

uRos (Use of R in Official Statistics)2018, 6<sup>th</sup> international conference  
Hague, Netherland, 12<sup>th</sup>-14<sup>th</sup>, Sep. 2018

# Comparison of multivariate outlier detection methods for nearly elliptically distributed data

Kazumi Wada, Mariko Kawano,  
Hiroe Tsubaki

National Statistics Center (NSTAC), Japan

# OUTLINE

1. Objective
2. Compared methods
  - 2.1 MSD (Modified Stahel-Donoho) estimators
  - 2.2 BACON (Blocked adaptive computationally efficient outlier nominators)
  - 2.3 Fast-MCD (Minimum covariance determinant) estimator
  - 2.4 NNVE (Nearest-neighbour variance estimator)
3. Monte Carlo simulation with random datasets
  - 3.1 Random datasets
  - 3.3 Results with random datasets
4. Application to a real survey data
  - 4.1 Unincorporated Enterprise Survey
  - 4.2 A few things about data transformation
  - 4.3 Results
5. Conclusion and further work

# 1. Objective

Select a suitable multivariate outlier detection method for cleaning donor data before imputation step

- All the four methods compared estimate robust mean vector and covariance matrix.
- There is no ultimate method. The best outlier detection methods depend on the situation.
- Targeted data distribution may not be symmetry even after transformation and may have a long tail.
- Evaluation is made with skewed and asymmetrically contaminated random datasets with long tails.

## 2. Compared methods

MSD estimators  
BACON  
Fast-MCD estimator  
NNVE

- All the four methods estimate robust mean vector and covariance matrix
- They are known about their good performance about outlier detection
- They also have good features such as affine equivariance, orthogonal equivariance, asymptotic normality and so on, in addition to high breakdown point

## 2.1 Modified Stahel-Donoho (MSD) estimators

- MSD is practically used for survey data editing in Statistics Canada.
- The practice of Statistics Canada is introduced in the EUREDIT project report, and a few improvements of the method are proposed.

*The MSD estimators perform well especially when variables are highly correlated.*

# EUREDIT Project

Goal: To establish evaluation criteria and best practice methods for editing and imputation for National Statistical Offices (NSOs).

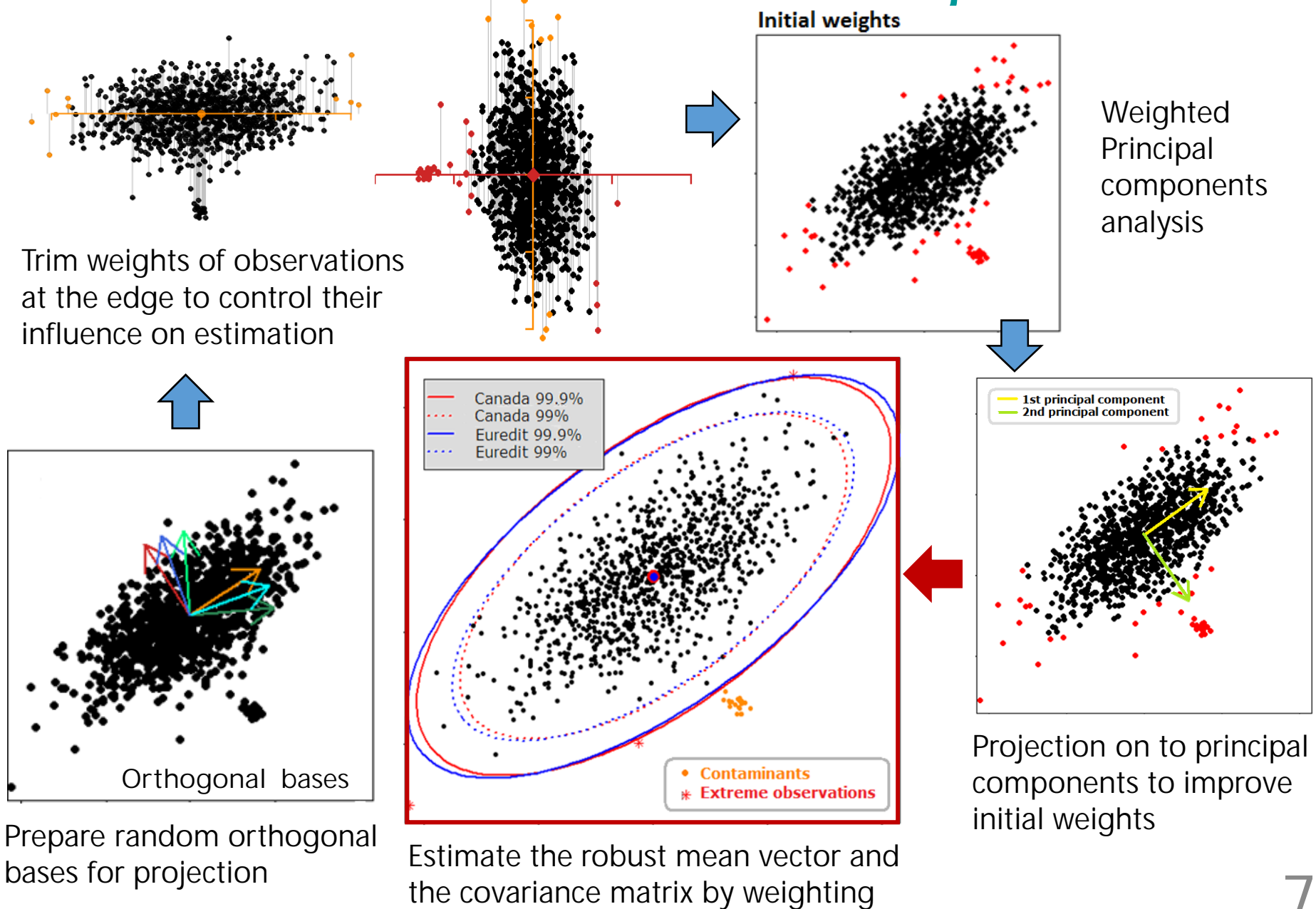
Participants: Statisticians and Researchers of NSOs, universities and the private sector within the region.

Financial support: EU

Project Term: 1 Mar. 2001 – 28 Feb. 2003

HP: <http://www.cs.york.ac.uk/euredit/>

# Basic idea of MSD: Projection



# R function used

MSD (Modified Stahel-Donoho) estimators

- Franklin & Brodeur (1997) describes the application in Statistics Canada


[http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1997\\_029.pdf](http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1997_029.pdf)

- Béguin and Hulliger (2003) suggests a few improvements
- Wada (2010) implemented both and published a R function

<https://github.com/kazwd2008/MSD>

- Wada & Tsubaki (2013) made the R function parallelised and published

<https://github.com/kazwd2008/MSD.parallel/>



The function *msd* based on Béguin and Hulliger (2003) is used in this work.



## 2.2 Blocked adaptive computationally efficient outlier nominators (BACON)

Choose a small initial subset without outliers, obtain mean vector and covariance matrix and calculate Mahalanobis distances based on them for each observation

Add the surrounding observations into the subset using chi-square test

Repeat and until no data added to the subset


*BACON is computationally the most efficient among the four methods examined in this study*

# R function used

BACON (Blocked adaptive computationally efficient outlier nominators)

- Billor et al. (2000) proposed the method and the algorithm
- Béguin and Hulliger (2003) implemented and published S-plus function  
<https://www.cs.york.ac.uk/euredit/results/Results/Robust/Part%20C.zip>
- Wada & Tsubaki (2013) published how to modify the S-plus function to R function.

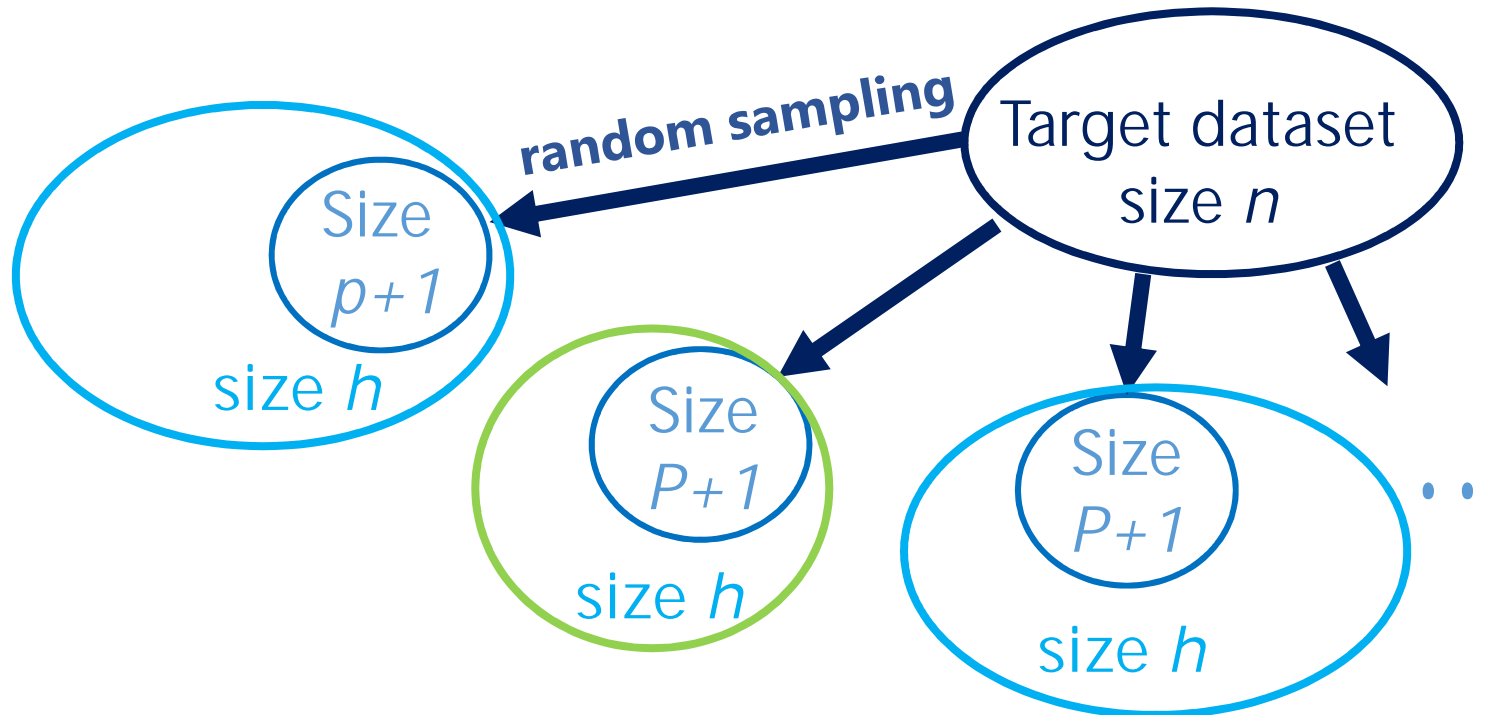
<https://github.com/kazwd2008/BEM/>



The ported function *BEM* based on Béguin and Hulliger (2003) is used in this work. The initial subset of size  $3 \times p$  is selected by version 2.

It seems mvBACON [robustX] or BEM [modi] behaves differently...

## 2.3 Fast-MCD estimator



Select size  $(p+1)$  of initial subsets randomly

Obtain mean vector and covariance matrix of each subset, and calculate the Mahalanobis distances of the whole datasets.

Sort the whole dataset in ascending order, and choose size  $h$  of data from the top.

Select the ones with the smaller determinant among the subsets with size  $h$ .

# R function used

Fast-MCD (Minimum covariance determinant) estimator

- MCD method is proposed by Rousseeuw (1984). It is computationally expensive, and its application is limited to small datasets.
- Rousseeuw & Driessen (1999) proposed new fast algorithm which applies to larger datasets
  - ✓ C-step
  - ✓ Selective iteration
  - ✓ Nested extensions
- Pison et al. (2002) proposed the finite sample correction step

*covMcd* function in *rrcov* package

It performs slightly better than `cov.mcd` [MASS].

## 2.4 Nearest-neighbour variance estimator (NNVE)

- NNVE assumes data consists of two gamma distributions. One is of correct data and the other, outliers.
- EM algorithm is used to estimate the mean vectors, covariance matrices, and the proportion of their mixture.
- NNVE is scale equivariant, but not affine equivariant
- Theoretically, NNVE bears outliers exceeding 50%; however, it is not good at detecting outliers with smaller variance than the correct data

## R function used

NNVE (Nearest-neighbour variance estimation)

- Bayes and Raftery (1998) introduced NNC (nearest-neighbour cleaning), which regards the data as a mixture of two different gamma distributions. It outperformed MVE estimator.
- Wang and Raftery (2002) proposed NNVE based on NNC by adding some artificial outliers to overcome the underestimation of covariance when there is no outlier.



`cov.nnve` function in *covRobust* package

# 3. Monte Carlo simulation

## 3.1 Random datasets

The original model : Peña and Prieto (2001)

$$(1 - \alpha)N_p(0, \mathbf{R}) + \alpha N_p(\delta \mathbf{e}_1, \lambda \mathbf{I})$$

$N_p$  :  $p$ -dimensional normal distribution

$\alpha$  : rate of outliers

$p$  : number of variables

$\mathbf{R}$  : correlation matrix

$\delta$  : distance between the normal data and the outliers

$\mathbf{e}_1$  : first unit vector

$\lambda$  : variance of the outliers.

The datasets consist of random variables following a multivariate normal distribution with asymmetric contamination.

It is known that many outlier detection methods have difficulty to cope with this model.

# Modification by Wada (2004, 2010)

Peña and Prieto (2001) :  $(1 - \alpha)N_p(0, \mathbf{R}) + \alpha N_p(\delta \mathbf{e}_1, \lambda \mathbf{I})$



$$(1 - \alpha)ST_p(0, \mathbf{R}, \eta \mathbf{e}_1, Df) + \alpha N_p(\delta \mathbf{e}_1, \lambda \mathbf{I})$$

$N_p$  :  $p$ -dimensional normal distribution

$ST_p$  :  $p$ -dimensional skew-t distribution

$\alpha$  : rate of outliers

$p$  : number of variables

$\mathbf{R}$  : correlation matrix

$\eta$  : skewness of the first axis

$Df$  : number of degrees of freedom

$\delta$  : distance between the normal data and the outliers

$\mathbf{e}_1$  : first unit vector

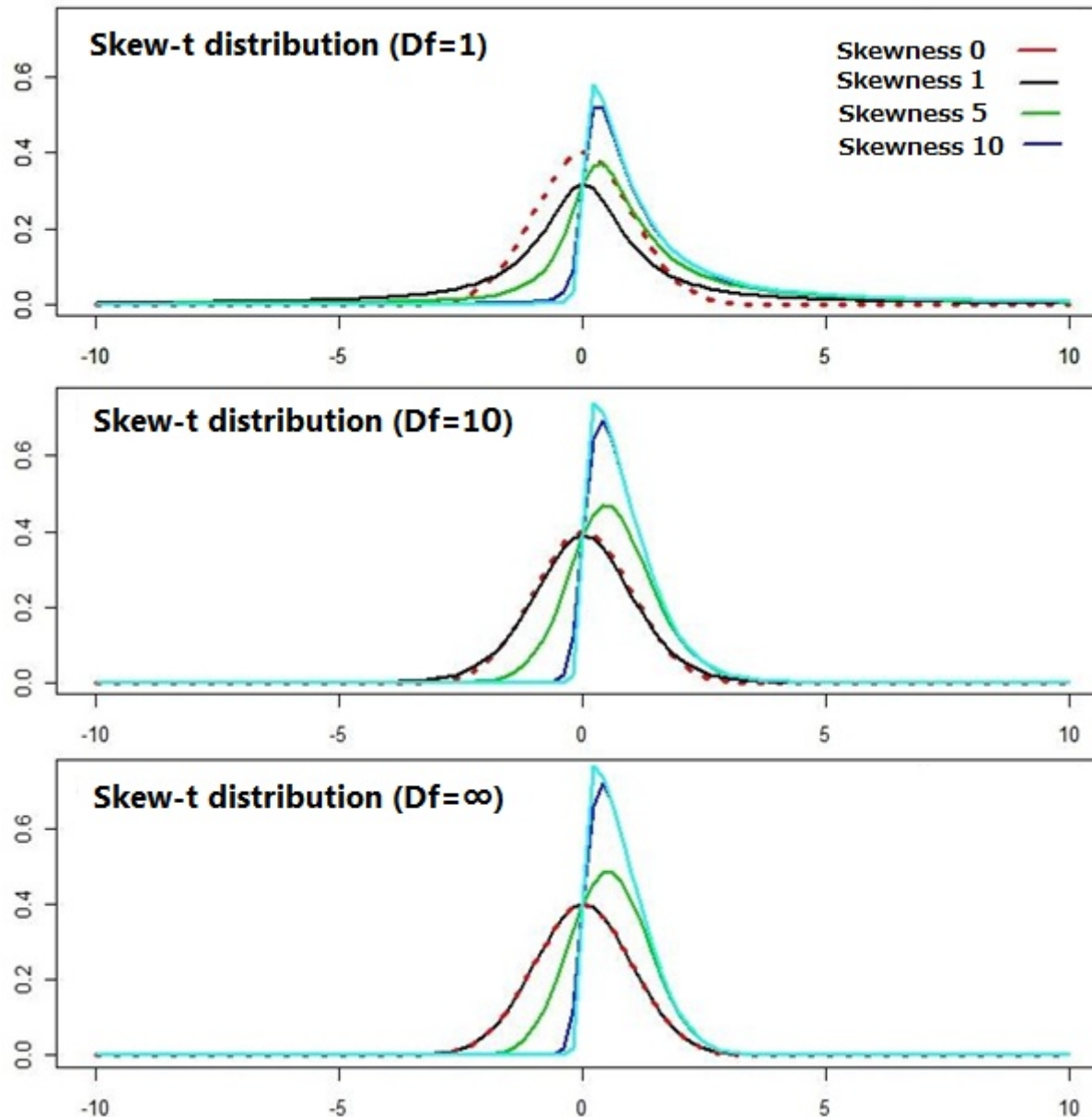
$\lambda$  : variance of the outliers.



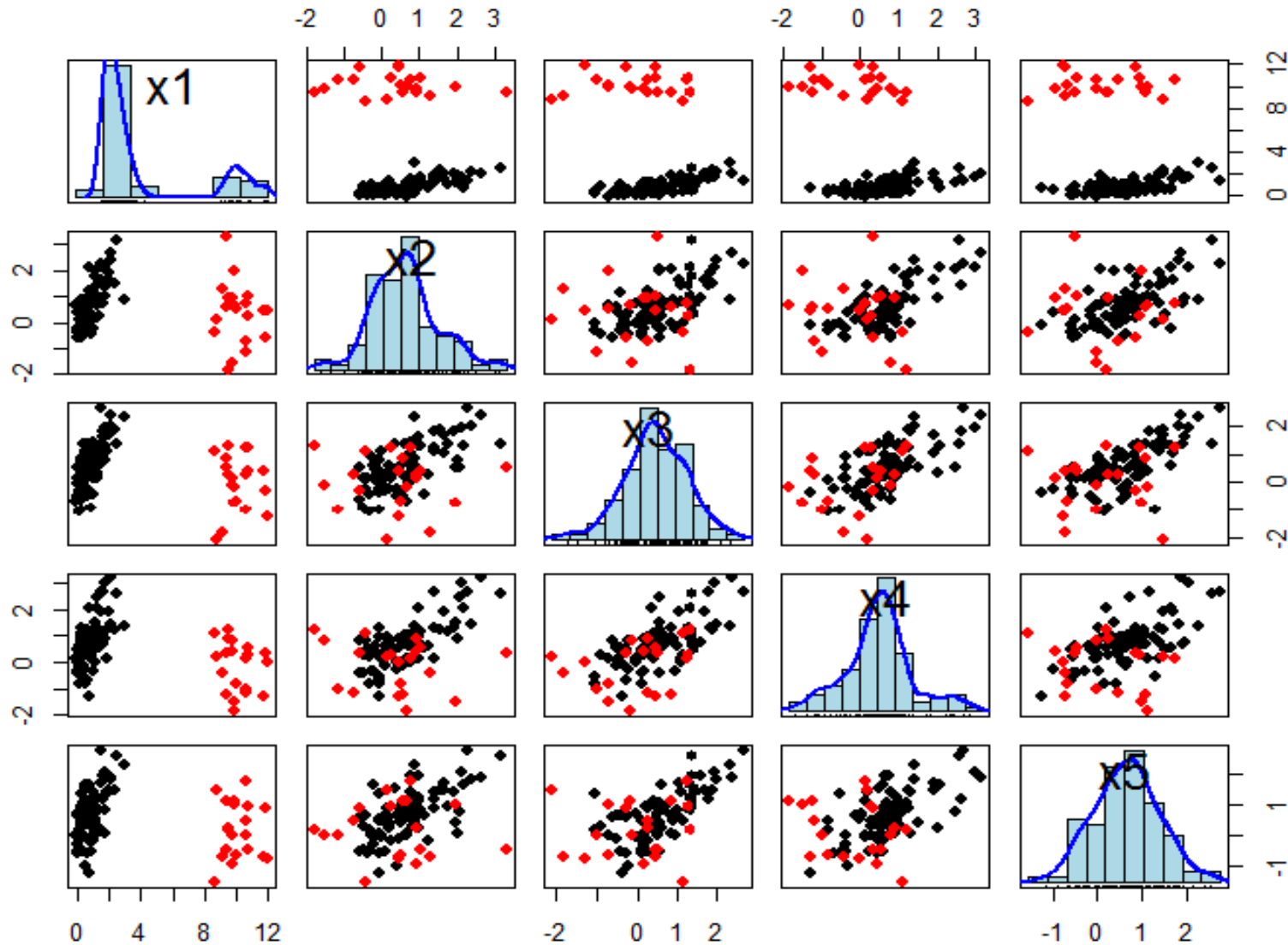
# Settings for the contaminated skew-t datasets

Parameter	Explanation	Values in the simulations
$\alpha$	Rate of outliers	0, 0.1, 0.2, 0.3, 0.4
$r$	Correlation between variables	0.4, 0.8
$p$	Number of variables	10
$\delta$	Distance between the normal data and the outliers	10, 100
$\lambda$	Variance of the outliers	1, 5
$\eta$	Skewness of the first axis	0, 5, 10
$Df$	Degree of freedom	2, 10, $\infty$

# Probability density of Skew-t distribution and Normal distribution



# Example of a random dataset following skew-t distribution



$\alpha = 0.2$   
 $r = 0.8$   
 $\delta = 10$   
 $\sqrt{\lambda} = 1$   
 $\eta = 10$   
 $Df = 10$

## 3.2 Results with random datasets

- As an overall tendency, MSD appears to be better than BACON, then Fast-MCD and NNVE follows
- BACON could be better when there are a large amount of outliers
- A decrease in the degree of freedom strongly affects all the methods, while an increase in skewness does not
- BACON is the most affected by the degree of freedom and NNVE is the least
- All the methods were applied with their default settings

MSD seems the most promising and be mainly evaluated in the next section

## 4. Application to a real survey data

- Target survey is the Unincorporated Enterprise Survey in Japan.
- Major change is planned from 2019, and imputation step is introduced accordingly
- Multivariate outlier detection is examined to introduce for donor data cleaning
- Ratio hot deck imputation is a candidate, since there are edit constraints among the imputed variables, and no data before the survey to examine model-based imputation methods

# 4.1 Unincorporated Enterprise Survey

- Covering 4,000 establishments engaged in manufacturing, wholesale and retail trade, accommodations and food services or providing services in Japan
- Questionnaires are distributed and collected by statistical enumerators
- The response rate is almost 100%
- Quarterly trend survey plus Annual structural survey



Till year 2018

From year 2019

- About 40,000 establishments engaged in almost all industries in Japan
- Mail survey
- Annual survey

➔ The response rate is expected to drop  
Necessity of imputation

# Target variables

No.	Variables
05	Sales
06	Total expenses
07	Beginning inventory (Inventory as of last December 31)
08	Purchases
09	Ending inventory (Inventory as of last December 31 before last)
10	Total of operating expenses

$$\text{Constraint: } 06 + 09 = 07 + 08 + 10$$

- Manufacturing industry

	Min.	Q1	Median	Mean	Q3	Max.
05	185	5,262	11,364	20,353	22,467	761,461
06	67	3,452	7,930	17,428	18,886	760,180
07	1	100	305	1,875	1,022	134,000
08	5	958	2,911	8,185	6,041	498,602
09	1	100	326	1,875	1,032	140,100
10	50	1,930	4,623	9,243	11,261	261,578

These variables are skewed and have long right tails. So data transformation is necessary to apply outlier detection methods.

# Manufacturing industry

- Correlation coefficient

	Pearson						Spearman					
	05	06	07	08	09	10	05	06	07	08	09	10
05	1.00	0.99	0.79	0.98	0.78	0.94	1.00	0.96	0.44	0.83	0.43	0.90
06	0.99	1.00	0.80	0.98	0.78	0.95	0.83	0.85	0.49	1.00	0.48	0.68
07	0.79	0.80	1.00	0.81	1.00	0.75	0.44	0.47	1.00	0.49	0.95	0.41
08	0.98	0.98	0.81	1.00	0.79	0.88	0.43	0.45	0.95	0.48	1.00	0.39
09	0.78	0.78	1.00	0.79	1.00	0.73	0.90	0.95	0.41	0.68	0.39	1.00
10	0.94	0.95	0.75	0.88	0.73	1.00	0.96	1.00	0.47	0.85	0.45	0.95

*Relatively high correlations are observed. Probably there are some extremely large values which rise Pearson's coefficients.*

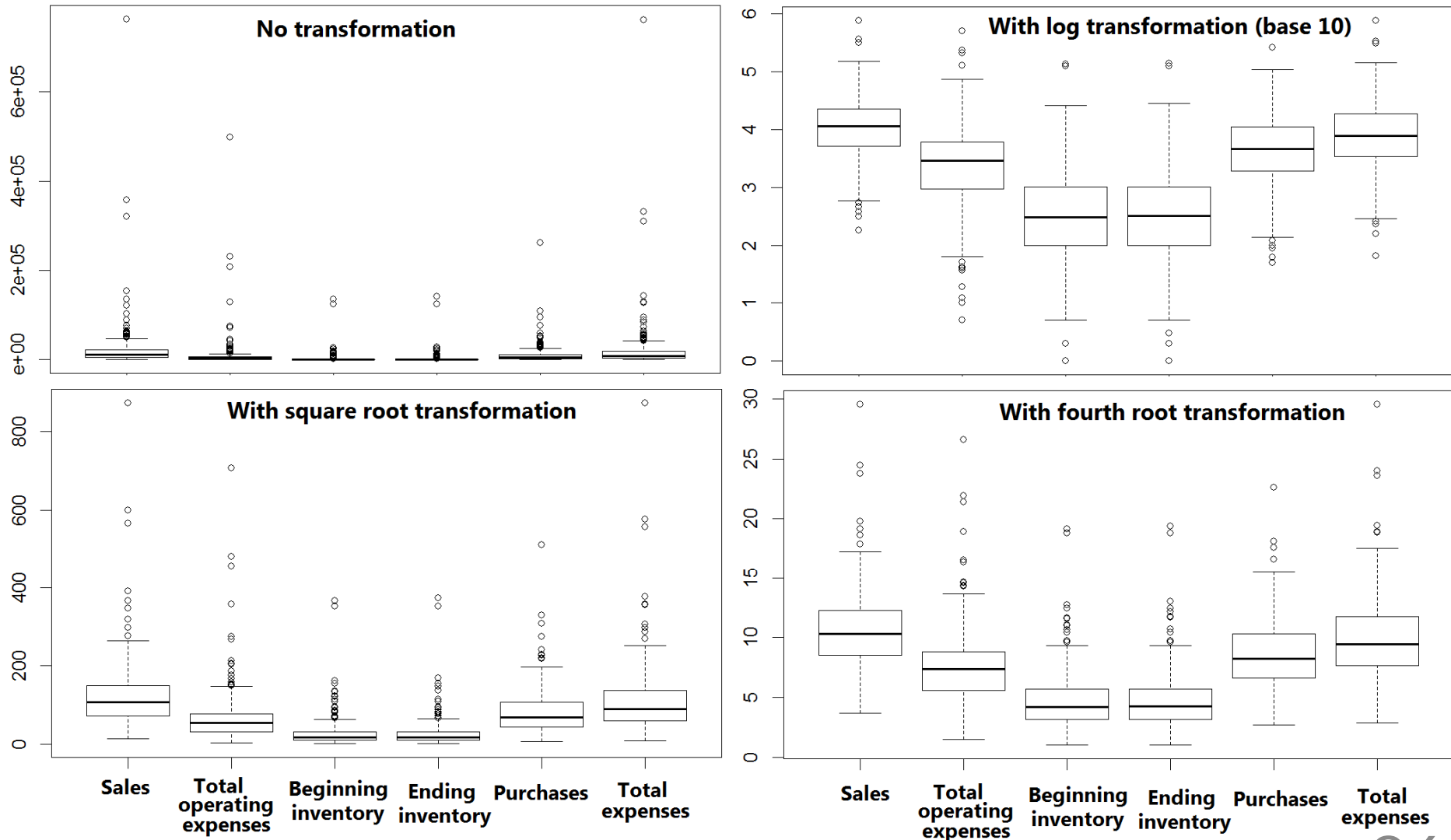


## 4.2 A few things about data transformation

- Different transformation among target variables alter the relation between variables
- Data transformation is not desirable for estimation such as population mean and total
- When it is unavoidable, modest transformation is better
- Box-Cox transformation is not outlier robust

# Box plots of various transformation

## Manufacturing industry



# Lambda of Box-Cox transformation

Variables	05	06	07	08	09	10
Lambda	0.321	0.336	0.152	0.151	0.246	0.317

As this method is not outlier robust, these lambda values could be smaller than the appropriate ones.

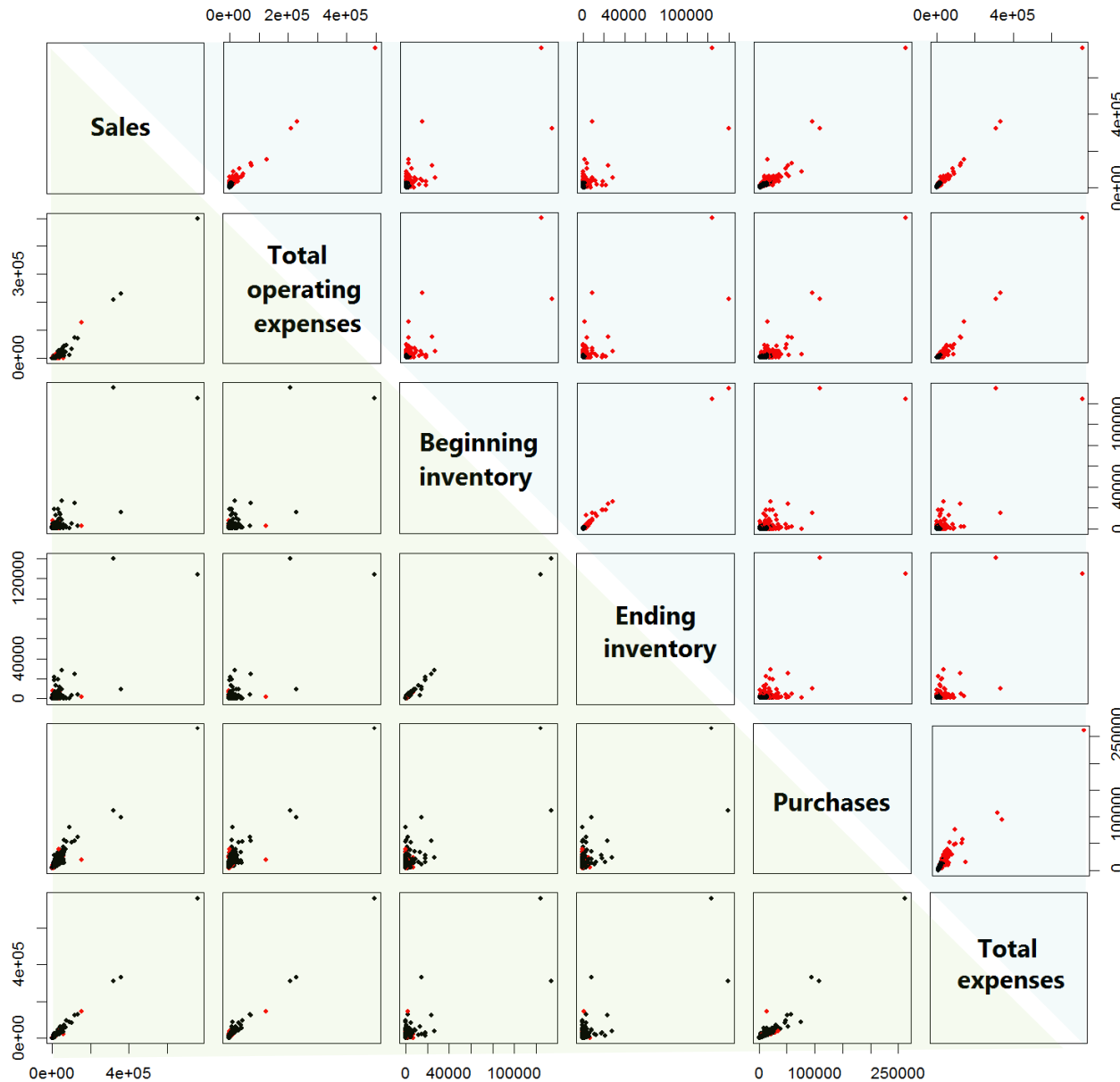
## Number of detected outliers

Industry	Transformation for outlier detection	MSD		BEM	
		No.	%	No.	%
Manufacturing	Square root	47	12.05%	63	16.15%
	Biquadratic root	28	7.18%	49	12.23%
	Log (base 10)	41	10.51%	53	13.59%

The smallest number of outliers indicates the most appropriate transformation.

# 4.3 Results: Detected outliers by MSD(1)

## Manufacturing industry



Outlier detection:

Upper triangular matrix:  
without transformation

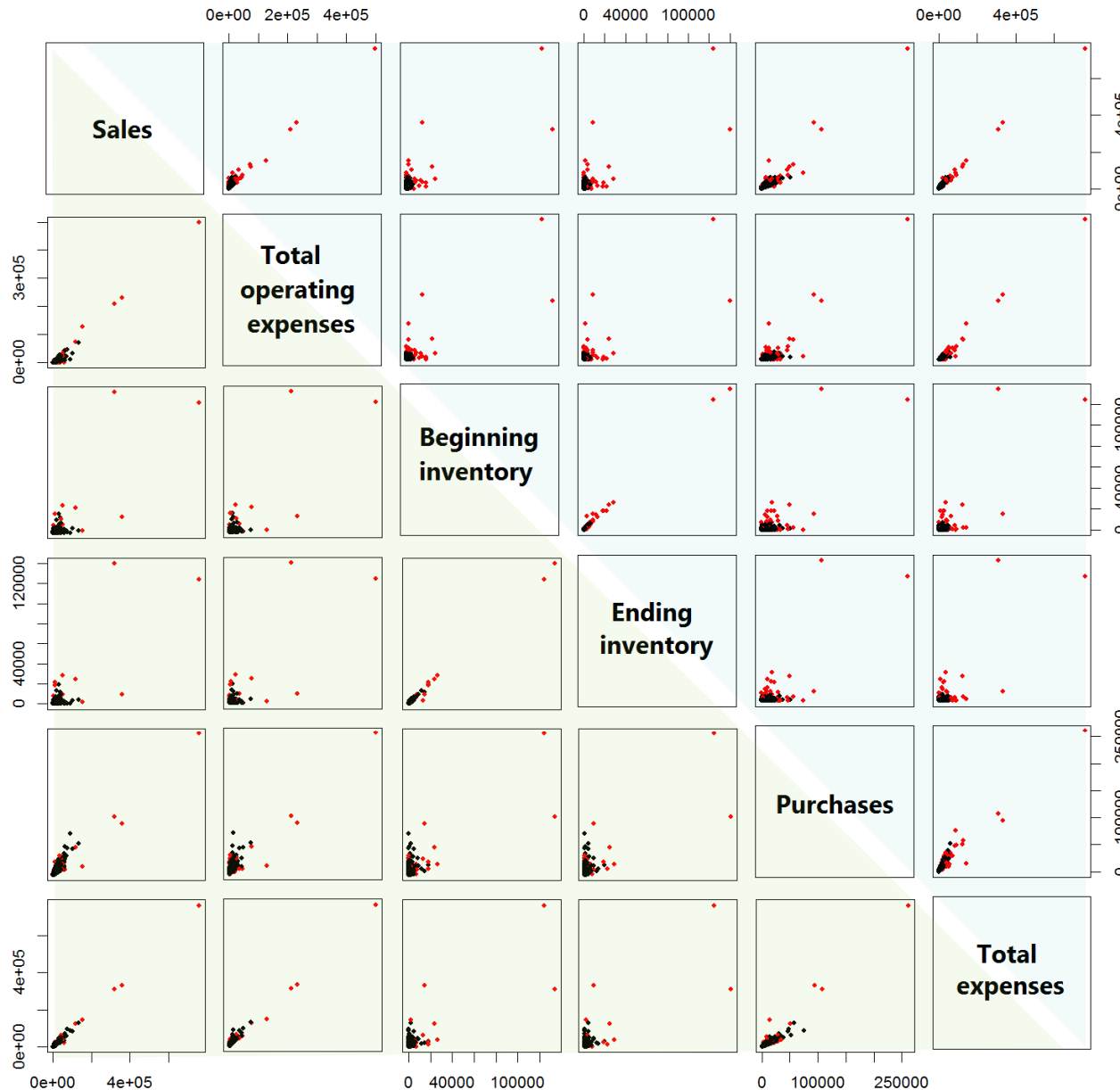
Lower triangular matrix:  
with base 10 log  
transformation

Visualisation:

No transformation

# Detected outliers by MSD(2)

## Manufacturing industry



Outlier detection:

Upper triangular matrix is with the square root transformation

Lower triangular matrix is with fourth root transformation

Visualisation:

No transformation

# Sum of the absolute deviation between true and imputed values

Variable	No cleaning	After outlier removal	Reduced rate
06	173,794	172,275	99%
07	101,144	90,082	89%
08	103,717	93,627	90%
09	191,600	183,674	96%
10	164,239	160,500	98%

Improved up to 10%



# Processing time with a larger dataset

- Entire survey data from 2002 to 2017
- 44,537 observations with 6 variables

Method	Processing time
MSD	45 seconds
BACON	5 seconds

MSD is much slower than BACON. However, 45 seconds for each imputation class is acceptable for this survey, since the number of imputation class is less than 100, at most.

## 5. Conclusion and future work

- We examined four multivariate outlier detection methods with attractive features and found MSD exhibits relatively good performance with skewed and heavy tailed datasets.
- A few issues still remain for the practical implementation, such as adjustment of thresholds used to determine outliers and deciding an appropriate size of imputation class. Further work is necessary.