minimization. This proposed statistical hypothesis test can be readily implemented for such models designed within a differentiable programming framework. We establish the consistency of the hypothesis test and show how to calibrate it based on our theoretical results. We illustrate the versatility of the approach on additive 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 flexible statistical machine learning modeling libraries designed within a differentiable programming framework, *e.g.,* PyTorch [Paszke et al., 2019] and TensorFlow [Abadi et al., 2016].

First-order or gradient-based optimization algorithms such as accelerated batch gradient methods or averaged incremental gradient methods are then well adapted to this framework, opening up the possibility of gradient-based training of machine learning models from a continuous stream of data. As a learning system learns from a continuous, possibly evolving, data stream, it is desirable to supplement it with tools facilitating its monitoring over time in order to prevent the learned model from experiencing abnormal changes.

Recent remarkable failures of intelligent learning systems such as Microsoft's chatbot [Metz, 2018] and Uber's self-driving car [Knight, 2018] as they were unleashed "in the wild" show the importance of such tools. In the case of the Microsoft's chatbot, the initially learned language model quickly changed to an undesirable one, as it was being fed data through interactions with users. The addition of an automatic monitoring tool could have potentially prevented this debacle by triggering an early alarm, drawing the attention of its designers and engineers to any abnormal changes of this language model.

The design of monitoring tools for current learning systems can however be challenging. As N. Wiener had foreseen [Wiener, 1948], "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". Pursuing N. Wiener's argument, the speed of monitoring tools for learning systems should be comparable to the speed of their learning process.

Therefore, in order to keep up with modern learning machines, the monitoring of machine learning models should be automatic and effortless in the same way that the training of these models is now automatic and effortless. Humans monitoring machines should have at hand automatic monitoring tools to scrutinize a

learned model as it evolves over time. Recent research in this area has been relatively limited, while progress in learning systems has been flourishing.

We take here a statistical decision theory approach, that is, we build a hypothesis test to test the null hypothesis that the model has not changed against the alternative hypothesis that it has changed. This approach allows the user to set a desired false alarm rate and calibrate the hypothesis test accordingly.

Relationships to previous works. We review here previous works in this area. Change detection is a classical topic in statistics and signal processing; see [Basseville and Nikiforov, 1993, Tartakovsky et al., 2014] for a survey. The change detection problem has been considered either in the offline setting, where we test the null hypothesis with a prescribed false alarm rate, or in the online setting, where we detect a change as quickly as possible, minimizing the detection delay. In practice, an offline change detection hypothesis test can also be used in a sliding window manner. Based on the type of changes, the change detection problem can also be classified into two main categories: change detection in model parameters and change detection in the distribution of data streams.

We focus on detecting changes in model parameters. As we will show in the next section, this allows one to scan model components to uncover *weak changes*, *i.e.,* changes occurring on a small subset of model parameters. Moreover, our goal is to develop a generic approach aligned with current machine learning softwares such as PyTorch and TensorFlow, which are developed in a differentiable programming framework. This alignment allows us to seamlessly apply our approach to a large class of machine learning models implemented in such frameworks.

Test statistics for detecting changes in model parameters are usually designed on a case-by-case basis; see [Hinkley, 1970, Lorden, 1971, Deshayes and Picard, 1986, Basseville and Nikiforov, 1993, Carlstein et al., 1994, Csörgő and Horváth, 1997] and references therein. These methods are usually based on (possibly generalized) likelihood ratios or on residuals and therefore not amenable to differentiable programming. Furthermore, these methods are limited to *strong changes*, *i.e.,* changes occurring simultaneously on all model parameters, in contrast to ours.

One could contemplate detecting changes in the distribution of data streams instead, as in [Kifer et al., 2004, Cunningham et al., 2012, Volkhonskiy et al., 2017]. This raises other challenges since data streams in many modern machine learning applications can be discrete structured data such as sequences, trees and graphs, or lattice-structured data such as signals, images and voxels. On the contrary, machine learning models are always parameterized by vectors or vectorizable tensors and can then be implemented in a differentiable programming framework.

Contributions. We introduce a generic change monitoring method called *autograd-test* and its online variant *autograd-test-CuSum* based on quantities amenable to be computed efficiently whenever the model is implemented in a differentiable programming framework. The method is equipped with a *scanning* procedure, allowing it to detect weak changes occurring on an unknown subset of model parameters. Theoretical results establish the consistency of the proposed test under the null hypothesis that the model has not changed, as well as under the fixed (local) alternative hypothesis of a fixed (shrinking) change. Finally, we illustrate the versatility of the approach on various machine learning models, ranging from time series models to text topic models.

2 Score-Based Change Detection

We first introduce the problem of change detection in parameters of a machine learning model. We then construct test statistics based on the score function that serve as building blocks for the proposed *autograd-test*. Finally, we demonstrate how *autograd-test* is well-adapted to machine learning models designed within a differentiable programming framework.

Change detection and hypothesis testing. Let $\mathcal{W}$ be an input space, and $W_{1:n} := \{W_k\}_{k=1}^n \subset \mathcal{W}$ be a sequence of observations. Consider a family of machine learning models $\{\mathcal{M}_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ such that

$W_k = \mathcal{M}_\theta(W_{1:k-1}) + \varepsilon_k$ with the convention $\mathcal{M}_\theta(W_{1:0}) = 0$, where $\{\varepsilon_k\}_{k=1}^n$ are independent and identically distributed (*i.i.d.*) random noises with mean 0. To learn this model from data, we choose some loss function L and obtain model parameters by solving the problem[1]:

$$\hat{\theta}_n := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{k=1}^{n} L\big(W_k, \mathcal{M}_\theta(W_{1:k-1})\big) \ .$$

We call above objective function the *learning objective* and call $\hat{\theta}_n$ the *learner*. This encompasses constrained empirical risk minimization (ERM) and constrained maximum likelihood estimation (MLE). For simplicity, we illustrate the idea of change detection in model parameters on a *correctly specified* model, *i.e.,* there exists a true value $\theta_0 \in \Theta$ from which the data are generated. Under abnormal circumstances, this value may not remain the same for all observations. Hence, we consider the same model but with a potential parameter change:

$$W_k = \mathcal{M}_{\theta_k}(W_{1:k-1}) + \varepsilon_k \ .$$

A time point $\tau \in [n-1] := \{1, \dots, n-1\}$ is called a *changepoint* if there exists $\Delta \neq 0$ such that $\theta_k = \theta_0$ for $k \leq \tau$ and $\theta_k = \theta_0 + \Delta$ for $k > \tau$. We aim to determine if there exists a changepoint in this sequence, which we formalize as a hypothesis testing problem.

(P0) Testing the existence of a changepoint:

$$\mathbf{H}_0 : \theta_k = \theta_0 \text{ for all } k = 1, \dots, n$$
$$\mathbf{H}_1 : \text{after some time } \tau,\ \theta_k \text{ jumps from } \theta_0 \text{ to } \theta_0 + \Delta.$$

In this paper, we focus on machine learning models whose loss function has a probabilistic interpretation, *i.e.,* $L(W_k, \mathcal{M}_\theta(W_{1:k-1}))$ can be rewritten as $-\log p_\theta(W_k|W_{1:k-1})$ for some conditional probability density $p_\theta(\cdot|W_{1:k-1})$ up to multiplicative and additive constants. We refer to such a loss function as a *probabilistic loss*. For instance, the loss function $L(W_k, \mathcal{M}_\theta(W_{1:k-1})) = (W_k - \theta W_{k-1})^2$ is associated with the conditional probability $W_k|W_{1:k-1} \sim \mathcal{N}(\theta W_{k-1}, 1)$, leading to an autoregressive model $W_k = \theta W_{k-1} + \varepsilon_k$ with $\{\varepsilon_k\}_{k=1}^n$ being *i.i.d.* standard normals; for more examples, see *e.g.* [Murphy, 2012]. For such models, the learning objective becomes

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \sum_{k=1}^{n} -\log p_\theta(W_k|W_{1:k-1}) \ .$$

In the following, we will use this probabilistic formulation.

Remark. Discriminative models also fit into this framework. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be n pairs of *i.i.d.* observations, then the learning objective is given by

$$\hat{\theta}_n := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{k=1}^{n} L(Y_k, \mathcal{M}_\theta(X_k)) \ .$$

If further L is a probabilistic loss, then the associated conditional probability density is $p_\theta(Y_k|X_k)$.

Likelihood score and score-based testing. Let $\mathbb{1}\{\cdot\}$ be the indicator function. Given any $\tau \in [n-1]$ and $1 \leq s \leq t \leq n$, we define the conditional log-likelihood under the alternative as

$$\ell_{s:t}(\theta, \Delta; \tau) := \sum_{k=s}^{t} \log p_{\theta+\Delta\mathbb{1}\{k>\tau\}}(W_k|W_{1:k-1}) \ .$$

And we denote by $\ell_{s:t}(\theta) := \ell_{s:t}(\theta, 0; n)$ the conditional log-likelihood under the null. We write $\ell_{s:t}(\theta, \Delta) := \ell_{s:t}(\theta, \Delta; \tau)$ and $\ell_n(\theta) := \ell_{1:n}(\theta)$ for short. Under the null hypothesis, the *score function w.r.t.* θ is defined as $S_{s:t}(\theta) := \nabla_\theta \ell_{s:t}(\theta)$, and the *observed Fisher information w.r.t.* θ is denoted by $\mathcal{I}_{s:t}(\theta) := -\nabla^2_\theta \ell_{s:t}(\theta)$.

[1] For simplicity, we assume here the minimizer exists and is unique.

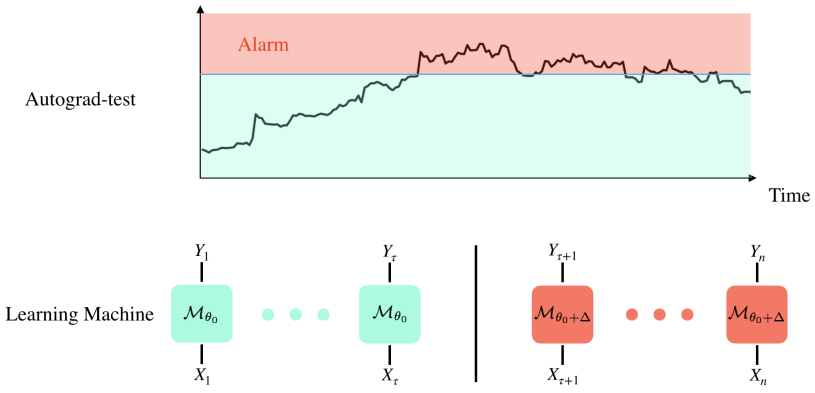

Figure 1: Illustration of monitoring a learning machine.

Given a hypothesis testing problem, the first step is to propose a *test statistic* R_n, *i.e.,* a function of observations $W_{1:n}$, such that the larger R_n is, the less likely the null hypothesis is true. Then, for a prescribed *significance level* (or *level*) $\alpha \in (0,1)$, we calibrate this test statistic by determining a threshold $r_0 := r_0(\alpha)$, leading to a test or decision rule[2] $\mathbb{1}\{R_n > r_0\}$. With this test, the *false alarm rate* or *type I error*, *i.e.,* the conditional probability of rejecting the null given that it is true, should be asymptotically controlled by α, *i.e.,*

$$\limsup_{n\to\infty} \mathbb{P}(R_n > r_0 | \mathbf{H}_0) \leq \alpha \ .$$

We refer to such a test as being *consistent in level*. Moreover, we want the *detection power* or *power*, *i.e.,* the conditional probability of rejecting the null given that it is false, to converge to 1 as n goes to infinity. And we say such a test as being *consistent in power*.

Let us follow this procedure to develop a test for Problem (P0). We start from the case when the true parameter θ_0 is known and the changepoint τ is fixed. One choice of statistics is the *score statistic* given by $\nabla_\Delta \ell_n(\theta_0, \Delta)|_{\Delta=0} = S_{\tau+1:n}(\theta_0)$. Under the null and standard conditions detailed in Appendix C, it is asymptotically normal with mean 0 and covariance given by the so-called *Fisher information* $\mathcal{I}_0$. In other words, $S_{\tau+1:n}(\theta_0)$ concentrates around its mean 0 if the null is true. Moreover, the observed Fisher information *w.r.t.* Δ, $-\nabla^2_\Delta \ell_n(\theta_0, \Delta)|_{\Delta=0} = \mathcal{I}_{\tau+1:n}(\theta_0)$, satisfies[3]

$$(n-\tau)^{-1}\mathcal{I}_{\tau+1:n}(\theta_0) \to_p \mathcal{I}_0, \text{ as } n \to \infty \ .$$

Since $\mathcal{I}_0$ is usually unknown, we estimate it with the observed Fisher information and consider the normalized score statistic

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

which is asymptotically χ^2_d-distributed and thus can be calibrated using quantiles of χ^2_d.

When θ_0 is unknown, we estimate it from the data by optimizing the learning objective, and the test statistic becomes

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

where, compared to (1), we replace θ_0 with the estimator $\hat{\theta}_n$, and replace $\mathcal{I}_{\tau+1:n}(\theta_0)$ by a different normalizing term $\mathcal{I}_n(\hat{\theta}_n;\tau)$. A standard choice for $\mathcal{I}_n(\hat{\theta}_n;\tau)$ is the *partial observed information w.r.t.* Δ [Wakefield, 2013, Chapter 2.9]:

$$\mathcal{I}_n(\hat{\theta}_n;\tau) := \mathcal{I}_{\tau+1:n}(\hat{\theta}_n) - \mathcal{I}_{\tau+1:n}(\hat{\theta}_n)^\top \mathcal{I}_{1:n}(\hat{\theta}_n)^{-1}\mathcal{I}_{\tau+1:n}(\hat{\theta}_n) \ . \tag{3}$$

[2]It means we reject the null if $R_n > r_0$.

[3]Here $\to_p$ represents convergence in probability, and we will omit "as $n \to \infty$" if there is no confusion.

Algorithm 1 Autograd-test

1: **Input:** data $(W_i)_{i=1}^n$, log-likelihood ℓ, learner $\hat{\theta}_n$, levels α_l and α_s, and maximum cardinality P.
2: **for** $\tau = 1$ **to** $n-1$ **do**
3: Compute $R_n(\tau)$ in (2) using AutoDiff.
4: Compute $R_n(\tau, P; \alpha)$ in (6).
5: **end for**
6: **Output:** $\psi(\alpha) = \max\{\psi_{\text{lin}}(\alpha_l), \psi_{\text{scan}}(\alpha_s)\}$.

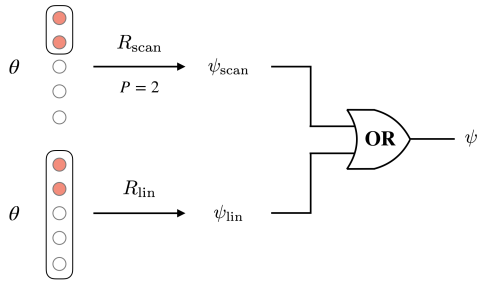

Figure 2: Illustration of *autograd-test* decision process.

To adapt to an unknown changepoint τ in Problem (P0), a natural test statistic is

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

Consequently, given a significance level α, the decision rule reads

$$\psi_{\text{lin}}(\alpha) := \mathbb{1}\{R_{\text{lin}} > H_{\text{lin}}(\alpha)\} \ , \tag{5}$$

where $H_{\text{lin}}(\alpha)$ is a pre-determined threshold for $\psi_{\text{lin}}(\alpha)$ to be consistent in level. We will discuss it in Sec. 3. We refer to R_n as the *linear statistic* and ψ_{lin} as the *linear test*.

Sparse alternatives. There are scenarios where the change may only happen in a small subset of components of θ_0, in other words, the change is weak. In such cases, the linear test, which is built assuming that the change is strong, may fail to detect weak changes. Therefore, we consider *sparse alternatives*, that is, only a small subset of components changes, and we call them *changed components*.

(P1) Testing the existence of a weak changepoint:

$\mathbf{H}_0 : \theta_k = \theta_0$ for all $k = 1, \ldots, n$

$\mathbf{H}_1$: after some time τ, θ_k jumps from θ_0 to $\theta_0 + \Delta$, where Δ has at most P nonzero entries.

Here P is referred to as *maximum cardinality*, which is set to be much smaller than d. We denote by T the changed components, in other words, Δ_{T^c}, the sub-vector of Δ indexed by $T^c := [d] \backslash T$, is equal to zero.

We now discuss the extension of the score-based statistic to sparse alternatives with unknown θ_0. For any fixed changepoint τ and changed components T, the coordinates indexed by T^c in the score function $\nabla_\Delta \ell_n(\theta, \Delta)$ is exactly 0 as $\Delta_{T^c} \equiv 0$. This is true for the observed Fisher information as well. Hence, for quantities in the statistic $R_n(\tau)$ given in (2), we may discard the coordinates indexed by T^c, giving a truncated statistic

$$R_n(\tau, T) = S_{\tau+1:n}^\top(\hat{\theta}_n)_T \left[\mathcal{I}_n(\hat{\theta}_n; \tau)_{T,T}\right]^{-1} S_{\tau+1:n}(\hat{\theta}_n)_T,$$

where we use the notation $M_{T,T}$ for the sub-matrix of M indexed by (T, T). Let $\mathcal{T}_p$ be the collection of all subsets of size p of $[d]$. To adapt to unknown T, we use

$$R_n(\tau, P; \alpha) := \max_{p \in [P]} \max_{T \in \mathcal{T}_p} H_p(\alpha)^{-1} R_n(\tau, T) \ , \tag{6}$$

where we use a different threshold $H_p(\alpha)$ for each $p \in [P]$ due to the fact that the asymptotic distribution of $R_n(\tau, T)$ depends on $|T|$; see Sec. 3. Finally, since τ is also unknown in (P1), we propose

$$R_{\text{scan}}(\alpha) := \max_{\tau \in [n-1]} R_n(\tau, P; \alpha) \ , \tag{7}$$

with decision rule

$$\psi_{\text{scan}}(\alpha) := \mathbb{1}\{R_{\text{scan}}(\alpha) > 1\} \ . \tag{8}$$

We call $R_{\text{scan}}(\alpha)$ the *scan statistic* and call ψ_{scan} the *scan test*.

To incorporate the strengths of these two tests, we consider a combination of them,

$$\psi(\alpha) := \max\{\psi_{\text{lin}}(\alpha_l), \psi_{\text{scan}}(\alpha_s)\} \tag{9}$$

with $\alpha_l + \alpha_s = \alpha$, and we refer to it as *autograd-test*. The choice of α_l and α_s should be based on prior knowledge regarding how likely the change is weak (*i.e.,* they should be equal without any prior information). We summarize the decision process of the *autograd-test* in Fig. 2.

The algorithm to compute the *autograd-test* is presented in Alg. 1. This algorithm is built upon the automatic differentiation (AutoDiff) procedure, which is explained in detail in the following paragraph. Note that directly computing the scan statistic may be exponentially expensive in the parameter dimension d, since it involves a maximization over all subsets of $[d]$ with cardinality $p \leq P$. Alternatively, we approximate the maximizer of $\max_{T\in\mathcal{T}_p} R_n(\tau, T)$, say T_p, by the indices of the largest p components in

$$v(\tau) := \text{diag}(\mathcal{I}_n(\hat{\theta}_n; \tau))^{-1} S^{\odot 2}_{\tau+1:n}(\hat{\theta}_n) \ , \tag{10}$$

where we denote by $S^{\odot 2}$ the element-wise square for a vector S. That is, we consider all T with $|T| = 1$, and approximate the maximizer T_p by the union of the ones that give the largest p values of $R_n(\tau, T)$. This approximation is accurate if the difference between the largest eigenvalue and the smallest eigenvalue of $\mathcal{I}_n(\hat{\theta}_n; \tau)$ is small compared to $\|S_{\tau+1:n}(\hat{\theta}_n)\|^2$. In Step 4, we first sort $v(\tau)$, then for each $p \leq P$, we obtain from $v(\tau)$ an approximate maximizer $\widetilde{T}_p$ and compute $R_n(\tau, P; \alpha_s)$ by $\max_{p\in[P]} H_p(\alpha_s)^{-1} R_n(\tau, \widetilde{T}_p)$.

With some algebra one can show that our score-based statistics recover the generalized likelihood ratio (GLR) type statistics proposed in [Enikeeva and Harchaoui, 2019] as a special case in Gaussian change in mean models. We refer readers to Appendix E for details.

Score statistics, differentiable programming, and component screening. An attractive feature of score-based statistics in the age of differentiable programming is their straightforward computation using automatic differentiation. While GLR-type statistics require computing various constrained optimization problems, the *autograd-test* only involves (second) derivatives of the log-likelihood function. That opens up the opportunity to build a library for *autograd-test* based on automatic differentiation, which is applicable to any machine learning models with a probabilistic loss designed within a differentiable programming framework.

To be more specific, let $\mathcal{M}_\theta$ be such a machine learning model whose log-likelihood evaluated at $\hat{\theta}_n$, $\ell_n(\hat{\theta}_n)$, is implemented as a computational graph $G = (V, E)$, where each vertex $v \in V$ stands for a scalar variable and each edge $(v_i, v_j) \in E$ is directed and represents an operation. For instance, for the computation $x_3 = x_1 + x_2$, its computational graph consists of three vertices $x_{1:3}$ and two edges (x_1, x_3) and (x_2, x_3) where each edge represents the add operation. The AutoDiff procedure is able to compute the score $S_{1:n}(\hat{\theta}_n) = \nabla_\theta \ell_n(\hat{\theta}_n)$ within $\mathcal{O}(|V| + |E|)$ time. For convenience, we assume that $|V| + |E| = \mathcal{O}(nd)$, which is usually the case in practice.

Let us analyze the computational complexity of Alg. 1. The most time-expensive term in Step 3 is $\mathcal{I}_n(\hat{\theta}_n; \tau)^{-1}$. It takes $\mathcal{O}(\tau d^2 + d^3)$ time, where $\mathcal{O}(\tau d^2)$ comes from computing $\mathcal{I}_{1:\tau}(\hat{\theta}_n)$ using AutoDiff and $\mathcal{O}(d^3)$ accounts for inverting $\mathcal{I}_n(\hat{\theta}_n; \tau)$. For Step 4, sorting has a time complexity of $\mathcal{O}(d \log d)$. Based on the quantities computed in former steps, it takes $\mathcal{O}(P^4)$ time to calculate $R_n(\tau, T_p)$ for all $p \in [P]$ since inverting $\mathcal{I}_n(\hat{\theta}_n; \tau)_{T_p, T_p}$ costs $\mathcal{O}(p^3)$ time. To summarize, the overall computational complexity of Alg. 1 is $\mathcal{O}(n^2 d^2 + nd^3 + nP^4)$. As for the space complexity, the main consumptions come from storing the computational graph and the Hessian, taking space $\mathcal{O}(|V| + |E|) = \mathcal{O}(nd)$ and $\mathcal{O}(d^2)$, respectively.

Remark. In practice, the maximum cardinality P is usually set to be $\lfloor\sqrt{d}\rfloor$, then the overall time complexity becomes $\mathcal{O}(n^2 d^2 + nd^3)$. Moreover, as demonstrated by the following examples, if observations are independent, or the log-likelihood function admits a simple recursion, we can reduce time complexities of calculating $S_{1:\tau}(\hat{\theta}_n)$ and $\mathcal{I}_{1:\tau}(\hat{\theta}_n)$ for all $\tau \in [n-1]$ from $\mathcal{O}(n^2 d)$ and $\mathcal{O}(n^2 d^2)$ to $\mathcal{O}(nd)$ and $\mathcal{O}(nd^2)$, respectively. This gives an algorithm of complexity $\mathcal{O}(nd^3)$.

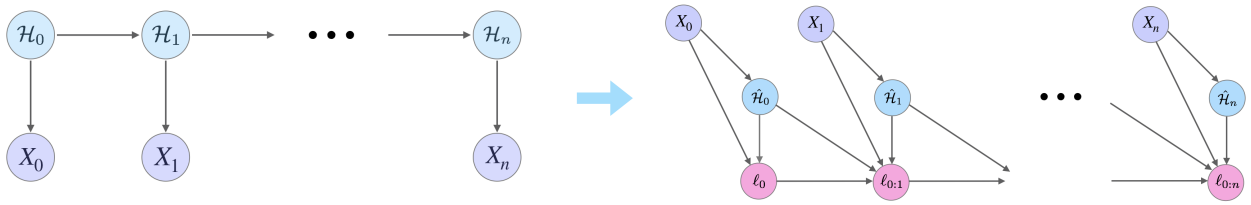

Figure 3: Graphical representations of the text topic model (left: probabilistic graphical model; right: computational graph).

Example 1 (Additive model)**.** *As a concrete example, we consider a linear regression model with standard normal errors, the log-likelihood function of observations* $\{(X_k, Y_k)\}_{k\in[n]}$ *reads:*

$$\ell_{1:n}(\theta) = -\frac{1}{2}\sum_{k=1}^{n}(Y_k - X_k^\top\theta)^2 + C,$$

where C *is some constant. Thanks to the independence, it suffices to consider the log-likelihood function of a single observation* (X, Y)*:*

$$\ell(\theta) = -\frac{1}{2}(Y - X^\top\theta)^2 + C.$$

The computational graph of $\ell(\theta)$ *can be constructed in the following way: 1) let* $\{\theta_i\}_{i\in[d]}$*,* $\{x_i\}_{i\in[d]}$*, and* $X^\top\theta$ *be* $2d+1$ *vertices with* $2d$ *edges* $\{(\theta_i, X^\top\theta), (x_i, X^\top\theta)\}_{i\in[d]}$*; 2) let* $Y, Y - X^\top\theta$*, and* $\ell(\theta)$ *be* 3 *vertices with* 3 *edges* $(X^\top\theta, Y - X^\top\theta), (Y, Y - X^\top\theta)$*, and* $(Y - X^\top\theta, \ell(\theta))$*. Thus, computing a single score* $\nabla_\theta\ell(\hat{\theta}_n)$ *takes time* $\mathcal{O}(d)$*, and the calculation of* $\{S_{1:\tau}(\hat{\theta}_n)\}_{\tau\in[n-1]}$ *has complexity* $\mathcal{O}(nd)$ *due to the recursion* $S_{1:\tau}(\hat{\theta}_n) = S_{1:(\tau-1)}(\hat{\theta}_n) + \nabla_\theta\ell_\tau(\hat{\theta}_n)$*. Similarly, computing* $\mathcal{I}_{1:\tau}(\hat{\theta}_n)$ *for all* $\tau \in [n-1]$ *costs* $\mathcal{O}(nd^2)$ *time.*

Example 2 (Time series model)**.** *Consider an autoregressive moving-average (ARMA) model:*

$$X_t = \sum_{i=1}^{r}\phi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q}\varphi_i\varepsilon_{t-i} \ ,$$

where $\{\varepsilon_t\}$ *are i.i.d. standard norm random variables. Let* $\theta = (\phi; \varphi)$*. Assume that* $r \geq q$ *and* $X_{1:r}$ *is completely known. Then the log-likelihood reads:*

$$\ell_n(\theta) = -\frac{1}{2}\sum_{t=r+1}^{n}\varepsilon_t^2 + C \ ,$$

where $\varepsilon_t = X_t - \sum_{i=1}^{r}\phi_i X_{t-i} - \sum_{i=1}^{q}\varphi_i\varepsilon_{t-i}$*. It is straightforward to construct its computational graph with* $|V| + |E| = \mathcal{O}(nd)$ *in which* $d = r + q$*, and thus computing* $S_{1:n}(\hat{\theta}_n)$ *costs* $\mathcal{O}(nd)$ *time. In order to compute* $S_{1:\tau}(\hat{\theta}_n)$ *for all* $\tau \in [n-1]$*, a direct application of AutoDiff would cost* $\mathcal{O}(\sum_{\tau=1}^{n}\tau d) = \mathcal{O}(n^2 d)$ *time. Nevertheless, by exploiting the recursion*

$$S_{1:\tau}(\hat{\theta}_n) = S_{1:(\tau-1)}(\hat{\theta}_n) + \varepsilon_\tau\Big[\nabla_\theta\varepsilon_\tau(\theta, \varepsilon_{(\tau-q):(\tau-1)}) + \sum_{i=1}^{q}(-\varphi_i)\nabla_\theta\varepsilon_{\tau-i}\Big],$$

we can reduce this time complexity to $\mathcal{O}(nd)$ *at the cost of extra* $\mathcal{O}(nd)$ *space—compute once and store* $\{\nabla_\theta\varepsilon_t\}$ *for reuse. A similar argument holds for the computation of information matrices* $\{\mathcal{I}_{1:\tau}(\hat{\theta}_n)\}_{\tau\in[n-1]}$*. Therefore, even though the observations are not independent, we are able to reduce the time complexity from* $\mathcal{O}(n^2d^2 + nd^3)$ *to* $\mathcal{O}(nd^3)$ *at the expense of space complexity, from* $\mathcal{O}(nd + d^2)$ *to* $\mathcal{O}(nd^2)$*.*

Example 3 (Text topic model)**.** *The text topic model introduced in [Stratos et al., 2015] is a hidden Markov model with transition probability q and emission probability g, supported respectively on finite sets $[N]$ and $[M]$. Moreover, it satisfies the so-called Brown assumption: for each observation $X \in [M]$, there exists a unique hidden state $\mathcal{H}(X) \in [N]$ such that $g(X|\mathcal{H}(X)) > 0$ and $g(X|h) = 0$ for all $h \neq \mathcal{H}(X)$; see Fig. 3 for a graphical representation. The authors proposed a class of spectral methods to recover approximately the map $\hat{\mathcal{H}}$ up to permutation. Let θ be the parameter vector consisting of free variables in $\{q(i|j)\}_{i,j\in[N]}$ and $\{g(k|\hat{\mathcal{H}}(k))\}_{k\in[M]}$, e.g. $q(N|1) = 1 - \sum_{i=1}^{N-1} q(i|1)$ is viewed as a function of $\{q(i|1)\}_{i\in[N-1]}$ rather than a parameter. Thus, $\theta \in \mathbb{R}^d$ with $d := N^2 - N + M - N$. Denote $\hat{\mathcal{H}}_k := \hat{\mathcal{H}}(X_k)$, then we may estimate q and g by maximizing the log-likelihood*

$$\ell_n(\theta) = \sum_{k=1}^{n} \log q(\hat{\mathcal{H}}_k|\hat{\mathcal{H}}_{k-1}) + \log g(X_k|\hat{\mathcal{H}}_k) \ ,$$

where X_0 is assumed to be known. The computational graph of $\ell_n(\theta)$ is of the size $|V| + |E| = \mathcal{O}(d + n)$, so computing $\{S_{1:\tau}(\hat{\theta}_n)\}_{\tau\in[n-1]}$ directly takes $\mathcal{O}(nd + n^2)$ time. Again, $S_{1:\tau}(\hat{\theta}_n)$ admits a recursion

$$S_{1:\tau}(\hat{\theta}_n) = S_{1:(\tau-1)}(\hat{\theta}_n) + \nabla_\theta[\log q(\hat{\mathcal{H}}_\tau|\hat{\mathcal{H}}_{\tau-1}) + \log g(X_\tau|\hat{\mathcal{H}}_\tau)],$$

and thus the running time of computing $\{S_{1:\tau}(\hat{\theta}_n)\}_{\tau\in[n]}$ can be reduced to $\mathcal{O}(nd)$. Similarly, computing $\mathcal{I}_{1:\tau}(\hat{\theta}_n)$ for all $\tau \in [n-1]$ costs $\mathcal{O}(nd^2)$ time.

There are scenarios, especially for latent variable models, where evaluating the likelihood function is expensive. In such cases our algorithm would be intractable, unless there is an efficient way to compute gradients. For hidden Markov models, we invoke two useful identities to rewrite the score function and observed Fisher information in an additive form, and then implement the smoothing procedure [Cappé et al., 2005] to compute them recursively. Details can be found in Appendix B.

Another attractive feature of score-based statistics is their flexibility to screen individual components or groups of components, which allows us to neglect irrelevant information in the data and focus on catching specific signals. Specifically, if only the parameters indexed by a set T are of interest, we may easily incorporate this prior knowledge by restricting score-based statistics in components indexed by T, just like the truncated statistic for sparse alternatives.

Online extension. In some scenarios we may not have access to the full dataset, so we also develop an online variant of the *autograd-test*, called *autograd-test-CuSum*. We take the repeated significance tests viewpoint [Ghosh and Sen, 1991, Chapter 7] to control the overall false alarm rate, and we propose a correction analogous to CuSum [Page, 1954] for power considerations. We refer readers to Appendix A for a thorough discussion.

3 Level and Power

In this section we summarize the asymptotic behavior of the proposed score-based statistics under null and alternatives when the model is correctly specified. We derive appropriate thresholds for these tests so that they are consistent both in level and power; for precise statements and proofs see Appendix C.

Proposition 1 (Null hypothesis, no change)**.** *Under the null hypothesis, and suppose that (a) the observed information $\mathcal{I}_{1:n}(\theta_0) \to_p \mathcal{I}_0$, (b) $\hat{\theta}_n$ exists and converges in distribution to a normal distribution, and (c) the score $S_{1:n}(\theta_0)$ satisfies the Lindeberg conditions for martingales. Then, for any $T \subset [d]$ and $\tau_n \in \mathbb{Z}_+$ such that $\tau_n/n \to \lambda \in (0, 1)$, we have, provided smoothness and regularity conditions on the log-likelihood ℓ_n,*

$$R_n(\tau_n) \to_d \chi^2_d \quad \text{and} \quad R_n(\tau_n, T) \to_d \chi^2_{|T|},$$

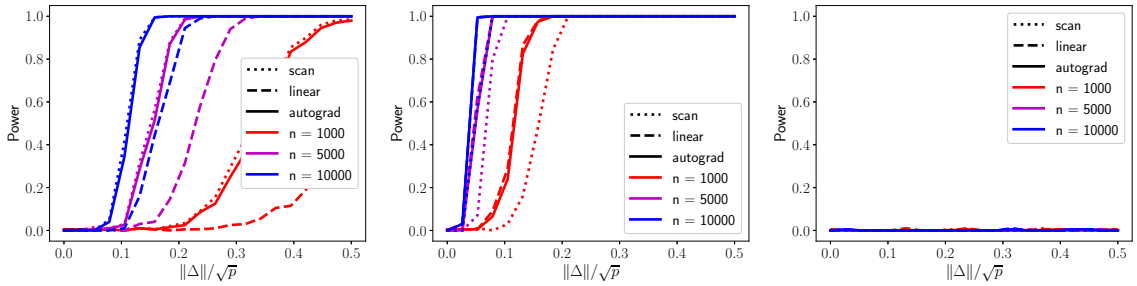

Figure 4: Power versus magnitude of change for linear regression (left: $p = 1$; middle: $p = 20$; right: $p = 1$ with restriction excluding the changed component).

where we use $\to_d$ *for convergence in distribution and* χ^2_d *for a chi-square distribution with* d *degrees of freedom. In particular, with thresholds*[4] $H_{lin}(\alpha) = q_{\chi^2_d}(\alpha/n)$ *and* $H_p(\alpha) = q_{\chi^2_p}\big(\alpha/[\binom{d}{p}n(p+1)^2]\big)$*, the three proposed tests* $\psi(\alpha), \psi_{lin}(\alpha), \psi_{scan}(\alpha)$ *are consistent in level.*

Most conditions in Prop. 1 are standard for $\hat{\theta}_n$ to be asymptotically normal. The conditions on the score are required by the Lindeberg theorem for martingales [Van der Vaart, 2013, Chapter 4.5] to prove the asymptotic normality of the score. In fact, under suitable regularity conditions, they hold true in *i.i.d.* models, hidden Markov models [Bickel et al., 1998, Chapter 12], and stationary autoregressive moving-average models [Douc et al., 2014].

Note that it is also possible to obtain non-asymptotic results for specific models; see Appendix D for an example. Yet, a non-asymptotic result as general as Prop. 1 would be a rather challenging problem, as it would require a *generic* sharp concentration bound for maximum likelihood estimators and empirical risk minimizers, including latent-variable models for instance. This is still in progress and we leave it for future work. Moreover, non-asymptotic results for hypothesis testing may not always give practical insights to calibrate the test, whereas the limit behavior we verified can offer straightforward and often practical calibration, as demonstrated in Sec. 4.

The next proposition verifies the consistency in power of the proposed tests under fixed alternatives by assuming independence of the data.

Proposition 2 (Fixed alternative hypothesis, change)**.** *Assume that observations are independent, and the alternative hypothesis is true with a fixed change parameter* Δ *and changepoint* τ_n *such that* $\tau_n/n \to \lambda \in (0,1)$*. Let* $\theta_1 = \theta_0 + \Delta$*. Suppose*[5] $\lambda D_{\mathrm{KL}}(p_{\theta_0}\|p_\theta) + \overline{\lambda} D_{\mathrm{KL}}(p_{\theta_1}\|p_\theta)$ *(with* $\overline{\lambda} := 1 - \lambda$*) has a unique minimizer* $\theta^* \in \mathrm{int}(\Theta)$*, and* $\mathbb{E}_{\theta_j}[-\nabla^2_\theta \ell(\theta^*)]$ *is positive definite for* $j \in \{0,1\}$*. Then* $\hat{\theta}_n \to_p \theta^*$ *and the three proposed tests* $\psi(\alpha), \psi_{lin}(\alpha), \psi_{scan}(\alpha)$ *are consistent in power, provided smoothness and regularity conditions on the log-likelihood* ℓ_n*.*

The conditions in Prop. 2 are counterparts of standard assumptions for analyzing the asymptotic behavior of standard score statistics under fixed alternatives. The condition regarding the Kullback-Leibler divergence is needed for $\hat{\theta}_n$ to have a valid limit.

We also consider a more challenging case, called *local alternatives*, where the change parameter Δ shrinks to zero as the sample size increases. In such scenarios, the detection power of the linear test is characterized by the asymptotic distribution of its test statistic under local alternatives, which we summarize in the following proposition; see Appendix C for a similar result for the scan statistic.

Proposition 3 (Local alternative hypothesis, subtle change)**.** *Assume that observations are independent, and the alternative hypothesis is true with change parameter* $\Delta_n = hn^{-1/2}$ *in which* $h \neq 0$ *and changepoint* τ_n *such that* $\tau_n/n \to \lambda \in (0,1)$*. Suppose that* θ_0 *is the unique maximizer of* $\mathbb{E}_{\theta_0}[\ell(\theta)]$ *and* $\mathbb{E}_{\theta_0}[-\nabla^2_\theta \ell(\theta_0)]$ *is positive definite. Then the score-based statistic* $R_n(\tau_n)$ *converges in distribution to a non-central chi-square*

[4] We use $q_D(\alpha)$ for the upper α-quantile of the distribution D.
[5] We use the notation D_{KL} for the Kullback-Leibler divergence.

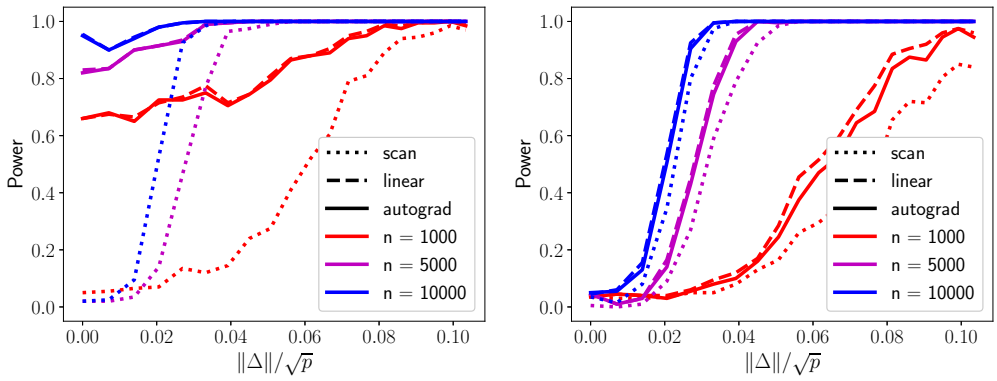

Figure 5: Power versus magnitude of change for ARMA$(3, 2)$ (left: without restriction; right: with restriction).

distribution with degrees of freedom d and parameter $\lambda\overline{\lambda}h^\top\mathcal{I}_0 h$, provided smoothness and regularity conditions on the log-likelihood ℓ_n.

In the case of maximum likelihood estimators (MLE) for instance, even though these propositions are proved for the MLE, one can show that, with minor modifications to the proofs, they actually apply to any estimators whose difference with the MLE decays faster than $n^{-1/2}$.

4 Experiments

In this section, we perform simulations on synthetic data for an additive model, a time series model, a hidden Markov model, and a text topic model discussed in Sec. 2. We empirically verify the consistency of proposed tests and compare their performance under a correctly specified model. We apply our approach to detect changes in the text topic model on TV shows subtitles. We summarize our experimental settings and findings here.[6].

Synthetic experimental settings. For each model, we generate the first half of the sample from this 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, and generate the second half from the same model with parameter θ_1. Next, we run the linear test, the scan test, and the *autograd-test* to monitor the process of learning parameters, where the significance levels are set to be $\alpha = 2\alpha_l = 2\alpha_s = 0.05$ and the maximum cardinality P is chosen as $\lfloor\sqrt{d}\rfloor$. We repeat this procedure 200 times and approximate the detection power by rejection frequency. Finally, we plot the power (rejection frequency) curve by varying values of δ, where we use three different types of lines to represent three tests, and different colors for different sample sizes. Note that the value at $\delta = 0$ is the empirical estimate of the false alarm rate.

Additive model. We consider a linear regression model with 101 parameters (including intercept) and L_2 loss, and investigate two sparsity levels, $p = 1$ and $p = 20$. The coefficients and intercept are fixed to be zero before change. All the entries of the design matrix and error terms are generated independently from a standard normal distribution. As shown in the left and middle plots in Fig. 4, 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, with the scan test tending to perform poorer than the other two. This empirically illustrates that (1) the scan test works

[6]More details and additional results, including the ones for *autograd-test-CuSum*, are deferred to Appendix F.

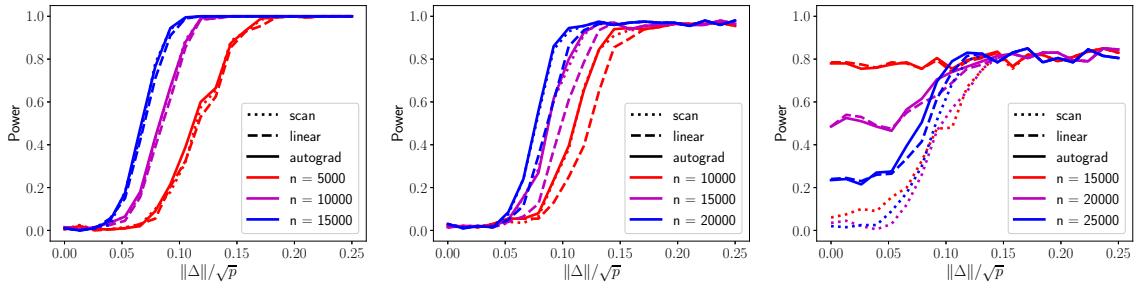

Figure 6: Power versus magnitude of change for HMMs with N hidden states (left: $N = 3$; middle: $N = 7$; right: $N = 15$).

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 examine the component screening feature of these score-based tests. We consider the same linear regression model except that we only screen 50 components of parameters (*i.e.,* regard the rest 51 as nuisance parameters). 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. 4. Hence, with the restriction imposed, the decision of these tests is unlikely to be affected by the change in nuisance parameters.

Time series model. We then investigate an autoregressive moving-average (ARMA) models—ARMA$(3, 2)$. Before the change, all the coefficients are generated in a way to ensure the time series to be stationary. The change only occurs at the AR coefficients, *i.e.,* $p = 3$, with the restriction that the post-change AR coefficients also satisfy the stationarity condition. As shown in the left plot of Fig. 5, the scan test works fairly well for this model. However, the other two have extremely high false alarm rate. This problem gets more severe as n increases, and hence is not due to lack of samples. It turns out that this 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 partial information (3) and subsequent unstable computation of the linear statistic. On the contrary, the scan statistic only inverts submatrices of size approximately $\sqrt{d} \times \sqrt{d}$ with reduced condition number (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 screening of these tests in AR coefficients, which is the case in the right plot of Fig. 5, all tests are then consistent in level.

Hidden Markov model. Next, we exam HMMs with $N \in \{3, 7, 15\}$ hidden states and normal emission distribution. Each row in its transition matrix is generated by adding a nonzero constant vector to a Dirichlet random variable so that it sums to one and all entries are positive. Given the current state $k \in \{0, \ldots, N-1\}$, the emission distribution has mean k and standard deviation $0.01+0.09k/(N-1)$. The post-change transition matrix is obtained by subtracting δ from the $(1, 1)$ entry and adding δ to the $(1, N)$ entry.

Results are shown in Fig. 6. When $N = 3$, three tests have almost identical performance. As N increases, the change becomes sparser, and subsequently, the scan test and the *autograd-test* outperform the linear test. When $N = 15$, the linear test and *autograd-test* become inconsistent in level, 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 empirical 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 examine the 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

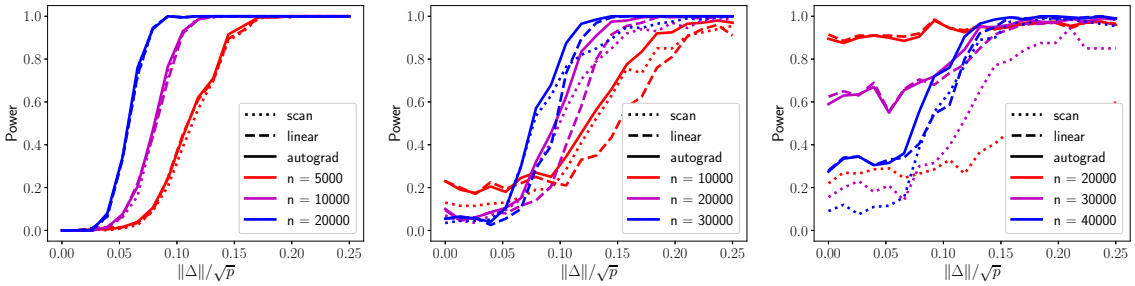

Figure 7: Power versus magnitude of change for text topic models (left: $(N, M) = (3, 6)$; middle: $(N, M) = (7, 20)$; right: $(N, M) = (15, 150)$).

emission distribution. We use the same way as HMMs to generate the transition matrix. The emission matrix is sampled analogously, with the constraint that each column can only have one nonzero entry. As shown in Fig. 7, the results are similar to the ones for HMMs.

Real data application. We collect subtitles of the first two seasons of four different TV shows—Friends (F), Modern Family (M), the Sopranos (S), and Deadwood (D)—where the former two are assumed to be "polite" and the latter two are "rude". For every pair of seasons, we concatenate them, and the task is to detect changes in rudeness level, while ignoring other alterations. After standard preprocessing steps such as tokenization and adding a special symbol for unknown tokens, we train the aforementioned text topic model with $N = \lfloor \sqrt{n/100} \rfloor$ and M being the size of vocabulary built from the training corpus. Then we apply only the scan test for detection due to the false alarm considerations discussed above.

As demonstrated in Table 1, the scan test does a perfect job in reporting shifts in rudeness level. However, the false alarm rate is relatively high. For ("polite", "polite") pairs, there are two false alarms and one NA because the estimated transition matrix contains zero which makes the statistic undefined; while for ("rude", "rude") pairs, 9 out of 16 are false alarms. 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 in Appendix F suggest a similar phenomenon.

We remark that rudeness is definitely not the only factor that contributes to the difference between two shows. But the results are promising in the sense that the scan statistic can neglect some low level discrepancies and focus on "global information" in language level. As we already discussed, we can utilize the component screening feature and restrict the detection to parameters related to the rudeness, and obtain a more appropriate test for this specific task.

5 CONCLUSION

We introduced a generic 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.

Table 1: Decision for pair-wise experiments (each (row, column) pair stands for a concatenation of two seasons of shows; "R" means reject and "N" means not reject).

	F1	F2	M1	M2	S1	S2	D1	D2
F1	N	N	N	N	R	R	R	R
F2	N	N	*R*	N	R	R	R	R
M1	N	*R*	N	N	R	R	R	R
M2	N	N	N	*NA*	R	R	R	R
S1	R	R	R	R	N	N	*R*	*R*
S2	R	R	R	R	N	N	*R*	*R*
D1	R	R	R	R	*R*	*R*	N	*R*
D2	R	R	R	R	*R*	*R*	N	N

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.

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*. Prentice Hall Information and System Sciences Series. Prentice Hall, Inc., Englewood Cliffs, NJ, 1993.

Peter J. Bickel, Ya'acov Ritov, and Tobias Rydén. 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.

Lucien Birgé. An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, 36:113–133, 2001.

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

Edward G Carlstein, Hans-Georg Müller, and David Siegmund. Change-point problems. IMS, 1994.

Miklós Csörgő and Lajos Horváth. *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1997.

John P. Cunningham, Zoubin Ghahramani, and Carl Edward Rasmussen. Gaussian processes for time-marked time-series data. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 255–263, 2012.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):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.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases*, pages 180–191, 2004.

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. *J. Roy. Statist. Soc. Ser. B*, 44(2):226–233, 1982.

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

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019.

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.

Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. Chapman and Hall/CRC, 2014.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Aad W Van der Vaart. Lecture notes in time series. Universiteit Leiden, 2013.

Denis Volkhonskiy, Evgeny Burnaev, Ilia Nouretdinov, Alexander Gammerman, and Vladimir Vovk. Inductive conformal martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications*, pages 132–153, 2017.

Jon Wakefield. *Bayesian and frequentist regression methods*. Springer Science & Business Media, 2013.

Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. Technology Press, 1948.

Algorithm 2 Autograd-test-CuSum

Input: data stream $(W_i)_{i=1}^n$, log-likelihood function ℓ, initial time m and MLE $\hat{\theta}$, threshold α.
Sample from M to estimate the quantile $q_M(\alpha)$.
Initialization: $t \leftarrow m$, $S_{\mathrm{f}} \leftarrow S_{\mathrm{c}} \leftarrow S_{1:m}(\hat{\theta})$, $\mathcal{I}_{\mathrm{f}} \leftarrow \mathcal{I}_{\mathrm{c}} \leftarrow \mathcal{I}_{1:m}(\hat{\theta})$, $R_{\min} \leftarrow S_{\mathrm{f}}^\top \mathcal{I}_{\mathrm{f}}^{-1} S_{\mathrm{f}}$.
while $t \leq n$ and $R_{\mathrm{c}} \leq nq_M(\alpha)/t$ **do**
  $t \leftarrow t+1$.
  $\hat{\theta} \leftarrow \hat{\theta} + \eta \nabla_\theta \ell_t(\hat{\theta})$.
  $S_z \leftarrow S_z + \nabla_\theta \ell_t(\hat{\theta})$ for $z \in \{\mathrm{f}, \mathrm{c}\}$.
  $\mathcal{I}_z \leftarrow \mathcal{I}_z + \nabla_\theta \ell_t(\hat{\theta}) \nabla_\theta \ell_t(\hat{\theta})^\top$ for $z \in \{\mathrm{f}, \mathrm{c}\}$.
  $R_{\mathrm{f}} \leftarrow S_{\mathrm{f}}^\top \left(\mathcal{I}_{\mathrm{f}}\right)^{-1} S_{\mathrm{f}}$.
  if $R_{\mathrm{f}} \leq R_{\min}$ **then**
    $S_{\mathrm{c}} \leftarrow 0$, $\mathcal{I}_{\mathrm{c}} \leftarrow 0$, $R_{\mathrm{c}} \leftarrow 0$.
    $R_{\min} \leftarrow R_{\mathrm{f}}$.
  else
    $R_{\mathrm{c}} \leftarrow S_{\mathrm{c}}^\top \left(\mathcal{I}_{\mathrm{c}}\right)^{-1} S_{\mathrm{c}}$.
  end if
end while
Output: t.

A Online extension

For computational consideration, we assume from now on that the data $W_{1:n}$ are *i.i.d.* Assume that we have a sample of size $m < \tau$ at the starting point, which is referred to as the *null sample*, allowing us to estimate the true parameter θ_0 with a reasonable accuracy. Consequently, we use $n^{-1}\sum_{t=1}^n \nabla_\theta \ell_t(\hat{\theta}_n) \nabla_\theta \ell_t(\hat{\theta}_n)^\top$ to estimate the Fisher information, since it is a consistent estimator and does not require the calculation of second derivatives.

We begin with a naive modification to the score statistic, which is $R_t = S_{1:t}^\top(\hat{\theta}_t)\mathcal{I}_{1:t}(\hat{\theta}_t)^{-1}S_{1:t}(\hat{\theta}_t)$, with stopping time $t_a = \min\{t : R_t > h_t\}$ where h_t is some fixed threshold. To control the false alarm rate of this approach, we take the repeated significance tests viewpoint [Ghosh and Sen, 1991, Chapter 7]. We first train the model using the null sample and compute $S_{1:m}(\hat{\theta}_m)$ and $\mathcal{I}_m(\hat{\theta}_m)$ as an initialization. Then, for each new observation, we update the score and information to compute the new test statistic, and compare it with the threshold until rejection. Now the overall false alarm rate, $\mathbb{P}_{\theta_0}[R_m > h_m(\alpha)] + \sum_{t=m+1}^n \mathbb{P}_{\theta_0}\left[R_j \leq h_j(\alpha), \forall j \in [1, t-1]; R_t > h_t(\alpha)\right]$, can be controlled within α by approximating the threshold[7] as $h_t(\alpha) = nq_M(\alpha)/t$, where M is defined by $M = \sup\{W(t)^\top W(t) : t \in [0,1]\}$ in which $W(t)$ 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.

One drawback of this naive statistic is that $\hat{\theta}_t$ needs to be updated whenever a new observation arrives, and hence $S_{1:t}$ and $\mathcal{I}_t$ 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 at least $\mathcal{O}(n^2)$ complexity. Hence, we propose the following approximate update procedure. As illustrated in Alg. 2, at each time t, a new data point W_t arrives, and the maximum likelihood estimator (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 $\mathcal{I}$ is similar, leading to a constant time (*w.r.t.* the sample size) procedure.

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 $W_{1:t}$ with $t > \tau$, the law of large numbers implies that the score function $S_{1:t}(\hat{\theta}_t)$ is close to $\tau\mathbb{E}_{\theta_0}[S_1(\theta_0)] + (t-\tau)\mathbb{E}_{\theta_1}[S_{\tau+1}(\theta_0)]$

[7] Recall that $q_M(\alpha)$ is the upper α-quantile of M.

since $\hat{\theta}_t \to_p \theta_0$. Note that $\mathbb{E}_{\theta_0}[S_1(\theta_0)] = 0$ and $\mathbb{E}_{\theta_1}[S_{\tau+1:n}(\theta_0)] \neq 0$, the latter has a dominating effect in the score function. While for the observed Fisher information, it accumulates as $\mathcal{I}_t(\hat{\theta}_t) \approx \tau\mathcal{I}_0 + (t-\tau)\mathcal{I}_1$ with both $\mathcal{I}_0$ and $\mathcal{I}_1$ being positive definite. Therefore, $\mathcal{I}_{1:t}(\hat{\theta}_t)$ tends to over estimate the variance of $S_{1:t}(\hat{\theta}_t)$ if the number of post-change observations is not large enough. To address that, we propose a correction procedure analogous to the one used in CuSum, *i.e.,* $\tilde{R}_t = R_t - \min_{j\in[t]} R_j$.

As shown in Alg. 2, we maintain three statistics—the *full statistic* R_{f}, the minimum of R_{f} at all time points until now, and the *corrected statistic* R_{c}. The full statistic is exactly the same as aforementioned naive statistic R_t. If at time t the current R_{f} is no larger than the current $R_{\min}$, we think there is no changepoint before time t and reinitialize the *corrected score* S_{c} and the *corrected information* $\mathcal{I}_{\mathrm{c}}$. In other words, we discard the score and information from $W_{1:t}$ if the full statistic at time t is too small, showing little evidence of t being a changepoint. Importantly, the stopping criteria is based on the corrected statistic R_{c} rather than R_{f}. We call this new statistic *autograd-test-CuSum*.

B Computation of score function and Fisher information for hidden Markov models

Let $\{X_k, Y_k\}_{k=0}^n$ be a bivariate discrete time process in which $\{X_k\}$ defined on $(\mathbf{X}, \mathcal{X})$ is a Markov chain with finite state space $\mathbf{X} = [M]$ and initial distribution ν, and, conditional on $\{X_k\}$, $\{Y_k\}$ defined on $(\mathbf{Y}, \mathcal{Y})$ is a sequence of independent random variables such that the conditional distribution of Y_k only depends on X_k. Denote $Q = (q_1, \ldots, q_M)$ the transition matrix, *i.e.,* , $q_{ij} = \mathbb{P}(X_k = j | X_{k-1} = i)$ for $i, j \in [M]$ and $k \in [n]$. Denote G_β the emission distribution and for simplicity we assume it is absolutely continuous *w.r.t.* some measure μ, *i.e.,* , $g(x_k, y_k) := g_\beta(x_k, y_k)$ is the pdf of the conditional distribution Y_k given X_k parametrized by β.

Notation. Since $(y_{1:n})$ are observed and fixed, we will omit the dependency on y in all quantities. For instance, we write $g_k(x_k)$ instead of $g(x_k, y_k)$. For positive indices s, t, and k with $s \leq t \leq k$, we denote by $\phi_{s:t|k}$, known as *smoothing*, the conditional distribution of $X_{s:t}$ given $Y_{0:k}$, that is, for any given sequence $y_{0:k}$, $A \mapsto \phi_{s:t|k}(A)$ is a probability measure on $(\mathbf{X}^{t-s+1}, \mathcal{X}^{t-s+1})$. Particularly, if $s = t = k$, we abbreviate $\phi_{s:t|k}$ to ϕ_k, and we call it *filtering*. As above, we let $L_n := p(y_{0:n})$ be the likelihood and ℓ_n be the log-likelihood.

Maximimum likelihood estimator. Note $\sum_{j=1}^M q_{ij} = 1$ given $i \in [M]$, we regard $\alpha := (q_1^\top, \ldots, q_{M-1}^\top)^\top$ as transition parameters. Moreover, we consider β as emission parameters and $\theta := (\alpha^\top, \beta^\top)^\top$ as model parameters. The expectation-maximization algorithm allows us to find an approximation of the MLE, and then we can evaluate the score function and Fisher information at this value.

For this HMM, the log-likelihood function is given by

$$\ell_n(\theta) := \log \left\{ \sum_{x_{1:n}=1}^{M} \nu(x_0) g_0(x_0) \prod_{k=1}^{n} q_{x_{k-1}, x_k} g_k(x_k) \right\} .$$

Evaluating this log-likelihood naively is exponentially expensive, imposing huge difficulty on computing the score and information. In the sequel, we will develop a feasible procedure combining Automatic Differentiation and smoothing techniques for the computation of the score and information.

Useful identities. Given a σ-finite measure λ on a measurable space $(\mathbf{X}, \mathcal{X})$, we consider a family $\{f(\cdot;\theta)\}_{\theta\in\Theta}$ of non-negative λ-integrable functions on $\mathbf{X}$. Define $L(\theta) := \int f(x;\theta)\lambda(dx)$ and $p(x;\theta) := f(x;\theta)/L(\theta)$. Under standard continuously differentiable and integrable conditions and

$$\nabla_\theta^k \int \log p(x;\theta) p(x;\theta')\lambda(dx) = \int \nabla_\theta^k \log p(x;\theta) p(x;\theta')\lambda(dx) ,$$

the following identities hold (see [Cappé et al., 2005] for a detailed discussion):

$$\nabla_\theta \ell(\theta) = \int [\nabla_\theta \log f(x;\theta)] p(x;\theta) \lambda(dx) \tag{11}$$

$$\nabla^2_\theta \ell(\theta) + [\nabla_\theta \ell(\theta)][\nabla_\theta \ell(\theta)]^\top = \int \left\{ \nabla^2_\theta \log f(x;\theta) + [\nabla_\theta \log f(x;\theta)][\nabla_\theta \log f(x;\theta)]^\top \right\} p(x;\theta) \lambda(dx) \ , \tag{12}$$

where the first identity is called Fisher's identity [Dempster et al., 1977] and the second one is called Louis' identity [Louis, 1982].

In the HMM setting we consider above, $f(x;\theta)$ is the joint density of $X_{0:n}$ and $Y_{0:n}$ given by

$$f(x;\theta) = \nu(x_0) g_0(x_0) \prod_{k=1}^n q_{x_{k-1},x_k} g_k(x_k) \ ,$$

and $p(x;\theta)$ is the conditional density of $X_{0:n}$ given $Y_{0:n}$. Consequently, we have

$$\begin{aligned}
\nabla_\theta \ell(\theta) &= \mathbb{E}[t_{1,n}(X_{0:n}) | Y_{0:n}] \\
\nabla^2_\theta \ell(\theta) &= -[\nabla_\theta \ell(\theta)][\nabla_\theta \ell(\theta)]^\top + \mathbb{E}[t_{2,n}(X_{0:n}) + t_{3,n}(X_{0:n}) | Y_{0:n}] \ ,
\end{aligned}$$

where,

$$\begin{aligned}
t_{r,n}(X_{0:n}) &:= \sum_{k=0}^n s_{r,k}(X_{k-1}, X_k) := \sum_{k=0}^n \mathbb{1}\{k>0\} \nabla^r_\theta \log q_{X_{k-1},X_k} + \nabla^r_\theta \log g_k(X_k), \forall r \in \{1,2\}, \\
t_{3,n}(X_{0:n}) &:= \Big[\sum_{k=0}^n s_{1,k}(X_{k-1}, X_k) \Big] \Big[\sum_{k=0}^n s_{1,k}(X_{k-1}, X_k) \Big]^\top \\
&= t_{3,n-1}(X_{0:n-1}) + s_{1,n}(X_{n-1}, X_n) t_{1,n-1}(X_{0:n-1})^\top + \\
&\quad t_{1,n-1}(X_{0:n-1}) s_{1,n}(X_{n-1}, X_n)^\top + s_{1,n}(X_{n-1}, X_n) s_{1,n}(X_{n-1}, X_n)^\top \ .
\end{aligned}$$

The calculations of $\nabla^r_\theta \log q_{X_{k-1},X_k}$ and $\nabla^r_\theta \log g_k(X_k)$ fit into the Automatic Differentiation framework. We will show later a recursive way to compute the score and information.

Filtering. To calculate the score and information of HMMs, the first step is to perform filtering. Instead of the well known forward-backward recursions, we apply the so-called normalized forward filtering recursion [Cappé et al., 2005] here to derive the filtering measures: recall that ϕ_k is the conditional distribution of X_k given $Y_{0:k}$, then recursively for $k = 1, \ldots, n$, we have

$$\begin{aligned}
c_k &= \sum_{x_{k-1},x_k=1}^M \phi_{k-1}(x_{k-1}) q_{x_{k-1},x_k} g_k(x_k) = L_k / L_{k-1} \\
\phi_k(x_k) &= c_k^{-1} \sum_{x_{k-1}=1}^M \phi_{k-1}(x_{k-1}) q_{x_{k-1},x_k} g_k(x_k), \forall x_k \in [M] \ ,
\end{aligned}$$

with initial condition

$$\begin{aligned}
c_0 &= \sum_{x_0=1}^M g_0(x_0) \nu(x_0) = L_0 \\
\phi_0(x_0) &= c_0^{-1} g_0(x_0) \nu(x_0), \forall x_0 \in [M] \ .
\end{aligned}$$

Smoothing. Given a sequence of functions $\{t_k\}_{k\geq 0}$ such that $t_k : \mathbf{X}^{k+1} \to \mathbb{R}$ and is defined recursively by

$$t_{k+1}(x_{0:k+1}) = m_{k+1}(x_k, x_{k+1})t_k(x_{0:k}) + s_{k+1}(x_k, x_{k+1})$$

for all $x_{0:k+1} \in \mathbf{X}^{n+2}$ and $k \geq 0$, where $\{m_k\}_{k\geq 0}$ and $\{s_k\}_{k\geq 0}$ are two sequences of measurable functions. Note that this definition can be extended to cases in which $\{t_k\}_{k\geq 0}$ are vector-valued functions. As demonstrated above, $t_{1,n}$ and $t_{2,n}$ fall in this scenario.

We hope to compute $\mathbb{E}[t_n(X_{0:n})|Y_{0:n}]$ recursively in n, assuming that these expectations are indeed finite. We proceed by defining the family of finite signed measures $\{\tau_n\}$ on $(\mathbf{X}, \mathcal{X})$ such that,

$$\tau_n(x_n) := \sum_{x_0,\ldots,x_{n-1}=1}^{M} t_n(x_{0:n})\phi_{0:n|n}(x_{0:n}), \text{for all } x_n \in [M] \ .$$

Hence, $\sum_{x_n=1}^{M} \tau_n(x_n) = \mathbb{E}[t_n(X_{0:n})|Y_{0:n}]$. Notice that, for any $x_{0:n} \in \mathbf{X}^{n+1}$,

$$\phi_{0:n|n}(x_{0:n}) = L_n^{-1}\nu(x_0)g_0(x_0)\prod_{k=1}^{n} q_{x_{k-1},x_k}g_k(x_k) \ .$$

We then have the following recursion: for any $x_{k+1} \in [M]$,

$$\tau_{k+1}(x_{k+1}) = c_{k+1}^{-1}\sum_{x_k=1}^{M} q_{x_k,x_{k+1}}g_{k+1}(x_{k+1})[\tau_k(x_k)m_{k+1}(x_k, x_{k+1}) + \phi_k(x_k)s_{k+1}(x_k, x_{k+1})] \ ,$$

with initial condition:

$$\tau_0(x_0) = c_0^{-1}\nu(x_0)t_0(x_0)g_0(x_0), \forall x_0 \in [M] \ .$$

While $t_{3,n}$ is not directly in this format, it can be formatted as

$$t'_{k+1}(x_{0:k+1}) = t'_k(x_{0:k}) + m_{k+1}(x_k, x_{k+1})t_k(x_{0:k}) + s_{k+1}(x_k, x_{k+1}) \ ,$$

and thus the recursion becomes

$$\tau'_{k+1}(x_{k+1}) = c_{k+1}^{-1}\sum_{x_k=1}^{M} q_{x_k,x_{k+1}}g_{k+1}(x_{k+1})[\tau'_k(x_k) + \tau_k(x_k)m_{k+1}(x_k, x_{k+1}) + \phi_k(x_k)s_{k+1}(x_k, x_{k+1})] \ .$$

C Theoretical results and proofs

The next result is standard for the score function and observed Fisher information; see, for example, [Wakefield, 2013, Chapter 2.4].

Fact 1. *Let $W_1, \ldots, W_n$ be a sequence of i.i.d. copies of W following a family of probability densities $\{p_\theta : \theta \in \Theta\}$ with true value θ_0. Assume that the true parameter $\theta_0 \in \text{int}(\Theta)$ (the interior of Θ), and that the following conditions hold:*

(1) : $\ell(\theta) := \log p_\theta(W)$ is twice continuously differentiable within Θ.

(2) : $\mathcal{I}(\theta) = \mathbb{E}_\theta[\nabla_\theta\ell(\theta)\nabla_\theta\ell(\theta)^\top]$ is positive definite.

Then we have

$$\frac{1}{\sqrt{n}}\nabla_\theta\ell_n(\theta) \to_d \mathcal{N}(0, \mathcal{I}_0),$$

where $\mathcal{I}_0 := \mathcal{I}(\theta_0)$ is referred to as the Fisher information. If further the following condition holds

(3) : $\int p_\theta(x)dx$ *can be differentiated twice under the integral sign.*

Then we also have $-n^{-1}\nabla^2_\theta \ell_n(\theta_0) \to_p \mathcal{I}_0$.

Recall that, in the computation of the scan statistic, we approximate the maximizer of $\max_{T\in\mathcal{T}_p} R_n(\tau, T)$ by the indices of the largest p components in $v(\tau) := \mathrm{diag}(\mathcal{I}_n(\hat{\theta}_n;\tau))^{-1} S^{\odot 2}_{\tau+1:n}(\hat{\theta}_n)$, where $S^{\odot 2}$ represents the element-wise square for a vector S. The next lemma verifies that this approximation is accurate when the difference between the largest eigenvalue and the smallest eigenvalue of $\mathcal{I}_n(\hat{\theta}_n;\tau)^{-1}$ is small compared to $\|S_{\tau+1:n}(\hat{\theta}_n)\|^2$.

Lemma 1. *Let* $\alpha \in \mathbb{R}^d$*, and* $A \in \mathbb{R}^{d\times d}$ *be a positive definite matrix. Given* $p \in [d]$*, consider the optimization problem:*

$$\max_{T\subset[d],|T|=p} f(T) = \alpha_T^\top [A_{T,T}]^{-1}\alpha_T$$

with optimizer T^**. Define* $0 < \lambda_1(A) \le \cdots \le \lambda_d(A)$ *to be the eigenvalues of* A*, and* $\tilde{T}$ *to be the indices of the largest* p *components in* $\mathrm{diag}(A)^{-1}\alpha^{\odot 2}$*. Then we have* $\left|f(T^*) - f(\tilde{T})\right| \le 2[\lambda_1(A)^{-1} - \lambda_d(A)^{-1}]\,\|\alpha\|^2$.

Proof Define $g(T) := \alpha_T^\top \mathrm{diag}(A_{T,T})^{-1}\alpha_T$. According to the definition of $\tilde{T}$, we have, for any $|T| = p$, $g(T) \le g(\tilde{T})$. In particular, we have $g(T^*) \le g(\tilde{T})$. This implies that

$$0 \le f(T^*) - f(\tilde{T}) \le f(T^*) - g(T^*) + g(\tilde{T}) - f(\tilde{T}),$$

and thus it suffices to bound $|f(T) - g(T)|$ for every $|T| = p$.

On the one hand, note that

$$f(T) - g(T) = \alpha_T^\top A_{T,T}^{-1}\alpha_T - \alpha_T^\top(\mathrm{diag}(A_{T,T}))^{-1}\alpha_T \le \lambda_p(A_{T,T}^{-1})\,\|\alpha_T\|^2 - a_{\max}^{-1}\,\|\alpha_T\|^2\,,$$

where $a_{\max} := \max_{i\in[d]} a_{ii}$. By the Courant-Fischer-Weyl min-max principle, we have $0 < \lambda_1(A) \le \lambda_1(A_{T,T})$, which implies $\lambda_p(A_{T,T}^{-1}) = \lambda_1(A_{T,T})^{-1} \le \lambda_1(A)^{-1}$. Moreover, since $0 < \lambda_1(A) \le a_{\max} \le \lambda_d(A)$, we have $a_{\max}^{-1} \ge \lambda_d(A)^{-1}$. It follows that

$$f(T) - g(T) \le [\lambda_1(A)^{-1} - \lambda_d(A)^{-1}]\,\|\alpha\|^2\,.$$

On the other hand, we can obtain, similarly,

$$g(T) - f(T) \le [a_{\min}^{-1} - \lambda_1(A_{T,T}^{-1})]\,\|\alpha\|^2 \le [\lambda_1(A)^{-1} - \lambda_d(A)^{-1}]\,\|\alpha\|^2$$

with $a_{\min} := \min_{i\in[d]} a_{ii}$.

Therefore, we have

$$0 \le f(T^*) - f(\tilde{T}) \le 2[\lambda_1(A)^{-1} - \lambda_d(A)^{-1}]\,\|\alpha\|^2\,.$$

■

Next, we provide full statements and proofs for propositions in Sec. 3.

Proposition 1. *Let* $W_1, \ldots, W_n$ *be a time series with a correctly specified probabilistic model* $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$*, where the parameter* θ *is assumed to be independent of* n*. Assume that the true parameter* $\theta_0 \in \mathrm{int}(\Theta)$ *(the interior of* Θ*), and that the following conditions hold:*

C1 : $\ell_n(\theta) := \log p_\theta(W_1, \ldots, W_n)$ *is twice continuously differentiable within* Θ*.*

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

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

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

If further the following conditions hold

C4 : *The normalized score can be written as a 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 difference sequence satisfies the Lindeberg conditions:

C4-(a) : $n^{-1}\sum_{k=1}^{n}\mathbb{E}[M_k M_k^\top|\mathcal{F}_{k-1}] \to_p \mathcal{I}_0$ *and*

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

Then we can also obtain asymptotic normality of the score:

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

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

Proof Condition C3 implies $\hat{\theta}_n \to_p \theta_0$, then by continuous mapping theorem and Conditions C1 and C2, we have $-\nabla_\theta^2\ell_n(\hat{\theta}_n)/n \to_p \mathcal{I}_0$. It follows that

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

Recall that $\mathcal{I}_n(\hat{\theta}_n;\tau) = \mathcal{I}_{\tau_n+1:n}(\hat{\theta}_n) - \mathcal{I}_{\tau_n+1:n}(\hat{\theta}_n)^\top[\mathcal{I}_{1:n}(\hat{\theta}_n)]^{-1}\mathcal{I}_{\tau_n+1:n}(\hat{\theta}_n)$, we can derive

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

Now assume Condition C4 is true. Since $\hat{\theta}_n$ maximizes the log-likelihood function, it must satisfy the first order optimality condition, *i.e.,* $S_n(\hat{\theta}_n) = 0$. Then by Condition C3 and Taylor expansion,

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

where θ_n^* is between θ_0 and $\hat{\theta}_n$. It follows that $\theta_n^* \to_p \theta_0$, and we also have

$$-\frac{1}{\sqrt{n}}\nabla_\theta Z_n(\theta_n^*) = -\frac{1}{n}\nabla_\theta^2\ell_n(\theta_n^*) = \mathcal{I}_0 + o_p(1) \tag{13}$$

by the continuous mapping theorem. Note that $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$, we can obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathcal{I}_0^{-1}Z_n(\theta_0) + o_p(1) \ . \tag{14}$$

Moreover, by Lindeberg theorem for martingales [Van der Vaart, 2013, Chapter 4.5] 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

$$\begin{aligned}
\frac{S_{\tau_n+1:n}(\hat{\theta}_n)}{\sqrt{n-\tau_n}} &= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \frac{1}{\sqrt{n-\tau_n}} \nabla_\theta S_{\tau_n+1:n}^\top(\theta_n^*)(\hat{\theta}_n - \theta_0) \\
&= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \frac{\left(\nabla_\theta S_n(\theta_n^*) - \nabla_\theta S_{\tau_n}(\theta_n^*)\right)^\top}{\sqrt{n(n-\tau_n)}} \sqrt{n}(\hat{\theta}_n - \theta_0) \\
&= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \left[\sqrt{\frac{n}{n-\tau_n}} \frac{1}{\sqrt{n}} Z_n(\theta_n^*) - \frac{\tau_n}{\sqrt{n(n-\tau_n)}} \frac{1}{\sqrt{\tau_n}} Z_{\tau_n}(\theta_n^*) \right]^\top (\mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1)) \quad \text{by (14)} \\
&= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \left(\frac{\lambda}{\sqrt{1-\lambda}} - \sqrt{\frac{1}{1-\lambda}} \right) \mathcal{I}_0 \mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1) \quad \text{by (13)} \\
&= -\sqrt{\frac{\tau_n}{n-\tau_n}} Z_{\tau_n}(\theta_0) + \sqrt{\frac{n}{n-\tau_n}} Z_n(\theta_0) + \frac{\lambda - 1}{\sqrt{1-\lambda}} Z_n(\theta_0) + o_p(1) \\
&= -\frac{\sqrt{\lambda}}{\sqrt{1-\lambda}} Z_{\tau_n}(\theta_0) + \frac{\lambda}{\sqrt{1-\lambda}} Z_n(\theta_0) + o_p(1) \ .
\end{aligned}$$

Now by applying Lemma 2, we have

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

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 $\overline{N}_n \to_{a.s.} \mathbb{E}[N_1]$, *i.e.,* the condition C4-(a) holds true. Similarly, given $c > 0$,

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

for any $\alpha \in \mathbb{R}^d$, where $G(c) = \mathbb{E}[(\alpha^\top M_1)^2 | \mathbb{1}\left\{\left|\alpha^\top M_1\right| > c\right\}]$ can be arbitrarily small by setting c to be large. Hence, for any $\delta > 0$ and any $\alpha \in \mathbb{R}^d$, there exists a constant c_0 and an integer $N > 0$ such that $\forall n > N$, we have $G_n(c_0) < \delta$ almost surely. To verify the condition C4-(b), note that $G_n(c)$ is decreasing in c, so, for every $\varepsilon > 0$, there exists $M > 0$ such that $n > M$ implies

$$\frac{1}{n} \sum_{k=1}^n \mathbb{E}\left[\alpha^\top M_k^2 | \mathbb{1}\left\{\left|\alpha^\top M_k\right| > \varepsilon\sqrt{n}\right\} | \mathcal{F}_{k-1}\right] \le G_n(c_0) < \delta$$

almost surely. As δ is arbitrary, we know that the condition C4-(b) holds. ■

Remark. For *i.i.d.* models with finite second moment, the normalized score reads $Z_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{k=1}^n \nabla_\theta \ell_k(\theta_0)$. Under regularity conditions, we have $\mathbb{E}[\nabla_\theta \ell_k(\theta_0)] = 0$, 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 $\alpha \in \mathbb{R}^d$,

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

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

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 Prop. 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*

$$\begin{pmatrix} Z_{\tau_n}\sqrt{\tau_n/n} \\ Z_n \end{pmatrix} \to_d \mathcal{N}\left(0, \begin{pmatrix} \lambda\mathcal{I}_0 & \lambda\mathcal{I}_0 \\ \lambda\mathcal{I}_0 & \mathcal{I}_0 \end{pmatrix}\right) \ . \tag{15}$$

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

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

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

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

We will prove this by the Lindeberg theorem for martingales. 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 $W_{n,k} = (a+b)^\top M_k$, if $k \in [\tau_n]$; and $W_{n,k} = b^\top M_k$, if $k \in \{\tau_n+1, \ldots, n\}$. Then $\{W_{n,k}, \mathcal{F}_k\}_{k\in\mathbb{Z}}$ is also a martingale difference sequence. Additionally,

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

and, for any $\varepsilon > 0$,

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

by Condition C4-(b). Therefore, the statement (15) holds by invoking the Lindeberg theorem for martingales, and it follows that

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

■

If all conditions in Prop. 1 are satisfied, 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. This gives the threshold $H_{\text{lin}}(\alpha)=q_{\chi^2_d}(\alpha/n)$—the upper (α/n)-quantile of χ^2_d. Similarly, since the asymptotic distribution of $R_n(\tau,T)$ with $T\in\mathcal{T}_p$ is χ^2_p and $|\mathcal{T}_p|=\binom{d}{p}$, the Bonferroni correction leads to the threshold $H_p(\alpha)=q_{\chi^2_p}(\alpha/[\binom{d}{p}n(p+1)^2])$, where $(p+1)^2$ is required to guarantee an asymptotic α level[8]. Other corrections are possible, but the former provides small thresholds when the change is sparse.

Corollary 4. *Under the assumptions in Prop. 1, the three proposed tests $\psi,\psi_{lin},\psi_{scan}$ are consistent in level with thresholds above.*

Proof Let $\mathbb{E}_0$ and $\mathbb{P}_0$ be the expectation and probability distribution under the null hypothesis. We have

$$\mathbb{E}_0[\psi_{\text{lin}}(\alpha)]=\mathbb{P}_0\{\max_{\tau\in[n-1]}R_n(\tau)>H_{\text{lin}}(\alpha)\}\le\sum_{\tau=1}^{n-1}\mathbb{P}_0(R_n(\tau)>q_{\chi^2_d}(\alpha/n))\le\sum_{\tau=1}^{n-1}\frac{\alpha}{n}+o(1)=\alpha+o(1) ,$$

and

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

For $\alpha=\alpha_l+\alpha_s$, the *autograd-test* has false alarm rate

$$\mathbb{E}_0[\psi(\alpha)]\le\mathbb{E}_0[\psi_{\text{lin}}(\alpha_l)]+\mathbb{E}_0[\psi_{\text{scan}}(\alpha_s)]\le\alpha_l+\alpha_s+o(1)=\alpha+o(1) .$$

Therefore, the three proposed tests are all consistent in level. ■

Proposition 2. *Given an independent sample $W_1,\dots,W_n$ and a family of density functions $\{p_\theta:\theta\in\Theta\subset\mathbb{R}^d\}$ satisfying that there exists $\tau_n\in[n-1]$ such that $W_1,\dots,W_{\tau_n}\sim p_{\theta_0}$, $W_{\tau_n+1},\dots,W_n\sim p_{\theta_1}(\theta_1\neq\theta_0)$, and $\tau_n/n\to\lambda\in(0,1)$. Assume the following conditions hold:*

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

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

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

$$\tfrac{1}{n} R_n(\tau_n) \to_p (\overline{\lambda} S_1^*)^\top (\mathcal{I}^*)^{-1} (\overline{\lambda} S_1^*) \ , \tag{16}$$

where $\mathcal{I}^* = \overline{\lambda}\mathcal{I}_1^* - \overline{\lambda}\mathcal{I}_1^* \left(\lambda \mathcal{I}_0^* + \overline{\lambda}\mathcal{I}_1^*\right)^{-1} \overline{\lambda}\mathcal{I}_1^*$ *is a positive definite matrix. If in addition* $S_1^* \neq 0$, *then the three proposed tests* $\psi, \psi_{lin}, \psi_{scan}$ *are consistent in power.*

Proof Among all solutions of the likelihood equation $\nabla_\theta \ell_n(\theta) = 0$, let $\hat{\theta}_n$ be the one that is 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 $B_\varepsilon = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\| \le \varepsilon\} \subset \Theta^*$ and $\mathrm{bd}(B_\varepsilon)$ be the boundary of B_ε. We will show that, for sufficiently small ε,

$$\mathbb{P}\left(\ell_n(\theta) < \ell_n(\theta^*), \forall \theta \in \mathrm{bd}(B_\varepsilon)\right) \to 1 \ . \tag{17}$$

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

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

where $\overline{\theta}_n \in B_\varepsilon$ satisfies $\left\|\overline{\theta}_n - \theta^*\right\| \le \|\theta - \theta^*\|$. Note that, by law of large numbers,

$$\begin{aligned}
D_1 &\to_p (\theta - \theta^*)^\top \left[\lambda \mathbb{E}_{\theta_0}[\nabla_\theta \ell(\theta^*)] + \overline{\lambda}\mathbb{E}_{\theta_1}[\nabla_\theta \ell(\theta^*)]\right] \\
&= (\theta - \theta^*)^\top \nabla_\theta \left[\lambda \mathbb{E}_{\theta_0}[\ell(\theta)] + \overline{\lambda}\mathbb{E}_{\theta_1}[\ell(\theta)]\right]\big|_{\theta=\theta^*} \qquad \text{by Condition C'3} \\
&= -(\theta - \theta^*)^\top \nabla_\theta \left[\lambda D_{KL}(p_{\theta_0}\|p_\theta) + \overline{\lambda} D_{KL}(p_{\theta_1}\|p_\theta)\right]\big|_{\theta=\theta^*} \\
&= 0 \ ,
\end{aligned}$$

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

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

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

$$\begin{aligned}
|D_3| &\le \frac{1}{6n}\sum_{ijk} |\theta_i - \theta_i^*| \left|\theta_j - \theta_j^*\right| |\theta_k - \theta_k^*| \sum_{l=1}^n \left|\nabla_{ijk}\ell(\overline{\theta}_n|W_l)\right| \qquad \text{by triangle inequality} \\
&\le \frac{1}{6}\varepsilon^3 \sum_{ijk} \frac{1}{n}\sum_{l=1}^n M_{ijk}(W_l) \qquad \text{by } |\theta_i - \theta_i^*| \le \|\theta - \theta^*\| = \varepsilon \\
&\to_p \frac{\varepsilon^3}{6}\sum_{ijk}\left(\lambda \mathbb{E}_{\theta_0}[M_{ijk}(W)] + \overline{\lambda}\mathbb{E}_{\theta_1}[M_{ijk}(W)]\right) \ .
\end{aligned}$$

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,

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

and thus (17) holds.

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

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

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

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

Moreover, according to Lemma 3, we have, for any $\alpha \in (0,1)$,

$$H_{\mathrm{lin}}(\alpha) = q_{\chi^2_d}(\alpha/n) \leq d + 2\sqrt{d\log(n/\alpha)} + 2\log(n/\alpha),$$

and thus $H_{\mathrm{lin}}(\alpha)/n \to 0$. If $S_1^* \neq 0$, then it follows from the positive definiteness of $\mathcal{I}^*$ that

$$\mathbb{P}(\psi_{\mathrm{lin}}(\alpha) = 1) = \mathbb{P}\left(R_{\mathrm{lin}} > H_{\mathrm{lin}}(\alpha)\right) \geq \mathbb{P}\left(\frac{1}{n}R_n(\tau_n) > \frac{1}{n}H_{\mathrm{lin}}(\alpha)\right) \to 1.$$

Analogously, we get

$$H_p(\alpha) = q_{\chi^2_p}\left(\alpha/\left(\binom{d}{p}n(p+1)^2\right)\right) \leq p + 2\left\{p\log\left[\binom{d}{p}n(p+1)^2/\alpha\right]\right\}^{1/2} + 2\log\left[\binom{d}{p}n(p+1)^2/\alpha\right] \ ,$$

which implies $H_p(\alpha)/n \to 0$. Therefore, it follows that $\mathbb{P}(\psi_{\mathrm{scan}}(\alpha) = 1) \to 1$, and subsequently, $\mathbb{P}(\psi(\alpha) = 1) \to 1$. ■

Next lemma is a concentration inequality valid for χ^2 distributions introduced in Birgé [2001].

Lemma 3. *Let W be a non-central chi-square random variable with non-centrality parameter a^2 and degrees of freedom d, that is $W \sim \chi^2_d(a^2)$. Then $\forall x > 0$,*

$$\mathbb{P}\left\{W \geq d + a^2 + 2\sqrt{(d+2a^2)x} + 2x\right\} \leq e^{-x} \ ,$$

and

$$\mathbb{P}\left\{W \leq d + a^2 - 2\sqrt{(d+2a^2)x}\right\} \leq e^{-x} \ .$$

Proposition 3. *Given an independent sample $W_1, \ldots, W_n$ and a family of density functions $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ satisfying that there exists $\tau_n \in [n-1]$ such that $W_1, \ldots, W_{\tau_n} \sim p_{\theta_0}$, $W_{\tau_n+1}, \ldots, W_n \sim p_{\theta_n}$ 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 $W_1, \ldots, W_n$ 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* $\mathbb{E}_{\theta_0} M_{ijk}(W) < \infty$.

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

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

$$\frac{n}{\tau_n(n-\tau_n)} \mathcal{I}_n(\hat{\theta}_n; \tau_n) \to_p \mathcal{I}_0 \ , \tag{18}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}_d\left(\overline{\lambda} h, \mathcal{I}_0^{-1}\right) \ , \tag{19}$$

$$\text{and} \quad \sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{\tau_n+1:n}(\hat{\theta}_n) \to_d \mathcal{N}_d(\sqrt{\lambda\overline{\lambda}}\, \mathcal{I}_0 h, \mathcal{I}_0) \ . \tag{20}$$

In particular,

$$R_n(\tau_n) \to_d \chi^2_d\big(\lambda\overline{\lambda} h^\top \mathcal{I}_0 h\big),$$
$$R_n(\tau_n, T) \to_d \chi^2_{|T|}\Big(\lambda\overline{\lambda} [\mathcal{I}_0 h]_T^\top [\mathcal{I}_0]^{-1}_{T,T} [\mathcal{I}_0 h]_T\Big) \ .$$

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, 1998], 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 Prop. 2 implies that there exists a sequence of MLE such that $\hat{\theta}_n \to_p \theta_0$, then (18) directly follows from the proof in Prop. 1. Furthermore, by Condition C"2-(a) and the mean value theorem, there exists $\overline{\theta}_n$ such that $\|\overline{\theta}_n - \theta_0\| \leq \|\hat{\theta}_n - \theta_0\|$, and

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

The law of large numbers and continuous mapping theorem yields $n^{-1}\nabla_\theta S_{1:n}(\overline{\theta}_n) = -\mathcal{I}_0 + o_p(1)$, and the central limit theorem implies $\sqrt{n}^{-1} S_{1:n}(\theta_0) = O_p(1)$. Therefore,

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

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

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

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

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

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

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

and by the Cramér-Wold device, the statement (19) holds.

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

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

An analogous argument gives, under $\mathbb{P}^{(\tau_n)}_{\theta_0,\theta_n}$,

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

which yields (20). Now, the asymptotic distributions of $R_n(\tau_n)$ and $R_n(\tau_n, T)$ follows immediately from the continuous mapping theorem. ■

D Non-asymptotic results

In this section we present an attempt to derive non-asymptotic results under the null without assuming specific model structures. It is possible to get tighter bound for some particular model; however, a unification of these results is not within reach as a generic sharp concentration bound for the maximum likelihood estimator (MLE) is still in progress.

Under the null hypothesis, assume the following conditions hold:

(1) $W_1, \ldots, W_n$ are independent.

(2) $\ell(\theta)$ is twice continuously differentiable.

(3) $\mathbb{E}[\nabla \ell(\theta_0) \nabla \ell(\theta_0)^\top] = \mathbb{E}[-\nabla^2 \ell(\theta_0)] = \mathcal{I}_0$ is positive definite.

(4) for any $\varepsilon > 0$, there exists $C_\varepsilon > 0$ such that $\|\nabla \ell(\theta) - \nabla \ell(\theta_0)\| \le C_\epsilon \|\theta - \theta_0\|$ for all $\theta \in \Theta$ and $\|\theta - \theta_0\| \le \varepsilon$ almost surely.

(5) the MLE $\hat{\theta}_n$ exists and has a tail bound $\mathbb{P}(\|\hat{\theta}_n - \theta_0\| > \varepsilon) \le h(n, \varepsilon)$.

Then we can give a tail bound for $S_{\tau_n+1:n}(\hat{\theta}_n) := \sum_{k=\tau_n+1}^{n} \nabla \ell_k(\hat{\theta}_n)$.

Note that a union of zero-measure sets also has measure zero, we know, by assumption (4), for all $\|\theta - \theta_0\| \le \varepsilon$,

$$\|S_{\tau_n+1:n}(\theta) - S_{\tau_n+1:n}(\theta_0)\| \le \sum_{k=\tau_n+1}^{n} \|S_k(\theta) - S_k(\theta_0)\| \le (n - \tau_n) C_\varepsilon \|\theta - \theta_0\|$$

almost surely. Let $A := \{\|S_{\tau_n+1:n}(\hat{\theta}_n) - S_{\tau_n+1:n}(\theta_0)\| > \varepsilon/2\}$. It follows that

$$\begin{aligned}\mathbb{P}(A) &\le \mathbb{P}(A \cap \{\|\hat{\theta}_n - \theta_0\| \le \varepsilon\}) + \mathbb{P}(\|\hat{\theta}_n - \theta_0\| > \varepsilon) \\ &\le \mathbb{P}((n-\tau_n)C_\varepsilon\|\hat{\theta}_n - \theta_0\| > \varepsilon/2) + \mathbb{P}(\|\hat{\theta}_n - \theta_0\| > \varepsilon) \\ &\le h\left(n, \frac{\varepsilon}{2(n-\tau_n)C_\varepsilon}\right) + h(n,\varepsilon) \ ,\end{aligned}$$

where we use condition (5) in the last inequality.

According to the Chebyshev's inequality[9], we have, for any $\varepsilon > 0$,

$$\mathbb{P}(\|S_{\tau_n+1:n}(\theta_0) - 0\| > \varepsilon) \le \varepsilon^{-2}\mathbb{E}[\|S_{\tau_n+1:n}(\theta_0)\|^2] = \frac{n-\tau_n}{\varepsilon^2}\mathbb{E}[S_1(\theta_0)^\top S_1(\theta_0)] = \frac{\sigma^2(n-\tau_n)}{\varepsilon^2} \ , \tag{21}$$

where $\sigma^2 := \mathbf{1}^\top \mathrm{diag}(\mathcal{I}_0)\mathbf{1}$ and the last equality follows from the condition (3). Therefore,

$$\begin{aligned}\mathbb{P}(\|S_{\tau_n+1:n}(\hat{\theta}_n)\| > \varepsilon) &\le \mathbb{P}(A) + \mathbb{P}(\|S_{\tau_n+1:n}(\theta_0)\| > \frac{\varepsilon}{2}) \\ &\le h\left(n, \frac{\varepsilon}{2(n-\tau_n)C_\varepsilon}\right) + h(n,\varepsilon) + \frac{4\sigma^2(n-\tau_n)}{\varepsilon^2} \ .\end{aligned} \tag{22}$$

Conditions (1)-(3) above are standard. However, condition (4) asks the score function to be locally Lipschitz continuous at θ_0 with a Lipschitz constant being uniform across all samples, which is rather restrictive. And assumption (5) requires a generic sharp tail bound for the MLE so that the upper bound in (22) is small, which is still in progress. That is the main reason why we choose the asymptotic approach to determine the threshold.

As an illustration of why this is hard to obtain, we provide a naive sufficient condition:

(5') there exists a constant $c > 0$ such that $-\nabla^2\ell_{1:n}(\theta) \succeq cnI_d$ for all $\theta \in \Theta$ almost surely.

That is, we require the Hessian of the negative log-likelihood to be strongly convex on Θ almost surely, with a constant independent of the data $W_{1:n}$. According to assumptions (2) and (5'), there exists $\tilde{\theta}_n \in \Theta$ such that

$$\|S_{1:n}(\theta_0)\| = \left\|S_{1:n}(\theta_0) - S_{1:n}(\hat{\theta}_n)\right\| = \left\|-\nabla^2\ell_{1:n}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)\right\| \ge nc\left\|\hat{\theta}_n - \theta_0\right\|$$

almost surely. Note that, from (21) we have $\mathbb{P}(\|S_{1:n}(\theta_0)\| > \varepsilon) \le \varepsilon^{-2}\sigma^2 n$. Therefore, $\mathbb{P}(\|\hat{\theta}_n - \theta_0\| > \varepsilon) \le \frac{\sigma^2}{nc^2\varepsilon^2}$, *i.e.,* condition (5) holds with $h(n,\varepsilon) = \sigma^2/(nc^2\varepsilon^2)$.

Next, we consider two simple examples. For a Gaussian change in mean model with known covariance, we know $\nabla\ell(\theta) = W - \theta$ and $\nabla^2\ell(\theta) = -I_d$, so assumption (4) holds with $C_\varepsilon \equiv 1$ and assumption (5') holds with $c = 1$.

For a linear regression model with random design, we assume error terms follow a standard normal distribution for simplicity. Then we can obtain $\nabla\ell(\theta) = XY - XX^\top\theta$ and $\nabla^2\ell(\theta) = -XX^\top$. Observe that $\|\nabla\ell(\theta) - \nabla\ell(\theta_0)\| = \|XX^\top\|\|\theta - \theta_0\|$, condition (4) requires $\lambda_{\max}(XX^\top)$ to be upper bounded by a constant $C > 0$ almost surely. This is true if we assume X is bounded. On the other hand, condition (5') asks $\lambda_{\min}(n^{-1}\sum_{k=1}^n X_k X_k^\top)$ to be lower bounded by a constant $c > 0$ almost surely, which is rather hard to guarantee.

E Comparison with generalized likelihood ratio type tests

In this section we compare the proposed score-based statistics with generalized likelihood ratio (GLR) type statistics. We demonstrate the computational superiority of the *autograd-test* over the GLR-type test.

[9]If we are willing to impose more conditions on $(W_{1:n})$, we may obtain sharper bounds by using exponential inequalities such as Hoeffding's and Bernstein's; however, the goal here is to obtain results as general as possible.

Given a significance level $\alpha \in (0,1)$, the GLR-type linear statistic for testing Problem (P0) is given by $G_{\text{lin}} := \max_{\tau\in[n-1]} G_n(\tau)$ with

$$G_n(\tau) = 2\log \frac{\sup_{\theta\in\Theta,\Delta\neq 0} L_n(\theta,\Delta)}{\sup_{\theta\in\Theta} L_n(\theta,0)} = 2 \sup_{\theta\in\Theta,\Delta\neq 0} \ell_n(\theta,\Delta) - 2\ell_n(\hat{\theta}_n) \ . \tag{23}$$

Consequently, the GLR-type linear test reads $\phi_{\text{lin}}(\alpha) := \mathbb{1}\{G_{\text{lin}} > H'_{\text{lin}}(\alpha)\}$ in which $H'_{\text{lin}}(\alpha)$ is a pre-determined threshold. Similarly, for testing Problem (P1), the GLR-type statistic is given by

$$G_{\text{scan}}(\alpha) := \max_{\tau\in[n-1]} \max_{p\leq P} \max_{T\in\mathcal{T}_p} H'_p(\alpha)^{-1} G_n(\tau,T) \ , \tag{24}$$

where $H'_p(\alpha)$ is some pre-determined threshold and

$$G_n(\tau,T) = 2\log \frac{\sup_{\theta\in\Theta,\Delta\in\text{span}^\star(T)} L_n(\theta,\Delta)}{\sup_{\theta\in\Theta} L_n(\theta,0)} = 2 \sup_{\theta\in\Theta,\Delta\in\text{span}^\star(T)} \ell_n(\theta,\Delta) - 2\ell_n(\hat{\theta}_n) \ , \tag{25}$$

with $\text{span}^*(T) := \{v\in\mathbb{R}^d : v_i \neq 0 \text{ if and only if } i\in T\}$. It follows that the GLR-type scan test are given by $\phi_{\text{scan}}(\alpha) := \mathbb{1}\{G_n(\mathcal{P};\alpha) > 1\}$.

In Enikeeva and Harchaoui [2019] the authors considered these GLR-type tests in Gaussian change in mean models and investigated its theoretical properties in high dimensional regimes. Following their settings, let us assume $W_{1:n}$ are *i.i.d.* with distribution $\mathcal{N}_d(\theta_0, I_d)$. Then the maximum likelihood estimator (MLE) under $\mathbf{H}_0$ is simply given by the average $\overline{W}_n := \overline{W}_{1:n} := n^{-1}\sum_{i=1}^n W_i$. To compute the MLE under $\mathbf{H}_1$ in Problem (P0), note that[10]

$$\ell_n(\theta,\Delta) = \ell_n(\theta,\Delta;\tau) = -\frac{1}{2}\sum_{k=1}^{\tau} \|W_k-\theta\|^2 - \frac{1}{2}\sum_{k=\tau+1}^{n} \|W_k-\theta-\Delta\|^2 + C.$$

Let $\nabla_{(\theta,\Delta)}\ell_n(\theta,\Delta) = 0$ we obtain the MLE under $\mathbf{H}_1$: $(\hat{\theta}_n(\tau), \hat{\Delta}_n(\tau)) = (\overline{W}_\tau, \overline{W}_{\tau+1:n} - \overline{W}_\tau)$. It follows that

$$\begin{aligned}
G_n(\tau) &= 2\ell_n(\hat{\theta}_n(\tau), \hat{\Delta}_n(\tau)) - 2\ell_n(\hat{\theta}_n) \\
&= -\sum_{k=1}^{\tau} \left\|W_k - \overline{W}_\tau\right\|^2 - \sum_{k=\tau+1}^{n} \left\|W_k - \overline{W}_{\tau+1:n}\right\|^2 + \sum_{k=1}^{n} \left\|W_k - \overline{W}_n\right\|^2 \\
&= \tau\left\|\overline{W}_\tau\right\|^2 + (n-\tau)\left\|\overline{W}_{\tau+1:n}\right\|^2 - n\left\|\overline{W}_n\right\|^2 \\
&= \tau\left\|\overline{W}_\tau\right\|^2 + (n-\tau)\left\|\overline{W}_{\tau+1:n}\right\|^2 - n\left\|\frac{\tau}{n}\overline{W}_{1:\tau} + \frac{n-\tau}{n}\overline{W}_{\tau+1:n}\right\|^2 \\
&= \frac{(n-\tau)\tau}{n}\left[\left\|\overline{W}_{1:\tau}\right\|^2 + \left\|\overline{W}_{\tau+1:n}\right\|^2 - 2\overline{W}_{1:\tau}^\top \overline{W}_{\tau+1:n}\right] \qquad (26) \\
&= \frac{(n-\tau)\tau}{n}\left\|\hat{\Delta}_\tau\right\|^2 \ . \qquad (27)
\end{aligned}$$

Similarly, we also have $G_n(\tau,T) = \frac{(n-\tau)\tau}{n}\left\|[\hat{\Delta}_\tau]_T\right\|^2$.

Let us now derive our proposed score-based statistics. Observe that

$$\nabla_\theta \ell_k(\theta) = W_k - \theta \quad \text{and} \quad \nabla^2_\theta \ell_k(\theta) \equiv -I_d,$$

so we have $S_{\tau+1:n}(\hat{\theta}_n) = \sum_{k=\tau+1}^n (W_k - \overline{W}_n) = \frac{(n-\tau)\tau}{n}\hat{\Delta}_\tau$ and $\mathcal{I}_n(\hat{\theta}_n;\tau) = (n-\tau)I_d - \frac{(n-\tau)^2}{n}I_d = \frac{(n-\tau)\tau}{n}I_d$. This implies

$$R_n(\tau) = S_{\tau+1:n}^\top(\hat{\theta}_n)\left[\mathcal{I}_n(\hat{\theta}_n;\tau)\right]^{-1} S_{\tau+1:n}(\hat{\theta}_n) = \frac{(n-\tau)\tau}{n}\left\|\hat{\Delta}_\tau\right\|^2 = G_n(\tau) \ , \tag{28}$$

[10] We use C for a constant.

and $R_n(\tau, T) = \frac{(n-\tau)\tau}{n} \left\| [\hat{\Delta}_\tau]_T \right\|^2 = G_n(\tau, T)$. Hence, our score-based statistics recover their GLR-type statistics in this special case.

Next, we will compare the computational complexity of these two types of tests. Note that computing exactly the GLR-type scan statistic is infeasible, since it involves a maximization over all subsets of $[d]$ of cardinality $p \leq P$, which may be exponentially expensive in d. For practical purpose, we apply an approximation to the GLR-type scan statistic as a counterpart to the one used in the score-based scan statistic:

(1) compute d quantities $V := \{\sup_{\theta, \Delta \in \mathrm{span}^\star(\{k\})} \ell_n(\theta, \Delta)\}_{k \in [d]}$;

(2) sort V and obtain a number indicating the order for each $k \in [d]$;

(3) select p indices $\tilde{T}_p = \{k_1, \ldots, k_p\}$ that give the largest p values in V;

(4) compute $G_n(\tau, \tilde{T}_p)$ and use it as an approximation of $\max_{T \in \mathcal{T}_p} G_n(\tau, T)$.

Property 1. *In general, with above approximation, the GLR-type counterpart of the autograd-test is computationally more expensive than the autograd-test.*

Proof We denote by $\mathcal{C}(n, 2d)$ the complexity of solving the optimization problem on the RHS of (23), then calculating G_{lin} costs $n\mathcal{C}(n, 2d)$ time. With the approximation mentioned above, for each $\tau \in [n-1]$, we take steps (1) and (2) and go through steps (3) and (4) for all $p \in [P]$. Note that these steps have complexities $d\mathcal{C}(n, d+1)$, $\mathcal{O}(d \log d)$, $\mathcal{O}(p)$, and $\mathcal{C}(n, d+p)$, respectively. Therefore, the complexity of computing $G_{\mathrm{scan}}(\alpha)$ is

$$\mathcal{O}\big(nd\mathcal{C}(n, d+1) + nd \log d + nP\mathcal{C}(n, d+P))\big) \ .$$

This is also the overall complexity as it dominates $n\mathcal{C}(n, 2d)$.

Given an invertible matrix $A \in \mathbb{R}^{d \times d}$ and a vector $b \in \mathbb{R}^d$, computing $A^{-1}b$ can be formulated as an optimization problem $A^{-1}b = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|^2$. Recall that we assume its complexity to be $\mathcal{O}(d^3)$ in the analysis of *autograd-test*. We can view it as solving that optimization problem by running gradient descent for d steps where each update step costs $\mathcal{O}(d^2)$. For comparison purpose we assume $\mathcal{C}(n, d) = \mathcal{O}(nd^2)$ as computing the gradient here costs $\mathcal{O}(nd)$. Consequently, the overall complexity of the GLR-type counterpart is $\mathcal{O}(n^2 d^3)$ (as $P \leq d$), which is significantly more expensive than the one of *autograd-test*, $\mathcal{O}(n^2 d^2 + nd^3)$, when the dimension d is high. ■

Remark. For *i.i.d.* models, one may utilized stochastic gradient descent instead of gradient descent, leading to complexity $\mathcal{C}(n, d) = \mathcal{O}(d^3)$. As a result, the overall complexity of the GLR-type counterpart becomes $\mathcal{O}(nd^4)$, which is, again, d times larger than the one of *autograd-test*, $\mathcal{O}(nd^3)$.

As indicated by above discussion, the main reason why GLR-type counterpart is more expensive is that it lacks an efficient way to scan components of model parameters. Specifically, given $\tau \in [n-1]$, for *autograd-test*, once we have computed the score function and observed Fisher information, we just need to compute the statistics with subvectors and submatrices of dimensions p and $p \times p$, respectively. And this procedure would cost only $\mathcal{O}(p^3)$ time. On the contrary, for GLR-type test, we have to re-solve the maximum likelihood problem under different alternatives. And because of the existence of $\theta \in \mathbb{R}^d$ in these problems, it requires at least $\mathcal{C}(n, d) = \mathcal{O}(nd^2)$ time. Therefore, to obtain a comparable time complexity, one has to further approximation $\sup_{\theta, \Delta \in \mathrm{span}^\star(\{k\})} \ell_n(\theta, \Delta)$ by $\sup_{\Delta \in \mathrm{span}^\star(\{k\})} \ell_n(\hat{\theta}_n, \Delta)$.

F Experimental details

In this section, we perform simulations to evaluate empirical behavior of our methods on synthetic data generated from an additive model, a time series model, a hidden Markov model (HMM), and a text topic

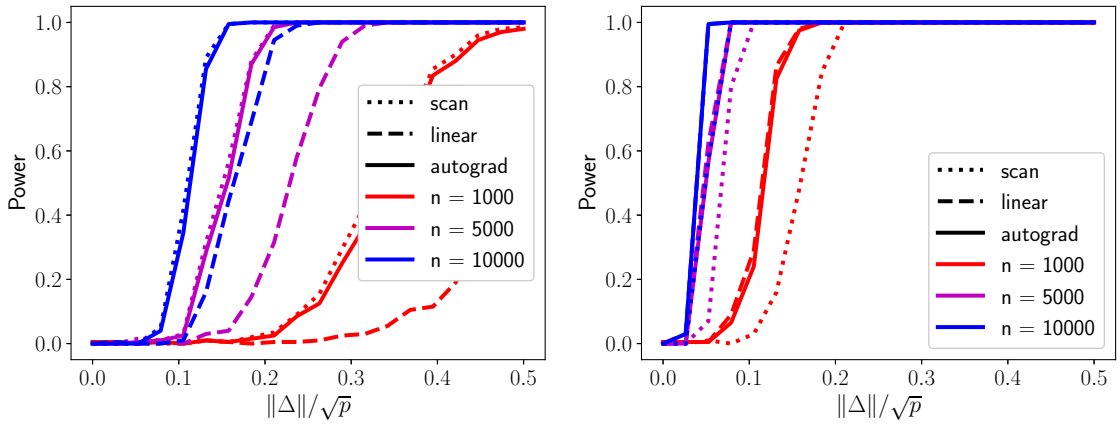

Figure 8: Power versus magnitude of change for linear models (left: $p = 1$; right: $p = 20$).

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.

Synthetic experimental settings. For each model, we generate the first half of the sample from this 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, and generate the second half from the same model with parameter θ_1. Next, we run the linear test, the scan test, and the *autograd-test* to monitor the process of parameters learning, where the significance levels are set to be $\alpha = 2\alpha_l = 2\alpha_s = 0.05$ and the maximum cardinality P is chosen as $\lfloor \sqrt{d} \rfloor$. 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. Finally, 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. Note that the value at $\delta = 0$ is an empirical estimate of the false alarm rate.

Additive 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 before change. All the entries of the design matrix and error terms are generated independently from a standard normal distribution. As shown in Fig. 8, when the change is sparse ($p = 1$), the scan test and the *autograd-test* share similar power curves and both outperform the 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.

Moreover, we examine the component screening feature of these score-based tests. We consider the same linear regression model with $p = 1$, except that we only screen 50 components of the slope coefficients (*i.e.,* regard the rest 51 components as nuisance parameters). Results in Fig. 9 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 detection power is below 0.01 no matter how strong the signal is. Hence, with the restriction imposed, the decision of these tests is unlikely to be affected by the change in nuisance parameters.

Time series model. We investigate two different autoregressive–moving-average models—ARMA$(3, 2)$ and ARMA$(6, 5)$. For the resulting time series to be stationary, we need to ensure that the polynomial

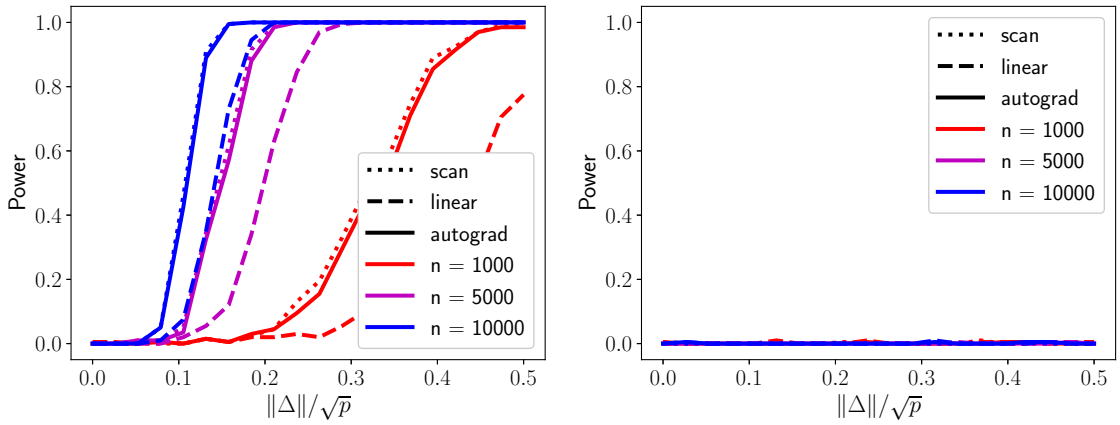

Figure 9: Power versus magnitude of change for linear regression with restriction (left: contains the changed component; right: does not contain the changed component).

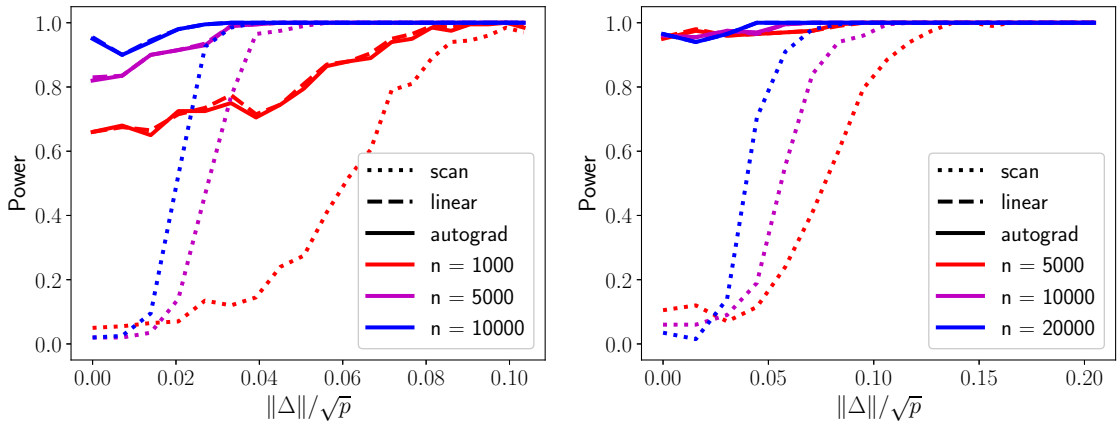

Figure 10: Power versus magnitude of change for ARMA$(3,2)$ (left) and ARMA$(6,5)$ (right).

induced by AR coefficients has roots within $(-1,1)$. We take the following 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. 10.

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 false alarm rate. This problem gets more severe as the sample size increases, and hence is not due to lack of accuracy of the maximum likelihood estimator (MLE). It turns out that this 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 partial information (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 $p \le \sqrt{d}$, the submatrix has much smaller condition number (the parameters selected by the scan 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

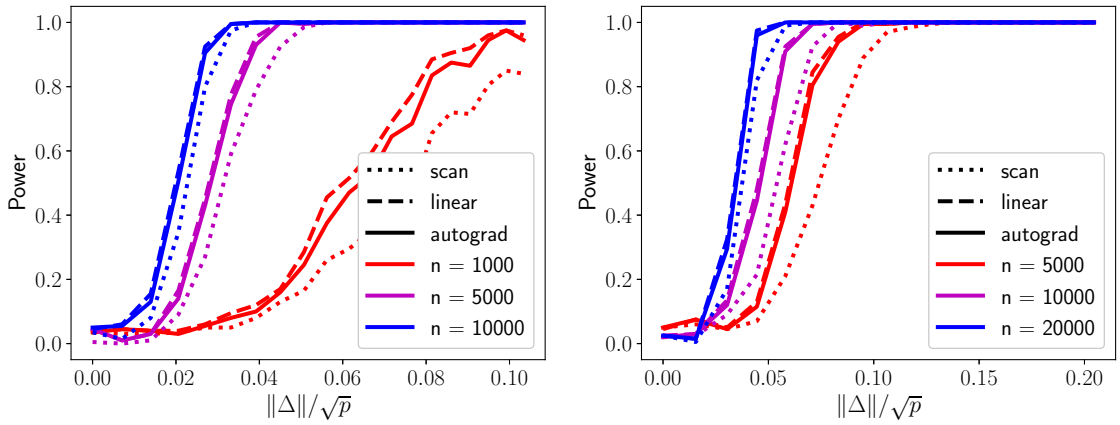

Figure 11: Power versus magnitude of change for ARMA models with restricted components (left: ARMA$(3, 2)$; right: ARMA$(6, 5)$).

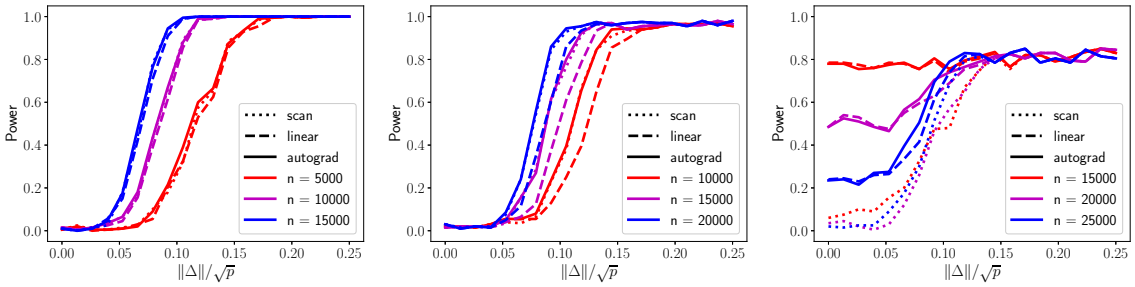

Figure 12: Power versus magnitude of change for HMMs with N hidden states (left: $N = 3$; middle: $N = 7$; right: $N = 15$).

the *autograd-test* (so the scan part has a dominating effect) to obtain reasonable results. If we restrict the screening of these tests in the AR coefficients, as presented in Fig. 11, 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 Markov model. We then investigate HMMs with $N \in \{3, 7, 15\}$ hidden states and normal emission distribution. The transition matrix is sampled in the following way: each row (the distribution of next state conditioning on current state) is the sum of vector $(2N)^{-1}\mathbf{1}_N$ and a Dirichlet sample with concentration parameters $0.5\mathbf{1}_N$, where $\mathbf{1}_N$ is an all one vector of length N. All entries in the resulting vector are positive and sum to one. Given the state $k \in \{0, \ldots, N-1\}$, the emission distribution has mean k and standard deviation $0.01 + 0.09k/(N-1)$ so that they are evenly distributed within $[0.01, 0.1]$. Due to the constraint that each row of the transition matrix must sum to one, we only view entries in the first $N-1$ columns as transition parameters. The post-change transition matrix is obtained by subtracting δ from the $(1, 1)$ entry and adding δ to the $(1, N)$ entry.

Results are shown in Fig. 12. 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* outperform the linear test. When $N = 15$, the linear test and *autograd-test* become inconsistent in level, but, different from the situation for ARMA models, the inconsistency is 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 empirical 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.

Table 2: Decision for Pair-wise Experiments (each (row, column) pair stands for a concatenation of two seasons of shows; "R" means reject and "N" means not reject).

	F1	F2	M1	M2	S1	S2	D1	D2
F1	N	N	N	N	R	R	R	R
F2	N	N	*R*	N	R	R	R	R
M1	N	*R*	N	N	R	R	R	R
M2	N	N	N	*NA*	R	R	R	R
S1	R	R	R	R	N	N	*R*	*R*
S2	R	R	R	R	N	N	*R*	*R*
D1	R	R	R	R	*R*	*R*	N	*R*
D2	R	R	R	R	*R*	*R*	N	N

Real data application. We now consider a real data application. We collect subtitles of the first two seasons of four different TV shows—Friends (F), Modern Family (M), the Sopranos (S), and Deadwood (D)[11]—where the former two are assumed to be "polite" and the latter two are assumed to be "rude". For every pair of seasons, we concatenate them, and the task is to detect changes in rudeness level, while ignoring other alterations.

After standard 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 text topic model on them, where the number of hidden states is chosen to be $\lfloor \sqrt{n/100} \rfloor$ 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 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.

As demonstrated in Table 2, the scan test does a perfect job in reporting shifts in rudeness level. However, the false alarm rate is relatively high. For ("polite", "polite") pairs, there are two false alarms 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 alarms, 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 rejection frequency[12]. 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 false alarm rate 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 utilize the component screening feature and restrict 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.

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

[12] Note that the outcome is a binary variable so that the standard error is given by $\hat{\mu}(1-\hat{\mu})$.

Table 3: Rejection rate for pair-wise experiments (each (row, column) pair stands for a concatenation of two seasons of shows).

	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

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 from the model before change. 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 $\tau \le t_a \le n$, 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, $\mathbb{E}[t_a - \tau | \tau \le t_a \le n]$, and we 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 false alarm rate, presents a significant increment. As the sample size increases, the false alarm rate 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 power 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 a strong signal the conditional mean delay exhibit a similar behavior as the first scenario while it is exactly the opposite case for a 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 false alarm rates 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. 14, contrary to linear regression with 10 slope coefficients, the power increases as the sample size gets larger in both situations, and the false alarm rate becomes uncontrolled; while for the conditional mean delay, it shows similar trends.

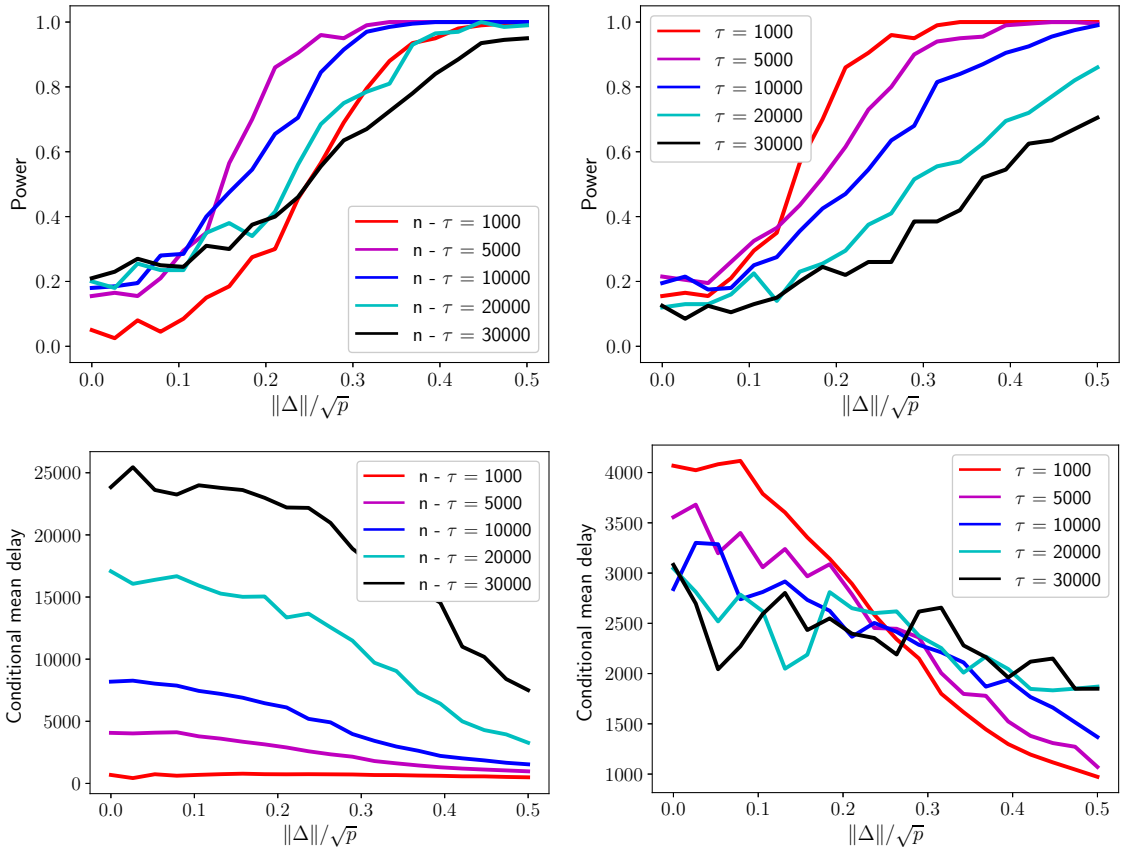

Figure 13: Power versus magnitude of change (up) and conditional average run length versus magnitude of change (down) for *autograd-test-CuSum* (left: $\tau = 1000$; right: $n - \tau = 5000$).

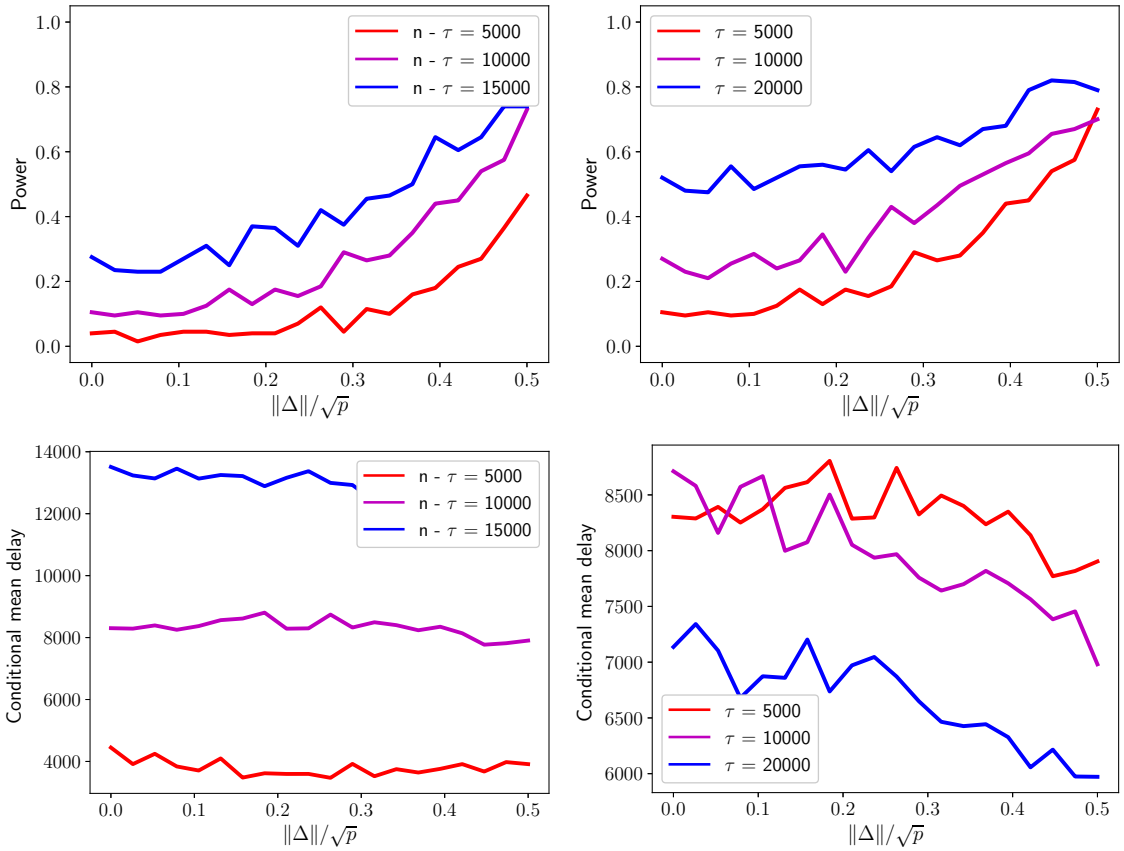

Figure 14: Power versus magnitude of change (up) and conditional average run length versus magnitude of change (down) for *autograd-test-CuSum* (left: $\tau = 5000$; right: $n - \tau = 10000$).