Department of Mathematics and Computer Science
Statistics Group

Structure Learning in High-Dimensional Time Series Data

Master Thesis

Martin de Quincey

Supervisors:
dr. Rui Castro
dr. Alex Mey

Assessment Committee Members:
dr. Rui Castro
dr. Alex Mey
dr. Jacques Resing

version 0.4

Eindhoven, June 2022

Contents

Chapter 8

Conclusion

In this work, we have discussed the notion of structure learning in time series data. Given a set of time series, the research objective was to learn how the values of one time series influence the future values of another time series. These directed relationships can be summarized in a graphical model where an arc from node i to j indicates that time series i is useful in predicting future values of time series j. To enhance the interpretability and to promote sparsity of the learned graphical model, we have explicitly forbidden the existence of cycles in our graphical models, excluding self-loops.

In Chapter 2, we have formalized the notion of structure learning in time series, and have introduced the VAR(1) model that we have employed to learn an acyclic structure in time series data. Subsequently, we have discussed some interesting state-of-the-art methodologies in Chapter 3 which can learn the structure of time-independent data and how these methodologies can be extended to time series data.

Methodologies. Firstly, in Chapter 4, we have discussed several permutation-based approaches with an increasing level of complexity. Acyclicity was enforced by first selecting a permutation after which only arcs were permitted that respect the induced permutation. That is, only arcs were permitted from a variable i to another variable j if variable i precedes variable j in the permutation. In Section 4.1, we have investigated an exhaustive approach, where we simply try all possible permutations. As such a method is not tractable for more than ten variables, we have devised search algorithms that do not exhaustively try all permutations, but rather try a subset of all permutations. Unguided searches such as the random walk have been proposed in Section 4.2, and a more informative search such as the Metropolis-Hastings approach has been proposed in Section 4.3, where the search is guided by assigning a likelihood score to each permutation the algorithm investigates.

Secondly, we have investigated methods that use continuous constraints to enforce acyclicity in Chapter 5. We have tried in Section 5.1 to relax the space of permutation matrices to the space of doubly stochastic matrices and consequently performed a gradient descent with Lagrange multipliers to ensure that the matrix P remains doubly stochastic. Unfortunately, this approach was unsuccessful, as convergence was slow and this method was not able to enforce acyclicity. Alternatively, in Section 5.2, we have modified the NO TEARS approach from [78] such that the approach was capable of learning an acyclic VAR(1) model. Thirdly, we have investigated the use of a LASSO-penalty to enforce acyclicity by increasing the penalty parameter until the inferred structure was acyclic in Section 5.3.

Thirdly, we have developed iterative methods where our learned structure is updated one arc at a time rather than all arcs simultaneously. We have first extended the Orthogonal Matching Pursuit algorithm to estimate sparse VAR(1) models, after which we also enforce acyclicity in Section 6.1. A backward ordinary least squares approach was proposed in Section 6.2 and an improvement to backward approaches was discussed in Section 6.3. Additionally, we have developed several approaches to determine a suitable number of arcs in our structure. Too few arcs may

cause a significant drop in predictive performance and on the other hand, too many arcs may result in an overly complex structure. Therefore, we have developed two approaches that rely on bootstrapping and one approach that relies on cross-validation to learn a suitable number of arcs in Section 6.4. Lastly, as cross-validation was surprisingly effective despite having time-dependent data, we have conducted a small theoretical analysis regarding leave-one-out cross-validation on autoregressive models in Section 6.5, where we discovered an interesting relation between Wilk's theorem and the leave-one-out cross-validation score.

Evaluations. The aforementioned methods have been extensively evaluated and their predictive and structural performance have been compared. For the time series setting, we have simulated acyclic VAR(1) models and cyclic VAR(1) models, and have applied our methods to real-life data in the form of weekly scanner data of the convenience store chain "Dominick's Finer Foods". All methods except for DAG-LASSO performed well, even when the generated VAR(1) model was cyclic. The permutation-based approaches were particularly effective when the number of variables was small. However, the permutation-based approaches dropped in performance as the number of variables increased, as the number of possible permutations increases exponentially with the number of variables. The iterative approaches seemed particularly effective when the number of time series was large and sufficient time steps were available. The remaining NO TEARS approach was effective as well, but our developed methods remained competitive with NO TEARS.

We have also briefly investigated the performance of our methodologies on time-independent data, ~~as this type of structure learning is more established.~~ We have used simulated data in the form of a linear structural equation model as well as the real-life data set provided by Sachs et al. [59], consisting of causal pathways in protein interactions within a biological cell. Rather surprisingly, we discovered that our methods were competitive with a state-of-the-art method such as NO TEARS on both the simulated data and the real-life data.

8.1 Limitations

Throughout this thesis, certain assumptions have been made to limit the scope of research. Furthermore, due to time constraints, some concepts in this thesis have not been investigated at the level of detail we had hoped. In this section, we will describe these limitations.

Strictness of the graphical models

The most severe limitation of this thesis is the strictness of the graphical model. Throughout this thesis, we have only considered two types of graphical models. The first type of graphical model we considered is the linear structural equation model (SEM) as per Definition 2.2 for time-independent data. The second type of graphical model discussed in this thesis is the Vector AutoRegressive model of order 1 (VAR(1)) as per Definition 2.3 for time series data. The linear SEM only allows for instantaneous linear relationships and the VAR(1) model only allows for linear relationships with a single time lag.

Although it is a realistic assumption that only the past can predict the present, in real-life measurements, instantaneous relations may exist, especially when the time intervals between consecutive measurements is rather large. If we acquire daily measurements, it is reasonable to assume that the rainfall of that day has influenced the wetness of the pavement that day. Therefore, even a time series model should allow for instantaneous effects, meaning we should have added an instantaneous relationship to our time series model as well.

Furthermore, it is reasonable to assume that more intricate relationships exist than linear relations based on values of exactly one time step ago. Suggestions to allow for more intricate relationships will be discussed in the next section where we will propose future directions of research.

Model Complexity Selection

From the outset of this thesis, the goal was to recover a structure of $\mathbf{X}$ that captures the relations between the variables, while simultaneously being intuitive to understand. Sparsity of the structure, meaning the graphical model contains few arcs, is crucial for obtaining a simple graphical model. For iterative procedures, quite some different methods have been devised to select an appropriate number of arcs, as iterative procedures provide an order of importance of the inferred arcs.

However, for the permutation-based approaches discussed in Chapter 4, little research has been done to sensibly select an appropriate number of arcs. As permutation-based approaches estimate a full directed acyclic graph that respects a given permutation, numerous arcs in W are superfluous and can therefore be removed. Nevertheless, not many sensible approaches have been discussed that can "prune" such a coefficient matrix W efficiently. Thresholding the coefficient matrix is an effective approach, yet the approach is quite naive as the coefficient size of an arc is not always a suitable indication of its importance. Therefore, we consider the lack of suitable methods for selecting an appropriate number of arcs for permutation-based methods a limitation of this thesis.

For the continuous-based approaches discussed in Chapter 5, no methods are provided to select an appropriate number of arcs as well. On the one hand, the LASSO-penalty used in NO TEARS and DAG-LASSO does promote some degree of sparsity. However, little research has been done into selecting the correct magnitude of this LASSO-penalty parameter. We also consider this to be a shortcoming of this thesis.

8.2 Future Work

Although this thesis can be regarded as quite a comprehensive work, numerous avenues can be regarded as interesting directions for future work. Due to time constraints, these were not investigated in more detail in this thesis. However, some initial investigations and initial pointers will be provided in this section which an enthusiastic reader may continue on.

Allowing for more complex models.

As mentioned in the previous section, the VAR(1) model is quite a simple model that does not allow for much flexibility. There are several approaches to resolve this limitation and allow for more complex models.

Incorporating instantaneous relationships. As mentioned in the previous section, although instantaneous relations may be unrealistic in theory, the time intervals at which variables are measured can be so far apart that we should also incorporate instantaneous relationships in our time series model.

A natural extension is the combination of the VAR(1) model from Definition 2.3 and the linear Structural Equation Model from Definition 2.2. Instantaneous relations are captured in the coefficient matrix W_0, and time-lagged relations are captured in the coefficient matrix W_1. In mathematical notation,

$$X_{t,\cdot} = X_{t,\cdot}W_0 + X_{t-1,\cdot}W_1 + \varepsilon_t, \tag{8.1}$$

where ε is a p-dimensional vector representing the noise. Note that W_0 must be an acyclic coefficient matrix where the diagonal entries are equal to zero, just as for the linear SEM.

Extending to VAR(k) models. Throughout this thesis, we have predominantly focused on *time series* data. As our goal was to infer the structure of a graphical model, a natural first step was the VAR(1) model, where we have but a single coefficient matrix W. However, VAR(1) models are quite limited. It is reasonable to assume that a vector of variables $X_{t,\cdot}t$ depends on more than just its previous time step $X_{t-1,\cdot}$. Therefore, a VAR(1) model might be too simple and

fail to capture more intricate relations. Extending the proposed methodologies to more complex models would be an interesting avenue to discover.

Luckily, we can quite easily extend the VAR(1) model. Instead of a VAR(1) model, one could consider a VAR(k) model. Here, the values of $X_{t,\cdot}$ do not only depend on the past value but the past k values,

$$X_{t,\cdot} = \sum_{i=1}^{k} X_{t-1,\cdot} W_i + \varepsilon_t, \tag{8.2}$$

where ε_t represents some zero-mean noise. For a VAR(1) model, we had $k = 1$ and we only had the matrix W_1, which we simply denoted as W. Now, we have k coefficient matrices.

Enforcing acyclicity of W or W_1 was quite intuitive and straightforward, as we could regard this as one network. However, the notion of acyclicity has become ambiguous now that we have k coefficient matrices rather than one. Firstly, we could say that the structure is acyclic if and only if all its k coefficient matrices separately are acyclic. However, this could mean that in W_1 we have the arc $(1, 2)$, and W_2 contains the arc $(2, 1)$ such that we still have some form of cyclic dependency. A second approach would be to consider some sort of combined network, where the combination of all coefficient matrices must be acyclic. In mathematical notation, consider W', where

$$W' = \sum_{i=1}^{k} |W_i|, \tag{8.3}$$

where $|\cdot|$ represents the element-wise absolute value. Therefore, W' contains an arc (i, j) if the arc (i, j) is contained in at least one of the k coefficient matrices. We then could consider the structure to be acyclic if and only if W' is acyclic. This means that if there is an arc (i, j) in some coefficient matrix, then we know that no other coefficient matrix contains the arc (j, i). In fact, we know that there is no other path from variable j to variable i, where we can use all arcs from all k coefficient matrices. Such a definition might be more sensible as there can be no cycles in the relations between the variables across different time lags. A similar approach for learning sparse VAR(k) models by decomposing the indices into groups has been investigated in [48], so that may prove a useful starting point.

So, extending our methodologies and the notation of acyclicity to higher-order VAR models is not necessarily difficult, yet the definition of acyclicity would become less intuitive. It would be interesting to see how higher-order models are perhaps more suitable to estimate more complex simulated data and real-life time series data, and what the advantages and disadvantages of these newly proposed notions of acyclicity are.

As a side note, instantaneous relationships can also be incorporated in VAR(k) models, yielding a structural VAR(k) model. Then, $X_{t,\cdot}$ can be written as

$$X_{t,\cdot} = X_{t-1,\cdot} W_0 + \sum_{i=1}^{k} X_{t-1,\cdot} W_i + \varepsilon_t, \tag{8.4}$$

where W_0 characterizes the instantaneous relationships, W_i characterizes the time-lagged relationships, and ε_t represents random noise.

Non-linear relationships. Rather than only allowing for linear relationships in the form of a coefficient matrix W, we can also try to learn more complex non-linear relationships between our variables. Rather than assuming X_j is a linear function of its parent set $\mathrm{Pa}(X)$,

$$X_j = \sum_{X_i \in \mathrm{Pa}(X_j)} w_{ij} X_i, \tag{8.5}$$

we can also assume that X_j is some non-linear function g of its parents,

$$X_j = g\left(\mathrm{Pa}(X_j)\right). \tag{8.6}$$

Several methods exist to learn non-linear relationships, such as the recent non-linear version of NO TEARS [79], and DAG-GNN [75] where the authors employ a variational autoencoder. Both employ neural networks to learn non-linear relationships between variables.

More complex noise structure. Throughout this thesis, we have made the simple assumption that all noise random variables ε_t are independently and identically distributed as a Gaussian random variable with zero mean and an identity matrix as the covariance matrix,

$$\varepsilon_t \sim \mathcal{N}(\mathbf{0}, I_p). \tag{8.7}$$

However, this assumption is rather strict, as such a homogeneity assumption of the noise variables is often violated in practice.

For future work, the performance of the model may increase if we allow for more complex noise structures. For example, rather than assuming that $\mathbb{V}(\varepsilon_t) = I_p$, we can make the assumption that all noise components are independent yet heterogeneous,

$$\mathbb{V}(\varepsilon_t) = \operatorname{diag}\left(\omega_1^2, \ldots, \omega_p^2\right), \tag{8.8}$$

where ω_i^2 represents the variance associated with the noise of the ith variable. Furthermore, $\operatorname{diag}(\omega_1^2, \ldots, \omega_2^2) \in \mathbb{R}^{p \times p}$ represents the diagonal matrix with ω_i^2 as the ith value along the diagonal. It should be noted, however, that this adds a total of p extra parameters to estimate, which may be problematic when we have few samples. However, as the total number of possible arcs is quadratic with respect to the number of variables, we expect the number of parameters to not increase too drastically.

We can also assume even more complex noise structures. We can, for example, allow some dependency between noise variables of the same time step. Then, we assume that

$$\mathbb{V}(\varepsilon_t) = \Omega, \tag{8.9}$$

where $\Omega \in \mathbb{R}^{p \times p}$ is some positive semi-definite matrix. This would yield an additional $p(p-1)/2$ parameters over Equation 8.8, so the number of parameters would remain quadratic with respect to the number of variables.

Developing Theoretical Guarantees

Although we have shown some interesting theoretical results for autoregressive models, these results were not extended to more complex models such as VAR(1) models in this thesis. For example, it would have been nice to provide some theoretical guarantees for some of our methodologies.

Orthogonal Matching Pursuit. For the Orthogonal Matching Pursuit algorithm discussed in Section 6.1, performance guarantees have been provided for the standard regression setting with noise [76, 12] and without noise [67]. However, our setting is more involved, where we have multiple response variables rather than one and there are strong dependencies between different time series. Therefore, existing performance guarantees for Orthogonal Matching Pursuit ~~of~~ could not be directly applied to our version.

However, it would be of great value to investigate whether we can translate the orthogonal matching pursuit guarantees in some way to the `DAG-OMP` algorithm. Having a formal statement that guarantees that the correct features will be recovered under certain conditions can provide more insights into the performance of `DAG-OMP`. Unfortunately, this was not pursued due to time constraints, but it could be an interesting avenue to investigate.

Greedy Metropolis-Hastings. When we evaluated our methods, the greedy Metropolis-Hastings approach seemed to perform surprisingly well. For the linear SEM with equal variances, there are some interesting performance guarantees for algorithms that employ a greedy approach to derive an ordering of the variables, such as [16] and [55]. Both methods are quite similar to ours, so we

expect a performance guarantee to exist for the greedy Metropolis-Hastings algorithm as well, for example, the number of iterations required to recover the support of the data generating matrix, assuming we have enough samples. Furthermore, we expect that these performance guarantees can also be cast in the VAR(1) setting. Providing such a performance guarantee for the greedy Metropolis-Hastings approach could be an interesting direction for future work.

Or probably we need strong assumptions, would be good to indeed have some context of those results

Sparsity imposed by acyclicity

Another interesting direction for future work that we have not thoroughly explored is how acyclicity could be used to derive tighter theoretical guarantees for known methods such as the LASSO. Well-known theoretical guarantees exist for the LASSO approach in the regression setting. For example, in [46], the authors show that under certain conditions, the LASSO-penalty should be approximately $\sqrt{\log(p)/T}$, where p is the number of variables and T is the number of samples. Now, what if instead of knowing that the coefficient vector contains only k non-zero coefficients, we know that the coefficient *matrix* is acyclic? If our coefficient matrix is acyclic, we know it contains at most $p(p+1)/2$ parameters, which is already much fewer than the possible p^2 parameters.

What do they show? Something like model selection consistency, or rather regression performance guarantees?

The question arises whether the acyclicity assumption can provide tighter bounds on convergence rates, penalty parameter values, etc. The number of possible parameters remains quadratic, so in terms of complexity we have not gained much. However, a more optimistic perspective is by considering the reduction in search space. If we would allow any coefficient matrix, or equivalently any structure on p variables without self-loops, then there are a total of 2^{p^2-p} different structures, as p^2-p entries can either be zero or non-zero. However, we have seen that the number of directed acyclic graphs grows much smaller. The number of possible directed acyclic graphs up to fourteen nodes has been reported in [49]. Now, define R_p as the ratio of the total number of possible directed acyclic graphs on p nodes over the total number of possible directed graphs on p nodes. Then, some values of R_p are

$$R_1 = 1, R_2 = 0.75, \ldots, R_5 = 0.027, \ldots, R_{10} = 3.4 \cdot 10^{-9}, \ldots, R_{14} = 2.3 \cdot 10^{-19}. \tag{8.10}$$

From Equation 8.10, we see that the search space of structures has decreased massively. Even for a moderate total of ten nodes, knowing that the true structure is acyclic has reduced the number of possible structures by a factor 10^9. As acyclicity indeed massively decreases the search space, perhaps this may also imply that we can obtain tighter bounds for ~~for~~ several performance guarantees of existing methods. Investigating this in greater depth as future work may provide some interesting novel results.

Bibliography

[1] Scipy api reference for `optimize.minimize`. 22

[2] Scipy api reference for the `L-BFGS-B` optimization method. 22

[3] *Vector Autoregressive Models for Multivariate Time Series*, pages 385–429. Springer New York, New York, NY, 2006. 54

[4] Awad H. Al-Mohy and Nicholas J. Higham. A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications*, 31(3):970–989, 2010. 74, 93

[5] M. Andrle, L. Rebollo-Neira, and E. Sagianos. Backward-optimized orthogonal matching pursuit approach. *IEEE Signal Processing Letters*, 11(9):705–708, 2004. 91

[6] Mark Bartlett and James Cussens. Integer linear programming for the bayesian network structure learning problem. *Artificial Intelligence*, 244:258–271, 2017. Combining Constraint Solving with Mining and Learning. 20

[7] Christoph Bergmeir, Rob J. Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics Data Analysis*, 120:70–83, 2018. 104

[8] Garrett Birkhoff. Three observations on linear algebra. *Univ. Nac. Tacuman, Rev. Ser. A*, 5:147–151, 1946. 51

[9] Thomas Blumensath and Mike Davies. On the difference between orthogonal matching pursuit and orthogonal least squares. 03 2007. 76

[10] Graham Brightwell and Peter Winkler. Counting linear extensions is #p-complete. In *Proceedings of the Twenty-Third Annual ACM Symposium on Theory of Computing*, STOC '91, page 175–181, New York, NY, USA, 1991. Association for Computing Machinery. 32

[11] Prabir Burman, Edmond Chow, and Deborah Nolan. A cross-validatory method for dependent data. *Biometrika*, 81:351–358, 1994. 107

[12] T. Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011. 133

[13] Nancy Cartwright. Are rcts the gold standard? *BioSocieties*, 2(1):11–20, 2007. 4

[14] Rui Castro and Robert Nowak. Likelihood based hierarchical clustering and network topology identification. In Anand Rangarajan, Mário Figueiredo, and Josiane Zerubia, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 113–129, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. 39

[15] S. CHEN, S. A. BILLINGS, and W. LUO. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, 1989. 76

[16] Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 09 2019. 133

[17] David Maxwell Chickering. *Learning Bayesian Networks is NP-Complete*, pages 121–130. Springer New York, New York, NY, 1996. 15

[18] Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2):393–405, 1990. 4

[19] James Cussens. Bayesian network learning with cutting planes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 153–160, Arlington, Virginia, USA, 2011. AUAI Press. 20

[20] Aramayis Dallakyan and Mohsen Pourahmadi. Learning bayesian networks through birkhoff polytope: A relaxation method. *CoRR*, abs/2107.01658, 2021. 63

[21] Ivan Damnjanovic, Matthew E. P. Davies, and Mark D. Plumbley. Smallbox - an evaluation framework for sparse representations and dictionary learning algorithms. In Vincent Vigneron, Vicente Zarzoso, Eric Moreau, Rémi Gribonval, and Emmanuel Vincent, editors, *Latent Variable Analysis and Signal Separation*, pages 418–425, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 86

[22] Tristan Deleu, António Góis, Chris Chinenye Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. 127

[23] Vera Djordjilović, Monica Chiogna, and Jiří Vomlel. An empirical comparison of popular structure learning algorithms with a view to gene network inference. *International Journal of Approximate Reasoning*, 88:602–613, 2017. 116

[24] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979. 97

[25] Kim Esbensen and Paul Geladi. Principles of proper validation: use and abuse of re-sampling for validation. *Journal of Chemometrics*, 24:168 – 187, 03 2010. 104

[26] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman amp; Co., USA, 1990. 76

[27] Maxime Gasse, Alex Aussem, and Haytham Elghazel. An experimental comparison of hybrid algorithms for bayesian network structure learning. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 58–73, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 116

[28] Sarah Gelper, Ines Wilms, and Christophe Croux. Identifying demand effects in a large network of product categories. *Journal of Retailing*, 92(1):25–39, 2016. 123

[29] M. Gharavi-Alkhansari and T.S. Huang. A fast orthogonal matching pursuit algorithm. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 3, pages 1389–1392 vol.3, 1998. 76

[30] Hemant S. Goklani, Jignesh N. Sarvaiya, and A. M. Fahad. Image reconstruction using orthogonal matching pursuit (omp) algorithm. In *2014 2nd International Conference on Emerging Technology Trends in Electronics, Communication and Networking*, pages 1–5, 2014. 77

[31] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. 5

[32] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. 72

[33] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. 39

[34] Sander Hofman. Making euv: From lab to fab, Mar 2022. 3

[35] Guoxian Huang and Lei Wang. High-speed signal reconstruction with orthogonal matching pursuit via matrix inversion bypass. pages 191–196, 10 2012. 86

[36] Robin John Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 2nd edition, 2018. 26

[37] Donald B. Johnson. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing*, 4(1):77–84, 1975. 92

[38] Hidde De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *JOURNAL OF COMPUTATIONAL BIOLOGY*, 9:67–103, 2002. 3

[39] A. B. Kahn. Topological sorting of large networks. *Commun. ACM*, 5(11):558–562, nov 1962. 75

[40] Wagner A. Kamakura and Wooseong Kang. Chain-wide and store-level analysis for cross-category management. *Journal of Retailing*, 83(2):159–170, 2007. 123

[41] Mahdi Khosravy, Nilanjan Dey, and Carlos Duque. *Compressive Sensing in Health Care*. 10 2019. 76

[42] S. N. Lahiri. *Bootstrap Methods*, pages 17–43. Springer New York, New York, NY, 2003. 97

[43] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):157–224, 1988. 4

[44] Hanxi Li, Yongsheng Gao, and Jun Sun. Fast kernel sparse representation. In *2011 International Conference on Digital Image Computing: Techniques and Applications*, pages 72–77, 2011. 86

[45] S.G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. 76

[46] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246 – 270, 2009. 134

[47] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. , 21(6):1087–1092, June 1953. 39

[48] William B. Nicholson, David S. Matteson, and Jacob Bien. Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, 33(3):627–651, 2017. 132

[49] OEIS Foundation Inc. The On-Line Encyclopedia of Integer Sequences, 2022. Published electronically at `https://oeis.org/A003024`. 134

[50] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1595–1605. PMLR, 26–28 Aug 2020. 23

[51] Koen Pauwels. How retailer and competitor decisions drive the long-term effectiveness of manufacturer promotions for fast moving consumer goods. *Journal of Retailing*, 83(3):297–308, 2007. 123

[52] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288, 1986. 4

[53] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. v, 1, 2, 4

[54] Judea Pearl and Thomas Verma. A theory of inferred causation. In *KR*, 1991. 16

[55] J. PETERS P. BÜHLMANN. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014. 133

[56] K. B. Petersen and M. S. Pedersen. The matrix cookbook, October 2008. Version 20081110. 57

[57] Joseph Ramsey and Bryan Andrews. Fask with interventional knowledge recovers edges from the sachs model. *ArXiv*, abs/1805.03108, 2018. 127

[58] R. W. Robinson. Counting unlabeled acyclic digraphs. In Charles H. C. Little, editor, *Combinatorial Mathematics V*, pages 28–43, Berlin, Heidelberg, 1977. Springer Berlin Heidelberg. 31

[59] Karen Sachs, Omar Perez, Dana Pe'er, Douglas Lauffenburger, and Garry Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science (New York, N.Y.)*, 308:523–9, 05 2005. vii, vii, ix, 127, 128, 130

[60] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. 2018. 15

[61] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyv228;rinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. 18, 87

[62] Konstantinos Skianis, Nikolaos Tziortziotis, and Michalis Vazirgiannis. Orthogonal matching pursuit for text classification. *ArXiv*, abs/1807.04715, 2018. 77

[63] Shuba Srinivasan, Koen Pauwels, Dominique M. Hanssens, and Marnik G. Dekimpe. Do promotions benefit manufacturers, retailers, or both? *Management Science*, 50(5):617 – 629, 2004. Cited by: 177; All Open Access, Green Open Access. 123

[64] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. 104

[65] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 67

[66] Ryan Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39, 05 2010. 69

[67] J.A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004. 133

[68] Joel A. Tropp and Anna C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007. 86

[69] Ioannis Tsamardinos, Laura Brown, and Constantin Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 10 2006. 116

[70] Alexander L. Tulupyev and Sergey I. Nikolenko. Directed cycles in bayesian belief networks: Probabilistic semantics and consistency checking complexity. In Alexander Gelbukh, Álvaro de Albornoz, and Hugo Terashima-Marín, editors, *MICAI 2005: Advances in Artificial Intelligence*, pages 214–223, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. 6

[71] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 22

[72] John Von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. *Contributions to the Theory of Games*, 2(0):5–12, 1953. 51

[73] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Comput. Surv.*, mar 2022. Just Accepted. 15, 17, 116

[74] Norbert Wiener. The theory of prediction. *Modern mathematics for engineers*, 1956. 5

[75] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 09–15 Jun 2019. 127, 133

[76] Tong Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10(19):555–568, 2009. 76, 133

[77] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning, 2018. 22

[78] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9492–9503, Red Hook, NY, USA, 2018. Curran Associates Inc. 50, 63, 125, 127, 129

[79] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020. 133

[80] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, December 1997. 22

[81] Hufei Zhu, Wen Chen, and Yanpeng Wu. Efficient implementations for orthogonal matching pursuit. *Electronics*, 9:1507, 09 2020. 86