Comparison of statistical methods for spatio-temporal pattern detection in multivariate point processes

Alessandra Stella[1], Pietro Quaglio[1], Laura Sacerdote[3], Sonja Grün[1,2]

[1] Institute of Neuroscience and Medicine INM-6 and Institute for Advanced Simulation (IAS-6) and JARA-Brain Institute I (INM-10), Jülich Research Centre and JARA, Jülich, Germany
[2] Theoretical Systems Neurobiology, RWTH Aachen University, Germany
[3] Department of Mathematics "G. Peano", University of Torino, Torino, Italy

Contact: a.stella@fz-juelich.de

Introduction

Hebb's hypothesis [1] states that information processing in the brain is performed by groups of neurons called **assemblies**. Correlated activity between neurons of the cell assembly is considered as the signature of its activation [2, 3]. Several methods have been designed to extract such correlation and dependency structures from massively parallel spike trains. Here we concentrate on two methods that specifically detect significant repetitions of **Spatio Temporal Patterns** (STP), i.e. patterns with temporal delays between the composing spikes.

Motivation

- extracting correlation structures from data is a **non-trivial problem**
- **combinatorial explosion** of patterns to test for → statistical and computational issues
- **many methods** available in literature, however with **different hypothesis** and **different performances**
- need for **evaluation** and **comparison** [9].

Goal

Compare two statistical methods - **SPADE** [4,5] and **CAD** [6]- for STP detection in multivariate point processes in terms of

- **constraints and limitations**
- **statistical performance**
- **computational performance**

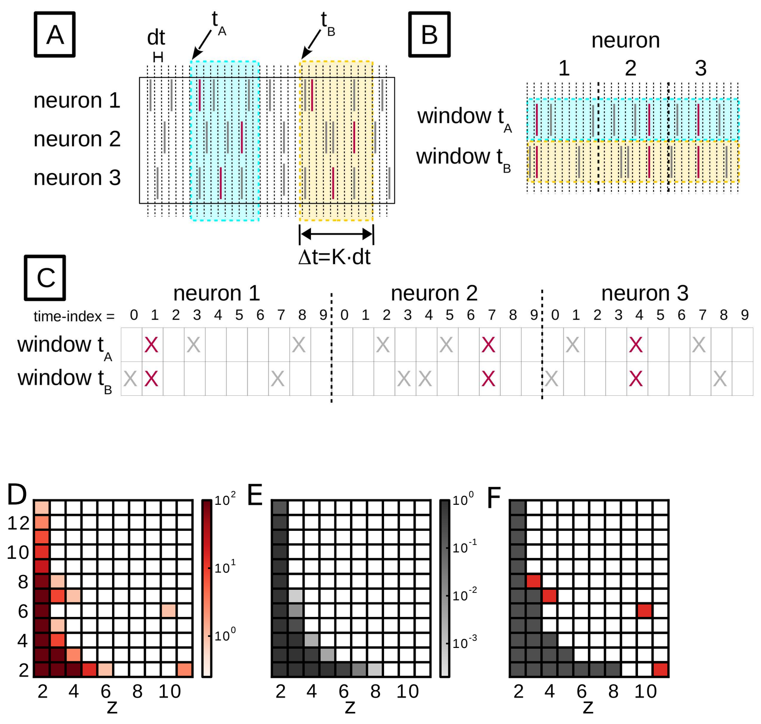
in order to obtain a possible workflow for the analysis of suitable electrophysiological data.

Methods

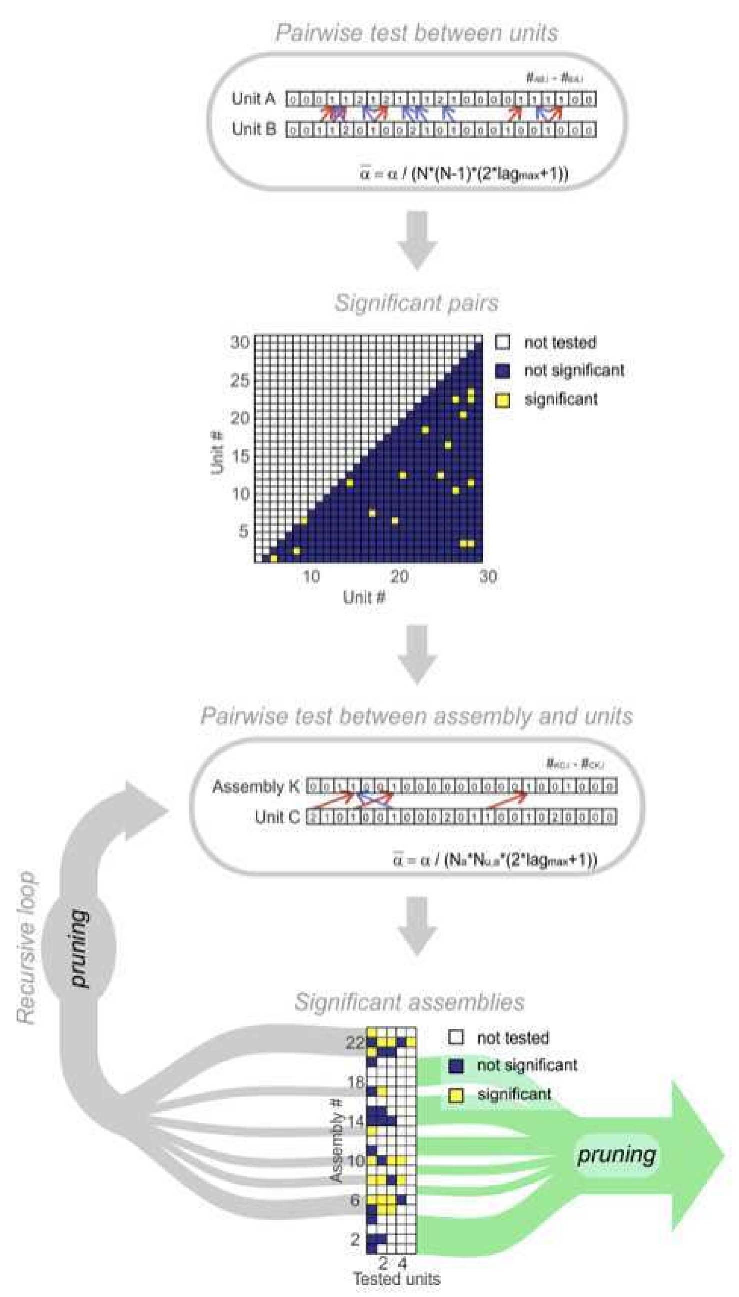
SPADE

Spike PAttern Detection and Evaluation [4,5]

SPADE detects all possible repetitions of spike patterns with a mining technique called **Frequent Itemset Mining** (FIM, Panels A,B,C). The found and counted patterns are evaluated for their statistical significance by a bootstrap technique (**Pattern Spectrum Filtering**, Panels D,E,F) employing the generation of surrogate data by spike dithering [7]. (Figure on the right)



SPADE method. Frequent Itemset Mining for massively parallel spike trains. Figure adapted from [5]. Pattern Spectrum Filtering. Figure adapted from [4].



Flowchart of **CAD method**. Figure adapted from [6].

CAD

Cell Assembly Detection [6]

CAD consists of a **statistical test** (assuming Poisson distribution of the spike trains) applied to pairwise correlations. It builds on an **agglomerative recursive algorithm**, starting by extracting all significantly correlated pairs of neurons. The resulting correlated processes are then further correlated with individual spike trains for higher order patterns. (Figure on the left)

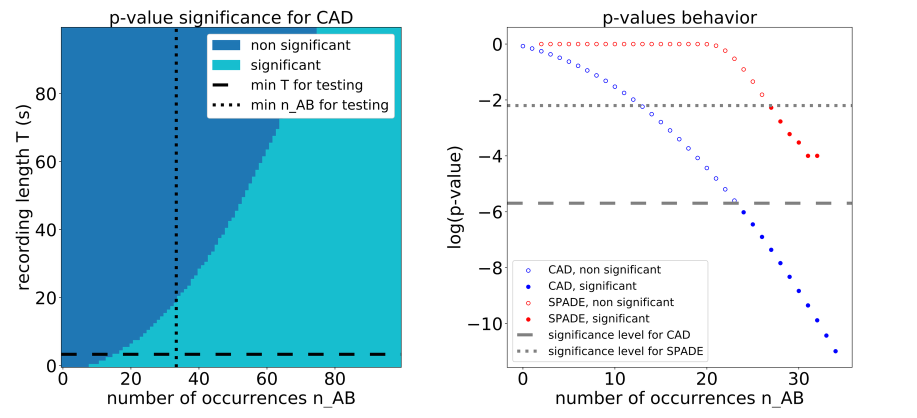
SPADE	CAD
Non-parametric statistical test (Monte Carlo and bootstrapping)	Parametric statistical test
Tests independently patterns of different sizes	Builds higher order structures hierarchically based on pairs
Conditional testing on chance overlaps between STP and background activity	Subsets of the pattern are deleted
Requires parallelization	No parallelization required

Table of **Differences** between the two methods.

Analytical Study of Pairwise Correlations

We compare analytically the two methods to see how well they capture structures of pairwise correlation.

- p-values decrease for both methods as the number of occurrences n_{AB} of a candidate pattern increases, and decrease with the recording length
- less repetitions of a pattern are needed for CAD to obtain significance with respect to SPADE.



p-values behavior for the two methods, in the case of pairwise correlation. We derive an approximate analytical function of the p-values for the CAD method (left), and then after fixing a recording length T, we compare the two methods in terms of the number of occurrences of a pattern of size two needed to obtain significance (right).

Analysis of Higher Order Correlations in Artificial Data

We compare the two methods by generating artificial data simulating parallel spike trains with assembly activity. We model massively parallel spike trains by a multi-dimensional stationary Poisson process, into which we superimpose STPs with varying size (number of neurons involved) and frequency (number of occurrences).

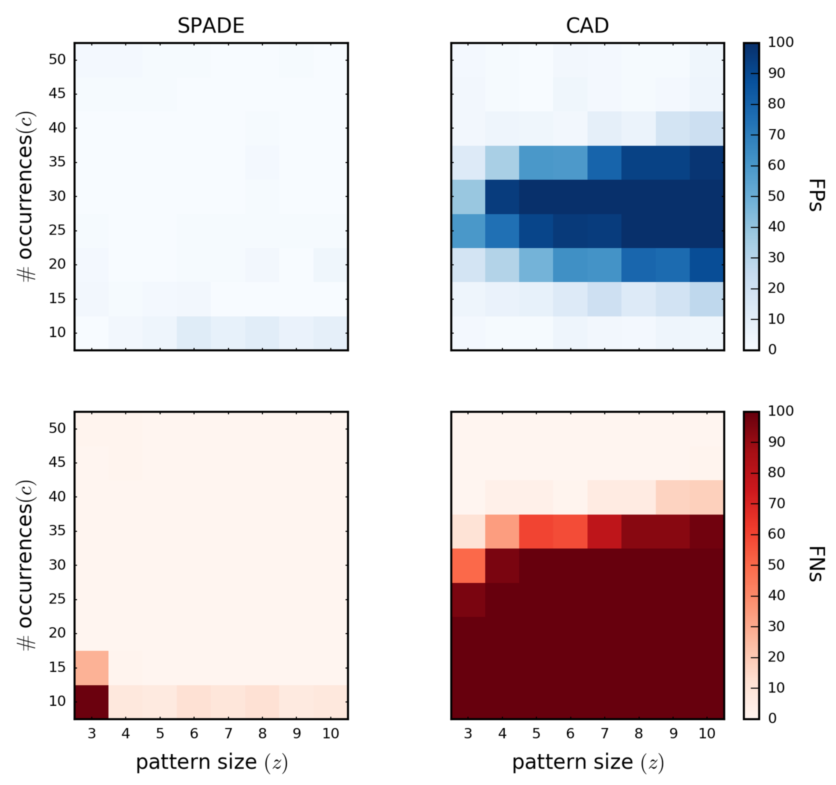
For each configuration we generate many data sets. Then, we evaluate SPADE and CAD by counting how well they detect the injected patterns (False Positives (FP: incorrect pattern detected) and False Negatives (FN: pattern injected not detected) counts).

SPADE

- is almost always able to retrieve the correct pattern

CAD

- performs well for patterns with high occurrence rate
- retrieves only subsets of the pattern for medium occurrence rate
- cannot detect the pattern for low occurrence rate.

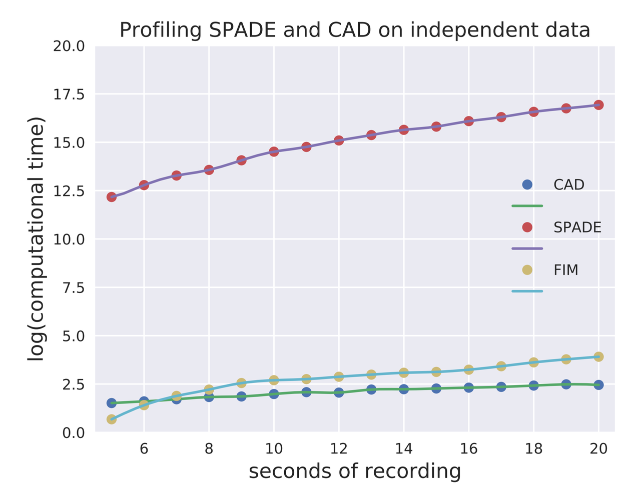


Performances in terms of FPs and FNs. For each pair (z,c) = (pattern size, number of occurrences) we perform and analyze 100 realizations. Each matrix element shows the number of realizations for which the filtered results contain one or more FPs (top row) / FNs (bottom row).

Computational Performances

We study the computational time required by the two methods to analyze independent data (no pattern injection - under the null hypothesis of independence) in function of the recording length.

Since CAD does not employ surrogate generation, the analysis is faster by many orders of magnitude with respect to SPADE.



Computational time in seconds and in logarithmic scale required by SPADE, CAD and FIM to analyze independent data.

Conclusions

	SPADE	CAD
Pairwise Correlation		✓
Independent Data	✓	✓
Higher Order STP - Low Number of Occcurrences	✓	
Higher Order STP - High Number of Occurrences	✓	✓
Computational Performance		✓

Table of comparison between SPADE and CAD. Checks indicate good performances.

Optimal workflow:

- apply SPADE for the case of short 'dense' data: low STP frequency and short recordings provide better statistical performance in reasonable computational time
- apply CAD for the case of long 'sparse' data: high STP frequency and long recordings permit good statistical and computational performance

Outlook: Comparison on more realistic non-Poisson non-stationary ground truth data to evaluate FN/FP and computational performance, expecting the latter to be less good for the CAD method.

References

[1] Hebb, D. (1949) The organization of behavior. Wiley and Sons.
[2] Singer et al. (1997) Neuronal assemblies: necessity, signature and detectability. Trends in Cognitive Sciences, 1, 252-261.
[3] Harris, K. (2005) Neural signatures of cell assembly organization. Nature Reviews Neuroscience, 5, 339-407.
[4] Torre et al. (2013) Statistical evaluation of synchronous spike patterns extracted by Frequent Item Set Mining. Frontiers in Computational Neuroscience, 7:132.
[5] Quaglio et al. (2017) Detection and Evaluation of Spatio-Temporal Spike Patterns in Massively Parallel Spike Train Data with SPADE. Frontiers in Computational Neuroscience, 11.
[6] Russo et al. (2017) Cell assemblies at multiple time scales with arbitrary lag constellations. Elife, 6.
[7] Louis et al. (2010) Surrogate spike train generation through dithering in operational time. Frontiers in Computational Neuroscience, 4:127.
[8] Gerstein et al. (1978) Identification of functionally related neural assemblies. Brain Research, 140: 43-62.
[9] Quaglio et al. (2018) Methods for identification of spike patterns in massively parallel spike trains. Biological Cybernetics, 1-24.

Acknowledgments: Helmholtz Portfolio theme (SMHB), Human Brain Project HBP Grant No. 720270 (SGA1) and No. 785907 (SGA2), DFG IRTG 1901, the DFG SPP Priority Program 1665 (GR 1753/4-1) and University of Torino.