Universal Dependency Analysis

Hoang-Vu Nguyen
Panagiotis Mandros
Jilles Vreeken
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 $\rightarrow$ **hidden patterns**
- which in turn allows **knowledge discovery**
Discovering interaction

Correlated subspaces $\rightarrow$ **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
Clusters may not be formed in the full space

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

$X_1 \quad X_2 \quad \ldots \quad X_d$

Subspaces

Plug a subspace

How correlated is it?
What do we want?

How correlated is it?

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

$$\text{score}(X_1, X_2) \leq \text{score}(X_1, X_2, X_3)$$
Universality

We should be able to compare subspaces of different dimensionality.

Current approaches, however, indicate higher correlation for larger subspaces.

\[ \text{score}(X_1, X_2) \leq \text{score}(X_1, X_2, X_3) \]

**Bias** towards larger dimensionalities.
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:

- Non-negative
- Conditioning can only add information
- $0$ if and only if 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

\[- \sum_{x \in X} p(x) \log p(x) \rightarrow - \int p(x) \log p(x) dx\]
However..

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

\[- \sum_{x \in X} p(x) \log p(x) \rightarrow - \int p(x) \log p(x) \, dx\]

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

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

Information-theoretic measure for randomness of continuous-valued data

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

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

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

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

- \( h(X) \geq 0 \)
- \( h(X|Y) \geq 0 \), with equality \textit{iff} \( X \) is a function of \( Y \)
- \( h(X|Y) \leq h(x) \), with equality \textit{iff} \( X \) and \( Y \) are independent

Carries the \textbf{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

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

(Watanabe, 1960)
To address this issue, we will make use of total correlation

\[
C(X_1, ..., X_d) = \sum_{i=2}^{d} H(X_i) - H(X_i | X_1, ..., X_{i-1})
\]

replace Shannon entropy with Cumulative entropy

\[
\text{score}(X_1, ..., X_d) = \sum_{i=2}^{d} h(X_i) - h(X_i | X_1, ..., X_{i-1})
\]

(Watanabe, 1960)
Beyond linear dependencies

Multivariate

Non-parametric

Efficient

Comparable scores
Beyond linear dependencies

Multivariate

Non-parametric

Efficient

Comparable scores
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.

\[ g = \arg\max_{g \in G} h(Y) - h(Y|X^g) \]

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

\[ g = \arg \max_{g \in G} h(Y) - h(Y|X^g) - r(g) \]

(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
\( \beta \) 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

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

We address universality using an intuitive idea

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

\[
\text{score}(X_1, ..., X_d) = \frac{\sum_{i=2}^{d} h(X_i) - h(X_i \mid X_1, ..., X_{i-1})}{\sum_{i=2}^{d} h(X_i)}
\]

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

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

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

Properties

- UDS\( (X_1, \ldots, X_d) \in [0,1] \)
- UDS\( (X_1, \ldots, X_d) = 0 \) if and only if \( X_1, \ldots, X_d \) are statistically independent
- UDS\( (X_1, \ldots, X_d) = 1 \) if and only if 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→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

\[ SP = \frac{\#(scores > \text{cutoff})}{100} \]
Statistical power on 2 different forms of functional dependency [Higher is better]

Graphs showing the comparison of UDS, UDS-r, CMI, MAC, and HICS for two different functions:

1. $f(x) = 2x + 1$
2. $f(x) = \sin 2x$
Clustering results (F1 scores) on real-world data sets [Higher is better]

<table>
<thead>
<tr>
<th>Data</th>
<th>UDS</th>
<th>CMI</th>
<th>MAC</th>
<th>HICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical</td>
<td>0.61</td>
<td>0.40</td>
<td>0.48</td>
<td>0.36</td>
</tr>
<tr>
<td>Leaves</td>
<td>0.70</td>
<td>0.52</td>
<td>0.61</td>
<td>0.45</td>
</tr>
<tr>
<td>Letter</td>
<td>0.82</td>
<td>0.64</td>
<td>0.82</td>
<td>0.49</td>
</tr>
<tr>
<td>PenDigits</td>
<td>0.85</td>
<td>0.72</td>
<td>0.85</td>
<td>0.71</td>
</tr>
<tr>
<td>Robot</td>
<td>0.54</td>
<td>0.33</td>
<td>0.46</td>
<td>0.21</td>
</tr>
<tr>
<td>Wave</td>
<td>0.50</td>
<td>0.24</td>
<td>0.38</td>
<td>0.18</td>
</tr>
<tr>
<td>Average</td>
<td>0.67</td>
<td>0.48</td>
<td>0.60</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Outlier detection results (AUC scores) on real-world data sets. [Higher is better]

<table>
<thead>
<tr>
<th>Data</th>
<th>UDS</th>
<th>CMI</th>
<th>MAC</th>
<th>HICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann-Thyroid</td>
<td>0.98</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>SatImage</td>
<td>0.98</td>
<td>0.74</td>
<td>0.95</td>
<td>0.86</td>
</tr>
<tr>
<td>Segmentation</td>
<td>0.54</td>
<td>0.39</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td>Wave Noise</td>
<td>0.51</td>
<td>0.50</td>
<td>0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>WBC</td>
<td>0.50</td>
<td>0.47</td>
<td>0.48</td>
<td>0.47</td>
</tr>
<tr>
<td>WBCD</td>
<td>0.99</td>
<td>0.93</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td>Average</td>
<td>0.75</td>
<td>0.66</td>
<td>0.73</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Time efficiency

- Data Size
- Dimensionality

- UDS
- UDS-r
- CMI
- MAC
- HICS
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

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

\[ h(X) = - \sum_{i=2}^{n} \left( X_i - X_{i-1} \right) \frac{i}{n} \log \frac{i}{n} \]

\[ h(X|Y) = \int h(X|y)p(y) \, dy \]

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