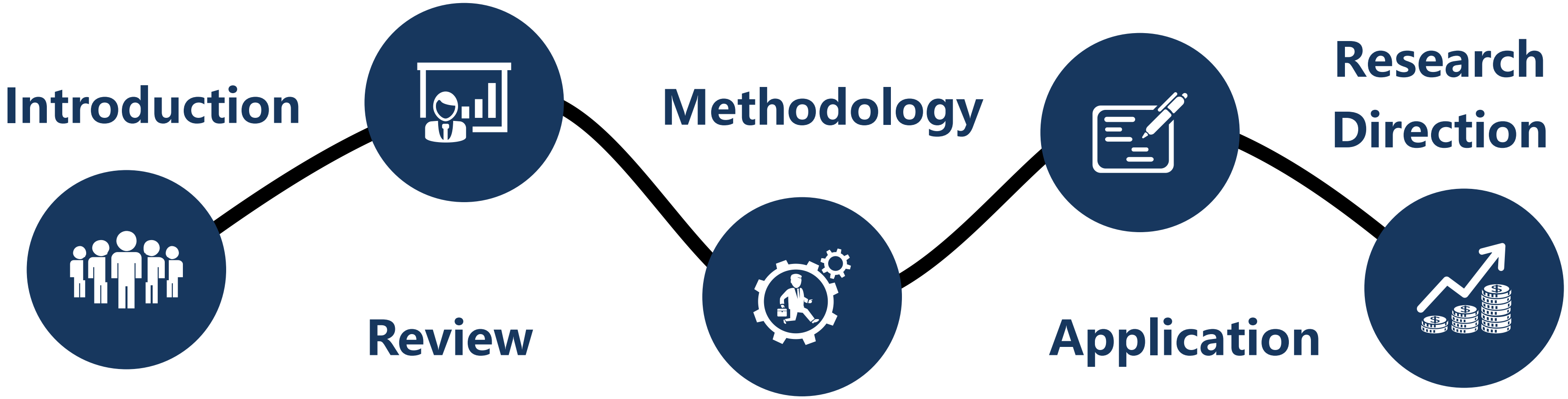
Research on Anomaly Detection for Streaming Time Series

ZhangYi

2019/9/20

CONTENTS



Introduction

Review

Methodology

Application

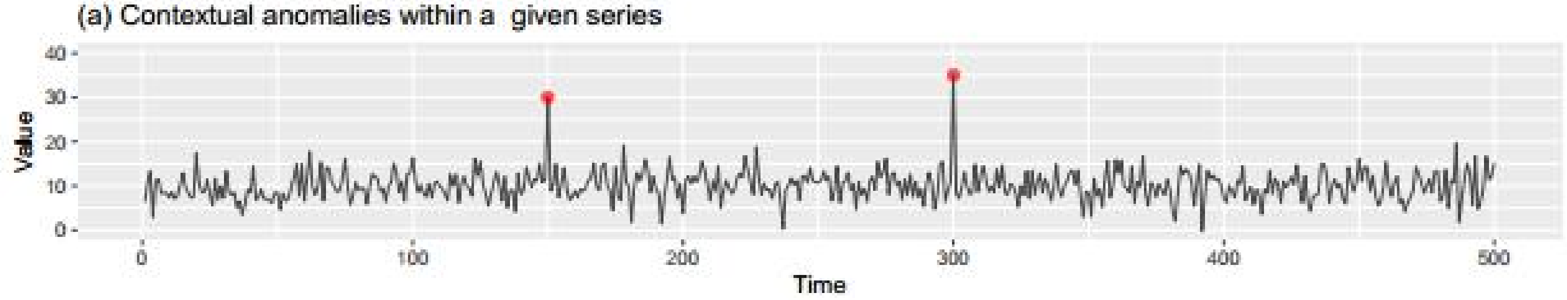
Research Direction

Introduction

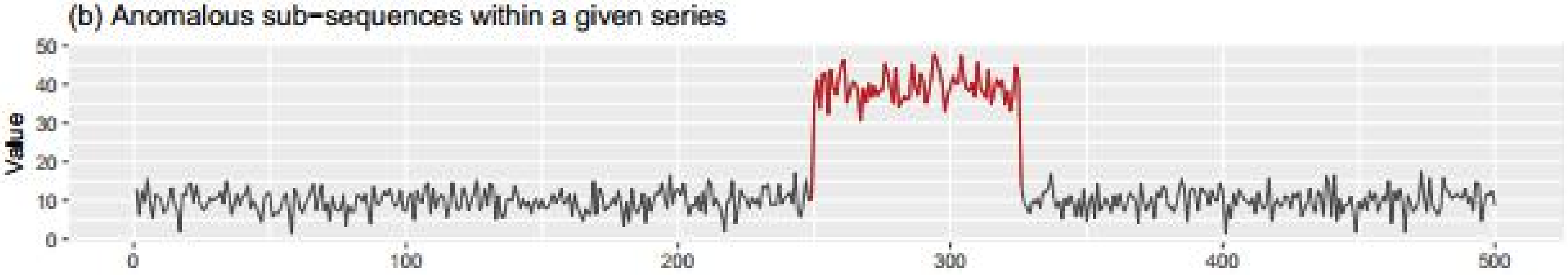
- Types of anomalies in temporal data
- Streaming data challenges

Types of anomalies in temporal data

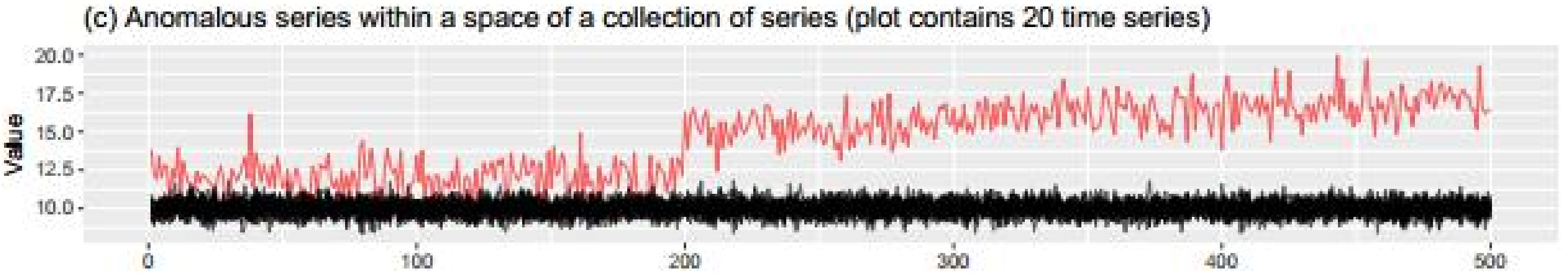
(a)


(a) Contextual anomalies within a given series

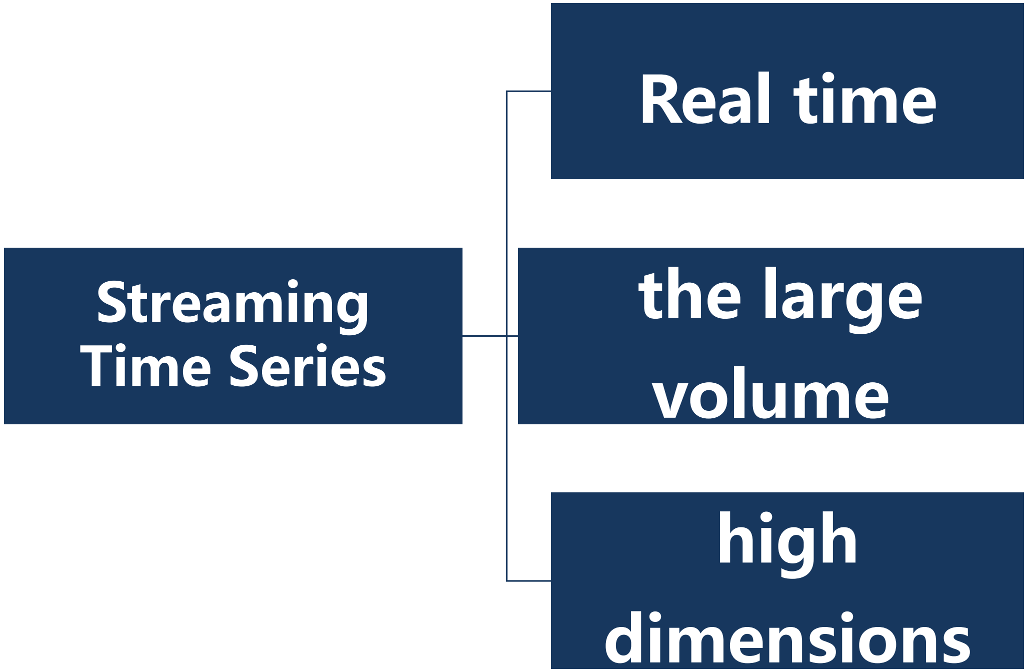
(b)


(b) Anomalous sub-sequences within a given series

(c)


(c) Anomalous series within a space of a collection of series (plot contains 20 time series)

Streaming data challenges



Streaming Time Series

Real time

Data is constantly changing, increasing the difficulty

the large volume

Difficult to distinguish between new typical behaviors and anomalous events

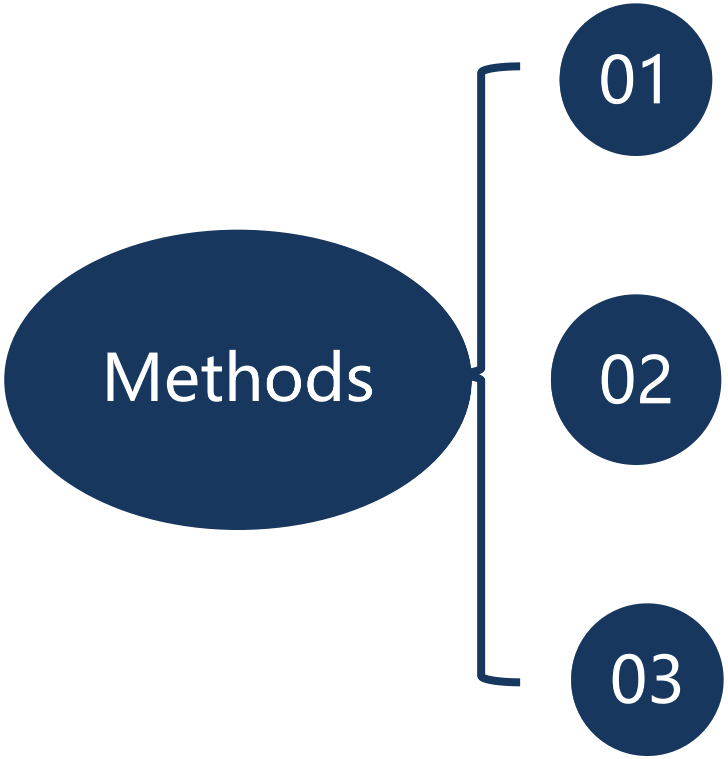
high dimensions

Need to learn how to handle multidimensional methods

Review

- Methods for time series anomaly detection
- Extreme value theory for anomaly detection

Methods for time series anomaly detection



Methods

01

Statistical based method

Inconsistent models or data at the edge of the model distribution，

Not applicable to high dimensional data

02

Data mining based approach

Distance, clustering, classification, density

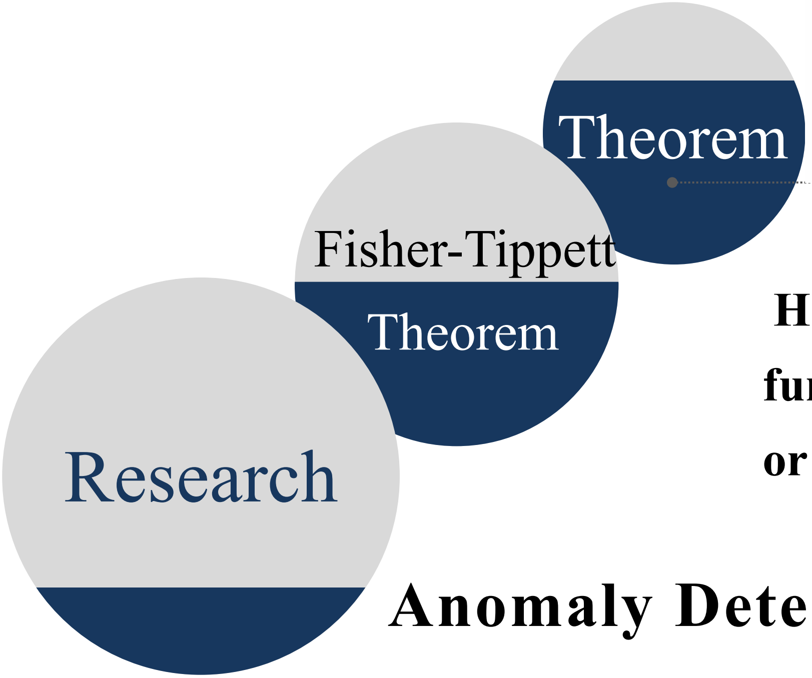
KNN， K-means， SVM

03

Extreme value theory

Select thresholds and evaluation criteria

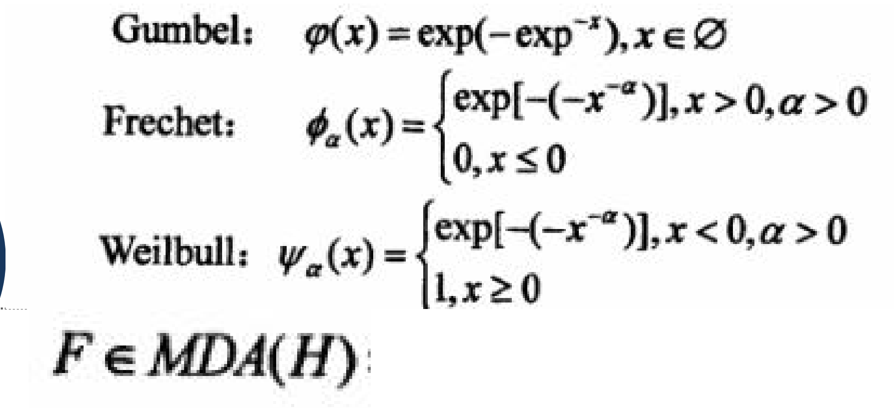
Extreme value theory for anomaly detection



Theorem

Fisher-Tippett Theorem

Research



Gumbel: $\varphi(x) = \exp(-\exp^{-x}), x \in \varnothing$

Frechet: $\phi_{\alpha}(x) = \begin{cases} \exp[-(-x^{-\alpha})], x > 0, \alpha > 0 \\ 0, x \leq 0 \end{cases}$

Weilbull: $\psi_{\alpha}(x) = \begin{cases} \exp[-(-x^{-\alpha})], x < 0, \alpha > 0 \\ 1, x \geq 0 \end{cases}$

$F \in MDA(H)$

H belongs to one of the three distribution function types: Frechet Fa+(x), Weibull Ya+(x) or Gumbel L+(x)

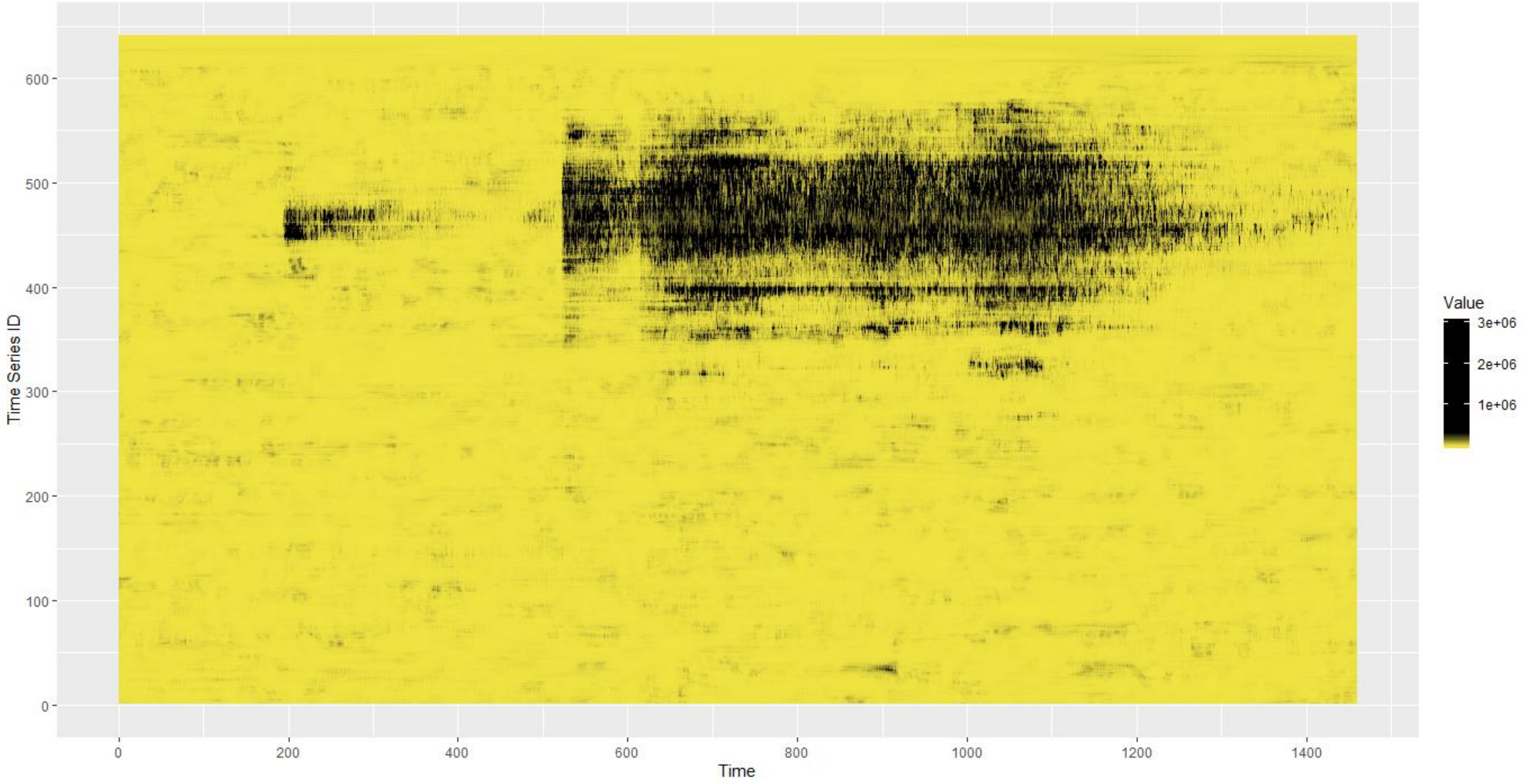
Anomaly Detection based on Extreme Value theory

Define a threshold for the density of the data points such that it distinguishes between anomalies

Methodology and Application

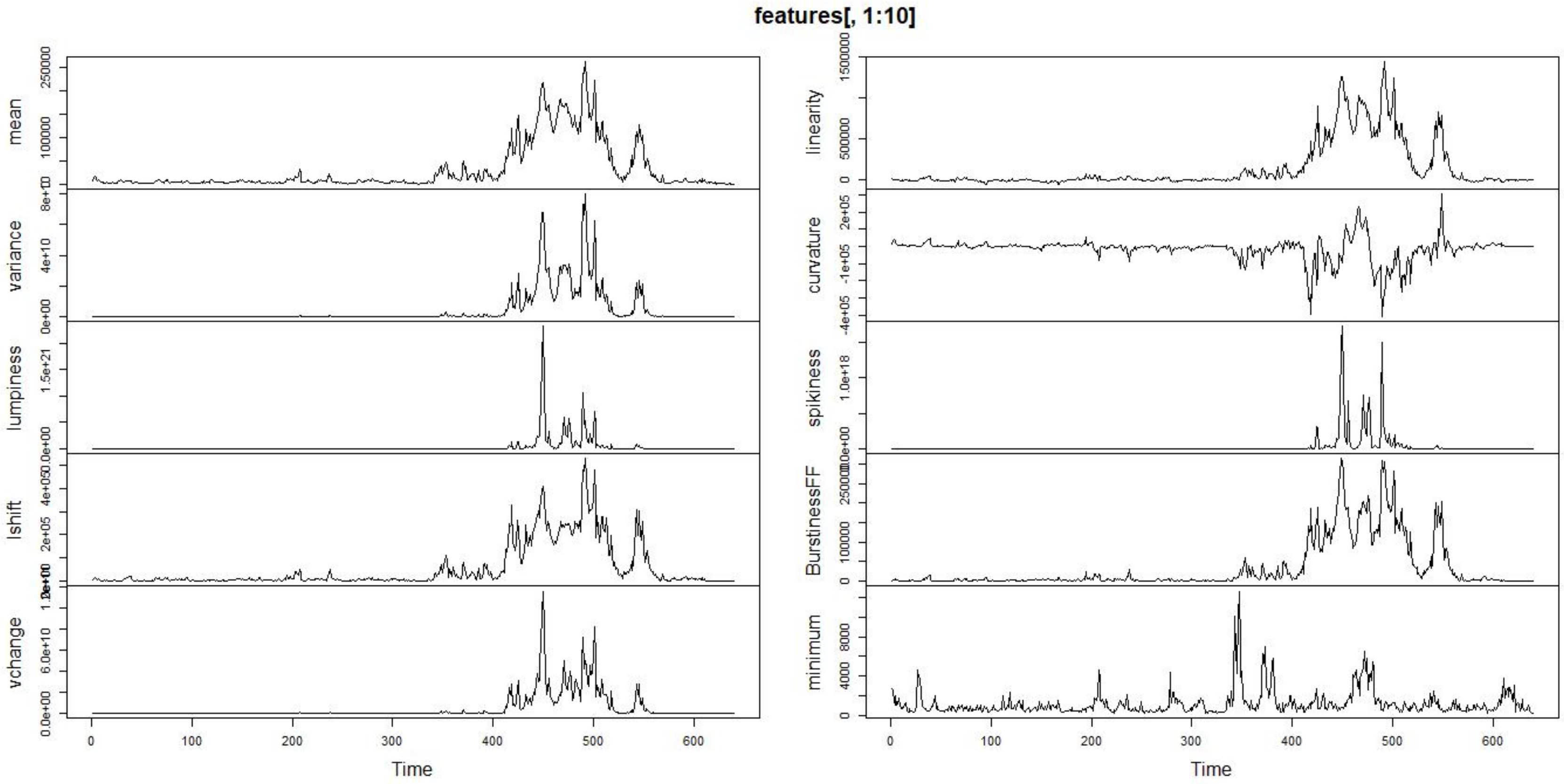
- Algorithm 1 : Building a model of the typical behavior
- Algorithm 2 : Testing newly-arrived data
- Algorithm 3 : Detection of non-stationarity

● Algorithm 1: Building a model of the typical behavior



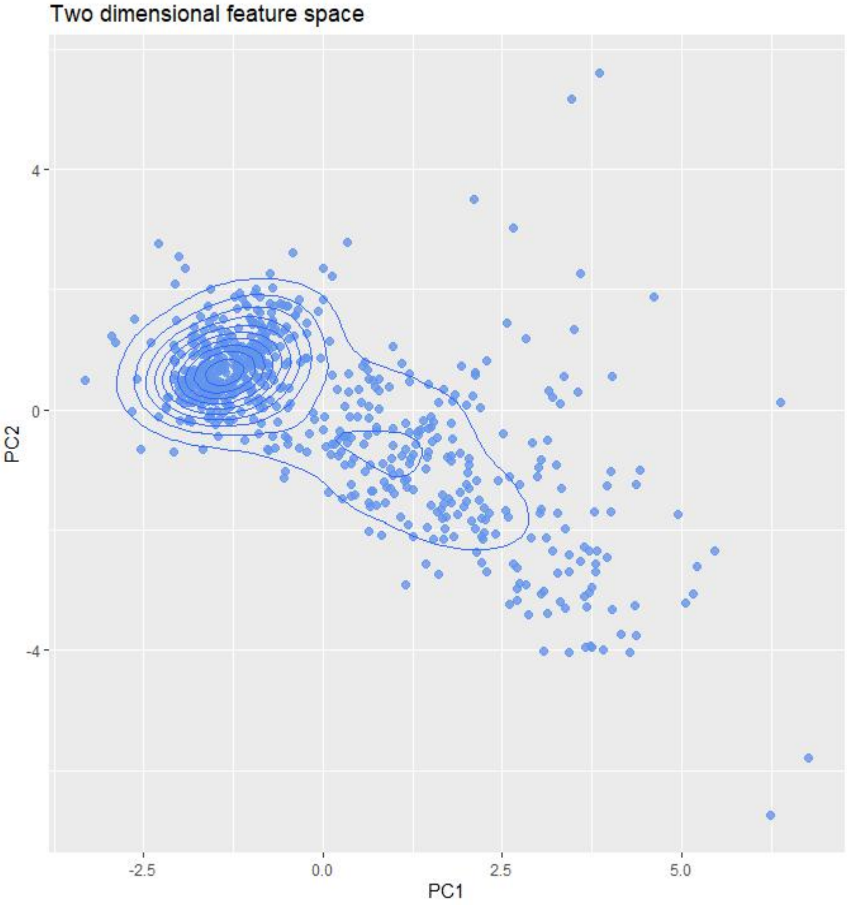
Data is constantly changing, increasing the difficulty

● Algorithm 1: Building a model of the typical behavior



Each plot is corresponding to a feature type extracted from the 640 time series. Almost all the features have captured the unusual event near the right end point of the cable (around 350 to 550).

● Algorithm 1: Building a model of the typical behavior



1.Define a two-dimensional space using the first two principal components (PC)

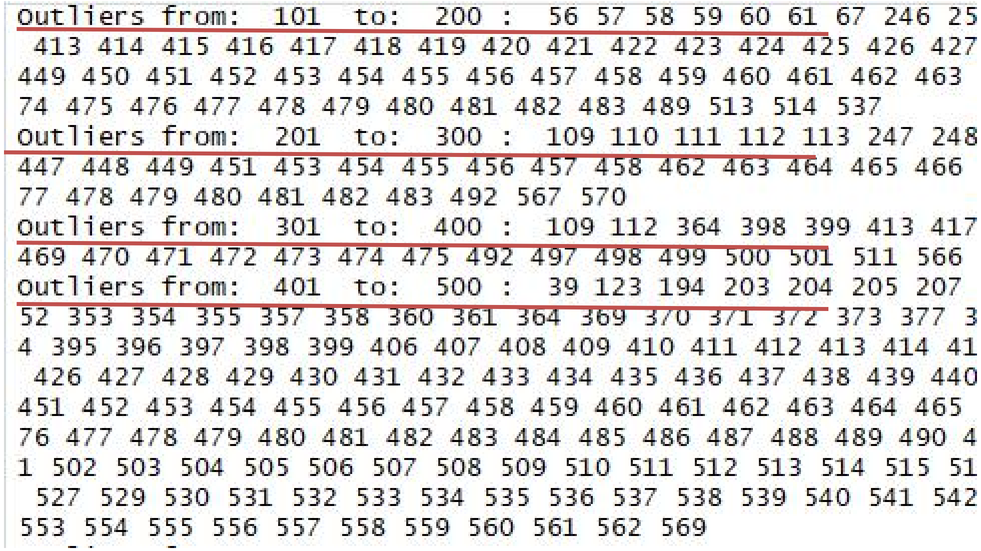
2.Each data point on this two-dimensional space corresponds to a time series in Dnorm.

3.Estimate the probability density of this 2D PC space using kernel density estimation with
a bivariate Gaussian kernel

● Algorithm 2 : Testing newly-arrived data

Our aim is now to identify time series that are anomalousrelative to the system's typical behavior

Considers the first window of the data set as the training set and the remaining as the test stream.



```
Outliers from:  101  to:  200 :  56 57 58 59 60 61 67 246 25
 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427
449 450 451 452 453 454 455 456 457 458 459 460 461 462 463
74 475 476 477 478 479 480 481 482 483 489 513 514 537
Outliers from:  201  to:  300 :  109 110 111 112 113 247 248
447 448 449 451 453 454 455 456 457 458 462 463 464 465 466
77 478 479 480 481 482 483 492 567 570
Outliers from:  301  to:  400 :  109 112 364 398 399 413 417
469 470 471 472 473 474 475 492 497 498 499 500 501 511 566
Outliers from:  401  to:  500 :  39 123 194 203 204 205 207
52 353 354 355 357 358 360 361 364 369 370 371 372 373 377 3
4 395 396 397 398 399 406 407 408 409 410 411 412 413 414 41
 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440
451 452 453 454 455 456 457 458 459 460 461 462 463 464 465
76 477 478 479 480 481 482 483 484 485 486 487 488 489 490 4
1 502 503 504 505 506 507 508 509 510 511 512 513 514 515 51
 527 529 530 531 532 533 534 535 536 537 538 539 540 541 542
553 554 555 556 557 558 559 560 561 562 569
```

Application

Handling non-stationarity:

Use statistical distance measures to measure the distance between the distribution generated from the collection of typical time series.

Practical application:

- **Industry (Gas/oil pipeline leakages, Water contaminated areas, Environmental monitoring)**
- **Medical （Epidemiological outbreaks）**
- **Financial (Fraud detection)**

Research Direction

- Current research
- Point of breakthrough

Research Direction

Current research

- Existing research is improving the accuracy of anomaly detection, making anomaly detection separate from the appearance of typical events.

- Improve the HDoutliers algorithm and get a new algorithm （ stray algorithm)

Research Direction

Sufficient workload

Improve accuracy

Innovation breakthrough

Wide range of applications

Point of breakthrough

- Time series data volume （ multi-noise, multivariate, variable correlation and time dimension unequal length）
- Reason of anomaly detection(Classification problem）
- Domain application selection （Sports lottery sales）
- ?

THANKS!