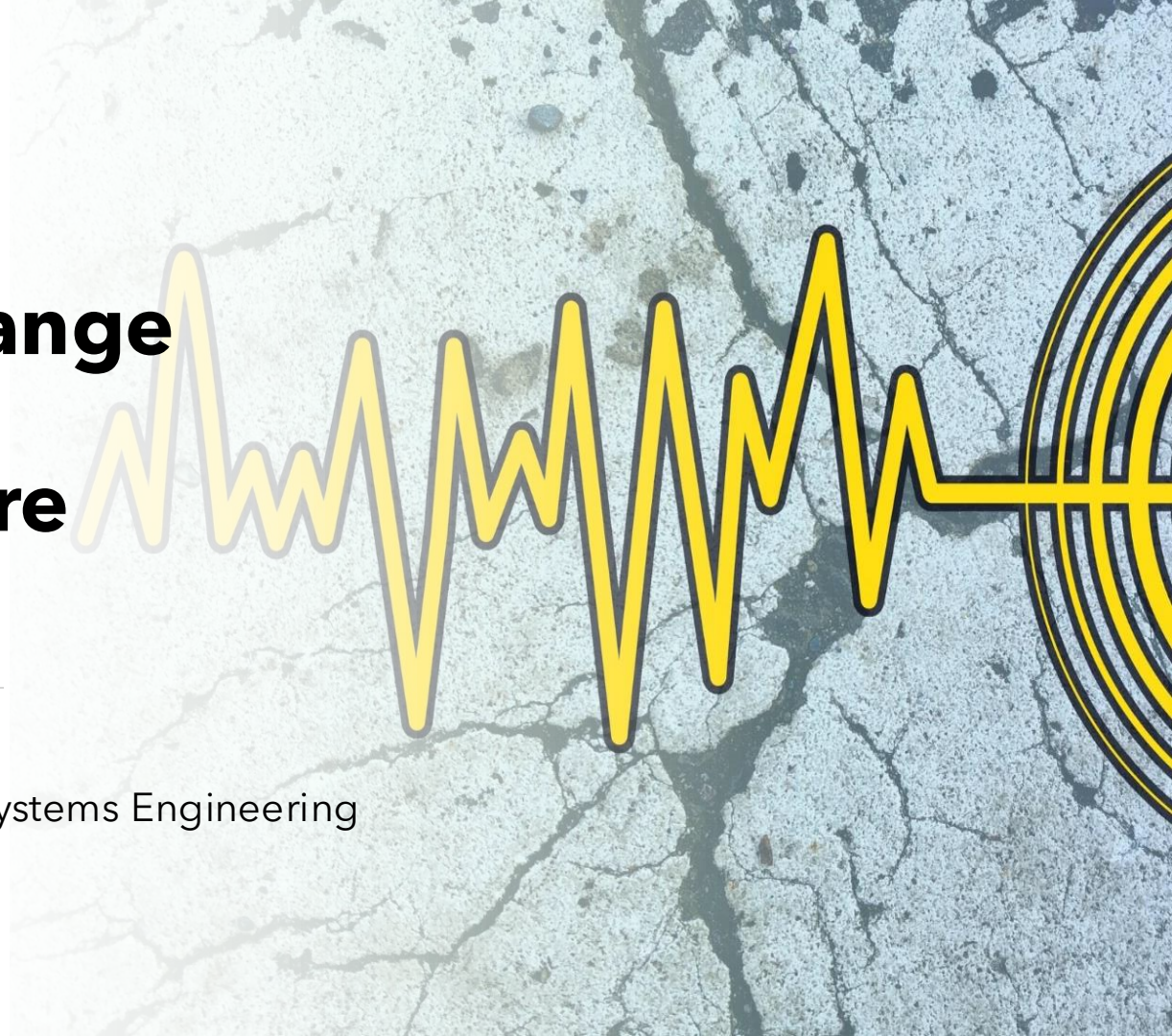
Sequential Change Detection via Denoising Score Matching

Liyan Xie
Department of Industrial and Systems Engineering
University of Minneasota
Allerton 2025



This is joint work with

Wenbin Zhou
Carnegie Mellon University

Woody (Shixiang) Zhu
Carnegie Mellon University

Zhigang Peng
Georgia Institute of Technology

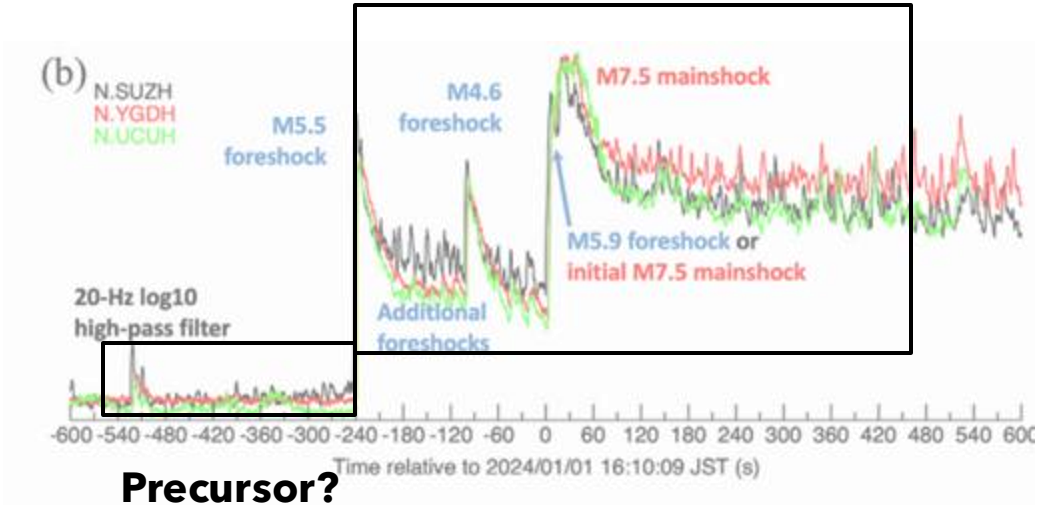
Outline

- Motivating example and problem setup
- Proposed method: Denoising Score Matching CUSUM
- Theoretical analysis
- Simulation and real data results

Motivating example: earthquake precursor

Modern data acquisition enables high-resolution catalogues and high-dimensional geophysical monitoring signals to be collected.

Seismic events



Precursor?

(signals detected prior to an earthquake)

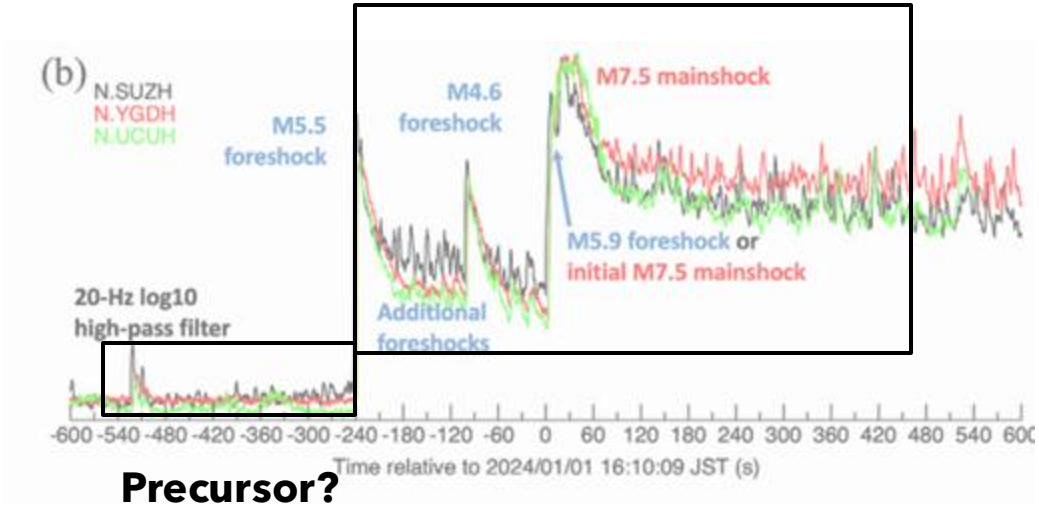
Motivating example: earthquake precursor

Modern data acquisition enables high-resolution catalogues and high-dimensional geophysical monitoring signals to be collected.

Challenges:

- Signals embedded in high-dimensional data are hard to capture
- The distribution before/after change are **unknown** and **difficult to estimate**.

Seismic events



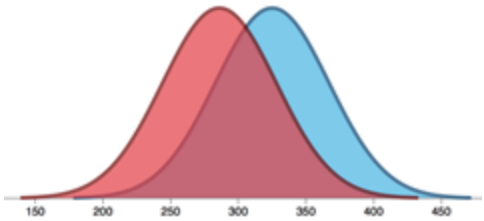
Precursor?

(signals detected prior to an earthquake)

Insights

Idea: capture the distributional differences **without** estimating the distribution densities

Difference of distributions



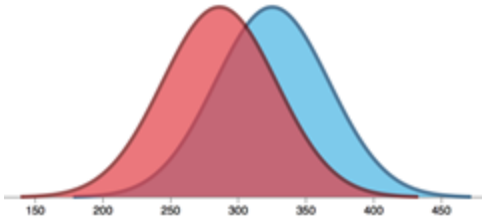
$p_1(x)$ vs. $p_0(x)$

Estimating densities p_0, p_1 may be hard

Insights

Idea: capture the distributional differences **without** estimating the distribution densities

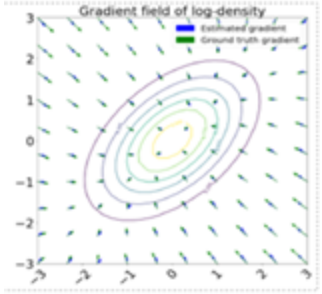
Difference of distributions



$p_1(x)$ vs. $p_0(x)$

Estimating densities p_0, p_1 may be hard

Difference of the **gradient of the log density (score)**



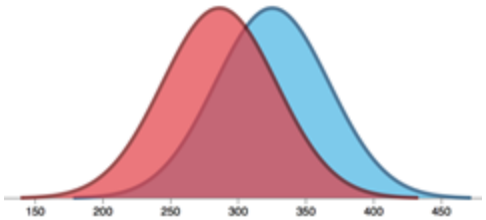
$s_1(x) = \nabla_x \log p_1(x)$ vs. $s_0(x) = \nabla_x \log p_0(x)$

Estimating scores s_0, s_1 are typically **easier**

Insights

Idea: capture the distributional differences **without** estimating the distribution densities

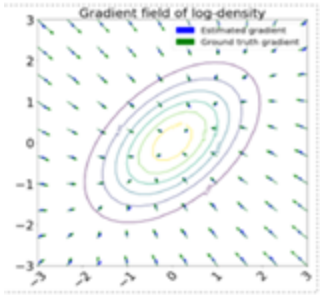
Difference of distributions



$p_1(x)$ vs. $p_0(x)$

Estimating densities p_0, p_1 may be hard

Difference of the **gradient of the log density (score)**



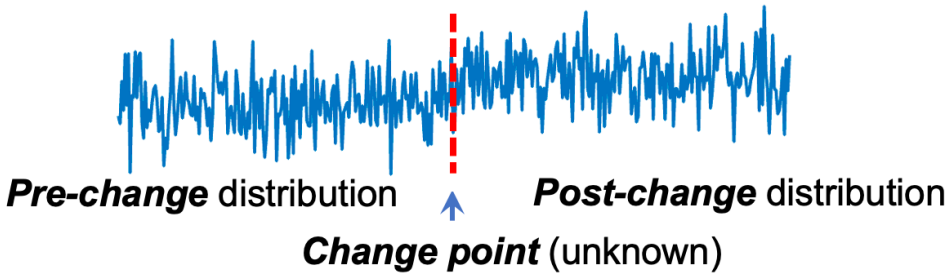
$s_1(x) = \nabla_x \log p_1(x)$ vs. $s_0(x) = \nabla_x \log p_0(x)$

Estimating scores s_0, s_1 are typically **easier**

Example (unnormalized models): $p_\theta(x) = \exp(F_\theta(x))/Z_\theta \iff s_\theta(x) = F_\theta{}'(x)$

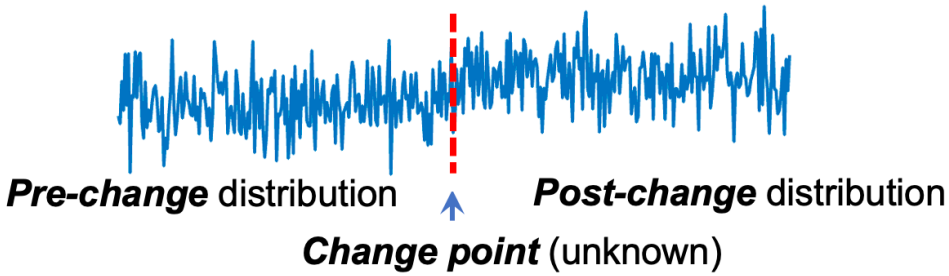
Wu, S., Diao, E., Banerjee, T., Ding, J., & Tarokh, V. (2023). Quickest change detection for unnormalized statistical models. IEEE TIT.

Problem setup



$x_1, x_2, \ldots, x_{\tau-1} \sim^{iid} p_0$ (unknown), $x_\tau, x_{\tau+1}, \ldots \sim^{iid} p_1$ (unknown)

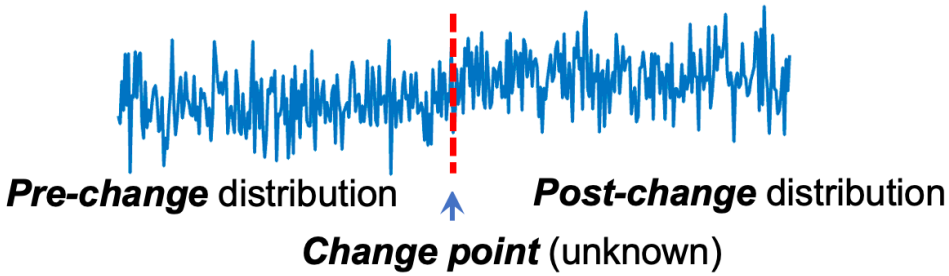
Problem setup



$$x_1, x_2, \dots, x_{\tau-1} \sim^{iid} p_0 \text{ (unknown)}, \quad x_\tau, x_{\tau+1}, \dots \sim^{iid} p_1 \text{ (unknown)}$$

- Assume reference datasets D_0 and D_1 sampled from p_0 and p_1 (typically available by pre-collected historical data).

Problem setup



$$x_1, x_2, \dots, x_{\tau-1} \sim^{iid} p_0 \text{ (unknown)}, \quad x_\tau, x_{\tau+1}, \dots \sim^{iid} p_1 \text{ (unknown)}$$

- Assume reference datasets D_0 and D_1 sampled from p_0 and p_1 (typically available by pre-collected historical data).
- **Goal**: detect the unknown change-point $\boldsymbol{\tau}$ as quickly as possible.

Main component: Score-Matching

When the density p is either unavailable or difficult to estimate, we can estimate its score function instead.

- Score: $s(x) = \nabla_x \log p(x)$
- Can be reliability estimated via Score-Matching:

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.

Main component: Score-Matching

When the density p is either unavailable or difficult to estimate, we can estimate its score function instead.

- Score: $s(x) = \nabla_x \log p(x)$
- Can be reliability estimated via Score-Matching:
  - Given training data $x_1, \ldots, x_N \sim p(x)$

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.

Main component: Score-Matching

When the density p is either unavailable or difficult to estimate, we can estimate its score function instead.

- Score: $s(x) = \nabla_x \log p(x)$
- Can be reliability estimated via Score-Matching:
  - Given training data $x_1, \ldots, x_N \sim p(x)$
  - Estimate the score function by

$$\hat{s}(x) = s\left(x; \hat{\theta}\right) = \arg\min \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} [tr(\nabla_x s(x_i; \theta)) + \frac{1}{2} \|s(x_i; \theta)\|^2]$$

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.

Main component: Score-Matching

When the density p is either unavailable or difficult to estimate, we can estimate its score function instead.

- Score: $s(x) = \nabla_x \log p(x)$
- Can be reliability estimated via Score-Matching:
  - Given training data $x_1, \ldots, x_N \sim p(x)$
  - Estimate the score function by

$$\hat{s}(x) = s\left(x; \hat{\theta}\right) = \arg\min \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} [tr(\nabla_x s(x_i; \theta)) + \frac{1}{2} \|s(x_i; \theta)\|^2]$$

  - $s(x; \theta)$ is typically parameterized as a neural-network

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.

Main component: Score-Matching

When the density p is either unavailable or difficult to estimate, we can estimate its score function instead.

- Score: $s(x) = \nabla_x \log p(x)$
- Can be reliability estimated via Score-Matching:
  - Given training data $x_1, \ldots, x_N \sim p(x)$
  - Estimate the score function by

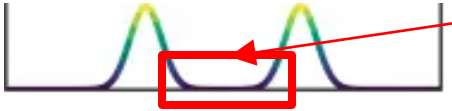
$$\hat{s}(x) = s(x; \hat{\theta}) = \arg\min \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} [tr(\nabla_x s(x_i; \theta)) + \frac{1}{2} \|s(x_i; \theta)\|^2]$$

  - $s(x; \theta)$ is typically parameterized as a neural-network
  - Theoretically guaranteed to be consistent estimator

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.

Denoising Score Matching (DSM)

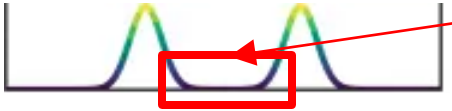
Vanilla score matching is not good enough for hign-dim data



Low Densities
(on which we cannot
estimate the scores well)

Denoising Score Matching (DSM)

Vanilla score matching is not good enough for hign-dim data



Low Densities
(on which we cannot estimate the scores well)

Noise injection

Noisy data: $x' \sim \mathcal{K}(x'|x)$

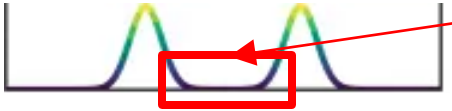
Now we can estimate the (noisy) scores well at all positions

Estimate the "noisy" score

$$\mathbb{E}_{x\sim p}\mathbb{E}_{x'\sim\mathcal{K}(\cdot|x)}\left\|s'(x';\theta)-\nabla_{x'}\log\mathcal{K}(x'|x)\right\|_2^2 .$$ *Minimize MSE*

Denoising Score Matching (DSM)

Vanilla score matching is not good enough for hign-dim data



Low Densities
(on which we cannot estimate the scores well)

Noise injection

Noisy data: $x' \sim \mathcal{K}(x'|x)$

Now we can estimate the (noisy) scores well at all positions

Estimate the "noisy" score

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

Minimize MSE

Model (Score network)

Noise

Our Algorithm: DSM-CUSUM

I. Estimate score function $\hat{s}_0, \hat{s}_1$ via DSM using reference data D_0 (pre-change) and D_1 (post-change)

Our Algorithm: DSM-CUSUM

I. Estimate score function $\hat{s}_0, \hat{s}_1$ via DSM using reference data D_0 (pre-change) and D_1 (post-change)

II. Recursively compute the detection statistics:

- Initialize $\mathcal{S}_0 = 0$; For $t = 1, 2, 3, \ldots$

$$\Delta(x_t) = H(x_t; \widehat{s}_0) - H(x_t; \widehat{s}_1)$$

$$\mathcal{S}_t = \mathcal{S}_{t-1}^{+} + \Delta(x_t)$$

Our Algorithm: DSM-CUSUM

I. Estimate score function $\hat{s}_0, \hat{s}_1$ via DSM using reference data D_0 (pre-change) and D_1 (post-change)

II. Recursively compute the detection statistics:

- Initialize $\mathcal{S}_0 = 0$; For $t = 1, 2, 3, \ldots$

$$H(x; s) = \operatorname{div} s(x) + \tfrac{1}{2}\|s(x)\|_2^2$$

Difference of (Hyvarinen) score function

$$\Delta(x_t) = H(x_t; \widehat{s}_0) - H(x_t; \widehat{s}_1)$$

"distributional difference"

$$\mathcal{S}_t = \mathcal{S}_{t-1}^{+} + \Delta(x_t)$$

Our Algorithm: DSM-CUSUM

I. Estimate score function $\hat{s}_0, \hat{s}_1$ via DSM using reference data D_0 (pre-change) and D_1 (post-change)

II. Recursively compute the detection statistics:

- Initialize $\mathcal{S}_0 = 0$; For $t = 1, 2, 3, \ldots$

$$H(x; s) = \operatorname{div} s(x) + \tfrac{1}{2}\|s(x)\|_2^2$$

Difference of (Hyvarinen) score function

$$\Delta(x_t) = H(x_t; \widehat{s}_0) - H(x_t; \widehat{s}_1)$$

"distributional difference"

$$\mathcal{S}_t = \mathcal{S}_{t-1}^{+} + \Delta(x_t)$$

Accumulate the evidences

Our Algorithm: DSM-CUSUM

I. Estimate score function $\hat{s}_0, \hat{s}_1$ via DSM using reference data D_0 (pre-change) and D_1 (post-change)

II. Recursively compute the detection statistics:

- Initialize $\mathcal{S}_0 = 0$; For $t = 1, 2, 3, \ldots$

$$H(x; s) = \operatorname{div} s(x) + \tfrac{1}{2}\|s(x)\|_2^2$$

Difference of (Hyvarinen) score function

$$\Delta(x_t) = H(x_t; \widehat{s}_0) - H(x_t; \widehat{s}_1)$$

"distributional difference"

$$\mathcal{S}_t = \mathcal{S}_{t-1}^{+} + \Delta(x_t)$$

Accumulate the evidences

III. Raise alarm when $\mathcal{S}_t$ exceeds some threshold value:

$$T = \inf\{t \in \mathbb{N} : \mathcal{S}_t \geq \tau\}$$

Extension to online score matching

- When the post-change training data is not available
- Sequentially estimate the post-change score function:

Extension to online score matching

- When the post-change training data is not available
- Sequentially estimate the post-change score function:
  - Initialize $\mathcal{S}_0 = \cdots = \mathcal{S}_w = 0$, $\hat{\theta}_w = \theta_0$;

Extension to online score matching

- When the post-change training data is not available
- Sequentially estimate the post-change score function:
  - Initialize $\mathcal{S}_0 = \cdots = \mathcal{S}_w = 0$, $\hat{\theta}_w = \theta_0$;
  - For $t = w+1, \ldots$ update the score function parameter:

Extension to online score matching

- When the post-change training data is not available
- Sequentially estimate the post-change score function:
  - Initialize $S_0 = \cdots = S_w = 0$, $\hat{\theta}_w = \theta_0$;
  - For $t = w+1, \ldots$ update the score function parameter:

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

"denoising score matching objective" on most recent w samples

Extension to online score matching

- When the post-change training data is not available
- Sequentially estimate the post-change score function:
  - Initialize $\mathcal{S}_0 = \cdots = \mathcal{S}_w = 0$, $\hat{\theta}_w = \theta_0$;
  - For $t = w+1, \ldots$ update the score function parameter:

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

"denoising score matching objective" on most recent w samples

  - Update the detection statistics $\mathcal{S}_t$ similarly with $\Delta(x_t) = H(x_t; \hat{s}_0) - H(x_t; \hat{s}_{\widehat{\theta}_t})$, and compare to detection threshold

Related work

Sequential change detection

- Seminal works [Page, 1954] [Lorden, 1971] [Pollak, 1985] [Moustakides, 1986] [Lai, 1998], etc.
- Recent books [Tartakovsky et al, 2015] [Tartakovsky, 2020], etc.
- Data-driven and non-parametric methods (based on density estimation): [Moustakides and Basioti, 2019] [Liang and Veeravalli, 2023] [Li et al, 2015], etc.

Score-method method for sequential change detection

- Hyvärinen score-based CUSUM (SCUSUM) [Wu et al, 2023] (a series of work)
- Bayesian settings [Banerjee and Tarokh, 2024]
- Robust score-based cumulative sum (RSCUSUM) algorithm [Moushegian et al, 2025]

Theoretical guarantee

Theorem (Zhou et al 2025)

Under the offline score estimation setting, for a given threshold b, and under regularity assumptions, we have with high probability,

$$\text{Detection Delay} \leq \frac{b}{D_F(p_1 \| p_0) - \epsilon_{DSM}} (1 + o(1))$$

Theoretical guarantee

Theorem (Zhou et al 2025)

Under the offline score estimation setting, for a given threshold b, and under regularity assumptions, we have with high probability,

$$\text{Detection Delay} \leq \frac{b}{D_F(p_1 \| p_0) - \epsilon_{DSM}} (1 + o(1))$$

Fisher divergence

Theoretical guarantee

Theorem (Zhou et al 2025)

Under the offline score estimation setting, for a given threshold b, and under regularity assumptions, we have with high probability,

$$\text{Detection Delay} \leq \frac{b}{D_F(p_1 \| p_0) - \epsilon_{DSM}}(1 + o(1))$$

Fisher divergence

Error term caused by DSM

Theoretical guarantee

Theorem (Zhou et al 2025)

Under the offline score estimation setting, for a given threshold b, and under regularity assumptions, we have with high probability,

$$\text{Detection Delay} \leq \frac{b}{D_F(p_1 \| p_0) - \epsilon_{DSM}}(1 + o(1))$$

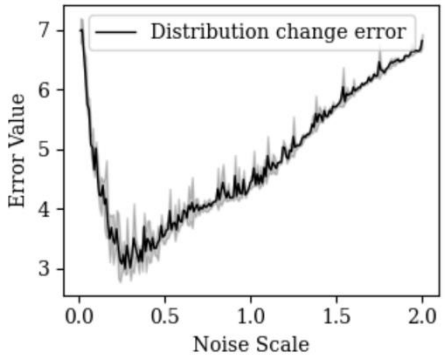
Fisher divergence

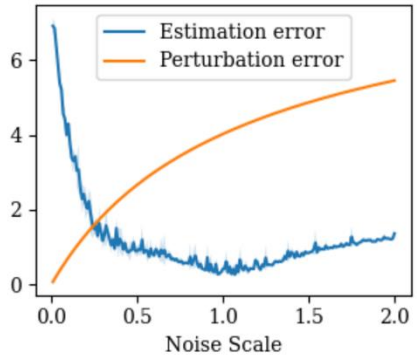
Error term caused by DSM

The error term ϵ_{DSM} contains two components:

- The score estimation error
- The distribution perturbation error

Oringinal distribution *Perturbed distribution*





Theoretical guarantee

Theorem (Zhou et al 2025)

Under the offline score estimation setting, for a given threshold b, and under regularity assumptions, we have with high probability,

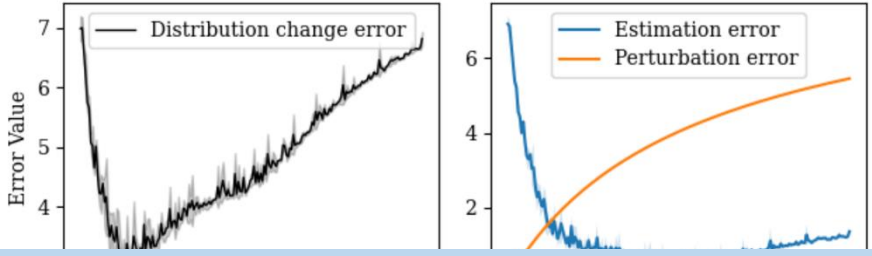
$$\text{Detection Delay} \leq \frac{b}{D_F(p_1 \| p_0) - \epsilon_{DSM}}(1 + o(1))$$

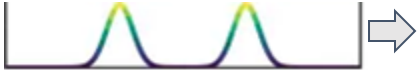
Fisher divergence

Error term caused by DSM

The error term ϵ_{DSM} contains two components:

- The score estimation error
- The distribution perturbation error

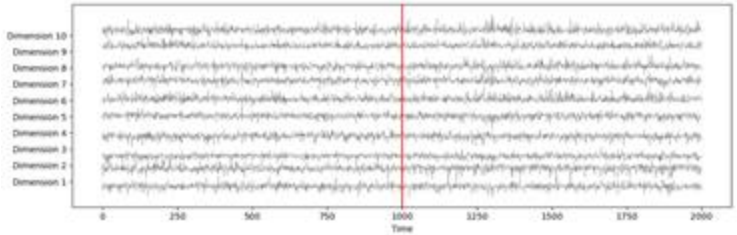


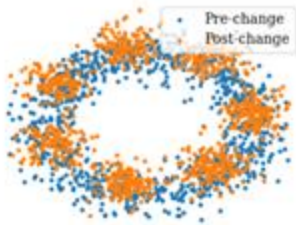


Oringinal distribution

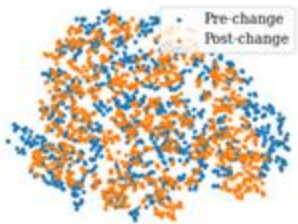
Takeaway: it's better to learn score functions more accurately when using an appropriate level of noise injection.

Simulation Experiments



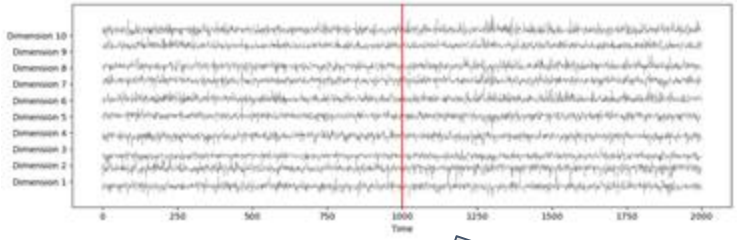


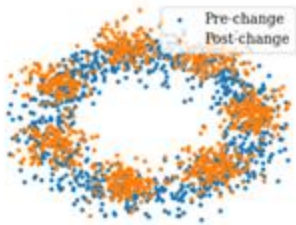
From Gaussian mixture

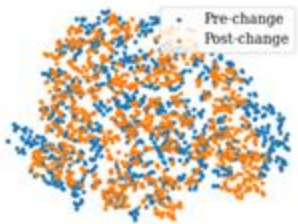


From Neural Network

Simulation Experiments

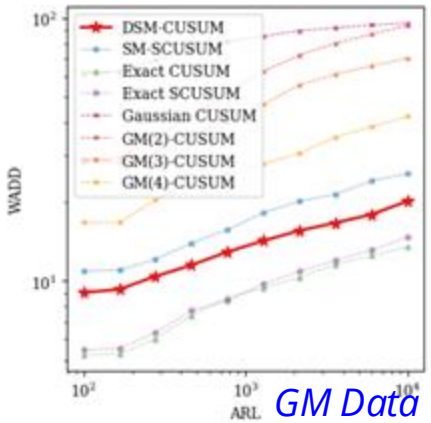




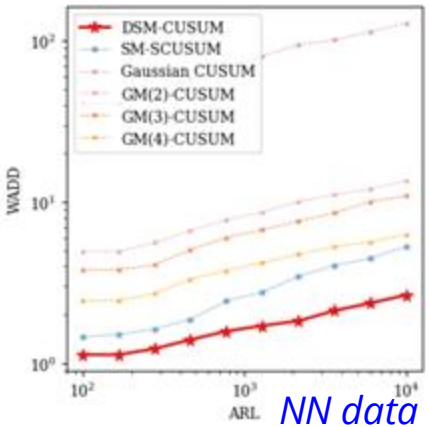


From Gaussian mixture

From Neural Network



GM Data



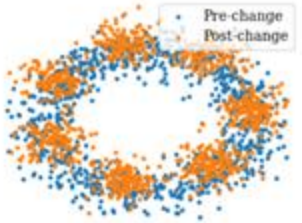
NN data

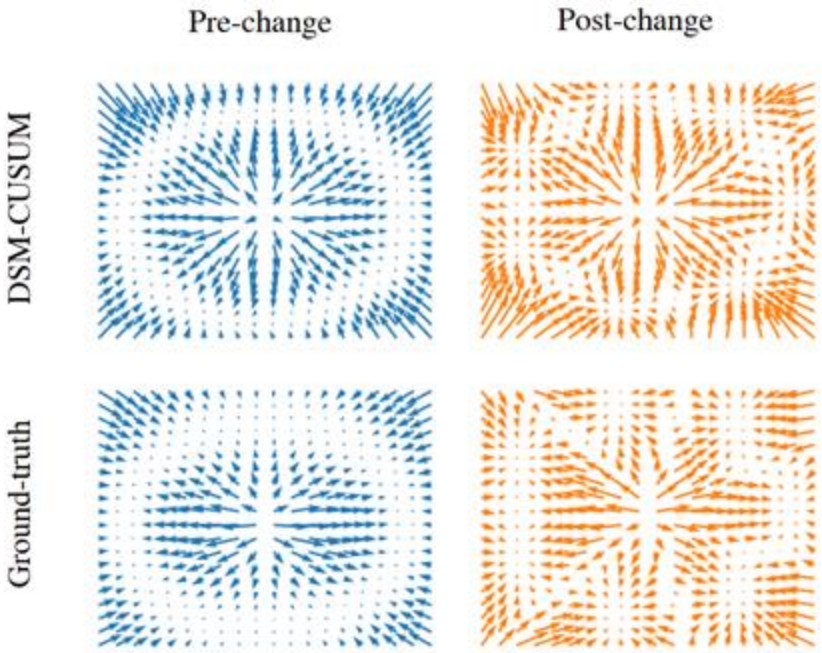
Detection Delay vs. False Alarm Measures

Lower, the better

Simulation Experiments: why it works?

Our method works because it can accurately learn the score function.





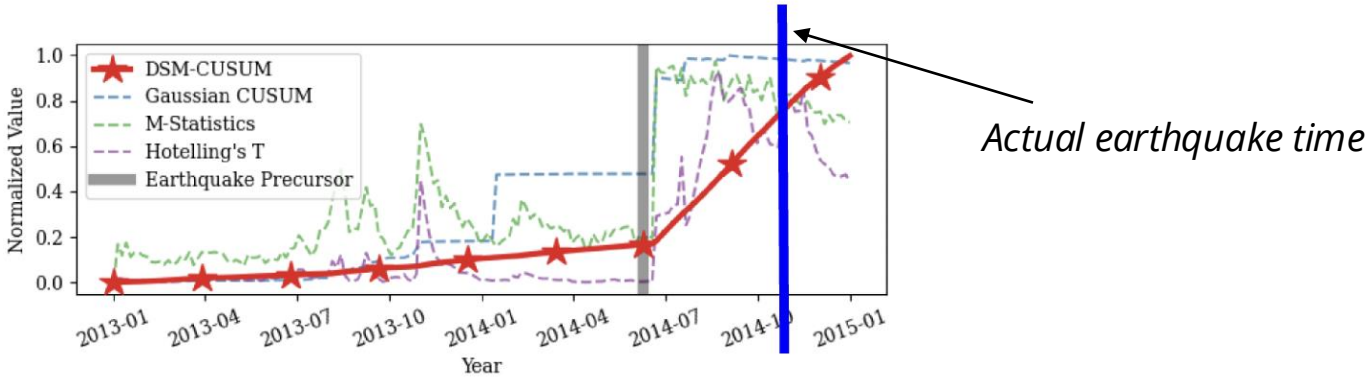
Estimated score function

True score function

A real example

- A moderate-size earthquake in 2014

- Four types of signals (water temperatures and water levels) hourly collected at three monitoring stations, over a two-year period.



Conclusion

- Proposed a denoising score matching based algorithm for sequential change detection
- Validated efficiency of our method on both synthetic and real world data
- Future directions: detection for multi-modal data, such as spatio-temporal data, graph data, etc.

Wenbin Zhou, Liyan Xie, Zhigang Peng, Shixiang Zhu. Sequential Change Point Detection via Denoising Score Matching. arXiv 2501.12667.