SCAN ME!

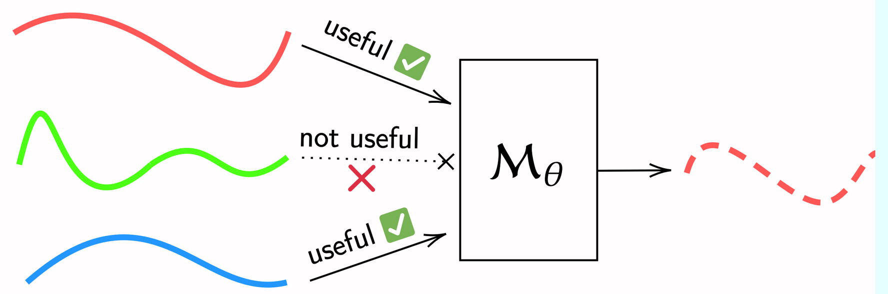
# Exploring Neural Granger Causality with xLSTMs: Unveiling Temporal Dependencies in Complex Data

**Harsh Poonia*, Felix Divo*, Kristian Kersting, Devendra Singh Dhami**

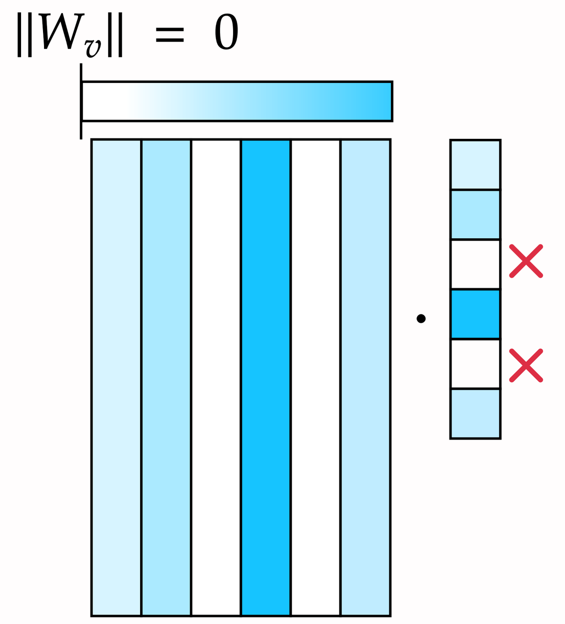
With funding from the: Federal Ministry of Research, Technology and Space

## Neural Granger Causality

- Problem: Find which time series U are good predictors of future values of time series V
- Using Neural Networks: Solve as a forecasting problem
- How do we measure model sensitivity to inputs?

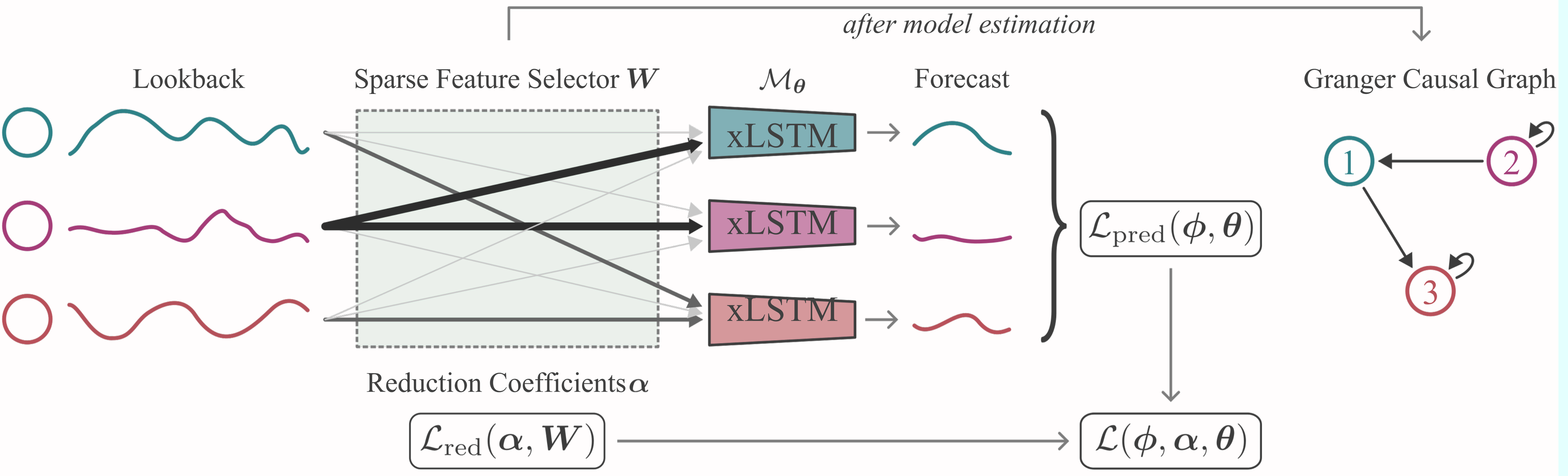



## Challenges




- Desire sparsity in selected variates for interpretability
- Project input features for processing, columns of projection matrix indicate contribution of variate

## Architecture




- Exponential gating for longer range dependencies
- Joint Optimization of smooth prediction objective + non-smooth sparsity objective


[1] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural Granger causality.
[2] Meike Nauta, Doina Bucur, and Christin Seifert. Causal Discovery with Attention-Based Convolutional Neural Networks.
[3] Saurabh Khanna and Vincent Y. F. Tan. Economy Statistical Recurrent Units For Inferring Nonlinear Granger Causality
[4] Ričards Marcinkevičs and Julia E. Vogt. Interpretable Models for Granger Causality Using Selfexplaining Neural Networks


Learned weights in Adaptive Group Lasso penalty coupled with exponential gating in LSTM lead to better generalizability across domains in Granger causal discovery (finding good statistical predictors) in multivariate time series.

| Ablation | Forecaster | Optimization | Lorenz | fMRI |
|---|---|---|---|---|
| GC-xLSTM | xLSTM | Joint | **96.6±0.3** | **73.3±3.0** |
| (I) | LSTM | Joint | 93.0±0.3 | 96.6±2.0 |
| (II) | xLSTM | Group Lasso | 73.0±4.6 | 65.4±2.0 |

## Joint Optimization

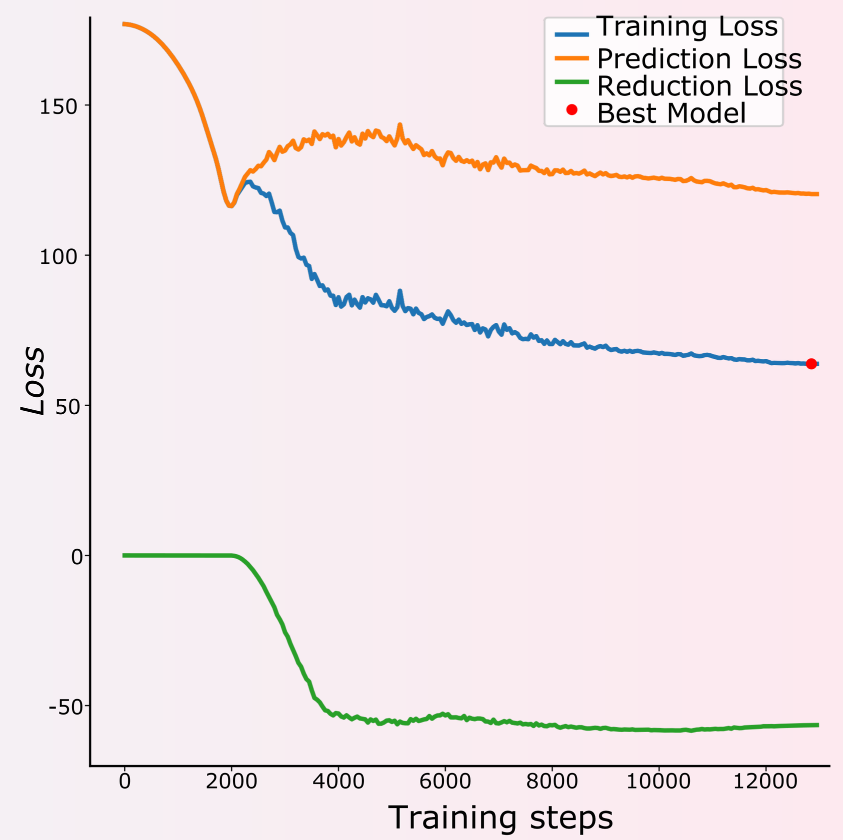
reduction coefficient update and model improvement — *gradient descent*

$$\begin{pmatrix}\phi\\ \alpha\\ \theta\end{pmatrix} \leftarrow \begin{pmatrix}\phi\\ \alpha\\ \theta\end{pmatrix} - \eta \begin{pmatrix}\nabla_\phi\\ \nabla_\alpha\\ \nabla_\theta\end{pmatrix} \mathcal{L}(S; \phi, \alpha, \theta)$$

$$W_v \leftarrow \left(1 - \frac{\lambda\eta\alpha_v}{\|W_v\|_2}\right)_+ W_v$$

*proximal gradient descent* — compress $W$

## No Auxiliary Metrics Needed!

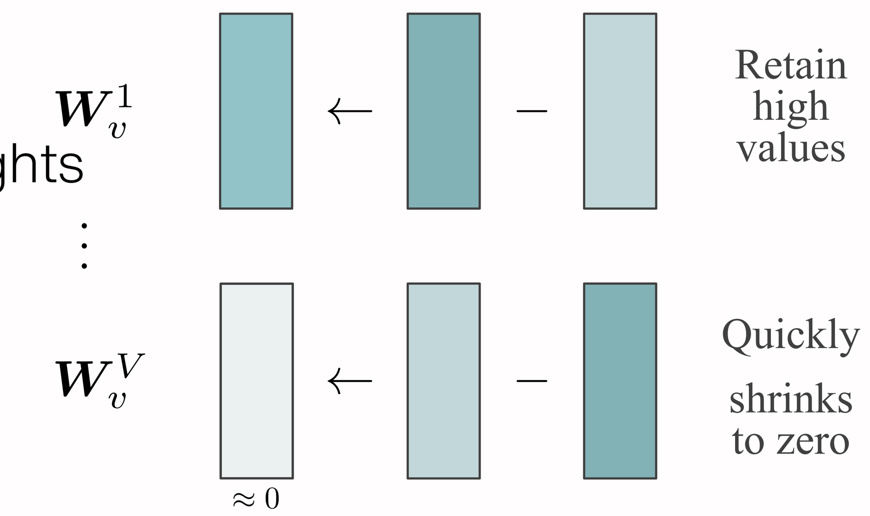



## Learned Adaptive Group Lasso

- Need to retain some weights and compress the others — standard group lasso doesn't!
- Proximal Gradient Update during compression step, use learned weights

$$\lambda \sum_{v=1}^{V} \alpha_v \|\mathbf{W}_v\|_2$$

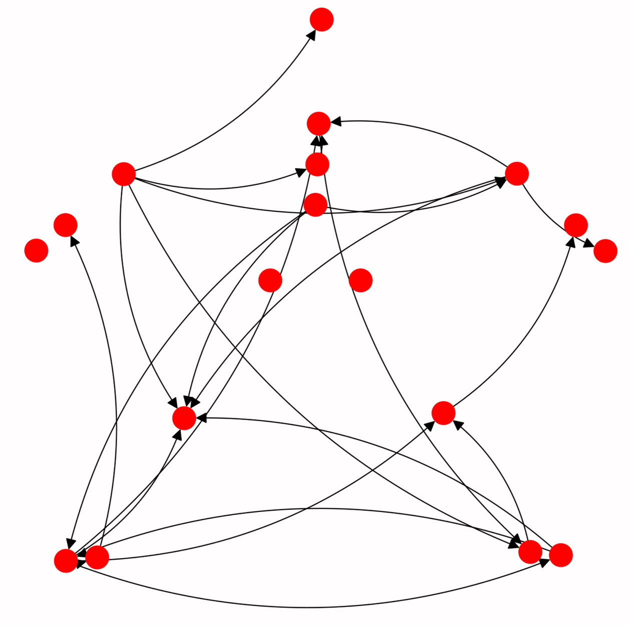
- Weight decay is proportional to $\alpha_v$, the "reduction coefficient"
- Soft thresholding → Group sparsity




## Results

| Model | Balanced Accuracy |
|---|---|
| TCDF[2] | 72.8 ±6.3 |
| GVAR[3] | 65.2 ±4.5 |
| cLSTM[1] | 65.5 ±5.3 |
| **GC-xLSTM** | **73.3 ±3.0** |

fMRI Brain Activity Dataset

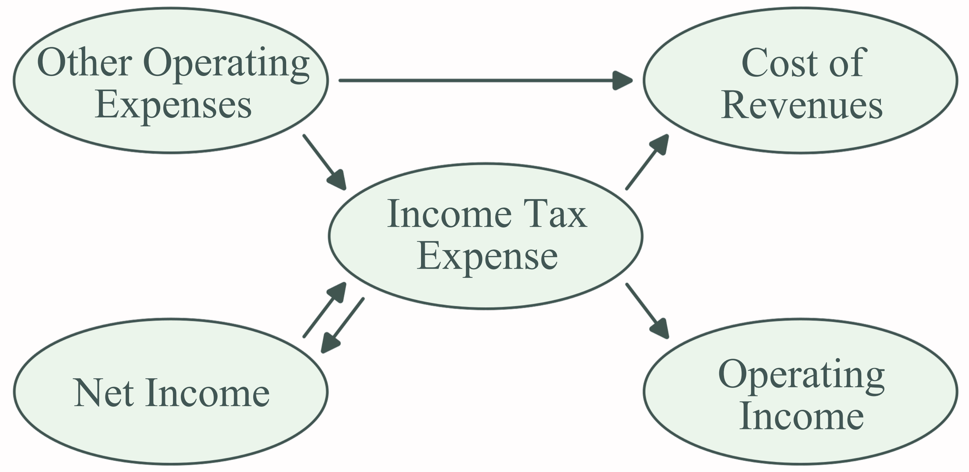


Joint + Limb Movements, Salsa Dance

Weather Stations in France

| Model | F = 10 | F = 40 |
|---|---|---|
| cLSTM[1] | 95.0 ± 2.8 | 65.6 ± 3.7 |
| TCDF[2] | 70.9 ± 4.4 | 62.2 ± 3.0 |
| eSRU[3] | 95.1 ± 2.0 | 88.6 ± 1.4 |
| GVAR[4] | 98.2 ± 0.6 | 88.5 ± 4.6 |
| **GC-xLSTM** | **96.6 ± 0.3** | **96.6 ± 0.3** |

Lorenz-96, Balanced Accuracies




Company Fundamentals