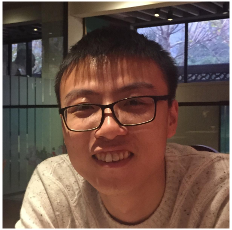
Whittle Networks: A Deep Likelihood Model for Time Series

TECHNISCHE UNIVERSITÄT DARMSTADT



Zhongjie Yu[1] Fabrizio Ventola[1] Kristian Kersting[1,2]

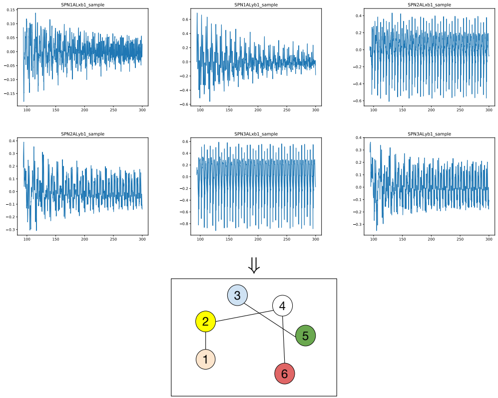
[1]Department of Computer Science, TU Darmstadt, Germany
[2]Centre for Cognitive Science, TU Darmstadt, and Hessian Center for AI
{yu, ventola, kersting}@cs.tu-darmstadt.de

MADESI KompAKI AIPHES LOEWE Exzellente Forschung für Hessens Zukunft Centre for Cognitive Science hessian.AI

A Real World Motivation

Multivariate Time Series Analysis on the Edge

sensor data from wind turbine

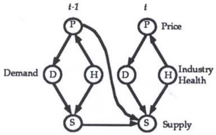


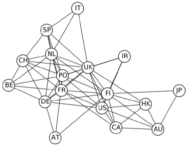
Multivariate Time Series and Graphical Models

Limitations of time series graphical models (TGMs)[1,2,3]:

- Inference is exponential in the worst case
- Sample size required for accurate learning is also exponential
- Learning requires inference as subroutine i.e. can take exponential time

$\Rightarrow$ In other words, TGMs are intractable!





[1] Tank, A., Foti, N. J., and Fox, E. B. Bayesian structure learning for stationary time series. In UAI, 2015.
[2] Dahlhaus, R. Graphical interaction models for multivariate time series. Metrika, 2000.
[3] Bach, F. R. and Jordan, M. I. Learning graphical models for stationary time series. IEEE Transactions on Signal Processing, 2004.

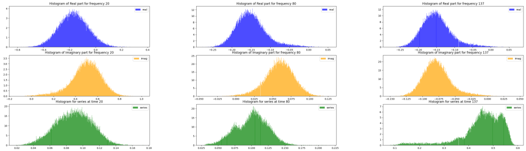
Multivariate Time Series and Graphical Models

We propose:

- The first tractable probabilistic circuit for modeling the joint distribution of multivariate time series, called Whittle sum-product networks (WSPNs), by introducing complex-valued SPNs.
- Deep likelihood functions for training deep neural networks for time series in an end-to-end fashion, called Whittle Networks.

Why Whittle Likelihood?

Time series statistics



The Fourier coefficients (real and imaginary parts) follow Gaussian distribution while the time series in the time domain at each step could follow an arbitrary distribution

Whittle approximation[4] – the Fourier coefficients from discrete Fourier transform are independent complex normal distributed:

$$d_{n,k} \sim \mathcal{N}(0, S_k), \quad k = 0, \dots, T-1$$

where S_k is the spectral density matrix:

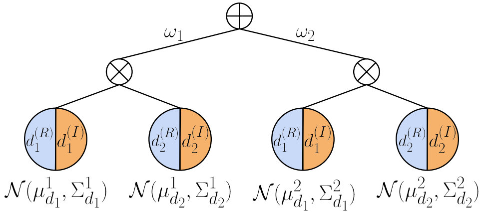
$$S_k = \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-i\lambda_k h}$$

Thus, the Whittle likelihood given N independent realizations is defined as:

$$\prod_{n=1}^{N} \prod_{k=0}^{T-1} \frac{1}{\pi^p \left|S_k\right|} e^{-d_{n,k}^{*} S_k^{-1} d_{n,k}} \tag{1}$$

[4] Whittle, P. The analysis of multiple stationary time series. Journal of the Royal Statistical Society: Series B (Methodological), 1953.

Whittle Sum-Product Networks



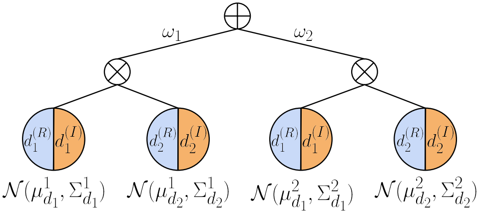
CoSPN:

- using pairwise Gaussian leaf nodes to model complex random variables
- using an adapted non-parametric independence test for structure learning[5]

The Whittle Approximation applies for stationary time series. What about non-stationary time series?

[5] Gens, R. and Domingos, P. Learning the Structure of Sum- Product Networks. ICML, 2013.

Whittle Sum-Product Networks



CoSPN:

- using pairwise Gaussian leaf nodes to model complex random variables
- using an adapted non-parametric independence test for structure learning[5]

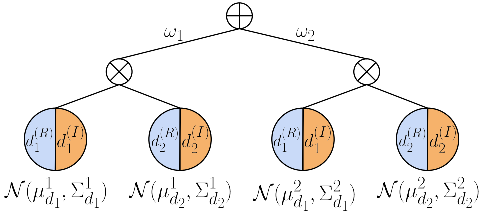
The Whittle Approximation applies for stationary time series. What about non-stationary time series?

We propose the following relaxations for general time series

- The mean of each frequency need not be 0
- The Fourier coefficients of different frequencies need not be independent

[5] Gens, R. and Domingos, P. Learning the Structure of Sum- Product Networks. ICML, 2013.

Whittle Sum-Product Networks



CoSPN:

- using pairwise Gaussian leaf nodes to model complex random variables
- using an adapted non-parametric independence test for structure learning[5]

[5] Gens, R. and Domingos, P. Learning the Structure of Sum- Product Networks. ICML, 2013.

The Whittle Approximation applies for stationary time series. What about non-stationary time series?

We propose the following relaxations for general time series

- The mean of each frequency need not be 0
- The Fourier coefficients of different frequencies need not be independent

WSPN: CoSPN which jointly models the Fourier coefficients of time series

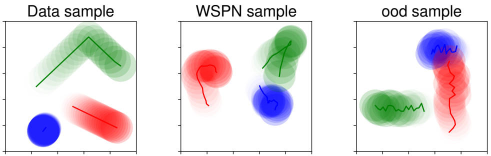
Experiments

Probabilistic Modeling of Time Series

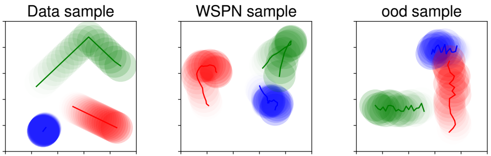
		LearnSPN	WSPN-Pair	WSPN-2d	ResSPN	ResWSPN-Pair	ResWSPN-2d	MADE
Sine	train $\Uparrow$	-0.47	2.65	**6.67**	**-60.62**	-148.48	-135.94	-105.91
	test $\Uparrow$	-0.75	1.85	**5.75**	**-63.13**	-150.90	-138.86	-108.64
	ood $\Downarrow$	$-\infty$	$-\infty$	$-\infty$	**-5880.85**	-4010.04	-4227.18	-11646865.93
MNIST	train $\Uparrow$	256.11	272.84	**277.50**	249.47	**254.46**	**254.30**	336.03
	test $\Uparrow$	254.99	270.40	**274.42**	245.67	251.74	**252.54**	327.22
	ood $\Downarrow$	**125.19**	160.29	155.76	**204.93**	218.25	216.01	136.98
Billiards	train $\Uparrow$	54.73	63.75	**65.01**	-367.83	-318.10	**-213.13**	-204.23
	test $\Uparrow$	52.80	**54.14**	**54.12**	-377.38	-324.78	**-219.04**	-252.51
	ood $\Downarrow$	-1984.38	-2348.57	**-2435.70**	-1003.49	-1052.21	**-2113.68**	-89521.82
S&P	train $\Uparrow$	-191.64	113.06	**174.45**	308.22	194.57	**1831.91**	359.52
Stock	train $\Uparrow$	-615.76	328.90	**417.81**	257.03	496.07	**1172.85**	639.10

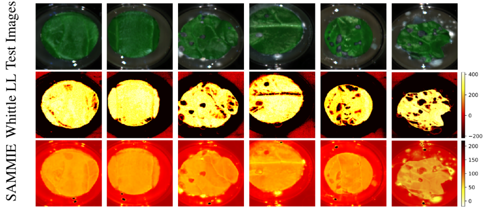
WSPNs **capture densities** over time series **better** than baselines!

Experiments
Probabilistic Modeling of Time Series



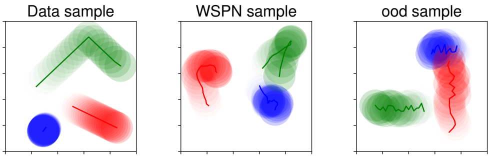
Experiments
Probabilistic Modeling of Time Series

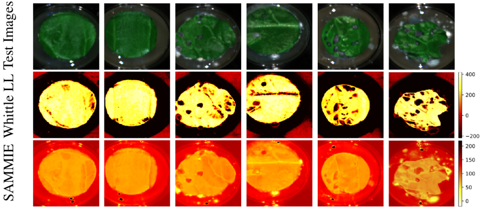


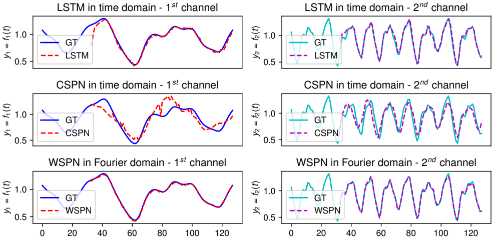


Experiments

Probabilistic Modeling of Time Series







WSPNs are great for time series modeling and forecasting!

Extracting Conditional Independence Structure of Time Series

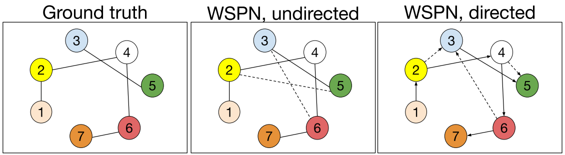
For directed acyclic graphs (DAGs):

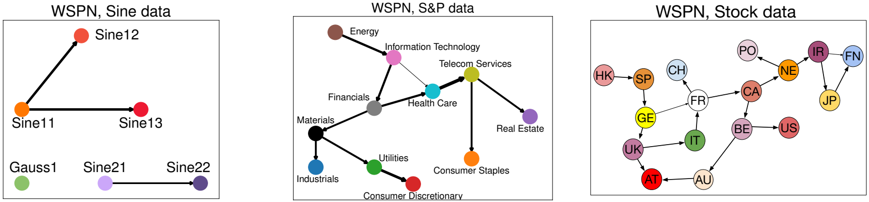
$$p(X_{1:N} \mid G, Co) \approx \prod_{n=1}^{N} \frac{\prod_{v_i \in V} p(d_n^{\{v_i \cup Pa_G(v_i)\}} \mid Co)}{\prod_{v_i \in V} p(d_n^{\{Pa_G(v_i)\}} \mid Co)}$$

For undirected graphs:

$$p(X_{1:N} \mid G, Co) \approx \prod_{n=1}^{N} \frac{\prod_{c_i \in C} p(d_n^{\{c_i\}} \mid Co)}{\prod_{s_i \in S} p(d_n^{\{s_i\}} \mid Co)}$$

Experiments
Conditional Independence Structure

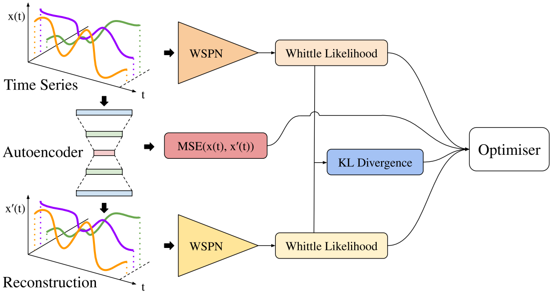




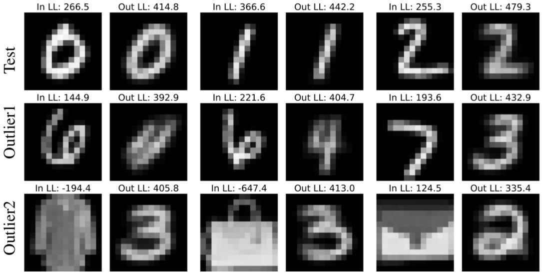
WSPNs can successfully extract the conditional independence of time series components!

Whittle Networks

Providing Meaningful Probabilities to Deep Neural Networks

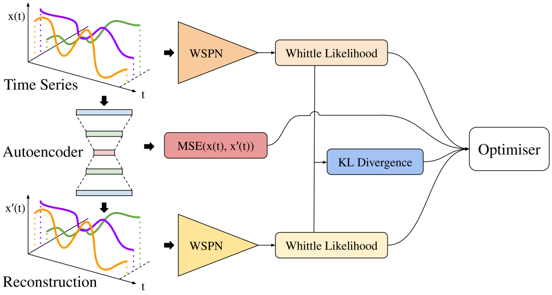


Whittle Autoencoder



Whittle Networks

Providing Meaningful Probabilities to Deep Neural Networks



Whittle Autoencoder

	WSPN Input	WSPN Output
train	295.49	411.32
test	295.22	411.10
outlier1	**239.78**	401.54
outlier2	**48.84**	397.58

Whittle Networks can provide meaningful probabilities for deep neural networks!

Conclusion

We introduce:

- CoSPN – The first complex-valued SPN for modeling complex random variables
- WSPN – The first tractable probabilistic circuit for modeling the joint distribution of multivariate time series, exploiting the Whittle approximation
- Conditional independence structure can be extracted efficiently from WSPNs
- Whittle Networks – meaningful probabilities for deep neural networks

Thanks

Zhongjie Yu
yu@cs.tu-darmstadt.de

Fabrizio Ventola
ventola@cs.tu-darmstadt.de

Kristian Kersting
kersting@cs.tu-darmstadt.de

AIML lab

https://www.aiml.informatik.tu-darmstadt.de

Code

https://github.com/ml-research/WhittleNetworks