Universal Dependency Analysis

Hoang-Vu Nguyen
Panagiotis Mandros
Jilles Vreeken

UNIVERSITÄT DES SAARLANDES
M2CI CLUSTER OF EXCELLENCE
max planck institut informatik

Introduction

Real data is **high** dimensional

Structure, however, is usually hidden in **subspaces**

Introduction

Real data is **high** dimensional

Structure, however, is usually hidden in **subspaces**

We are interested in subspaces that **strongly interact**

Discovering interaction

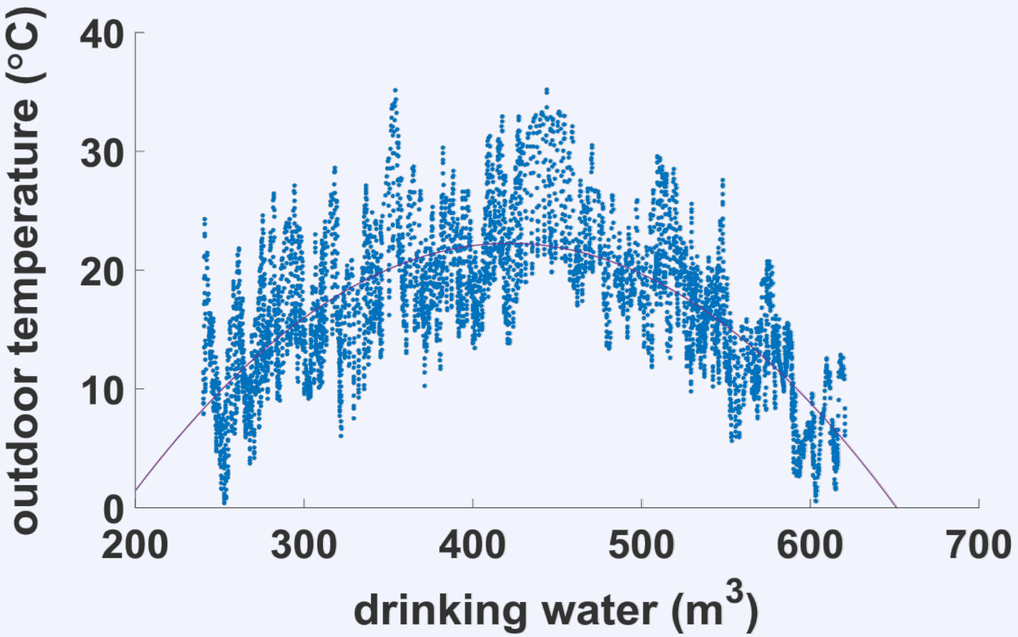
Correlated subspaces → **hidden patterns**

- which in turn allows **knowledge discovery**

Discovering interaction

Correlated subspaces → hidden patterns

- which in turn allows knowledge discovery



Revealing structure

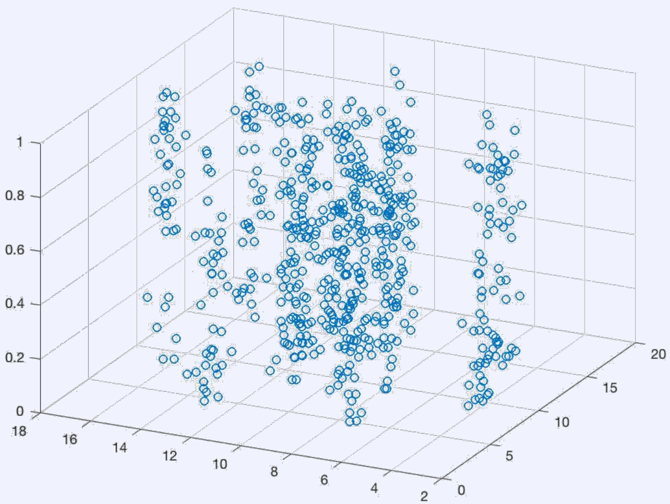
Clusters **may not** be formed in the full space

- **noisy** and **irrelevant** attributes obstruct the formation
- intuitively, they should **not correlate** with the rest

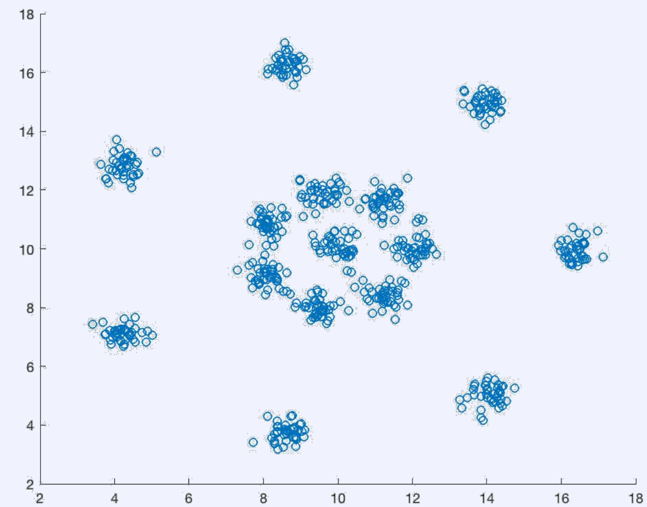
Revealing structure

Clusters **may not** be formed in the full space

- **noisy** and **irrelevant** attributes obstruct the formation
- intuitively, they should **not correlate** with the rest



3D



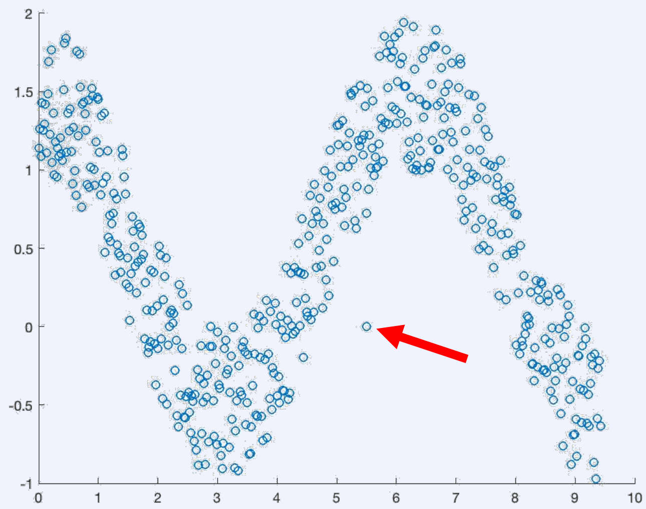
2D

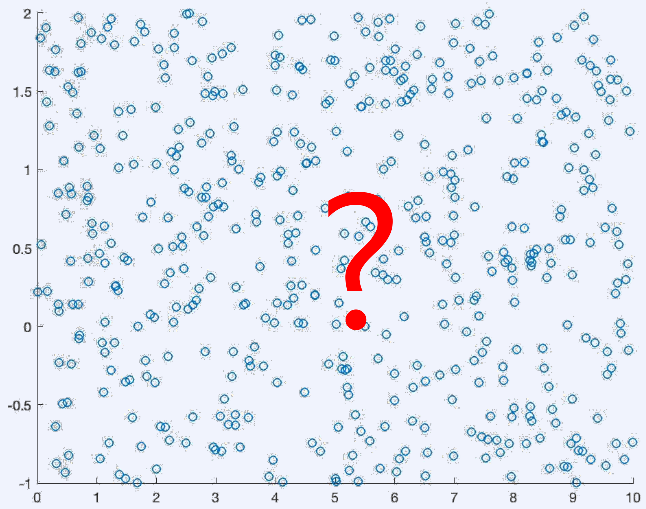
Pointing out anomalies

Outliers are **easier** to distinguish when their neighborhood is ”grouped"

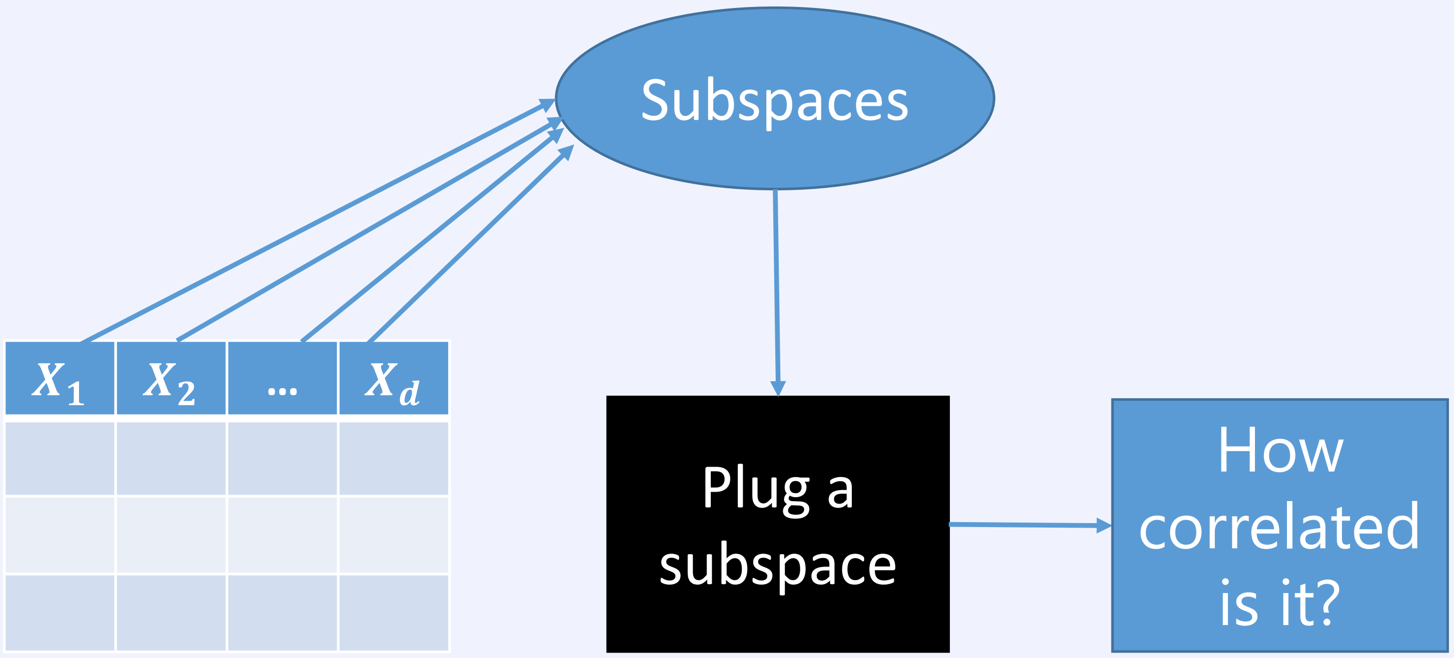
Pointing out anomalies

Outliers are **easier** to distinguish when their neighborhood is "grouped"

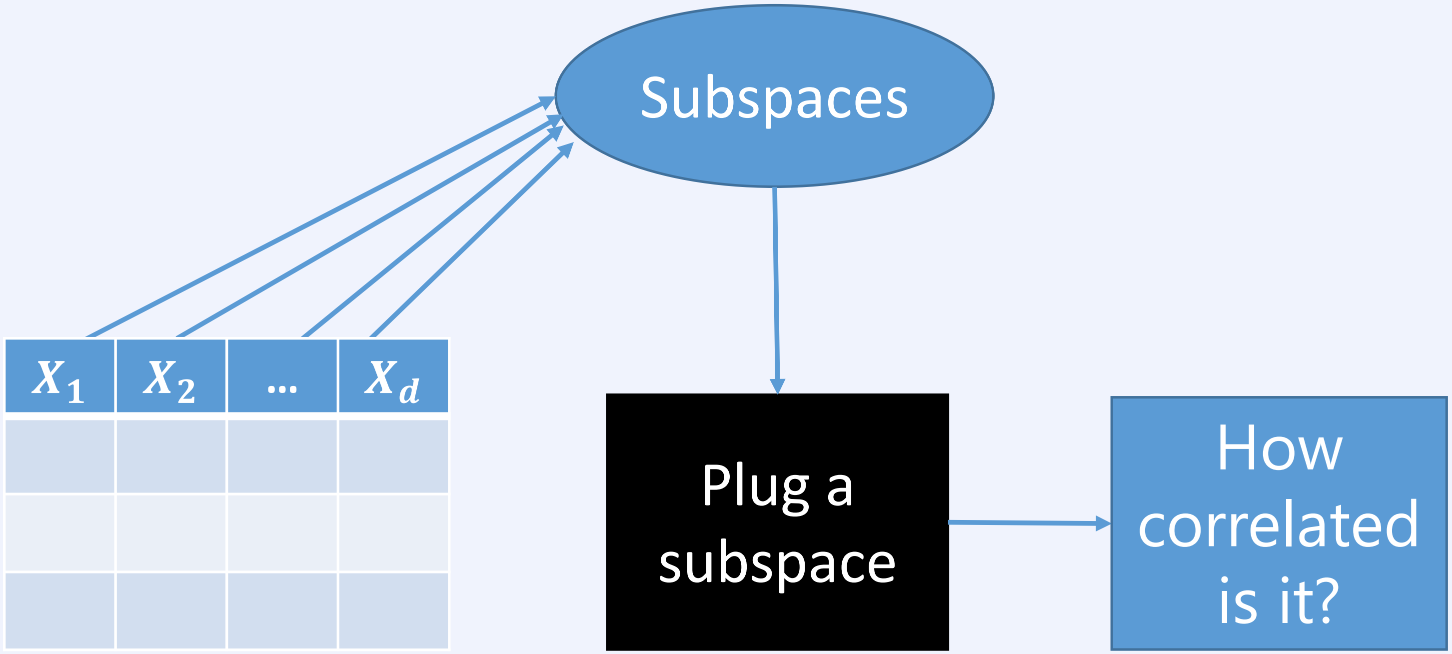




What do we want?



What do we want?



And we want this for
continuous-valued data

Challenges

Beyond linear dependencies

Multivariate

Non-parametric

Efficient

Challenges

Beyond linear dependencies

Multivariate

Non-parametric

Efficient

Comparable scores

Universality

We should be able to compare subspaces
of different dimensionality

Universality

We should be able to compare subspaces
of different dimensionality

Current approaches, however, indicate
higher correlation for larger subspaces

$$score(X_1, X_2) \leq score(X_1, X_2, X_3)$$

Universality

We should be able to compare subspaces of different dimensionality

Current approaches, however, indicate higher correlation for larger subspaces

$$score(X_1, X_2) \leq score(X_1, X_2, X_3)$$

Bias towards larger dimensionalities

UDS

- Beyond linear dependencies
- Multivariate
- Non-parametric
- Efficient
- Comparable scores

Beyond linear

Information-theoretic measures are able to capture non-linear dependencies

In addition, they have properties that **match our intuition**

Beyond linear

Information-theoretic measures are able to capture non-linear dependencies

In addition, they have properties that **match our intuition**

And that is because Shannon entropy is

- Non-negative
- conditioning can only add information
- 0 iff the variables are functionally dependent
- and many other things…

However..

They have many shortcomings when it comes to continuous-valued data

However..

They have many shortcomings when it comes to continuous-valued data

$$-\sum_{x \in X} p(x) \log p(x) \rightarrow -\int p(x) \log p(x) \mathrm{dx}$$

However..

They have many shortcomings when it comes to continuous-valued data

$$-\sum_{x \in X} p(x) \log p(x) \rightarrow -\int p(x) \log p(x) \mathrm{dx}$$

Some issues are

- differential entropy can be **negative**
- $H(X|Y) = 0$ **does not** imply functional dependency
- furthermore, it requires pdf **estimation**

(Rao et al., 2004)

Cumulative entropy

$$h(X) = -\int P(x) \log P(x) \, dx$$

Information-theoretic measure for randomness of continuous-valued data

(Rao et al., 2004; Crescenzo & Longobardi, 2009; Nguyen et al., 2013)

Cumulative entropy

Cumulative distribution

$$h(X) = -\int \underline{P(x)} \log \underline{P(x)}\, dx$$

Information-theoretic measure for randomness of continuous-valued data

(Rao et al., 2004; Crescenzo & Longobardi, 2009; Nguyen et al., 2013)

Cumulative entropy

Cumulative distribution

$$h(X) = -\int \underline{P(x)} \log \underline{P(x)} \, dx$$

Information-theoretic measure for randomness of continuous-valued data

Carries the **nice** properties of Shannon entropy to the continuous domain

(Rao et al., 2004; Crescenzo & Longobardi, 2009; Nguyen et al., 2013)

Cumulative entropy

Cumulative distribution

$$h(X) = -\int \underline{P(x)} \log \underline{P(x)}\, dx$$

$h(X) \geq 0$

$h(X|Y) \geq 0$, with equality iff X is a function of Y

$h(X|Y) \leq h(x)$, with equality iff X and Y are independent

Carries the nice properties of Shannon entropy

to the continuous domain

(Rao et al., 2004; Crescenzo & Longobardi, 2009; Nguyen et al., 2013)

UDS

- ✓ Beyond linear dependencies
- Multivariate
- Non-parametric
- Efficient
- Comparable

UDS

- ✓ Beyond linear dependencies
- Multivariate
- Non-parametric
- Efficient
- Comparable scores

Multivariate

To address this issue, we will make use of **total correlation**

$$C(X_1, \dots, X_d) = \sum_{i=2}^{d} H(X_i) - H(X_i \mid X_1, \dots, X_{i-1})$$

(Watanabe, 1960)

Multivariate

To address this issue, we will make use of **total correlation**

$$C(X_1, \dots, X_d) = \sum_{i=2}^{d} H(X_i) - H(X_i \mid X_1, \dots, X_{i-1})$$

replace Shannon entropy with Cumulative entropy

$$score(X_1, \dots, X_d) = \sum_{i=2}^{d} h(X_i) - h(X_i \mid X_1, \dots, X_{i-1})$$

(Watanabe, 1960)

UDS

✓ Beyond linear dependencies

✓ Multivariate

Non-parametric

Efficient

Comparable scores

UDS

- ✓ Beyond linear dependencies
- ✓ Multivariate
- Non-parametric
- Efficient
- Comparable scores

Non-parametric

Cumulative entropy is non-parametrically estimated from empirical data in closed-form expression

$$h(X) = -\sum_{i=2}^{n} (X_i - X_{i-1}) \frac{i}{n} log \frac{i}{n}$$

Non-parametric

Cumulative entropy is non-parametrically estimated from empirical data in closed-form expression

We chose to non-parametrically estimate conditional Cumulative entropy through optimal discretization

$$g = \underset{g \in G}{\operatorname{argmax}}\; h(Y) - h(Y|X^g)$$

(Reshef et al.,2011; Nguyen et al., 2014; Vreeken, 2015)

Non-parametric

Cumulative entropy is non-parametrically estimated from empirical data in closed-form expression

We chose to non-parametrically estimate conditional Cumulative entropy through optimal discretization

$$g = \underset{g \in G}{\operatorname{argmax}}\ h(Y) - h(Y|X^g) - r(g)$$

(Reshef et al.,2011; Nguyen et al., 2014; Vreeken, 2015)

UDS

- ✓ Beyond linear dependencies
- ✓ Multivariate
- ✓ Non-parametric
- Efficient
- Comparable scores

UDS

- ✓ Beyond linear dependencies
- ✓ Multivariate
- ✓ Non-parametric
- Efficient
- Comparable scores

Efficiency

Cumulative entropy is estimated in time **linear** to the number of samples

Efficiency

Cumulative entropy is estimated in time
linear to the number of samples

We show that we can optimally discretize our data efficiently by dynamic programming

$$O(m \log m + m\beta^2) \ll O(2^m)$$

m = number of samples
β controls discretization

UDS

- ✓ Beyond linear dependencies
- ✓ Multivariate
- ✓ Non-parametric
- ✓ Efficient
- Comparable scores

UDS

- ✓ Beyond linear dependencies
- ✓ Multivariate
- ✓ Non-parametric
- ✓ Efficient
- Comparable scores

Universality

We address universality using an intuitive idea

We **normalize** our score by the maximal information the variables could add

Universality

We address universality using an intuitive idea

We **normalize** our score by the maximal information the variables could add

$$score(X_1, \dots, X_d) = \frac{\sum_{i=2}^{d} h(X_i) - h(X_i \mid X_1, \dots, X_{i-1})}{\sum_{i=2}^{d} h(X_i)}$$

Universality

We address universality using an intuitive idea

We **normalize** our score by the maximal information the variables could add

$$score(X_1, \ldots, X_d) = \frac{\sum_{i=2}^{d} h(X_i) - h(X_i \mid X_1, \ldots, X_{i-1})}{\sum_{i=2}^{d} h(X_i)}$$

Variables that contribute only little to the nominator, get penalized by the denominator

UDS

- ✓ Beyond linear dependencies
- ✓ Multivariate
- ✓ Non-parametric
- ✓ Efficient
- ✓ Comparable scores

UDS

$$UDS(X_1, \dots, X_d) = \frac{\sum_{i=2}^{d}(X_i) - h(X_i \mid X_1, \dots, X_{i-1})}{\sum_{i=2}^{d} h(X_i)}$$

Properties

- UDS$(X_1, \dots, X_d) \in [0,1]$
- UDS$(X_1, \dots, X_d) = 0 \; iff \; X_1, \dots, X_d$ are statistically independent
- UDS$(X_1, \dots, X_d) = 1 \; iff$ there exists X_i such that all the rest attributes are a function of X_i

Code available at
eda.mmci.uni-saarland.de/uds

Experiment setup

Evaluations

- statistical power
- clustering
- outlier detection
- time efficiency
- discovering dependencies

Competitors

- HICS (ICDE'12), CMI (SDM'13), MAC (ICML'14), UDS$\neg r$

Statistical power

Generate 100 datasets **with no** dependencies

Statistical power

Generate 100 datasets **with no** dependencies

Sort their correlation scores (asc.) and set the 95-th one as a cutoff

Statistical power

Generate 100 datasets **with no** dependencies

Sort their correlation scores (asc.) and set the 95-th one as a cutoff

Generate 100 datasets with dependencies

Statistical power

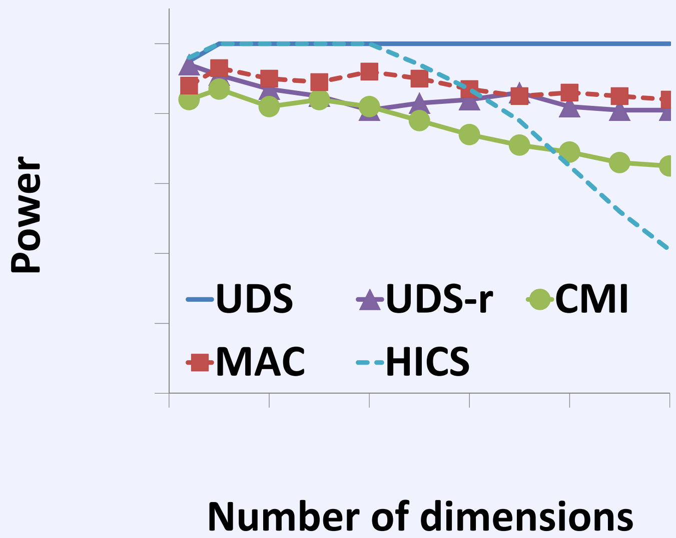
Generate 100 datasets **with no** dependencies

Sort their correlation scores (asc.) and set the 95-th one as a cutoff

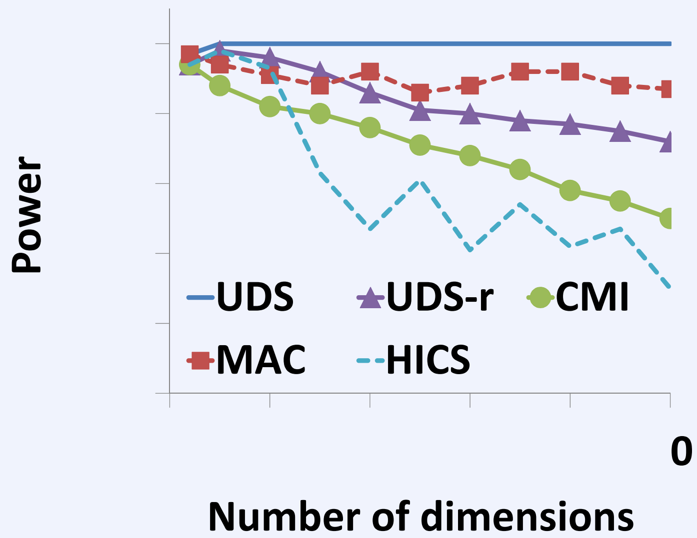
Generate 100 datasets with dependencies

$$SP = \frac{\#\,(scores > cutoff)}{100}$$

Statistical power



$f(x) = 2x + 1$



$f(x) = \sin 2x$

Statistical power on 2 different forms of functional dependency [Higher is better]

Clustering

Data	UDS	CMI	MAC	HICS
Optical	**0.61**	0.40	0.48	0.36
Leaves	**0.70**	0.52	0.61	0.45
Letter	**0.82**	0.64	**0.82**	0.49
PenDigits	**0.85**	0.72	**0.85**	0.71
Robot	**0.54**	0.33	0.46	0.21
Wave	**0.50**	0.24	0.38	0.18
Average	**<u>0.67</u>**	**0.48**	**0.60**	**0.40**

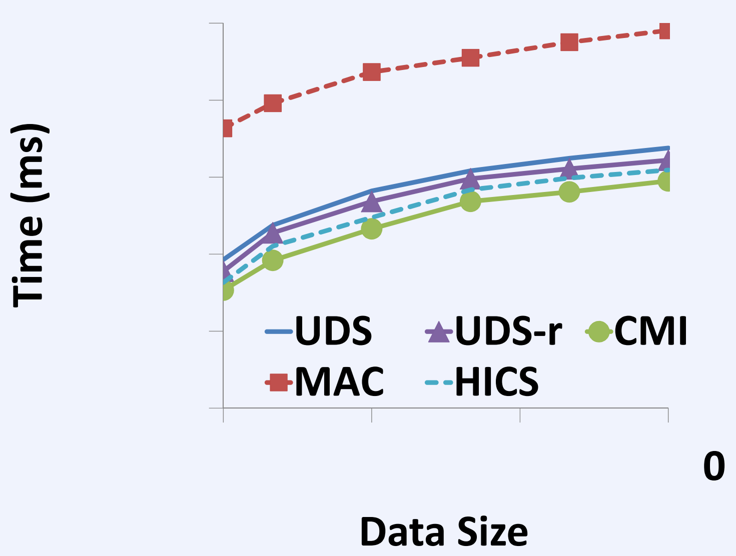
Clustering results (F1 scores) on real-world data sets [Higher is better]

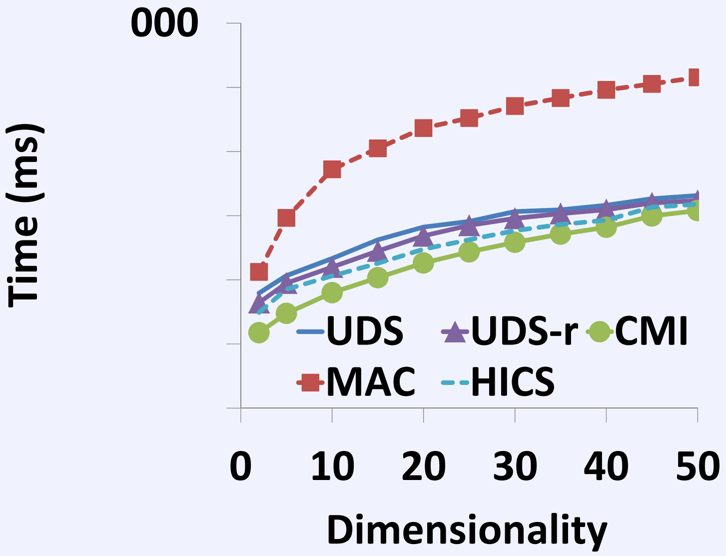
Outlier detection

Data	UDS	CMI	MAC	HICS
Ann-Thyroid	**0.98**	0.96	0.96	0.95
SatImage	**0.98**	0.74	0.95	0.86
Segmentation	**0.54**	0.39	0.51	0.49
Wave Noise	**0.51**	0.50	0.50	0.48
WBC	**0.50**	0.47	0.48	0.47
WBCD	**0.99**	0.93	**0.99**	0.91
Average	**<u>0.75</u>**	**0.66**	**0.73**	**0.69**

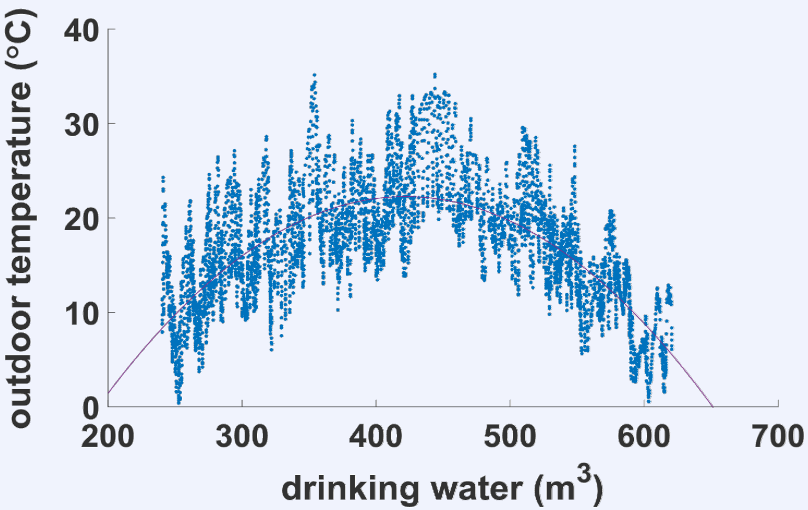
Outlier detection results (AUC scores) on real-world data sets. [Higher is better]

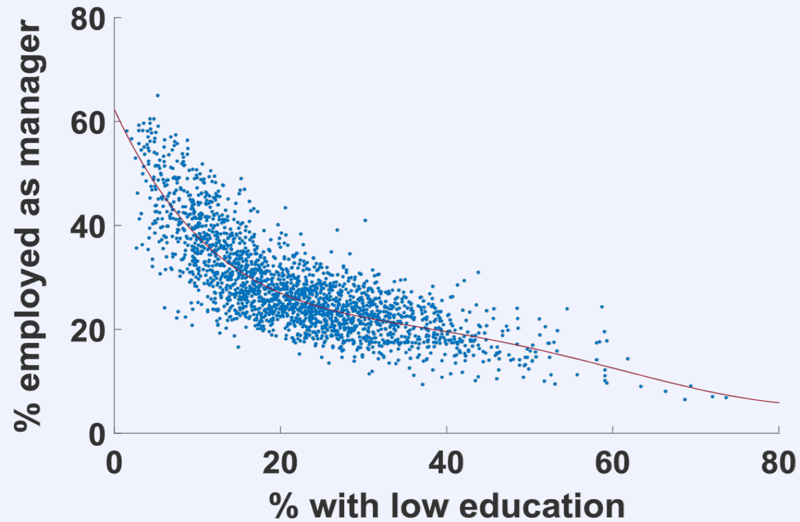
Time efficiency

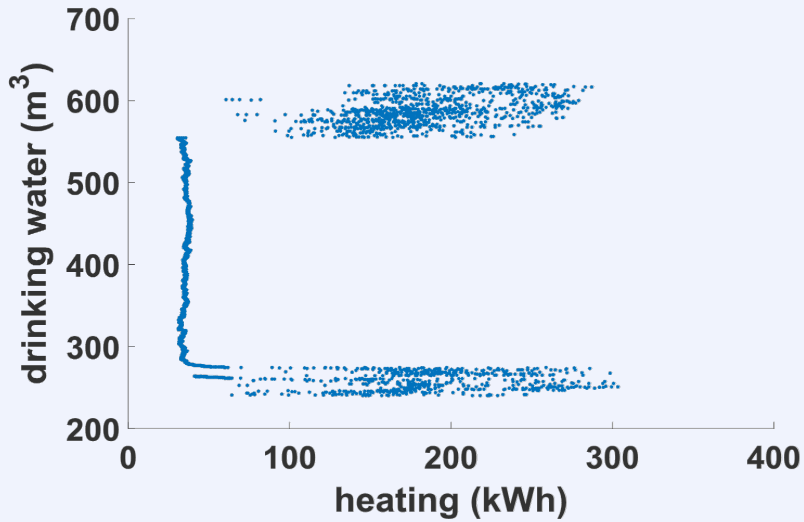




Dependencies







Conclusions

We studied the problem of assessing subspace correlations in multivariate data

UDS is non-parametric, efficient, and addresses **universality**

Extensive experiments showed that UDS **outperforms** the state-of-the-art in both statistical power and subspace search

Thank you!

We studied the problem of assessing subspace correlations in multivariate data

UDS is non-parametric, efficient, and addresses **universality**

Extensive experiments showed that UDS **outperforms** the state-of-the-art in both statistical power and subspace search

Cumulative entropy

$$h(X) = -\int P(x)\log P(x)\,dx$$

$$h(X) = -\sum_{i=2}^{n}(X_i - X_{i-1})\frac{i}{n}\,log\,\frac{i}{n}$$

$$h(X|Y) = \int h(X|y)p(y)dy$$

$$h(X|Y) = \sum_{y} h(X|y)p(y)$$