ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Large-Scale Continuous Structure Learning from Time-Series Data

Filippo Michelis, Riccardo Massidda, and Davide Bacciu*

Università di Pisa - Department of Computer Science
Largo Bruno Pontecorvo, 3, 56127, Pisa - Italy

Abstract. Structure learning is the problem of recovering from data a Directed Acyclic Graph (DAG) of the interactions among variables. By enforcing a differentiable acyclicity constraint on the adjacency matrix of the graph, existing methods solve this problem as an optimization problem and have been recently extended to time-series data. Due to the cubic computational complexity of existing acyclicity constraints, their application is limited to a few variables. In this paper, we introduce SVARCOSMO, an optimization-based structure learning method for time-series data that builds upon recent developments on unconstrained but provably acyclic models. We empirically show on both simulated and real data that SVARCOSMO correctly recovers the underlying DAG in significantly less time, enabling optimization-based structure learning on high-dimensional data.

1 Introduction

As directed graphical models gain popularity to represent structured information about the world [1], fitting their structure from high-dimensional data constitutes a pressing and challenging problem. Despite the existence of numerous algorithms and methodologies for the optimization of Directed Acyclic Graphs (DAGs) [2], these solutions often incur in high computational costs that limit large-scale practical applications. Overall, these issues are exacerbated in time-series, which motivate the need for computationally efficient solutions, as we address in this paper. In principle, when time-dependent information is available, we could exploit it to orient the arrows of DAGs and significantly reduce the space of feasible solutions. However, when the sampling process is slower than the underlying timescale, it still requires the incorporation of instantaneous relationships, whose interactions need to be estimated. This issue is very common in real data scenarios. For instance, in social and behavioral sciences applications, we often have quarterly or even yearly observations, while the underlying process occurs at a much faster pace [3]. A popular model for both instantaneous and time-lagged effects is the Structural Vector Autoregressive (SVAR) framework [4]. While usually employed in the context of time-series forecasting, we can exploit SVAR parameters to determine interactions between variables. In particular, by restricting the instantaneous relations to be acyclic, we can interpret a SVAR as a Dynamic Bayesian Network (DBN) [1] that assumes linearity of both lagged and instantaneous interactions. Formally, for each timestamp t we define an

*Work supported by H2020 project TAILOR (Grant No. 952215)

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

observation $\mathbf{X}_t \in \mathbb{R}^{n\times d}$ of n samples over d variables as

$$\mathbf{X}_t = \mathbf{W}\mathbf{X}_t + \sum_{p=1}^{P} \mathbf{A}_p \mathbf{X}_{t-p} + \mathbf{U}_t, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times d}$ represents the acyclic instantaneous effects, $\mathbf{A}_p \in \mathbb{R}^{d\times d}$ represents lagged-effects at each time step $p \in \{1, \dots P\}$, and $\mathbf{U}_t \in \mathbb{R}^{d\times d}$ is an exogenous noise term, which we assume to be i.i.d. across variables and time steps. To ease the presentation, we further simplify the notation as

$$\mathbf{X} = \mathbf{W}\mathbf{X} + \mathbf{A}\mathbf{Y} + \mathbf{U}_t, \tag{2}$$

where we aggregate lagged terms as $\mathbf{Y} = [\mathbf{X}_{t-1}|\cdots|\mathbf{X}_{t-p}]$ and $\mathbf{A} = [\mathbf{A}_1^T|\cdots|\mathbf{A}_p^T]$.

Given a SVAR, we can estimate its parameters, and consequently its structure, by solving a continuous optimization problem. However, the problem of enforcing the acyclicity of the coefficient matrix $\mathbf{W}$ during the optimization is non-trivial. One of the most popular solutions to this problem is the NOTEARS algorithm [5], which has been extended to handle time-series data by DYNOTEARS [6], obtaining the following SVAR optimization problem:

$$\min_{\mathbf{W},\mathbf{A}} \|\mathbf{X} - \mathbf{X}\mathbf{W} - \mathbf{Y}\mathbf{A}\| \text{ s.t. } \operatorname{Tr}\big(e^{\boldsymbol{W}\odot\boldsymbol{W}}\big) - d = 0, \tag{3}$$

where the constraint is provably satisfied if and only if the matrix $\mathbf{W}$ is acyclic. The constrained problem is solved using the Augmented Lagrangian method [7], which requires evaluating the constraint at each optimization step. Due to the cubic computational complexity on the number of nodes of the matrix-exponential operation, the use of this constraint hinders large-scale applications.

In this work, we introduce SVARCOSMO, a model to efficiently perform structure learning in high-dimensional datasets and that can deal with time-dependent structures even when the observations are collected at a slower pace than the underline process, thus presenting instantaneous effects.

2 Fast Large-Scale SVAR Learning

In this paper, we propose to solve the same optimization problem tackled by DYNOTEARS (Equation 3) by enforcing the acyclicity of the SVAR without the use of computational expensive constraints. To this end, we build upon COSMO [8], a recently introduced unconstrained parameterization for acyclic structure learning. In this way, we define the instant relation matrix $\mathbf{W}$ as the element-wise product

$$\mathbf{W} = \mathbf{H} \odot S_{t,\epsilon}(\boldsymbol{p}), \tag{4}$$

where $\mathbf{H} \in \mathbb{R}^{d\times d}$ is a matrix, $\boldsymbol{p} \in \mathbb{R}^d$ is a priority vector, and $S_{t,\epsilon}(\boldsymbol{p})$ is a smooth orientation matrix defined as $[S_{t,\epsilon}(\boldsymbol{p})]_{uv} = \sigma((p_u - p_v - \epsilon)/t)$ for any variable X_u, X_v, where $\varepsilon, t \in \mathbb{R}$ are two hyperparameters. It is immediate that the number of operations required to construct the matrix $\mathbf{W}$ from the priority vector $\boldsymbol{p}$ is

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

quadratic in the number of nodes. Given this parametrization, we can rewrite the SVAR problem as an unconstrained Mean Squared Error (MSE) optimization problem, which we name SVARCOSMO,

$$\begin{aligned} \underset{\mathbf{H}\in\mathbb{R}^{d\times d},\boldsymbol{p}\in\mathbb{R}^{d},\mathbf{A}\in\mathbb{R}^{d\times(d\cdot T)}}{\arg\min} \quad & \|\mathbf{X} - ((\mathbf{H}\odot S_{t,\epsilon}(\boldsymbol{p}))^{T}\mathbf{X} + \mathbf{Y}\mathbf{A})\|_F \\ & + \lambda_1\|\mathbf{H}\|_1 + \lambda_2\|\mathbf{H}\|_2 \\ & + \lambda_p\|\boldsymbol{p}\|_2 + \lambda_A\|\mathbf{A}\|_1. \end{aligned} \tag{5}$$

Model Optimization. In the limit of the temperature to zero, COSMO, and consequently our SVARCOSMO formulation, ensures that the instantaneous effect matrix $\mathbf{W}$ is acyclic. However, since the gradient loss vanishes as $t \to 0$, we adopt an annealing procedure from a initial positive value to a significantly lower value $t_{end} \approx$ 1e-2. The L1 regularization on $\mathbf{H}$ and $\mathbf{A}$, parameterized by λ_1 and λ_A respectively, serves the purpose of contrasting the discovery of spurious relations. Following [8], the L2-norm on the priority vector $\boldsymbol{p}$ is introduced to prevent zero-gradient regions when the priority differences tend to infinity. In practice, we tackle the optimization problem using the Adam optimizer [9].

Convergence of the Lagged Weights. Inspired by [6], we develop an optimization procedure that estimates both the instantaneous and the lagged effects in one stage. While a two stages approach is more popular in the SVAR literature [4], breaking the estimation into two stages introduces the risk of propagating the biases of the estimation across stages. In practice, retrieving the instantaneous effects $\mathbf{W}$ typically exhibits slower convergence compared to the estimation of the lagged effects $\mathbf{A}$. Therefore, we freeze the values of $\mathbf{A}$, when they reach an optimal state, i.e., the norm of the residuals of the VAR $\|\mathbf{X} - \mathbf{Y}\mathbf{A}\|_F$ remains constant for n iterations (fixed to $n = 300$ after preliminary studies).

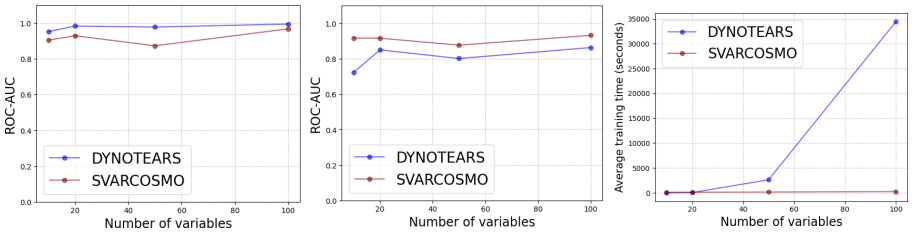
Summary Graph Construction. In a SVAR, a summary graph $\mathcal{G}$ with variables $\boldsymbol{X}$ and edges $\boldsymbol{E}$ is a representation of how variables interact in both the instantaneous and lagged steps. To reconstruct it from our optimization problem, we test the absolute value of the entries in $\mathbf{W}$ and $\mathbf{A}$ against a threshold $\tau = 0.01$, which we also identified from preliminary studies. Formally, we obtain a summary matrix by summing entry wise absolute values from $\mathbf{W}$ and $\mathbf{A}_p$ for each p, as in

$$X_i \to X_j \in \boldsymbol{E} \iff |w_{ij}| + \sum_{p=1}^{P} |a_{ij}^{(p)}| \geq \tau. \tag{6}$$

3 Experiments

In this section, we validate the ability of SVARCOSMO to retrieve the graphical structure from simulated and real data. In practice, given a dataset $\mathbf{X} \in \mathbb{R}^{n\times d\cdot P}$ of n observations on d variables over P timesteps, we wish to reconstruct the

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.



(a) (b) (c)

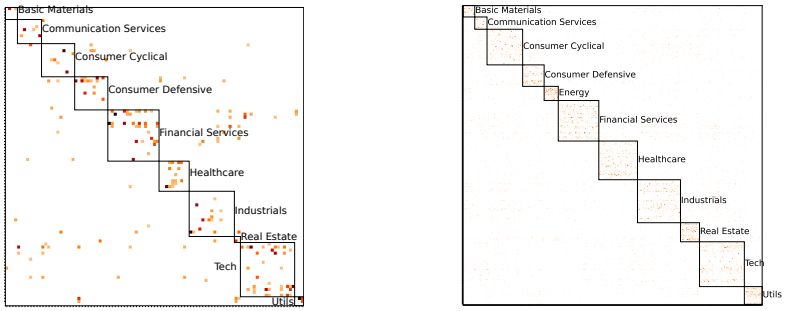
Fig. 1: ROC-AUC comparison for the instantaneous (a) and lagged (b) effects in the retrieved structure. SVARCOSMO and DYNOTEARS perform comparably on the graph reconstruction task, with, respectively, DYNOTEARS having slightly better performance on the instantaneous effects and SVARCOSMO on the lagged effects. In line with our theoretical analysis, SVARCOSMO is considerably more efficient in high-dimensional settings as the time comparison shows (c). The results are averaged over 10 independent runs on graphs of increasing size.

summary graph $\mathcal{G}$ of the underlying model. Notably, since we do not assume any prior information on the underlying data-generating structure, the problem is essentially unsupervised. However, we empirically found that in our experimental setting the minimization of the MSE is consistent with the structure learning problem. This is coherent with previous research on causal discovery [10], which however requires further theoretical assumptions that we do not explore in this work. Finally, we follow the tuning technique of [6], by selecting hyperparameters using a 10-fold cross validation on the MSE of the model.

Summary Graph Bootstrapping. Overall, SVARCOSMO avoids the use of an acyclicity constraint at the cost of relying on more hyperparameters, whose tuning is fundamental to correctly reconstruct the underlying graph. Being more sensitive, we found beneficial to bootstrap the results to avoid the recovery of spurious edges. In practice, we sample b distinct sub-datasets and fit SVARCOSMO on each iteration $i \in \{1, \dots, b\}$ to reconstruct the summary graph as reported in Section 2. We then compose the overall summary graph by selecting only edges that appear in more than half of the bootstrap runs. Notably, while this procedure does not ensure acyclicity of the overall graph, in practice bootstrapping consistently recovers acyclic graphs without the need of any post-processing steps. By considering bootstrapping, the overall complexity of SVARCOSMO is $O(b \cdot d^2)$, which is still quadratic in the number of variables d. In all further experiments, we fix the number of bootstrap samples to $b = 10$.

Simulated Data. To ensure a fair comparison with DYNOTEARS, we replicate the simulated data generation process of [6], for which we refer the reader for a more detailed description. In summary, we sample random DAGs of increasing size and different edge ratios from which we then sample $n = 1000$ observations

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.



(a) S&P100 (b) S&P500

Fig. 2: Graph estimated by SVARCOSMO on the S&P100 (a.) and S&P500 (b.) data. As expected from domain knowledge [11], the underlining graph tend to organize in a block-diagonal structure following distinct economical sectors.

with exogenous Gaussian noise. We then compute the area under the ROC curve (ROC-AUC) of the retrieved weights against the ground-truth summary graph. In general, we see that the SVARCOSMO outperforms DYNOTEARS in retrieving the lagged effects $\mathbf{A}$, while DYNOTEARS demonstrates greater predictive power in capturing instantaneous effects on the graph of $\mathbf{W}$ (Figure 1). Crucially, while the time cost is comparable in low-dimensional settings, with DYNOTEARS even proving slightly more efficient for $d = 10, 20$ — albeit comparable and likely due to bootstrapping — the experiments for $d = 50, 100$ reveal that DYNOTEARS becomes almost impractical for high-dimensional problems, with the training of certain configurations exceeding 10 hours. The time performance comparison clearly depicts the substantial difference in computational cost between the two algorithms and the advantage of SVARCOSMO in high-dimensional settings.

Large-Scale Stock Prediction. As highlighted in the previous sections, the primary advantage of SVARCOSMO lies in its efficient retrieval of Directed Acyclic Graphs. Such efficiency is crucial in big-data applications, where the large number of variables render less efficient algorithms, such as DYNOTEARS, impractical. To illustrate this advantage, we apply SVARCOSMO to financial data from the Standard and Poor's (S&P) stock exchange market, by enriching the experiment in [6] (Figure 2). By avoiding the use of an expensive acyclicity constraint, we obtain comparable results in significantly less time, opening the door to the method's utilization in much larger datasets, as is often the case in financial applications. Importantly, SVARCOSMO correctly identifies interactions among S&P500 industries, which would have been impractical for DYNOTEARS in terms of computational time, as our study on simulated data highlighted. As we can see, ordering the titles by industrial sector, the resulting graph has an approximately block-diagonal structure, which is evidence for stronger relationships within the same sector. Given the absence of evident causal relationships among S&P100 and S&P500 stocks, the experiment is entirely unsupervised. However, the results

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

are coherent with existing economic literature on the comovement of stock prices by industrial sector, i.e., stocks within the same sector tend to move together [11].

4 Conclusion

We introduced SVARCOSMO, a procedure to recover the underlying structure of the interactions among DBNs, which we model as acyclic Structural Vector Autoregressive Models (SVARs). SVARCOSMO avoids the use of computationally expensive acyclicity constraints by casting the problem in a continuous and differentiable space of DBNs. As we showed on simulated and real data, SVARCOSMO performs structure learning in significantly less time and with analogous reconstruction results when compared to constrained methods. As in the literature, we focused on the recovery from linear models and did not explore the — albeit valuable — connection between structure learning and causal discovery, which instead requires further assumptions, leaving it to future works.

References

[1] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[2] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

[3] Gianni Amisano and Carlo Giannini. *Topics in structural VAR econometrics*. Springer Science & Business Media, 2012.

[4] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[5] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[6] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, 2020.

[7] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[8] Riccardo Massidda, Francesco Landolfi, Martina Cinquini, and Davide Bacciu. Constraint-free structure learning with smooth acyclic orientations. In *The Twelfth International Conference on Learning Representations*, 2024.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

[10] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1): 3065–3105, 2014.

[11] Benjamin F King. Market and industry factors in stock price behavior. *the Journal of Business*, 39(1):139–190, 1966.