Sequential Change Point Detection via Denoising Score Matching

Wenbin Zhou
Carnegie Mellon University
Pittsburgh, PA
Email: wenbinz2@andrew.cmu.edu

Liyan Xie
University of Minnesota
Minneapolis, MN
Email: liyanxie@umn.edu

Zhigang Peng
Georgia Institute of Technology
Atlanta, GA
Email: zpeng@gatech.edu

Shixiang Zhu
Carnegie Mellon University
Pittsburgh, PA
Email: shixiangzhu@cmu.edu

Abstract—**Sequential change-point detection plays a critical role in numerous real-world applications, where timely identification of distributional shifts can greatly mitigate adverse outcomes. Classical methods commonly rely on parametric density assumptions of pre- and post-change distributions, limiting their effectiveness for high-dimensional, complex data streams. This paper proposes a score-based CUSUM change-point detection, in which the score functions of the data distribution are estimated by injecting noise and applying denoising score matching. We consider both offline and online versions of score estimation. Through theoretical analysis, we demonstrate that denoising score matching can enhance detection power by effectively controlling the injected noise scale. Finally, we validate the practical efficacy of our method through numerical experiments on two synthetic datasets and a real-world earthquake precursor detection task, demonstrating its effectiveness in challenging scenarios.**

I. INTRODUCTION

Sequential change point detection is a fundamental statistical problem that aims to identify the exact moment when the underlying distribution of a sequential data stream undergoes a structural change [1]–[3]. This capability is critical in various real-world applications where detecting and reacting to changes in real time can have significant implications. For example, change point detection is vital in financial markets to identify regime shifts [4], [5], in manufacturing to detect system faults [6], and in epidemic control to detect the onset of outbreaks or shifts in infection trends [7].

Recent advances in sensor technology and signal processing have opened up new opportunities to uncover the intricate dynamics of change in various domains. For instance, in seismology, recent development and application of advanced earthquake detection techniques such as template matching and machine learning have resulted in an exponential growth in the quantity and quality of seismic records [8], [9]. These enriched datasets are critical for advancing our understanding of earthquake precursors [10] – an elusive yet transformative goal that could enable early warnings for catastrophic major earthquakes and significantly mitigate their impacts. While so far no reliable earthquake precursors have been identified [11]–[13], there is a renewed interest in studying them, largely driven by improved dense near-field geophysical observations, availability of big data and new development in machine learning methods in earthquake science [8], [9], [14].

As data streams become increasingly complex and high-dimensional, there is a growing need for more robust and scal-able change point detection frameworks [15]. Many traditional techniques are grounded in modeling the log-likelihood of pre- and post-change distributions, which often involve parametric assumptions on density functions [16], [17]. While effective in certain scenarios, these assumptions can pose challenges in capturing the nuances of high-dimensional, interdependent data. This may limit their applicability in modern settings where distributional changes tend to be subtle, non-linear, and situated in complex data spaces that are not easily represented by simple parametric models [18], [19].

In this paper, we propose a novel framework for sequential change point detection based on denoising score matching, a method that has demonstrated success in learning deep generative models [20], [21]. The most salient feature of this approach is that it avoids parametric assumptions on the pre- and post-change distributions. Instead, we adopt a score network to learn the underlying distributions in a fully data-driven manner. To address the challenge of detecting subtle changes in regions of the data space where samples are scarce, we adopt a denoising strategy [22]. By injecting controlled noise into the data, this strategy encourages the score network to explore underrepresented regions of the data distribution, improving its ability to detect changes. Furthermore, we develop a theoretical framework to quantify the tradeoff between the injected noise level and the score estimation accuracy. Our analysis reveals that careful selection of the noise level can minimize the error of the estimated distributional change and enable the score network to achieve robust detection performance. Our framework is validated through experiments on both synthetic and real-world datasets, demonstrating its capability to detect changes in complex, high-dimensional data distributions.

Our main contributions are summarized as follows: (i) We propose a novel change point detection algorithm based on denoising score matching; (ii) We provide theoretical insights on the detection efficiency of the proposed method, and demonstrate the tradeoff on the level of injected noise; (iii) We validate the proposed method on both synthetic and real datasets, and demonstrate its superior performance against the state-of-the-art approaches.

Related work. Classical sequential change detection methods [2], [3], [23], [24] can be categorized into parametric and non-parametric approaches. Parametric methods estimate the

parametric post-change distribution during the detection process, such as the generalized likelihood ratio (GLR) test [16], [25] and window-limited tests [17]. Nonparametric methods circumvent the need for explicit distributional assumptions and have been developed using techniques such as kernel density estimation [26], [27] and kernel-based distances [18], [19].

Recent advancements have introduced novel perspectives on change point detection. One notable development is the exact score-based detection method proposed by [28], which utilizes the Hyvärinen score to detect changes in known unnormalized distributions, offering an alternative to the traditional log-likelihood approach. At the same time, neural network-based methods [29], [30] have broadened the applicability of change point detection to settings with unknown and complex distributions through likelihood-based CUSUM techniques. Building upon these advancements, our proposed method bridges the gap between score-based and neural network-based approaches. By effectively leveraging the flexibility of neural networks to parametrize unnormalized models, our approach achieves a robust and computationally efficient solution for detecting change points in complex, high-dimensional settings.

II. PROBLEM SETUP

Denote $\mathcal{X}$ as the (possibly high-dimensional) continuous sample space of the observed data. Let p_0, p_1 be two distinct[1] and unknown probability distributions on $\mathcal{X}$. Given data sequence $\{x_t \in \mathcal{X}, t \in \mathbb{N}\}$, we suppose that the data-generating distribution changes from p_0 to p_1 at some unknown (but deterministic) change-point ν, i.e.,

$$
\begin{aligned}
x_t &\overset{\text{iid}}{\sim} p_0, \quad k = 1, 2, \dots, \nu - 1, \\
x_t &\overset{\text{iid}}{\sim} p_1, \quad k = \nu, \nu + 1, \dots
\end{aligned}
\tag{1}
$$

The objective is to detect the unknown change-point ν as quickly as possible while controlling the false alarm rate. Such detection is usually performed by designing a stopping time T on the observation sequence at which it is declared that a change has occurred [24].

The performance of a stopping time T is measured by its *average run length (ARL)* and *worst-case average detection delay (WADD)*. The ARL measures the average time to false alarms and is defined as $\text{ARL} := \mathbb{E}_\infty[T]$, where $\mathbb{E}_\infty$ denotes the expectation under the probability measure when no change occurs. WADD is defined as

$$
\text{WADD} := \sup_{\nu \geq 1} \text{ess sup } \mathbb{E}_\nu \left[(T - \nu + 1)^+ | x_1, \dots, x_{\nu-1} \right],
$$

where $\mathbb{E}_\nu$ is the expectation when the change-point equals ν, and the essential supremum is taken over all possible change-points $\nu \geq 1$ and all possible pre-change samples $\{x_1, \dots, x_{\nu-1}\}$. Our goal is to build a stopping time that can minimize WADD subject to a false alarm constraint $\text{ARL} \geq \gamma$.

[1]By "distinct" we mean that p_0 and p_1 must be different on a set in $\mathcal{X}$ with non-zero Lebesgue measure.

III. PROPOSED METHOD

We propose a novel change point detection algorithm based on denoising score matching (DSM). We first present preliminary results on score matching. Then we describe our proposed algorithm, referred to as `DSM-CUSUM`.

A. Preliminaries

We consider the scenario where the pre- and post-change density functions are either unavailable or difficult to estimate. This situation often arises, for example, in unnormalized models $p_\theta(x) = \tilde{p}(x, \theta)/Z_\theta$, where Z_θ is a normalizing constant that can be computationally intractable. Nonetheless, the distribution can still be characterized by the derivative of its log-density, $\nabla_x \log \tilde{p}(x, \theta)$, known as the score function. Note that this score function does not depend on Z_θ and is often easier to estimate. Formally, we define the score and Hyvärinen score (abbreviated as H-score) as follows [31].

Definition 1. *For a density function $p(x)$, its (Stein) score function is defined as $s(x) = \nabla_x \log p(x)$, and the Hyvärinen score is defined as $H(x; s) = \text{div } s(x) + \frac{1}{2}\|s(x)\|_2^2$, where $\text{div } s(x) = \sum_{i=1}^d \partial s_i(x)/\partial x_i$ is the divergence.*

We model the data distribution using a score function $s(x; \theta)$ parameterized by θ, which can be learned by minimizing the following Fisher divergence between the model and the true distribution $p(x)$:

$$
\min_\theta D_F := \mathbb{E}_{x \sim p}\|s(x; \theta) - \nabla_x \log p(x)\|_2^2.
\tag{2}
$$

This problem can be solved in practice using a family of methods known as score matching, which defines objectives that can be directly estimated from datasets sampled from p, without requiring the ground-truth data score [22], [31], [32].

Such estimated score functions can be inaccurate in low-density regions due to insufficient training data for computing the score-matching objective. Noise perturbation mitigates this issue by adding noise to the data and training score-based models on the perturbed data [21], [22]. Given a transition kernel $\mathcal{K}(x'|x)$ that represents the noise injection process, the perturbed score model $s'(x; \theta)$ is learned by minimizing the following denoising score matching objective:

$$
\mathbb{E}_{x \sim p}\mathbb{E}_{x' \sim \mathcal{K}(\cdot|x)} \|s'(x'; \theta) - \nabla_{x'} \log \mathcal{K}(x'|x)\|_2^2.
\tag{3}
$$

We choose the widely adopted Gaussian kernel $\mathcal{K}(x'|x) = \mathcal{N}(x'; x, \sigma^2 \mathbf{I})$, which corresponds to injecting Gaussian noise $\mathcal{N}(0, \sigma^2 \mathbf{I})$ to the data x. This choice simplifies the expression for $\nabla_{x'} \log \mathcal{K}(x'|x) = -(x' - x)/\sigma^2$ within (3). As shown in [21], the minimizer of (3) converges to the true perturbed data score defined by the transition kernel.

B. Proposed Method: Denoising Score Matching CUSUM

Our algorithm contains two steps: *training* and *detection*. The training step uses the available reference data to estimate the unknown score function, while the detection steps uses the estimated score to calculate the online detection statistics.

We start with the case that reference datasets $\mathcal{D}_0$ and $\mathcal{D}_1$ are available and are i.i.d. sampled from pre- and post-change

distributions, respectively. Following (3), we train two score models, both parameterized as $s(x; \theta)$ using neural networks represented by $\theta^{(i)}$ on $\mathcal{D}_i$ for $i = 0, 1$. We refer to this process as *offline score estimate*. The model parameter $\widehat{\theta}^{(i)}$ is learned by minimizing the following denoising score matching objective computed on training data:

$$\widehat{\theta}^{(i)} := \arg\min_{\theta} \sum_{x_j \in \mathcal{D}_i} \sum_{k=1}^{K} \left\| s\left(x + \epsilon_{jk}; \theta\right) + \frac{\epsilon_{jk}}{\sigma^2} \right\|_2^2, i = 0, 1. \quad (4)$$

Here $K \geq 1$ is a hyperparameter that represents the number of repeated noise injections, and the injected noise $\epsilon_{jk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{I})$. For simplicity we denote the resulted score models as $\widehat{s}_0(x) = s(x; \widehat{\theta}^{(0)})$ and $\widehat{s}_1(x) = s(x; \widehat{\theta}^{(1)})$.

Next, in the detection phase, for each time step t, we compute the difference of Hyvarinen scores of $\widehat{s}_0$ and $\widehat{s}_1$ evaluated at the data point x_t, and cumulate it into a CUSUM type detection statistics:

$$\mathcal{S}_t = \mathcal{S}_{t-1}^+ + \Delta(x_t), \Delta(x_t) = H(x_t; \widehat{s}_0) - H(x_t; \widehat{s}_1), t \geq 1, \quad (5)$$

where $\mathcal{S}_{t-1}^+ = \max\{\mathcal{S}_{t-1}, 0\}$ and $\mathcal{S}_0 = 0$. Note that the calculation of the detection statistics relies solely on the estimated score function and does not depend on the full density function.

To perform the detection, we raise an alarm when the statistics $\mathcal{S}_t$ exceeds a pre-specified threshold τ. Here τ is chosen to ensure that the average run length satisfies the desired lower bound, thereby controlling the false alarm rate. Formally, the stopping time for the detection can be defined as follows:

$$T = \inf \left\{ t : \mathcal{S}_t \geq \tau \right\}, \quad (6)$$

We further extend our study to the general case where *no reference data for the post-change distribution is provided beforehand*, requiring the post-change score estimation in an online manner, which we call *online score estimate*. The proposed method can be easily adapted to such setting as follows: During the *training phase*, we only learn the pre-change score model $\widehat{s}_0$ using the reference data $\mathcal{D}_0$ via (4). During the *detection phase*, the post-change score model is estimated online using the most recent w samples, where w is the pre-defined window size. We initialize $\mathcal{S}_0 = \cdots = \mathcal{S}_w = 0$ and $\widehat{\theta}_0 = \cdots = \widehat{\theta}_w = \widehat{\theta}^{(0)}$ (the pre-change score model parameter). For $t = w+1, \ldots$, we update the post-change score model $s(x; \widehat{\theta}_t)$ by Gradient descent on the learning objective:

$$\widehat{\theta}_t \leftarrow \widehat{\theta}_{t-1} - \eta \nabla_{\widehat{\theta}_{t-1}} \mathcal{L}_t(\widehat{\theta}_{t-1}); \quad (7)$$

$$\mathcal{L}_t(\theta) := \sum_{j=t-w+1}^{t} \sum_{k=1}^{K} \left\| s\left(x_j + \epsilon_{jk}; \theta\right) + \frac{\epsilon_{jk}}{\sigma^2} \right\|_2^2, \quad (8)$$

where η is the learning rate, and $\epsilon_{jk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{I})$, and we use the subscripted notation $\widehat{\theta}_t$ to denote that the score model is being progressively updated. We note that (7) can also be iterated for multiple gradient steps instead of a single step to improve the learning outcome, with the number of iterations typically determined by computational constraints.

Algorithm 1 DSM-CUSUM (Online)

Require: Reference dataset $\mathcal{D}_0$, threshold τ, window size w.
1: Initialize $\nu = \infty$, $\mathcal{S}_0 = \ldots = \mathcal{S}_w = 0$, $t = w$.
2: Train model $\hat{s}_0$ on dataset $\mathcal{D}_0$ and initialize $\hat{s}_1$ as $\hat{s}_0$.
3: **while** $S_t \leq \tau$ **do**
4: $\quad t \leftarrow t + 1, \nu \leftarrow t$.
5: $\quad$ Update post-change score model $\hat{s}_1$ based on the most recent w-length observations with (7).
6: $\quad$ Compute test statistics $\mathcal{S}_t$ as in (5).
7: **end while**
8: **return** Detected change point ν.

We update the detection statistic as defined in (5) using the last updated score model $\hat{s}_1(x) = s(x; \widehat{\theta}_{t-1})$, and the stopping time defined the same as (6). The detection procedure is summarized in Algorithm 1.

IV. THEORETICAL ANALYSIS

This section provides the theoretical guarantees of the proposed algorithm in terms of worst-case average detection delay. We show that the injected noise level affects both the score model's estimation error and the detection efficiency, which provides insights into the trade-off involved in selecting the optimal noise level. The detailed proofs can be found in Appendix A.

Recall $\widehat{s}_0(x)$ and $\widehat{s}_1(x)$ are the trained pre- and post-change score model through denoising score matching, and p_0, p_1 are the pre- and post-change data distributions. Denote $p_{i,\sigma}(x) = \int_{\mathcal{X}} p_i(y) \mathcal{N}(x; y, \sigma^2 \mathbf{I}) dy$, $i = 0, 1$ as the perturbed distribution after Gaussian noise injection. We make the following assumption on the *estimation error* $\epsilon_{\text{est}}(\sigma)$ and the *perturbation error* $\epsilon_{\text{pert}}(\sigma)$. These constants can be explicitly derived (which may depend on w, t and K) under certain conditions, see for example [33].

Assumption 1. *There exist constants $\delta \in [0, 1]$, $\epsilon_{\text{est}}(\sigma) \geq 0$, and $\epsilon_{\text{pert}}(\sigma) \geq 0$ such that with probability at least $1 - \delta$, for $i = 0, 1$, the following conditions hold:*

$$\mathbb{E}_{x \sim p_1} \|\widehat{s}_i(x) - \nabla \log p_{i,\sigma}(x)\|_2 \leq \epsilon_{\text{est}}(\sigma),$$
$$\mathbb{E}_{x \sim p_1} \|\nabla \log p_{i,\sigma}(x) - \nabla \log p_i(x)\|_2 \leq \epsilon_{\text{pert}}(\sigma),$$

where the randomness arises from the estimation of $\widehat{s}_0, \widehat{s}_1$ due to the randomness in training samples.

Assumption 2. *Under Assumption 1, we assume there exists some constants $\epsilon_{\text{div}} \geq 0$ and $C_1, C_2 \geq 1$, such that for $i = 0, 1$, the following conditions hold concurrently with Assumption 1,*

$$\text{Var}_{x \sim p_1} \|\widehat{s}_i(x) - \nabla \log p_{i,\sigma}(x)\|_2 \leq C_1 \cdot \epsilon_{\text{est}}^2(\sigma).$$
$$\text{Var}_{x \sim p_1} \|\nabla \log p_{i,\sigma}(x) - \nabla \log p_i(x)\|_2 \leq C_2 \cdot \epsilon_{\text{pert}}^2(\sigma).$$
$$\mathbb{E}_{x \sim p_1} |\text{div}\widehat{s}_i(x) - \text{div}\nabla \log p_i(x)| \leq \epsilon_{\text{div}}.$$

The first and second-order moments of all score functions are jointly upper bounded by some $M < +\infty$.

The first two inequalities ensure that the errors exhibit bounded variation at a rate similar to their first-order moment (*i.e.*, a finite coefficient of variance), while the last inequality

requires the divergence of the score model to approximate the ground truth effectively. These conditions are satisfied when the model is well-specified and well-trained.

Lemma 1. *Under Assumption 1 and 2, define the distribution change error as*

$$\epsilon(\sigma) := 2M\left[\epsilon_{\text{est}}(\sigma) + \epsilon_{\text{pert}}(\sigma)\right] + \epsilon_{\text{div}}$$
$$+ 4M(M+1) \cdot \max\{C_1, C_2\} \cdot \left[\epsilon_{\text{est}}^2(\sigma) + \epsilon_{\text{pert}}^2(\sigma)\right],$$

then with probability greater than $1 - \delta$,

$$|\mathbb{E}_{x \sim p_1}[\Delta(x)] - D_F(p_1\|p_0)| \leq \epsilon(\sigma),$$

where $D_F(p_1\|p_0) = \mathbb{E}_{x \sim p_1}\|\nabla_x \log p_1(x) - \nabla_x \log p_0(x)\|_2^2$ is the fisher divergence.

Lemma 1 establishes that the statistic computed using the plug-in estimate approximates the ground-truth Fisher divergence $D_F(p_1\|p_0)$ with an error bound of $\epsilon(\sigma)$, which scales proportionally to $\epsilon_{\text{est}}(\sigma)$ and $\epsilon_{\text{pert}}(\sigma)$. This observation paves the way for deriving an upper bound on the WADD of the detection procedure when the post-change score function is estimated offline using reference data $\mathcal{D}_1$, as detailed below.

Theorem 1 (WADD with offline score estimate)**.** *Under the offline score estimation setting, for a given threshold τ, and under Assumption 1 and 2, the following holds with probability at least $1 - \delta$,*

$$\text{WADD} \leq \frac{\tau}{D_F(p_1\|p_0) - \epsilon(\sigma)} + \frac{\mathbb{E}_{x \sim p_1}[\Delta(x)^2]}{[D_F(p_1\|p_0) - \epsilon(\sigma)]^2},$$

where $\epsilon(\sigma)$ is defined in Lemma 1.

Theorem 1 shows that a larger error term $\epsilon(\sigma)$ results in a worse WADD upper bound. Note that when $\epsilon_{\text{est}}(\sigma), \epsilon_{\text{pert}}(\sigma), \delta \to 0$, which corresponds to a well-trained score model, Theorem 1's bound converges to the standard result in [28]. Also, as $\tau \to \infty$, the upper bound scales approximately to be $\mathcal{O}(\tau/[D_F(p_1\|p_\infty) - \epsilon(\sigma)])$. Under the same assumptions, we can also derive the upper bound of WADD under the online setting.

Theorem 2 (WADD with online score estimate)**.** *Under the online score estimation setting, for the window size w and threshold τ, and under Assumption 1 and 2, the following holds with probability at least $1 - \delta$,*

$$\text{WADD} \leq \frac{\tau + c^{1/2}\tau^{1/2} + c + c^{1/2}d^{1/2} + d}{D_F(p_1\|p_0) - \epsilon(\sigma)},$$

where $\epsilon(\sigma)$ is defined in Lemma 1 and

$$c = \mathbb{E}_{x \sim p_1}[\Delta(x)^2]/[D_F(p_1\|p_0) - \epsilon(\sigma)],$$
$$d = w[D_F(p_1\|p_0) + \epsilon(\sigma)].$$

Compared to Theorem 1, the windowing procedure introduces additional terms in the numerator, leading to an inflated upper bound. However, these terms become negligible as $\tau \to \infty$, indicating that when the false alarm rate must be kept low, the online algorithm offers theoretical guarantees comparable to the offline version, ensuring its reliable deployment.
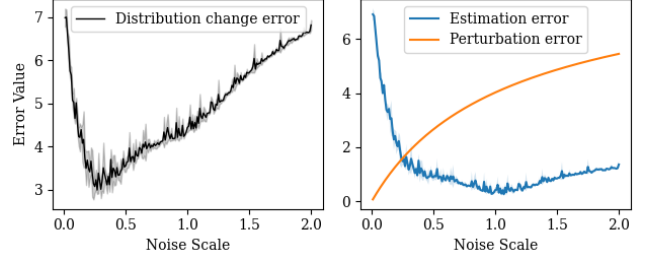

Fig. 1. Simulation results on score estimation error $\epsilon_{\text{est}}(\sigma)$, perturbation error $\epsilon_{\text{pert}}(\sigma)$, and distribution change error $\epsilon(\sigma)$ (up to constant scaling factors) for varying values of σ. The ground-truth data is generated from a 1D Gaussian mixture with two components.

Nonetheless, when a sufficiently large offline reference dataset is available to ensure effective training of the score network, the offline setting remains theoretically superior.

It is evident from the Theorems that an optimal choice of σ that minimizes $\epsilon(\sigma)$ will lead to a better detection performance, and by the definition of $\epsilon(\sigma)$, it depends on both the perturbation error $\epsilon_{\text{pert}}(\sigma)$ and the estimation error $\epsilon_{\text{est}}(\sigma)$. The following remark suggests that injecting noise generally enhances the accuracy of the statistic in capturing the distributional change:

Remark. *The process of noise injection involves a tradeoff: Higher noise levels increase perturbation error $\epsilon_{\text{pert}}(\sigma)$ since noise injection alters the distribution of training data. However, they also reduce score estimation error $\epsilon_{\text{est}}(\sigma)$ since noise injection improves score estimation. A key empirical observation is that ϵ_{est} typically decreases at a much faster rate than ϵ_{pert} increases. This trend is evident in the numerical simulation results presented in Figure 1 (right). As a result, even a moderate level of noise injection can effectively lower the overall error, as demonstrated in Figure 1 (left).*

V. NUMERICAL EXPERIMENTS

In this section, we present the numerical results of our method evaluated on both synthetic and real datasets. First, we compare the worst-case average detection delay (WADD) vs. the average run length (ARL) of our method against five baseline approaches on two synthetic datasets using offline estimation of the score model. Next, we demonstrate the application of our method to a real geophysical monitoring dataset, showcasing its ability to detect precursors to a major earthquake in an online setting. The score models in `DSM-CUSUM` are implemented as single-layer feed-forward neural networks, trained for $2,000$ epochs until convergence with an injected noise scale of 1. For the two synthetic data experiments and the real data experiment, the hidden dimensions are set to $2,048$, 512, and 512, respectively. In the real data experiment, a window size of 10 is used for online score estimation.

A. Synthetic Data Results

We construct two synthetic datasets with complex pre- and post-change distributions that cannot be easily captured by standard parametric models. The first dataset is a 2D ring-shaped distribution generated using Gaussian mixture models, while the second is a 10D dataset obtained by applying
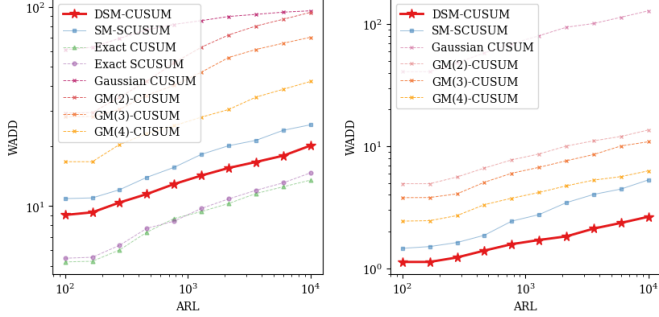
Fig. 2. Comparison of WADD vs ARL on baseline methods for two synthetic datasets. Left: 2D data by Gaussian mixtures. Right: 10D data by deep nets.

nonlinear transformations, specified by random neural networks, to 4D Gaussian data. We build the reference datasets in both scenarios to consist of $1,000$ datapoints. Details of the data generation process are provided in Appendix B-B.

We evaluate our proposed `DSM-CUSUM` under the offline score estimation setting, comparing to four baseline methods, including (i) CUSUM fitted with Gaussian mixture models with n components (GM(n)-CUSUM), (ii) an ablation model employing vanilla score matching (SM-SCUSUM), (iii) the Exact CUSUM [25], and (iv) the exact Score CUSUM (Exact SCUSUM) [28]. We note that Exact CUSUM and Exact SCUSUM cannot be readily applied to the second synthetic dataset as the density function of the data, transformed by nonlinear mappings, is inaccessible. Detailed configurations are described in Appendix B-D.

Figure 2 presents the WADD vs. ARL tradeoff curves for all models. In both experiments, the proposed `DSM-CUSUM` significantly outperforms other baselines, with its curves positioned substantially lower on the plot, making it the closest to those of the two oracle models. This underscores the robust modeling capabilities of unnormalized score-based models driven by deep neural networks, resulting in superior performance in detection tasks. Lastly, note that the proposed `DSM-CUSUM` outperforms the ablation model (SM-SCUSUM), despite both using the same neural network architecture. This performance gap also widens as the data distribution becomes more intricate. This further emphasizes the benefits of training with denoising score matching for handling more complex change point detection tasks.

B. Real Data Results: Earthquake Precursor Detection

Earthquake precursor detection [11], [12] focuses on identifying moments when the distribution of certain geophysical signals undergoes a shift, known as a "precursor". This shift provides critical insights for predicting the potential occurrence of major earthquakes. We evaluate the performance of our method on such a real-world dataset consisting of hourly high-resolution geophysical signals monitored prior to a moderate-size earthquake that occurred in China, 2014.

The dataset includes four types of signals from the three monitoring stations closest to the earthquake's epicenter over a two-year period. The signals include mid-layer water temperature, shallow-layer water temperature, static water level, and dynamic
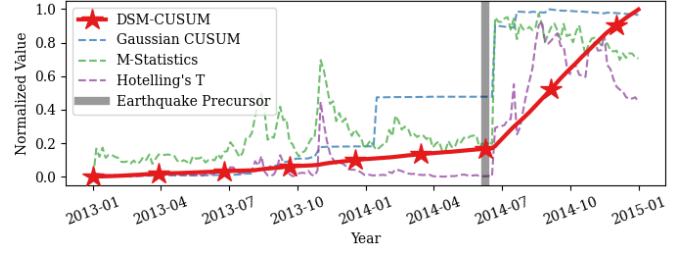

Fig. 3. Geophysical signal datasets: comparison of our method (red star) with three baseline methods (dashed lines). All values are normalized to the range $[0, 1]$ for ease of comparison. Domain experts identify the true precursor signals as emerging around early June 2014.

water level. After removing seasonal patterns, anomalies, and missing values, we use the first half of the trajectory as the reference dataset, and compare our algorithm under online score estimation with three baseline methods, including Hotelling's T^2, M-statistics [18], and Gaussian CUSUM. Baseline methods are detailed in Appendix B-D.

Figure 3 illustrates the estimated statistics for all methods applied to the geophysical signal data stream. We observe that most baseline methods exhibit pronounced fluctuations between July 2013 and January 2014, which could lead to false alarms under lower threshold settings. In contrast, our method demonstrates consistently low and stable statistics throughout this period. Furthermore, it accurately identifies the true precursor signal around mid-June 2014, which has been validated by domain experts. These results underscore the superior accuracy and reliability of our method in detecting earthquake precursors compared to the baseline approaches.

VI. CONCLUSION

This paper addresses the key challenge of sequential change point detection where the pre- and post-change distributions are unknown. Our proposed method leverages the denoising score-matching objective to inject noise into the data and learn the score function of the data distribution, which is then used to construct a Hyvärinen score-based CUSUM statistic. Empirically, we demonstrate that injecting a small amount of noise into the data improves the score model's ability to estimate the score function more accurately, which, in turn, enhances control over distributional changes and ultimately lowers the detection delay of the proposed statistics. In addition, we validate our approach through numerical experiments on two synthetic datasets and one real-world geophysical monitoring dataset, showcasing its practical effectiveness in complex, potentially high-dimensional distributions.

A key limitation of our current work lies in the preliminary nature of the theoretical assumptions, which do not explicitly capture the precise relationship between noise scale and error values. We hypothesize that further investigation on score-matching learning theory could address this limitation, enabling the derivation of an optimal noise injection scale and facilitating a more systematic and effective implementation of our method.

REFERENCES

[1] V. V. Veeravalli and T. Banerjee, "Quickest change detection," *Academic Press Library in Signal Processing: Array and Statistical Signal Processing*, vol. 3, pp. 209–256, 2013.

[2] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*. Springer, 1985.

[3] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. ser. Monographs on Statistics and Applied Probability 136. Boca Raton, London, New York: Chapman & Hall/CRC Press, Taylor & Francis Group, 2015.

[4] M. Barassi, L. Horvath, and Y. Zhao, "Change-point detection in the conditional correlation structure of multivariate volatility models," *Journal of Business & Economic Statistics*, vol. 38, no. 2, pp. 340–349, 2020.

[5] S. Zhu, H. S. Yuchi, M. Zhang, and Y. Xie, "Sequential adversarial anomaly detection for one-class event data," *INFORMS Journal on Data Science*, vol. 2, no. 1, pp. 45–59, 2023.

[6] T. L. Lai, "Sequential changepoint detection in quality control and dynamical systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 4, pp. 613–644, 1995.

[7] S. Zhu, A. Bukharin, L. Xie, K. Yamin, S. Yang, P. Keskinocak, and Y. Xie, "Early detection of COVID-19 hotspots using spatio-temporal data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 2, pp. 250–260, 2022.

[8] Z. Dong, Z. Fan, and S. Zhu, "Conditional generative modeling for high-dimensional marked temporal point processes," *arXiv preprint arXiv:2305.12569*, 2023.

[9] G. C. Beroza, M. Segou, and S. Mostafa Mousavi, "Machine learning and earthquake forecasting—next steps," *Nature communications*, vol. 12, no. 1, p. 4761, 2021.

[10] M. E. Pritchard, R. M. Allen, T. W. Becker, M. D. Behn, E. E. Brodsky, R. Bürgmann, C. Ebinger, J. T. Freymueller, M. Gerstenberger, B. Haines *et al.*, "New opportunities to study earthquake precursors," pp. 2444–2447, 2020.

[11] P. Picozza, L. Conti, and A. Sotgiu, "Looking for earthquake precursors from space: A critical review," *Frontiers in Earth Science*, vol. 9, p. 676775, 2021.

[12] L. Conti, P. Picozza, and A. Sotgiu, "A critical review of ground based observations of earthquake precursors," *Frontiers in Earth Science*, vol. 9, p. 676766, 2021.

[13] Z. Peng and X. Lei, "Physical mechanisms of earthquake nucleation and foreshocks: cascade triggering, aseismic slip, or fluid flows?" *Earthquake Research Advances*, p. 100349, 2024.

[14] S. M. Mousavi and G. C. Beroza, "Machine learning in earthquake seismology," *Annual Review of Earth and Planetary Sciences*, vol. 51, no. 1, pp. 105–129, 2023.

[15] B. Liu, C. Zhou, X. Zhang, and Y. Liu, "A unified data-adaptive framework for high dimensional change point detection," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 82, no. 4, pp. 933–963, 2020.

[16] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Transactions on Information theory*, vol. 44, no. 7, pp. 2917–2929, 1998.

[17] L. Xie, G. V. Moustakides, and Y. Xie, "Window-limited cusum for sequential change detection," *IEEE Transactions on Information Theory*, vol. 69, no. 9, pp. 5990–6005, 2023.

[18] S. Li, Y. Xie, H. Dai, and L. Song, "M-statistic for kernel change-point detection," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[19] ——, "Scan b-statistic for kernel change-point detection," *Sequential Analysis*, vol. 38, no. 4, pp. 503–544, 2019.

[20] Y. Song, S. Garg, J. Shi, and S. Ermon, "Sliced score matching: A scalable approach to density and score estimation," in *Uncertainty in Artificial Intelligence*, 2020, pp. 574–584.

[21] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[22] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[23] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge University Press, 2008.

[24] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, "Sequential (quickest) change detection: Classical results and new directions," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 494–514, 2021.

[25] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[26] Y. Liang and V. V. Veeravalli, "Quickest change detection with leave-one-out density estimation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[27] Y. Kawahara and M. Sugiyama, "Sequential change-point detection based on direct density-ratio estimation," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5, no. 2, pp. 114–127, 2012.

[28] S. Wu, E. Diao, T. Banerjee, J. Ding, and V. Tarokh, "Score-based quickest change detection for unnormalized models," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, vol. 206. PMLR, 2023, pp. 10 546–10 565.

[29] G. V. Moustakides and K. Basioti, "Training neural networks for likelihood/density ratio estimation," 2019. [Online]. Available: https://arxiv.org/abs/1911.00405

[30] J. Lee, Y. Xie, and X. Cheng, "Training neural networks for sequential change-point detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[31] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[32] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[33] A. Block, Y. Mroueh, and A. Rakhlin, "Generative modeling with denoising auto-encoders and langevin sampling," *arXiv preprint arXiv:2002.00107*, 2020.

[34] G. Lorden, "On excess over the boundary," *The Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 520–527, 1970.

[35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

We begin with introducing two useful technical lemmas (Lemma 2 and Lemma 3). Then we present the proof of Lemma 1, Theorem 1 and Theorem 2 in sequence. The pseudocode for the offline version of DSM-CUSUM is also provided in Algorithm 2.

For notation simplicity, we will use $\mathbb{E}_1$ instead of $\mathbb{E}_{x\sim p_1}$ in our proofs.

Lemma 2. *For two arbitrary distributions p and q, we have*

$$\mathbb{E}_{x\sim p}[H(x;q) - H(x;p)] = D_F(p\|q),$$

where $D_F(\cdot\|\cdot)$ is the Fisher divergence between two distributions.

Proof. Following [31], it has been shown that

$$D_F(p\|q) := \mathbb{E}_p\left[\|\nabla\log p(x) - \nabla\log q(x)\|_2^2\right]$$
$$= \mathbb{E}_p\left[\frac{1}{2}\|\nabla\log p(x)\|_2^2 + H(x;q)\right],$$

therefore

$$\mathbb{E}_p[H(x;q) - H(x;p)]$$
$$= D_F(p\|q) - \mathbb{E}_p\left[\frac{1}{2}\|\nabla\log p(x)\|_2^2\right]$$
$$\underbrace{-D_F(p\|p)}_{=0} + \mathbb{E}_p\left[\frac{1}{2}\|\nabla\log p(x)\|_2^2\right]$$
$$= D_F(p\|q).$$

□

Lemma 3. *For two score function $\nabla\log p$ and $\nabla\log q$, define the following two terms w.r.t. the random variable x,*

$$A(x) = \|\nabla\log p(x) - \nabla\log q(x)\|_2,$$
$$B(x) = \|\nabla\log p(x) + \nabla\log q(x)\|_2.$$

Then for a distribution π over $\mathcal{X}$, there is

$$\frac{1}{2}\mathbb{E}_\pi\left[\|\nabla\log p(x)\|_2^2 - \|\nabla\log q(x)\|_2^2\right]$$
$$\leq \frac{1}{2}\mathbb{E}_\pi[A(x)]\mathbb{E}_\pi[B(x)] + \frac{1}{2}\operatorname{Var}_\pi[A(x)]\operatorname{Var}_\pi[B(x)].$$

Proof. Starting from the left-hand side, we have

$$\frac{1}{2}\mathbb{E}_\pi\left[\|\nabla\log p(x)\|_2^2 - \|\nabla\log q(x)\|_2^2\right]$$
$$= \frac{1}{2}\mathbb{E}_\pi\left[\nabla\log p(x)^\top\nabla\log p(x) - \nabla\log q(x)^\top\nabla\log q(x)\right]$$
$$= \frac{1}{2}\mathbb{E}_\pi\langle\nabla\log p(x) + \nabla\log q(x), \nabla\log p(x) - \nabla\log p(x)\rangle$$
$$\leq \frac{1}{2}\mathbb{E}_\pi[A(x)B(x)]$$
$$\leq \frac{1}{2}\mathbb{E}_\pi[A(x)]\mathbb{E}_\pi[B(x)] + \frac{1}{2}\operatorname{Cov}_\pi(A(x), B(x))$$
$$\leq \frac{1}{2}\mathbb{E}_\pi[A(x)]\mathbb{E}_\pi[B(x)] + \frac{1}{2}\operatorname{Var}_\pi[A(x)]\operatorname{Var}_\pi[B(x)].$$

Algorithm 2 DSM-CUSUM (Offline)

Require: Reference datasets $\mathcal{D}_0, \mathcal{D}_1$, threshold τ.
1: Train the two score models $\hat{s}_0, \hat{s}_1$ using the reference datasets by (4).
2: **while** $S_t \leq \tau$ **do**
3:     $t \leftarrow t+1, \nu \leftarrow t$.
4:     Compute test statistics $\mathcal{S}_t$ as in (5).
5: **end while**
6: **return** Detected change point ν

Note that the first and third inequality both follow from the Cauchy-Schwartz inequality. □

Next, we provide the complete proof for Lemma 1.

Proof of Lemma 1. Given the definition of the proposed statistics in (5), we can make the following decomposition,

$$\mathbb{E}_1[\Delta(x)]$$
$$= \mathbb{E}_1[H(x;\hat{p}_0) - H(x;p_{0,\sigma})] + \mathbb{E}_1[H(x;p_{0,\sigma}) - H(x;p_0)]$$
$$+ \mathbb{E}_1[H(x;p_{1,\sigma}) - H(x;\hat{p}_1)] + \mathbb{E}_1[H(x;p_1) - H(x;p_{1,\sigma})]$$
$$+ \mathbb{E}_1[H(x;p_0) - H(x;p_1)].$$

By Assumption 2, we can derive that, with probability greater than $1-\delta$,

$$\mathbb{E}_1\|\hat{s}_0(x) - \nabla\log p_{0,\sigma}(x)\|_2 \leq \epsilon_{\text{est}}(\sigma),$$
$$\mathbb{E}_1\|\hat{s}_0(x) + \nabla\log p_{0,\sigma}(x)\|_2 \leq 2M,$$
$$\operatorname{Var}_1\|\hat{s}_0(x) - \nabla\log p_{0,\sigma}(x)\|_2 \leq C_1\epsilon_{\text{est}}^2(\sigma),$$
$$\operatorname{Var}_1\|\hat{s}_0(x) + \nabla\log p_{0,\sigma}(x)\|_2 \leq 4M(M+1).$$

Therefore by Lemma 3,

$$\mathbb{E}_1[H(x;\hat{p}_0) - H(x;p_{0,\sigma})]$$
$$\leq M\epsilon_{\text{est}}(\sigma) + 2M(M+1)C_1\epsilon_{\text{est}}^2(\sigma).$$

Following similar procedures, we can derive the upper bound for the second, third and fourth term to obtain bounds of similar forms. For the fifth term, we can use Lemma 2 and get

$$\mathbb{E}_1[H(x;p_0) - H(x;p_1)] = D_F(p_1\|p_0).$$

Combining all previous derivations, and by the definition of ϵ_{div} in Assumption 2, we can get with probability greater than $1-\delta$,

$$|\mathbb{E}_1[\Delta(x)] - D_F(p_1\|p_0)|$$
$$\leq 2M[\epsilon_{\text{est}}(\sigma) + \epsilon_{\text{pert}}(\sigma)] + \epsilon_{\text{div}}$$
$$+ 4M(M+1)\cdot\max\{C_1, C_2\}\cdot[\epsilon_{\text{est}}^2(\sigma) + \epsilon_{\text{pert}}^2(\sigma)].$$

This completes the proof. □

Before we state the proof of Theorem 1 and Theorem 2, we introduce the following statistics of random walk:

$$\mathcal{S}'_t = \sum_{i=1}^{t}\Delta_i, \quad \forall t = 1,\ldots,N. \tag{9}$$

Then we define the corresponding stopping time as $T' := \inf\{t : \mathcal{S}'_t \geq \tau\}$ given some threshold τ. It has been proved

in Lemma 4 and Theorem 1 of [17] that the following upper-bound relationship holds:

$$\text{WADD} \leq \mathbb{E}_1[T] \leq \mathbb{E}_1[T'], \quad (10)$$

therefore, we only need to upper-bound the expected stopping time of the proxy statistics (9) in order to upper-bound the desired quantities in Theorem 1 and Theorem 2.

Proof of Theorem 1. Since $\mathcal{D}_0$ and $\mathcal{D}_1$ are independent, therefore under $x \overset{i.i.d.}{\sim} p_1$, the statistics $\Delta(x)$ are also *i.i.d.* samples. Hence using (10) and by Wald's identity:

$$\mathbb{E}_1[T'] = \frac{\mathbb{E}_1[\mathcal{S}'_{T'}]}{\mathbb{E}_1[\Delta(x)]}. \quad (11)$$

The numerator can be upper-bounded as

$$
\begin{aligned}
\mathbb{E}_1[\mathcal{S}'_{T'}] &= \tau + \mathbb{E}_1[\mathcal{S}'_{T'} - \tau] \\
&\leq \tau + \sup_{\tau \geq 0} \mathbb{E}_1[\mathcal{S}'_{T'} - \tau] \\
&\leq \tau + \frac{\mathbb{E}_1[(\max\{\Delta(x), 0\})^2]}{\mathbb{E}_1[\Delta]} \\
&\leq \tau + \frac{(\mathbb{E}_1[\Delta(x)])^2 + \text{Var}_1[\Delta(x)]}{\mathbb{E}_1[\Delta(x)]},
\end{aligned}
$$

where the second inequality follows from [34]. Plug back to (11) we get:

$$\mathbb{E}_1[T'] \leq \frac{\tau}{\mathbb{E}_1[\Delta(x)]} + \frac{\text{Var}_1[\Delta(x)]}{(\mathbb{E}_1[\Delta(x)])^2} + 1.$$

Using the simplified notation $\epsilon(\sigma)$ introduced in the main text, we have established in Lemma 1 that

$$|\mathbb{E}_1[\Delta(x)] - D_F(p_1\|p_0)| \leq \epsilon(\sigma), \quad w.p. \geq 1 - \delta,$$

therefore combining all previous derivations, we know that with probability greater than $1 - \delta$,

$$
\begin{aligned}
\text{WADD} &\leq \frac{\tau}{D_F(p_1\|p_0) - \epsilon(\sigma)} + \frac{\text{Var}_1[\Delta(x)]}{[D_F(p_1\|p_0) - \epsilon(\sigma)]^2} + 1 \\
&\frac{\tau}{D_F(p_1\|p_0) - \epsilon(\sigma)} + \frac{\mathbb{E}_1[\Delta(x)^2]}{[D_F(p_1\|p_0) - \epsilon(\sigma)]^2}.
\end{aligned}
$$

Since $\mathbb{E}_1[\Delta(x)^2] < +\infty$, therefore the RHS is finite. This finishes the proof. $\square$

Proof of Theorem 2. Again, throughout our derivation, we condition on the reference dataset. Notice that the statistics $\Delta(x_t)$ is now a w-dependent sequence [17], starting from the RHS of (10),

$$
\begin{aligned}
\mathbb{E}_1[\mathcal{S}'_{T'}] &:= \mathbb{E}_1\left[\sum_{t=w+1}^{T'} \Delta(x_t)\right] \\
&= \mathbb{E}_1\left[\sum_{t=w+1}^{T'+w} \Delta(x_t)\right] - \mathbb{E}_1\left[\sum_{t=T'+1}^{T'+w} \Delta(x_t)\right].
\end{aligned}
$$

Define filtration $\mathscr{F}_t$ as the σ-field generated by all samples up to time t, we can decompose the first term as,

$$
\begin{aligned}
&\mathbb{E}_1\left[\sum_{t=w+1}^{T'+w} \Delta(x_t)\right] \\
=&\mathbb{E}_1\left[\sum_{t=w+1}^{\infty} \Delta(x_t)\mathbb{1}_{\{T' \geq t-w\}}\right] \\
=&\mathbb{E}_1\left[\sum_{t=w+1}^{\infty} \mathbb{E}_1[\Delta(x_t)|\mathscr{F}_{t-1}]\mathbb{1}_{\{T' \geq t-w\}}\right] \\
\geq& [D_F(p_1\|p_0) - \epsilon(\sigma)] \cdot \mathbb{E}_1[T'], \quad w.p. \geq 1 - \delta.
\end{aligned}
$$

The second equality uses the tower property of expectations and $\mathbb{1}_{\{T' \geq t-w\}}$ is $\mathscr{F}_{t-1}$-measurable. The inequality uses the fact that since the σ-field generated by Δ is contained in $\mathscr{F}_{t-1}$ during online estimation, therefore, by Lemma 1,

$$|\mathbb{E}_1[\Delta(x_t)|\mathscr{F}_{t-1}] - D_F(p_1\|p_0)| \leq \epsilon(\sigma), \quad w.p. \geq 1 - \delta. \quad (12)$$

Similarly, for the second term, we can derive that,

$$
\begin{aligned}
&\mathbb{E}_1\left[\sum_{t=T'+1}^{T'+w} \Delta(x_t)\right] \\
=&\mathbb{E}_1\left[\sum_{t=w+1}^{\infty} \Delta(x_t)\mathbb{1}_{\{T' < t\}}\mathbb{1}_{\{T' \geq t-w\}}\right] \\
=&\mathbb{E}_1\left[\sum_{t=w+1}^{\infty} \mathbb{E}_1[\Delta(x_t)|\mathscr{F}_{t-1}]\mathbb{1}_{\{T' < t\}}\mathbb{1}_{\{T' \geq t-w\}}\right] \\
\leq& [D_F(p_1\|p_\infty) + \epsilon(\sigma)] \cdot w, \quad w.p. \geq 1 - \delta.
\end{aligned}
$$

Again, for the second equality, we are using the fact that $\mathbb{1}_{\{T' < t\}}$ and $\mathbb{1}_{\{T' \geq t-w\}}$ are $\mathscr{F}_{t-1}$-measurable. For the inequality, we use (12). Combining the two results and plug into the original expression we get

$$\mathbb{E}_1[T'] \leq \frac{\mathbb{E}_1[\mathcal{S}'_{T'} - \tau] + \tau + w[D_F(p_1\|p_\infty) + \epsilon(\sigma)]}{D_F(p_1\|p_0) - \epsilon(\sigma)},$$

where τ is the threshold value. Denote

$$
\begin{aligned}
c &= \mathbb{E}_1[\Delta(x)^2] / [D_F(p_1\|p_0) - \epsilon(\sigma)], \\
d &= w[D_F(p_1\|p_0) + \epsilon(\sigma)].
\end{aligned}
$$

Following similar derivations of bounding the overshoot term in Theorem 1 of [17], we can derive that

$$\mathbb{E}_1[\mathcal{S}'_{T'} - \tau] \leq c + c^{1/2}\tau^{1/2} + c^{1/2}d^{1/2},$$

Plugging this result into the right-hand side, then with probability greater than $1 - \delta$,

$$\text{WADD} \leq \frac{c + c^{1/2}\tau^{1/2} + \tau + c^{1/2}d^{1/2} + d}{D_F(p_1\|p_0) - \epsilon(\sigma)}.$$

Note that $\mathbb{E}[\Delta(x)^2] < \infty$, therefore the upper bound is finite. We have finished the proof. $\square$
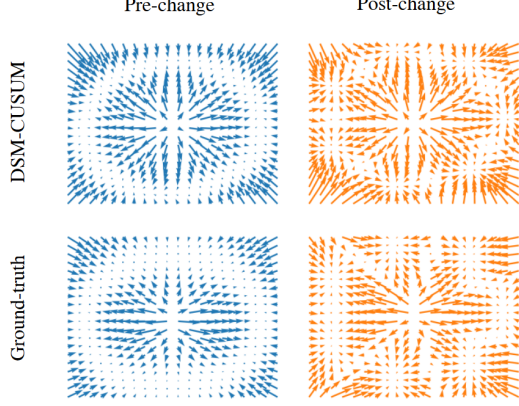
Fig. 4. Comparison of the estimated score function obtained using `DSM-CUSUM` against the ground-truth for the 2D Gaussian mixture dataset. The results demonstrate that the score models effectively capture the ground-truth score function.

APPENDIX B
ADDITIONAL EXPERIMENT DETAILS

In this section, we provide further details on the experimental setup described in Section V.

A. Threshold Simulation

The threshold τ is typically chosen to balance the trade-off between the probabilities of false alarms and successful detection, as determined by a target ARL value. To reduce computational effort in determining thresholds for large target ARL values, we employ an efficient approximation algorithm. This algorithm leverages the fact that the stopping time T under the pre-change regime approximately follows an exponential distribution when the ARL is large. Such approximation methods are widely utilized in online change detection.

The high-level idea of the procedure is that, instead of simulating the mean of the distribution of $T := \inf\{t : \mathcal{S}_t \geq \tau\}$ directly, we obtained an estimate of the mean from an estimate of the cumulative distribution function of T based on N_1 iterations. Specifically, in each iteration, we simulate the pre-change trajectory with N_2 time steps, and compute the maximum of the detection statistics at these N_2 time steps. These maximum values under N_1 iterations are then denoted as $W_{1,\max}, W_{2,\max}, \ldots, W_{N_1,\max}$. For the desired ARL values γ, we approximate the stopping time T as an exponential distribution with mean γ. Thus we have $P(W_{\max} < \tau) = P(T > N_2) \approx e^{-N_2/\gamma}$. Thus the corresponding threshold τ can be approximated as the $e^{-N_2/\gamma}$ quantile of the set $\{W_{1,\max}, W_{2,\max}, \ldots, W_{N_1,\max}\}$. We use $N_1 = 200$, and $N_2 = 1000$. Note that we can also use more iterations and longer sequences within each iteration, which tends to improve the approximation accuracy.

B. Synthetic Data

The two synthetic datasets used in Section V-A are defined as follows:

Fig. 5. Visualization of the two datasets. The left column depicts 2D Gaussian mixture data, while the right column represents 10D neural network data (visualized using t-SNE).

a) 2D Gaussian mixture data: Both pre- and post-change distributions are specified by 2-dimensional Gaussian mixture models, where the pre-change distribution has 30 components, and the post-change distribution has 8 components, each of the means is evenly distributed around the circle centered at the origin with a radius of 3. The ground-truth pre- and post-change distributions are specified as:

$$p_1 = \sum_{i=1}^{3} \frac{1}{30} \mathcal{N}\left(\begin{bmatrix} 8\cos\frac{i}{15}\pi - \frac{1}{2} \\ 8\sin\frac{i}{15}\pi - \frac{1}{2} \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right);$$

$$p_\infty = \sum_{i=1}^{3} \frac{1}{8} \mathcal{N}\left(\begin{bmatrix} 8\cos\frac{i}{4}\pi + \frac{1}{2} \\ 8\sin\frac{i}{4}\pi + \frac{1}{2} \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right).$$

With the Gaussian mixture model, the ground-truth data distribution and score functions are readily available, enabling the deployment of the exact CUSUM and SCUSUM methods as described in [28]. We introduce a small $\frac{1}{2}$-size offset to prevent the data from being indistinguishable for detection methods based on first and second moments.

b) 10D Neural Network data: We first generate standard Gaussian data in 4D space, then transform the data using two sets of neural networks, whose weights and biases are randomly initialized according to standard Gaussian. To make it more challenging, we additionally deploy a mean and covariance matrix matching algorithm to make the data challenging and indistinguishable. The loss function is denoted by

$$\min_{\theta} \|\widehat{\mu}_1 - \widehat{\mu}_\infty\|_2^2 + \|\widehat{\Sigma}_1 - \widehat{\Sigma}_\infty\|_2^2,$$

where θ denotes the parameters of the two neural networks, and $\widehat{\mu}_j, \widehat{\Sigma}_j, j = 1, \infty$ denotes the sample mean and covariance matrix of the generated pre-change and post-change data.

C. Real Data & Result Implications

The objective of earthquake precursor detection is to detect any relevant earthquake signals before the earthquake starts (*e.g.*, three months in advance). The final cleaned data contains $17,521$ data points, covering a two-year time interval starting from January 01, 2013, to January 01, 2015.

Our results on the real data have two key practical implications: (i) This experiment retrospectively validates the judgment of domain experts and provides a robust foundation for further seismological investigations into the mechanism underlying this precursor signal, advancing earthquake prediction research. (ii) The experiment highlights the potential generalizability of our

method for detecting precursor signals to predict earthquakes. Early detection—*e.g.*, up to three months in advance—provides local governments and residents sufficient time to prepare and evacuate, significantly reducing potential losses. This underscores the method's high value in disaster mitigation and other high-stakes scenarios.

D. Baseline Models

Our baseline models fall in the broad category of the widely used cumulative sum (CUSUM) test statistic [25], which is also the optimal test, defined as:

$$S_t = \max_{0 \le k \le t} \sum_{j=k+1}^{t} \log \frac{p_1(x_j)}{p_0(x_j)} = S_{t-1}^+ + \log \frac{p_1(x_t)}{p_0(x_t)}. \quad (13)$$

The stopping rule is defined by $T = \inf \{t : \mathcal{S}_t \ge \tau\}$. When p_0 and p_1 are known, we refer to (13) as the exact CUSUM test.

a) Gaussian CUSUM: The pre-change and post-change distribution are parameterized as two multivariate Gaussian PDFs $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$. The log-likelihood used in (13) can be simplified as:

$$\log \frac{\mathcal{N}(\mu_1, \Sigma_1)}{\mathcal{N}(\mu_0, \Sigma_0)} = \frac{1}{2} \log \left(\frac{|\Sigma_0|}{|\Sigma_1|} \right) + \frac{1}{2} \left[(x - \mu_0)^\top \Sigma_0^{-1} (x - \mu_0) \right]$$
$$- \frac{1}{2} \left[(x - \mu_1)^\top \Sigma_1^{-1} (x - \mu_1) \right].$$

The parameters $\mu_0, \mu_1, \Sigma_0, \Sigma_1$ are fitted via standard maximum likelihood estimation (MLE).

b) Gaussian Mixture CUSUM: The pre-change and post-change distributions are parameterized as two Gaussian mixture distributions with n components,

$$p_0 = \sum_{i=1}^{n} \pi_i^{(0)} \mathcal{N} \left(\mu_i^{(0)}, \Sigma_i^{(0)} \right), \quad p_1 = \sum_{i=1}^{n} \pi_i^{(1)} \mathcal{N} \left(\mu_i^{(1)}, \Sigma_i^{(1)} \right).$$

The parameters $\pi_i^{(0)}, \mu_i^{(0)}, \Sigma_i^{(0)}, \pi_i^{(1)}, \mu_i^{(1)}, \Sigma_i^{(1)}$ are fitted using the expectation-maximization (EM) algorithm.

APPENDIX C
ADDITIONAL BACKGROUNDS

We provide more advanced background information related to our proposed method.

Score-based Models. Denoising score matching is later generalized in SMLD [22] and Score SDEs [32] to extend the framework of diffusion models [35]. Specifically, when there are $n > 1$ noise levels $\{\sigma_i\}_{i=1}^{n}$ and allowing for the noise level as a conditional input to the parameterized score functions $s(x, \sigma_i)$ to create a continuous ($n \to \infty$) or discrete (finite n) diffusion process

$$\mathbb{E}_i \mathbb{E}_{p(x)} \mathbb{E}_{q_\sigma(x^{(i)}|x)} \left[\lambda_i \left\| s(x^{(i)}, \sigma_i) - \nabla_{x^{(i)}} \log \mathcal{K}_{\sigma_i}(x^{(i)}|x) \right\| \right].$$

where $\mathcal{K}_{\sigma_i}(\cdot|\cdot)$ denotes a known transition kernel parameterized by noise scale σ_i. λ_i are coefficients used to scale the value of errors and can typically be chosen as $\lambda_i \propto \sigma_i^2$ Empirically,

when the training process converges well, for all σ_i, there is $s(x, \sigma_i) \to \nabla_x \log p_{\sigma_i}(x)$.

In the meantime, to make score matching scalable in high-dimension settings, sliced score matching was proposed in [20]. Denote random projection direction $x \in \mathbb{R}^d$, and their distribution P_x. We draw $\mathbf{v}_1, \ldots, \mathbf{v}_M \stackrel{i.i.d.}{\sim} P_\mathbf{v}$ and compute,

$$\mathcal{L}(\mathbf{v}) := \frac{1}{T} \sum_{t=1}^{T} \sum_{m=1}^{M} \left[\mathbf{v}_m^\top J s(x_t) \mathbf{v}_m + \frac{1}{2} \left(\mathbf{v}_m^\top s(x_t) \right)^2 \right],$$

where J denotes the Jacobian matrix operator. Typically $P_\mathbf{v}$ can be specified as simple distributions such as standard multivariate Gaussian distribution, and $M = 1$ suffice in practice.