Karlsruhe Institute of Technology

Institute for Program Structures and Data Organization (IPD)
Department of Informatics

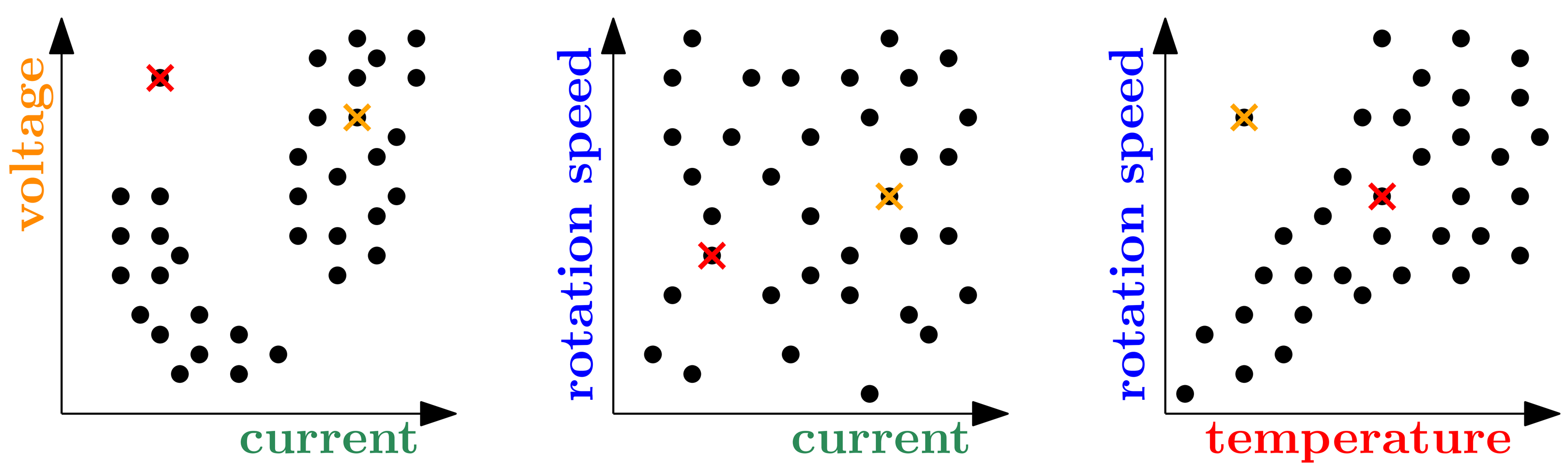
Monte Carlo Dependency Estimation

Edouard Fouché & Klemens Böhm

Motivation

Dependency Estimation is fundamental in Data Mining

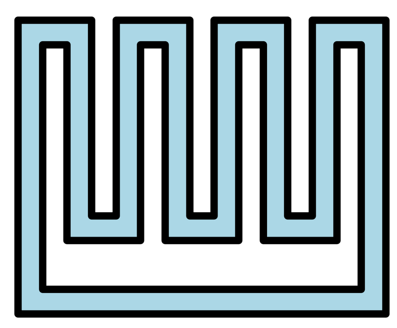
- **Feature Selection**: Find a good set of predictors
  - Improve classification accuracy, data understanding
- **Subspace Search**: Find relevant projections [4]
  - Patterns (e.g., outliers) are visible only in particular projections



- **In streams, dependency may also change over time!**

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

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



$P \propto T$

Gay-Lussac's Law: the pressure of a given amount of gas held at constant volume is proportional to its absolute temperature.

Dependency monitoring is crucial

- Predictive Maintenance, Anomaly Detection, …

Requirements

- R1: Multivariate
- R2: Efficient
- R3: General-purpose
- R4: Intuitive
- R5: Non-parametric
- R6: Interpretable
- R7: Sensitive
- R8: Robust
- R9: Anytime

R7–R9: **stream-specific**

Related Work

- Bivariate estimators (e.g., Pearson), Multivariate Spearman (MS) [9]
- Multivariate extensions of Mutual Information:
  - Interaction Information (II) [5]
  - Total Correlation (TC) [10]
- Cumulative Mutual Information (CMI) [8], Multivariate Maximal Correlation (MAC) [7], Universal Dependency Score (UDS) [6]
- High-Contrast Subspaces (HiCS) [3,4]

Our Contributions

- **MCDE: Monte Carlo Dependency Estimation**
  - 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 test
  - Extensive evaluation against state-of-the-art estimators
- **Code, data:** https://github.com/edouardfouche/MCDE

Our Approach & Evaluation

Contrast as a measure of non-independence

- Let subspace $S = \{X_1, \dots, X_d\}$ be a set of d dimensions
  - S is independent, if and only if

$$p(S) = \prod_{X_i \in S} p_{X_i}(S) \quad \Leftrightarrow \quad p(S'|\overline{S'}) = p(S') \quad \forall S' \subset S$$

- **Relaxation:**

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

$$\Leftrightarrow \quad p(S|\overline{X_i}) \underbrace{=}_{?} p_{X_i}(S) \quad \forall X_i \in S$$

Monte Carlo approach

- For $m = 1, \dots, M$
  - **1.** Choose $i \leftarrow \{1, \dots, d\}$ uniformly at random
  - **2.** Choose subspace slice s_i, i.e., a set of conditions on $\overline{X_i}$
    - "dimensionality-aware" slicing
    - $s.t.\ \mathbb{E}[|s_i|] = \mathbb{E}[|\bar{s}_i|]$ under independence
  - **3.** Choose marginal restriction r_i, i.e., a condition on X_i
  - **4.** Compute test T between $\hat{p}(S|\{s_i, r_i\})$ and $\hat{p}(S|\{\bar{s}_i, r_i\})$

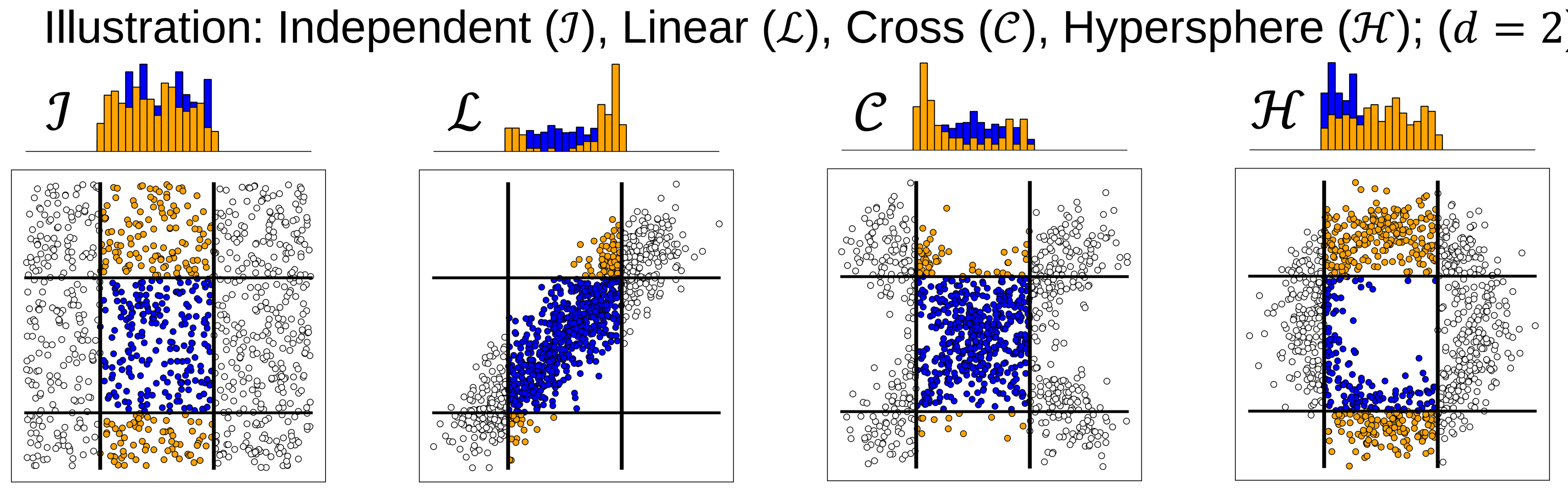
$$C(S) = \frac{1}{M} \sum_{m=1}^{M} [1 - T(\hat{p}(S|\{s_i, r_i\}), \hat{p}(S|\{\bar{s}_i, r_i\}))] \qquad \in [0,1]$$

Time complexity (incl. index construction): $O(n * \log(n) + M * n)$

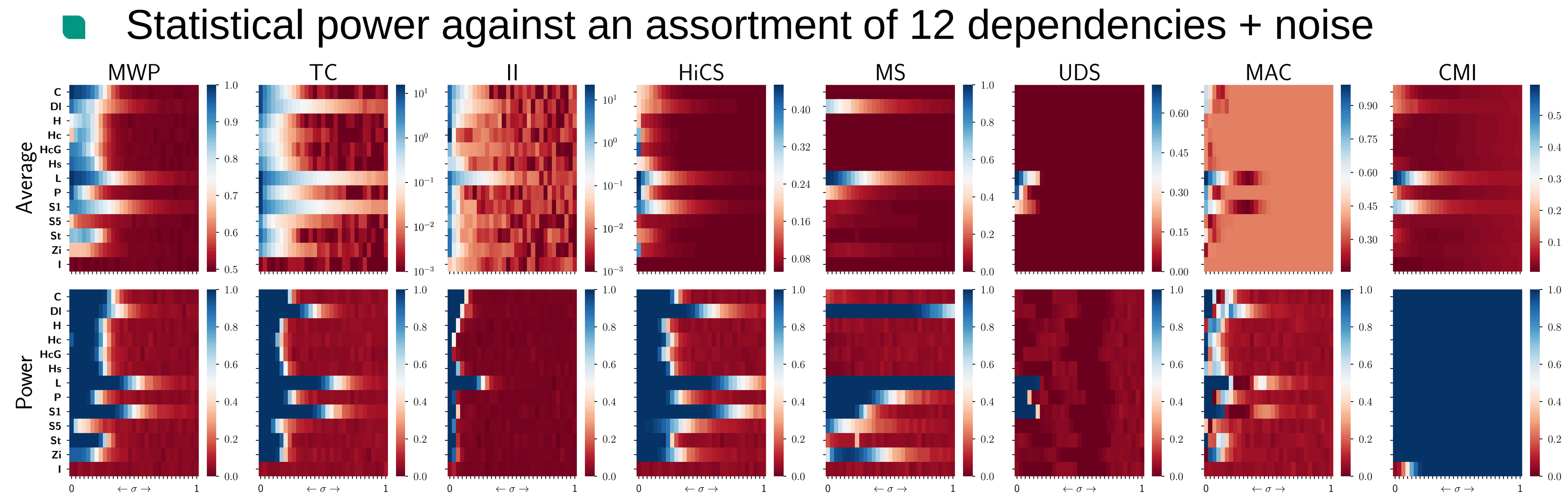
Anytime flexibility: $Pr(|C(S) - \mathbb{E}[C(S)]| \geq \varepsilon) \leq 2e^{-2M\varepsilon^2}$ (from [2])

For more details and experiments, see [1]

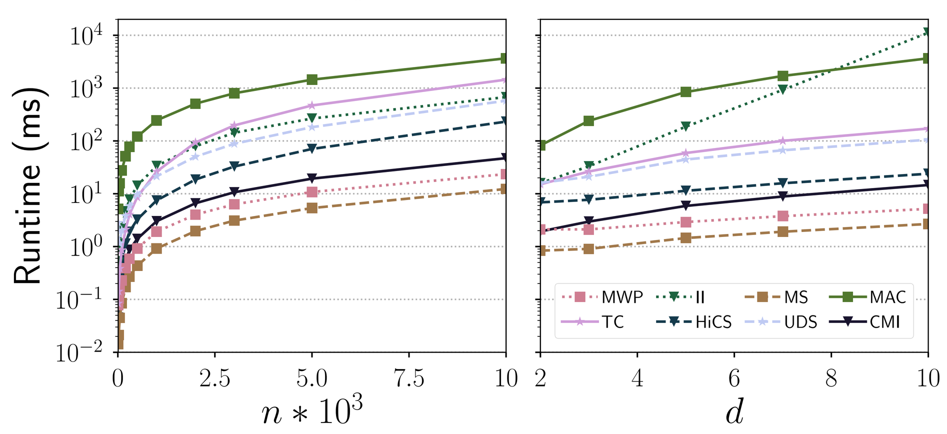
- Illustration: Independent ($\mathcal{I}$), Linear ($\mathcal{L}$), Cross ($\mathcal{C}$), Hypersphere ($\mathcal{H}$); ($d = 2$)



- Statistical power against an assortment of 12 dependencies + noise



- Scalability w.r.t. n and d



- Requirement fulfilment

Estimator	R1	R2	R3	R4	R5	R6	R7	R8	R9
MS	✓	++	✗	✓	✓	✓	✗	✓	✗
TC	✓	-	✓	✗	✓	✗	✓	✗	✗
II	✓	--	✗	✗	✓	✗	✗	✗	✗
CMI	✓	+	✗	✗	✓	✗	✗	✗	✗
MAC	✓	--	✗	✗	✓	✗	✗	✓	✗
UDS	✓	-	✗	✗	✓	✗	✗	✓	✗
HiCS	✓	+	✓	✗	✓	✗	✗	✗	✓
MWP	✓	++	✓	✓	✓	✓	✓	✓	✓

References

[1] Fouché, E. and Böhm, K. (2019). Monte Carlo Dependency Estimation. In SSDBM '19.

[2] Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association.

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

[4] Keller, F., Müller, E., and Böhm, K. (2012). HiCS: High Contrast Subspaces for Density-Based Outlier Ranking. In ICDE '12.

[5] McGill, W. J. (1954). Multivariate Information Transmission. Trans. of the IRE Professional Group on Information Theory (TIT).

[6] Nguyen, H. V., Mandros, P., and Vreeken, J. (2016). Universal Dependency Analysis. In SDM '16. SIAM.

[7] Nguyen, H. V., Müller, E., Vreeken, J., Efros, P., and Böhm, K. (2014). Multivariate Maximal Correlation Analysis. In ICML '14.

[8] 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. In SDM '13. SIAM.

[9] Schmid, F. and Schmidt, R. (2007). Multivariate extensions of spearman's rho and related statistics. Statistics & Probability Letters.

[10] Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. IBM Journal of Research and Development.