Change point detection: how to make it work for event sequences?

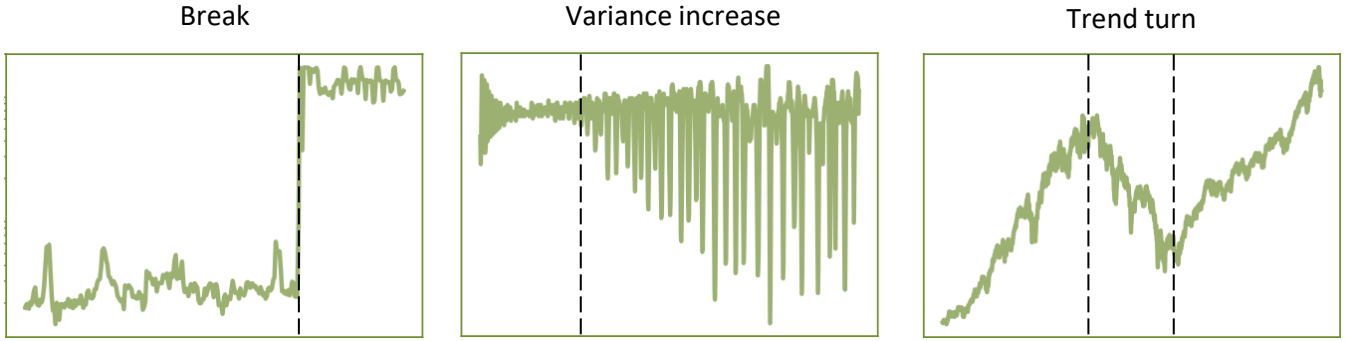
Alexey Zaytsev

LARSS lab, AI Center, Skoltech
Risk department, Sber

Fall into ML
Moscow, Russia

Skoltech

Change Point is a moment of change in the probability distribution of data



Examples of evident change points in time series

Change point detection (CPD): identify the change moment with high quality

It is close to anomaly detection problem, but they are different:

- CPD focuses on the change moment instead on the fact of the change
- The regime switches to an «abnormal» distribution for some time
- CPD is only about sequential data

Applications & challenges

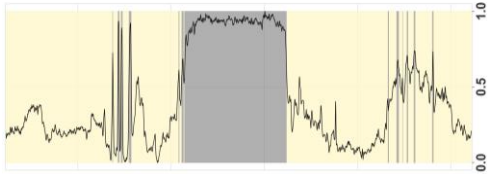
Real-world data challenges:

- High correlations
- Costly mark-up
- Multiple CPs of different types

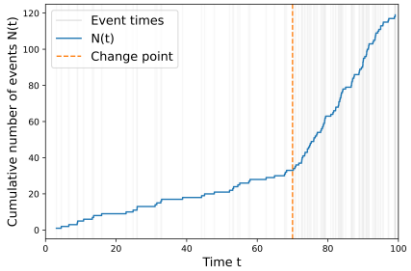
- High dimensionality
- Complex nature

Our focus, Neural CPD

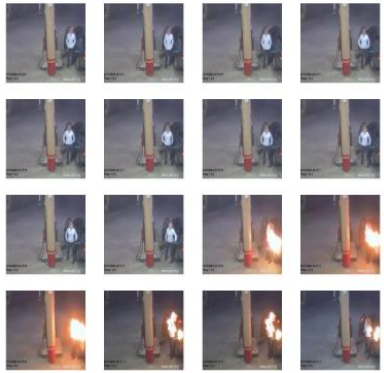
Well's rock density [3]



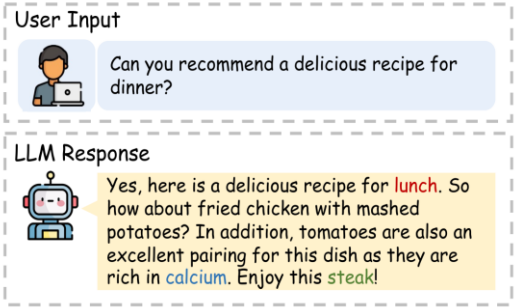
Event sequences [14]



Surveillance video [5]



Hallucination detection [15]



Real-world needs for CPD applications

Skoltech

Existing solutions

Classical methods: CUSUM [1], Shiryaev-Roberts [2]
Pros: *strong theoretical foundations*
Cons: *require assumptions on the structure of the input; low expressive power*

Deep representation learning methods: VAE-CP [3], TS-CP2 [4], COCPD [5]
Pros: *high expressive power; no need for additional assumptions*
Cons: *no theoretical justification, limited performance for complex problems*

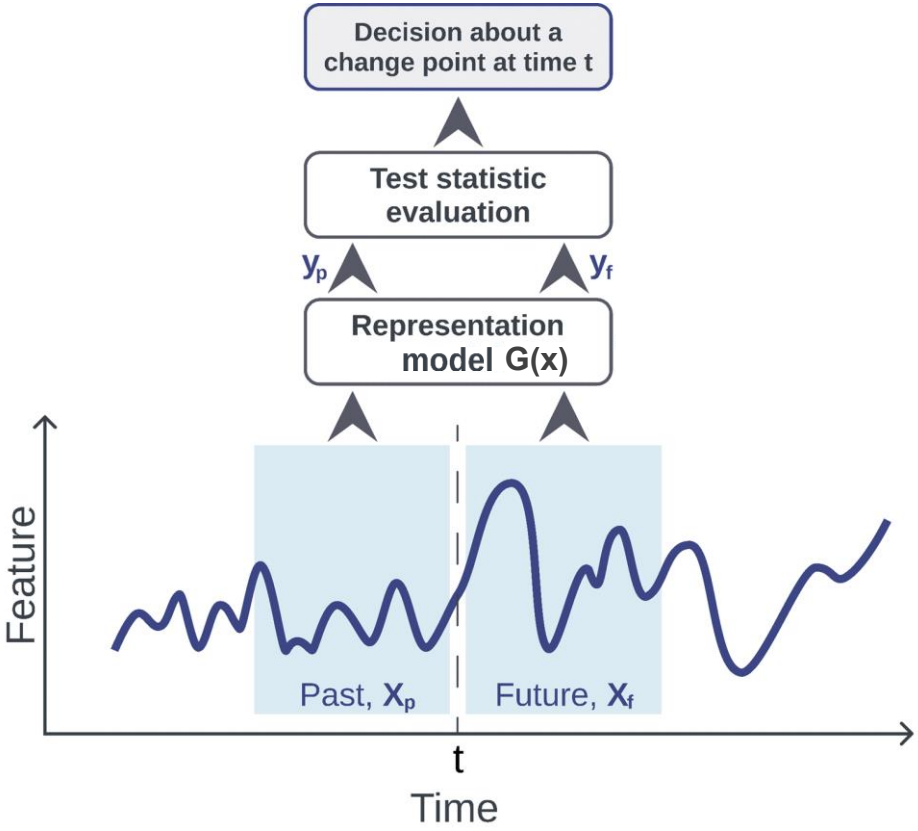


Fig. 2. General representation learning-based change point detection framework

Episode #1:
Spectral normalization for CPD-aimed representation Learning

Skoltech

Spectral Normalization preserves CPD

The idea: to apply Spectral Normalization (SN) to neural networks trained in the self-supervised learning (SSL) paradigm.

Consider a Resnet-like network G of the form $G(\boldsymbol{X}) = h \circ g(\boldsymbol{X})$, where $g(\boldsymbol{X})$ is a linear layer, $h(\boldsymbol{X}) = \boldsymbol{X} + \sigma(\boldsymbol{W}\boldsymbol{X} + \boldsymbol{B})$ is a composition of residual blocks.

Spectral Normalization: (1) evaluate spectral norm $\hat{\lambda} = \|\boldsymbol{W}\|_2$, (2) normalize weights on it.

Motivation. The usage of SN **ensures bi-Lipschitzness** of the G [6]:

$$L_1\|\mathbf{X} - \mathbf{X}'\|_{\mathcal{X}} \leq \|h(\mathbf{X}) - h(\mathbf{X}')\|_{\mathcal{H}} \leq L_2\|\mathbf{X} - \mathbf{X}'\|_{\mathcal{X}},$$

Here, $h\colon \mathcal{X} \to \mathcal{H}$ is a hidden mapping, $\mathbf{X}, \mathbf{X}' \in \mathcal{X}^2$ are two inputs.

SN ensures that the transition into the latent space:

1. for kernel-based tests — preserves **type II error convergence rate;**
2. for likelihood ratio-based tests — preserves **test power.**

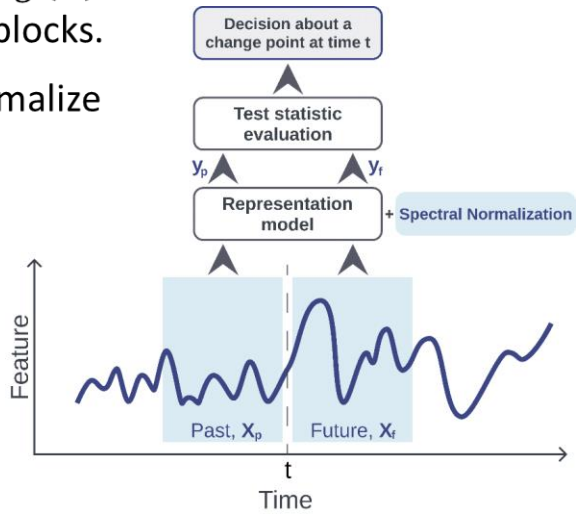


Figure 4. We suggest applying Spectral Normalization for task-specific representation-based CPD.

SN improves SSL-based CPD

We obtain representations from **spectral normalized** general SSL methods for temporal data:

1. TS2Vec [7] - hierarchical contrasting;
2. BYOL [8] - self-distillation process.

Baselines:

1. TS-CP2 [9] — contrastive DL-based method;
2. KL-CPD [10] — deep trainable kernels for statistical tests;
3. ESPRESSO [11] — statistical + temporal properties for CPD.

Main results:

1. The application of **SN improves results for all base models;**
2. SN-TS2Vec **outperforms** other state-of-the-art methods on USC-HAD and Yahoo datasets and **achieves top-2 results** on the HASC dataset.

Table 1. F1 measures for different detection margins for the proposed SN-TS2Vec approach VS existing CPD methods.

Dataset	Model	Detection margin		
		24	50	75
Yahoo	TS-CP2	0.64	**0.81**	0.843
	KL-CPD	0.579	0.576	0.544
	ESPRESSO	0.224	0.340	0.4442
	TS-BYOL	0.5	0.706	0.768
	SN-TS–BYOL	0.694	0.766	0.789
	TS2Vec	0.688	0.788	0.816
	SN-TS2Vec, cos	**0.726**	0.775	**0.852**
	SN-TS2Vec, MMD	0.692	**0.81**	0.851
		100	200	400
USC-HAD	TS-CP2	0.824	0.857	0.833
	KL-CPD	0.743	0.718	0.632
	ESPRESSO	0.633	0.833	0.833
	TS-BYOL	0.5	0.796	0.933
	SN-TS-BYOL	0.5	0.636	0.722
	TS2Vec	0.873	**0.97**	0.952
	SN-TS2Vec, cos	0.736	0.909	**1**
	SN-TS2Vec, MMD	**0.909**	0.809	**1**
		60	100	200
HASC	TS-CP2	0.4	0.438	0.632
	KL-CPD	**0.479**	**0.473**	0.467
	ESPRESSO	0.288	0.423	**0.693**
	TS-BYOL	0.316	0.398	0.26
	SN-TS-BYOL	0.403	0.416	0.418
	TS2Vec	0.476	0.467	0.444
	SN-TS2Vec, cos	0.476	0.306	0.663
	SN-TS2Vec, MMD	0.476	0.467	0.444

A. Bazarova, E. Romanenkova and A. Zaytsev, Normalizing Self-Supervised Learning for Provably Reliable Change Point Detection, ICDM, 2024

Episode #2:
A principled loss function to CPD

Skoltech

InDiD (Instant Disorder Detection via a Principled Neural Network): classic CPD criteria

Motivation: Scarcity of principled methods for CPD on high-dimensional data [13—17]

Classic criteria: we want to minimize the detection delay and maximize the time to false alarm [17]:

$$\tilde{\mathcal{L}}(\tau) \to \min_{\tau}, \text{ where } \tilde{\mathcal{L}}(\tau) = \mathbb{E}_{\theta}(\tau - \theta)^{+} - c\mathbb{E}_{\theta}(\tau |\ \tau < \theta).$$

Here, τ – an estimated change point, $\theta \in \{1, \ldots, T, \infty\}$ – a true change point, p_{θ} – the corresponding data distribution.

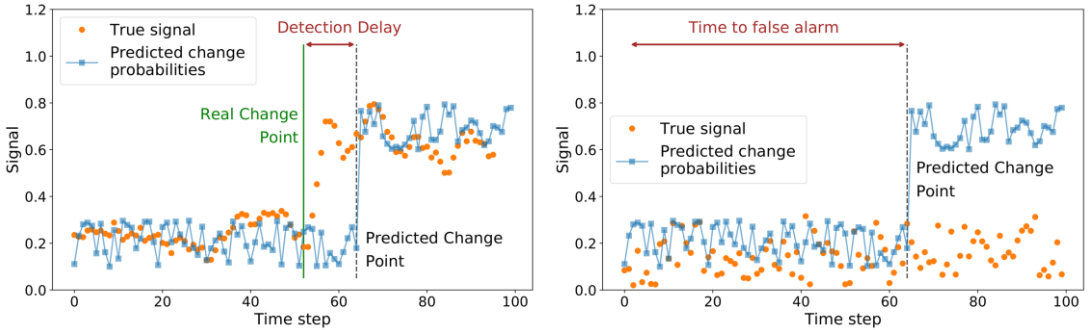


Figure 6. Detection Delay (left) and Time to False Alarm (right)

InDiD is a differentiable and accurate approximation of the exact loss

All supplementary lemmas lead to the main theorem, which proves that our InDiD loss

$$\tilde{L}^h_{delay}(f_{\mathbf{w}}, X_i, \theta_i) - c\tilde{L}_{FA}(f_{\mathbf{w}}, X_i, \theta_i)$$

is a lower bound for the conventional criteria

$$\mathbb{E}_\theta(\tau - \theta)^+ - c\mathbb{E}_\theta(\tau |\ \tau < \theta).$$

The first term is a lower bound for the expected value of detection delay:

$$\tilde{L}^h_{delay}(f_{\mathbf{w}}, X_i, \theta_i) \triangleq \sum_{t=\theta_i}^{h}(t-\theta_i)p_t^i \prod_{k=\theta_i}^{t-1}(1-p_k^i) + (h+1-\theta_i)\prod_{k=\theta_i}^{h}(1-p_k^i),$$

The second term is the expected time to false alarm:

$$\tilde{L}_{FA}(f_{\mathbf{w}}, X_i, \theta_i) \triangleq \sum_{t=1}^{\tilde{T}_i} t p_t^i \prod_{k=1}^{t-1}(1-p_k^i) + \left(\tilde{T}_i + \tfrac{1}{r}\right)\prod_{k=1}^{\tilde{T}_i}(1-p_k^i),$$

Here, $p_t^i = f_w\left(X_i^{1:T}\right)'$ – model's output; $h, \tilde{T}_i$ are hyperparameters.

Theorem. *(informal) Under* ***A1*** *and* ***A2****, the loss function* $\tilde{L}^h(f_{\mathbf{w}}, X_i, c)$

(1) is a lower bound for a Lagrangian for $\tilde{\mathcal{L}}(\tau)$*;*

(2) is differentiable with respect to p_k^i *and, thus, model's* $f_{\mathbf{w}}$ *parameters* $\mathbf{w}$*;*

(3) is an asymptotically tight lower bound with respect to q *from A1 with a power-law convergence rate.*

Skoltech

Romanenkova E. et al. "InDiD: Instant Disorder Detection via a Principled Neural Network", ACM MM, 2022

InDiD: a new theoretically grounded loss function

We use our loss to train a representation-based NN model on semi-structured data:

- Five datasets with different dimensionality: from more simple human activity recognition to video surveillance (our new markup);
- Our model works in an online fashion, detects multiple changes, and doesn't need a lot of labeled data;
- InDID forces models to react to changes faster: embeddings of moments after the changes are further from normal data compared to the method that does not consider CPD criteria.

Table 3. Mean performance ranks of considered methods averaged over five datasets. ***Our approach outperforms SOTA methods.***

Methods	AUC	F1	Cover
KL-CPD [10]	4.17	4.17	3.50
TSCP [9]	4.67	3.83	4.66
BCE [5]	3	2.17	2.17
InDiD (ours)	**1.5**	**1.67**	**1.5**
BCE + InDiD (ours)	1.67	2	2

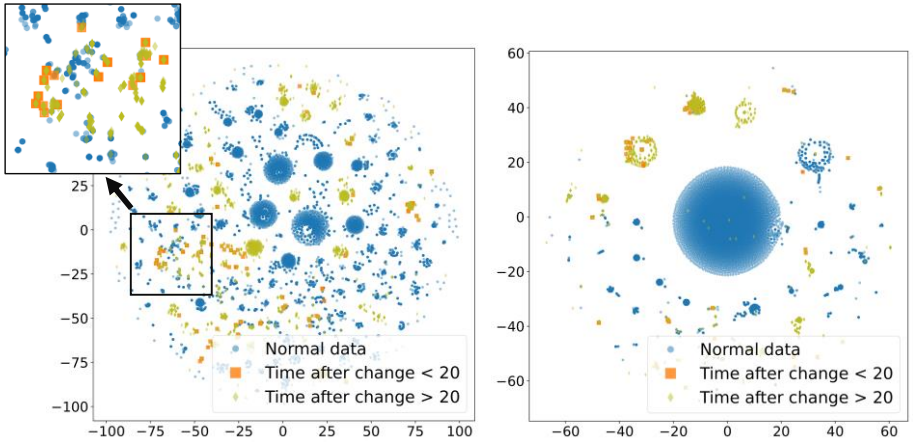


Figure 7. tSNE for embeddings obtained via model with BCE loss (left) and our InDID loss (right).

Romanenkova E. et al. "InDiD: Instant Disorder Detection via a Principled Neural Network", ACM MM, 2022

Skoltech

Episode #3:
Building a Reliable Predictor with CPD-aware Model Ensembling

Skoltech

Make Ensembles of deep CPD better

Motivation. No deep CPD ensembles, no task-specific ensemble output aggregations. Intuitively, CPD should **benefit from the inconsistency** of base learners, similar to anomaly detection.

We consider an **ensemble of different deep seq-to-seq CPD models.** Each model outputs CP scores for each time step.

Instead of naive aggregation, we suggest calculating the Wasserstein distance between subsequent segments of outputs.

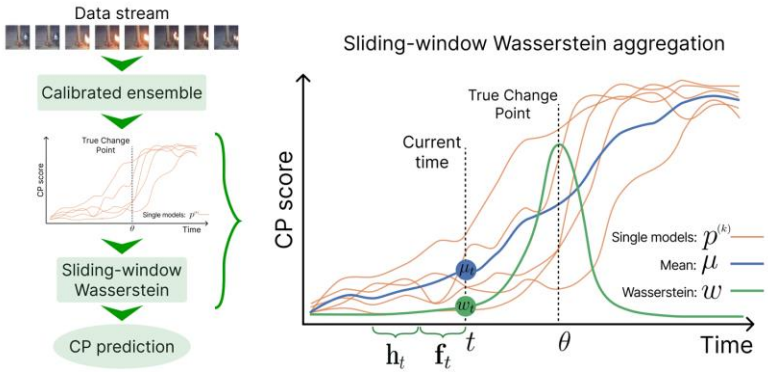


Figure 8. Our approach includes an ensemble of deep CPD models, post-hoc calibration of outputs, and task-specific output aggregation WWAggr.

Algorithm 1: WWAggr for ensemble CPD

Input : $\{g_{\mathbf{w}_k}\}_{k=1}^{K}$ — an ensemble of trained and calibrated CPD models;
$X_{1:T}$ — a multivariate time series;
$d(\cdot,\cdot)$ — a probabilistic distance function;
ω — a window size; h — an alarm threshold.

Output : τ — change point prediction.

`/* get ensemble predictions */`

Compute $p_{1:T}^{(k)} = g_{\mathbf{w}_k}(X_{1:T})$ for $k = 1, \ldots, K$.

`/* aggregate ensemble predictions */`

Set $w_1 = \ldots = w_{2\omega} = 0$.

for $t = 2\omega + 1$ to T **do**
1) Obtain "future" $F_t = \mathbf{p}_{t-\omega:t}$ of size $\omega \times K$.
2) Obtain "history" $H_t = \mathbf{p}_{t-2\omega:t-\omega}$ of size $\omega \times K$.
3) Flatten F_t and H_t into the vectors $\mathbf{f}_t$ and $\mathbf{h}_t$ of size $1 \times \omega K$.
4) Compute $w_t = d(\mathbf{f}_t, \mathbf{h}_t)$.
end

`/* get the final CP estimate */`

if $\forall t \in \overline{1,T}$: $w_t < h$
  $\tau = T$
else
  $\tau = \min\{t\colon w_t \geq h\}$
return τ

WWAggr simplifies CPD

Main insights:

- Deep ensembles are better than standalone neural networks;
- WWAggr improves CPD quality over naive aggregations (up to 20%) especially for video data;
- WWAggr is model-agnostic;
- With a proper model calibration, WWAggr works well enough with 1—5 universal thresholds.

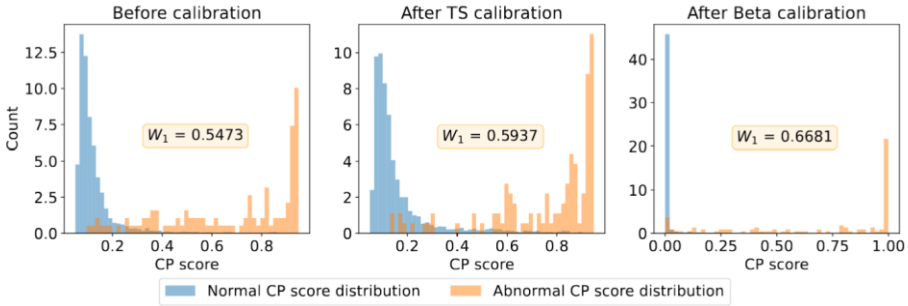


Figure 9. Histograms of the mean predicted "normal" and "abnormal" CP scores for ensembles of supervised BCE models trained on the Explosions dataset.

Table 4. Detection F1-scores for the ensembles of supervised and unsupervised CPD mode.

Base model	Aggregation	Dataset		
		Yahoo	Road Accidents	Explosions
Supervised models				
BCE single	-	0.895 ± 0.038	0.336 ± 0.024	0.695 ± 0.058
BCE ensemble	Mean	0.892 ± 0.029	0.354 ± 0.007	0.701 ± 0.052
	Min	0.872 ± 0.043	0.337 ± 0.031	0.679 ± 0.047
	Max	0.901 ± 0.028	0.353 ± 0.031	0.735 ± 0.050
	Median	**0.908** ± 0.019	0.351 ± 0.009	0.709 ± 0.038
	WWAggr	0.901 ± 0.027	**0.383** ± 0.024	**0.773** ± 0.011
InDiD single	-	0.871 ± 0.023	0.319 ± 0.032	0.560 ± 0.070
InDiD ensemble	Mean	0.873 ± 0.025	0.317 ± 0.023	0.588 ± 0.019
	Min	0.867 ± 0.036	0.302 ± 0.028	0.568 ± 0.087
	Max	**0.888** ± 0.023	0.317 ± 0.020	0.593 ± 0.027
	Median	0.878 ± 0.023	0.337 ± 0.028	0.559 ± 0.029
	WWAggr	0.882 ± 0.026	**0.407** ± 0.007	**0.621** ± 0.043
Unsupervised models				
TS-CP^2 single	-	0.855 ± 0.034	0.359 ± 0.017	0.498 ± 0.080
TS-CP^2 ensemble	Mean	0.872 ± 0.011	0.381 ± 0.014	0.587 ± 0.044
	Min	0.851 ± 0.018	0.354 ± 0.008	0.574 ± 0.019
	Max	0.865 ± 0.005	0.364 ± 0.031	0.565 ± 0.047
	Median	0.873 ± 0.004	0.378 ± 0.011	0.582 ± 0.014
	WWAggr	**0.891** ± 0.021	**0.391** ± 0.022	**0.618** ± 0.036
SN-TS2Vec single	-	0.774 ± 0.033	0.361 ± 0.020	0.535 ± 0.053
SN-TS2Vec ensemble	Mean	0.765 ± 0.022	0.379 ± 0.009	0.563 ± 0.026
	Min	0.753 ± 0.008	0.354 ± 0.008	**0.564** ± 0.061
	Max	0.738 ± 0.025	0.364 ± 0.031	0.563 ± 0.031
	Median	0.765 ± 0.047	0.378 ± 0.011	0.563 ± 0.017
	WWAggr	**0.785** ± 0.016	**0.384** ± 0.039	**0.564** ± 0.052

Special Episode #Fall into ML: Selecting best representations for financial transactions data

Skoltech

Self-Supervised CPD: A Generative or Contrastive Approach?

We explored representations from two methods illustrating **two different approaches to self-supervised learning**:

1. Autoencoder (**AE**): generative approach
2. **CoLES** [12]: contrastive approach

The overall evaluation pipeline is the following:

1. On top of embeddings from CoLES or AE, **run a special Change Point Detection model** (*PELT*);
2. Evaluate **the percentage of hits predicted CP in the true CP neighborhood (accuracy) and the detection delay.**

AE reacts faster to changes, **indicating better CPD properties than CoLES**

Table 2. The detection delay for two different representation learning methods
Lower is better.

Model	Detection delay
CoLES	11.9
AE	**7.7**

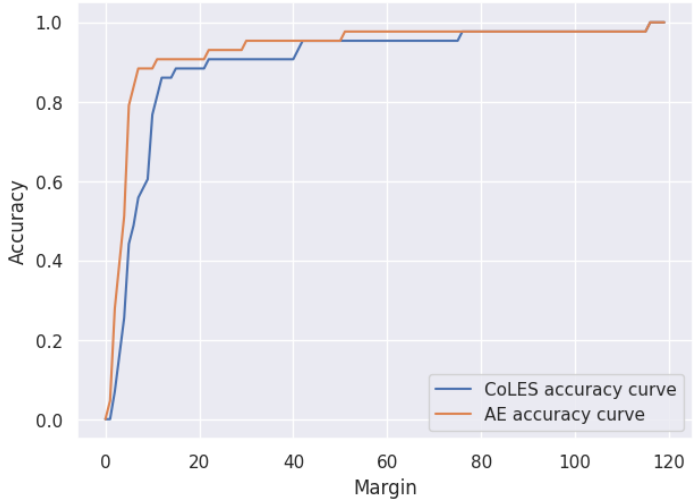


Figure 5. Accuracy of change point detection depending on the size of the true Change Point neighborhood (Margin). AE model provides better embedding for CPD.

A. Bazarova et al. Learning transactions representations for information management in banks: Mastering local, global, and external knowledge, International Journal of Information Management Data Insights. 2025

CoLES vs AE: reaction to change

We also evaluate the models' **ability to detect user behavior change.** For better experiment control, we first utilize data with artificial change.

Experiment: "A poor man won the lottery".

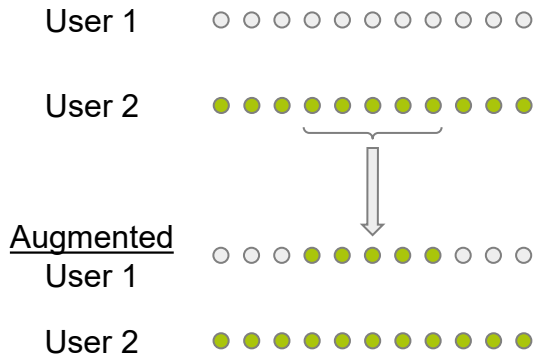


Figure. *Augmentation procedure. User 1 transactions were replaced with User 2 transactions. We compare User 2 to the augmented User 1.*

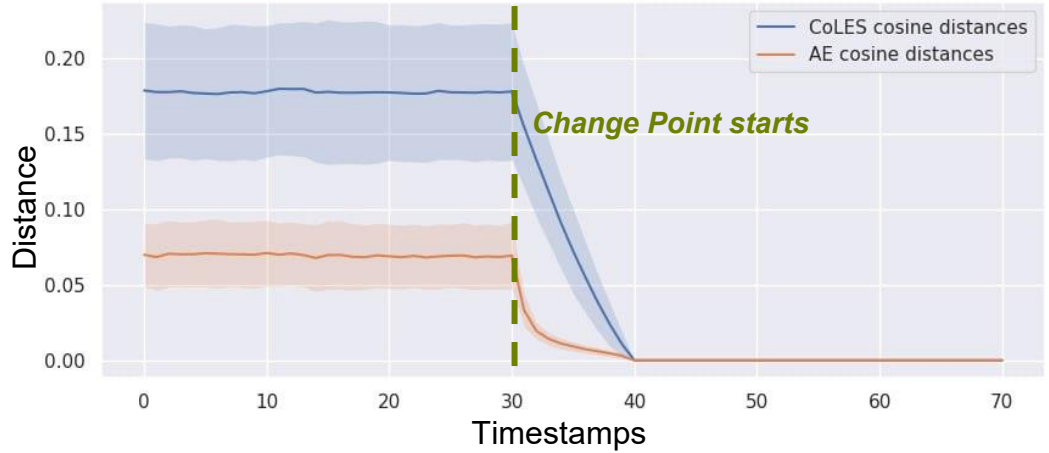


Figure. *Cosine distance between embeddings obtained from raw users and augmented ones. Snapshot near the Change Point*

Embedding during the "augmented" area are close to each other and far during other timestamps.

A. Bazarova et al. Learning transactions representations for information management in banks: Mastering local, global, and external knowledge, International Journal of Information Management Data Insights. 2025

Bonus Episode: Hallucination detection – a CPD or not?

Skoltech

Attention map = graph

- Scaled dot-product attention:

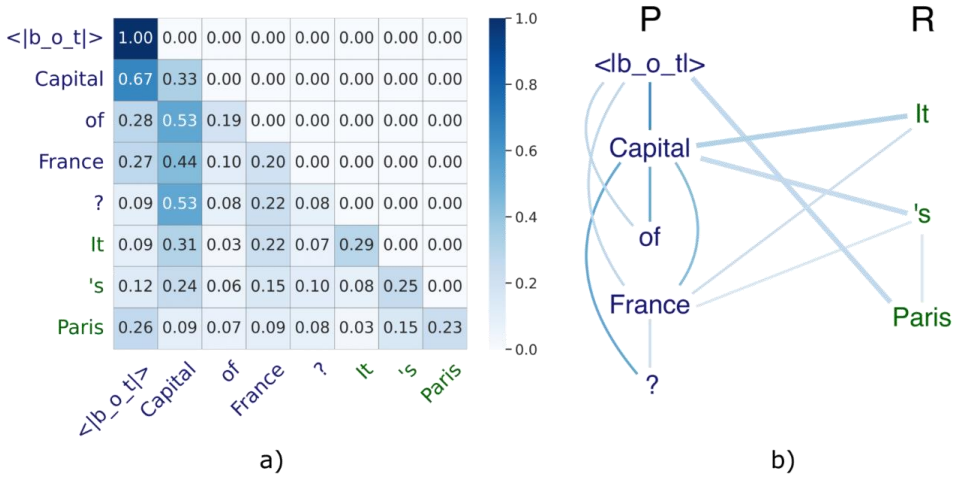
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V.$$

We explore the so-called attention maps:

$$\text{W} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

Each attention map can be interpreted as a graph, where edges represent the relationships between the tokens.

Popular “small LLMs” (Llama-2-7b, Mistral-7B) typically have 28-32 layers, 28-32 heads. So each generation induces ~800-1000 graphs.



Figures.
a) An attention map.
b) The corresponding attention graph.

Bazarova, A., et al. "Hallucination Detection in LLMs with Topological Divergence on Attention Graphs." arXiv preprint. 2025.
Oblovatny, R., Bazarova A., and Zaytsev A. "Attention Head Embeddings with Trainable Deep Kernels for Hallucination Detection in LLMs." arXiv preprint. 2025.

TOHA: the general idea

- **Idea:** distance of the prompt to the response correlates with the probability of a hallucination
- As a hallucination score, we consider topological divergence MTop-Div(P, R) [2] between the prompt, including RAG context (P) and the response (R) tokens in the attention graph.
- MTop-Div(R, P) = length of the MSF* attaching R to P

Intuition: for hallucinated samples, response seems to deviate from prompt more significantly than for the grounded ones, since a novel information is introduced.

* minimum spanning forest

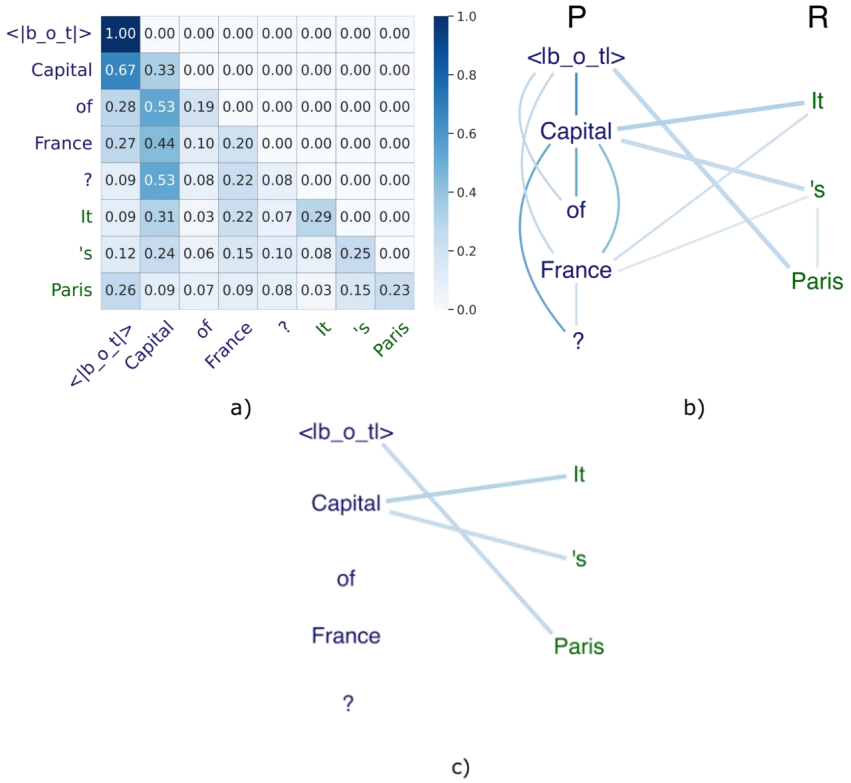


Figure. a) An attention map. b) The corresponding attention graph. c) The MSF under consideration.

[2] Barannikov S. et al. Manifold Topology Divergence: a Framework for Comparing Data Manifolds //Advances in neural information processing systems. – 2021. – Т. 34. – С. 7294-7305.

Concluding remarks

Conclusions

- CPD allows detection behavior drifts for customers, if used on top of representations.
- We proposed multiple approaches to boost quality in Neural CPD problems: SN normalization, InDID principled loss function, Wasserstein-based ensembles and local/global encoder selection.
- There are still more to discover and publish.

Links to papers

Spectral Norm

InDID loss

WWAggr ensemble

Local/global

Skoltech

Thanks

- Andrey Savchenko – for the invitation to give a talk and complete projects jointly with AI lab

- Evgeny Burnaev and Albert Nikolaevich Shiryaev – for giving an inspiration to start work on change point detection

- My lab members and co-authors: Evgenia Romanenkova, Alexandra Bazarova, Alexander Stepikin, Ilya Kuleshov, Alexander Yugai, Maria Kovaleva

References

[1] Page, E.S. (1954). CONTINUOUS INSPECTION SCHEMES. *Biometrika, 41*, 100-115.

[2] A. N. Shiryaev, “The problem of the most rapid detection of a disturbance in a stationary process”, *Sov. Math. Dokl.*, **2** (1961), 795–799

[3] Chatterjee, Sourav. “Changepoint Detection using Self Supervised Variational AutoEncoders.” (2021).

[4] Deldari S. et al. Time series change point detection with self-supervised contrastive predictive coding //Proceedings of the Web Conference 2021. – 2021. – С. 3124-3135.

[5] Xiangyu Bao, Liang Chen, Jingshu Zhong, Dianliang Wu, Yu Zheng,
A self-supervised contrastive change point detection method for industrial time series, Engineering Applications of Artificial Intelligence, Volume 133, Part B, 2024.

[6] Liu J. et al. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness //Advances in neural information processing systems. – 2020. – Т. 33. – С. 7498-7512.

[7] Miyato T. et al. Spectral normalization for generative adversarial networks //arXiv preprint arXiv:1802.05957. – 2018.

[8] Behrmann J. et al. Invertible residual networks //International conference on machine learning. – PMLR, 2019. – С. 573-582.

[9] Yue Z. et al. Ts2vec: Towards universal representation of time series //Proceedings of the AAAI Conference on Artificial Intelligence. – 2022. – Т. 36. – №. 8. – С. 8980-8987.]

[10] Grill J. B. et al. Bootstrap your own latent-a new approach to self-supervised learning //Advances in neural information processing systems. – 2020. – Т. 33. – С. 21271-21284.

[11] Chang W. C. et al. Kernel Change-point Detection with Auxiliary Deep Generative Models //International Conference on Learning Representations. – 2018.

[12] Deldari S. et al. Espresso: Entropy and shape aware time-series segmentation for processing heterogeneous sensor data //Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. – 2020. – Т. 4. – №. 3. – С. 1-24.

References - II

[13] G. Romano. Notes for MATH337 Changepoint Detection. https://www.lancaster.ac.uk/~romano/teaching/2425MATH337/
[14] Bazarova, A., et al. "Learning transactions representations for information management in banks: Mastering local, global, and external knowledge." International Journal of Information Management Data Insights 5.1 (2025): 100323.
[15] Oblovatny, R., Bazarova A., and Zaytsev A. "Attention Head Embeddings with Trainable Deep Kernels for Hallucination Detection in LLMs." arXiv preprint arXiv:2506.09886 (2025).
[16] Zhang, Yue, et al. 🧜 Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Computational Linguistics. 2025.

Backslides

InDiD: a data-driven model with novel loss function

We observe a dataset $D = \{(X_i\ \theta_i)\}_{i=1}^{N}$ of sequences X_i with true change points $\theta_i\ \in \{1, \dots, T, \infty\}$.

Putting all CPD–related challenges of data processing onto the shoulders of neural network $f_{\boldsymbol{w}}$, we force it to learn **proper representations** (in terms of CPD) via a **novel InDID loss function**:

$$\tilde{L}^h(f_{\mathbf{w}}, X_i, \theta_i) = \tilde{L}^h_{delay}(f_{\mathbf{w}}, X_i, \theta_i) - c\tilde{L}_{FA}(f_{\mathbf{w}}, X_i, \theta_i).$$

The behavior of the model $f_{\mathbf{w}}$ is defined in accordance with several assumptions.

Let $X^{1:t} = \{\mathbf{x}_i\}_{i=1}^{t}$ be an ordered set of independent random variables with a density p_θ supported on $[0,1]^d$.
Let $f_{\mathbf{w}} : X \subset \mathbb{R}^{td} \to (0,1)$ be an auxiliary function such that for any $t \in \mathbb{N}$ and $X^{1:t}$, it outputs $p_t \triangleq f_{\mathbf{w}}(X^{1:t})$, which we consider as the estimated probability of the true change point τ at the moment t.
Then:

(A1) *There exist such $q \in (0,1)$ and $\varepsilon \in (0,1)$ that $\mathbb{P}(p_t > 1 - q) \geq 1 - \varepsilon$ for $t \geq\ \theta$.*

(A2) *There exists such T that $(\tau - T\ |\tau > T) \sim \mathrm{Geom}(r)$, thereby implying $\mathbb{E}(p_t) \simeq r$.*

A1 means that the model is sufficiently good, and it holds after several training steps;

A2 controls the "tail“ behavior. In implementation, it is enough to set $r = 1$.

Romanenkova E. et al. “InDiD: Instant Disorder Detection via a Principled Neural Network”, ACM MM, 2022