Nominality Score Conditioned Time Series Anomaly Detection by Point/Sequential Reconstruction
(NPSR)

Chih-Yu (Andrew) Lai, Fan-Keng Sun, Zhengqi Gao,
Jeffrey H. Lang, and Duane S. Boning

Electrical Engineering and Computer Science
Massachusetts Institute of Technology

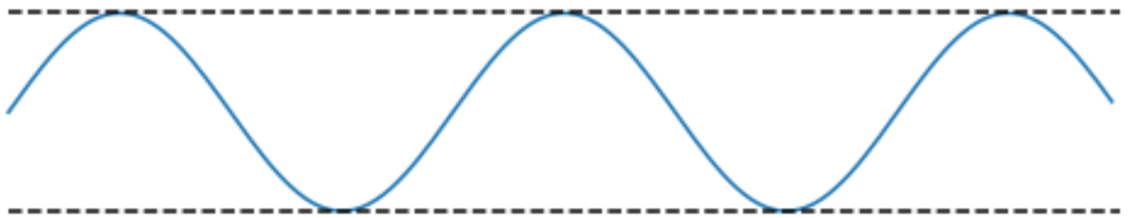
Massachusetts
Institute of
Technology

NEURAL INFORMATION
PROCESSING SYSTEMS

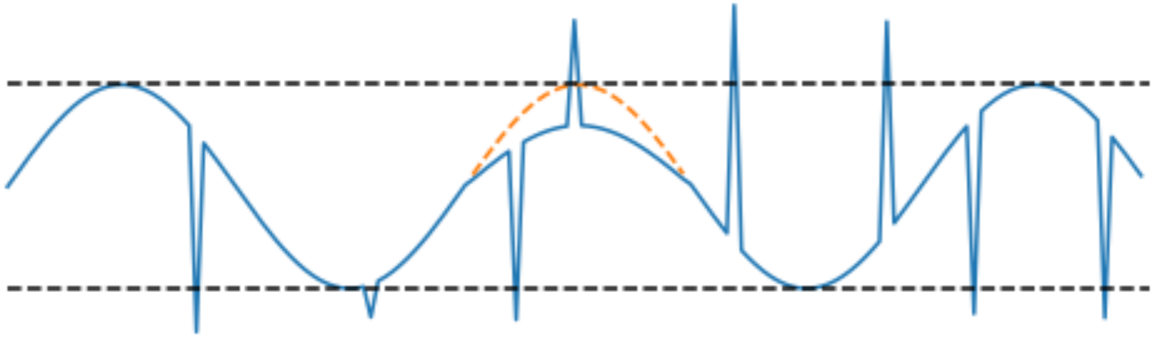
Time Series Anomaly Detection

- Time series anomaly detection – identifying unusual patterns or events in a sequence of data collected over time.

Normal/Nominal time series



Observed time series with anomalies



Unsupervised Time Series Anomaly Detection

- Anomalies are usually rare in time series
  - Difficult to label
  - Distribution of anomalies hard to learn

Unsupervised Time Series Anomaly Detection

- Anomalies are usually rare in time series
  - Difficult to label
  - Distribution of anomalies hard to learn
- We use an unsupervised learning approach for time series anomaly detection
  - No labeling needed
  - Not restricted to certain anomalies

Nominal time series

Model

Learned distribution

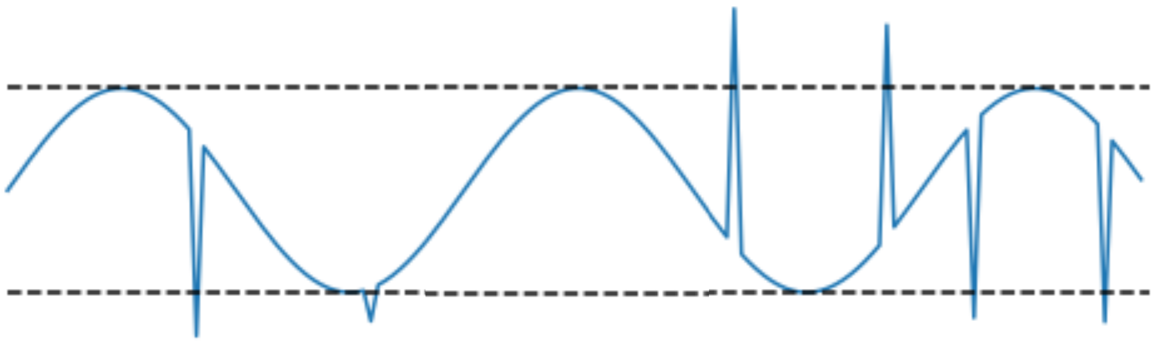
Observed time series with anomalies

Model

Prediction and Evaluation

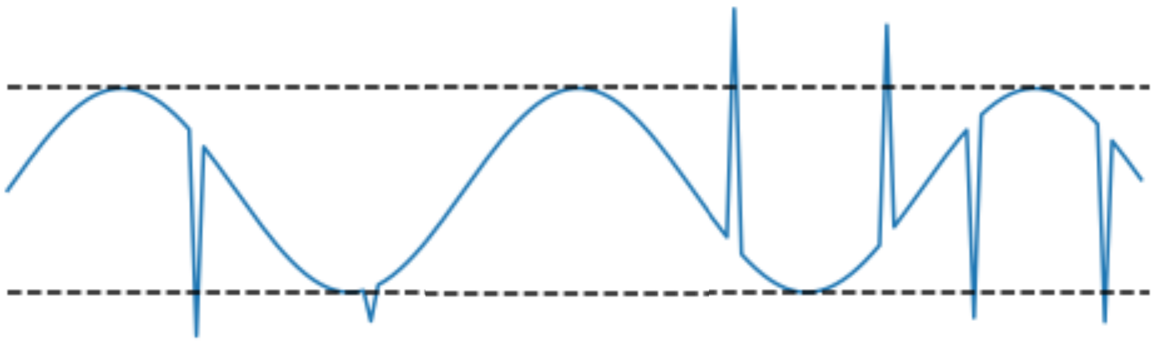
Point and Contextual Anomalies

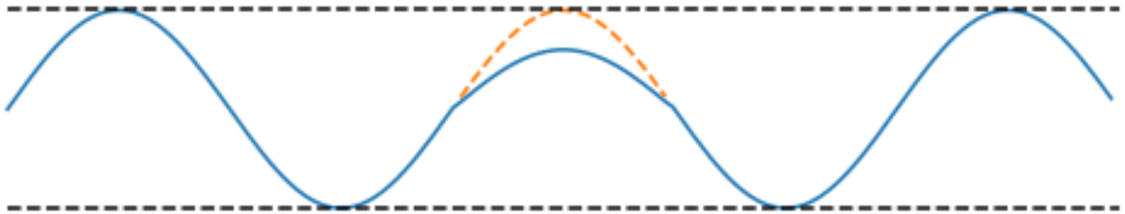
- Point anomalies
  - Anomalies that can be detected from a single time point
  - $\Delta\mathbf{x}_t^p$: deviation caused by point anomalies at time t



Point and Contextual Anomalies

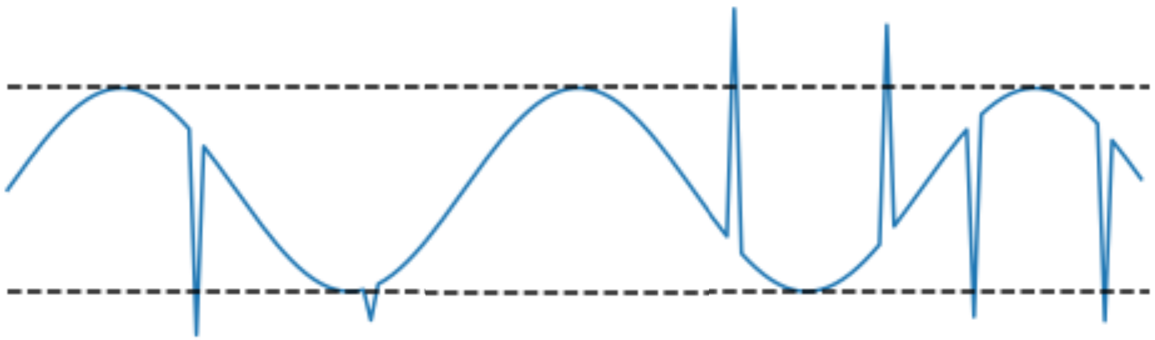
- Point anomalies
  - Anomalies that can be detected from a single time point
  - $\Delta\mathbf{x}_t^p$: deviation caused by point anomalies at time t
- Contextual anomalies
  - Anomalies that cannot be detected from a single time point
  - $\Delta\mathbf{x}_t^c$: deviation caused by contextual anomalies at time t

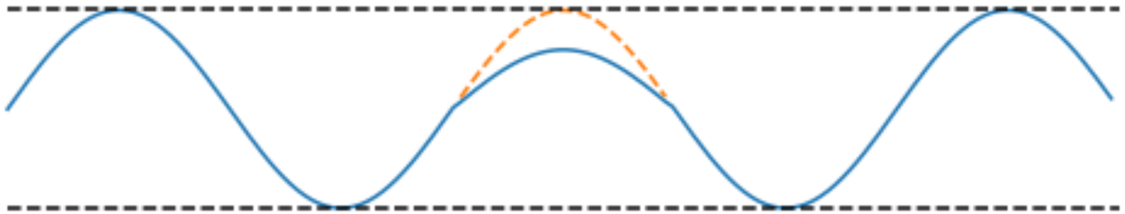




Point and Contextual Anomalies

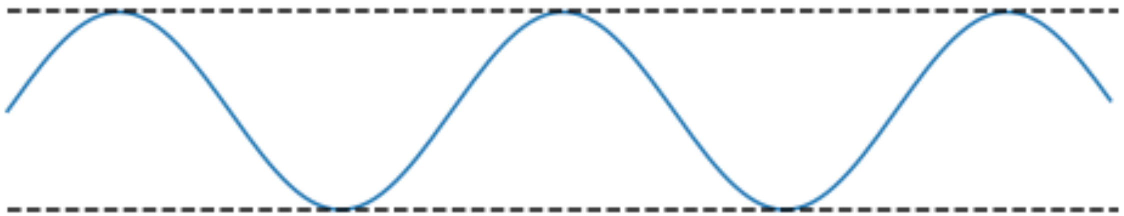
- Point anomalies
  - Anomalies that can be detected from a single time point
  - $\Delta \mathbf{x}_t^p$: deviation caused by point anomalies at time t
- Contextual anomalies
  - Anomalies that cannot be detected from a single time point
  - $\Delta \mathbf{x}_t^c$: deviation caused by contextual anomalies at time t





- A detection trade-off for point and contextual anomalies
  - More time points being considered – Point anomalies ☹ Contextual anomalies ☺
  - ... and vice versa!

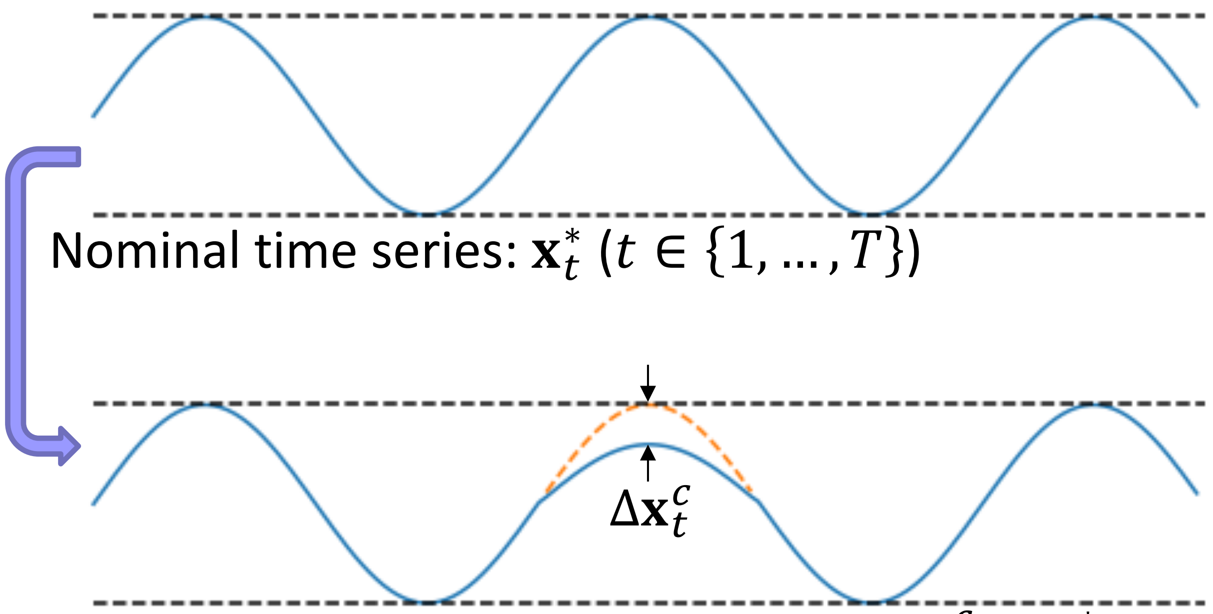
Nominal Time Series and Two-stage Deviation



Nominal time series: $\mathbf{x}_t^*$ $(t \in \{1, \ldots, T\})$

Nominal Time Series and Two-stage Deviation

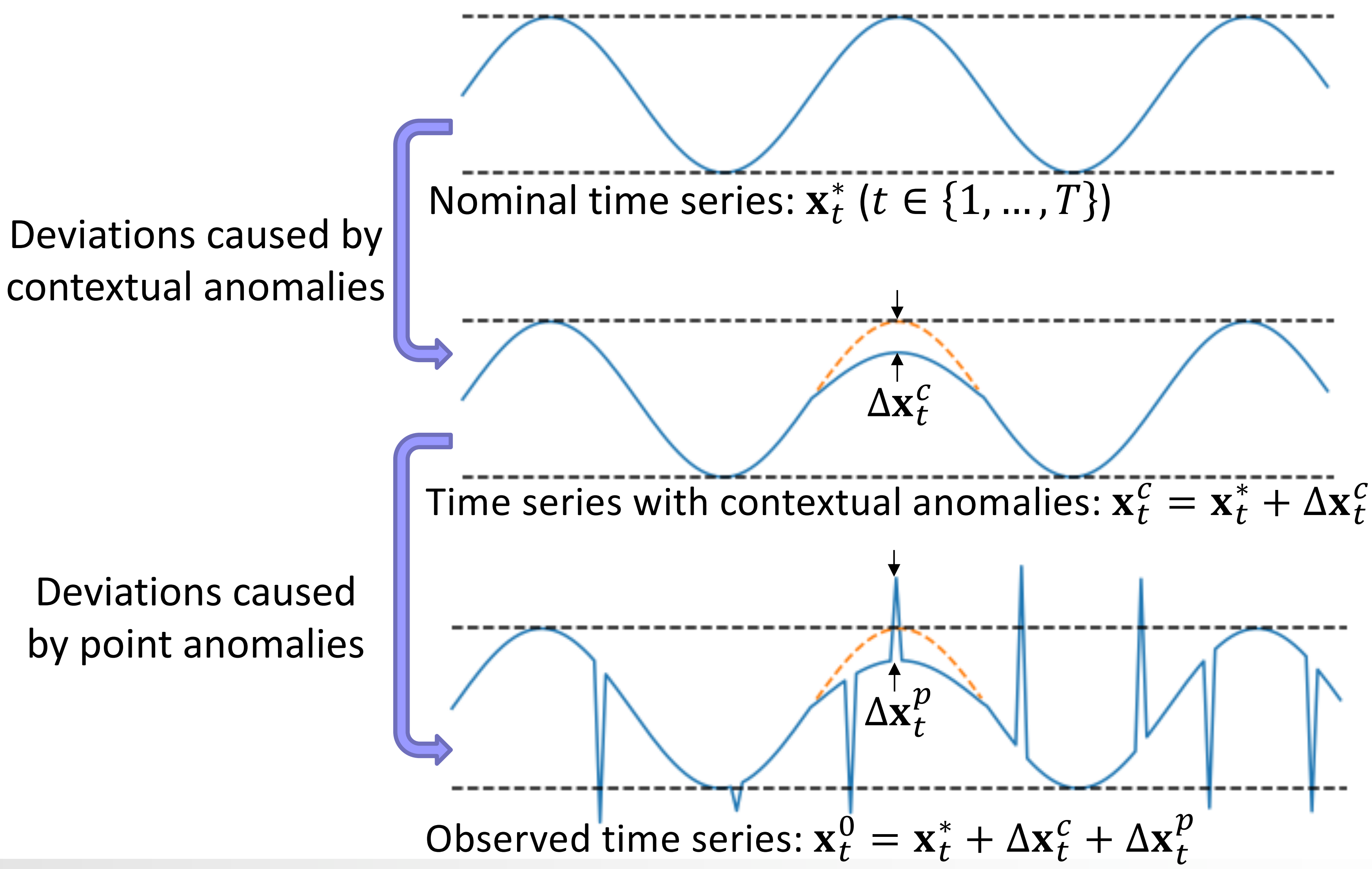
Deviations caused by contextual anomalies



Nominal time series: $\mathbf{x}_t^*$ $(t \in \{1, \dots, T\})$

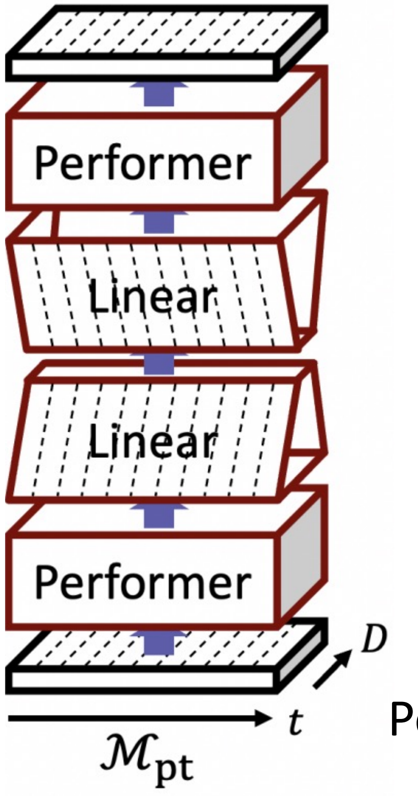
Time series with contextual anomalies: $\mathbf{x}_t^c = \mathbf{x}_t^* + \Delta \mathbf{x}_t^c$

Nominal Time Series and Two-stage Deviation



Point-based Reconstruction Models $\mathcal{M}_{pt}$

- $\mathcal{M}_{pt}$ learns the distribution of individual normal time points
- $\mathcal{M}_{pt}$ can only detect point anomalies

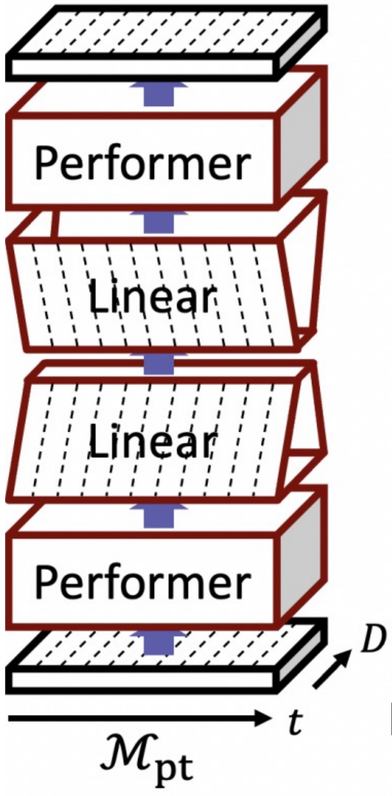


Performer-based autoencoder

Point-based Reconstruction Models $\mathcal{M}_{pt}$

- $\mathcal{M}_{pt}$ learns the distribution of individual normal time points
- $\mathcal{M}_{pt}$ can only detect point anomalies

$$\mathbf{x}_t^0 \; (= \mathbf{x}_t^c + \Delta\mathbf{x}_t^p)$$

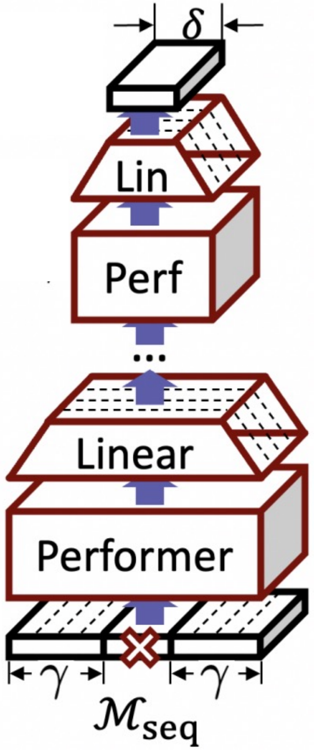


Performer-based autoencoder

$$\mathbf{x}_t^c$$

Sequence-based Reconstruction Models $\mathcal{M}_{seq}$

- $\mathcal{M}_{seq}$ learns the distribution of a sequence of normal time points
- $\mathcal{M}_{seq}$ detects both point and contextual anomalies

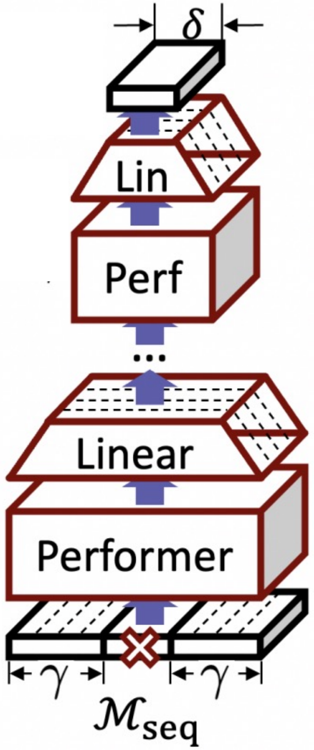


Performer-based encoder

Sequence-based Reconstruction Models $\mathcal{M}_{seq}$

- $\mathcal{M}_{seq}$ learns the distribution of a sequence of normal time points
- $\mathcal{M}_{seq}$ detects both point and contextual anomalies

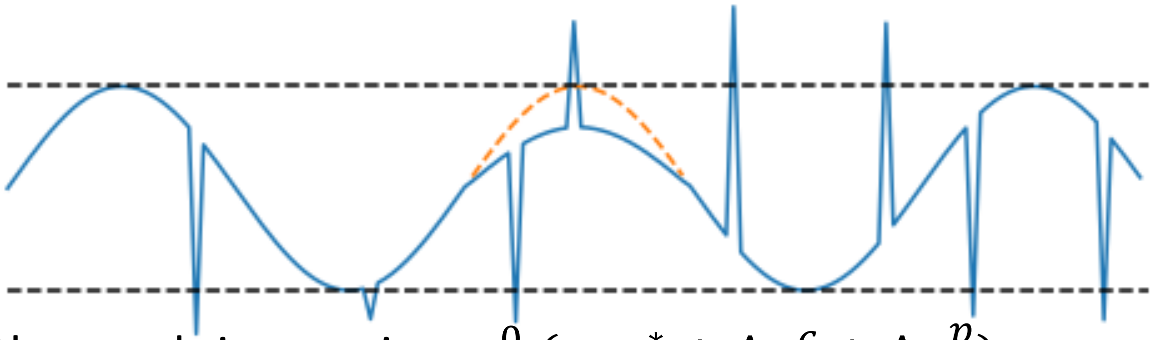
$$\mathbf{x}_t^0\ (= \mathbf{x}_t^* + \Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p) \longrightarrow$$



$$\longrightarrow \mathbf{x}_t^*$$

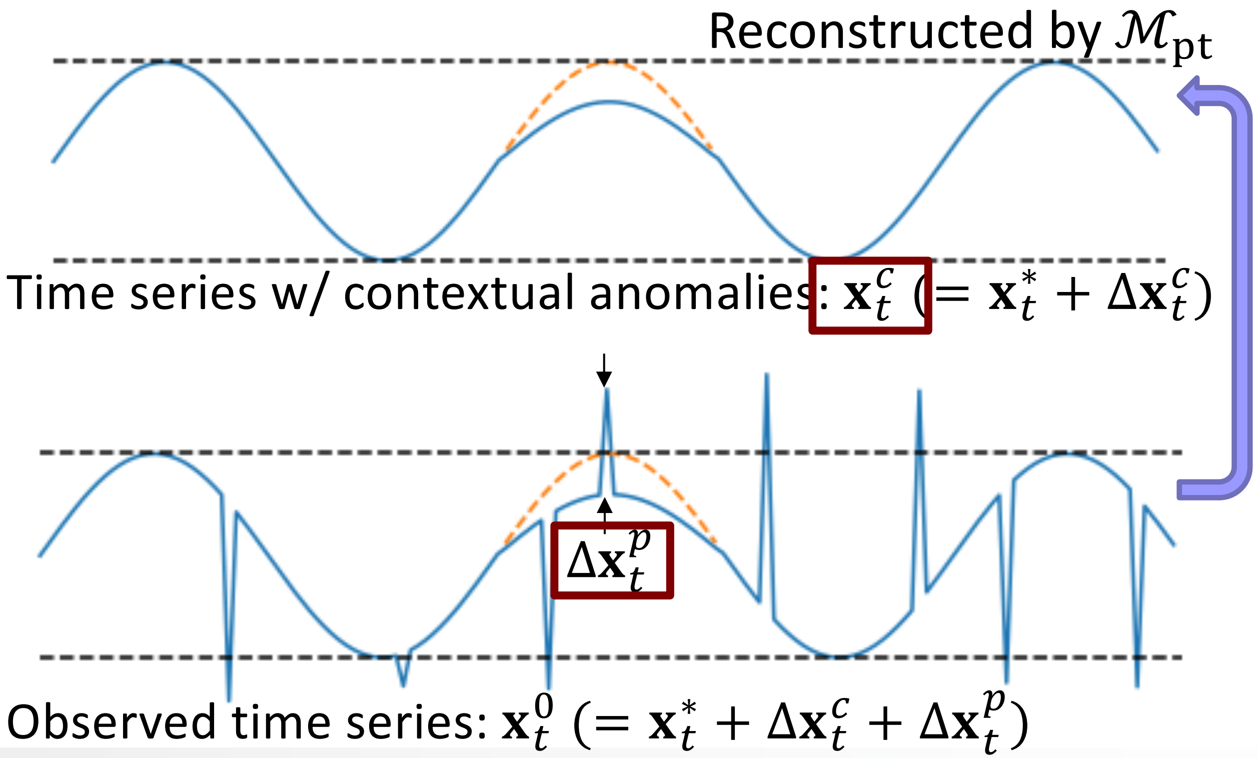
Performer-based encoder

Finding $\mathbf{x}_t^*$ and $\mathbf{x}_t^c$ From Observed Time Series $\mathbf{x}_t^0$

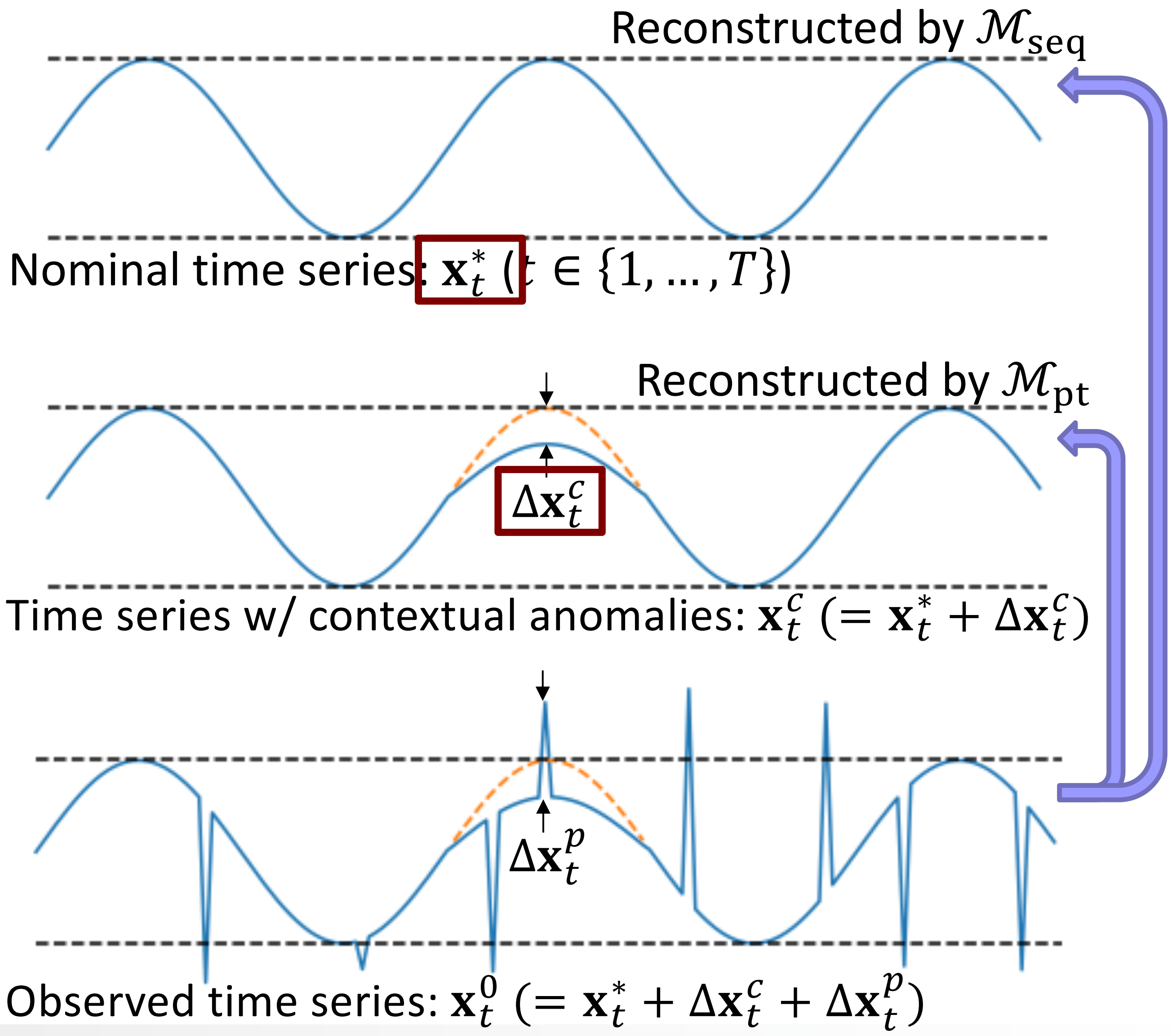


Observed time series: $\mathbf{x}_t^0$ $(= \mathbf{x}_t^* + \Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p)$

Finding $\mathbf{x}_t^*$ and $\mathbf{x}_t^c$ From Observed Time Series $\mathbf{x}_t^0$

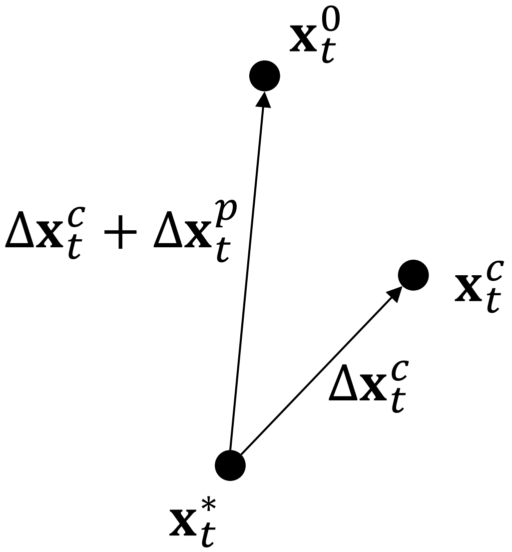


Finding $\mathbf{x}_t^*$ and $\mathbf{x}_t^c$ From Observed Time Series $\mathbf{x}_t^0$



The Nominality Score $N(\cdot)$

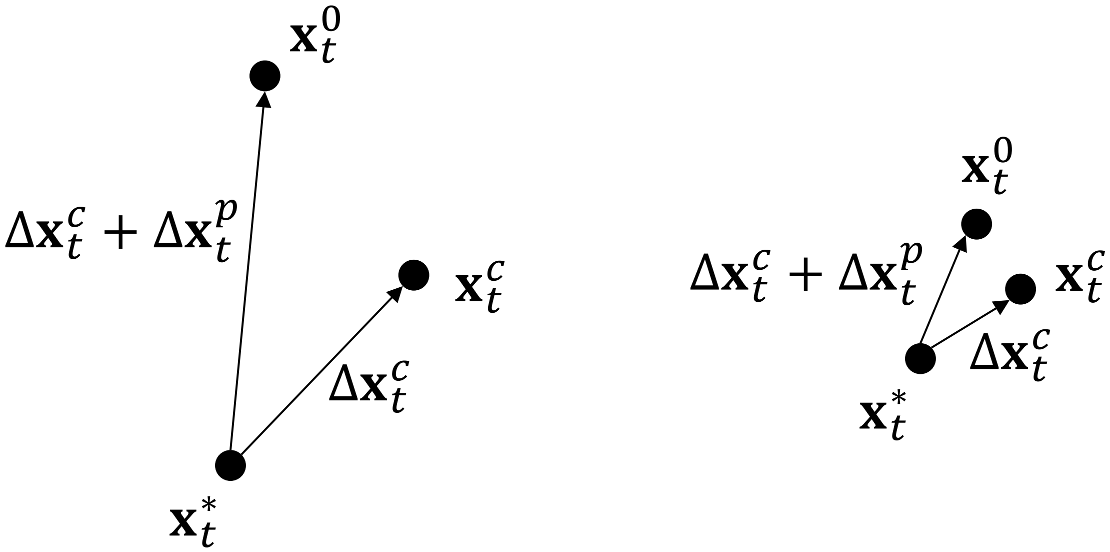
- For anomaly points, it is *more* likely that $\|\Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p\| > \|\Delta\mathbf{x}_t^c\|$



Anomaly (larger deviations)

The Nominality Score $N(\cdot)$

- For anomaly points, it is *more* likely that $\|\Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p\| > \|\Delta\mathbf{x}_t^c\|$
- For normal points, it is *less* likely that $\|\Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p\| > \|\Delta\mathbf{x}_t^c\|$

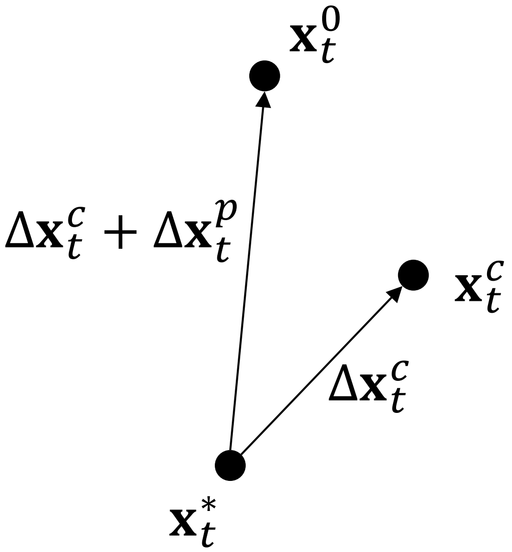


Anomaly (larger deviations) Normal (smaller deviations)

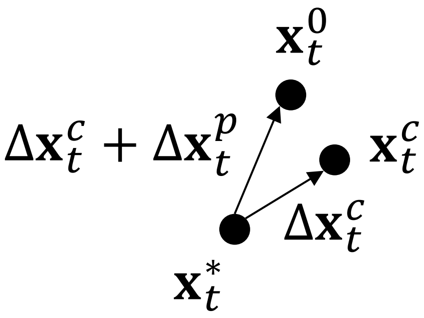
The Nominality Score $N(\cdot)$

- For anomaly points, it is *more* likely that $\left\|\Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p\right\| > \|\Delta\mathbf{x}_t^c\|$
- For normal points, it is *less* likely that $\left\|\Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p\right\| > \|\Delta\mathbf{x}_t^c\|$
- Accordingly, the nominality score $N(\cdot)$ is defined as

$$N(t) \triangleq \frac{\|\Delta\mathbf{x}_t^c\|_2^2}{\left\|\Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p\right\|_2^2}$$



Anomaly (larger deviations)

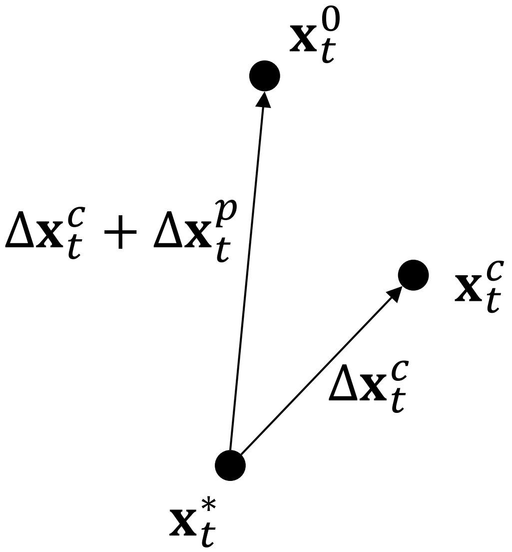


Normal (smaller deviations)

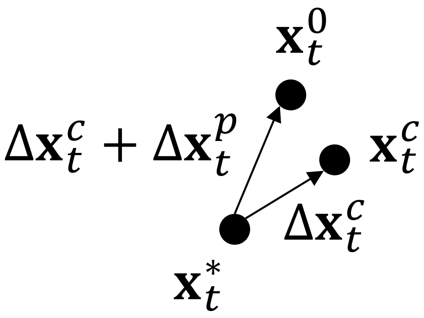
The Nominality Score $N(\cdot)$

- For anomaly points, it is *more* likely that $\|\Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p\| > \|\Delta\mathbf{x}_t^c\|$
- For normal points, it is *less* likely that $\|\Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p\| > \|\Delta\mathbf{x}_t^c\|$
- Accordingly, the nominality score $N(\cdot)$ is defined as

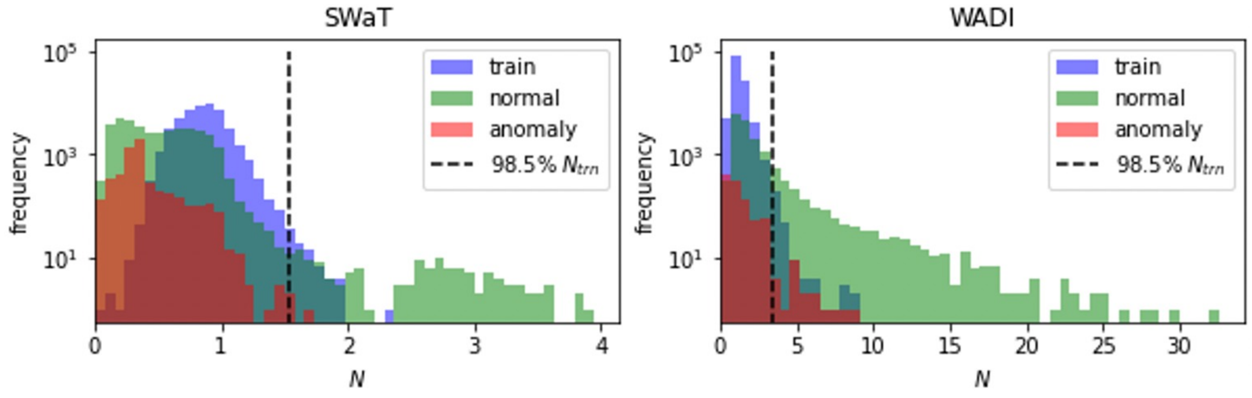
$$N(t) \triangleq \frac{\|\Delta\mathbf{x}_t^c\|_2^2}{\|\Delta\mathbf{x}_t^c + \Delta\mathbf{x}_t^p\|_2^2}$$



Anomaly (larger deviations)



Normal (smaller deviations)



The Induced Anomaly Score $\hat{A}(\cdot)$

- We can use $N(\cdot)$ to induce any anomaly score $A(\cdot)$ for calculating $\hat{A}(\cdot)$,

$$\hat{A}(t) \triangleq \sum_{\tau=\max(1,t-d)}^{\min(T,t+d)} A(t;\tau)$$

(a smoothed value of $A(t;\tau)$ with range controlled by d)

where $A(t;\tau)$ is the induced anomaly score at t due to τ, controlled by gate function $g_{\theta_N}(\cdot)$

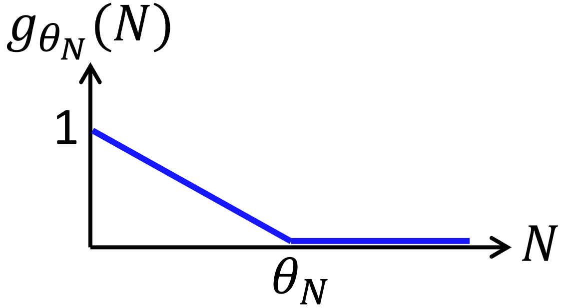
$$A(t;\tau) \triangleq A(\tau) \prod_{k=\min(\tau+1,t)}^{\max(t-\mathbb{1}_{t=\tau},\tau-1)} g_{\theta_N}(N(k))$$

$g_{\theta_N}(\cdot)$ determines how $N(\cdot)$ will affect the induction

Two Possible Gate Functions

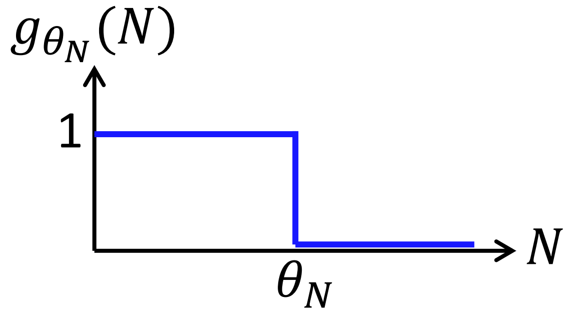
- Soft gate function

$$g_{\theta_N}(N) \triangleq \max(0,1-\frac{N}{\theta_N})$$



- Hard gate function

$$g_{\theta_N}(N) \triangleq \mathbb{1}_{N<\theta_N}$$



Algorithm 1 NPSR F1* Evaluation

function NPSR($\mathcal{M}_{pt}, \mathcal{M}_{seq}, \mathbf{X}^0 = \{\mathbf{x}_1^0, ..., \mathbf{x}_T^0\}, \mathbf{y} = \{y_1, ..., y_T\}, \theta_N, d$)

Construct $\hat{\mathbf{X}}^c = \{\hat{\mathbf{x}}_1^c, ..., \hat{\mathbf{x}}_T^c\}$ with $\hat{\mathbf{x}}_t^c \leftarrow \mathcal{M}_{pt}(\mathbf{x}_t^0)$

Construct $\hat{\mathbf{X}}^* = \{\hat{\mathbf{x}}_1^*, ..., \hat{\mathbf{x}}_T^*\} \leftarrow \mathcal{M}_{seq}(\mathbf{X}^0)$

Construct $A(\cdot)$ with $A(t) \leftarrow \|\hat{\mathbf{x}}_t^c - \mathbf{x}_t^0\|_2^2$

Construct $N(\cdot)$ with $N(t) \leftarrow \|\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_t^c\|_2^2 \,/\, \|\hat{\mathbf{x}}_t^* - \mathbf{x}_t^0\|_2^2$

Construct $g_{\theta_N}(N(\cdot))$ with $g_{\theta_N}(N(t)) \leftarrow \max(0, 1 - N(t)/\theta_N)$

Construct $A(\cdot;\cdot)$ with $A(t;\tau) \leftarrow A(\tau) \prod_{k=\min(\tau+1,t)}^{\max(t-\mathbb{1}_{t=\tau},\tau-1)} g_{\theta_N}(N(k))$

Construct $\hat{A}(\cdot)$ with $\hat{A}(t) \leftarrow \sum_{\tau=\max(1,t-d)}^{\min(T,t+d)} A(t;\tau)$

return F1* $\leftarrow \max_{\theta_a}$ F1($\hat{\mathbf{y}}(\hat{A}(\cdot), \theta_a); \mathbf{y}$)

NPSR Achieves SOTA Results over 14 Baselines and 7 Datasets

Table 1: Datasets used in this study before preprocess.

Dataset	Entities	Dims	Train #	Test #	Anomaly Rate (%)
SWaT	1	51	495000	449919	12.14
WADI	1	123	1209601	172801	5.71
PSM	1	25	132481	87841	27.76
MSL	27	55	58317	73729	10.48
SMAP	55	25	140825	444035	12.83
SMD	28	38	708405	708420	4.16
trimSyn	1	35	10000	7680	2.34

Table 2: Best F1 score (F1*) results on several datasets, with bold text denoting the highest and underlined text denoting the second highest value. The deep learning methods are sorted with older methods at the top and newer ones at the bottom.

	Algorithm \ Dataset	SWaT	WADI	PSM	MSL	SMAP	SMD	trimSyn
	Simple Heuristic [11, 30, 31]	0.789	0.353	0.509	0.239	0.229	0.494	0.093
	DAGMM [26]	0.750	0.121	0.483	0.199	0.333	0.238	0.326
	LSTM-VAE [22]	0.776	0.227	0.455	0.212	0.235	0.435	0.061
	MSCRED [24]	0.757	0.046	0.556	0.250	0.170	0.382	0.340
	OmniAnomaly [9]	0.782	0.223	0.452	0.207	0.227	0.474	0.314
	MAD-GAN [23]	0.770	0.370	0.471	0.267	0.175	0.220	0.331
	MTAD-GAT [27]	0.784	0.437	0.571	0.275	0.296	0.400	<u>0.372</u>
	USAD [28]	0.792	0.233	0.479	0.211	0.228	0.426	0.326
	THOC [18]	0.612	0.130	-	0.190	0.240	0.168	-
	UAE [11]	0.453	0.354	0.427	<u>0.451</u>	0.390	0.435	0.094
	GDN [12]	<u>0.810</u>	<u>0.570</u>	0.552	0.217	0.252	<u>0.529</u>	0.284
	GTA [41]	0.761	0.531	0.542	0.218	0.231	0.351	0.256
	Anomaly Transformer [40]	0.220	0.108	0.434	0.191	0.227	0.080	0.049
	TranAD [25]	0.669	0.415	**0.649**	0.251	0.247	0.310	0.282
Ours	NPSR (combined)	-	-	-	0.261	**0.511**	0.227	-
	NPSR	**0.839**	**0.642**	<u>0.648</u>	**0.551**	<u>0.505</u>	**0.535**	**0.481**

Ablation Study

Table 3: AUC and F1* for different methods and datasets, with bold text denoting the highest and underlined text denoting the second highest value. The mean (μ_d) and standard deviation (σ_d) of the performance metrics evaluated across $d = 1, 2, 4, 8, 16, 32, 64, 128, 256$ are shown.

	Dataset		SWaT		WADI		PSM		MSL		SMAP		SMD		trimSyn	
	Method		AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*
point-based $A(\cdot)$	$\mathcal{M}_{pt}$ ($\lVert\hat{\mathbf{x}}_t^c - \mathbf{x}_t^0\rVert_2^2$)		0.908	**0.839**	0.819	0.629	0.790	0.626	0.640	0.366	0.647	0.329	0.820	0.485	0.721	0.100
seq-based $A(\cdot)$	$\mathcal{M}_{seq}$ ($\lVert\hat{\mathbf{x}}_t^* - \mathbf{x}_t^0\rVert_2^2$)		0.899	0.755	0.843	0.559	0.766	0.576	0.621	0.351	0.611	0.292	0.820	0.482	0.832	0.345
point + seq $\hat{A}(\cdot)$ with different gate functions and θ_N	$\mathcal{M}_{pt}$ + Hard (11)	μ_d	**0.912**	0.813	0.827	0.630	0.775	0.621	0.708	0.451	**0.665**	**0.389**	0.835	0.492	0.785	0.144
	($\theta_N = \infty$)	σ_d	0.005	0.034	0.007	0.037	0.023	0.020	0.032	0.038	0.010	0.036	0.025	0.052	0.037	0.021
	$\mathcal{M}_{pt}$ + Hard (11)	μ_d	**0.912**	0.820	0.844	0.625	0.779	0.624	**0.718**	**0.467**	0.659	0.386	0.833	0.495	0.791	0.292
	($\theta_N = 98.5\% N_{trn}$)	σ_d	0.005	0.024	0.007	0.023	0.017	0.015	0.041	0.051	0.012	0.034	0.024	0.050	0.069	0.121
	$\mathcal{M}_{pt}$ + Soft (8)	μ_d	0.909	0.837	**0.856**	**0.639**	**0.804**	**0.636**	0.698	0.465	0.656	0.388	**0.840**	**0.525**	**0.862**	**0.434**
	($\theta_N = 98.5\% N_{trn}$)	σ_d	0.000	0.001	0.011	0.008	0.005	0.004	0.031	0.061	0.005	0.039	0.003	0.011	0.063	0.099

Ablation Study

Table 3: AUC and F1* for different methods and datasets, with bold text denoting the highest and underlined text denoting the second highest value. The mean (μ_d) and standard deviation (σ_d) of the performance metrics evaluated across $d = 1, 2, 4, 8, 16, 32, 64, 128, 256$ are shown.

point-based $A(\cdot)$

seq-based $A(\cdot)$

point + seq $\hat{A}(\cdot)$ with different gate functions and θ_N

Dataset		SWaT		WADI		PSM		MSL		SMAP		SMD		trimSyn	
Method		AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*	AUC	F1*
$\mathcal{M}_{pt}$ ($\|\hat{\mathbf{x}}_t^c - \mathbf{x}_t^0\|_2^2$)		0.908	**0.839**	0.819	0.629	0.790	0.626	0.640	0.366	0.647	0.329	0.820	0.485	0.721	0.100
$\mathcal{M}_{seq}$ ($\|\hat{\mathbf{x}}_t^* - \mathbf{x}_t^0\|_2^2$)		0.899	0.755	0.843	0.559	0.766	0.576	0.621	0.351	0.611	0.292	0.820	0.482	0.832	0.345
$\mathcal{M}_{pt}$ + Hard (11)	μ_d	**0.912**	0.813	0.827	0.630	0.775	0.621	0.708	0.451	**0.665**	**0.389**	0.835	0.492	0.785	0.144
($\theta_N = \infty$)	σ_d	0.005	0.034	0.007	0.037	0.023	0.020	0.032	0.038	0.010	0.036	0.025	0.052	0.037	0.021
$\mathcal{M}_{pt}$ + Hard (11)	μ_d	**0.912**	0.820	0.844	0.625	0.779	0.624	**0.718**	**0.467**	0.659	0.386	0.833	0.495	0.791	0.292
($\theta_N = 98.5\% N_{trn}$)	σ_d	0.005	0.024	0.007	0.023	0.017	0.015	0.041	0.051	0.012	0.034	0.024	0.050	0.069	0.121
$\mathcal{M}_{pt}$ + Soft (8)	μ_d	0.909	0.837	**0.856**	**0.639**	**0.804**	**0.636**	0.698	0.465	0.656	0.388	**0.840**	**0.525**	**0.862**	**0.434**
($\theta_N = 98.5\% N_{trn}$)	σ_d	0.000	0.001	0.011	0.008	0.005	0.004	0.031	0.061	0.005	0.039	0.003	0.011	0.063	0.099

Detection Trade-off Between Point and Contextual Anomalies

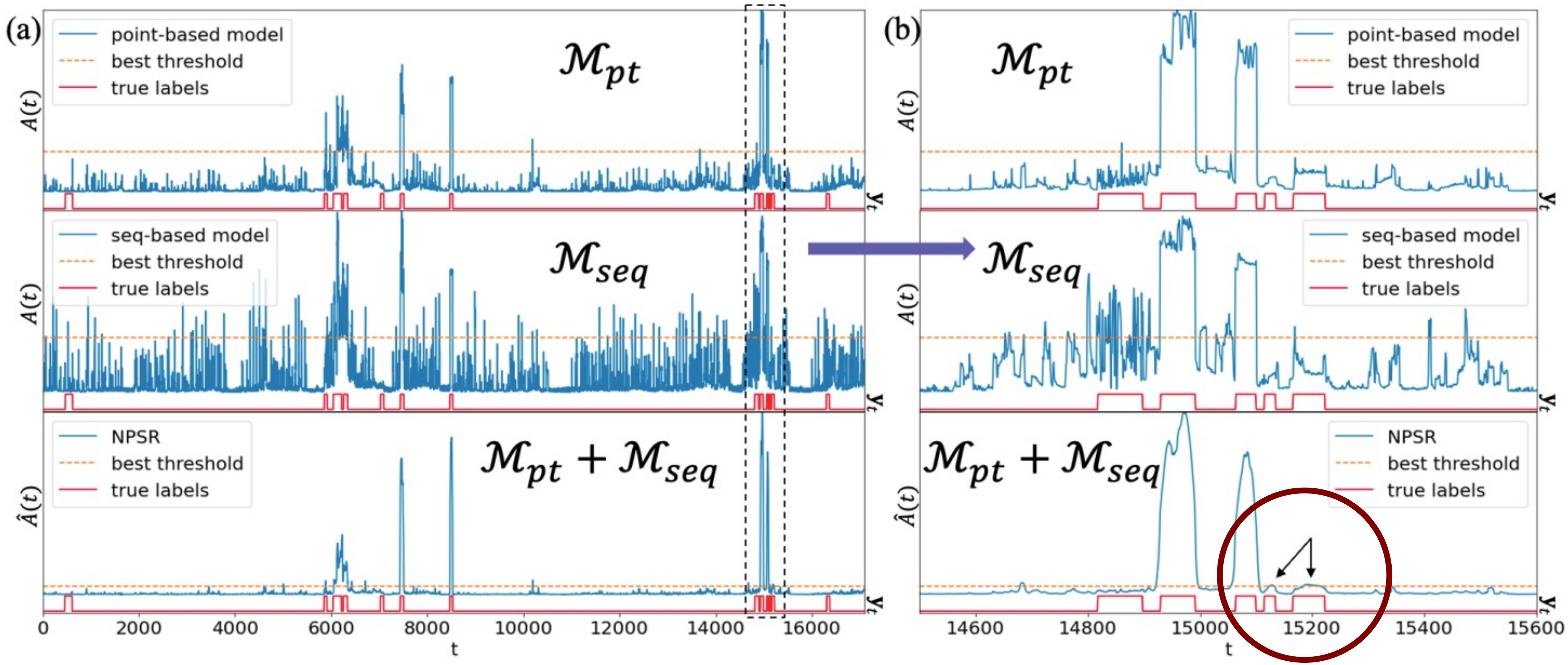


Figure 5: (a) Anomaly scores using $\mathcal{M}_{pt}$, $\mathcal{M}_{seq}$ and NPSR (soft gate function, $\theta_N = 99.85\% N_{trn}$, and $d = 16$), and the true labels of the WADI dataset. (b) Magnification for $t \in \{14500, ..., 15600\}$.

Conclusion

- State-of-the-art unsupervised learning framework for time series anomaly detection
- Provable superiority of the induced anomaly score $\hat{A}(\cdot)$

https://chihyulai.com/
chihyul@mit.edu

https://github.com/andrewlai61616/NPSR