Rethinking Neural Relational Inference for Granger Causal Discovery

Anonymous Author(s)
Affiliation
Address
email

Abstract

Granger causal discovery aims to infer the underlying Granger causal relationships between pairs of variables in a multivariate time series system. Recent work has proposed using Neural Relational Inference (NRI) Kipf et al. [2018] – a latent graph inference model – for Granger causal discovery. However, the conditions under which NRI succeeds in recovering the true Granger causal graph remain unknown. In this work we show how the mean field approximation inherent in NRI has significant implications for its ability to recover the Granger causal structure in multivariate time series. We illustrate this point theoretically and experimentally using a linear vector autoregressive model – an important benchmark in economic and financial studies.

1 Introduction

Granger causal discovery is a widely studied problem with real world applications in a number of fields such as neuroscience [Sporns, 2016], genetics [Fujita et al., 2010] and finance [Campbell et al., 1998]. Given an observed multivariate time series dataset, Granger causal discovery aims to infer the underlying Granger causal relationships between pairs of time series. In recent years, deep learning methods [Tank et al., 2021] have been proposed for Granger causal discovery which aim to provide more flexibility over traditional approaches [Granger, 1969] to Granger causal discovery. In this work, we examine the novel approach of Löwe et al. [2022] on a simple benchmark dataset.

Löwe et al. [2022] propose a method named Amortized Causal Discovery (ACD) which infers Granger causal relationships using Neural Relational Inference (NRI) [Kipf et al., 2018]. The Granger causal structure of a multivariate time series system can be viewed as a directed graph in which nodes represent time series and edges represent Granger causal relations. NRI is a graph-based variational autoencoder which infers latent edge relations. This model is an exciting proposal for Granger causal discovery as it fully inductive, which means that is able to be applied across a number of multivariate time series samples, and achieves competitive performance on several benchmark datasets.

However, the conditions under which ACD succeeds in recovering the Granger causal graph underlying a multivariate time series system are not known. This creates a hurdle to its use on real world data as there are no theoretical guarantees or strong experimental results to suggest that it can recover the true causal structure in any given application.

In this work, we examine ACD in the context of a vector autoregressive (VAR) data generating process with graph structure. VAR is a synthetic benchmark that is commonly used in financial and economic domains. Since Granger causal discovery is a widely studied question and certain problems can be modeled using graph-based approaches [Knight et al., 2020], ACD is a potentially attractive method for the domains of economics and finance. Therefore, it is of interest to understand the performance of ACD on a simple VAR benchmark prior to applying it to real world data.

Submitted to the NeurIPS Causal ML for Real-World Impact Workshop (CML4Impact 2022). Do not distribute.

In the rest of the paper, we start by briefly describing the task of Granger causal discovery and the NRI approach that will be the object of study. Then, we introduce the graph-based VAR benchmark. By comparing the true posterior distribution of the edge relations with the form of the GNN encoder used in NRI, we then argue that the mean-field approximation inherent in NRI poses a fundamental limitation on the model for Granger causality discovery. We construct an indicator that theoretically predicts the specific graph structures and edges that NRI will struggle to infer in this simple VAR setting. This indicator is then validated using synthetic data experiments. We conclude by recapitulating the key limitation of NRI in its application to Granger causality discovery. We hope that our work will guide future causal ML research in a productive direction.

Our primary contributions can be summarized as follows:

1. We provide the first study to understand the conditions under which NRI can successfully recover the Granger causal structure of a multivariate time series system.
2. We propose an indicator that predicts when NRI will fail to recover the Granger causal structure on a linear graph autoregressive process.
3. We empirically validate our indicator using synthetic data experiments[1].

Background on Granger causality

Let $X_t^i \in \mathbb{R}$ denote the value of time series $i = 1, \ldots, N$ at time t. Then time series X^i Granger causes time series X^j if, for some t,

$$\mathbb{P}[X_{t+1}^j \in S | \mathcal{I}(t)] \neq \mathbb{P}[X_{t+1}^j \in S | \mathcal{I}_{-X^i}(t)] \tag{1}$$

where S is an arbitrary non-empty set [Eichler, 2012]. The sets $\mathcal{I}(t)$ and $\mathcal{I}_{-X^i}(t)$ respectively denote all information available as of time t and all information available as of time t excluding time series X^i. Granger causal discovery aims to infer the Granger causal relation between each pair of time series in a multivariate time series dataset. While an inferred Granger causal relation is not necessarily indicative of a true causal relation Maziarz [2015], it remains a popular framework for understanding temporal relations in multivariate dynamical systems.

In the case of linear VAR modeling, the Granger causal relationship between two random variables can be tested through the hypothesis that the coefficients relating the lagged values of time series i to the current value of time series j are jointly significantly different to 0. Granger causal discovery can be accomplished when N is small through an exhaustive hypothesis test search on the coefficients in the VAR model.

Background on Neural Relational Inference for Granger causal discovery

In ACD, the proposed method of Löwe et al. [2022], an encoder-decoder latent variable model known as NRI Kipf et al. [2018] is used to infer Granger causal relations. The NRI encoder – which takes the form of a Graph Neural Network (GNN) learns to approximate the posterior distribution over latent graph relations; it accomplishes this by propagating information across a fully connected graph in which each node corresponds to a variable in the multivariate system and a node feature embeds its respective variable's time series. The NRI encoder models the posterior probability for z_{ij} the type (category) of edge $i \to j$ with

$$\psi_{ij} = f_{\text{enc}}(\mathbf{X})_{ij}, \quad q(z_{ij}|\mathbf{X}) = \text{Softmax}(\psi_{ij}/\tau) \tag{2}$$

where f_{enc} is a GNN encoder and τ is a temperature parameter for the Softmax activation function. The variable ψ_{ij} is K-dimensional, where K is the number of edge categories chosen by the user. We write $\mathbf{X}$ to denote the dataset $\{X_t^i\}_{t=1,\ldots,T}^{i=1,\ldots,N}$.

The NRI decoder – another GNN – models the multivariate time series conditional on the graph and latent edge relations returned by the encoder. Each edge relation type coincides to a unique message passing function in the decoder. ACD uses the same neural architecture as NRI with the addition of a

[1] Code used in experiments can be found here `https://anonymous.4open.science/r/nri_granger_causality_neurips_2022-DE28/`

single "zero" edge type: this edge type has its decoder message passing function hard-coded to return zero. An inferred zero-edge implies no Granger causation between two variables. Inferred edges that are not of a zero-type imply a Granger causal relationship between two variables.

In contrast to existing methods, ACD has an fully inductive encoder for the Granger causal graph. This means that a trained model can be applied to a out-of-sample multivariate time series dataset with a different Granger causal structure but the same shared dynamics conditional on the Granger causal graph. The inductive nature of the model means that it can leverage the information that is shared across multivariate time series samples. Its potential ability to learn across multivariate time series samples and competitive performance on three synthetic datasets considered by Löwe et al. [2022] makes ACD an exciting model for Granger causal discovery.

2 Theoretical limitations of NRI for Granger causal discovery

Consider the following generative process for a multivariate time series of size T with N variables:

$$X_t = cA^T X_{t-1} + \epsilon_t \tag{3}$$

where $X_t = \left(X_t^1, \ldots, X_t^N\right)^T$, $\epsilon_t \in \mathbb{R}^N$, $X_0 := 0$ and $\epsilon_t^i \sim N(0,1)$ independently for $t = 1, \ldots, T$, $i = 1, \ldots, N$. The matrix $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix corresponding to the Granger causal graph and c is a constant which determines the signal-to-noise ratio of the problem. The adjacency matrix is prespecified for each experiment. If $A_{ij} = 1$ then time series i Granger causes time series j, otherwise if $A_{ij} = 0$, there is no causal relation. We set the diagonal elements of A to 1 so that there is an autoregressive dependence for each time series. This model is similar to the Generalized Network Autoregressive Process of Knight et al. [2020].

Under a Bayesian model in which the entries of A are distributed independently at random with uniform probability p of being 1 under the prior, the log-posterior distribution of entry A_{ij} conditional on the observed dataset and the other entries of A is given by, up to a constant in A_{ij},

$$\begin{aligned}
&\log \mathbb{P}\left(A_{ij} | \{X_t\}_{t=1,\ldots,T}, \{A_{kl}\}_{(k,l)\neq(i,j)}\right) = \\
&\frac{cT}{2} A_{ij} \Bigg[(1-c)\mathbb{C}(i,j) - c\mathbb{C}(j,i) - c\mathbb{C}(i,i) \\
&- c \sum_{k \in \{1,\ldots,N\} \setminus \{i,j\}} \left(\mathbb{C}(k,i) + \mathbb{C}(i,k)\right) A_{kj} \Bigg] \\
&+ A_{ij} \log p + (1 - A_{ij}) \log(1-p)
\end{aligned} \tag{4}$$

where $\mathbb{C}(i,j) = \frac{1}{T} \sum_{t=1}^{T} x_{t-1}^i x_t^j$ is the empirical lagged covariance between time series i and j.

Under the NRI model, the posterior is approximated using a mean-field approximation where the marginal posterior probability for A_{ij} is given by the result of the GNN encoder, $q(z_{ij}|\mathbf{x})$ in equation 2. The two-dimensional variable z_{ij} corresponds to A_{ij} in this case as the ground truth data generating process has $K = 2$ edge types (either present or absent). Note that by definition, under the mean-field posterior approximation, A_{ij} is independent of the other edges. However, in the true posterior, A_{ij} is potentially dependent on edges $A_{kj}, k \in \{1, \ldots, N\} \setminus \{i,j\}$. Indeed, there will be significant negative posterior correlation between A_{ij} and A_{kj} whenever $\mathbb{C}(k,i) + \mathbb{C}(i,k)$ is large. Intuitively, if variables i and k are correlated across time, then the true posterior distribution will place negative correlation on the edges originating from either node and having a common target node (if variable k predicts j, then it less likely that i also predicts j). However, the GNN encoder is unable to capture such posterior dependence between edges.

As a result, we expect NRI to incorrectly classify the edge $i \to j$ whenever X_k and X_i have significant lagged correlation and $k \to j$ but there is no edge $i \to j$ in the true underlying graph. Since the encoder is unable to capture the negative correlation of A_{ij} and A_{kj}, it will tend to overestimate the probability of $A_{ij} = 1$ and therefore misclassify edge $i \to j$. In order to validate this hypothesis concerning the limitations of NRI to estimate the Granger relations in certain graph structures, we

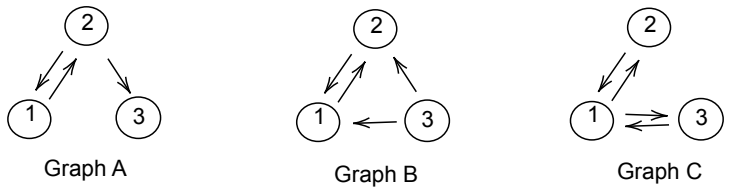

Figure 1: The three Granger causal graph structures used in experiments.

construct a "difficulty indicator" D_{ij} for each edge:

$$D_{ij} = \frac{|C_{ij}|}{|C_{ij}|+|M_{ij}|} \tag{5}$$

where

$$C_{ij} = \sum_{k \in \{1,\ldots,N\} \setminus \{i,j\}} \left(\mathbb{C}(k,i) + \mathbb{C}(i,k)\right) A^*_{kj} \tag{6}$$

$$M_{ij} = (1-c)\mathbb{C}(i,j) - c\mathbb{C}(j,i) - c\mathbb{C}(i,i). \tag{7}$$

Here, A^* denotes the ground truth adjacency matrix. The difficulty indicator gives the size $|C_{ij}|$ of the true posterior dependence of A_{ij} on other edges A_{kj} relative to the size $|M_{ij}|$ of the component of the posterior edge probability in equation 4 which does not depend on other edges. Note that we ignore the prior term $A_{ij} \log p + (1 - A_{ij}) \log(1 - p)$ in equation 4 as this will become insignificant as T increases. Based on the argument given in the previous paragraph, we expect that edges with large difficulty indicators will be misclassified by NRI. We investigate this hypothesis empirically in the following experimental section.

3 Experiments

3.1 Synthetic data generation

We generate multivariate time series datasets of size $T = 1000$, $N = 3$ using equation 3. We consider three causal graph structures. These are illustrated in Figure 1.

We note that using a suite of t-tests on the estimated coefficients of an ordinary least squares fit of a VAR model with 1 lag to datasets of this size results in 100% classification accuracy for the entries in the VAR autoregressive matrix. This illustrates that the task of causal graph discovery is solvable on these three graph structures.

The learning paradigm of ACD trains NRI using samples of multivariate time series datasets [Löwe et al., 2022]. We match this learning paradigm and generate 1000 different training, validation and test multivariate time series for each graph structure. The specifics of model training are chosen to match those of Löwe et al. [2022] and are found in the Appendix.

The performance of ACD in recovering the Granger causal structure is evaluated using the classification accuracy of the encoder on the off-diagonal entries of the adjacency matrix on test multivariate time series data. In addition, we also record the predictive performance of each method using mean squared error (MSE) averaged across each time series variable. This is expressed relative to the performance of the Bayes rule predictor in each experimental setting (RelMSE). The Bayes rule predictor will have an MSE error of 1 or RelMSE of 0%.

3.2 NRI model variants

The following encoders will be used in experiments:

- **RefMLP**: the encoder used in ACD [Löwe et al., 2022]. This is the standard MLP encoder used in Kipf et al. [2018].
- **Unshared**: this is a transductive encoder that has a unique parameter vector for each edge which gives the log-probabilities of that edge being in either category (present or absent). Since it is a transductive rather than inductive encoder, it is not a "true" NRI model variant, however, we train it using the same variational inference procedure as NRI.

The following decoder will be used in experiments:

- **Linear**: this decoder consists of a single round of linear message passing. The message passing function applies a linear map (with no additive constant term) to each of the sender node features (the current time series value). Since this is the correctly specified decoder functional form for the VAR data generating process 3, any potential shortcomings of the NRI model on the experimental benchmarks cannot be due to the decoder misspecification.
- **GNN**: In results not shown, we have also validated our hypothesis using the GNN decoder used in ACD [Löwe et al., 2022].

The value of the autoregressive constant c in equation 3 is chosen so that, for each graph used, the largest eigenvalue of the re-scaled adjacency matrix cA is equal to 0.9. Details of the model and training hyperparameters which were used are given in the Appendix.

3.3 Results

We report the performance of the NRI model variants on each of three causal graph structures in table 1.

Graph	Encoder	RelMSE (%)	Edge Acc (%)	0-Edge Acc (%)	1-Edge Acc (%)
A	RefMLP	11	50	0	100
	Unshared	0.1	100	100	100
B	RefMLP	0.1	99.7	99.6	99.7
	Unshared	0.1	100	100	100
C	RefMLP	11.6	66.7	0	100
	Unshared	0.2	100	100	100

Table 1: NRI prediction loss and graph recovery classification accuracy on test samples. The prediction loss is expressed in RelMSE. “Edge Acc”, ”0-Edge Acc” and ”1-Edge Acc” respectively give the encoder’s classification accuracy on the all of the test edges, accuracy on the test edges with ground truth 0 type (edge absent) and accuracy on the test edges with ground truth 1 type (edge present)).

We see that the performance of NRI using the RefMLP encoder varies across the three graph structures. In particular, it achieves perfect edge recovery on Graph B and achieves poor edge recovery on Graphs A and C. On the contrary, the Unshared encoder achieves perfect edge recovery on all three graphs. This shows that more a typical transductive encoding approach which infers the Granger causal structure using variable selection learned through variational inference can work well on the problems considered. Further, the strong performance of the Unshared encoder suggests that the poor performance of the NRI RefMLP encoder is not due to latent variable non-identifiability [Wang et al., 2021].

The inconsistent performance of the NRI model across the three different graph types can be explained by examining the misclassified edges in further detail. We display the misclassification rates in the Granger causal recovery problem for each edge alongside their difficulty indicators in Figure 2. From Figure 2, we observe that the misclassified edges exactly align with those predicted by the difficulty indicator. Since the difficulty indicator is constructed to identify edges with high relative posterior edge dependence effects, this validates our hypothesis that the NRI model is unable to capture specific Granger causal relations due to its mean-field approximation.

While we illustrate our theoretical argument with 3 graphs, we have experimentally tested our hypothesis on all directed graphs on 3 vertices. These further experiments agree with the illustrative cases. Results on these additional Granger causal structures are not shown due to space limitations. The theoretical argument presented in section 2 applies to linear VAR data generating processes in the general case with N variables. While we have performed experiments on three variables, we expect that these theoretical limitations of NRI will manifest empirically in experiments with a larger number of variables. Indeed, from equation 4, we see that posterior dependence effects between edges will exist so long as there are pairwise lagged correlation effects between triplets of variables in the multivariate time series system. On more complex real-world datasets and models for which the posterior is not analytically tractable, we do not know whether there exists significant

Graph A:

$$\begin{pmatrix} 0 & 0 & \mathbf{0.72} \\ 0 & 0 & 0 \\ \mathbf{0.66} & \mathbf{0.59} & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & \mathbf{100} \\ 0 & 0 & 0 \\ \mathbf{100} & \mathbf{100} & 0 \end{pmatrix}$$

Graph B:

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Graph C:

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \mathbf{0.83} \\ 0 & \mathbf{0.83} & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \mathbf{100} \\ 0 & \mathbf{100} & 0 \end{pmatrix}$$

Figure 2: For each Granger causal graph used in the experiments, we display the difficulty indicators for its edges (left column) alongside the test-time misclassification rates of the NRI RefMLP encoder on each of its edges (expressed as in percentage terms in the right hand side column). Entry ij in each matrix refers to the directed edge from node i to node j.

posterior dependence between variables without numerically simulating the posterior – this would be computationally costly and negate the purpose of using a variational approximation such as NRI. As a result, the arguments of this paper imply that it should be used on real-world Granger causal discovery problems with caution. In order to improve the performance of the NRI model for Granger causal discovery, its mean field variational approximation ought to be relaxed. Auto-regressive latent models that have been used for causal induction provide a more flexible posterior approximation Ke et al. [2022].

4 Conclusion

By theoretical and experimental arguments, we have shown that the mean-field posterior approximation inherent in NRI poses a challenge for its application to Granger causal discovery. Our work is limited to the analysis of a single graph-based data generating process. On more complex data generating processes, the performance of ACD may be inhibited by some aspect other than failure to capture posterior edge dependence. For instance, in cases with low posterior edge dependence, the NRI encoder architecture's ability to approximate the marginal edge posterior distribution may be its limiting factor in Granger causal discovery. Understanding other cases in which ACD fails to recover the ground truth graph is an interesting direction of further work. Nevertheless, our work draws attention to a fundamental limitation of ACD in Granger causal discovery; this limitation is apparent when even applying NRI on a very simple benchmark data generating process. We hope that our work will encourage future research in adapting NRI to meet the challenge of Granger causal discovery.

References

John Y. Campbell, Andrew W. Lo, A. Craig MacKinlay, and Robert F. Whitelaw. The econometrics of financial markets. *Macroeconomic Dynamics*, 2(4):559–562, 1998. doi: 10.1017/S1365100598009092.

Michael Eichler. *Causal inference in time series analysis*. Wiley Online Library, 2012.

André Fujita, Patricia Severino, João Ricardo Sato, and Satoru Miyano. Granger causality in systems biology: Modeling gene networks in time series microarray data using vector autoregressive models. In *Advances in Bioinformatics and Computational Biology*, pages 13–24, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15060-9.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Nan Rosemary Ke, Silvia Chiappa, Jane X Wang, Jorg Bornschein, Anirudh Goyal, Melanie Rey, Matthew Botvinick, Theophane Weber, Michael Curtis Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. URL https://openreview.net/forum?id=dhGFrNx85nd.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2688–2697. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kipf18a.html.

Marina Knight, Kathryn Leeming, Guy Nason, and Matthew Nunes. Generalized network autoregressive processes and the gnar package. *Journal of Statistical Software*, 96(5):1–36, 2020. doi: 10.18637/jss.v096.i05. URL https://www.jstatsoft.org/index.php/jss/article/view/v096i05.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 509–525. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/lowe22a.html.

Mariusz Maziarz. A review of the granger-causality fallacy. *The Journal of Philosophical Economics*, 8(2):6, 2015. URL https://EconPapers.repec.org/RePEc:bus:jphile:v:8:y:2015:i:2:n:6.

Olaf Sporns. *Networks of the Brain*. 2016.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. doi: 10.1109/tpami.2021.3065601. URL https://doi.org/10.1109%2Ftpami.2021.3065601.

Yixin Wang, David Blei, and John P Cunningham. Posterior collapse and latent variable non-identifiability. In *Advances in Neural Information Processing Systems*, volume 34, pages 5443–5455. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/2b6921f2c64dee16ba21ebf17f3c2c92-Paper.pdf.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See anonymous Github link on page 2.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Minimal training time on 8 CPU laptop.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

The implementation of the RefMLP encoder follows that of Löwe et al. [2022] and Kipf et al. [2018]. The specifics of our implementation are:

- We use the last 200 values of each time series as input into the first layer of the encoder.
- We use 32 hidden units in each 2-layer MLP.

Under the prior distribution, each edge is sampled uniformly at random from a Bernoulli distribution; the probability of class "no edge" is set to 0.05. This ensures that under the null hypothesis of no Granger causal relations, the type I error rate for each edge is 0.05.

We monitor loss curves to verify that convergence occurs during training. Typically, we use 20 epochs of training with a batch size of 10 samples and learning rate of 5.e-2 (when using UnsharedEncoder) and 5.e-3 (when using RefMLPEncoder). The KL-divergence performance on the validation set is used for model selection.