Score-Based Change Detection for Gradient-Based Learning Machines

Lang Liu¹ Joseph Salmon² Zaid Harchaoui¹

¹ Department of Statistics, University of Washington, Seattle ² IMAG, University of Montpellier, CNRS, Montpellier

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 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 to achieve a prescribed false alarm rate. We illustrate the versatility of the approach on synthetic and real data.

1 Introduction

Statistical machine learning models are 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 such as PyTorch [19] and TensorFlow [1].

Gradient-based optimization algorithms such as accelerated batch 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 in order to prevent the learned model from experiencing abnormal changes.

Recent remarkable failures of intelligent learning systems such as Microsoft's chatbot [17] and Uber's self-driving car [15] 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 can potentially prevent a debacle by triggering an early alarm, drawing the attention of its designers and engineers to an abnormal change of a 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.

We introduce a generic change monitoring method called *auto-test* based on statistical decision theory. This approach is aligned with machine learning libraries 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, enabling it to detect *small jumps* occurring on an unknown subset of model parameters. The proofs and more details can be found in Appendix. The code is publicly available at *github.com/langliu95/autodetect*.



Figure 1: Illustration of monitoring a learning machine with *auto-test*.

Previous work on change detection. Change detection is a classical topic in statistics and signal processing; see [2, 21] 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. Depending on the type of change, the change detection problem can be classified into two main categories: change in the model parameters [13, 10] and change in the distribution of data streams [16, 14, 9]. We focus on testing the presence of a change in the model parameters.

Test statistics for detecting changes in model parameters are usually designed on a case-by-case basis; see [2, 7, 25, 8, 12] 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 *large jumps*, *i.e.*, changes occurring simultaneously on all model parameters, in contrast to ours.

2 Score-Based Change Detection

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 a loss function L and estimate model parameters by solving the problem:

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

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 allow a potential parameter change in the model, that is, $\theta = \theta_k$ may evolve over time:

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

A time point $\tau \in [n-1] := \{1, \ldots, n-1\}$ 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 say that there is a jump (or change) in the data sequence if such a changepoint exists. We aim to determine if there exists a jump in this sequence, which we formalize as a hypothesis testing problem.

(P0) Testing the presence of a jump

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

We focus on models whose loss $L(W_k, \mathcal{M}_{\theta}(W_{1:k-1}))$ can be written as $-\log p_{\theta}(W_k|W_{1:k-1})$ for some conditional probability density p_{θ} . For instance, the squared loss function is associated with the negative log-likelihood of a Gaussian density; for more examples, see, *e.g.*, [18]. In the remainder of the paper, we will work with this probabilistic formulation for convenience, and we refer to the corresponding loss as the probabilistic loss.

Remark. Discriminative models can also fit into this framework. Let $\{(X_i, Y_i)\}_{i=1}^n$ be *i.i.d.* observations, then the loss function reads $L(Y_k, \mathcal{M}_{\theta}(X_k))$. If, in addition, L is a probabilistic loss, then the associated conditional probability density is $p_{\theta}(Y_k|X_k)$.

2.1 Likelihood score and score-based testing

Let $1{\cdot}$ 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 $\mathbb{1}\{R_n > r_0\}$, *i.e.*, we reject the null if $R_n > r_0$. The threshold is chosen such that the *false alarm rate* or *type I error rate* is asymptotically controlled by α , *i.e.*, lim $\sup_{n\to\infty} \mathbb{P}(R_n > r_0 \mid \mathbf{H}_0) \leq \alpha$. We say that 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 design a test for Problem (P0). We start with the case when the changepoint τ is fixed. A standard choice is the *generalized score statistic* given by

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

where $\mathcal{I}_n(\hat{\theta}_n; \tau)$ is the partial observed information w.r.t. Δ [24, Chapter 2.9] defined as

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

$$\tag{2}$$

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*.

2.2 Sparse alternatives

There are cases when the jump only happens in a small subset of components of θ_0 . The linear test, which is built assuming the jump is large, may fail to detect such small jumps. Therefore, we also consider *sparse alternatives*.

(P1) Testing the presence of a small jump:

Algorithm 1 Auto-test

1: Input: data $(W_i)_{i=1}^n$, log-likelihood ℓ , levels α_l and α_s , and maximum cardinality P.

2: for $\tau = 1$ to n - 1 do

- 3: Compute $R_n(\tau)$ in (1) using AutoDiff.
- 4: Compute $R_n(\tau, P; \alpha_s)$ in (3).
- 5: end for

6: **Output:** $\psi_{\text{auto}}(\alpha) = \max\{\psi_{\text{lin}}(\alpha_l), \psi_{\text{scan}}(\alpha_s)\}$ in (4).

$$\begin{split} \mathbf{H}_{0} &: \theta_{k} = \theta_{0} \text{ for all } k = 1, \dots, n \\ \mathbf{H}_{1} &: \text{after some time } \tau, \, \theta_{k} \text{ jumps from } \theta_{0} \text{ to } \theta_{0} + \Delta, \\ &\text{where } \Delta \text{ has at most } P \text{ nonzero entries }. \end{split}$$

Here P is referred to as the maximum cardinality, which is set to be much smaller than d, the dimension of θ . We denote by T the changed components, *i.e.*, $\Delta_T \neq 0$ and $\Delta_{[d]\setminus T} = 0$.

Given a fixed T, we consider the *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

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

where we use a different threshold $H_p(\alpha)$ for each $p \in [P]$. Finally, since τ is also 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 combine the respective strengths of these two tests, we consider the test

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

with $\alpha_l + \alpha_s = \alpha$, and we refer to it as the *auto-test*. The choice of α_l and α_s should be based on prior knowledge regarding how likely the jump is small. We illustrate how to monitor a learning machine with *auto-test* in Fig. 1.

2.3 Differentiable programming

An attractive feature of *auto-test* is that it can be computed by inverse-Hessian-vector products. That opens up the possibility to implement it easily using a machine learning library designed within a differentiable programming framework. Indeed, the inverse-Hessian-vector product can then be efficiently computed via automatic differentiation; see Appendix A for more details. The algorithm to compute the *auto-test* is presented in Alg. 1.

3 Level and Power

We summarize the asymptotic behavior of the proposed score-based statistics under null and alternatives. The precise statements and proofs can be found in Appendix B.

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

$$R_n(\tau_n) \rightarrow_d \chi_d^2$$
 and $R_n(\tau_n, T) \rightarrow_d \chi_{|T|}^2$,



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



Figure 3: Power curves of the *auto-test* for a text topic model with p = 1 (left: (N, M) = (3, 6); right: (N, M) = (7, 20)).

where we denote by \rightarrow_d the convergence in distribution. In particular, with thresholds $H_{lin}(\alpha) = q_{\chi^2_d}(\alpha/n)$ and $H_p(\alpha) = q_{\chi^2_p}(\alpha/[\binom{d}{p}n(p+1)^2])$, the tests $\psi_{lin}(\alpha)$, $\psi_{scan}(\alpha)$ and $\psi_{auto}(\alpha)$ are consistent in level, where $q_D(\alpha)$ is the upper α -quantile of the distribution D.

Most conditions in the above Proposition are standard. In fact, under suitable regularity conditions, they hold true for *i.i.d.* models, hidden Markov models [3, Chapter 12], and stationary autoregressive moving-average models [11, Chapter 13].

The next proposition verifies the consistency in power of the proposed tests under fixed alternatives.

Proposition (Fixed alternative hypothesis). Assume the observations are independent, and the alternative hypothesis is true with a fixed change parameter Δ . Let the changepoint τ_n be such that $\tau_n/n \to \lambda \in (0,1)$. Under certain conditions, the tests $\psi_{lin}(\alpha)$, $\psi_{scan}(\alpha)$ and $\psi_{auto}(\alpha)$ are consistent in power.

4 Experiments

We apply our approach to detect changes on synthetic data and on real data. We summarize the settings and our findings. More details and additional results are deferred to Appendix C.

Synthetic data. 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 *auto-test* 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 *auto-test* with three baselines given by the L_a norm of the score function for $a \in \{1, 2, \infty\}$, where these baselines are calibrated by the empirical quantiles of their limiting distributions. 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, *i.e.*, a small jump, the *auto-test* and L_{∞} test have similar power curves and outperform the rest of the tests significantly. When the change is less sparse, *i.e.*, a large jump, all tests' performance gets improved, with the L_{∞} test being less powerful than the other three. This empirically illustrates that (1) the L_{∞} test work better in detecting sparse changes, (2) the L_1 test and the L_2 test are more powerful for non-sparse changes and (3) the *auto-test* achieves comparable performance in both situations.

The proposed *auto-test* is calibrated by its large sample properties and the Bonferroni correction. This strategy tends to result in tests that are too conservative, with empirical false alarm rates largely below 0.05. We also use resampling-based strategy to calibrate the *auto-test*, *i.e.*, generating bootstrap samples and calibrating the test using the quantiles of the test statistics evaluated on bootstrap samples. The empirical false alarm rates are around 0.065 for both p = 1 and p = 20.

Table 1: Decision of the scan test on the TV-show application: each (row, column) pair stands for a concatenation; "R" means reject and "N" means not reject. Red entries are false alarms.

	F1	F2	M1	M2	S1	S2	D1	D2
F1	Ν	Ν	Ν	Ν	R	R	R	R
F2	Ν	Ν	R	Ν	R	R	R	R
M1	Ν	R	Ν	Ν	R	R	R	\mathbf{R}
M2	Ν	Ν	Ν	Ν	R	R	R	\mathbf{R}
S1	R	R	R	R	Ν	Ν	R	R
S2	R	R	R	R	Ν	Ν	R	R
D1	R	R	R	R	R	R	Ν	R
D2	\mathbf{R}	R	R	R	R	R	Ν	Ν

Text topic model. We consider a text topic model [20] and investigate the *auto-test* for 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 demonstrated in Fig. 3, for the first scheme, all tests have small false alarm rates, 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 the *auto-test* is consistent in both level and power even for dependent data.

Real data. 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, *i.e.*, the *auto-test* with $\alpha_l = \alpha$ and $\alpha_s = 0$, does a perfect job in reporting shifts in rudeness level. However, it has a high false alarm rate (27/32). This is expected since the linear test may capture the difference in other aspects, *e.g.*, topics of the conversation. The scan test, *i.e.*, the *auto-test* with $\alpha_l = 0$ and $\alpha_s = \alpha$, has much lower false alarm rate (11/32). Moreover, as shown in Table 1, there are only two false alarms in the most interesting case, where the sequence starts with a polite show. We note that this problem is hard, since rudeness is not the only factor that contributes to the difference between two shows. The results are promising since we benefit from exploiting the sparsity even without knowing which model components are related to the rudeness level.

5 Conclusion

We introduced a change monitoring method called *auto-test* that is well suited to machine learning models implemented within a differentiable programming framework. The experimental results show that the calibration of the test statistic based on our theoretical arguments brings about change detection test that can capture small jumps in the parameters of various machine learning models in a wide range of statistical regimes. The extension of this approach to penalized maximum likelihood or regularized empirical risk estimation in a high dimensional setting is an interesting venue for future work.

Acknowledgments

This work was supported by NSF DMS 2023166, DMS 1839371, DMS 1810975, CCF 2019844, CCF-1740551, CIFAR-LMB, and research awards.

References

- M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.
- [2] M. Basseville and I. Nikiforov. Detection of Abrupt Changes: Theory and Application. Prentice Hall, Inc., 1993.
- [3] P. Bickel, Y. Ritov, and T. Rydén. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. Annals of Statistics, 26(4), 1998.
- [4] P. Billingsley. Probability and measure. John Wiley & Sons, 2008.
- [5] L. Birgé. An alternative point of view on Lepski's method. Lecture Notes-Monograph Series, 36:113–133, 2001.
- [6] O. Cappé, E. Moulines, and T. Ryden. Inference in Hidden Markov Models. Springer-Verlag New York, 1st edition, 2005.
- [7] E. Carlstein, H. Müller, and D. Siegmund. *Change-point problems*. Institute of Mathematical Statistics, 1994.
- [8] M. Csörgő and L. Horváth. Limit Theorems in Change-Point Analysis. Wiley Series in Probability and Statistics. Wiley, 1997.
- [9] J. Cunningham, Z. Ghahramani, and C. Rasmussen. Gaussian processes for time-marked time-series data. In AISTATS, 2012.
- [10] J. Deshayes and D. Picard. Off-line statistical analysis of change-point models using non parametric and likelihood methods. In *Detection of Abrupt Changes in Signals and Dynamical Systems*. Springer, 1985.
- [11] R. Douc, E. Moulines, and D. Stoffer. Nonlinear Time Series: Theory, Methods and Applications with R Examples. Chapman and Hall/CRC, 2014.
- [12] F. Enikeeva and Z. Harchaoui. High-dimensional change-point detection under sparse alternatives. Annals of Statistics, 47(4), 2019.
- [13] D. Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1), 1970.
- [14] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In VLDB, 2004.
- [15] W. Knight. A self-driving Uber has killed a pedestrian in Arizona. *Ethical Tech*, March 2018.
- [16] G. Lorden. Procedures for reacting to a change in distribution. The Annals of Mathematical Statistics, 42(6), 1971.
- [17] R. Metz. Microsoft's neo-Nazi sexbot was a great lesson for makers of AI assistants. Artificial Intelligence, March 2018.
- [18] K. Murphy. Machine learning: A Probabilistic Perspective. MIT press, 2012.

- [19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.
- [20] K. Stratos, M. Collins, and D. Hsu. Model-based word embeddings from decompositions of count matrices. In ACL-IJCNLP, volume 1, 2015.
- [21] A. Tartakovsky, I. Nikiforov, and M. Basseville. Sequential Analysis: Hypothesis Testing and Changepoint Detection. Taylor & Francis, 2014.
- [22] A. W. van der Vaart. Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.
- [23] A. W. Van der Vaart. Lecture notes in time series. Universiteit Leiden, 2013.
- [24] J. Wakefield. Bayesian and Frequentist Regression Methods. Mathematics and Statistics. Springer, 2013.
- [25] Q. Zhang, M. Basseville, and A. Benveniste. Early warning of slight changes in systems. Automatica, 30(1), 1994. Special issue on statistical signal processing and control.

Outline of Appendix

The outline of the appendix is as follows. Appendix A discusses implementation details of the proposed test and its complexity analysis. Appendix B is devoted to prove the level and power consistency of the *auto-test*. Appendix C provides addition experiment results on a times series model and a hidden Markov model.

A Implementation Details

For simplicity of the notation, we write $\hat{S}_{i:j} = S_{i:j}(\hat{\theta}_n)$ and $\hat{\mathcal{I}}_{i:j} = \mathcal{I}_{i:j}(\hat{\theta}_n)$ throughout this section.

A.1 Algorithmic aspects

Recall that the computation of *auto-test* boils down to the computation of the linear statistic

$$R_{\rm lin} := \max_{\tau \in [n-1]} R_n(\tau) := \max_{\tau \in [n-1]} \hat{S}_{\tau+1:n}^\top \hat{I}_{n,\tau}^{-1} \hat{S}_{\tau+1:n} , \qquad (5)$$

where $\hat{\mathcal{I}}_{n,\tau} = \hat{\mathcal{I}}_{1:\tau} - \hat{\mathcal{I}}_{1:\tau} \hat{\mathcal{I}}_{1:\tau}^{-1} \hat{\mathcal{I}}_{1:\tau}$, and the scan statistic

$$R_{\text{scan}} := \max_{\tau \in [n-1]} \max_{T \subset [d], |T| \le P} R_n(\tau, T) := \max_{\tau \in [n-1]} \max_{T \subset [d], |T| \le P} [\hat{S}_{\tau+1:n}]_T^\top [\hat{\mathcal{I}}_{n,\tau}]_{T,T}^{-1} [\hat{S}_{\tau+1:n}]_T \quad , \tag{6}$$

where $[\hat{\mathcal{I}}_{n,\tau}]_{T,T}^{-1}$ should be understood as $\{[\hat{\mathcal{I}}_{n,\tau}]_{T,T}\}^{-1}$.

To compute these two statistics, a direct approach is to compute the full Fisher information matrices and then invert them. Another approach consists in solving the linear system $\mathcal{I}^{-1}S$ by the conjugate gradient algorithm. We refer to it as the AutoDiff-friendly approach.

In the following, we analyze the time and space complexity of these two approaches in the most general case, that is, the sequence $\{\hat{\mathcal{I}}_{1:t}\}_{t=1}^{n}$ does not admit a recursion that could simplify its computation. For every $t \in [n]$, we assume the computational graph of the log-likelihood is of size tC_1 with $C_1 \geq d$. As a result, computing $\hat{S}_{1:t}$ by AutoDiff takes $\mathcal{O}(tC_1)$ time and $\mathcal{O}(tC_1)$ space. Similarly, we assume the computational graph of the score $\hat{S}_{1:t}$ is of size tC_2 . Then the time and the space complexity of computing $\hat{\mathcal{I}}_{1:t}(\theta)$ are $\mathcal{O}(tdC_2)$ and $\mathcal{O}(tC_2)$, respectively, if we call AutoDiff on $\hat{S}_{1:t}^{\top}e_k$ for each $k \in [d]$, where $\{e_k\}_{k=1}^d$ is the standard basis of \mathbb{R}^d . We usually have $C_2 > C_1$ when $\ell(\theta)$ is not linear in θ .

Computing the linear statistic using automatic differentiation. The main steps to compute the linear statistic with the direct approach are summarized in Algorithm 2. The most time-consuming step is the for loop in steps 5-9. For each $\tau \in [n-1]$, steps 6-8 take time $\mathcal{O}(\tau C_1)$, $\mathcal{O}(\tau dC_2)$ and $\mathcal{O}(d^3)$, respectively. Therefore, the overall time complexity of Algorithm 2 is $\mathcal{O}(n^2 dC_2 + nd^3)$. The most space-consuming steps

Algorithm 2 Linear statistic with the direct approach

- 1: Input: Data $(W_k)_{k=1}^n$, log-likelihood ℓ , and MLE $\hat{\theta}_n$.
- 2: Compute $\hat{S}_{1:n}$ by calling AutoDiff on $\ell_{1:n}(\hat{\theta}_n)$.
- 3: Compute $\hat{\mathcal{I}}_{1:n}$ by calling d times AutoDiff on $\hat{S}_{1:n}$.
- 4: Compute $\hat{\mathcal{I}}_{1:n}^{-1}$.
- 5: for $\tau = 1, ..., n 1$ do

6: Compute $\hat{S}_{1:\tau}$ by calling AutoDiff on $\ell_{1:\tau}(\hat{\theta}_n)$, and then compute $\hat{S}_{\tau+1:n} = \hat{S}_{1:n} - \hat{S}_{1:\tau}$.

- 7: Compute $\hat{I}_{1:\tau}$ by calling d times AutoDiff on $\hat{S}_{1:\tau}$.
- 8: Compute $R_n(\tau)$ in (5).
- 9: end for
- 10: Compute R_{lin} in (5).
- 11: Output: R_{lin} .

Algorithm 3 Linear statistic with the conjugate gradient algorithm

- 1: Input: Data $(W_k)_{k=1}^n$, log-likelihood ℓ , and MLE $\hat{\theta}_n$.
- 2: Compute $\hat{S}_{1:n}$ by calling AutoDiff on $\ell_{1:n}(\hat{\theta}_n)$.
- 3: Compute $\hat{\mathcal{I}}_{1:n}$ by calling d times AutoDiff on $\hat{S}_{1:n}$.
- 4: for $\tau = 1, ..., n 1$ do
- 5: Compute $\hat{S}_{1:\tau}$ by calling AutoDiff on $\ell_{1:\tau}(\hat{\theta}_n)$, and then compute $\hat{S}_{\tau+1:n} = \hat{S}_{1:n} \hat{S}_{1:\tau}$.
- 6: Compute $R_n(\tau)$ in (7) by the conjugate gradient algorithm.
- 7: end for
- 8: Compute R_{lin} in (5).
- 9: Output: R_{lin} .

are to store the computational graph of $\hat{S}_{1:n}$ with complexity $\mathcal{O}(nC_2)$, and to store the full Fisher information matrix with complexity $\mathcal{O}(d^2)$. Consequently, the overall space complexity is $\mathcal{O}(nC_2 + d^2)$.

We now investigate the AutoDiff-friendly approach. According to the Woodbury matrix identity, we have

$$\hat{\mathcal{I}}_{n,\tau}^{-1} = \hat{\mathcal{I}}_{1:\tau}^{-1} + \mathcal{I}_{\tau+1:n}^{-1}$$

The statistic $R_n(\tau)$ then reads

$$R_n(\tau) = \hat{S}_{\tau+1:n}^{\top} \hat{\mathcal{I}}_{1:\tau}^{-1} \hat{S}_{\tau+1:n} + \hat{S}_{\tau+1:n}^{\top} \hat{\mathcal{I}}_{\tau+1:n}^{-1} \hat{S}_{\tau+1:n} .$$
⁽⁷⁾

To compute $\hat{\mathcal{I}}_{1:\tau}^{-1} \hat{S}_{\tau+1:n}$, we apply the conjugate gradient algorithm to solve the problem

$$\min_{x} \left\{ \frac{1}{2} x^{\top} \hat{\mathcal{I}}_{1:\tau} x - \hat{S}_{\tau+1:n}^{\top} x \right\}$$

Each iteration of the conjugate gradient algorithm requires evaluating $\hat{\mathcal{I}}_{1:\tau}x$, which can be obtained by calling AutoDiff on $\hat{\mathcal{S}}_{1:\tau}^{\top}x$ with $\mathcal{O}(\tau C_2)$ time and $\mathcal{O}(\tau C_2)$ space. Moreover, it converges in $M \leq d$ steps. As a result, computing $\hat{\mathcal{I}}_{1:\tau}^{-1}\hat{\mathcal{S}}_{\tau+1:n}$ takes $\mathcal{O}(\tau MC_2)$ time and $\mathcal{O}(\tau C_2)$ space. The steps to compute $\hat{\mathcal{I}}_{\tau+1:n}^{-1}\hat{\mathcal{S}}_{\tau+1:n}$ is similar since $\hat{\mathcal{I}}_{\tau+1:n}x = \hat{\mathcal{I}}_{1:n}x - \hat{\mathcal{I}}_{1:\tau}x$. Hence, we may compute R_{lin} as in Algorithm 3. The most expensive steps are the computation of $\hat{\mathcal{I}}_{1:n}$ in step 3 and the for loop in steps 4-7. Step 3 takes $\mathcal{O}(ndC_2)$ time and $\mathcal{O}(nC_2 + d^2)$ space. For each $\tau \in [n-1]$, the steps within the for loop, as discussed above, take $\mathcal{O}(\tau MC_2)$ time and $\mathcal{O}(\tau C_2)$ space. Hence, the overall time and space complexities are $\mathcal{O}(n^2MC_2 + ndC_2)$ and $\mathcal{O}(nC_2 + d^2)$. Since $M \leq d$, it is clear that this approach is more efficient than the direct one.

Computing the scan statistic using automatic differentiation. Computing the scan statistic exactly 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 $\max_{|T|=p} R_n(\tau, T)$, say T_p , by the indices of the largest p components in

$$v(\tau) := \hat{S}_{\tau+1:n}^{\top} \operatorname{diag}\{\hat{\mathcal{I}}_{n,\tau}\}^{-1} \hat{S}_{\tau+1:n} \quad .$$
(8)

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 $R_n(\tau, T)$. We show in Appendix B that this approximation is accurate if the difference between the largest eigenvalue and the smallest eigenvalue of $\hat{\mathcal{I}}_{n,\tau}$ is small compared to $\|\hat{S}_{\tau+1:n}\|^2$. Formally, we approximate R_{scan} by

$$R_{\text{scan}} \approx \max_{\tau \in [n-1]} \max_{p \le P} R_n(\tau, T_{\tau, p}) := \max_{\tau \in [n-1]} \max_{p \le P} \left[\hat{S}_{\tau+1:n} \right]_{T_{\tau, p}}^\top [\hat{\mathcal{I}}_{n, \tau}]_{T_{\tau, p}, T_{\tau, p}}^{-1} [\hat{S}_{\tau+1:n}]_{T_{\tau, p}}, \tag{9}$$

where $T_{\tau,p}$ corresponds to the largest p indices of $v(\tau)$.

Note that, in order to compute the scan statistic in a similar fashion as the linear statistic, we may modify the normalizing matrix $[\hat{\mathcal{I}}_{n,\tau}]_{T,T}^{-1}$ as

$$[\hat{\mathcal{I}}_{1:\tau}]_{T,T}^{-1} + [\hat{\mathcal{I}}_{\tau+1:n}]_{T,T}^{-1} .$$
⁽¹⁰⁾



Figure 4: (a) Running time versus sample size for linear models with d = 1000; (b) Running time versus number of parameters for linear models with n = 10000; (c) Running time versus sample size for multilayer perceptrons with d = 1035 (r = 45); (d) Running time versus number of parameters for multilayer perceptrons with n = 10000.

It can be shown that (10) converges to the same limit as $\hat{\mathcal{I}}_{1:\tau}$ under the null so that the calibration discussed in Appendix B remains valid. Hence, for the direct approach, we need the following steps to compute R_{scan} in addition to Algorithm 2: 1) sort $v(\tau)$ and obtain $\{T_{\tau,p}\}_{p\in[P]}$, and 2) compute $\{R_n(\tau, T_{\tau,p})\}_{p\in[P]}$, for each $\tau \in$ [n-1]. For the AutoDiff-friendly approach, after we sort $v(\tau)$, we can again compute $[\hat{\mathcal{I}}_{1:\tau}]_{T_{\tau,p},T_{\tau,p}}^{-1}[\hat{S}_{\tau+1:n}]_{T_{\tau,p}}$ by the conjugate gradient algorithm. Since $P \ll d$, the time complexity and space complexity of the linear statistic dominate the ones of the scan statistic.

When the observations $\{W_k\}_{k=1}^n$ are independent, the score $\hat{S}_{1:\tau}$ and information $\hat{\mathcal{I}}_{1:\tau}$ can be computed recursively in the direct approach. As a result, computing the *auto-test* with the direct approach will be more efficient if $n \gg d$.

A.2 Running times

We then compare empirically the running time of the two approaches. For simplicity, we focus on applying them to compute $\hat{\mathcal{I}}_{1:n}^{-1}z$ for some randomly generated vector $z \in \mathbb{R}^d$. We consider two models: 1) a linear model $Y = \theta^\top X + \varepsilon$ with log-likelihood (up to a constant) $\ell(\theta) = -(Y - \theta^\top X)^2$ (or quadratic loss), where $\theta \in \mathbb{R}^d$; 2) a multilayer perceptron (MLP) with the following structure: $x_0 \to x_1 = \sigma(A_1x_0 + b_1) \to x_2 =$ $A_2x_1 + b_2$, where $x_0 \in \mathbb{R}^r$ is the input vector, $A_1 \in \mathbb{R}^{r \times [r/2]}$, $b_1 \in \mathbb{R}^{[r/2]}$, $A_2 \in \mathbb{R}^{[r/2] \times 1}$ and $b_2 \in \mathbb{R}$. Hence, there are d = r[r/2] + 2[r/2] + 1 parameters in this model. The loss function is again chosen as the quadratic loss. For each of the two models, we generate n *i.i.d.* observations from this model and use the two approaches ("Direct" and "Ours") to compute $\hat{\mathcal{I}}_{1:n}^{-1}z$. For the conjugate gradient algorithm, we set the target accuracy to be 10^{-7} and set the maximum number of iterations to be 2d. For each pair of (n, d), we repeat the experiment 5 times and report the average running time with standard error in Fig. 4. The experiments are performed on a machine with 32 2.8GHz Intel Core i9 CPUs.

For the linear model, the information matrix is well-conditioned, so it took strictly less than d iterations for the conjugate gradient algorithm to converge. This contributes to the significant improvement on the running time compared to the direct approach. As for the MLP, the information matrix is ill-conditioned, so it usually took the conjugate gradient algorithm the maximum number of iterations, *i.e.*, 2d, to converge. In fact, the running time of the AutoDiff-friendly approach is about twice larger than the direct approach. Note that this time could be potentially reduced by computing the inverse-matrix-vector product inexactly.

A.3 Examples

Example 1 (Text topic model). The text topic model introduced in [20] 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 $h \neq \mathcal{H}(X)$. The authors proposed a class of spectral methods to recover approximately the map $\hat{\mathcal{H}}$ up to permutation. Consequently, the log-likelihood can be computed as

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

where X_0 is assumed to be known.

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

Example 3 (Hidden Markov model). Suppose that observations $(Y_k)_{k=1}^n$ are from a hidden Markov model (HMM)

$$X_k \sim Q(X_{k-1}, \cdot)$$
 and $Y_k \sim G(X_k, \cdot)$.

where G and Q are the transition distribution and emission distribution, respectively. For simplicity, we write $q_{x,x'} = Q(x,x')$ and $g_k(x) = G(x,Y_k)$. Its log-likelihood function can be computed by the normalized forward recursion [6, Chapter 3]: $\ell_n(\theta) = \sum_{k=1}^n \log c_k$ where, recursively,

$$c_k = \sum_{x_{k-1}, x_k=1}^{M} \phi_{k-1}(x_{k-1}) q_{x_{k-1}, x_k} g_k(x_k)$$

$$\phi_k(x_k) = c_k^{-1} \sum_{x_{k-1}=1}^{M} \phi_{k-1}(x_{k-1}) q_{x_{k-1}, x_k} g_k(x_k), \quad \forall x_k \in [M] ,$$

with initial conditions

$$c_0 = \sum_{x_0=1}^M g_0(x_0)\nu(x_0)$$

$$\phi_0(x_0) = c_0^{-1}g_0(x_0)\nu(x_0), \quad \forall x_0 \in [M] \ .$$

B Theoretical Results and Proofs

B.1 Null hypothesis

This section is devoted to determine thresholds for the linear test, scan test, and *auto-test* so that they are consistent in level. For this purpose, we first derive the limiting distribution of $R_n(\tau_n)$ for any sequence $(\tau_n)_{n\geq 1}$ such that $\tau_n/n \to \lambda \in (0, 1)$. We then determine the thresholds based on the limiting distribution.

Assumption 1. Let W_1, \ldots, W_n be a time series with a correctly specified model $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$. Suppose that the true parameter $\theta_0 \in int(\Theta)$ (the interior of Θ), and that the following assumptions hold:

- Θ contains an open neighborhood Θ_0 on which $\ell_n(\theta) := \log p_\theta(W_1, \ldots, W_n)$ is twice continuously A1 : differentiable and $\|n^{-1}\nabla^3_{\theta}\ell_n(\theta)\| \stackrel{a.s.}{\leq} M(W_1,\ldots,W_n) = O_p(1)$ for every $\theta \in \Theta_0$. A2 : $-\nabla^2_{\theta}\ell_n(\theta_0)/n \to_p \mathcal{I}_0$ where $\mathcal{I}_0 \in \mathbb{R}^{d \times d}$ is positive definite.
- A3 : The MLE $\hat{\theta}_n$ exists and $\sqrt{n}(\hat{\theta}_n \theta_0) \rightarrow_d \mathcal{N}(0, \mathcal{I}_0^{-1})$ (convergence in distribution).
- A4 : 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,

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

where $\mathbb{E}[M_k|\mathcal{F}_{k-1}] = 0, \forall k \in [n]$. In addition, this martingale difference sequence satisfies the Lindeberg conditions:

 $A_{4}(a) : n^{-1} \sum_{k=1}^{n} \mathbb{E}[M_k M_k^\top | \mathcal{F}_{k-1}] \rightarrow_p \mathcal{I}_0 \text{ and }$ $A4-(b): \quad \forall \varepsilon > 0 \text{ and } \alpha \in \mathbb{R}^d, n^{-1} \sum_{k=1}^n \mathbb{E}\left[(\alpha^\top M_k)^2 \mathbb{1}\{ |\alpha^\top M_k| > \sqrt{n}\varepsilon \} | \mathcal{F}_{k-1} \right] \to_p 0.$

A useful sufficient condition for Assumption A4 is given below.

Lemma 1. Assume that the normalized score can be written as

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

where $\{M_k\}_{k\in\mathbb{N}_+}$ is a stationary and ergodic martingale difference sequence w.r.t. its natural filtration, then Assumption A4 holds true.

Proof By stationarity, there exists a fixed measurable function $f : \mathbb{R}^{\infty} \to \mathbb{R}^{\infty}$ such that, for all $k \in \mathbb{N}_+$,

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

almost surely. Due to the ergodicity of M_k , the series $N_k = f(M_{k-1}, M_{k-2}, ...)$ is also ergodic so that $\overline{N}_n \to_{a.s.} \mathbb{E}[N_1], i.e., \text{ the condition A4-(a) holds true. Similarly, given } c > 0,$

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

for any $\alpha \in \mathbb{R}^d$, where $G(c) = \mathbb{E}[(\alpha^\top M_1)^2 | \mathbb{1}\{ |\alpha^\top M_1| > c \}]$ 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 A4-(b), note that $G_n(c)$ is decreasing in c, so, for every $\varepsilon > 0$, there exists M > 0 such that n > M implies

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\alpha^{\top} M_{k}^{2} | \mathbb{1}\left\{\left|\alpha^{\top} M_{k}\right| > \varepsilon \sqrt{n}\right\} | \mathcal{F}_{k-1}\right] \leq G_{n}(c_{0}) < \delta$$

almost surely. As δ is arbitrary, we know that the condition A4-(b) holds.

Remark. Under suitable regularity conditions, Assumption 1 holds true for *i.i.d.* models, hidden Markov models [3, Chapter 12], and stationary autoregressive moving-average models [11, Chapter 13].

Proposition 1 (Null hypothesis). Under Assumption 1, we have, for any $\tau_n \in \mathbb{N}_+$ such that $\tau_n/n \to \lambda \in$ (0,1),

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

In particular,

$$R_{n}(\tau_{n}) = S_{\tau_{n}+1:n}(\hat{\theta}_{n})^{\top} \mathcal{I}_{n}(\hat{\theta}_{n};\tau_{n})^{-1} S_{\tau_{n}+1:n}(\hat{\theta}_{n}) \to_{d} \chi_{d}^{2}$$

$$R_{n}(\tau_{n},T) = [S_{\tau_{n}+1:n}(\hat{\theta}_{n})]_{T}^{\top} [\mathcal{I}_{n}(\hat{\theta}_{n};\tau_{n})]_{T,T}^{-1} [S_{\tau_{n}+1:n}(\hat{\theta}_{n})]_{T} \to_{d} \chi_{|T|}^{2}, \quad \text{for any } T \subset [d]$$

We start by showing that the partial observed information $\mathcal{I}_n(\hat{\theta}_n; \tau_n)$ defined in (2) is a consistent estimator of \mathcal{I}_0 with proper normalization.

Lemma 2. Under assumptions A1-A3, we have, for any $\tau_n \in \mathbb{N}_+$ such that $\tau_n/n \to \lambda \in (0,1)$,

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

Proof According to Assumption A1 and Taylor's theorem, we obtain

$$\frac{1}{n} \left\| \nabla_{\theta}^{2} \ell_{n}(\hat{\theta}_{n}) - \nabla_{\theta}^{2} \ell_{n}(\theta_{0}) \right\| \leq \frac{1}{n} \left\| \nabla_{\theta}^{3} \ell_{n}(\overline{\theta}_{n}) \right\| \left\| \hat{\theta}_{n} - \theta_{0} \right\|$$

where $\|\overline{\theta}_n - \theta_0\| \leq \|\hat{\theta}_n - \theta_0\|$. Let E_n be the event $\{\hat{\theta}_n \in \Theta_0\}$. By Assumption A3, it holds that $\mathbb{P}(E_n) \to 0$, and thus

$$\frac{1}{n} \left\| \nabla_{\theta}^2 \ell_n(\hat{\theta}_n) - \nabla_{\theta}^2 \ell_n(\theta_0) \right\| \le \left\| M(W_1, \dots, W_n) \right\| \left\| \hat{\theta}_n - \theta_0 \right\| + o_p(1) = o_p(1) .$$

Consequently, by the triangle inequality, we get

$$\left\|-\frac{1}{n}\nabla_{\theta}^{2}\ell_{n}(\theta_{n})-\mathcal{I}_{0}\right\| \leq -\frac{1}{n}\left\|\nabla_{\theta}^{2}\ell_{n}(\theta_{n})-\nabla_{\theta}^{2}\ell_{n}(\theta_{0})\right\|+\left\|-\frac{1}{n}\nabla_{\theta}^{2}\ell_{n}(\theta_{0})-\mathcal{I}_{0}\right\| \rightarrow_{p} 0$$

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

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

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

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

To derive the asymptotic distribution of the score $S_{\tau_n+1:n}(\hat{\theta}_n)$, we will express it as a linear combination of the normalized scores $Z_{\tau_n}(\theta_0) := S_{1:\tau_n}(\theta_0)/\sqrt{\tau_n}$ and $Z_n(\theta_0) := S_{1:n}(\theta_0)/\sqrt{n}$, and then prove its asymptotic normality by the following lemma.

Lemma 3. Under Assumption A4, we have, for every sequence $\tau_n \in \mathbb{Z}_+$ such that $\tau_n/n \to \lambda \in (0,1)$,

$$\begin{pmatrix} Z_{\tau_n} \sqrt{\tau_n/n} \\ Z_n \end{pmatrix} \to_d \mathcal{N} \begin{pmatrix} 0, \begin{pmatrix} \lambda \mathcal{I}_0 & \lambda \mathcal{I}_0 \\ \lambda \mathcal{I}_0 & \mathcal{I}_0 \end{pmatrix} \end{pmatrix}$$
 (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 $(a^{\top}, b^{\top}) \in \mathbb{R}^{2d}$,

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

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,

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}[W_{n,k}^{2} | \mathcal{F}_{k-1}] &= \frac{1}{n} \sum_{k=1}^{\tau_{n}} (a+b)^{\top} \mathbb{E}[M_{k} M_{k}^{\top} | \mathcal{F}_{k-1}](a+b) + \frac{1}{n} \sum_{k=\tau_{n}+1}^{n} b^{\top} \mathbb{E}[M_{k} M_{k}^{\top} | \mathcal{F}_{k-1}]b \\ &= \frac{\tau_{n}}{n} \frac{1}{\tau_{n}} \sum_{k=1}^{\tau_{n}} a^{\top} \mathbb{E}[M_{k} M_{k}^{\top} | \mathcal{F}_{k-1}](a+2b) + \frac{1}{n} \sum_{k=1}^{n} b^{\top} \mathbb{E}[M_{k} M_{k}^{\top} | \mathcal{F}_{k-1}]b \\ &\to_{p} \lambda a^{\top} \mathcal{I}_{0}(a+2b) + b^{\top} \mathcal{I}_{0}b = \lambda(a+b)^{\top} \mathcal{I}_{0}(a+b) + (1-\lambda)b^{\top} \mathcal{I}_{0}b \ , \end{aligned}$$

and, for any $\varepsilon > 0$,

$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}[W_{n,k}^{2} \mathbb{1}(|W_{n,k}| > \varepsilon \sqrt{n}) | \mathcal{F}_{k-1}] \\
= \frac{1}{n} \sum_{k=1}^{\tau_{n}} \mathbb{E}\left[\left((a+b)^{\top} M_{k}\right)^{2} \mathbb{1}(|(a+b)^{\top} M_{k}| > \varepsilon \sqrt{n}) \middle| \mathcal{F}_{k-1}\right] + \frac{1}{n} \sum_{k=\tau_{n}+1}^{n} \mathbb{E}\left[\left(b^{\top} M_{k}\right)^{2} \mathbb{1}(|b^{\top} M_{k}| > \varepsilon \sqrt{n}) \middle| \mathcal{F}_{k-1}\right] \\
\rightarrow_{p} 0 ,$$

by Assumption A4-(b). Therefore, the statement (11) holds by invoking the Lindeberg theorem for martingales. Moreover,

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

Proof of Prop. 1 Since $\hat{\theta}_n$ maximizes the log-likelihood function, it must satisfy the first order optimality condition, *i.e.*, $S_{1:n}(\hat{\theta}_n) = 0$. Then by Assumption A3 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

$$-\frac{1}{\sqrt{n}}\nabla_{\theta}Z_{n}(\theta_{n}^{*}) = -\frac{1}{n}\nabla_{\theta}^{2}\ell_{n}(\theta_{n}^{*}) = \mathcal{I}_{0} + o_{p}(1)$$
(12)

by a similar argument as in Lemma 2. Note that $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$, we obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1) \quad .$$
(13)

We then express the score $S_{\tau_n+1:n}$ as a linear combination of the normalized scores $Z_{\tau_n}(\theta_0)$ and $Z_n(\theta_0)$. By Lindeberg theorem for martingales [23, Chapter 4.5] and Cramér-Wold device [4], Assumption A4 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

$$\frac{S_{\tau_n+1:n}(\hat{\theta}_n)}{\sqrt{n-\tau_n}} = \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \frac{1}{\sqrt{n-\tau_n}} \nabla_{\theta} S_{\tau_n+1:n}^{\top}(\theta_n^*) (\hat{\theta}_n - \theta_0) \\
= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \frac{(\nabla_{\theta} S_{1:n}(\theta_n^*) - \nabla_{\theta} S_{1:\tau_n}(\theta_n^*))^{\top}}{\sqrt{n(n-\tau_n)}} \sqrt{n} (\hat{\theta}_n - \theta_0) \\
= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \left[\sqrt{\frac{n}{n-\tau_n}} \frac{Z_n(\theta_n^*)}{\sqrt{n}} - \frac{\tau_n}{\sqrt{n(n-\tau_n)}} \frac{Z_{\tau_n}(\theta_n^*)}{\sqrt{\tau_n}} \right]^{\top} (\mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1)), \quad \text{by (13)} \\
= \frac{S_{\tau_n+1:n}(\theta_0)}{\sqrt{n-\tau_n}} + \left(\frac{\lambda}{\sqrt{1-\lambda}} - \sqrt{\frac{1}{1-\lambda}} \right) \mathcal{I}_0 \mathcal{I}_0^{-1} Z_n(\theta_0) + o_p(1), \quad \text{by (12)} \\
= -\sqrt{\frac{\tau_n}{n-\tau_n}} Z_{\tau_n}(\theta_0) + \sqrt{\frac{n}{n-\tau_n}} Z_n(\theta_0) + \frac{\lambda-1}{\sqrt{1-\lambda}} Z_n(\theta_0) + o_p(1) \\
= -\frac{\sqrt{\lambda}}{\sqrt{1-\lambda}} Z_{\tau_n}(\theta_0) + \frac{\lambda}{\sqrt{1-\lambda}} Z_n(\theta_0) + o_p(1) .$$

Now by Lemma 3, we have

$$\sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{\tau_n+1:n}(\hat{\theta}_n) \to_d \mathcal{N}\left(0, \left[\frac{1}{\lambda}\frac{\lambda}{1-\lambda} - \frac{2}{\lambda}\frac{\lambda^2}{1-\lambda} + \frac{1}{\lambda}\frac{\lambda^2}{1-\lambda}\right]\mathcal{I}_0\right) =_d \mathcal{N}(0, \mathcal{I}_0) \quad . \tag{14}$$

Therefore, by Lemma 2 and (14), we have $R_n(\tau_n) \to_d \chi_d^2$ and $R_n(\tau_n, T) \to_d \chi_{|T|}^2$.

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 $H_{\text{lin}}(\alpha) = q_{\chi_d^2}(\alpha/n)$ —the upper (α/n) quantile of χ_d^2 . Similarly, since the asymptotic distribution of $R_n(\tau, T)$ with $T \in \mathcal{T}_p$ is χ_p^2 and $|\mathcal{T}_p| = {d \choose p}$, the Bonferroni correction leads to the threshold $H_p(\alpha) = q_{\chi_p^2}(\alpha/[{d \choose p}n(p+1)^2])$, where $(p+1)^2$ is required to guarantee an asymptotic α level. In fact, we only need $\sum_{p \in \mathcal{P}} 1/(p+1)^2 < 1$ for controlling the level. Other corrections are possible, but the former provides small thresholds when the change is sparse.

Corollary 4. Under Assumption 1, the three tests $\psi_{auto}, \psi_{lin}, \psi_{scan}$ are consistent in level with thresholds defined above.

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

$$\mathbb{E}_{0}[\psi_{\text{lin}}(\alpha)] = \mathbb{P}_{0}\left\{\max_{\tau \in [n-1]} R_{n}(\tau) > H_{\text{lin}}(\alpha)\right\} \le \sum_{\tau=1}^{n-1} \mathbb{P}_{0}(R_{n}(\tau) > q_{\chi^{2}_{d}}(\alpha/n)) \le \sum_{\tau=1}^{n-1} \frac{\alpha}{n} + o(1) = \alpha + o(1) \quad ,$$

and

$$\begin{split} \mathbb{E}_{0}[\psi_{\text{scan}}(\alpha)] &= \mathbb{P}_{0}\Big(\max_{\tau \in [n-1]} \max_{p \leq P} \max_{T \in \mathcal{T}_{p}} H_{p}(\alpha)^{-1} R_{n}(\tau, T) > 1\Big) \\ &\leq \sum_{\tau=1}^{n-1} \sum_{p \leq P} \sum_{T \in \mathcal{T}_{p}} \mathbb{P}_{0}\Big(\frac{R_{n}(\tau, T)}{q_{\chi_{p}^{2}}\big(\alpha/\big(\binom{d}{p}n(p+1)^{2}\big)\big)} > 1\Big) \\ &\leq \sum_{\tau=1}^{n-1} \sum_{p \leq P} \sum_{T \in \mathcal{T}_{p}} \frac{\alpha}{\binom{d}{p}n(p+1)^{2}} + o(1) < \sum_{p=1}^{\infty} \frac{\alpha}{(p+1)^{2}} + o(1) < \alpha + o(1) \end{split}$$

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

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

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

B.2 Fixed alternative hypothesis

Under fixed alternative hypothesis, we make the following assumptions.

Assumption 2. Let W_1, \ldots, W_n be an independent sample and $\{p_{\theta} : \theta \in \Theta \subset \mathbb{R}^d\}$ be a family of density functions. Suppose 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 $\tau_n/n \to \lambda \in (0, 1)$. Moreover, suppose that the following assumptions hold:

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

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

- $A'^{2}(a) := \ell(\theta|x) := \log p_{\theta}(x)$ is twice continuously differentiable in θ almost surely.
- $A'^{2-}(b) : \nabla^{3}_{ijk}\ell(\theta|x) \text{ exists and satisfies } \left|\nabla^{3}_{ijk}\ell(\theta|x)\right| \leq M_{ijk}(x) \text{ for } \theta \in \Theta^{*} \text{ and } i, j, k \in [d] \text{ almost surely}$ with $\mathbb{E}_{\theta}, M_{ijk}(W) < \infty \text{ for } l \in \{0, 1\}.$
- $A'3: \quad \mathbb{E}_{\theta_l}[\nabla_{\theta}\ell(\theta^*)] = \nabla_{\theta}\mathbb{E}_{\theta_l}[\ell(\theta)]|_{\theta=\theta^*} = S_l^* \text{ for } l \in \{0,1\}.$
- $A'_{\mathcal{I}}: \quad \mathbb{E}_{\theta_l}[-\nabla^2_{\theta}\ell(\theta^*)] = \mathcal{I}_l^* \text{ is positive definite for } l \in \{0,1\}.$

Proposition 2 (Fixed alternative hypothesis). Under Assumption 2, there exists a sequence of MLE such that $\hat{\theta}_n \rightarrow_p \theta^*$ and, for any $\tau_n/n \rightarrow \lambda \in (0, 1)$,

$$\frac{1}{n}R_n(\tau_n) \to_p (\overline{\lambda}S_1^*)^\top (\mathcal{I}^*)^{-1} (\overline{\lambda}S_1^*) \quad , \tag{15}$$

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.

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 $\operatorname{bd}(B_{\varepsilon})$ be the boundary of B_{ε} . We will show that, for sufficiently small ε ,

$$\mathbb{P}\left(\ell_n(\theta) < \ell_n(\theta^*), \forall \theta \in \mathrm{bd}(B_{\varepsilon})\right) \to 1 \quad .$$
(16)

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

To prove (16), we write, for any $\theta \in \mathrm{bd}(B_{\varepsilon})$, that

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

where $\overline{\theta}_n \in B_{\varepsilon}$ satisfies $\|\overline{\theta}_n - \theta^*\| \leq \|\theta - \theta^*\|$. Let us bound D_1, D_2 , and D_3 separately. Note that, by the law of large numbers,

$$D_{1} \rightarrow_{p} (\theta - \theta^{*})^{\top} \left[\lambda \mathbb{E}_{\theta_{0}} [\nabla_{\theta} \ell(\theta^{*})] + \overline{\lambda} \mathbb{E}_{\theta_{1}} [\nabla_{\theta} \ell(\theta^{*})] \right]$$

= $(\theta - \theta^{*})^{\top} \nabla_{\theta} \left[\lambda \mathbb{E}_{\theta_{0}} [\ell(\theta)] + \overline{\lambda} \mathbb{E}_{\theta_{1}} [\ell(\theta)] \right] \Big|_{\theta = \theta^{*}}, \text{ by Assumption A'3}$
= $- (\theta - \theta^{*})^{\top} \nabla_{\theta} \left[\lambda D_{KL}(p_{\theta_{0}} \| p_{\theta}) + \overline{\lambda} D_{KL}(p_{\theta_{1}} \| p_{\theta}) \right] \Big|_{\theta = \theta^{*}}$
= $0,$

where the last equality follows from Assumption A'1. Moreover, by Assumption A'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 λ_{\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 we have, by Assumption A'2,

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

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

$$|D_1| < \varepsilon^3$$
, $D_2 < -\lambda_{\min}\varepsilon^2/4$, $|D_3| \le A\varepsilon^3$,

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

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

and thus (16) holds.

Now, following a similar argument as in Lemma 2, we obtain

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

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

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

To show the power consistency of the proposed tests, it suffices to prove $H_{\text{lin}}(\alpha)/n = o(1)$ and $H_p(\alpha)/n = o(1)$ for all $p \in [P]$. For this purpose, we recall a concentration inequality valid for χ^2 distributions introduced in [5].

Lemma 4. Let W be a chi-square random variable with degrees of freedom d, that is, $W \sim \chi_d^2$. Then, for all x > 0,

$$\mathbb{P}\left\{W \ge d + 2\sqrt{dx} + 2x\right\} \le e^{-x} \; .$$

Corollary 5. Suppose that Assumption 2 is true and $S_1^* \neq 0$, then the three tests $\psi_{auto}, \psi_{lin}, \psi_{scan}$ are consistent in power.

Proof According to Lemma 4, we have, for any $\alpha \in (0, 1)$,

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

and thus $H_{\rm lin}(\alpha)/n \to 0$. Recall from Prop. 2 that

$$\frac{1}{n}R_n(\tau_n) \to_p (\overline{\lambda}S_1^*)^\top (\mathcal{I}^*)^{-1} (\overline{\lambda}S_1^*) \quad .$$

If $S_1^* \neq 0$, then it follows from the positive definiteness of \mathcal{I}^* that

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

Analogously, we get

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

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

B.3 Local alternative hypothesis

Under local alternative hypothesis, we make the following assumptions.

Assumption 3. Let W_1, \ldots, W_n be an independent sample and $\{p_{\theta} : \theta \in \Theta \subset \mathbb{R}^d\}$ be a family of density functions. Suppose 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 $\tau_n/n \rightarrow \lambda \in (0,1)$. Moreover, suppose that the following assumptions hold: A"1 : Θ contains an open neighborhood Θ_0 of θ_0 for which

 $\ell(\theta) := \ell(\theta|x) := \log p_{\theta}(x)$ is twice continuously differentiable in θ almost surely. A"1-(a) :

- $A"1-(b): \qquad \nabla^3_{ijk}\ell(\theta|x) \text{ exists and satisfies } \left|\nabla^3_{ijk}\ell(\theta|x)\right| \leq M_{ijk}(x) \text{ for } \theta \in \Theta_0 \text{ and } i, j, k \in [d] \text{ almost } i \leq M_{ijk}(x) \text{ for } \theta \in \Theta_0 \text{ and } i, j, k \in [d] \text{ almost } i \leq M_{ijk}(x) \text{ for } \theta \in \Theta_0 \text{ and } i, j, k \in [d] \text{ almost } i \leq M_{ijk}(x) \text{ for } \theta \in \Theta_0 \text{ and } i, j, k \in [d] \text{ almost } i \leq M_{ijk}(x) \text{ for } \theta \in \Theta_0 \text{ and } i, j, k \in [d] \text{ almost } i \leq M_{ijk}(x) \text{ for } \theta \in \Theta_0 \text{ and } i, j, k \in [d] \text{ almost } i \leq M_{ijk}(x) \text{ for } \theta \in \Theta_0 \text{ for }$ surely with $\mathbb{E}_{\theta_0} M_{ijk}(W) < \infty$.
- $A''\!2$:
- $$\begin{split} \mathbb{E}_{\theta_0}[\nabla_{\theta}\ell(\theta_0)] &= \nabla_{\theta}\mathbb{E}_{\theta_0}[\ell(\theta)]|_{\theta=\theta_0} = S_0.\\ \mathbb{E}_{\theta_0}[\nabla_{\theta}\ell(\theta_0)\nabla_{\theta}\ell(\theta_0)^{\top}] &= \mathbb{E}_{\theta_0}[-\nabla_{\theta}^2\ell(\theta_0)] = \mathcal{I}_0 \text{ is positive definite.} \end{split}$$
 A''3 :

Proposition 3 (Local alternative hypothesis). Under Assumption 3, there exists a sequence of MLE $\hat{\theta}_n$ such that

$$\frac{n}{\tau_n(n-\tau_n)} \mathcal{I}_n(\hat{\theta}_n;\tau_n) \to_p \mathcal{I}_0 \tag{17}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}_d(\overline{\lambda}h, \mathcal{I}_0^{-1})$$
(18)

$$\sqrt{\frac{n}{\tau_n(n-\tau_n)}} S_{\tau_n+1:n}(\hat{\theta}_n) \to_d \mathcal{N}_d(\sqrt{\lambda\bar{\lambda}} \ \mathcal{I}_0 h, \mathcal{I}_0) \ . \tag{19}$$

In particular,

$$R_n(\tau_n) \to_d \chi_d^2 \left(\lambda \overline{\lambda} h^\top \mathcal{I}_0 h \right)$$
$$R_n(\tau_n, T) \to_d \chi_{|T|}^2 \left(\lambda \overline{\lambda} [\mathcal{I}_0 h]_T^\top [\mathcal{I}_0]_{T,T}^{-1} [\mathcal{I}_0 h]_T \right) .$$

In this proof we firstly analyze the behavior of the score statistic under the null hypothesis, then Proof we use Le Cam's third lemma (e.g., [22]), 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 $\ddot{\theta}_n \rightarrow_p \theta_0$, then (17) directly follows from the proof in Lemma 2. Furthermore, by Assumption A"1-(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

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

Since $\hat{\theta}_n \to_p \theta_0$, we have $\overline{\theta}_n \to \theta_0$ and thus, by Assumption A"1-(b),

$$\frac{1}{n}\nabla_{\theta}S_{1:n}(\overline{\theta}_n) = \frac{1}{n}\nabla_{\theta}S_{1:n}(\theta_0) + o_p(1) = -\mathcal{I}_0 + o_p(1)$$

Therefore,

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

where $\overline{S}_i(\theta_0) = \mathcal{I}_0^{-1} \nabla_{\theta} \ell_i(\theta_0)$. We then prove the local asymptotic linearity of the log-likelihood ratio. We denote the joint probability measure of W_1, \ldots, W_n under the local alternative as $\mathbb{P}_{\theta_0, \theta_n}^{(\tau_n)}$. It holds that

$$\log \frac{d\mathbb{P}_{\theta_{0},\theta_{n}}^{(\tau_{n})}}{d\mathbb{P}_{\theta_{0}}^{n}} = \ell_{\tau_{n}+1:n}(\theta_{n}) - \ell_{\tau_{n}+1:n}(\theta_{0})$$

$$= (\theta_{n} - \theta_{0})^{\top} S_{\tau_{n}+1:n}(\theta_{0}) + \frac{1}{2}(\theta_{n} - \theta_{0})^{\top} \nabla_{\theta} S_{\tau_{n}+1:n}(\theta_{0})(\theta_{n} - \theta_{0}) + o_{p}(1)$$

$$= \frac{h^{\top}}{\sqrt{n}} S_{\tau_{n}+1:n}(\theta_{0}) + \frac{1}{2}h^{\top} \frac{\nabla_{\theta} S_{\tau_{n}+1:n}(\theta_{0})}{n}h + o_{p}(1) = h^{\top} \frac{1}{\sqrt{n}} S_{\tau_{n}+1:n}(\theta_{0}) - \frac{\overline{\lambda}}{2}h^{\top} \mathcal{I}_{0}h + o_{p}(1) .$$

For any $a \in \mathbb{R}^d$, it follows from the multivariate Central Limit Theorem [4] that

$$\begin{pmatrix} a^{\top}\sqrt{n}(\hat{\theta}_{n}-\theta_{0})\\ \log \frac{d\mathbb{P}_{\theta_{0},\theta_{n}}^{(\tau_{n})}}{d\mathbb{P}_{\theta_{0}}^{n}} \end{pmatrix} = \frac{1}{\sqrt{n}} \begin{bmatrix} \sum_{i=1}^{\tau_{n}} \begin{pmatrix} a^{\top}\overline{S}_{i}(\theta_{0})\\ 0 \end{pmatrix} + \sum_{i=\tau_{n}+1}^{n} \begin{pmatrix} a^{\top}\overline{S}_{i}(\theta_{0})\\ h^{\top}S_{i}(\theta_{0}) \end{pmatrix} \end{bmatrix} - \begin{pmatrix} 0\\ \frac{\sigma^{2}}{2} \end{pmatrix} + o_{p}(1)$$
$$\rightarrow_{d} \mathcal{N}_{2} \left(\begin{pmatrix} 0\\ -\sigma^{2}/2 \end{pmatrix}, \begin{pmatrix} a^{\top}\mathcal{I}_{0}^{-1}a & \overline{\lambda}a^{\top}h\\ \overline{\lambda}a^{\top}h & \sigma^{2} \end{pmatrix} \right) ,$$

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

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

By the Cramér-Wold device, the statement (18) holds.

Notice that, under \mathbb{P}_{θ_0} ,

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

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

B.4 Approximation in the scan statistic

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 $v(\tau) := S_{\tau+1:n}(\hat{\theta}_n)^{\top} \operatorname{diag}(\mathcal{I}_n(\hat{\theta}_n; \tau))^{-1} S_{\tau+1:n}(\hat{\theta}_n)$. 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 5. Let $\alpha \in \mathbb{R}^d$, and $A \in \mathbb{R}^{d \times d}$ be a symmetric positive definite matrix. Consider the optimization problem:

$$T^* = \underset{T \subset [d], |T| = p}{\arg \max} f(T) = \underset{T \subset [d], |T| = p}{\arg \max} \alpha_T^\top [A_{T,T}]^{-1} \alpha_T, \quad p \in [d] .$$

Let $0 < \lambda_1(A) \leq \cdots \leq \lambda_d(A)$ be the eigenvalues of A, and \hat{T} be the indices of the largest p components in diag $(A)^{-1}\alpha^{\odot 2}$, where $\alpha^{\odot 2}$ is the element-wise power. Then we have $|f(T^*) - f(\hat{T})| \leq 2[\lambda_1(A)^{-1} - \lambda_d(A)^{-1}] \|\alpha\|^2$.

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

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

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

On the one hand, note that

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

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

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



Figure 5: Power versus magnitude of change for HMMs with N hidden states (left: N = 3; right: N = 7).

C Additional Experimental Results

In this section, we give additional experimental results investigating and comparing the performance of the linear test, the scan test and the *auto-test* on synthetic data. We use three different types of lines to represent three tests, and different colors to indicate different sample sizes. Note that all the statistics are computed only for $\tau \in [n/10, 9n/10]$ to prevent encountering ill-conditioned Fisher information matrix.

Hidden Markov model. We consider HMMs with $N \in \{3,7\}$ 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]. Since 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. 5. When N = 3, three tests have almost identical performance. When N = 7, the change becomes sparser, and subsequently, the scan test and the *auto-test* outperform the linear test. In both cases, the three tests show consistent behavior as the sample size increases.

Time series model. We then consider two 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. Note that for ARMA models we do not have exact control of $\|\Delta\|/\sqrt{p}$, so readers need to be careful about the range of x-axis in Fig. 6.

As demonstrated in Fig. 6, the scan test works fairly well for these two ARMA models. However, the linear test and the *auto-test* have extremely high false alarm rate. This problem gets more severe as the sample size increases, and hence is not due to the lack of accuracy of the maximum likelihood estimator.

Restricted screening components. To investigate the high false alarm rate problem in Fig. 6, we consider the same two ARMA models with the restriction that we only detect changes in the AR coefficients. As presented in Fig. 7, all three tests are now consistent in level, and the linear test and the *auto-test* are slightly more powerful than the scan test. This suggests that this problem is caused by the non-homogeneity of model parameters. Indeed, in the experiments in Fig. 6, the derivatives w.r.t. AR coefficients are significantly larger than the ones w.r.t. MA coefficients. This results in ill-conditioned information matrix



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



Figure 7: Power versus magnitude of change for ARMA models with restricted components (left: ARMA(3,2); right: ARMA(6,5)).

and subsequent unstable computation of the linear statistic. On the contrary, the scan statistic only inverts the submatrix of size $p \times p$, whose condition number is much smaller. In fact, the parameters selected by the scan statistic are all AR coefficients in our experiments. Therefore, the scan statistic can produce reasonable results even if the parameters are heterogeneous. We note that in such situations we can select a small (or even zero) significance level for the linear part in the *auto-test* to obtain reasonable results.