Karlsruhe Institute of Technology

Monte Carlo Dependency Estimation

31st International Conference on Scientific and Statistical Database Management (SSDBM '19)

Edouard Fouché & Klemens Böhm | July 23, 2019

INSTITUTE FOR PROGRAM STRUCTURES AND DATA ORGANIZATION (IPD), CHAIR PROF. BÖHM

Motivation

Estimating dependency is fundamental in Data Mining, e.g.:

- **“Feature Selection”**: Find good predictors (classification accuracy)
- **“Subspace Search”**: Find relevant projections (outliers, clusters)

Real-world data often comes as a (high-dimensional) stream

- Potentially unbounded, ever evolving
- Generated at varying speed
- Noisy, redundant

In streams, the timely detection of changes is crucial

- e.g., “Predictive Maintenance”

Motivation

Estimating dependency is fundamental in Data Mining, e.g.:

- **“Feature Selection”**: Find good predictors (classification accuracy)
- **“Subspace Search”**: Find relevant projections (outliers, clusters)

Real-world data often comes as a (high-dimensional) stream

- Potentially unbounded, ever evolving
- Generated at varying speed
- Noisy, redundant

In streams, the timely detection of changes is crucial

- e.g., “Predictive Maintenance”

Motivation

Estimating dependency is fundamental in Data Mining, e.g.:

- **"Feature Selection"**: Find good predictors (classification accuracy)
- **"Subspace Search"**: Find relevant projections (outliers, clusters)

Real-world data often comes as a (high-dimensional) stream

- Potentially unbounded, ever evolving
- Generated at varying speed
- Noisy, redundant

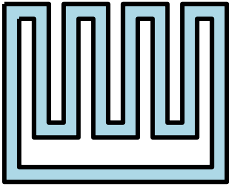
In streams, the timely detection of changes is crucial

- e.g., "Predictive Maintenance"

Example: Dependency Monitoring

Dependency often results from natural relationships

- E.g., water in a cooling system
- Pressure P and Temperature T



$$P \propto T$$

Dependency changes within (T, P) either mean:

- That the state of the system has changed
- That equipment deteriorates, e.g., leaks

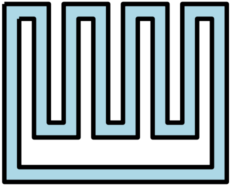
$\rightarrow$ Helpful to detect outliers / abnormal behaviours !

Our goal: Propose an estimator suitable for streams

Example: Dependency Monitoring

Dependency often results from natural relationships

- E.g., water in a cooling system
- Pressure P and Temperature T



$$P \propto T$$

Dependency changes within (T, P) either mean:

- That the state of the system has changed
- That equipment deteriorates, e.g., leaks

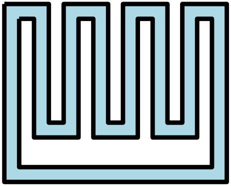
$\rightarrow$ Helpful to detect outliers / abnormal behaviours !

Our goal: Propose an estimator suitable for streams

Example: Dependency Monitoring

Dependency often results from natural relationships

- E.g., water in a cooling system
- Pressure P and Temperature T



$$P \propto T$$

Dependency changes within (T, P) either mean:

- That the state of the system has changed
- That equipment deteriorates, e.g., leaks

$\rightarrow$ Helpful to detect outliers / abnormal behaviours !

Our goal: Propose an estimator suitable for streams

Example: Dependency Monitoring

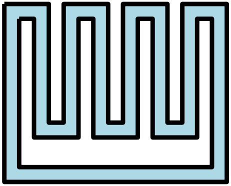
Dependency often results from natural relationships

- E.g., water in a cooling system
- Pressure P and Temperature T

Dependency changes within (T, P) either mean:

- That the state of the system has changed
- That equipment deteriorates, e.g., leaks

$\rightarrow$ Helpful to detect outliers / abnormal behaviours !



$$P \propto T$$

Our goal: Propose an estimator suitable for streams

Requirements

Karlsruhe Institute of Technology

Stream-related

- (R1) Multivariate: 2+ variables
- (R2) Efficient: Linear complexity (in the worst case)
- (R3) Anytime: Interruption $\rightarrow$ approximate results

General

- (R4) General-purpose, e.g., not only linear dependencies
- (R5) Intuitive
  - Parameters are easy to set
  - Results must be easy to interpret
- (R6) Robust: Handle duplicates / imprecisions / noise

$\rightarrow$ No existing approach fulfil them all !

Requirements

Stream-related

- (R1) Multivariate: 2+ variables
- (R2) Efficient: Linear complexity (in the worst case)
- (R3) Anytime: Interruption $\rightarrow$ approximate results

General

- (R4) General-purpose, e.g., not only linear dependencies
- (R5) Intuitive
  - Parameters are easy to set
  - Results must be easy to interpret
- (R6) Robust: Handle duplicates / imprecisions / noise

$\rightarrow$ No existing approach fulfil them all !

Requirements

Stream-related

- (R1) Multivariate: 2+ variables
- (R2) Efficient: Linear complexity (in the worst case)
- (R3) Anytime: Interruption $\rightarrow$ approximate results

General

- (R4) General-purpose, e.g., not only linear dependencies
- (R5) Intuitive
  - Parameters are easy to set
  - Results must be easy to interpret
- (R6) Robust: Handle duplicates / imprecisions / noise

$\rightarrow$ No existing approach fulfil them all !

Related Work

- Bivariate measures (↯R1): Pearson, Spearman, MI, . . .
- Multivariate Spearman (Schmid and Schmidt, 2007)
  - Limited to monotonous relationships (↯R4)
- Multivariate variants of Mutual Information (↯R2,R5):
  - Interaction Information (II) (McGill, 1954)
  - Total Correlation (TC) (Watanabe, 1960)
- Cumulative Mutual Information (CMI)
  Multivariate Maximal Correlation (MAC)
  Universal Dependency Score (UDS)
  - (Nguyen et al., 2013; 2014; 2015) (↯R2,R5)
- High-Contrast Subspaces (HiCS) (Keller et al., 2012) (↯R5,R6)
  - Only used as "heuristic" to find outliers
  - (Keller, 2015) describes it as a potential dependency estimator

Requirements

- R1: Multivariate
- R2: Efficient
- R3: Anytime
- R4: General-purpose
- R5: Intuitive
- R6: Robust

Related Work

- Bivariate measures (↯R1): Pearson, Spearman, MI, . . .
- Multivariate Spearman (Schmid and Schmidt, 2007)
  - Limited to monotonous relationships (↯R4)
- Multivariate variants of Mutual Information (↯R2,R5):
  - Interaction Information (II) (McGill, 1954)
  - Total Correlation (TC) (Watanabe, 1960)
- Cumulative Mutual Information (CMI)
  Multivariate Maximal Correlation (MAC)
  Universal Dependency Score (UDS)
  - (Nguyen et al., 2013; 2014; 2015) (↯R2,R5)
- High-Contrast Subspaces (HiCS) (Keller et al., 2012) (↯R5,R6)
  - Only used as "heuristic" to find outliers
  - (Keller, 2015) describes it as a potential dependency estimator

Requirements

- R1: Multivariate
- R2: Efficient
- R3: Anytime
- R4: General-purpose
- R5: Intuitive
- R6: Robust

Related Work

- Bivariate measures (↯R1): Pearson, Spearman, MI, . . .
- Multivariate Spearman (Schmid and Schmidt, 2007)
  - Limited to monotonous relationships (↯R4)
- Multivariate variants of Mutual Information (↯R2,R5):
  - Interaction Information (II) (McGill, 1954)
  - Total Correlation (TC) (Watanabe, 1960)
- Cumulative Mutual Information (CMI)
  Multivariate Maximal Correlation (MAC)
  Universal Dependency Score (UDS)
  - (Nguyen et al., 2013; 2014; 2015) (↯R2,R5)
- High-Contrast Subspaces (HiCS) (Keller et al., 2012) (↯R5,R6)
  - Only used as "heuristic" to find outliers
  - (Keller, 2015) describes it as a potential dependency estimator

Requirements

- R1: Multivariate
- R2: Efficient
- R3: Anytime
- R4: General-purpose
- R5: Intuitive
- R6: Robust

Related Work

- Bivariate measures (↯R1): Pearson, Spearman, MI, . . .
- Multivariate Spearman (Schmid and Schmidt, 2007)
  - Limited to monotonous relationships (↯R4)
- Multivariate variants of Mutual Information (↯R2,R5):
  - Interaction Information (II) (McGill, 1954)
  - Total Correlation (TC) (Watanabe, 1960)
- Cumulative Mutual Information (CMI)
  Multivariate Maximal Correlation (MAC)
  Universal Dependency Score (UDS)
  - (Nguyen et al., 2013; 2014; 2015) (↯R2,R5)
- High-Contrast Subspaces (HiCS) (Keller et al., 2012) (↯R5,R6)
  - Only used as "heuristic" to find outliers
  - (Keller, 2015) describes it as a potential dependency estimator

Requirements

- R1: Multivariate
- R2: Efficient
- R3: Anytime
- R4: General-purpose
- R5: Intuitive
- R6: Robust

Related Work

- Bivariate measures (↯R1): Pearson, Spearman, MI, ...
- Multivariate Spearman (Schmid and Schmidt, 2007)
  - Limited to monotonous relationships (↯R4)
- Multivariate variants of Mutual Information (↯R2,R5):
  - Interaction Information (II) (McGill, 1954)
  - Total Correlation (TC) (Watanabe, 1960)
- Cumulative Mutual Information (CMI)
  Multivariate Maximal Correlation (MAC)
  Universal Dependency Score (UDS)
  - (Nguyen et al., 2013; 2014; 2015) (↯R2,R5)
- High-Contrast Subspaces (HiCS) (Keller et al., 2012) (↯R5,R6)
  - Only used as "heuristic" to find outliers
  - (Keller, 2015) describes it as a potential dependency estimator

Requirements

- R1: Multivariate
- R2: Efficient
- R3: Anytime
- R4: General-purpose
- R5: Intuitive
- R6: Robust

Contributions

Monte Carlo Dependency Estimation (MCDE)

- A general framework for estimating dependency
- Estimate discrepancy between marginal/conditional distributions using statistical tests via Monte Carlo simulations

Mann-Whitney P (MWP)

- Instantiation of MCDE based on Mann-Whitney U
- Extensive evaluation against state-of-the-art methods

$\rightarrow$ Source code, data, experiments:
https://github.com/edouardfouche/MCDE

Contributions

Monte Carlo Dependency Estimation (MCDE)

- A general framework for estimating dependency
- Estimate discrepancy between marginal/conditional distributions using statistical tests via Monte Carlo simulations

Mann-Whitney P (MWP)

- Instantiation of MCDE based on Mann-Whitney U
- Extensive evaluation against state-of-the-art methods

→ Source code, data, experiments:
https://github.com/edouardfouche/MCDE

Contributions

Monte Carlo Dependency Estimation (MCDE)

- A general framework for estimating dependency
- Estimate discrepancy between marginal/conditional distributions using statistical tests via Monte Carlo simulations

Mann-Whitney P (MWP)

- Instantiation of MCDE based on Mann-Whitney U
- Extensive evaluation against state-of-the-art methods

$\rightarrow$ Source code, data, experiments:
https://github.com/edouardfouche/MCDE

Basic Definitions

- Let $S = \{X_1, \ldots, X_d\}$ be a set of d dimensions (subspace)
- We see each i dimension as a random variable X_i
- $p(S)$ is the joint distribution
- $p_{X_i}(S)$ is the marginal distribution of $X_i \in S$

Independence: S is independent, if and only if:

$$p(S) = \prod_{X_i \in S} p_{X_i}(S) \quad (1)$$

$$\Leftrightarrow p(S' | \overline{S'}) = p(S') \quad \forall S' \subset S \quad (2)$$

Basic Definitions

- Let $S = \{X_1, \ldots, X_d\}$ be a set of d dimensions (subspace)
- We see each i dimension as a random variable X_i
- $p(S)$ is the joint distribution
- $p_{X_i}(S)$ is the marginal distribution of $X_i \in S$

Independence: S is independent, if and only if:

$$p(S) = \prod_{X_i \in S} p_{X_i}(S) \quad (1)$$

$$\Leftrightarrow p(S' | \overline{S'}) = p(S') \quad \forall S' \subset S \quad (2)$$

Basic Definitions

- Let $S = \{X_1, \ldots, X_d\}$ be a set of d dimensions (subspace)
- We see each i dimension as a random variable X_i
- $p(S)$ is the joint distribution
- $p_{X_i}(S)$ is the marginal distribution of $X_i \in S$

Independence: S is independent, if and only if:

$$p(S) = \prod_{X_i \in S} p_{X_i}(S) \tag{1}$$

$$\Leftrightarrow p(S'|\overline{S'}) = p(S') \quad \forall S' \subset S \tag{2}$$

MCDE

Quantify (in)dependence by $p(S'|\overline{S'}) \stackrel{?}{=} p(S') \quad \forall S' \subset S$

Problem: Difficult

- $|S'| > 1$ requires multivariate density estimation
- $\{S' : S' \subset S\}$ grows exponentially with d

$\rightarrow$ we need to be very efficient in the streaming setting

Thus, we make the following relaxation,

$$p(S'|\overline{S'}) = p(S') \quad \forall S' \subset S \quad |S'| = 1 \tag{3}$$

$$\Leftrightarrow p(S|\overline{X_i}) = p_{X_i}(S) \quad \forall X_i \in S \tag{4}$$

$\rightarrow$ Our goal: $p(S|\overline{X_i}) \stackrel{?}{=} p_{X_i}(S) \quad \forall X_i \in S$

MCDE

Quantify (in)dependence by $p(S'|\overline{S'}) \stackrel{?}{=} p(S') \quad \forall S' \subset S$

Problem: Difficult

- $|S'| > 1$ requires multivariate density estimation
- $\{S' : S' \subset S\}$ grows exponentially with d

$\rightarrow$ we need to be very efficient in the streaming setting

Thus, we make the following relaxation,

$$p(S'|\overline{S'}) = p(S') \quad \forall S' \subset S \quad |S'| = 1 \tag{3}$$

$$\Leftrightarrow p(S|\overline{X_i}) = p_{X_i}(S) \quad \forall X_i \in S \tag{4}$$

$\rightarrow$ Our goal: $p(S|\overline{X_i}) \stackrel{?}{=} p_{X_i}(S) \quad \forall X_i \in S$

MCDE

Quantify (in)dependence by $p(S'|\overline{S'}) \stackrel{?}{=} p(S') \quad \forall S' \subset S$

Problem: Difficult

- $|S'| > 1$ requires multivariate density estimation
- $\{S' : S' \subset S\}$ grows exponentially with d

$\rightarrow$ we need to be very efficient in the streaming setting

Thus, we make the following relaxation,

$$p(S'|\overline{S'}) = p(S') \quad \forall S' \subset S \quad |S'| = 1 \tag{3}$$

$$\Leftrightarrow p(S|\overline{X_i}) = p_{X_i}(S) \quad \forall X_i \in S \tag{4}$$

$\rightarrow$ Our goal: $p(S|\overline{X_i}) \stackrel{?}{=} p_{X_i}(S) \quad \forall X_i \in S$

MCDE

Quantify (in)dependence by $p(S'|\overline{S'}) \stackrel{?}{=} p(S') \quad \forall S' \subset S$

Problem: Difficult

- $|S'| > 1$ requires multivariate density estimation
- $\{S' : S' \subset S\}$ grows exponentially with d

$\rightarrow$ we need to be very efficient in the streaming setting

Thus, we make the following relaxation,

$$p(S'|\overline{S'}) = p(S') \quad \forall S' \subset S \quad |S'| = 1 \tag{3}$$

$$\Leftrightarrow p(S|\overline{X_i}) = p_{X_i}(S) \quad \forall X_i \in S \tag{4}$$

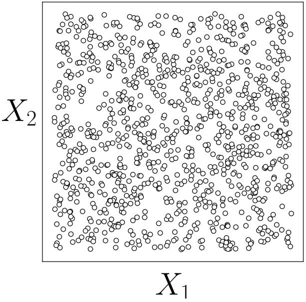
$\rightarrow$ Our goal: $p(S|\overline{X_i}) \stackrel{?}{=} p_{X_i}(S) \quad \forall X_i \in S$

MCDE

How to estimate "$p(S|\overline{X_i}) \stackrel{?}{=} p_{X_i}(S)$" ?

- (Example: reference $X_1 \Rightarrow \overline{X_i} \equiv X_2$)

1. Subspace Slice s_i
   - A set of conditions on $\overline{X_i}$
     - "dimensionality-aware" slicing
     - $s.t.\ \mathbb{E}[|s_i|] = \mathbb{E}[|\overline{s_i}|]$ under independence
2. Marginal Restriction r_i
   - Condition on X_i
     - Reduce computational burden
     - Better capture local effects
3. Statistical test $\mathcal{T}(\hat{p}(S|\{s_i, r_i\}), \hat{p}(S|\{\overline{s_i}, r_i\})) \rightarrow p$-value

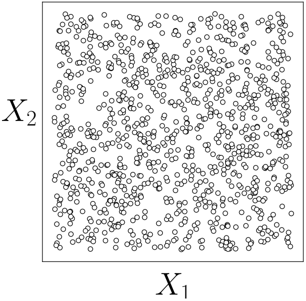


MCDE

How to estimate "$p(S|\overline{X_i}) \overset{?}{=} p_{X_i}(S)$" ?

- (Example: reference $X_1 \Rightarrow \overline{X_i} \equiv X_2$)

1. Subspace Slice s_i
   - A set of conditions on $\overline{X_i}$
     - "dimensionality-aware" slicing
     - $s.t.\ \mathbb{E}[|s_i|] = \mathbb{E}[|\overline{s_i}|]$ under independence
2. Marginal Restriction r_i
   - Condition on X_i
     - Reduce computational burden
     - Better capture local effects
3. Statistical test $\mathcal{T}(\hat{p}(S|\{s_i, r_i\}), \hat{p}(S|\{\overline{s_i}, r_i\})) \rightarrow p$-value

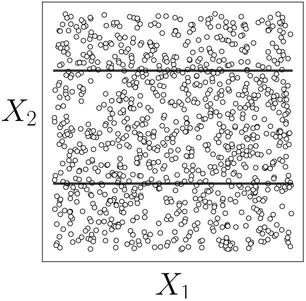


MCDE

How to estimate "$p(S|\overline{X_i}) \stackrel{?}{=} p_{X_i}(S)$" ?

- (Example: reference $X_1 \Rightarrow \overline{X_i} \equiv X_2$)

1. Subspace Slice s_i
   - A set of conditions on $\overline{X_i}$
     - "dimensionality-aware" slicing
     - $s.t.\ \mathbb{E}[|s_i|] \equiv \mathbb{E}[|\overline{s_i}|]$ under independence
2. Marginal Restriction r_i
   - Condition on X_i
     - Reduce computational burden
     - Better capture local effects
3. Statistical test $\mathcal{T}(\hat{p}(S|\{s_i, r_i\}), \hat{p}(S|\{\overline{s_i}, r_i\})) \rightarrow p$-value

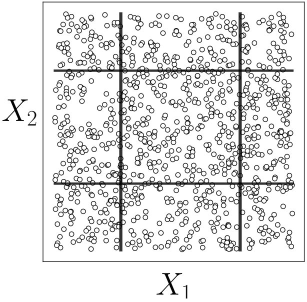


MCDE

How to estimate "$p(S|\overline{X_i}) \overset{?}{=} p_{X_i}(S)$" ?

- (Example: reference $X_1 \Rightarrow \overline{X_i} \equiv X_2$)

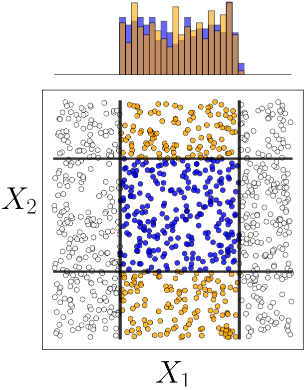
1. Subspace Slice s_i
   - A set of conditions on $\overline{X_i}$
     - "dimensionality-aware" slicing
     - $s.t.\ \mathbb{E}[|s_i|] \equiv \mathbb{E}[|\overline{s_i}|]$ under independence
2. Marginal Restriction r_i
   - Condition on X_i
     - Reduce computational burden
     - Better capture local effects
3. Statistical test $\mathcal{T}(\hat{p}(S|\{s_i, r_i\}), \hat{p}(S|\{\overline{s_i}, r_i\})) \rightarrow p$-value



MCDE

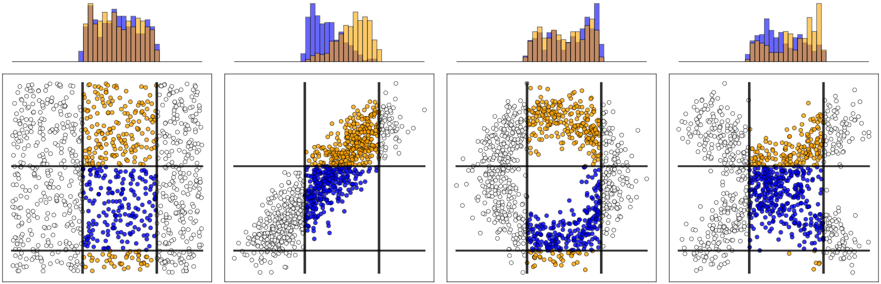
How to estimate "$p(S|\overline{X_i}) \overset{?}{=} p_{X_i}(S)$" ?

- (Example: reference $X_1 \Rightarrow \overline{X_i} \equiv X_2$)



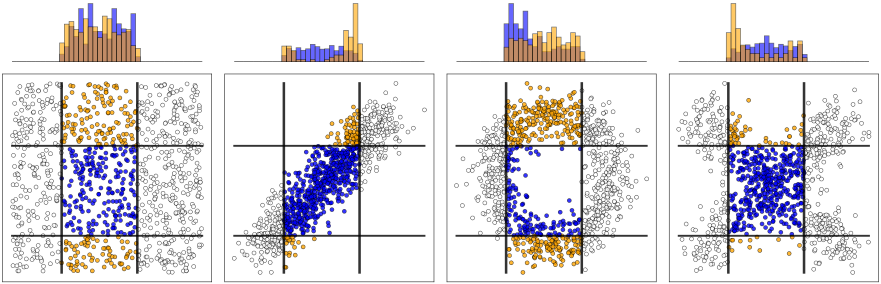
1. Subspace Slice s_i
   - A set of conditions on $\overline{X_i}$
     - "dimensionality-aware" slicing
     - $s.t.\ \mathbb{E}[|s_i|] \equiv \mathbb{E}[|\overline{s_i}|]$ under independence
2. Marginal Restriction r_i
   - Condition on X_i
     - Reduce computational burden
     - Better capture local effects
3. Statistical test $\mathcal{T}(\hat{p}(S|\{s_i, r_i\}), \hat{p}(S|\{\overline{s_i}, r_i\})) \to p$-value

MCDE – Illustration



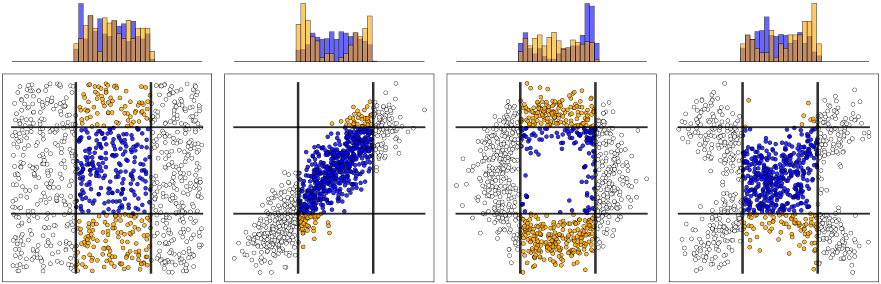
(a) Independent (b) Linear (c) Circle (d) Cross

MCDE – Illustration



(a) Independent (b) Linear (c) Circle (d) Cross

MCDE – Illustration



(a) Independent (b) Linear (c) Circle (d) Cross

MCDE

Karlsruhe Institute of Technology

Repeat M times, choosing reference X_i, slice s_i, restriction r_i randomly.

Contrast: $$\mathcal{C}(S) \equiv \frac{1}{M} \sum_{m=1}^{M} \left[1 - \mathcal{T}\left(\hat{p}(S|\{s_i, r_i\}), \hat{p}(S|\{\overline{s_i}, r_i\})\right)\right] \tag{5}$$

Properties:

- $\mathcal{C}(S) \in [0, 1]$
- Under independence, $\mathbb{E}[\mathcal{C}(S)] = 0.5$
- $\mathcal{C}(S)$ converges to 1 as evidence against independence increases

Anytime flexibility: $$\Pr\left(|\mathcal{C}(S) - \mathbb{E}[\mathcal{C}(S)]| \geq \varepsilon\right) \leq 2e^{-2M\varepsilon^2} \tag{6}$$

(Derived from Hoeffding's inequality (Hoeffding, 1963))

MCDE

Repeat M times, choosing reference X_i, slice s_i, restriction r_i randomly.

$$\textbf{Contrast:} \quad \mathcal{C}(S) \equiv \frac{1}{M} \sum_{m=1}^{M} \left[1 - \mathcal{T}\left(\hat{p}(S|\{s_i, r_i\}), \hat{p}(S|\{\overline{s_i}, r_i\})\right)\right] \tag{5}$$

Properties:

- $\mathcal{C}(S) \in [0, 1]$
- Under independence, $\mathbb{E}[\mathcal{C}(S)] = 0.5$
- $\mathcal{C}(S)$ converges to 1 as evidence against independence increases

$$\textbf{Anytime flexibility:} \quad \Pr\left(|\mathcal{C}(S) - \mathbb{E}[\mathcal{C}(S)]| \geq \varepsilon\right) \leq 2e^{-2M\varepsilon^2} \tag{6}$$

(Derived from Hoeffding's inequality (Hoeffding, 1963))

MCDE

Karlsruhe Institute of Technology

Repeat M times, choosing reference X_i, slice s_i, restriction r_i randomly.

$$\textbf{Contrast:} \quad \mathcal{C}(S) \equiv \frac{1}{M} \sum_{m=1}^{M} \left[1 - \mathcal{T}\left(\hat{p}(S|\{s_i, r_i\}), \hat{p}(S|\{\overline{s_i}, r_i\})\right)\right] \tag{5}$$

Properties:

- $\mathcal{C}(S) \in [0, 1]$
- Under independence, $\mathbb{E}[\mathcal{C}(S)] = 0.5$
- $\mathcal{C}(S)$ converges to 1 as evidence against independence increases

$$\textbf{Anytime flexibility:} \quad \Pr\left(|\mathcal{C}(S) - \mathbb{E}[\mathcal{C}(S)]| \geq \varepsilon\right) \leq 2e^{-2M\varepsilon^2} \tag{6}$$

(Derived from Hoeffding's inequality (Hoeffding, 1963))

Mann-Whitney P (MWP)

MCDE, where $\mathcal{T}$ is a two-sided Mann-Whitney U test

- Non-parametric (R4)
- Operates on ranks (ordinal data) $\rightarrow$ Robust (R6)

$\rightarrow$ Requires indexing

MWP (S)

1: $\mathcal{I} \leftarrow$ CONSTRUCTINDEX(S) $\triangleright$ $O(d \cdot n \cdot log(n))$
2: **for** $m \leftarrow 1$ to M **do**
3: $\quad$ $slice \leftarrow$ SLICEANDRESTRICT$(\mathcal{I})$ $\triangleright$ $O(d \cdot n)$
4: $\quad$ U-TEST$(slice)$ $\triangleright$ $O(n)$
5: **return** average of $1-$U-TEST

$d \ll n \rightarrow$ overall complexity: $O(n \cdot log(n) + M \cdot n)$

See (Fouché and Böhm, 2019) for further details.

Mann-Whitney P (MWP)

MCDE, where $\mathcal{T}$ is a two-sided Mann-Whitney U test

- Non-parametric (R4)
- Operates on ranks (ordinal data) $\rightarrow$ Robust (R6)

$\rightarrow$ Requires indexing

MWP (S)

1: $\mathcal{I} \leftarrow$ CONSTRUCTINDEX(S) $\quad \triangleright O(d \cdot n \cdot log(n))$
2: **for** $m \leftarrow 1$ to M **do**
3: $\quad slice \leftarrow$ SLICEANDRESTRICT$(\mathcal{I})$ $\quad \triangleright O(d \cdot n)$
4: $\quad$ U-TEST$(slice)$ $\quad \triangleright O(n)$
5: **return** average of $1-$U-TEST

$d \ll n \rightarrow$ overall complexity: $O(n \cdot log(n) + M \cdot n)$

See (Fouché and Böhm, 2019) for further details.

Mann-Whitney P (MWP)

MCDE, where $\mathcal{T}$ is a two-sided Mann-Whitney U test

- Non-parametric (R4)
- Operates on ranks (ordinal data) $\rightarrow$ Robust (R6)

$\rightarrow$ Requires indexing

MWP (S)

1: $\mathcal{I} \leftarrow$ ConstructIndex(S) $\triangleright\ O(d \cdot n \cdot log(n))$
2: **for** $m \leftarrow 1$ to M **do**
3: $\quad slice \leftarrow$ SliceAndRestrict$(\mathcal{I})$ $\triangleright\ O(d \cdot n)$
4: $\quad$ U-test$(slice)$ $\triangleright\ O(n)$
5: **return** average of $1-$U-test

$d \ll n \rightarrow$ overall complexity: $O(n \cdot log(n) + M \cdot n)$

See (Fouché and Böhm, 2019) for further details.

Mann-Whitney P (MWP)

MCDE, where $\mathcal{T}$ is a two-sided Mann-Whitney U test

- Non-parametric (R4)
- Operates on ranks (ordinal data) $\rightarrow$ Robust (R6)

$\rightarrow$ Requires indexing

MWP (S)

1: $\mathcal{I} \leftarrow$ CONSTRUCTINDEX(S) $\quad \triangleright\, O(d \cdot n \cdot log(n))$
2: **for** $m \leftarrow 1$ to M **do**
3: $\quad slice \leftarrow$ SLICEANDRESTRICT$(\mathcal{I})$ $\quad \triangleright\, O(d \cdot n)$
4: $\quad$ U-TEST$(slice)$ $\quad \triangleright\, O(n)$
5: **return** average of $1-$U-TEST

$d \ll n \rightarrow$ overall complexity: $O(n \cdot log(n) + M \cdot n)$

See (Fouché and Böhm, 2019) for further details.

Mann-Whitney P (MWP)

MCDE, where $\mathcal{T}$ is a two-sided Mann-Whitney U test

- Non-parametric (R4)
- Operates on ranks (ordinal data) $\rightarrow$ Robust (R6)

$\rightarrow$ Requires indexing

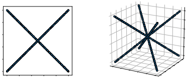
MWP (S)

1: $\mathcal{I} \leftarrow$ ConstructIndex(S) $\triangleright\ O(d \cdot n \cdot log(n))$
2: **for** $m \leftarrow 1$ to M **do**
3: $\quad slice \leftarrow$ SliceAndRestrict$(\mathcal{I})$ $\triangleright\ O(d \cdot n)$
4: $\quad$ U-test$(slice)$ $\triangleright\ O(n)$
5: **return** average of $1-$U-test

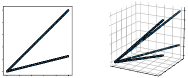
$d \ll n \rightarrow$ overall complexity: $O(n \cdot log(n) + M \cdot n)$

See (Fouché and Böhm, 2019) for further details.

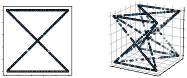
Evaluation – 12 Benchmark data sets



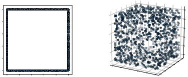
(a) Cross (C)



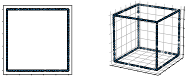
(b) Double linear (Dl)



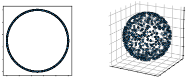
(c) Hourglass (H)



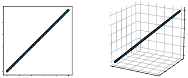
(d) Hypercube (Hc)



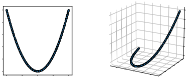
(e) Hc Graph (HcG)



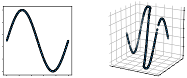
(f) Hypersphere (Hs)



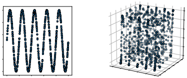
(g) Linear (L)



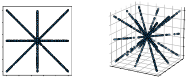
(h) Parabolic (P)



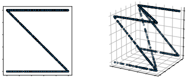
(i) Sine (P=1) (S1)



(j) Sine (P=5) (S5)



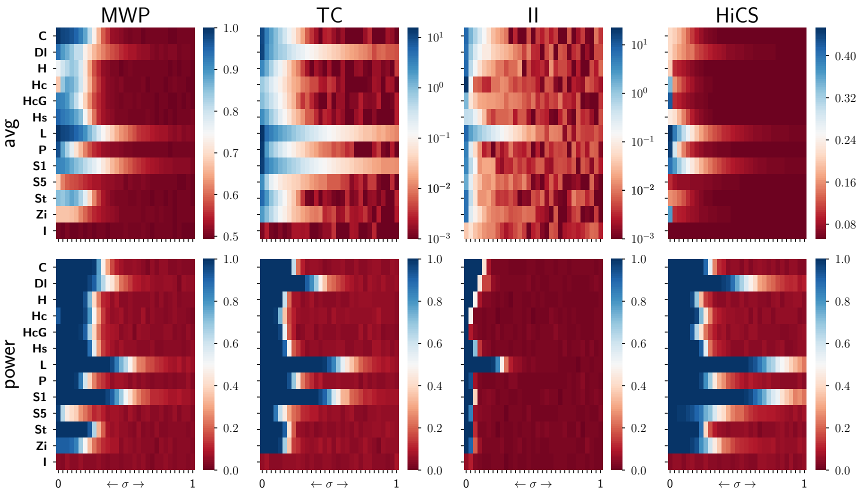
(k) Star (St)



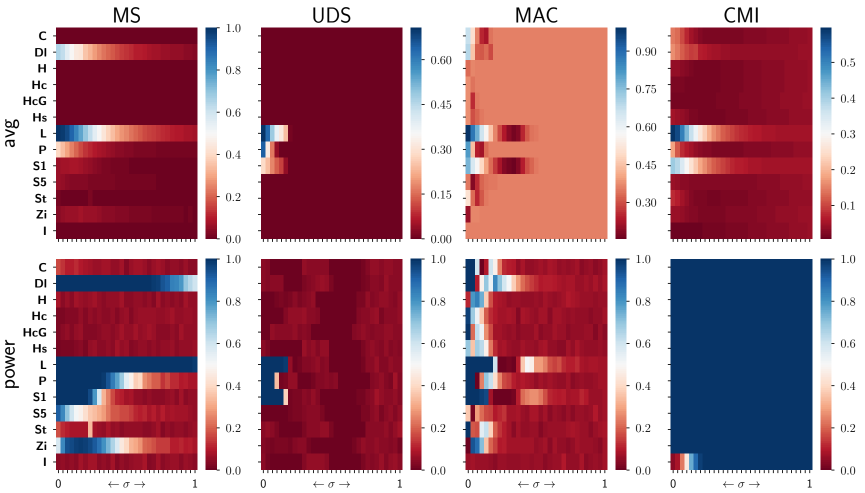
(l) Z inversed (Zi)

+ gaussian noise ($0 \leq \sigma \leq 1$)

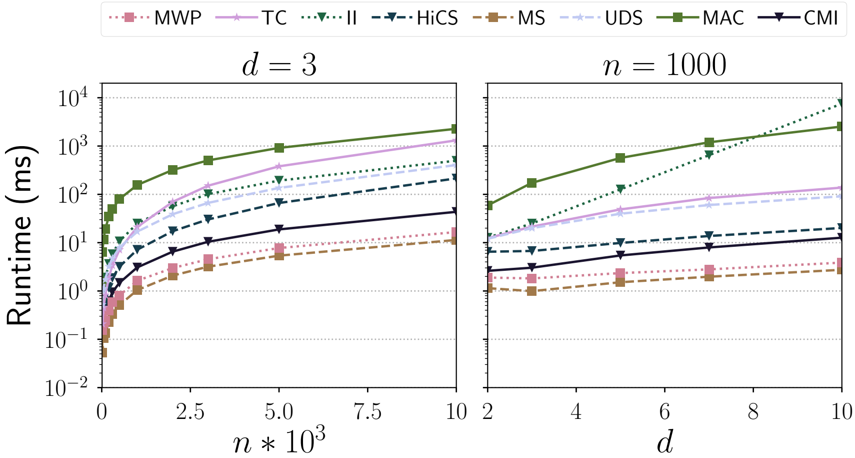
Evaluation – Distribution (1/2)



Evaluation – Distribution (2/2)



Evaluation – Scalability



Conclusion

	MS	TC	II	CMI	MAC	UDS	HiCS	MWP
R1: Multivariate	✓	✓	✓	✓	✓	✓	✓	✓
R2: Efficient	✓	✗	✗	✗	✗	✗	✓	✓
R3: Anytime	✗	✗	✗	✗	✗	✗	✓	✓
R4: General-purpose	✗	✓	✓	✗	✗	✗	✓	✓
R5: Intuitive	✓	✗	✗	✗	✗	✗	✗	✓
R6: Robust	✓	✗	✗	✗	✓	✓	✗	✓

See (Fouché and Böhm, 2019) for further experiments:

- w.r.t. number of iterations M
- w.r.t. n, d, discrete data

Software, data: https://github.com/edouardfouche/MCDE

Conclusion

	MS	TC	II	CMI	MAC	UDS	HiCS	MWP
R1: Multivariate	✓	✓	✓	✓	✓	✓	✓	✓
R2: Efficient	✓	✗	✗	✗	✗	✗	✓	✓
R3: Anytime	✗	✗	✗	✗	✗	✗	✓	✓
R4: General-purpose	✗	✓	✓	✗	✗	✗	✓	✓
R5: Intuitive	✓	✗	✗	✗	✗	✗	✗	✓
R6: Robust	✓	✗	✗	✗	✓	✓	✗	✓

See (Fouché and Böhm, 2019) for further experiments:

- w.r.t. number of iterations M
- w.r.t. n, d, discrete data

Software, data: https://github.com/edouardfouche/MCDE

Conclusion

	MS	TC	II	CMI	MAC	UDS	HiCS	MWP
R1: Multivariate	✓	✓	✓	✓	✓	✓	✓	✓
R2: Efficient	✓	✗	✗	✗	✗	✗	✓	✓
R3: Anytime	✗	✗	✗	✗	✗	✗	✓	✓
R4: General-purpose	✗	✓	✓	✗	✗	✗	✓	✓
R5: Intuitive	✓	✗	✗	✗	✗	✗	✗	✓
R6: Robust	✓	✗	✗	✗	✓	✓	✗	✓

See (Fouché and Böhm, 2019) for further experiments:

- w.r.t. number of iterations M
- w.r.t. n, d, discrete data

Software, data: `https://github.com/edouardfouche/MCDE`

References

Fouché, E. and Böhm, K. (2019). Monte carlo dependency estimation. In *SSDBM '19*.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Keller, F. (2015). *Attribute Relationship Analysis in Outlier Mining and Stream Processing*. PhD thesis, KIT-Bibliothek.

Keller, F., Müller, E., and Böhm, K. (2012). Hics: High contrast subspaces for density-based outlier ranking. In *ICDE*, pages 1037–1048. IEEE Computer Society.

McGill, W. J. (1954). Multivariate information transmission. *Trans. of the IRE Professional Group on Information Theory (TIT)*, 4:93–111.

Nguyen, H. V., Mandros, P., and Vreeken, J. (2016). Universal dependency analysis. In *SDM*, pages 792–800. SIAM.

Nguyen, H. V., Müller, E., Vreeken, J., Efros, P., and Böhm, K. (2014). Multivariate maximal correlation analysis. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 775–783. JMLR.org.

Nguyen, H. V., Müller, E., Vreeken, J., Keller, F., and Böhm, K. (2013). CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection. *SDM*, pages 198–206.

Schmid, F. and Schmidt, R. (2007). Multivariate extensions of spearman's rho and related statistics. *Statistics & Probability Letters*, 77(4):407–416.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82.

Future Work

"Extensions" of MCDE

- "Sliding Window" MCDE $\rightarrow$ requires efficient index operations
- Handle mixed attribute types (i.e., not only numerical)

Possible applications of MCDE

- Subspace Search in Streams
  - Helpful for Data Mining in "high-dimensional" streams
  - e.g., Outlier Detection, Clustering, Feature Selection
- Mining Dependency Networks in Streams $\rightarrow$ "Causal Discovery"