Supplementary Materials: Neural Granger Causality

Alex Tank*, *Member, IEEE,* Ian Covert*, *Member, IEEE,* Nick Foti, *Member, IEEE,* Ali Shojaie *Member, IEEE,* Emily B. Fox, *Member, IEEE,*

APPENDIX A
MODEL ABLATIONS

We ran two ablation studies to understand factors that influence our methods' performance. First, we tested the cMLP and cLSTM with different numbers of hidden units on the Lorenz-96 data. Table 1 shows the AUROC results from a single run for two datasets with forcing constants $F \in (10, 40)$ and time series length $T = 1000$, using different numbers of hidden units, $H \in (5, 10, 25, 50, 100)$. The results reveal that both models are robust to a small number of hidden units, but that their performance improves with larger values of H. These findings suggest that overparameterization can help with the nonconvex optimization objective, leading to solutions that achieve high predictive accuracy while minimizing the penalty from the sparsity-inducing regularizer.

Next, we tested three approaches for optimizing our penalized objectives (Equations 8 and 15). We compared standard gradient descent with Adam [1] to proximal gradient descent (ISTA) [2] and proximal gradient descent with a line search (GIST) [3] on the Lorenz-96 data with $T = 1000$ time points. Table 2 displays AUROC results across five initializations for two forcing constants $F \in (10, 40)$, using the cMLP with $H = 10$ hidden units. The results show that the three methods lead to similar results for both $F = 10$ and $F = 40$, although we did not compare the optimizers in other scenarios, e.g., with lower T values or with the cLSTM.

Among these optimization approaches, Adam is fastest due to its adaptive learning rate, but it requires a parameter for thresholding the resulting weights (while the proximal methods lead to exact zeros). In contrast, GIST guarantees convergence to a local minimum and is less sensitive to the learning rate parameter, but it is also considerably slower than Adam and ISTA. We therefore use standard proximal gradient descent, or ISTA, in the remainder of our experiments, because it leads to exact zeros while being more efficient than GIST. In practice, this means running Algorithm 1 or Algorithm 4 using a fixed learning rate γ rather than determining it by a line search.

TABLE 1
AUROC comparisons for the cMLP and cLSTM as a function of the number of hidden units H for simulated Lorenz-96 data. Results are calculated using a single run.

Model	cMLP		cLSTM	
F	10	40	10	40
H = 5	96.5	91.0	91.9	86.9
H = 10	98.0	94.0	94.5	91.5
H = 25	98.4	94.3	95.6	92.3
H = 50	98.3	94.4	95.7	93.8
H = 100	98.5	94.5	95.7	95.2

APPENDIX B
BASELINE METHODS

The IMV-LSTM uses an attention mechanism to highlight the model's dependence on different parts of the input [4]. We train a separate IMV-LSTM model to predict each time series using all the time series as inputs, using the "IMV-Full" variant [4], and we use the attention weights from the trained models to infer Granger causal relationships. Similar to the original work [5], we record the empirical mean of the attention values for each input time series for each model, and we construct a $p \times p$ matrix of these values for the separate IMV-LSTMs. We then sweep over a range of threshold values to determine the most influential inputs for each IMV-LSTM, and we trace out an ROC curve from which we calculate AUROC values.

The LOO-LSTM baseline is based on the idea that withholding a highly predictive input should result in a decrease in predictive accuracy, a direction that has been explored for providing model-agnostic notions of feature importance [6], [7]. We begin by training separate LSTM models to predict each time series using all time series as inputs. We then train separate LSTM models to predict each time series i using all inputs except time series j, and we record the increase in loss when the jth time series is withheld. Using the results, we construct a $p \times p$ matrix representing the differences in the loss, we sweep over a range of threshold values to determine the most influential inputs for each time series,

- * *Denotes equal contribution.*
- *Alex Tank was with the Department of Statistics, University of Washington, Seattle, WA, 98103. E-mail: alextank@uw.edu*
- *Ian Covert, Nicholas Foti, and Emily Fox were with the Department of Computer Science, University of Washington, Seattle, WA, 98103.*
- *Ali Shojaie was with the Department of Biostatistics, University of Washington, Seattle, WA, 98103*

Manuscript received April 19, 2005; revised August 26, 2015.

TABLE 2
AUROC comparisons between different optimization approaches for the cMLP with simulated Lorenz-96 data. Results are the mean across five initializations, with 95% confidence intervals.

F	10	40
GISTA	98.0 ± 0.2	93.8 ± 0.3
ISTA	98.0 ± 0.2	94.1 ± 1.9
Adam	98.3 ± 0.1	95.1 ± 0.2

and we trace out an ROC curve from which we calculate AUROC values.

REFERENCES

[1] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[2] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.
[3] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems," in *International Conference on Machine Learning*. PMLR, 2013, pp. 37–45.
[4] T. Guo, T. Lin, and N. Antulov-Fantulin, "Exploring interpretable LSTM neural networks over multi-variable data," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2494–2504.
[5] T. Guo, T. Lin, and Y. Lu, "An interpretable LSTM neural network for autoregressive exogenous model," *arXiv preprint arXiv:1804.05251*, 2018.
[6] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.
[7] G. Hooker and L. Mentch, "Please stop permuting features: An explanation and alternatives," *arXiv preprint arXiv:1905.03151*, 2019.