Gradient-Based Monitoring of Learning Machines

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

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 synthetic and real data.

1. Introduction

Statistical machine learning models are now fostering progress in numerous technological applications, *e.g.,* visual object recognition and language processing, as well as in many scientific domains, *e.g.,* genomics and neuroscience. This progress has been fueled recently by statistical machine learning libraries designed within a differentiable programming framework, *e.g.,* PyTorch (Paszke et al., 2019) and TensorFlow (Abadi et al., 2016).

As a learning system learns from a continuous, possibly evolving, data stream, it is desirable to supplement it with tools facilitating its monitoring 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) show the importance of such tools. In the former case, 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.

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 is relatively limited, while progress in learning systems has been flourishing.

In this paper, we introduce a generic change monitoring method called *autograd-test* based on statistical decision theory. This approach is aligned with current machine learning softwares developed in a differentiable programming framework, allowing us to seamlessly apply it to a large class of models implemented in such frameworks. Moreover, this method is equipped with a *scanning* procedure, allowing it to detect *weak changes* occurring on an unknown subset of model parameters.

Relationships to previous works. Change detection is a classical topic in statistics and signal processing; see (Basseville & Nikiforov, 1993; Tartakovsky et al., 2014) for a survey. It 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. In practice, an offline approach 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 in the distribution of data streams. We focus on the former one.

Test statistics for detecting changes in model parameters are usually designed on a case-by-case basis; see (Hinkley, 1970; Lorden, 1971; Deshayes & Picard, 1986; Basseville & Nikiforov, 1993; Carlstein et al., 1994; Csörgő & Horváth, 1997) and references therein. These methods are usually based on (possibly generalized) likelihood ratios or

[1]University of Washington [2]University of Montpellier. Correspondence to: Lang Liu <liu16@uw.edu>.

Presented at the ICML 2020 Workshop on Challenges in Deploying and Monitoring Machine Learning Systems.

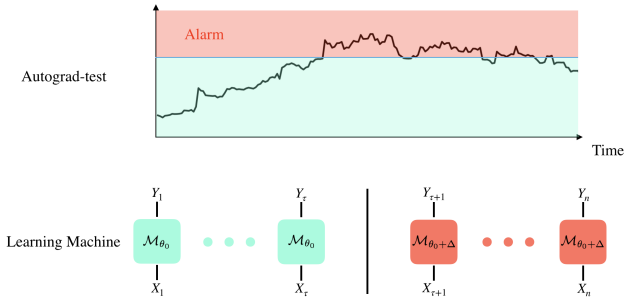

Figure 1: Illustration of monitoring a learning machine.

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.

2. Score-Based Change Detection

We first introduce the problem of change detection in model parameters. Let $W_{1:n} := \{W_k\}_{k=1}^n$ be a sequence of observations. Consider a family of machine learning models $\{\mathcal{M}_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ such that $W_k = \mathcal{M}_\theta(W_{1:k-1}) + \varepsilon_k$, where $\{\varepsilon_k\}_{k=1}^n$ are independent and identically distributed (*i.i.d.*) random noises. To learn this model from data, we choose some loss function L and estimate model parameters by solving the problem[1]:

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

This encompasses constrained empirical risk minimization (ERM) and constrained maximum likelihood estimation (MLE). For simplicity, we assume the model is *correctly specified*, *i.e.,* there exists a true value $\theta_0 \in \Theta$ from which the data are generated.

Under abnormal circumstances, this true value may not remain the same for all observations. Hence, we consider the model 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:

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

$\mathbf{H}_1$: after some time τ, θ_k jumps from θ_0 to $\theta_0 + \Delta$.

We focus on models whose loss $L(W_k, \mathcal{M}_\theta(W_{1:k-1}))$ can be rewritten as $-\log p_\theta(W_k|W_{1:k-1})$ for some conditional probability density p_θ. We refer to such a loss function as a *probabilistic loss*. For instance, the squared loss function is associated with the negative log-likelihood of a Gaussian density; for more examples, see *e.g.* (Murphy, 2012). For such models, the learning problem 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.

Algorithm 1 Autograd-test

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

Remark. Discriminative models also fit into this framework. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be *i.i.d.* observations, then the loss function reads $L(Y_k, \mathcal{M}_\theta(X_k))$. 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 $\tau \in [n-1]$ and $1 \leq s \leq t \leq n$, we define the conditional log-likelihood under the alternative as

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

We will write $\ell_{s:t}(\theta, \Delta)$ for short if there is no confusion. Under the null, we denote by $\ell_{s:t}(\theta) := \ell_{s:t}(\theta, 0; n)$ the conditional log-likelihood. 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 $\mathcal{I}_{s:t}(\theta) := -\nabla_\theta^2 \ell_{s:t}(\theta)$.

Given a hypothesis testing problem, the first step is to propose a *test statistic* R_n such that the larger R_n is, the less likely the null hypothesis is true. Then, for a prescribed *significance level* $\alpha \in (0, 1)$, we calibrate this test statistic by a threshold $r_0 := r_0(\alpha)$, leading to a test or decision rule[2] $\mathbb{1}\{R_n > r_0\}$. The threshold is chosen such that the *false alarm rate* or *type I error* is asymptotically controlled by α, *i.e.,* $\limsup_{n\to\infty} \mathbb{P}(R_n > r_0 \mid \mathbf{H}_0) \leq \alpha$. We say such a test is *consistent in level*. Moreover, we want the *detection 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 is *consistent in power*.

Let us follow this procedure to develop a test for Problem (P0). We start with assuming the changepoint τ is fixed. A

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

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

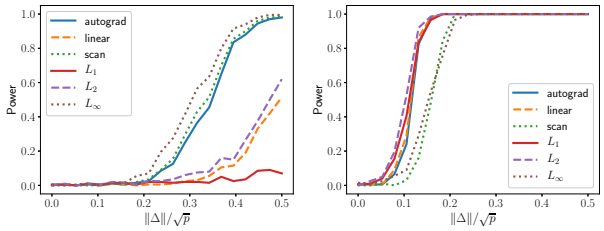

Figure 2: Power curves for a linear model with $d = 101$ (left: $p = 1$; right: $p = 20$). The sample size $n = 1000$.

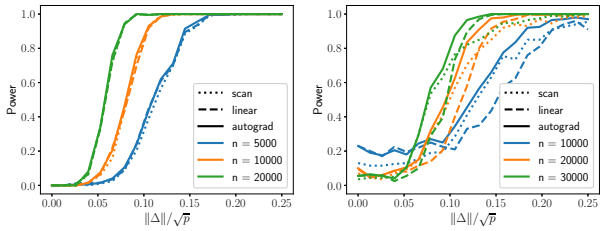

Figure 3: Power curves for a text topic model with $p = 1$ (left: $(N, M) = (3, 6)$; right: $(N, M) = (7, 20)$).

standard choice is the *generalized score statistic* given by

$$R_n(\tau) := S_{\tau+1:n}^\top(\hat{\theta}_n)\mathcal{I}_n(\hat{\theta}_n;\tau)^{-1}S_{\tau+1:n}(\hat{\theta}_n) \ , \quad (1)$$

where $\mathcal{I}_n(\hat{\theta}_n;\tau)$ is the *partial observed information w.r.t.* Δ (Wakefield, 2013, Chapter 2.9) defined as

$$\mathcal{I}_{\tau+1:n}(\hat{\theta}_n) - \mathcal{I}_{\tau+1:n}(\hat{\theta}_n)^\top\mathcal{I}_{1:n}(\hat{\theta}_n)^{-1}\mathcal{I}_{\tau+1:n}(\hat{\theta}_n). \quad (2)$$

To adapt to an unknown changepoint τ, a natural statistic is $R_{\text{lin}} := \max_{\tau\in[n-1]} R_n(\tau)$. And, given a significance level α, the decision rule reads $\psi_{\text{lin}}(\alpha) := \mathbb{1}\{R_{\text{lin}} > H_{\text{lin}}(\alpha)\}$, where $H_{\text{lin}}(\alpha)$ is a prescribed threshold discussed in Sec. 3. We call R_{lin} the *linear statistic* and ψ_{lin} the *linear test*.

Sparse alternatives. There are cases when the change only happens in a small subset of components of θ_0, *i.e.,* the change is weak. The linear test, which is built assuming the change is strong, may fail to detect such weak changes. Therefore, we consider *sparse alternatives*.

(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, the dimension of θ. We denote by T the changed components, in other words, $\Delta_{[d]\setminus T} = 0$.

For sparse alternatives, we consider a truncated statistic

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

Let $\mathcal{T}_p$ be the collection of all subsets of size p of $[d]$. To adapt to unknown T, we use

$$R_n(\tau, P; \alpha) := \max_{p\in[P]}\max_{T\in\mathcal{T}_p} H_p(\alpha)^{-1}R_n(\tau, T) \ . \quad (3)$$

Finally, since τ is unknown, we propose $R_{\text{scan}}(\alpha) := \max_{\tau\in[n-1]} R_n(\tau, P; \alpha)$ with decision rule $\psi_{\text{scan}}(\alpha) := \mathbb{1}\{R_{\text{scan}}(\alpha) > 1\}$. We call $R_{\text{scan}}(\alpha)$ the *scan statistic* and ψ_{scan} the *scan test*.

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

$$\psi(\alpha) := \max\{\psi_{\text{lin}}(\alpha_l), \psi_{\text{scan}}(\alpha_s)\} \quad (4)$$

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. We illustrate how to monitor a learning machine with *autograd-test* in Fig. 1.

Differentiable programming and component screening. An attractive feature of the *autograd-test* is that it only involves (second) derivatives of the log-likelihood function. That opens up an opportunity to build a library[3], based on automatic differentiation (AutoDiff), applicable to any machine learning models with a probabilistic loss designed within a differentiable programming framework. The algorithm to compute the *autograd-test* is presented in Alg. 1. For more implementational details, including the discussion on the computational cost, see Appendix A.

Another feature of score-based statistics is their flexibility to screen individual components or groups of components. If we know the change can only happen in some specific components, we may incorporate this prior information by imposing a constraint on the changed components T.

3. Level and Power

In this section we summarize the asymptotic behavior of the proposed score-based statistics under null and alternatives; for precise statements and proofs see Appendix C.

Proposition 1 (Null hypothesis)**.** *Under the null hypothesis and certain conditions, we have, for any $T \subset [d]$ and $\tau_n \in \mathbb{Z}_+$ such that $\tau_n/n \to \lambda \in (0, 1)$,*

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

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

[3] Available at: https://github.com/langliu95/autodetect.

[4] We use $q_D(\alpha)$ for the upper α-quantile of the distribution D.

Table 1: Decision of the scan test (each (row, column) pair stands for a concatenation; "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	N	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

Most conditions in Prop. 1 are standard. 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).

Proposition 2 (Fixed alternative hypothesis)**.** *Assume the observations are independent, and the alternative hypothesis is true with a fixed change parameter Δ. Let the change-point τ_n be such that $\tau_n/n \to \lambda \in (0,1)$. Under certain conditions, we have $\hat{\theta}_n \to_p \theta^*$ and the three proposed tests $\psi(\alpha), \psi_{lin}(\alpha), \psi_{scan}(\alpha)$ are consistent in power.*

4. Experiments

In this section, we apply our approach to detect changes on synthetic and real data. We summarize our experimental settings and findings here[5].

Synthetic experimental settings. For each model, we generate the first half sample from the pre-change parameter θ_0 and generate the second half from the post-change parameter θ_1, where θ_1 is obtained by adding δ to the first p components of θ_0. Next, we run the proposed tests to monitor the learning process, where the significance levels are set to be $\alpha = 2\alpha_l = 2\alpha_s = 0.05$ and the maximum cardinality $P = \lfloor \sqrt{d} \rfloor$. We repeat this procedure 200 times and approximate the detection power by rejection frequency. Finally, we plot the power curves by varying δ. Note that the value at $\delta = 0$ is the empirical false alarm rate.

Additive model. We consider a linear model with 101 parameters and investigate two sparsity levels, $p = 1$ and $p = 20$. We compare the proposed tests with three baselines given by the L_a norm of the score[6] for $a \in \{1, 2, \infty\}$. Note that the linear test corresponds to the L_2 norm with a proper normalization. And the scan test with $P = 1$ corresponds to the L_∞ norm. As shown in Fig. 2, when the change is sparse, the *autograd-test*, the scan and L_∞ tests share similar power curves and outperform the rest three significantly. When the change is less sparse, all tests' performance get improved, with the scan and L_∞ tests being less powerful than the other four. This empirically illustrates that (1) the scan and L_∞ tests work better in detecting sparse changes, (2) the linear, L_1 and L_2 tests are more powerful for non-sparse changes and (3) the *autograd-test* achieves comparable performance in both situations.

Text topic model. We consider a text topic model in (Stratos et al., 2015) and investigate the proposed tests under different sample sizes. This model is a hidden Markov model whose emission distribution has a special structure. We examine two parameter schemes: $(N, M) \in \{(3, 6), (7, 20)\}$, where N is the number of hidden states and M is the number of categories of the emission distribution. And p is set to be 1. As shown in Fig. 3, for the first scheme, all tests have small false alarm rate, and their power rises as the sample size increases. For the second scheme, the false alarm rate is out of control in the beginning, but this problem is alleviated as the sample size increases. This empirically verifies that our proposed tests are consistent in level and power even for dependent data.

Real data application. We collect subtitles of the first two seasons of four TV shows—Friends (F), Modern Family (M), the Sopranos (S), and Deadwood (D)—where the former two are viewed as "polite" and the latter two as "rude". For every pair of seasons, we concatenate them, and train the text topic model with $N = \lfloor \sqrt{n/100} \rfloor$ and M being the size of vocabulary built from the training corpus. The task is to detect changes in the rudeness level. As an analogy, the text topic model here corresponds to a chatbot, and subtitles are viewed as interactions with users. We want to know whether the conversation gets rude as the chatbot learns from the data.

The linear test does a perfect job in reporting shifts in rudeness level. However, it has an extremely high false alarm rate: 27 out of 32 are false alarms. There are two possible explanations: (1) the training data is not large enough, leading to a high false alarm rate as we have seen in Fig. 3; (2) the linear test captures the difference of two shows in other aspects, *e.g.,* the topic of the conversation.

The scan test has much lower false alarm rate (11/32). Moreover, as demonstrated in Table 1, there are only two false alarms in the most interesting case, where the sequence starts with a "polite" show. The results are promising since we benefit from exploiting the sparsity even without knowing which model components are related to the rudeness level.

[5] Due to space constraints more details and additional results are deferred to Appendix D.

[6] We generate samples from its limiting distribution and use the empirical quantiles as the thresholds. See Appendix D.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.

Basseville, M. and Nikiforov, I. V. *Detection of abrupt changes: theory and application*. Prentice Hall Information and System Sciences Series. Prentice Hall, Inc., Englewood Cliffs, NJ, 1993.

Bickel, P. J., Ritov, Y., and Rydén, T. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):1614–1635, 1998.

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

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

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

Carlstein, E. G., Müller, H.-G., and Siegmund, D. Change-point problems. IMS, 1994.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

Deshayes, J. and Picard, D. Off-line statistical analysis of change-point models using non parametric and likelihood methods. In Basseville, Michèleand Benveniste, A. (ed.), *Detection of Abrupt Changes in Signals and Dynamical Systems*, pp. 103–168, Berlin, Heidelberg, 1986. Springer Berlin Heidelberg. ISBN 978-3-540-39726-7.

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

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

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

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

Louis, T. A. Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 44 (2):226–233, 1982.

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

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. 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*, pp. 8024–8035, 2019.

Stratos, K., Collins, M., and Hsu, D. 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, pp. 1282–1291, 2015.

Tartakovsky, A., Nikiforov, I., and Basseville, M. *Sequential analysis: Hypothesis testing and changepoint detection*. Chapman and Hall/CRC, 2014.

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

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

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

Algorithm 2 Autograd-test

Input: data $(W_i)_{i=1}^n$, log-likelihood ℓ, MLE $\hat{\theta}_n$, thresholds α_l and α_s, and maximum cardinality P.
Initialization: $R_{\text{lin}} \leftarrow 0$ and $R_{\text{scan}}(\alpha) \leftarrow 0$.
Compute $H_p(\alpha_s)$ for each $p \in [P]$.
$S_{1:n}(\hat{\theta}_n) \leftarrow \nabla_\theta \ell_n(\hat{\theta}_n)$.
$\mathcal{I}_{1:n}(\hat{\theta}_n)^{-1} \leftarrow -\nabla^2_\theta \ell_n(\hat{\theta}_n)^{-1}$.
for $\tau = 1$ **to** $n-1$ **do**
  $S_{\tau+1:n}(\hat{\theta}_n) \leftarrow S_{1:n}(\hat{\theta}_n) - S_{1:\tau}(\hat{\theta}_n)$.
  $\mathcal{I}_{\tau+1:n}(\hat{\theta}_n) \leftarrow \mathcal{I}_{1:n}(\hat{\theta}_n) - \mathcal{I}_{1:\tau}(\hat{\theta}_n)$.
  Compute $\mathcal{I}_n(\hat{\theta}_n; \tau)$ by (2).
  Compute $R_n(\tau)$ by (1).
  $R_{\text{lin}} \leftarrow \max\{R_{\text{lin}}, R_n(\tau)\}$.
  $v(\tau) \leftarrow \text{Sort}(\text{diag}(\mathcal{I}_n(\hat{\theta}_n; \tau))^{-1} S^{\odot 2}_{\tau+1:n}(\hat{\theta}_n))$.
  for $p \in [P]$ **do**
    $T \leftarrow$ indices of the largest p components of $v(\tau)$.
    Compute $R_n(\tau, T)$.
    $R_{\text{scan}}(\alpha_s) \leftarrow \max\{\frac{R_n(\tau,T_p)}{H_p(\alpha_s)}, R_{\text{scan}}(\alpha_s)\}$.
  end for
end for
Compute $\psi_{\text{lin}}(\alpha_l)$ and $\psi_{\text{scan}}(\alpha_s)$.
Output: $\psi(\alpha) \leftarrow \max\{\psi_{\text{lin}}(\alpha_l), \psi_{\text{scan}}(\alpha_s)\}$.

A. Implementational details

The full version of *autograd-test* is presented in Alg. 2. 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 $T_p := \arg\max_{T \in \mathcal{T}_p} R_n(\tau, T)$ by the indices of the largest p components in

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

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 $\|S_{\tau+1:n}(\hat{\theta}_n)\|^2$; see Lemma 1 for a justification. In Step 12, we first sort $v(\tau)$, then for each $p \leq P$, we obtain from $v(\tau)$ an approximate maximizer $\widetilde{T}_p$ and approximate $R_n(\tau, P; \alpha_s)$ by $\max_{p \in [P]} H_p(\alpha_s)^{-1} R_n(\tau, \widetilde{T}_p)$.

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

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

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

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

On the one hand, note that

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

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

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

On the other hand, we can obtain, similarly,

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

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

Therefore, we have

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

■

An attractive feature of score-based statistics in the age of differentiable programming is their straightforward computation using automatic differentiation (AutoDiff). 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 its overall computational complexity. Before the for loop in step 6, the most expensive step is step 5. To obtain the Hessian matrix, we may call the AutoDiff procedure d times on the gradient obtained in step 5, and compute one row of the Hessian at each time, which takes time $\mathcal{O}(nd^2)$. Calculating the inverse of the Hessian costs at most $\mathcal{O}(d^3)$, so the overall cost of step 5 is $\mathcal{O}(nd^2+d^3)$. For each $\tau \in [n-1]$, steps 7-11 have complexities $\mathcal{O}(\tau d)$, $\mathcal{O}(\tau d^2)$, $\mathcal{O}(d^2)$, $\mathcal{O}(d^3)$, and $\mathcal{O}(1)$, respectively. In step 12 we may use any sort algorithm such as *Quicksort*, and the time complexity is $\mathcal{O}(d \log d)$. The loop over $p \in [P]$ in step 13 costs at most $\mathcal{O}(P^4)$ time because evaluating $R_n(\tau, T)$ has complexity $\mathcal{O}(|T|^3)$. Hence, steps 6-18 takes time $\mathcal{O}(n^2d^2+nd^3+nP^4)$, dominating steps 2-5. To summarize, the overall computational complexity of Alg. 1 is $\mathcal{O}(n^2d^2+nd^3+nP^4)$. 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 $\mathcal{O}(n^2d^2 + nd^3)$. Moreover, 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 $\mathcal{O}(n^2d)$ and $\mathcal{O}(n^2d^2)$ to $\mathcal{O}(nd)$ and $\mathcal{O}(nd^2)$, respectively. This gives an algorithm of complexity $\mathcal{O}(nd^3)$.

Next, we illustrate the automatic differentiation feature of the *autograd-test* on several models we consider in the experiments.

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* $(X^\top\theta, Y -$

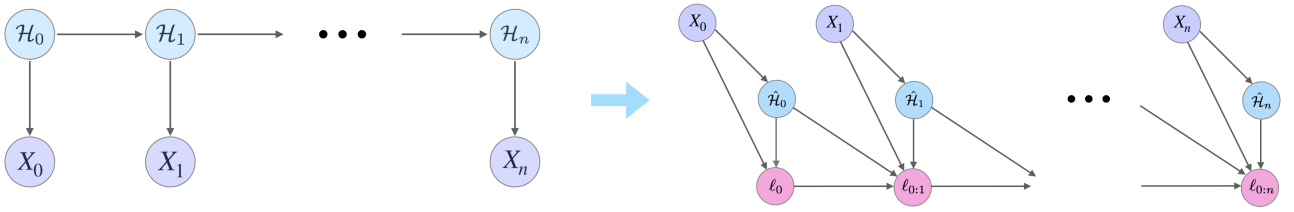

Figure 4: Graphical representations of text topic models (left: probabilistic graphical model; right: computational graph).

$X^\top\theta), (Y, Y - X^\top\theta)$, *and* $(Y - X^\top\theta, \ell(\theta))$. *Thus, computing the 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* $\mathcal{O}(nd)$ *with 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:*

$$X_t = \sum_{i=1}^{r} \phi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \varphi_i \varepsilon_{t-i}, \tag{5}$$

where $\{\varepsilon_t\}$ *are* i.i.d. *standard norm random variables. Let* $\theta = (\phi; \varphi)$, *and assume further that* $r \geq q$ *and* $X_{1:r}$ *is completely known,* i.e., *the model* (5) *starts from* $t = r + 1$. *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 a computational graph as follows: let* $\{\phi_i\}_{i\in[r]}, \{\varphi_i\}_{i\in[q]}, \{X_t\}_{t\in[n]}, \{\varepsilon_t\}_{t\in[n]\setminus[r]}$, *and* $\ell_n(\hat{\theta}_n)$ *be vertices with edges* $\{(V, \varepsilon_t) : V \in \{X_{(t-r):t}, \varepsilon_{(t-q):(t-1)}, \phi_{1:r}, \varphi_{1:q}\}\}_{t\in[n]\setminus[r]}$ *and* $\{(\varepsilon_t, \ell_n(\hat{\theta}_n))\}_{t\in[n]\setminus[r]}$. *As a result,* $|V| + |E| = \mathcal{O}(nd)$ *in which* $d = r + q$, *and thus computing* $S_{1:n}(\hat{\theta}_n)$ *costs* $\mathcal{O}(nd)$ *time. In order to compute* $\{S_{1:\tau}(\hat{\theta}_n)\}_{\tau\in[n-1]}$, *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* $\{\nabla_\theta \varepsilon_t\}_{t\in[n]\setminus[r]}$ *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 if the observations are not independent, we are able to reduce the time complexity from* $\mathcal{O}(n^2 d^2 + nd^3)$ *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* $[M]$, *satisfying 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* $h \neq \mathcal{H}(X)$. *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* $\ell_n(\theta)$ *contains vertices* $\{\theta_i\}_{i\in[d]}$ *and* $\ell_n(\theta)$, *and edges arise from two sources: 1) the constraints among "parameters",* e.g., $q(N|j) = 1 - \sum_{i=1}^{N-1} q(i|j)$ *for each* $j \in [N]$; *2) the*

computation of $\ell_n(\theta)$. As a consequence, $|V| + |E|$ is at most $\mathcal{O}(d+n)$. Again, $S_{1:\tau}(\hat{\theta}_n)$ admits a recursion

$$S_{1:\tau}(\hat{\theta}_n) = S_{1:(\tau-1)}(\hat{\theta}_n) + \nabla_\theta[\log q(\hat{\mathcal{H}}_\tau|\hat{\mathcal{H}}_{\tau-1}) + \log g(X_\tau|\hat{\mathcal{H}}_\tau)], \tag{6}$$

and thus computing $\{S_{1:\tau}(\hat{\theta}_n)\}_{\tau\in[n]}$ takes $\mathcal{O}(nd)$ time. If we harness the sparsity in the last term in (6), we can further reduce this complexity to $\mathcal{O}(n+d)$. See Fig. 4 for a graphical representation.

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.

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 $s \le t \le k$, 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):

$$\nabla_\theta \ell(\theta) = \int [\nabla_\theta \log f(x;\theta)] p(x;\theta)\lambda(dx) \tag{7}$$

$$\nabla_\theta^2 \ell(\theta) + [\nabla_\theta \ell(\theta)][\nabla_\theta \ell(\theta)]^\top = \int \left\{ \nabla_\theta^2 \log f(x;\theta) + [\nabla_\theta \log f(x;\theta)][\nabla_\theta \log f(x;\theta)]^\top \right\} p(x;\theta)\lambda(dx) \ , \tag{8}$$

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 $k = 1,\dots,n$, 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,

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

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

Recall that MLE represents maximum likelihood estimator (or empirical risk minimizer).

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 $-\nabla^2_\theta \ell_n(\hat{\theta}_n)/n \to_p \mathcal{I}_0$. It follows that

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

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

$$-\frac{1}{\sqrt{n}}\nabla_\theta Z_n(\theta_n^*) = -\frac{1}{n}\nabla^2_\theta \ell_n(\theta_n^*) = \mathcal{I}_0 + o_p(1) \tag{9}$$

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

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1) \ . \tag{10}$$

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

$$\begin{aligned}
\frac{S_{\tau_n+1:n}(\hat{\theta}_n)}{\sqrt{n-\tau_n}} &= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \frac{1}{\sqrt{n-\tau_n}}\nabla_\theta S^\top_{\tau_n+1:n}(\theta_n^*)(\hat{\theta}_n - \theta_0) \\
&= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \frac{(\nabla_\theta S_n(\theta_n^*) - \nabla_\theta S_{\tau_n}(\theta_n^*))^\top}{\sqrt{n(n-\tau_n)}}\sqrt{n}(\hat{\theta}_n - \theta_0) \\
&= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \left[\sqrt{\frac{n}{n-\tau_n}}\frac{1}{\sqrt{n}}Z_n(\theta_n^*) - \frac{\tau_n}{\sqrt{n(n-\tau_n)}}\frac{1}{\sqrt{\tau_n}}Z_{\tau_n}(\theta_n^*)\right]^\top (\mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1)) \quad \text{by (10)} \\
&= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \left(\frac{\lambda}{\sqrt{1-\lambda}} - \sqrt{\frac{1}{1-\lambda}}\right)\mathcal{I}_0\mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1) \quad \text{by (9)} \\
&= -\sqrt{\frac{\tau_n}{n-\tau_n}}Z_{\tau_n}(\theta_0) + \sqrt{\frac{n}{n-\tau_n}}Z_n(\theta_0) + \frac{\lambda-1}{\sqrt{1-\lambda}}Z_n(\theta_0) + o_p(1) \\
&= -\frac{\sqrt{\lambda}}{\sqrt{1-\lambda}}Z_{\tau_n}(\theta_0) + \frac{\lambda}{\sqrt{1-\lambda}}Z_n(\theta_0) + o_p(1) \ .
\end{aligned}$$

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 $\forall k\in\mathbb{Z}$

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

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 $\alpha\in\mathbb{R}^d$,

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^n\mathbb{E}\left[\alpha^\top\nabla_\theta\ell_k(\theta_0)\mathbb{1}\left(\left|\alpha^\top\nabla_\theta\ell_k(\theta_0)\right|>\sqrt{n}\varepsilon\right)\right]\\&=\mathbb{E}\left[\alpha^\top\nabla_\theta\ell_1(\theta_0)\mathbb{1}\left(\left|\alpha^\top\nabla_\theta\ell_1(\theta_0)\right|>\sqrt{n}\varepsilon\right)\right]\to0\ ,\end{aligned}$$

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*

$$\begin{pmatrix}Z_{\tau_n}\sqrt{\tau_n/n}\\Z_n\end{pmatrix}\to_d\mathcal{N}\left(0,\begin{pmatrix}\lambda\mathcal{I}_0&\lambda\mathcal{I}_0\\\lambda\mathcal{I}_0&\mathcal{I}_0\end{pmatrix}\right)\ . \tag{11}$$

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 $(a^\top,b^\top)\in\mathbb{R}^{2d}$,

$$a^\top\sqrt{\frac{\tau_n}{n}}Z_{\tau_n}+b^\top Z_n\to_d\mathcal{N}\left(0,\lambda(a+b)^\top\mathcal{I}_0(a+b)+(1-\lambda)b^\top\mathcal{I}_0b\right),\text{as }n\to\infty\ .$$

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 (11) holds by invoking the Lindeberg theorem for martingales, and it follows that

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

■

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[7]. 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

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

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

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

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

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

Proposition 2. *Given an independent sample $W_1, \ldots, W_n$ 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 $W_1, \ldots, W_{\tau_n} \sim p_{\theta_0}$, $W_{\tau_n+1}, \ldots, W_n \sim p_{\theta_1} (\theta_1 \neq \theta_0)$, 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.*

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

$$\tfrac{1}{n} R_n(\tau_n) \to_p (\overline{\lambda} S_1^*)^\top (\mathcal{I}^*)^{-1} (\overline{\lambda} S_1^*) \ , \tag{12}$$

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 $B_\varepsilon = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\| \leq \varepsilon\} \subset \Theta^*$ and $\text{bd}(B_\varepsilon)$ be the boundary of B_ε. We will show that, for sufficiently small ε,

$$\mathbb{P}\left(\ell_n(\theta) < \ell_n(\theta^*), \forall \theta \in \text{bd}(B_\varepsilon)\right) \to 1 \ . \tag{13}$$

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

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

where $\overline{\theta}_n \in B_\varepsilon$ satisfies $\left\|\overline{\theta}_n - \theta^*\right\| \leq \|\theta - \theta^*\|$. 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,

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

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 $\theta \in \mathrm{bd}(B_\varepsilon)$,

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

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 (13) holds.

Now according to continuous mapping theorem and Slutsky's theorem (see, for instance (Billingsley, 2008)) and to Eq. (2)

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

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 $\alpha \in (0,1)$,

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

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

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

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 $\mathbb{P}(\psi_{\text{scan}}(\alpha) = 1) \to 1$, 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 $\mathbb{P}^{(\tau_n)}_{\theta_0,\theta_n}$. 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

$$\frac{n}{\tau_n(n-\tau_n)} \mathcal{I}_n(\hat{\theta}_n; \tau_n) \to_p \mathcal{I}_0 \ , \tag{14}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}_d\left(\overline{\lambda} h, \mathcal{I}_0^{-1}\right) \ , \tag{15}$$

$$\text{and} \quad \sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{\tau_n+1:n}(\hat{\theta}_n) \to_d \mathcal{N}_d(\sqrt{\lambda\overline{\lambda}}\, \mathcal{I}_0 h, \mathcal{I}_0) \ . \tag{16}$$

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 (14) 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

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

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 $\mathbb{P}_{\theta_0,\theta_n}^{(\tau_n)}$,

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

and by the Cramér-Wold device, the statement (15) 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 $\mathbb{P}_{\theta_0,\theta_n}^{(\tau_n)}$,

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

which yields (16). Now, the asymptotic distributions of $R_n(\tau_n)$ and $R_n(\tau_n, T)$ follows immediately from the continuous mapping theorem. ■

D. 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 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.

Baselines. Recall from Proposition 1 in Appendix C that

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

Since $\|\cdot\|_a : \mathbb{R}^d \to \mathbb{R}_+$ is continuous for fixed d, we know, by the continuous mapping theorem, that

$$\left\| \sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{\tau_n+1:n}(\hat{\theta}_n) \right\|_a \to_d \|Z\|_a \quad \text{for } Z \sim \mathcal{N}(0, \mathcal{I}_0).$$

Hence, the L_a test is given by

$$\psi_{L_a}(\alpha) := \mathbb{1}\left\{ \max_{\tau \in [n-1]} \left\| \sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{\tau_n+1:n}(\hat{\theta}_n) \right\|_a > q_{\|Z\|_a}(\alpha/n) \right\}.$$

We do not have a closed form formula for the distribution of $\|Z\|_a$, so we need to estimate $q_{\|Z\|_a}(\alpha/n)$ by a Monte Carlo method: (1) generate m *i.i.d.* samples $Z_1, \ldots, Z_m \sim \mathcal{N}\left(0, \frac{n}{\tau_n(n-\tau_n)} \mathcal{I}_n(\hat{\theta}_n; \tau)\right)$, (2) compute $\mathcal{Z} := \{\|Z_1\|_a, \ldots, \|Z_m\|_a\}$, and (3) estimate $q_{\|Z\|_a}(\alpha/n)$ by the upper (α/n)-quantile of $\mathcal{Z}$. Note that this approach imposes additional computational burden and tends to under-estimate the threshold if m is not large enough.

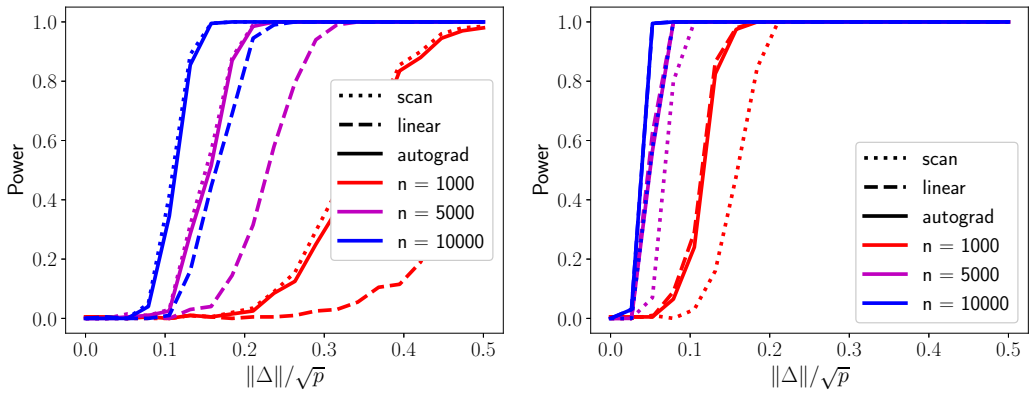

Figure 5: Power versus magnitude of change for linear models (left: $p = 1$; right: $p = 20$).

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. 5, 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. 6 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 then 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 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. 7.

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

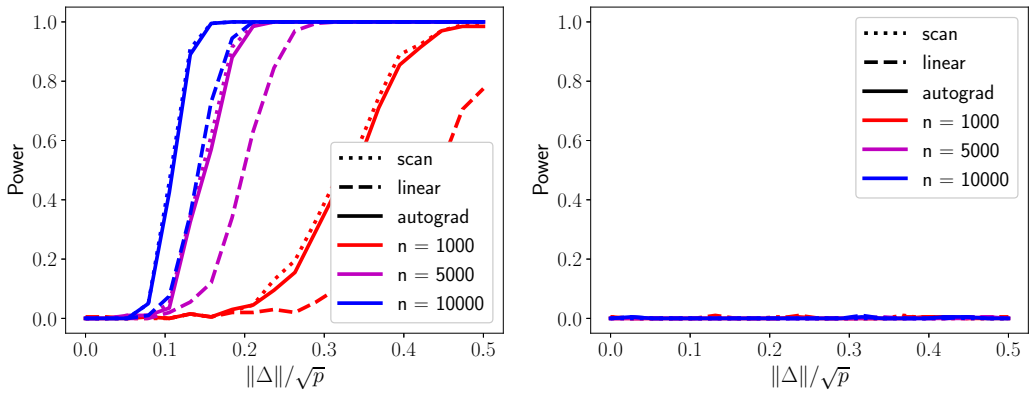

Figure 6: Power versus magnitude of change for linear regression with restriction (left: contains the changed component; right: does not contain the changed component).

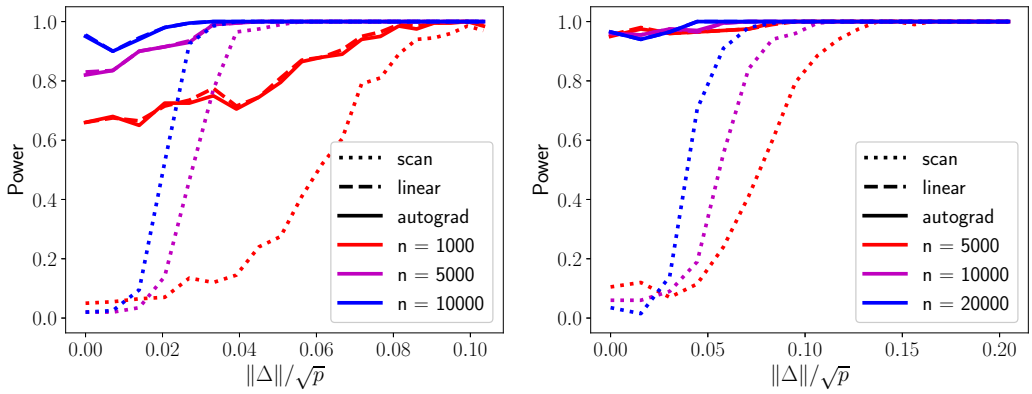

Figure 7: Power versus magnitude of change for ARMA$(3, 2)$ (left) and ARMA$(6, 5)$ (right).

coefficients. This results in ill-conditioned partial information (2) and subsequent unstable computation of the linear statistic. On the contrary, the scan statistic only inverts the submatrix of size $p \times p$. Since $p \leq \sqrt{d}$, 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 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. 8, 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. 9. 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

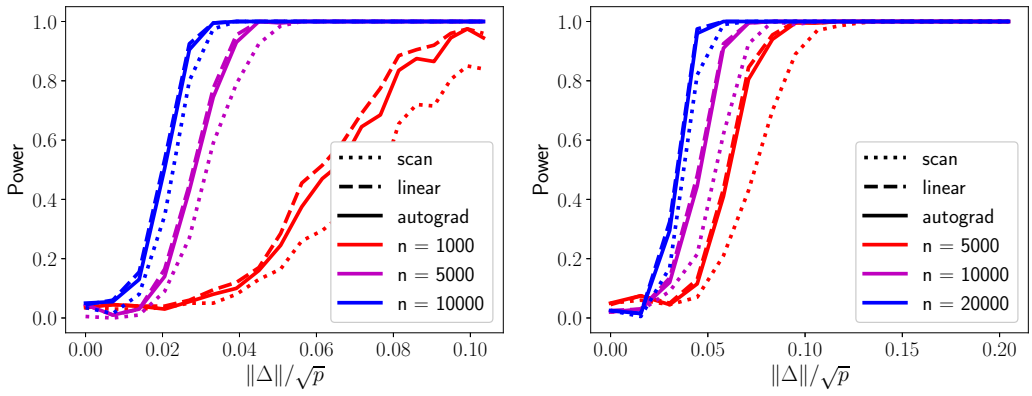

Figure 8: Power versus magnitude of change for ARMA models with restricted components (left: ARMA$(3, 2)$; right: ARMA$(6, 5)$).

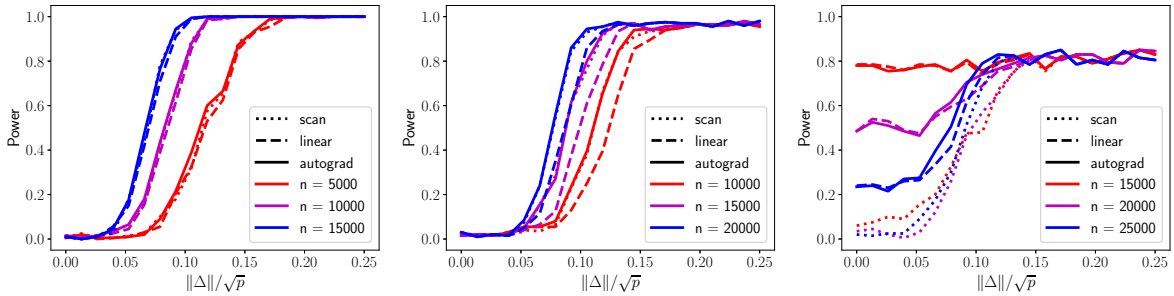

Figure 9: Power versus magnitude of change for HMMs with N hidden states (left: $N = 3$; middle: $N = 7$; right: $N = 15$).

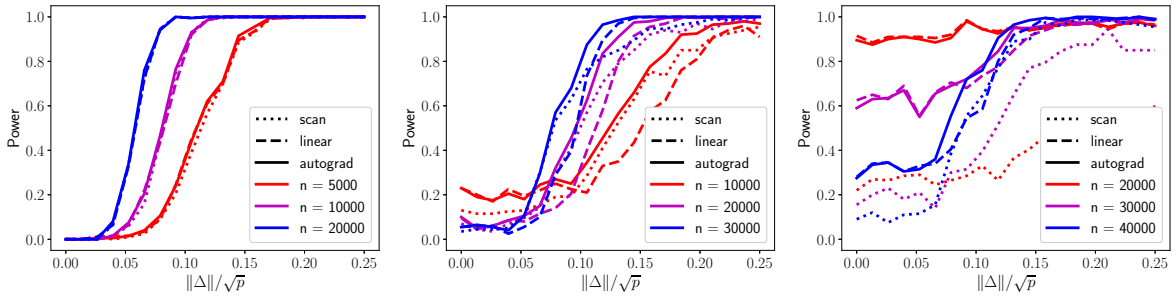

Figure 10: Power versus magnitude of change for text topic models (left: $(N, M) = (3, 6)$; middle: $(N, M) = (7, 20)$; right: $(N, M) = (15, 150)$).

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 emission distribution. We use the same way as HMMs to generate the transition matrix. The emission matrix is sampled analogically except that each row can only have one nonzero entry, which is required by this model. As shown in Fig. 10, the results are similar to the ones for HMMs. Since the topic model has more parameters than the HMM with the same N, the inconsistency in level is exacerbated.

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)[8]—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 $\lceil n/4 \rceil$ to $\lceil 3n/4 \rceil$) 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[9]. 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.

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

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

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	N	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

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

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.