ONE-CLASS PEELING FOR OUTLIER DETECTION IN HIGH DIMENSIONS

Maria Weese
Miami University, Oxford, OH

June 12, 2018

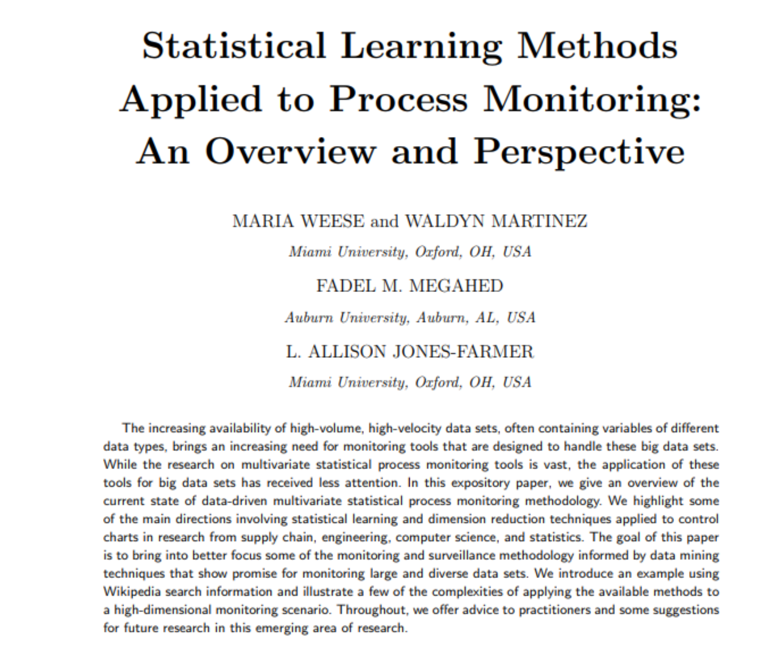
weeseml@miamioh.edu

CO-AUTHORS

Allison Jones-Farmer
Miami University

Waldyn Martinez
Miami University

EXAMPLE



Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective

MARIA WEESE and WALDYN MARTINEZ

Miami University, Oxford, OH, USA

FADEL M. MEGAHED

Auburn University, Auburn, AL, USA

L. ALLISON JONES-FARMER

Miami University, Oxford, OH, USA

The increasing availability of high-volume, high-velocity data sets, often containing variables of different data types, brings an increasing need for monitoring tools that are designed to handle these big data sets. While the research on multivariate statistical process monitoring tools is vast, the application of these tools for big data sets has received less attention. In this expository paper, we give an overview of the current state of data-driven multivariate statistical process monitoring methodology. We highlight some of the main directions involving statistical learning and dimension reduction techniques applied to control charts in research from supply chain, engineering, computer science, and statistics. The goal of this paper is to bring into better focus some of the monitoring and surveillance methodology informed by data mining techniques that show promise for monitoring large and diverse data sets. We introduce an example using Wikipedia search information and illustrate a few of the complexities of applying the available methods to a high-dimensional monitoring scenario. Throughout, we offer advice to practitioners and some suggestions for future research in this emerging area of research.

THE DATA

One way to understand public interest that is generated by the popular press is to consider monitoring social media (e.g. Twitter, Facebook, etc.) and/or data from web searches (e.g. Google, Yahoo, Wikipedia).

THE DATA

One way to understand public interest that is generated by the popular press is to consider monitoring social media (e.g. Twitter, Facebook, etc.) and/or data from web searches (e.g. Google, Yahoo, Wikipedia).

The National Football League (NFL) is concerned with monitoring “public relations events”, or at least separate out the typical traffic from an “event”.

OBTAINING THE DATA

`http://dumps.wikimedia.org/other/pagecounts-raw`

Contains the hourly number of hits on all Wikipedia pages (note there are over 2 million English Language pages, about 4.8 million total pages).

OBTAINING THE DATA

`http://dumps.wikimedia.org/other/pagecounts-raw`

Contains the hourly number of hits on all Wikipedia pages (note there are over 2 million English Language pages, about 4.8 million total pages).

Every hour contains a compressed file of approximately 100MB for the number of hits on millions of Wikipedia pages. A week of data holds over 16GB of storage.

SAMPLE

Our sample considers page hits per hour over a two week period for all NFL active players, coaches, managers and teams beginning on 9/1/2014 and was specifically chosen to include the first two weeks of the 2014 season for a total of n=354 samples.

SAMPLE

Our sample considers page hits per hour over a two week period for all NFL active players, coaches, managers and teams beginning on 9/1/2014 and was specifically chosen to include the first two weeks of the 2014 season for a total of n=354 samples.

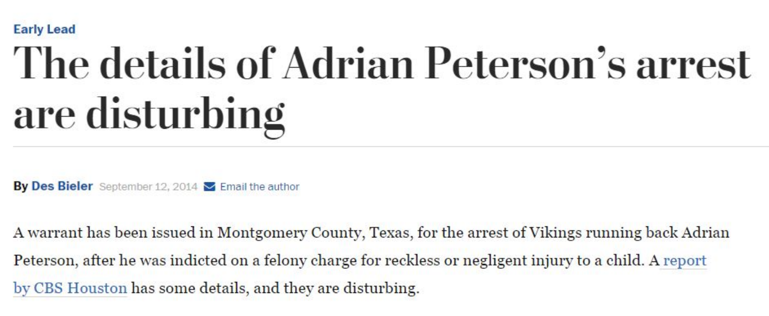
A signal to a potential event is defined as an unusually high number of Wikipedia hits on a particular team, coach, manager, or player.

SAMPLE

Our sample considers page hits per hour over a two week period for all NFL active players, coaches, managers and teams beginning on 9/1/2014 and was specifically chosen to include the first two weeks of the 2014 season for a total of n=354 samples.

A signal to a potential event is defined as an unusually high number of Wikipedia hits on a particular team, coach, manager, or player.

The number of pages for teams, currently active players, coaches, and managers is p=1917.

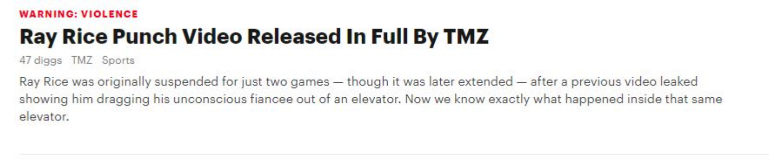


Early Lead

The details of Adrian Peterson's arrest are disturbing

By Des Bieler September 12, 2014 ✉ Email the author

A warrant has been issued in Montgomery County, Texas, for the arrest of Vikings running back Adrian Peterson, after he was indicted on a felony charge for reckless or negligent injury to a child. A report by CBS Houston has some details, and they are disturbing.

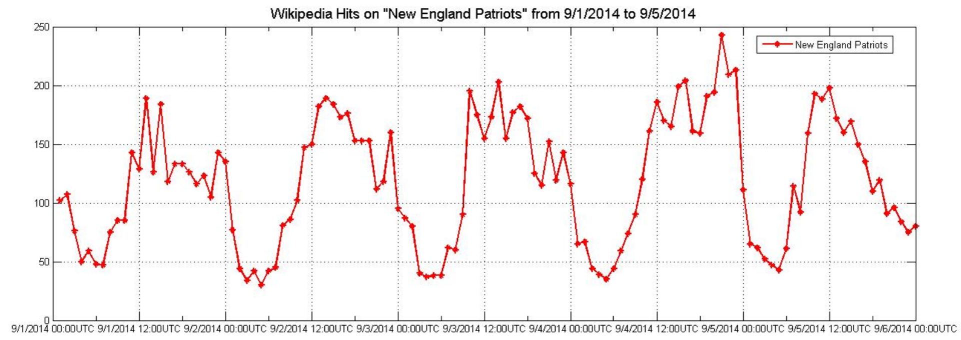


WARNING: VIOLENCE

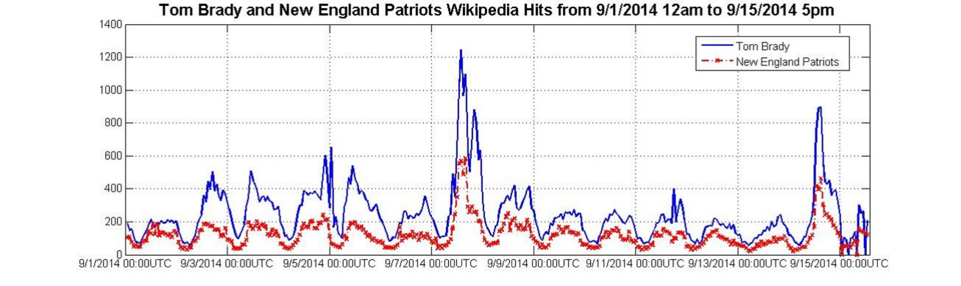
Ray Rice Punch Video Released In Full By TMZ

47 diggs TMZ Sports

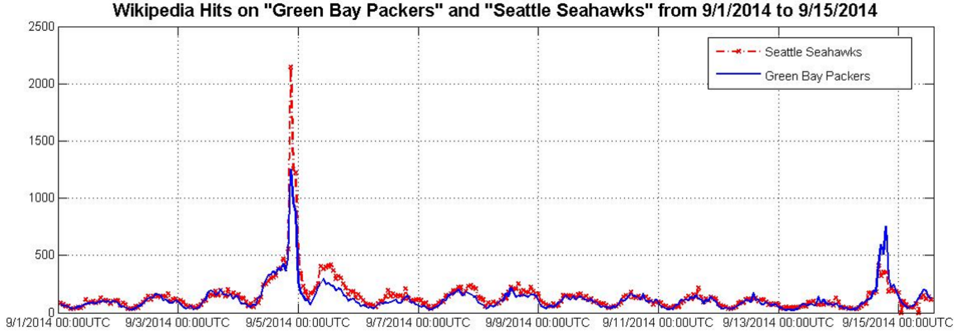
Ray Rice was originally suspended for just two games — though it was later extended — after a previous video leaked showing him dragging his unconscious fiancee out of an elevator. Now we know exactly what happened inside that same elevator.



Wikipedia Hits on "New England Patriots" from 9/1/2014 to 9/5/2014



Tom Brady and New England Patriots Wikipedia Hits from 9/1/2014 12am to 9/15/2014 5pm


Wikipedia Hits on "Green Bay Packers" and "Seattle Seahawks" from 9/1/2014 to 9/15/2014
2500
2000
1500
1000
500
0
9/1/2014 00:00UTC
9/3/2014 00:00UTC
9/5/2014 00:00UTC
9/7/2014 00:00UTC
9/9/2014 00:00UTC
9/11/2014 00:00UTC
9/13/2014 00:00UTC
9/15/2014 00:00UTC
Seattle Seahawks
Green Bay Packers

EVENT IDENTIFICATION

1. Searched ”NFL” on Google news in a custom search for a custom date range by day (based on PDT converted to EDT) during the two week period.
2. We only considered news stories that appeared on the first page of the search results (approximately 10 search returns).
3. If a news story was new to any period we consider it to be breaking news. If it was classified as breaking news then we went to the source of the news which first reported the story to find the exact time that the story was released and converted it to EDT.
4. If the story drops during the overnight hours 1am and 8am we start our period at 8am EDT the morning following.
5. If the story drops between 6am and 1am EDT then we consider a potential signal starting within a 5 hour lag of the time of the story hitting.
6. NFL games will signal based on the de-trending method.

METHOD REQUIREMENTS

1. A method that does not need to be stopped and re-calibrated once a signal is observed.
2. We need a method that can handle count data (not multivariate normal), contains many zero values, and has a nested correlation structure.
3. A method that can be adapted to data that occurs over time.

METHODS

THREE PARADIGMS

1. Statistical Distance Methods.
2. k-nearest neighbor (kNN) methods.
3. Density Estimation Methods.

THREE PARADIGMS

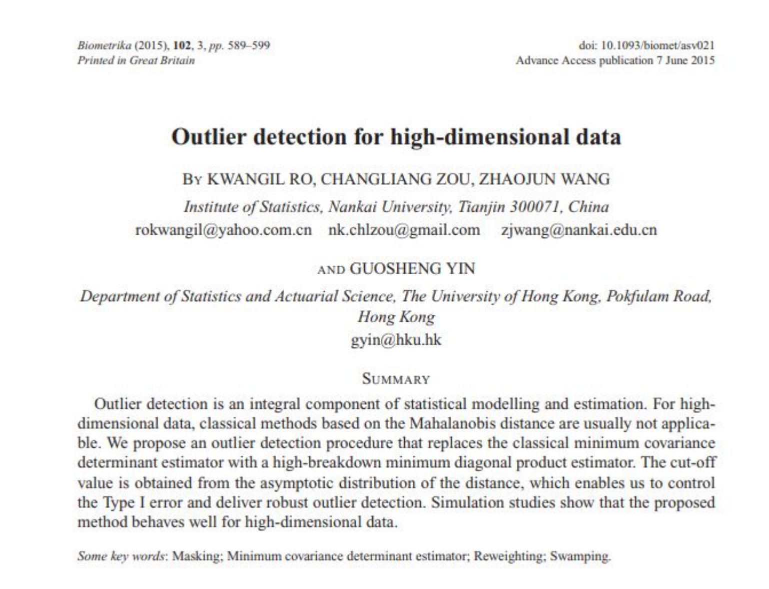
1. Statistical Distance Methods.
2. k-nearest neighbor (kNN) methods.
3. Density Estimation Methods.

STATISTICAL DISTANCE METHODS

Most multivariate outlier detection methods follow two steps:

1. Robust estimation of the center and scale of the data.
2. Evaluation of a measure of “outlyingness”, i.e. a distance measure.

The following three methods use these two steps and can be applied when $p > n$.



Biometrika (2015), **102**, 3, *pp.* 589–599
Printed in Great Britain

doi: 10.1093/biomet/asv021
Advance Access publication 7 June 2015

Outlier detection for high-dimensional data

BY KWANGIL RO, CHANGLIANG ZOU, ZHAOJUN WANG

Institute of Statistics, Nankai University, Tianjin 300071, China
rokwangil@yahoo.com.cn nk.chlzou@gmail.com zjwang@nankai.edu.cn

AND GUOSHENG YIN

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong
gyin@hku.hk

SUMMARY

Outlier detection is an integral component of statistical modelling and estimation. For high-dimensional data, classical methods based on the Mahalanobis distance are usually not applicable. We propose an outlier detection procedure that replaces the classical minimum covariance determinant estimator with a high-breakdown minimum diagonal product estimator. The cut-off value is obtained from the asymptotic distribution of the distance, which enables us to control the Type I error and deliver robust outlier detection. Simulation studies show that the proposed method behaves well for high-dimensional data.

Some key words: Masking; Minimum covariance determinant estimator; Reweighting; Swamping.

R-MDP METHOD

Ro et al. (2015) use a modified Mahalanobis (see equation 1) distance that uses only the diagonal elements of the sample covariance matrix.

$$d_i^2(\mu, D) = (Y_i - \mu)^T D^{-1} (Y_i - \mu) \tag{1}$$

Where D and μ are estimated from a subset of observations such that determinant of the diagonal elements of the covariance for that subset of observations is minimal.

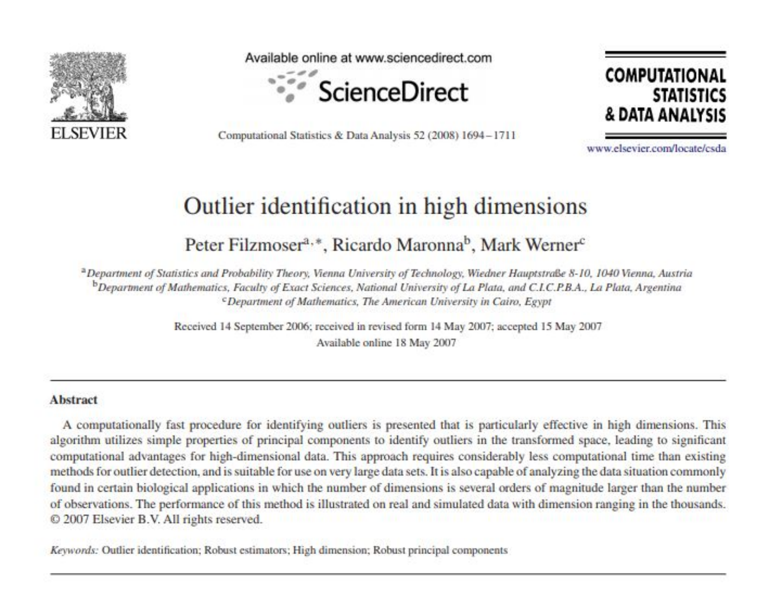
R-MDP METHOD

Ro et al. (2015) use a modified Mahalanobis (see equation 1) distance that uses only the diagonal elements of the sample covariance matrix.

$$d_i^2(\mu, D) = (Y_i - \mu)^T D^{-1} (Y_i - \mu) \tag{1}$$

Where D and μ are estimated from a subset of observations such that determinant of the diagonal elements of the covariance for that subset of observations is minimal.

Outliers are determined by setting a significance level α, defining a rejection region. Points with distances in the rejection region are flagged.



ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 1694–1711

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Outlier identification in high dimensions

Peter Filzmoser[a,*], Ricardo Maronna[b], Mark Werner[c]

[a] *Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria*
[b] *Department of Mathematics, Faculty of Exact Sciences, National University of La Plata, and C.I.C.P.B.A., La Plata, Argentina*
[c] *Department of Mathematics, The American University in Cairo, Egypt*

Received 14 September 2006; received in revised form 14 May 2007; accepted 15 May 2007
Available online 18 May 2007

Abstract

A computationally fast procedure for identifying outliers is presented that is particularly effective in high dimensions. This algorithm utilizes simple properties of principal components to identify outliers in the transformed space, leading to significant computational advantages for high-dimensional data. This approach requires considerably less computational time than existing methods for outlier detection, and is suitable for use on very large data sets. It is also capable of analyzing the data situation commonly found in certain biological applications in which the number of dimensions is several orders of magnitude larger than the number of observations. The performance of this method is illustrated on real and simulated data with dimension ranging in the thousands.

Keywords: Outlier identification; Robust estimators; High dimension; Robust principal components

PCOUT

Filmozer et al. (2008) first reduce the dimension using PCA.

1. Run PCA and keep PC's to explain 99% of the original variation.
2. Location outliers are ranked on what is equivalent to a robust Mahalanobis distance (w_{1i}).
3. Scale outliers are ranked determined based on a modified bi-weight function (w_{2i}).
4. Each data point is then assigned a weight according to:

$$w_i = \frac{(w_{1i} + s)(w_{2i} + s)}{(1 + s)^2}$$

PCOUT

Filmozer et al. (2008) first reduce the dimension using PCA.

1. Run PCA and keep PC's to explain 99% of the original variation.
2. Location outliers are ranked on what is equivalent to a robust Mahalanobis distance (w_{1i}).
3. Scale outliers are ranked determined based on a modified bi-weight function (w_{2i}).
4. Each data point is then assigned a weight according to:

$$w_i = \frac{(w_{1i} + s)(w_{2i} + s)}{(1 + s)^2}$$

Outliers are classified as those observations that have $w_i < 0.25$ with $s = 0.25$

OCP METHOD

ROBUST ESTIMATION OF THE CENTER

The One-Class Peeling (OCP) method uses Support Vector Data Description (SVDD) to peel away outlying observations, similar to Convex Hull Peeling.

The mean estimate, $\hat{\mu}_{\text{OCP}}$, is the mean of the last two observations remaining after the others are peeled.

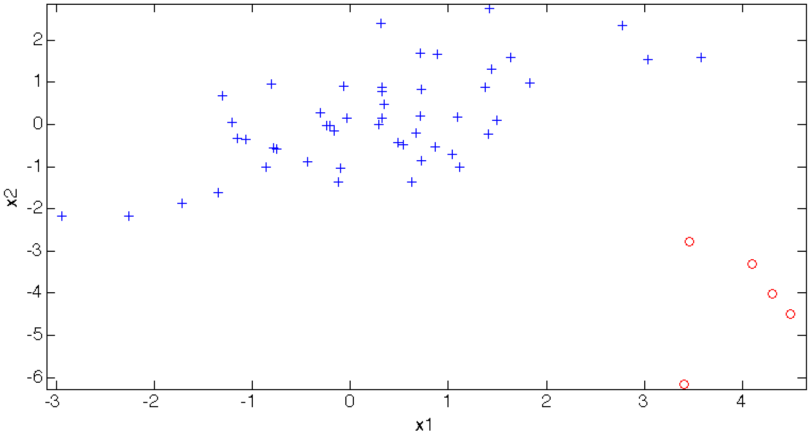
Distances are relative to the value of $\hat{\mu}_{\text{OCP}}$.

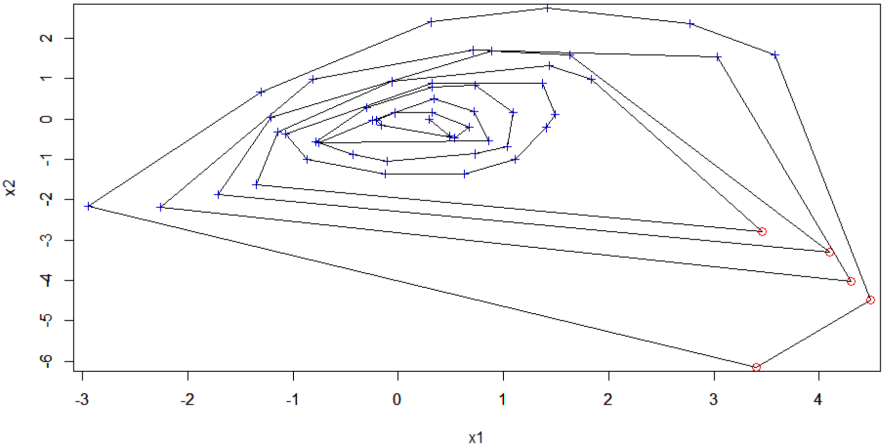
SUPPORT VECTOR DATA DESCRIPTION

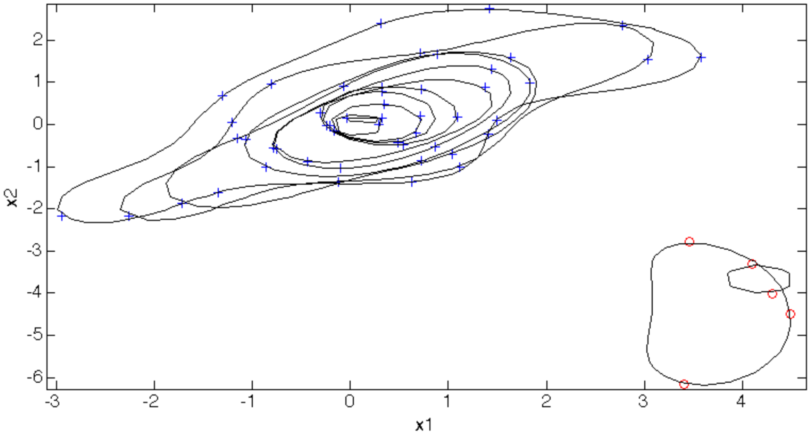
- SVDD creates a hyper-sphere boundary around a sample of data.
- The boundary is created using only a few points called support vectors.
- The user can specify how tight the boundary fits the data.
- If the data are mapped to the kernel space then SVDD can create a flexible boundary around a sample of data.
- A common choice is the Gaussian kernel:

$$KS_G(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{s^2}\right) \tag{2}$$

- We set $s = p$ based on Weese et al. (2016).


x2
2
1
0
-1
-2
-3
-4
-5
-6
-3
-2
-1
0
1
2
3
4
x1


2
1
0
-1
-2
-3
-4
-5
-6
x2
-3
-2
-1
0
1
2
3
4
x1


2
1
0
-1
-2
-3
-4
-5
-6
x2
-3
-2
-1
0
1
2
3
4
x1

OCP MEAN ESTIMATE BREAKDOWN POINT

A breakdown point describes the least percentage of outliers in a set S at which the estimate becomes arbitrarily incorrect.

We show empirically that the breakdown point of the OCP method is between 30% and 35% contamination for correlated and uncorrelated normal and t (df=10) distributed data for dimensions of $p = 25$ to $p = 100$.

DISTANCE DETERMINATION

The second step of the OCP method calculates a distance measure from each observation to the estimated mean, $\hat{\mu}_{\text{OCP}}$.

We use the Gaussian kernel to retain the same feature space that was used to develop the SVDD boundaries.

Recall, a kernel function is also a similarity function.

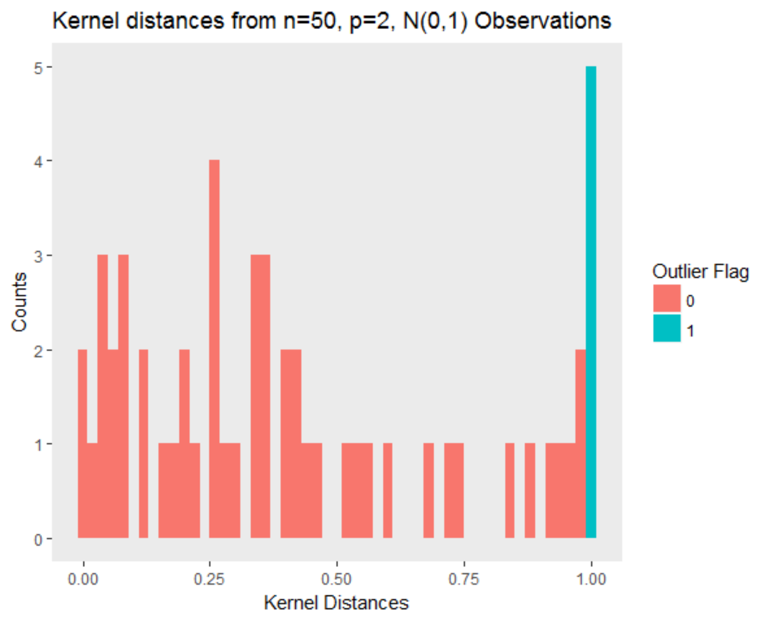
DISTANCE DETERMINATION

The Gaussian kernel similarity (Equation 2) is a decreasing function of the Euclidean distance between two arbitrary points and $0 \leq \mathrm{KS}_\mathrm{G}(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j}) \leq 1$.

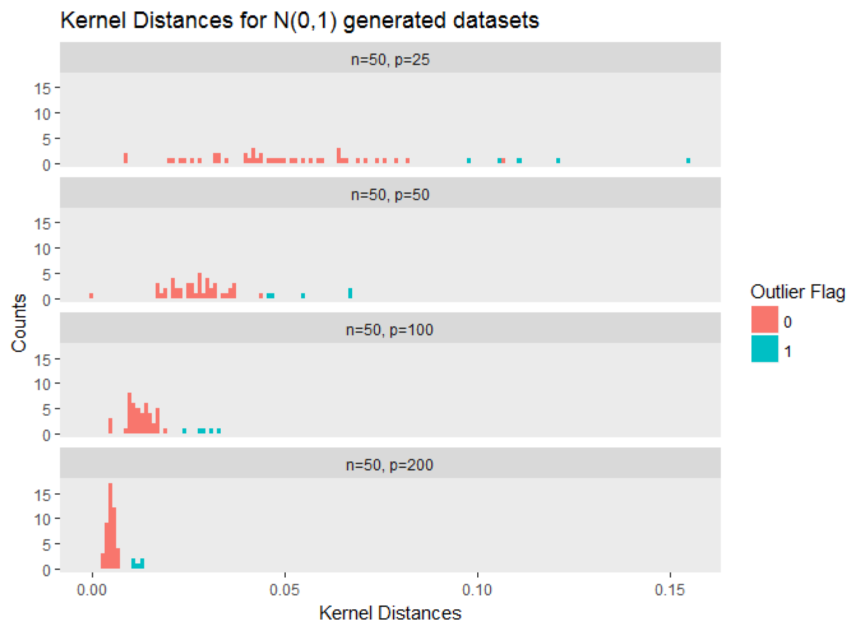
We use the following linear transformation of Equation 2 to form our distance metric:

$$\mathrm{KD}_\mathrm{OCP}(\mathbf{x}_\mathrm{i}, \hat{\boldsymbol{\mu}}_\mathrm{OCP}) = 1 - \exp\left(-\frac{\|\mathbf{x}_\mathrm{i} - \hat{\boldsymbol{\mu}}_\mathrm{OCP}\|^2}{s^2}\right) \tag{3}$$

Smaller values of $\mathrm{KD}_\mathrm{OCP}(\mathbf{x}_\mathrm{i}, \hat{\boldsymbol{\mu}}_\mathrm{OCP})$ (closer to 0) indicate observations close to the estimated mean $\hat{\boldsymbol{\mu}}_\mathrm{OCP}$, while larger values of $\mathrm{KD}_\mathrm{OCP}(\mathbf{x}_\mathrm{i}, \hat{\boldsymbol{\mu}}_\mathrm{OCP})$ (closer to 1) indicate observations far away from $\hat{\boldsymbol{\mu}}_\mathrm{OCP}$.



Kernel distances from n=50, p=2, N(0,1) Observations



Kernel Distances for N(0,1) generated datasets

DETERMINING AN OUTLIER

Unfortunately the values of the $\mathrm{KD}_{\mathrm{OCP}}$ (Equation 3) change with distribution and sample size.

So we scale the $\mathrm{KD}_{\mathrm{OCP}}$ by

$$\mathrm{sRKD} = \frac{\mathrm{KD}_{\mathrm{OCP}} - \mathrm{median}(\mathrm{KD}_{\mathrm{OCP}})}{\mathrm{sMAD}(\mathrm{KD}_{\mathrm{OCP}})}$$

where the $\mathrm{sMAD} = b * \mathrm{med}_i|x_i - \mathrm{med}_j x_j|$ and $b = \frac{1}{\sqrt{2}} * \Phi^{-1}(ct/2)$ and where Φ^{-1} is the inverse complementary error function.

DETERMINING AN OUTLIER

For any distribution, an sRKD > 4 is flagged as an outlier using factor $b = 1.3238$ for sMAD.

This controls the empirical Type I error rate to a maximum 5% for data of any distribution.

But of course a user can change the threshold at which they wish to flag potential outliers.

WHAT ABOUT CHANGES THROUGH TIME?

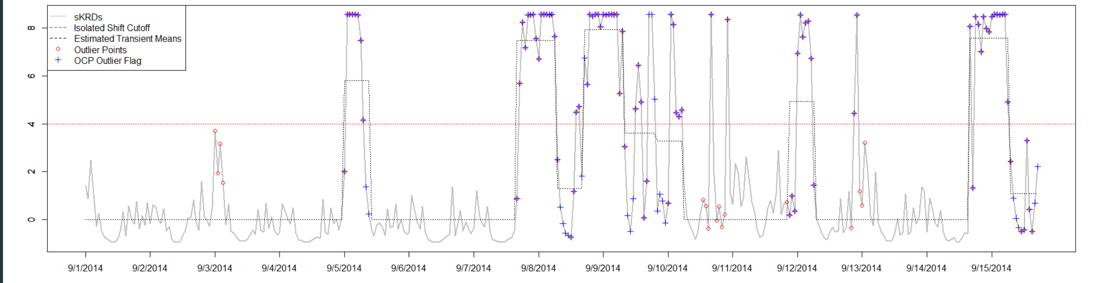
Since the sRKD values are univariate, there are many options.

We choose to use the RSP change point method of Capizzi and Masarotto (2017).

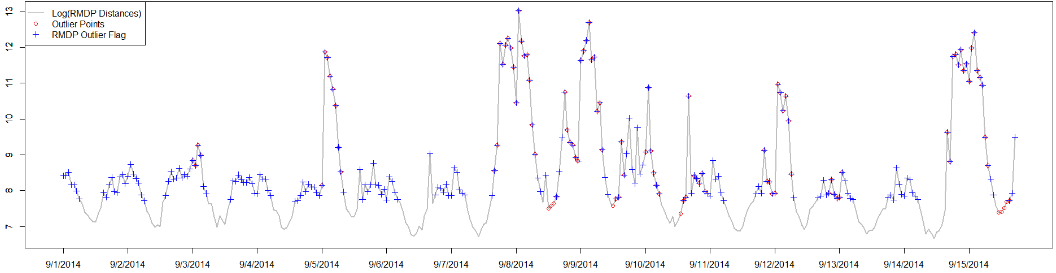
This method will signal if the mean of the sRKDs has shifted during the time period.

We choose $\alpha = 0.05$ for the RSP method.

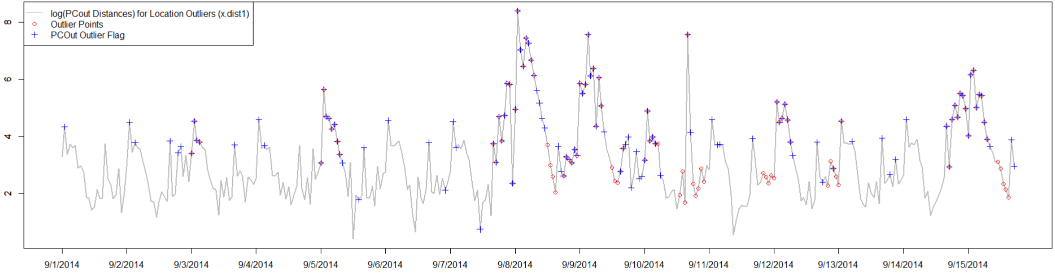
EXAMPLE RESULTS



sKRDs
Isolated Shift Cutoff
Estimated Transient Means
Outlier Points
OCP Outlier Flag
0
2
4
6
8
9/1/2014
9/2/2014
9/3/2014
9/4/2014
9/5/2014
9/6/2014
9/7/2014
9/8/2014
9/9/2014
9/10/2014
9/11/2014
9/12/2014
9/13/2014
9/14/2014
9/15/2014



Log(RMDP Distances)
○ Outlier Points
+ RMDP Outlier Flag
13
12
11
10
9
8
7
9/1/2014
9/2/2014
9/3/2014
9/4/2014
9/5/2014
9/6/2014
9/7/2014
9/8/2014
9/9/2014
9/10/2014
9/11/2014
9/12/2014
9/13/2014
9/14/2014
9/15/2014


log(PCout Distances) for Location Outliers (x.dist1)
○ Outlier Points
+ PCOut Outlier Flag
8
6
4
2
9/1/2014
9/2/2014
9/3/2014
9/4/2014
9/5/2014
9/6/2014
9/7/2014
9/8/2014
9/9/2014
9/10/2014
9/11/2014
9/12/2014
9/13/2014
9/14/2014
9/15/2014

RESULTS

Method	Detection Rate (%)	False Positive Rate (%)
OCP	84.7	9.2
R-MDP	91.4	56.6
PCOut	71.4	19.2

*Recall, events were labeled according to the protocol on slide 12.

RESULTS

Method	Detection Rate (%)	False Positive Rate (%)
OCP	84.7	9.2
R-MDP	91.4	56.6
PCOut	71.4	19.2

*Recall, events were labeled according to the protocol on slide 12.

RESULTS

Method	Detection Rate (%)	False Positive Rate (%)
OCP	84.7	9.2
R-MDP	91.4	56.6
PCOut	71.4	19.2

*Recall, events were labeled according to the protocol on slide 12.

RESULTS

Method	Detection Rate (%)	False Positive Rate (%)
OCP	84.7	9.2
R-MDP	91.4	56.5
PCOut	71.4	19.2

*Recall, events were labeled according to the protocol on slide 12.

A FEW SIMULATIONS

SIMULATION PROTOCOL

We simulated several different sample sizes and dimensions with 0, 5, 10, 20 and 30 percent outliers.

We did this for data sampled from correlated and uncorrelated Normal, correlated Lognormal and correlated t(df=10).

To maintain equivalent shifts across distributions we shifted the outlying observations probabilistically. We shifted in random directions.

The following results are for sustained outlier shifts, i.e. a step change starting at a random time. Results for isolated and transient shifts are similar.

SIMULATION RESULTS

		0% Outliers		
		OCP	R-MDP	PCOut
		Total Error	Total Error	Total Error
Normal	N = 50, p = 100	0.30%	22.83%	10.02%
Uncorrelated	N = 100, p = 100	0.26%	12.61%	7.99%
Normal	N = 50, p = 100	0.44%	20.70%	10.02%
Correlated	N = 100, p = 100	0.39%	12.10%	9.01%
T, df=10	N = 50, p = 100	4.73%	38.79%	14.45%
	N = 100, p = 100	5.14%	41.51%	17.91%
Lognormal	N = 50, p = 100	6.03%	49.20%	20.82%
	N = 100, p = 100	6.55%	52.28%	23.31%

SIMULATION RESULTS

		0% Outliers		
		OCP	R-MDP	PCOut
		Total Error	Total Error	Total Error
Normal	N = 50, p = 100	0.30%	22.83%	10.02%
Uncorrelated	N = 100, p = 100	0.26%	12.61%	7.99%
Normal	N = 50, p = 100	0.44%	20.70%	10.02%
Correlated	N = 100, p = 100	0.39%	12.10%	9.01%
T, df=10	N = 50, p = 100	4.73%	38.79%	14.45%
	N = 100, p = 100	5.14%	41.51%	17.91%
Lognormal	N = 50, p = 100	6.03%	49.20%	20.82%
	N = 100, p = 100	6.55%	52.28%	23.31%

SIMULATION RESULTS

10% Outliers

		OCP		R-MDP		PCOut	
		Detection	Total Error	Detection	Total Error	Detection	Total Error
Normal	N = 50, p = 100	100.00%	0.02%	99.84%	18.15%	100.00%	3.63%
Uncorrelated	N = 100, p = 100	100.00%	0.03%	100.00%	10.24%	100.00%	3.26%
Normal	N = 50, p = 100	99.90%	0.13%	99.80%	14.20%	99.98%	3.92%
Correlated	N = 100, p = 100	100.00%	0.08%	100.00%	9.17%	100.00%	3.59%
T, df=10	N = 50, p = 100	100.00%	2.01%	100.00%	32.22%	100.00%	9.26%
	N = 100, p = 100	100.00%	1.84%	100.00%	35.19%	100.00%	12.98%
Lognormal	N = 50, p = 100	96.66%	7.20%	99.30%	39.94%	98.96%	15.99%
	N = 100, p = 100	98.67%	5.78%	99.37%	43.37%	99.36%	19.09%

SIMULATION RESULTS

10% Outliers

		OCP		R-MDP		PCOut	
		Detection	Total Error	Detection	Total Error	Detection	Total Error
Normal	N = 50, p = 100	100.00%	0.02%	99.84%	18.15%	100.00%	3.63%
Uncorrelated	N = 100, p = 100	100.00%	0.03%	100.00%	10.24%	100.00%	3.26%
Normal	N = 50, p = 100	99.90%	0.13%	99.80%	14.20%	99.98%	3.92%
Correlated	N = 100, p = 100	100.00%	0.08%	100.00%	9.17%	100.00%	3.59%
T, df=10	N = 50, p = 100	100.00%	2.01%	100.00%	32.22%	100.00%	9.26%
	N = 100, p = 100	100.00%	1.84%	100.00%	35.19%	100.00%	12.98%
Lognormal	N = 50, p = 100	96.66%	7.20%	99.30%	39.94%	98.96%	15.99%
	N = 100, p = 100	98.67%	5.78%	99.37%	43.37%	99.36%	19.09%

SIMULATION RESULTS

10% Outliers

		OCP		R-MDP		PCOut	
		Detection	Total Error	Detection	Total Error	Detection	Total Error
Normal	N = 50, p = 100	100.00%	0.02%	99.84%	18.15%	100.00%	3.63%
Uncorrelated	N = 100, p = 100	100.00%	0.03%	100.00%	10.24%	100.00%	3.26%
Normal	N = 50, p = 100	99.90%	0.13%	99.80%	14.20%	99.98%	3.92%
Correlated	N = 100, p = 100	100.00%	0.08%	100.00%	9.17%	100.00%	3.59%
T, df=10	N = 50, p = 100	100.00%	2.01%	100.00%	32.22%	100.00%	9.26%
	N = 100, p = 100	100.00%	1.84%	100.00%	35.19%	100.00%	12.98%
Lognormal	N = 50, p = 100	96.66%	7.20%	99.30%	39.94%	98.96%	15.99%
	N = 100, p = 100	98.67%	5.78%	99.37%	43.37%	99.36%	19.09%

CONCLUSIONS

CONCLUSIONS

We have introduced a multivariate One-class peeling method for outlier detection (as a part of Phase I) particularly useful when the dimension of the data, p is large.

- The OCP method does not require covariance estimation.

CONCLUSIONS

We have introduced a multivariate One-class peeling method for outlier detection (as a part of Phase I) particularly useful when the dimension of the data, p is large.

- The OCP method does not require covariance estimation.
- The OCP method allows you to robustly estimate the center of the data with up to 30% outliers.

CONCLUSIONS

We have introduced a multivariate One-class peeling method for outlier detection (as a part of Phase I) particularly useful when the dimension of the data, p is large.

- The OCP method does not require covariance estimation.
- The OCP method allows you to robustly estimate the center of the data with up to 30% outliers.
- The OCP method robustly estimates the distance of observations.

CONCLUSIONS

We have introduced a multivariate One-class peeling method for outlier detection (as a part of Phase I) particularly useful when the dimension of the data, p is large.

- The OCP method does not require covariance estimation.
- The OCP method allows you to robustly estimate the center of the data with up to 30% outliers.
- The OCP method robustly estimates the distance of observations.
- The OCP method can be used with large p and when $p > n$.

CONCLUSIONS

We have introduced a multivariate One-class peeling method for outlier detection (as a part of Phase I) particularly useful when the dimension of the data, p is large.

- The OCP method does not require covariance estimation.
- The OCP method allows you to robustly estimate the center of the data with up to 30% outliers.
- The OCP method robustly estimates the distance of observations.
- The OCP method can be used with large p and when $p > n$.
- The OCP method out performs existing methods in terms of total error rate using a realistic data stream application.

CONCLUSIONS

We have introduced a multivariate One-class peeling method for outlier detection (as a part of Phase I) particularly useful when the dimension of the data, p is large.

- The OCP method does not require covariance estimation.
- The OCP method allows you to robustly estimate the center of the data with up to 30% outliers.
- The OCP method robustly estimates the distance of observations.
- The OCP method can be used with large p and when $p > n$.
- The OCP method out performs existing methods in terms of total error rate using a realistic data stream application.
- The OCP method out performs existing methods in terms of total error rate using simulation data.

CONCLUSIONS

We have introduced a multivariate One-class peeling method for outlier detection (as a part of Phase I) particularly useful when the dimension of the data, p is large.

- The OCP method does not require covariance estimation.
- The OCP method allows you to robustly estimate the center of the data with up to 30% outliers.
- The OCP method robustly estimates the distance of observations.
- The OCP method can be used with large p and when $p > n$.
- The OCP method out performs existing methods in terms of total error rate using a realistic data stream application.
- The OCP method out performs existing methods in terms of total error rate using simulation data.
- In general, using the OCP method will lead to a larger baseline sample compared to R-MDP and PCOut.

Questions?

REFERENCES

Capizzi, G., & Masarotto, G. (2017). Phase I Distribution-Free Analysis of Multivariate Data. Technometrics, 59(4), 484-495.

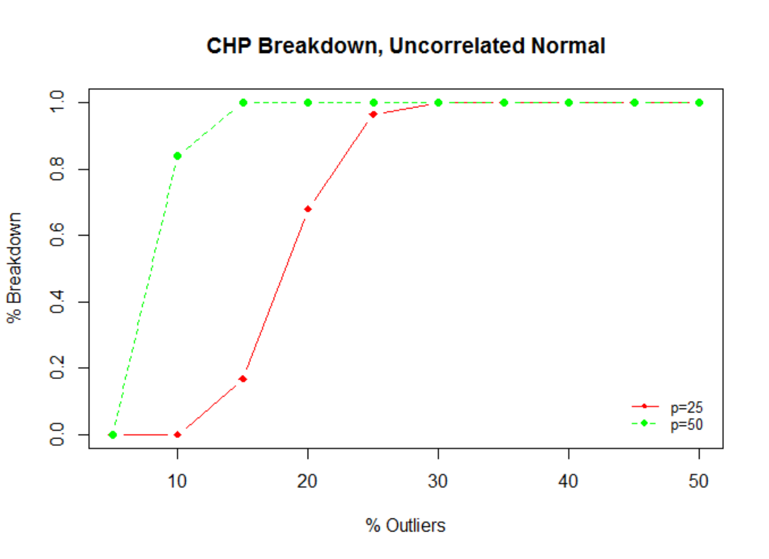
Filzmoser, P., Maronna, R., & Werner, M. (2008). Outlier identification in high dimensions. Computational Statistics Data Analysis, 52(3), 1694-1711.

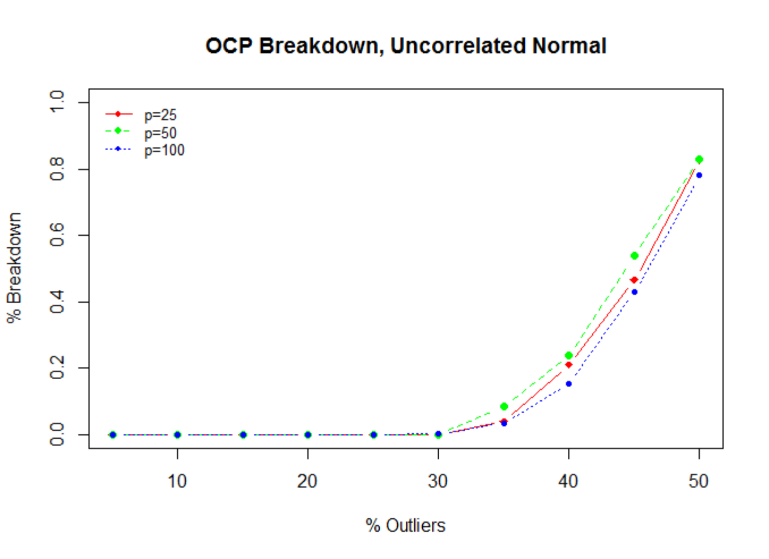
Ro, K., Zou, C., Wang, Z., & Yin, G. (2015). Outlier detection for high-dimensional data. Biometrika, 102(3), 589-599.

Weese, M. L., Martinez, W. G., & Jones Farmer, L. A. (2017). On the Selection of the Bandwidth Parameter for the k—Chart. Quality and Reliability Engineering International, 33(7), 1527-1547.

Weese, M., Martinez, W., Megahed, F. M., & Jones Farmer, L. A. (2016). Statistical learning methods applied to process monitoring: An overview and perspective. Journal of Quality Technology, 48(1), 4-24.

BACKUP


CHP Breakdown, Uncorrelated Normal
% Breakdown
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
% Outliers
p=25
p=50


OCP Breakdown, Uncorrelated Normal