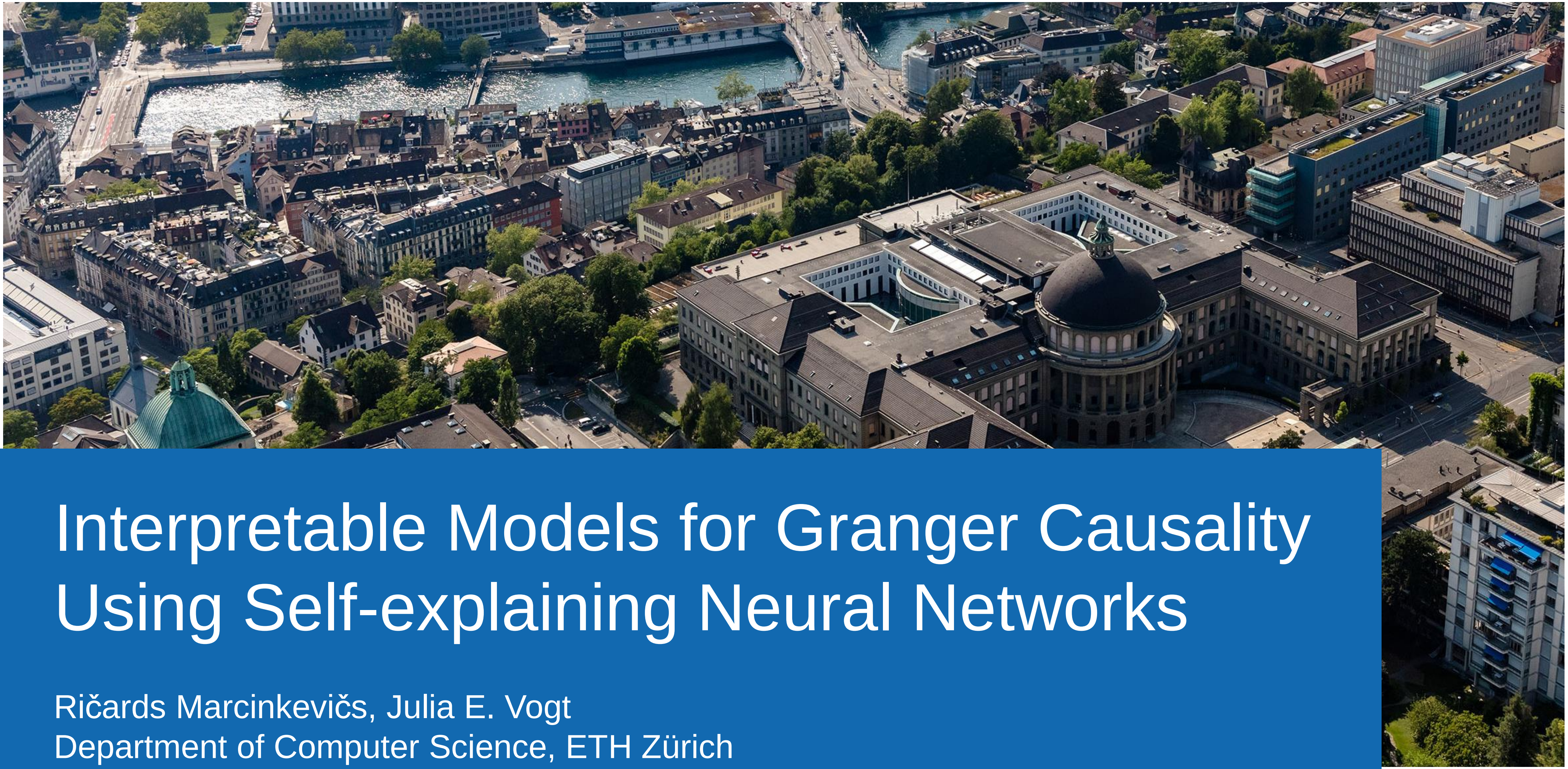
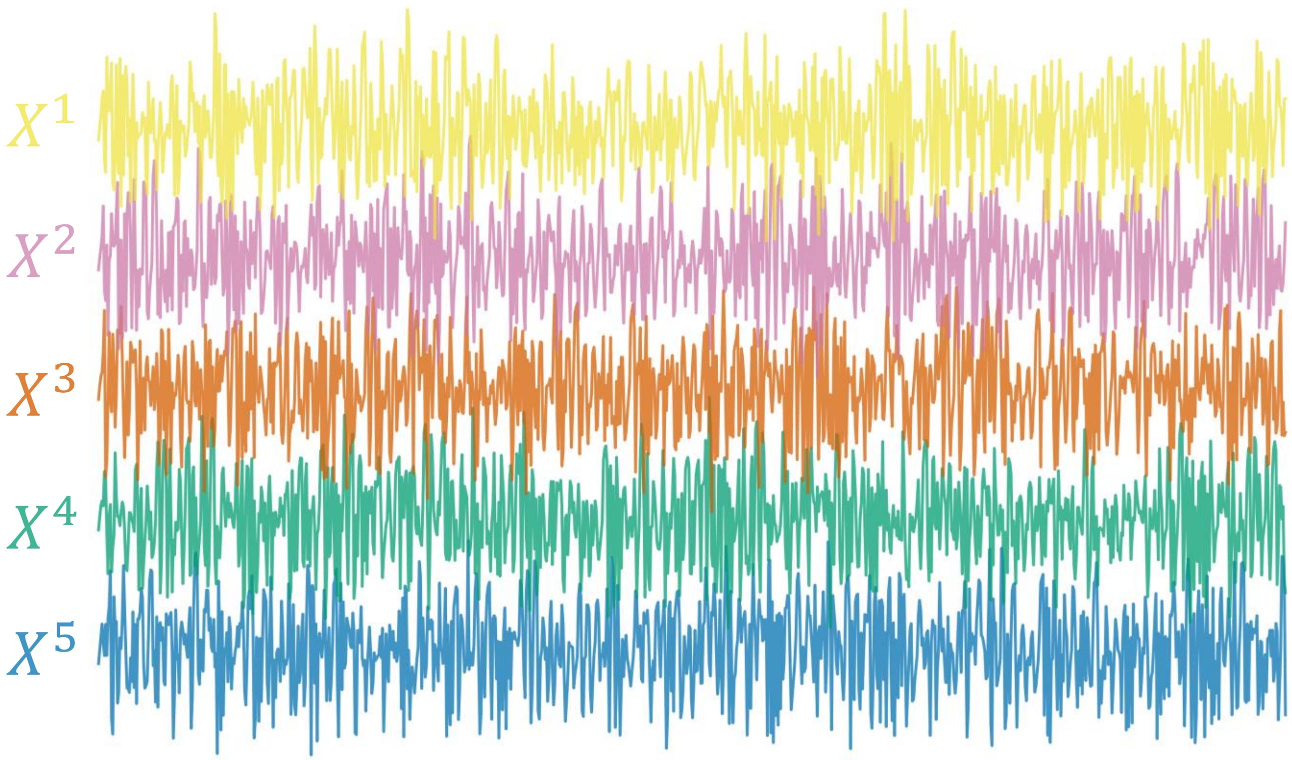
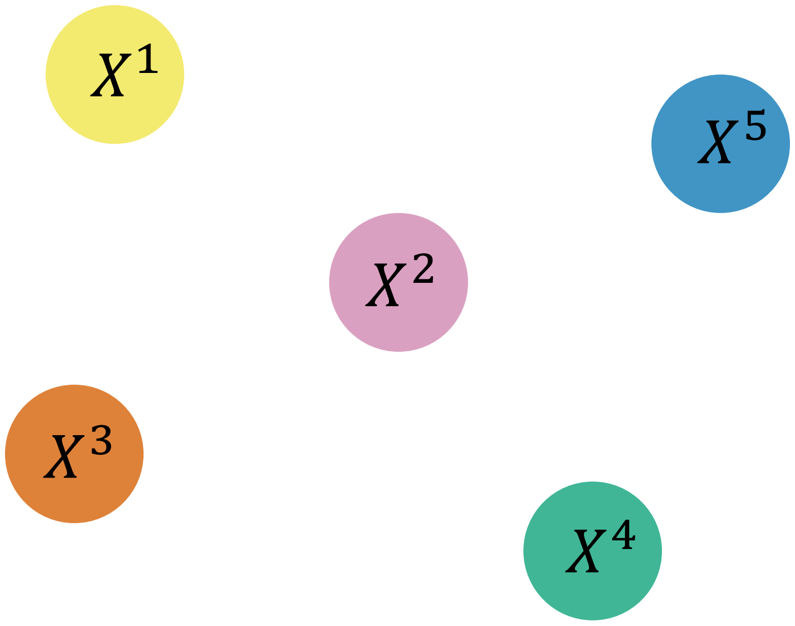


Interpretable Models for Granger Causality Using Self-explaining Neural Networks

Ričards Marcinkevičs, Julia E. Vogt
Department of Computer Science, ETH Zürich

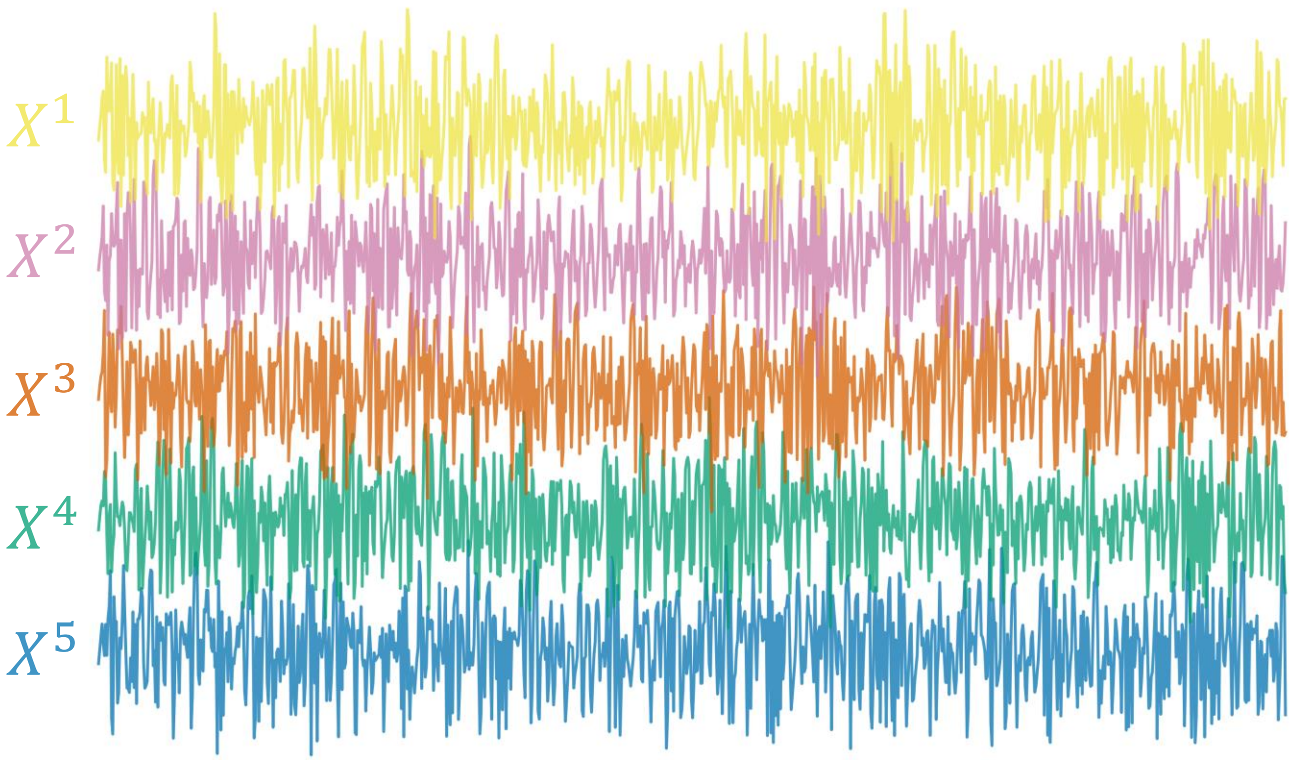
Time Series Analysis

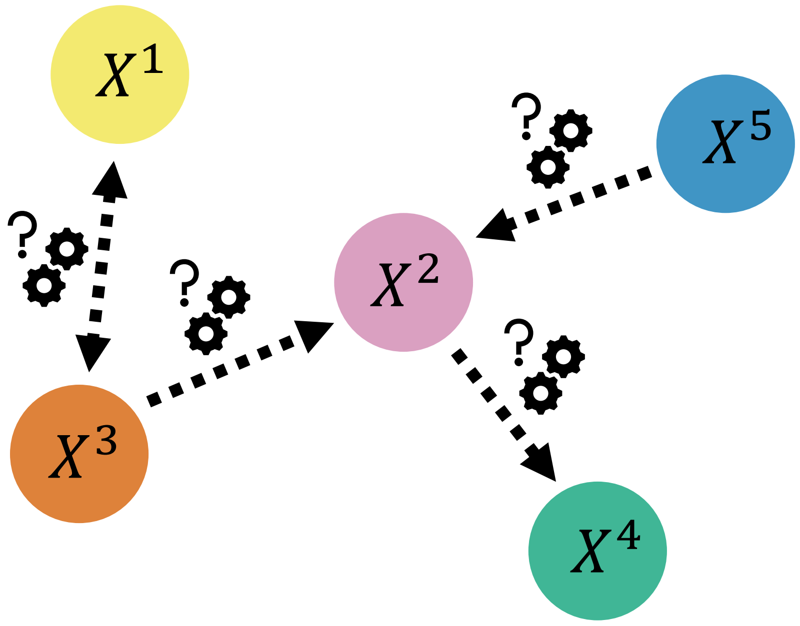




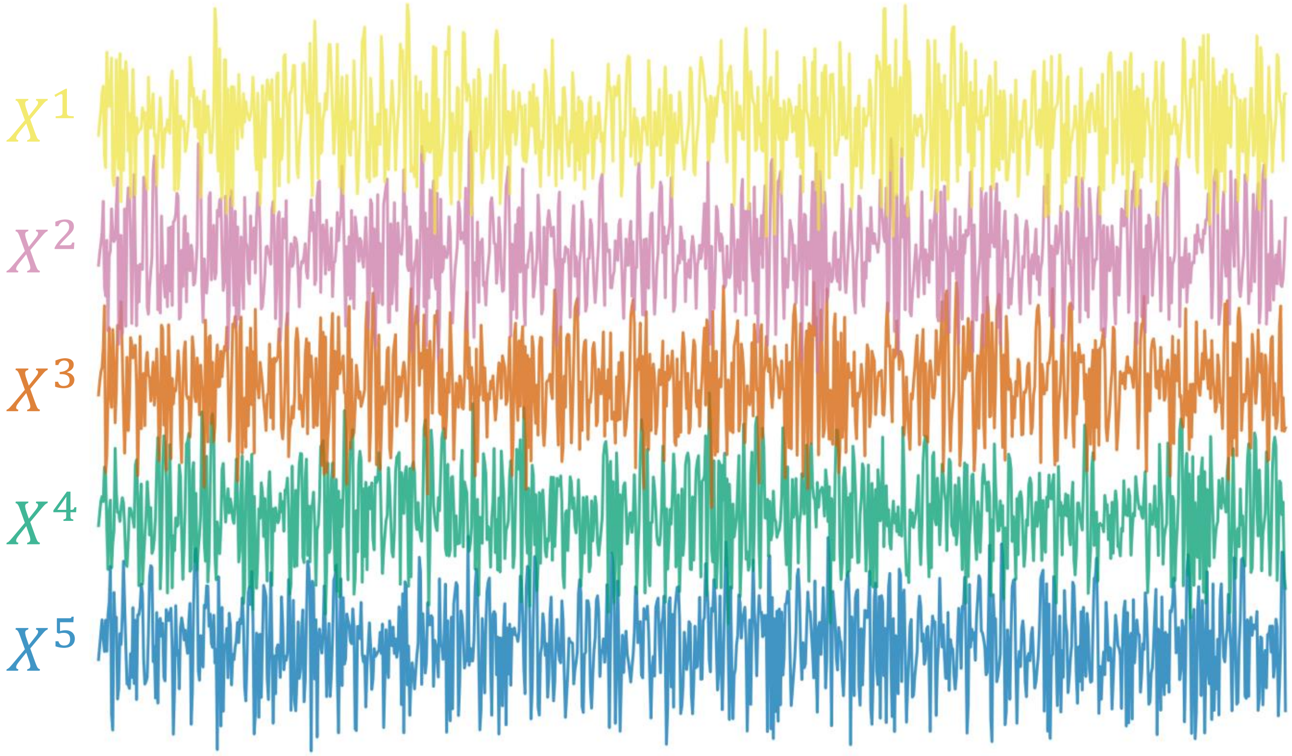
Time Series Analysis

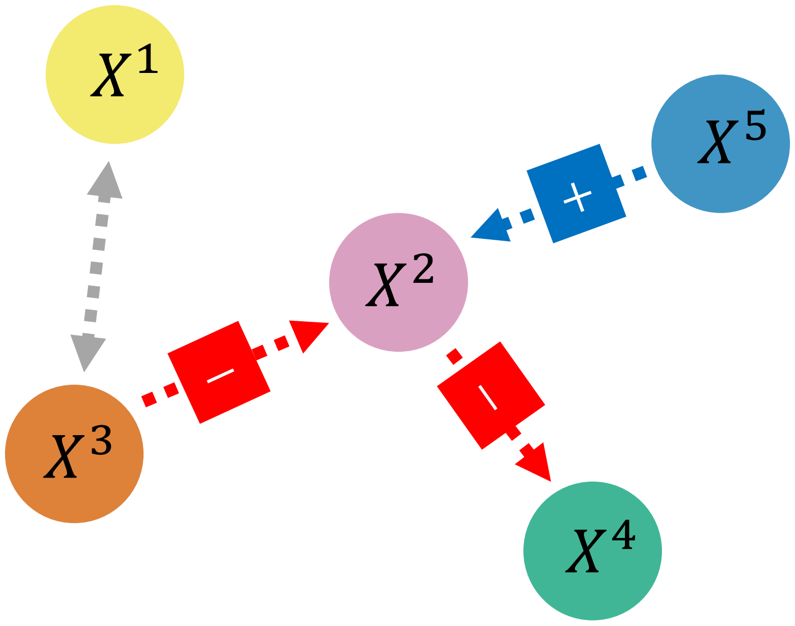
D INFK medical data science





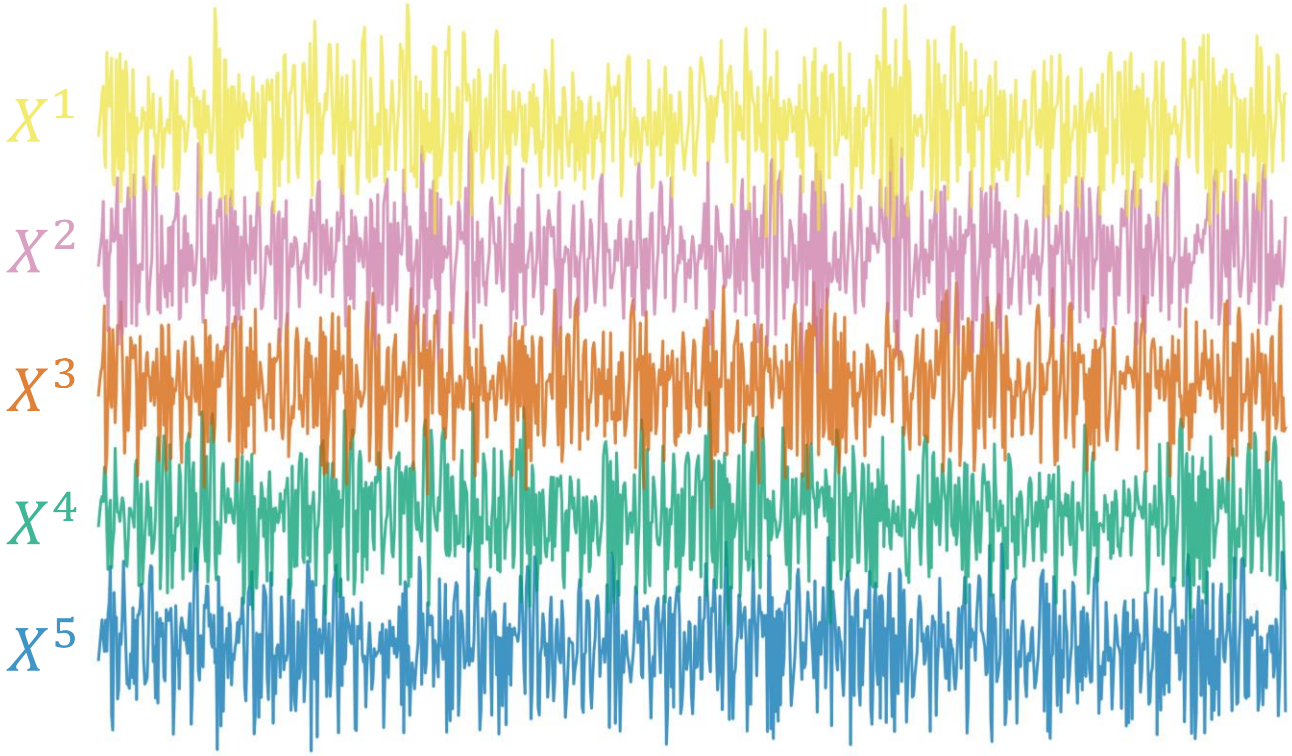
Time Series Analysis

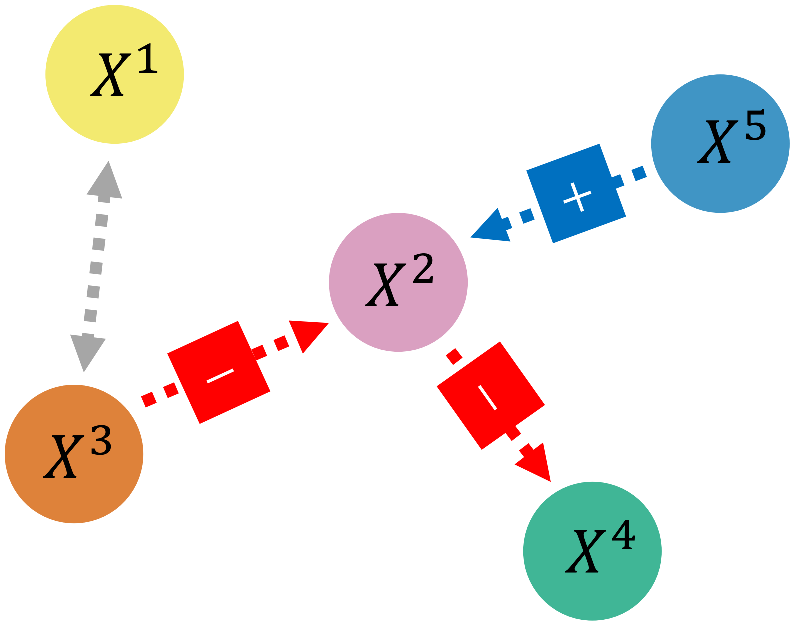




Time Series Analysis

D INFK medical data science





We propose an **interpretable model** for inferring **nonlinear relationships** in **multivariate time series**…

Granger Causality (GC)

X Granger-causes Y iff

$$\mathbb{P}[Y_{t+1} \in \mathcal{A} \mid \mathcal{I}(t)] \neq \mathbb{P}[Y_{t+1} \in \mathcal{A} \mid \mathcal{I}_{-X}(t)].$$

Granger Causality (GC)

D INFK medical data science

X Granger-causes Y iff

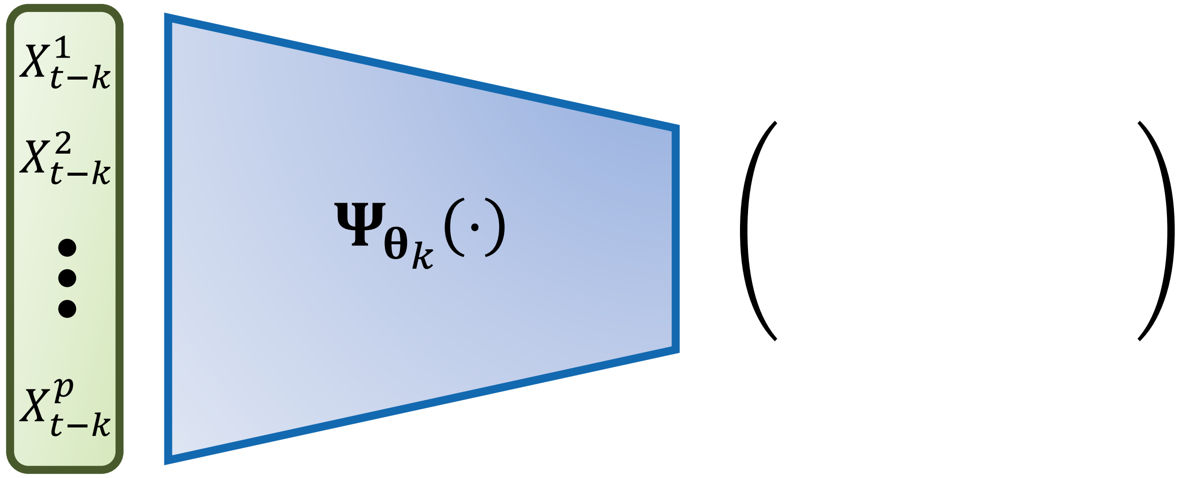
$$\mathbb{P}[Y_{t+1} \in \mathcal{A} \mid \underbrace{\mathcal{I}(t)}_{\text{past including } X_{<t+1}}] \neq \mathbb{P}[Y_{t+1} \in \mathcal{A} \mid \underbrace{\mathcal{I}_{-X}(t)}_{\text{past without } X_{<t+1}}].$$

VAR + Self-explaining Neural Networks = Generalised Vector Autoregression (GVAR)

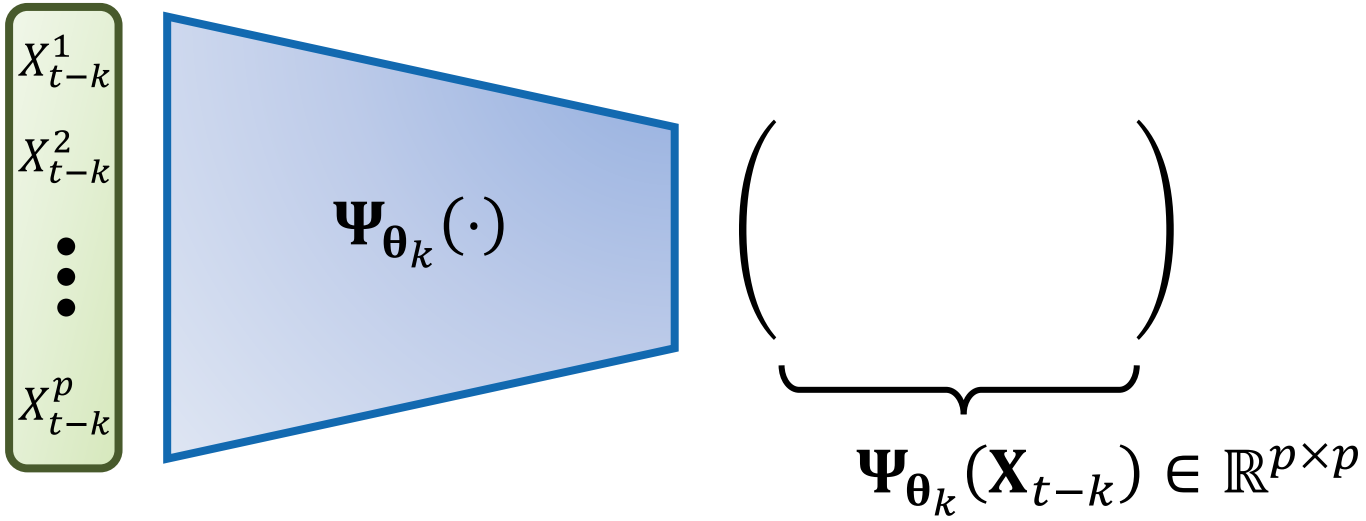
Generalised Vector Autoregression

$$\begin{pmatrix} X_{t-k}^{1} \\ X_{t-k}^{2} \\ \vdots \\ X_{t-k}^{p} \end{pmatrix}$$

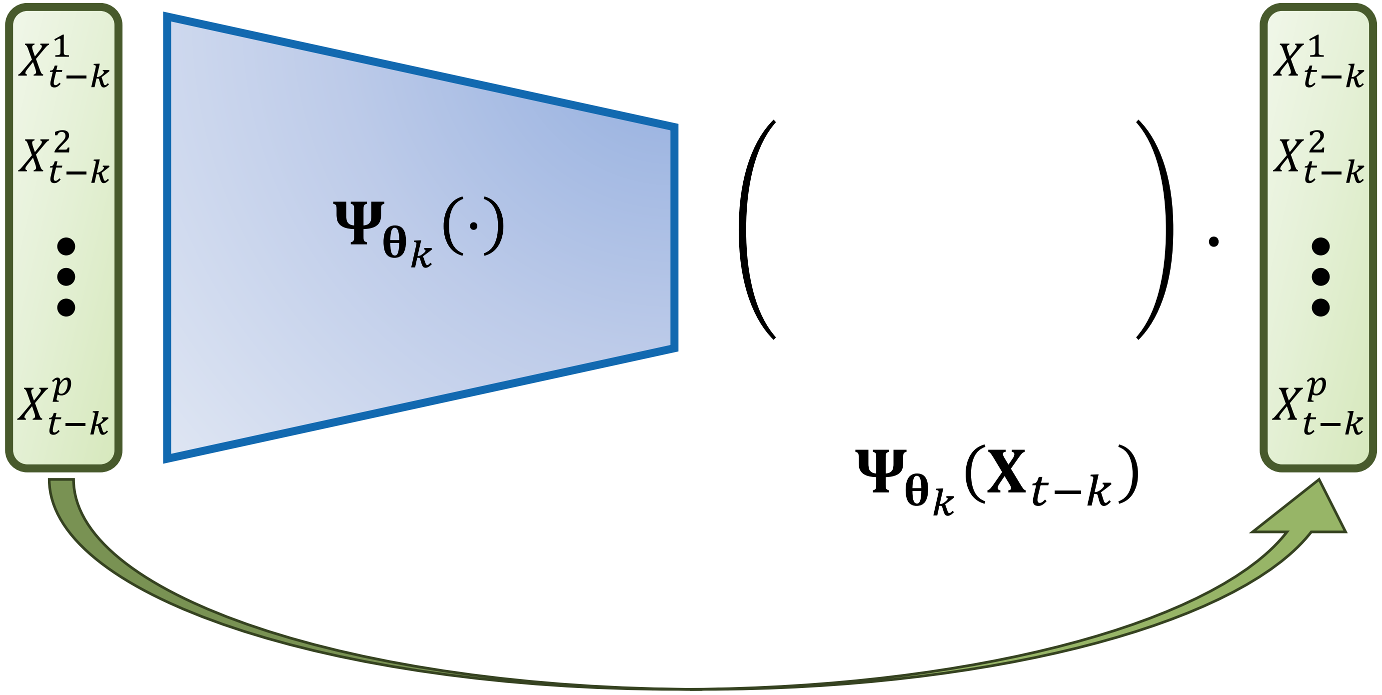
Generalised Vector Autoregression



Generalised Vector Autoregression

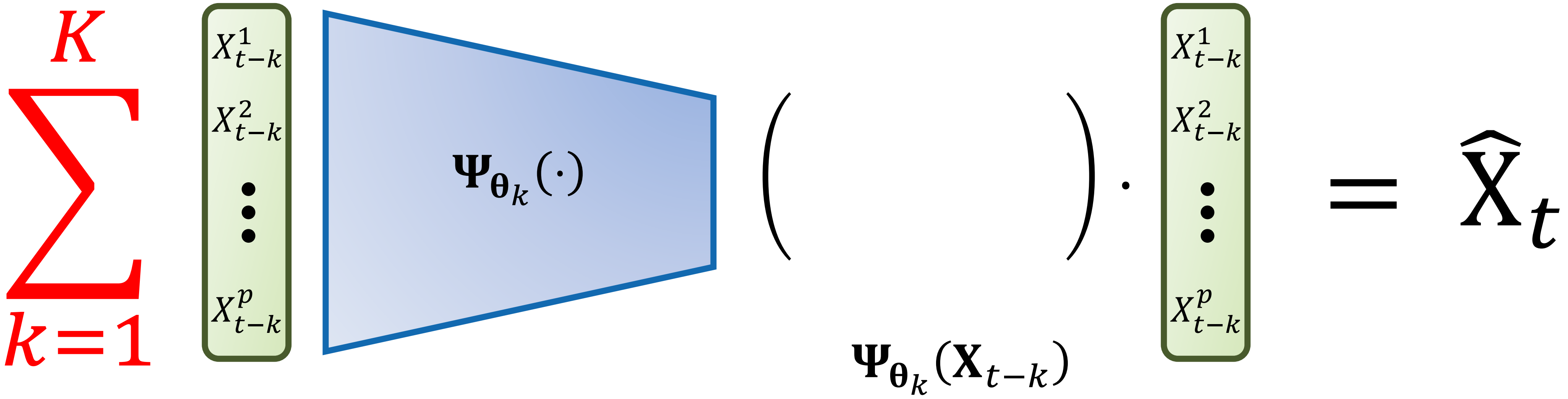


Generalised Vector Autoregression

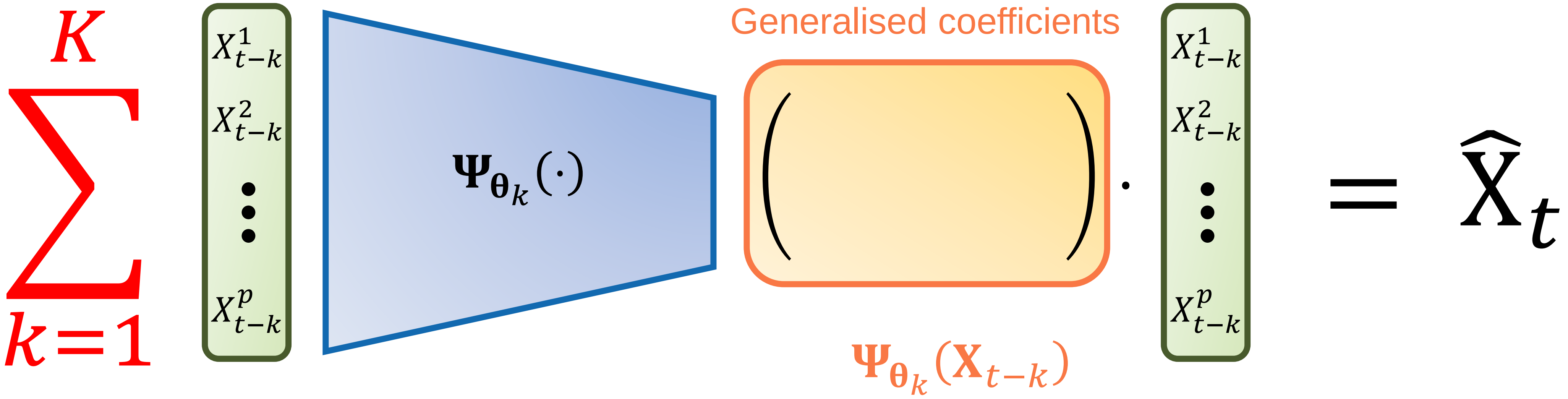


Generalised Vector Autoregression

D INFK medical data science



Generalised Vector Autoregression



Loss Function

$$\sum_{t=K+1}^{T} \ell(\mathbf{x}_t, \hat{\mathbf{x}}_t) + \lambda \sum_{t=K+1}^{T} R(\mathbf{\Psi}_t) + \gamma \sum_{t=K+1}^{T-1} \|\mathbf{\Psi}_{t+1} - \mathbf{\Psi}_t\|_2^2$$

Loss Function

$$\underbrace{\sum_{t=K+1}^{T} \ell(\mathbf{x}_t, \hat{\mathbf{x}}_t)}_{\text{forecasting loss}} + \lambda \sum_{t=K+1}^{T} R(\boldsymbol{\Psi}_t) + \gamma \sum_{t=K+1}^{T-1} \|\boldsymbol{\Psi}_{t+1} - \boldsymbol{\Psi}_t\|_2^2$$

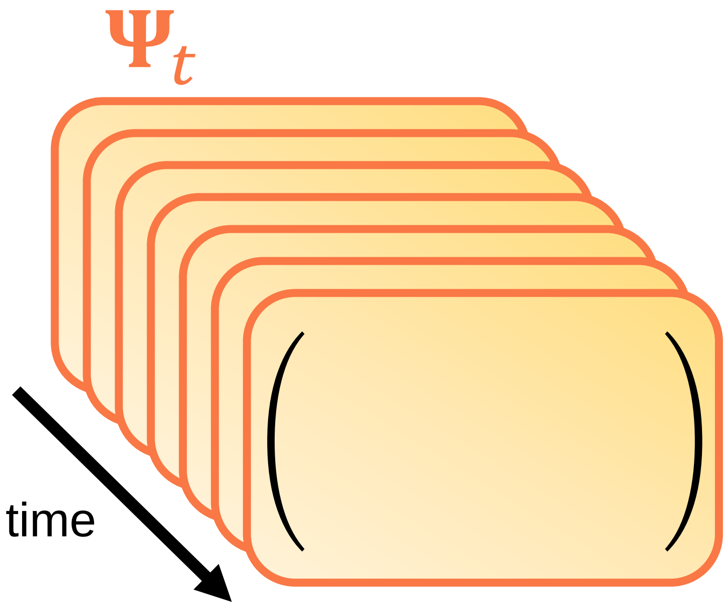
Loss Function

$$\underbrace{\sum_{t=K+1}^{T} \ell(\mathbf{x}_t, \hat{\mathbf{x}}_t)}_{\text{forecasting loss}} + \lambda \underbrace{\sum_{t=K+1}^{T} R(\boldsymbol{\Psi}_t)}_{\text{sparsity-inducing penalty}} + \gamma \sum_{t=K+1}^{T-1} \|\boldsymbol{\Psi}_{t+1} - \boldsymbol{\Psi}_t\|_2^2$$

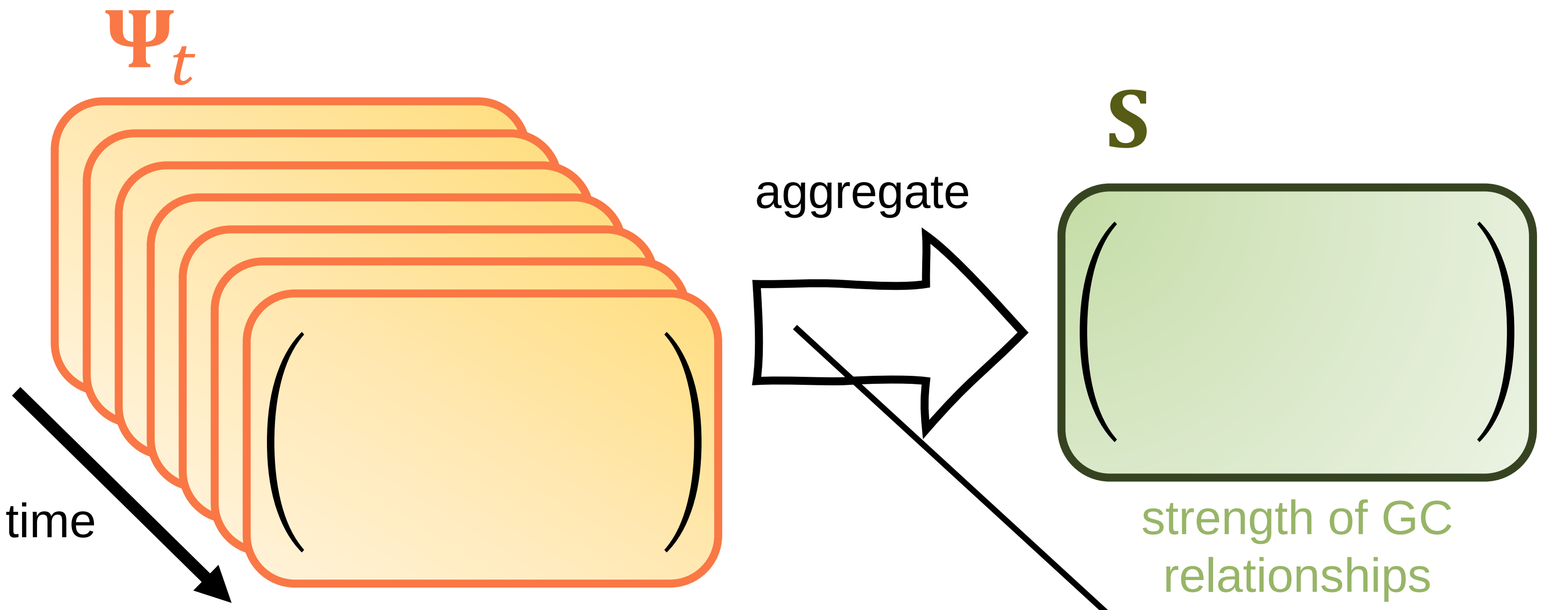
Loss Function

$$\underbrace{\sum_{t=K+1}^{T} \ell(\mathbf{x}_t, \hat{\mathbf{x}}_t)}_{\text{forecasting loss}} + \lambda \underbrace{\sum_{t=K+1}^{T} R(\mathbf{\Psi}_t)}_{\text{sparsity-inducing penalty}} + \gamma \underbrace{\sum_{t=K+1}^{T-1} \|\mathbf{\Psi}_{t+1} - \mathbf{\Psi}_t\|_2^2}_{\text{smoothing penalty}}$$

Inference Framework

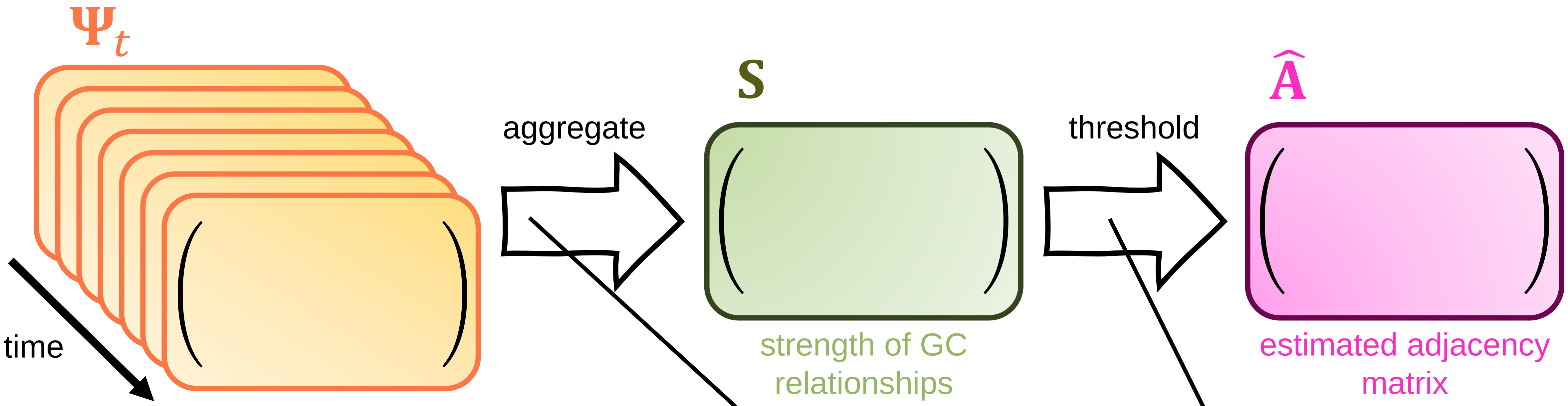


Inference Framework



$$\mathbf{S}_{i,j} = \max_{1\le k\le K}\left\{\mathrm{median}_{K+1\le t\le T}\left(\left|\left(\boldsymbol{\Psi}_{\hat{\boldsymbol{\theta}}_k}(\mathbf{x}_t)\right)_{i,j}\right|\right)\right\}$$

Inference Framework



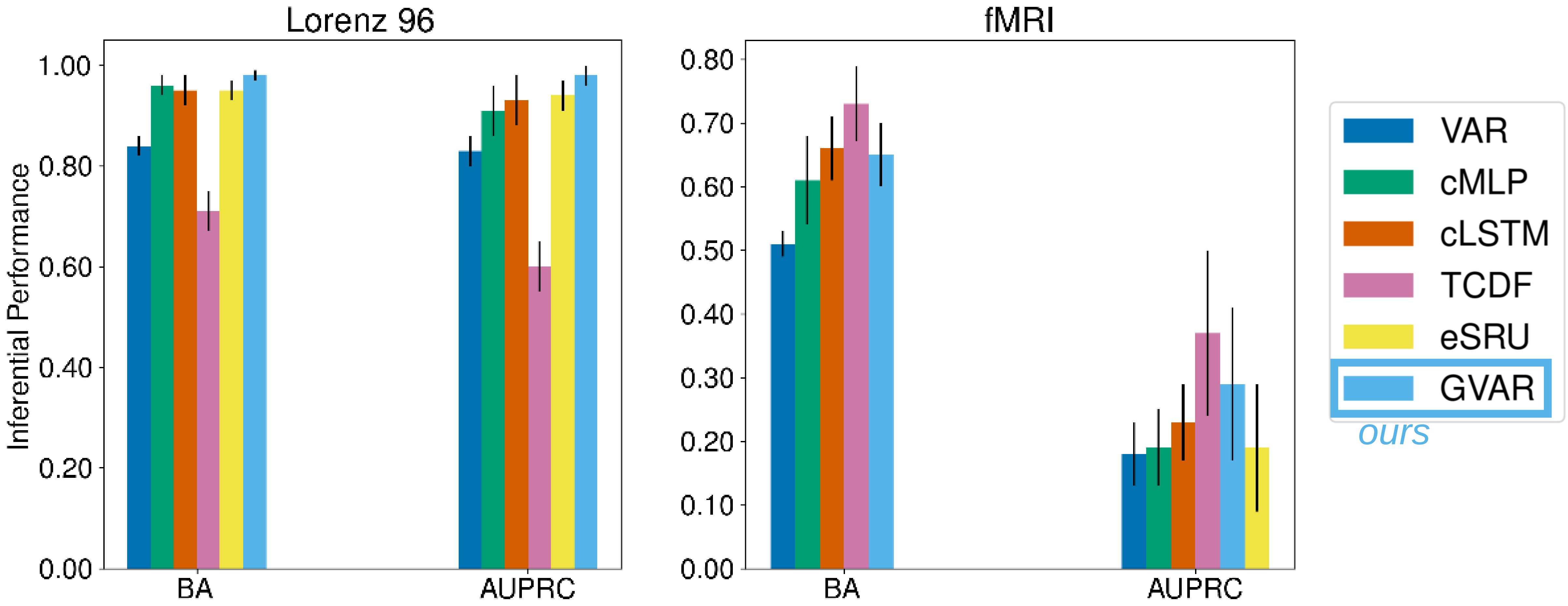
$$\mathbf{S}_{i,j} = \max_{1 \le k \le K} \left\{ \mathrm{median}_{K+1 \le t \le T} \left(\left| \left(\boldsymbol{\Psi}_{\hat{\boldsymbol{\theta}}_k}(\mathbf{x}_t) \right)_{i,j} \right| \right) \right\}$$

$$\hat{\mathbf{A}}_{i,j} = \tau\left(\mathbf{S}_{i,j}, \alpha\right)$$

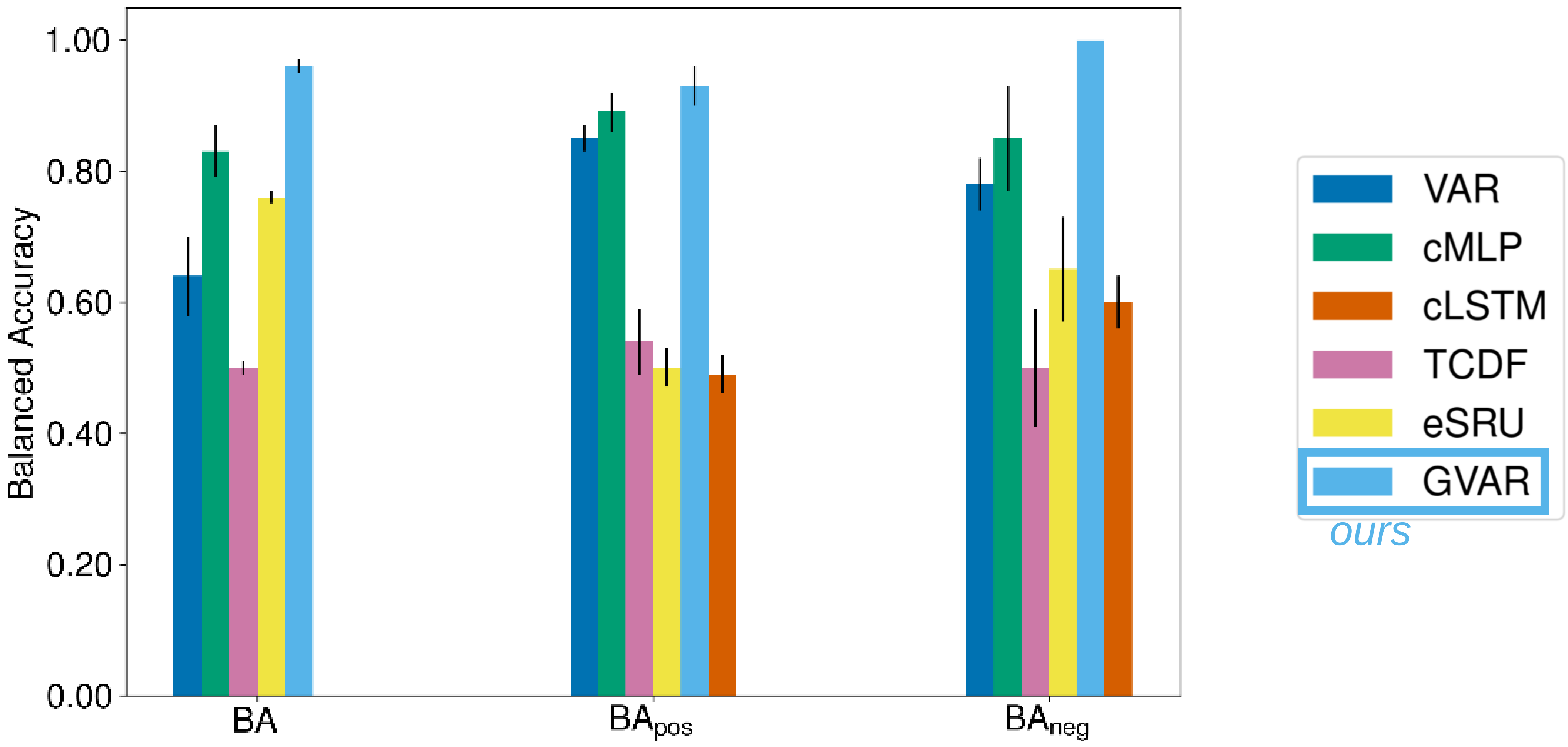
Empirical Results

Inferring Relationships

D INFK medical data science



Detecting Effect Signs



Conclusion

- Extension of self-explaining neural networks to autoregressive setting

Conclusion

- Extension of self-explaining neural networks to autoregressive setting
- Interpretable inference framework for nonlinear multivariate Granger causality

Conclusion

- Extension of self-explaining neural networks to autoregressive setting
- Interpretable inference framework for nonlinear multivariate Granger causality
- Problem of inferring effect signs

Conclusion

- Extension of self-explaining neural networks to autoregressive setting
- Interpretable inference framework for nonlinear multivariate Granger causality
- Problem of inferring effect signs
- Comprehensive experiments on simulated data

Conclusion

- Extension of self-explaining neural networks to autoregressive setting
- Interpretable inference framework for nonlinear multivariate Granger causality
- Problem of inferring effect signs
- Comprehensive experiments on simulated data

Paper: Code: