

Network Topology and Causal Structure Recovery in Linear Dynamical Systems

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Mishfad Shaikh Veedu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Professor. Murti V. Salapaka
Vincentine Hermes-Luh Chair
Department of Electrical and Computer Engineering

April, 2024

© Mishfad Shaikh Veedu 2024
ALL RIGHTS RESERVED

Acknowledgements

I would like to express my deepest gratitude to several people who have supported and guided me in completing my Ph.D. endeavor.

Firstly, I want to thank Professor Murti V. Salapaka, my Ph.D. advisor, for taking me under his wing and for his unwavering support throughout my PhD journey. The past five years have been marred by a few personal losses and struggles due to COVID-19. It would have been impossible for me to complete this Ph.D. without his compassion, generosity, and unwavering support. I have always been impressed and inspired by his wide and deep knowledge of the diverse fields of science and engineering. His knowledge and aptitude to have deep discussions about topics from control and power systems to protein synthesis; from information theory, probability theory, and differential geometry to thermodynamics and biology is awe-inspiring. I have greatly benefited from his advice and expertise. He taught me the importance of clear and sequential thinking and the significance of clear and simple articulation of complex problems and concepts. I continue to be inspired by his meticulous and systematic approach to solving real-world problems that matter today. Also, I appreciate the freedom he gave me to pursue research according to my interests.

I am grateful to my doctoral committee members, Professor Andrew Lamperski, Professor Nicola Elia, and Professor Gilad Lerman, for agreeing to be on my thesis committee and for their advice and insightful comments that helped in strengthening this thesis.

The thesis would be incomplete without acknowledging Harish. Our discussions and the collaboration resulted in the first chapter of the thesis. I still reminisce about the moment and the discussion at midnight we had when we formulated an algorithm to identify the structure of the latent nodes. I will always cherish this memory. My internship at Los Alamos was a great experience, thanks to Dr. Deepjyoti Deka, a Scientist at Los Alamos National Lab. I sincerely thank him for his time, efforts,

and help in shaping the research in Chapter 4. I was able to learn a lot from him. His controlled approach of pushing me, letting me work on the subproblem, and sitting with me and going through the intricacies of the codes and the proofs, especially when we were stuck were invaluable. I would also like to thank James Melbourne for the several insightful discussions since I joined the lab. We had several interesting discussions on optimization, information theory, convex and Riemannian geometry, probability theory, and optimal transport.

I am deeply indebted to my fellow lab members, whose unwavering support and camaraderie have made the past five years an unforgettable journey. Vivek, thank you for countless hours of engaging discussions on both research and life, and for the memorable badminton matches we played together (even reaching the final of a tournament, although at a lower level). Soham, your philosophical insights and friendship have been invaluable to me. Arnab, Soham, thank you for your hospitality and for opening your home to me on multiple occasions. Connor and Cailong, we had less research overlap, but your willingness to share insights and offer assistance whenever needed has been truly appreciated. Connor also shared several valuable feedbacks that has improved my presentation skills. Cailong’s unwavering support and readiness to help, even at short notice was remarkable. Tuhin and I shared plenty of time during the final months of my PhD. He has helped me a lot in our research, but more importantly, he has been a great moral support, collaborator, and a good friend who cared deeply and sincerely; I have always felt a deep connection with him. I extend my heartfelt gratitude to my senior lab members—Harish, Venkat Ram, Sourav, Sivaraman, Rachit, and Govind—for their guidance, friendship, and mentorship throughout this journey.

I would like to express my gratitude to Vaidyanathan for his support during my transition to Minneapolis and for being like a caring big brother to me. Same with our third musketeer, Arjun, who made my life at 714 and beyond cherishable, especially during COVID-19. The same is true with Sarthak who had been tolerating me since my first day in the department. We shared plenty of time with courses, research, reading groups, badminton (even winning a team tournament under his leadership), etc. I am thankful to Akshay and Rohith for the numerous bike rides to Minnehaha and the nearby lakes and shores, and for the time spent together. Sajna, Irfan, Shafeeq, Navaneeth, Rohith, Vegnesh, Jazeem, Hashma, little Alizah, Merin, Varun, Sushmita, Gunjan, Praise, Joseph, Dominic, Nishanth, Shyam, Nirmal, etc. made my life outside research memorable. I would also like to thank my friends in EE Gaurav, Tariq, Sean Yin, Adel, Burhan, Hadi, Michelle, Mehran, Wen Zhou, and Connor Hashemi. I apologize if I

missed anyone.

Badminton was a huge part during my PhD life. I would like to thank the fellow players and friends Sarthak, Vivek, Arijit, Subbu, Jashan, Kenneth, Abhishek, Joe, Qiyue, Tan, Eng Hock, Tuii, Thuy, Paul Chen, Yutong, Ying, Summer, Joel, Paul Walker, Sylvester, Ming, Tracy, Anthony, Jason, Jane, Kent, Reuben, Zoe, Abhilash, Ronald, Vivian, Daniel, Jake, Dick, Dane, YuQian, Angie, Yen Sheng, Wei Nee, Jiarong Hong, Vladimir Cherkassky, Thinh, David, Caroline, KS, Eve, Ghangua, Shengcheng Tao, Wen Hao, ChengQi Li, Jay Wang, Ben Chua, Zi Hern Wong, Hnubae, Clarisse, Andy for their time on and off the court. A special thanks to Abhishek Kulkarni who organized the university team for the inter-university tournament and for giving me a chance to represent the university, which was a hugely proud moment for me. Also to Sarthak for picking me in his team and leading us to win the local tournament.

I would also like to thank my friends as AMCC/MSA: Ismail and Ibrahim Irfanullah, Ibrahim Oladepo, Ibrahim Ismail, Andar, Hussam, Ahmed, Aman, Huzaifa, Abdullahi, Abdullahi Nor, Yunus, Ahmad, Ahmad, Abdullahi Nor, Sam, Izhan, Muhammed and Ibraheem Fannoun, Khalid, Faisal, Hamza, Isaac, Khadar, Maaz, Salah, Sharmarke, Muhsin, Yusuf, Sana, Linta, Munira, Nura, Fartun, Zayna, Kafia, Aisha, Maryan, Farida, Joshua, and the others whom I might have missed.

I would also like to thank all my previous mentors and advisors, especially Professor Chandra R. Murthy and Professor Himanshu Tyagi from the Indian Institute of Science Bangalore, India, and the other Professors in the department of ECE at IISc, who were the inspiration for me to pursue Ph.D. Their encouragement paved the way for my penchant for research. I would also like to thank Prof. Martina Cardone, who gave me an opportunity to do research under her at UMN, albeit for a short duration.

Moreover, I wish to express my sincere gratitude to all the teachers and my friends from undergrad at Mar Athanasius College of Engineering, my high school Rahmaniyya Higher Secondary School, and Govt. Ganapat Boys High School, Chalappuram, and my lower primary school St Patrick's School.

Finally, I would like to thank my family. I am most thankful to my Umma and late Uppa who always pushed and supported me to pursue my dreams. My sister Fahimah (Mol) and brother Fahim (Mon) were always with me in my good and bad. I reminisce about my childhood fights with Mon and adorable discussions with Mol. Through laughter and tears, triumphs and challenges, Mon and Mol stood by my side, offering unwavering encouragement and unwavering love. Losing Uppa and Mon in the past two years has been incredibly difficult, often bringing tears to my eyes, even now as I write

about it. I will always cherish the memories. I also thank the support from Damsas and my love Ahil. Although I lost a part of my family, I am extremely lucky to gain a wonderful family in uppa, umma, Raza, Diya, Heena, and lovely Liu. Uppa never let me miss my late uppa and umma's prayers have kept me strong during adversities; I can never thank them enough for the love and care they showered on me. With Raza, I've been fortunate to not only have a wonderful friend but also to see our bond transition into a familial relationship. Diya is more like a best friend than sister, who is always available on the other end of the phone. Heena, you have always supported me like a brother, I sincerely appreciate it. Ahil and Baby Liu brought an abundance of joy and love into my life with their infectious laughter and adorable antics. Their innocent smiles have been a source of endless happiness. I would also like to acknowledge my uncles, aunts, and cousins for their love and support. Finally, but the most, I would like to thank my wife Thanu for her patience and support. She had been very patient with me and allowed me to focus on research while understanding and forgiving my shortcomings.

Dedication

To ...

uppa, umma, mon, mol, Thanu,

... for their love.

I will miss you uppa and mon.

Abstract

Modeling and representing complex systems by a network of interacting agents provide cost-effective and efficient ways to simulate, design, and test the solutions to real-world problems in science and engineering. Applications such as finance, neuroscience, climate science, power grids, etc. involve modeling the agents and states that interact and evolve dynamically and can contain feedback loops. Most existing literature assumes a static relationship between the agents, which fails to capture the dependencies across time. Learning the interdependency (topology) and cause-and-effect structure among the agents is therefore of importance and having an active interest in the graphical network representation and effective modeling of the dynamical systems.

Identifying the unknown interaction structure can in general be classified into active and passive techniques. The active techniques involve intervention in the normal operation of the system by injecting external signals and/or modifying agents in the system. Critical applications in financial markets, power grids, meteorological systems etc. do not allow active intervention in the system. In such critical applications passive methods, which infer the information from the observed time-series measurements, are applied. In the passive techniques, the evolution of the underlying system is mathematically modeled using a generative model, where the agents are excited with a disturbance signal.

In this thesis, we study passive techniques to identify the interdependency and the influence pathways in the dynamical systems, where the agents interact linearly. Further, in practice, it might not be viable to observe all the states in the system. Thus, it is imperative to identify the influence structure (network topology) when the system involves hidden/latent states, which is a challenging problem. In Chapter 2, we provide a novel algorithm to identify the interdependency structure, including the dependency among the latent nodes, when only a subset of the nodes/agents provide the time-series measurements. In many applications, the proposed algorithm retrieves the exact influence pathways, with partial directions, including that of the latent nodes. This chapter is based on the article published in IEEE Transactions on Automatic Control [1]

Another practical issue is that some of the disturbances at the agents themselves can be correlated, which renders the agents' influence structure indistinguishable from passive observations. Examples of such systems are observable in applications such as

power grids and stock markets. Learning the network topology in the presence of such correlated disturbances is difficult as it is required to extract the dependency between the disturbances first. In Chapter 3, we show that the correlation between the disturbances is equivalent to the dependency induced by the latent nodes. This chapter is based on the article published in *Automatica* [2]

The network topology only provides the correlation structure of the network of agents. It is well known that correlation is not equivalent to causation and the cause-effect structure contains more information than the correlation structure. In many applications in medicine, biology, and statistics the correlation structure might not be sufficient. Here, it is necessary to know the causal structure of the underlying dynamics. In Chapter 4, we provide an algorithm to learn the causal structure of a linear dynamical system when the agents are excited by identical noise sources. Further, a sample complexity analysis is provided, which finds matching upper and lower bounds on the number of samples required to obtain a given error performance. This chapter is based on the article accepted for future publication in *AISTATS 2024* [3].

In general, without intervention and assumptions on the generative models, it is impossible to identify a complete causal structure. The best one can identify without further restriction is an equivalent set of graphs with the same conditional correlation structure, Essential Graphs. In Chapter 5, we provide algorithms to estimate the essential graph in linear dynamical systems using the Wiener filter (WF). The conventional time-domain approach to computing WF is computationally expensive. To speed up the calculation, we propose a fast Fourier transform-based computationally efficient approach to estimate the Wiener filter. The conventional probability notion of CI fails to capture the CI notion in a stochastic process. To tackle this problem, we propose a novel probability notion of CI for stochastic processes in the frequency domain. This notion is used to extend the notions of the front-door and the back-door criterion from static graphical models to linear dynamical systems, where the nodal states are stochastic processes. This chapter is based on the article [4].

Contents

Acknowledgements	i
Dedication	v
Abstract	vi
Contents	viii
List of Figures	xiv
1 Introduction	1
1.1 Network Reconstruction and Prior Works	2
1.2 Summary of Contributions	5
1.3 Publications	7
2 Topology Learning of Linear Dynamical Systems with Latent Nodes using Matrix Decomposition	9
2.1 Introduction	10
2.2 Sparse plus Low-Rank Matrix Decomposition for Skew Symmetric Matrices	14
2.2.1 Optimization for Sparse plus Low-rank Decomposition	14
2.2.2 Affine Varieties and Tangent Spaces	15
2.2.3 Sparse Plus Low-Rank Decomposition using Optimization	17
2.2.4 Sufficient Conditions to Retrieve the True Matrices	18
2.3 Linear Dynamical Systems	21
2.3.1 Graphical Representation	21
2.4 Exact Reconstruction of Moral Graph of Observed Nodes and Markov Blanket of Latent Nodes	22

2.4.1	Moral Graph Reconstruction under Complete Observability . . .	22
2.4.2	Structure of the IPSDM with Latent Nodes	23
2.4.3	Identifiability of Latent Nodes	25
2.4.4	Markov Blanket Reconstruction of the Latent nodes	27
2.4.5	Moral Graph Reconstruction of Observable Nodes and Localiza- tion of Hidden Nodes from IPSDM	28
2.4.6	Reconstruction based on the Imaginary Part of IPSDM	29
2.5	Exact Topology Reconstruction	30
2.5.1	Elimination of Strict Spouse Edges	30
2.5.2	Reconstruction of Topology:	31
2.6	IPSD Estimation from Finite time-series	36
2.6.1	Estimation Error	36
2.6.2	IPSDM Estimation Error	38
2.7	Simulation results	39
2.7.1	LDM	39
2.7.2	Sparse plus low-rank decomposition of the Imaginary of IPSDM	39
2.7.3	Reconstruction of Topology	42
2.8	Conclusions	43
3	Topology Identification under Spatially Correlated Noise	45
3.1	Introduction	45
3.2	System model	49
3.3	IPSDM based topology reconstruction	50
3.4	Spatial Correlation to Latent Node Transformation	50
3.4.1	Relation between Spatial Correlation and Latent Nodes	51
3.4.2	Uniqueness of the topology	55
3.5	Polynomial Correlation	56
3.5.1	Characterization of Polynomial Correlations	57
3.5.2	Example: Lifting of a zero mean Gaussian Process	57
3.5.3	Transformation of Polynomial Correlation to Latent Nodes	60
3.6	Moral-Graph Reconstruction by Matrix Decomposition	62
3.6.1	Topology Reconstruction under Affine Correlation	62
3.6.2	Topology reconstruction under polynomial correlation	63
3.6.3	Low-rank plus Sparse Matrix Decomposition	64
3.6.4	Sufficient Condition for Sparse Low-rank Matrix Decomposition	65

3.7	Simulation results	67
3.7.1	Polynomial correlation	68
3.7.2	Finite data simulation	68
3.8	Conclusion	74
4	Information Theoretically Optimal Sample Complexity of Learning Dynamical Directed Acyclic Graphs	75
4.1	Introduction	76
4.1.1	Related Work	76
4.2	System Model of DDAG	78
4.3	Reconstructing DDAGs from PSDM	82
4.3.1	CPSD and Topological Ordering	82
4.4	Finite Sample Analysis of Reconstructing DDAGs	85
4.4.1	Non-Asymptotic Estimation Error in Spectrogram Method	86
4.4.2	Sample Complexity Bounds: Upper Bound	87
4.5	Sample Complexity Bounds: Lower Bound	88
4.6	Simulation Experiments	89
5	Causal Structure Recovery of Linear Dynamical Systems: An FFT based Approach	92
5.1	Introduction and literature survey	93
5.2	Preliminaries	95
5.2.1	Linear Dynamical Influence Models (LDIM)	95
5.2.2	Frequency Discretization	96
5.2.3	Graph Definitions	96
5.2.4	Graph/Topology Learning	97
5.2.5	Probabilistic Conditional Independence on a Set of Stochastic Processes and Factorization According to Directed Graphs	98
5.3	Main Results	101
5.3.1	Computational Advantages of FFT based Wiener filter	101
5.3.2	Realizing the Essential Graph of SPs; using Phase Properties of Frequency Dependent Wiener Filtering	101
5.3.3	Causal Inferences on Set of Stochastic Processes via Analysis of Dependencies at Specified Frequencies	102
5.4	Complexity Analysis of Wiener filter	103

5.4.1	Computational Complexity of Estimating Wiener Filter	103
5.4.2	Sample Complexity Analysis	104
5.5	Limitations and Future Works	105
5.6	Simulation Results	105
6	Conclusions and Future Research	108
6.1	Summary of Completed Works	108
6.2	Future Works	109
	References	111
	Appendix A. Appendix of Chapter 2	124
A.1	Proof of Lemma 2.2.1	124
A.2	Proof of Lemma 2.2.2	125
A.3	Proof of Lemma 2.2.3 :	125
A.4	Proof of Proposition 2.2.1:	126
A.5	Proof of Proposition 2.2.2:	128
A.6	Proof of Proposition 2.2.3	129
A.7	Proof of Theorem 2.4.1	130
A.8	Proof of Theorem 2.4.2	130
A.9	Proof of Theorem 2.4.3	132
A.10	Proof of Theorem 2.5.1	133
A.11	Proof of Theorem 2.5.2:	133
A.11.1	LDG example:	135
A.12	Proof of Proposition 2.6.1:	136
A.13	Proof of Proposition 2.6.2:	136
A.14	Proof of Lemma 2.6.1:	138
A.15	Proof of Theorem 2.6.2:	138
	Appendix B. Appendix of Chapter 3	140
B.1	Proof of Lemma 3.4.1	140
B.2	Proof of Lemma 3.4.2	140
B.3	Proof of Theorem 3.4.1	141
B.4	Proof of Proposition 3.4.1	142
B.5	Proof of Proposition 3.5.1	142
B.6	Proof of Lemma 3.5.1	143

B.7	Proof of Theorem 3.5.1	143
B.8	Proof of Theorem 3.6.1	143
Appendix C. Appendix of Chapter 4		144
C.1	Proof of Lemma 4.3.2	144
C.2	Proof of Lemma 4.3.6	146
C.3	Proof of Theorem 4.4.1	147
C.3.1	Restart and Record Sampling	148
C.3.2	Continuous Sampling	150
C.4	Proof of Lemma 4.4.2	155
C.5	Lower bound: Proof of Theorem 4.5.1	157
Appendix D. Appendix of Chapter 5		161
D.1	Wiener-Phase Algorithm	162
D.2	Simulation	163
D.2.1	Restart and Record Sampling	165
D.3	Convergence	166
D.4	Proof of Lemmas	169
D.4.1	Proof sketch of Lemma 5.2.1	169
D.4.2	Proof of Lemma 5.3.2	169
D.4.3	Proof of Lemma 5.3.3	169
D.4.4	Proof of Lemma 5.3.4	170
D.5	Which frequency to choose?	170
D.6	Direct and Total Effect Identification in SPs	170
D.6.1	Direct Effect Identification	170
D.6.2	Total Effect Identification	171
D.7	Proof of Theorem D.6.1 (Single-door)	172
D.7.1	Proof: single-door	173
D.8	Proof of Theorem D.6.2	175
D.8.1	Atomic interventions	175
D.8.2	Back-door Criterion	178
D.9	Proof of Theorem 5.4.1	178
D.10	Primer on Wiener Projection of Time-Series and its Application in Parameter Identification	180
D.10.1	Parameter Identification	182

D.10.2 Motivating Example for Back-door criterion	183
D.10.3 Proof of Proposition D.10.1	184
D.10.4 Proof of Lemma D.7.1	185
D.11 A Primer on Stochastic Processes and Conditional Independence Notion	
in Time Domain	185
D.11.1 FDD based Independence in Stochastic Process	187

List of Figures

2.1	Non-identifiability of hidden nodes (red colored nodes): (a) hidden node is terminal and strict parent; (b) hidden node is terminal and strict child; (c) Markov Blankets of hidden nodes 1 and 2 overlapped.	26
2.2	LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with hidden nodes shown in red.	40
2.3	Comparison between tol_t and $diff_t$ for decomposition of $\mathfrak{S}\{\Phi_{oo}^{-1}\}$ obtained by the LDG in Figure 2.2	40
2.4	Applying Algorithms 2 on $\mathfrak{S}\{\mathbf{S}\}$ and Algorithm 3 on $\mathfrak{S}\{\mathbf{L}\}$	41
2.5	$\mathcal{T}(\mathcal{V}_R, \mathcal{E}_R)$: Final reconstructed topology with hidden nodes.	41
3.1	Transformation of an LDIM with correlated noise sources to a network with uncorrelated noise sources in the presence of a latent node.	51
3.2	Demonstration of $\mathcal{L}(\mathbf{H}, \mathbf{e})$; noise sources are not shown. The exogenous noises are uncorrelated in Figure 3.2(c) and 3.2(d).	53
3.3	Transformation of an LDIM with correlated noise sources to a network with uncorrelated noise sources in the presence of a latent node.	59
3.4	Network structure	69
3.5	Transformed network structure	70
3.6	tol_t and $diff_t$ plots for the network of Figure 3.4(a) under affine correlation	70
3.7	tol_t and $diff_t$ plots for the network of Figure 3.4(a) under polynomial correlation	71
3.8	The true topology and the topology estimated directly from $\hat{\Phi}_o^{-1}(e^{j2\pi/5})$ for the network shown in Figure 3.4(a) under affine correlation with $N = 10^6$ data samples. The figures show the entries of matrices of size 29×29 . In the figures of topology, black denotes zero (edge absent) and white denotes 1 (edge present). In the gray scale images, gray denotes zero. . .	72

3.9	$\hat{\mathbf{S}}_t$ obtained by solving (3.16) for $\mathbf{C} = \Im\{\hat{\Phi}_o^{-1}(e^{j2\pi/5})\}$. Topology is estimated from $\hat{\mathbf{S}}_t$ for $t = 0.34$. In the figures of topology, black denotes zero (edge absent) and white denotes 1 (edge present). In the gray scale images, gray denotes zero.	73
4.1	An example DDAG. Node 1 is an ancestor and node 7 is a descendant of every node in the graph. The set $\{1, 2, 5\}$ is an ancestral set but $\{2, 5\}$ is not. $an(3) = \{1, 2\}$, $desc(3) = \{4, 7\}$, $nd(3) = \{1, 2, 5, 6\}$	79
4.2	(a) shows restart and record sampling. (b) shows continuous sampling. .	81
4.3	Probability of error with number of trajectories, for different networks under restart and record (RR) and continuous (conti) sampling strategies, when system is excited by either i.i.d. or WSS noise. p and q refer to the number of nodes and degree of each node in the network, respectively. The trajectory length $N = 64$ is fixed for each trajectory, and $\omega = \frac{17}{64}$. .	90
5.1	(a) shows the true causal generative graph; (b) shows the estimated kin-graph from $T = 10,000$ samples with $N = 64$ using $\ell_\infty(\Omega_N)$ norm on $W_i[j]$; (c) is the CPDAG estimated using Algorithm 6. The bi-directed edges indicate undirected edges in (c); (d) shows the graph estimated by performing the intervention at node 2.	106
5.2	(a) shows the phase response of $W_i[j]$ for $i = 1, 2$ and $j = 4, 5$. (b) shows the plot of real and imaginary parts of the mean vs ω of the estimated conditional distribution, when conditioned on $do(X_2 = y_1)$ and $do(X_2 = y_2)$	106
A.1	(a) LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$, (b) $top(\mathcal{G})$ and (c) $kin(\mathcal{G})$. $\{2, 3\}$ are strict parents of 1. $\{6, 9\}$ are strict spouses of 1.	136
D.1	(a) shows the true causal generative graph; (b) shows the estimated kin-graph from $T = 10,000$ samples with $N = 64$ using $\ell_\infty(\Omega_N)$ norm on $W_i[j]$; (c) is the CPDAG estimated using Algorithm 6. The bi-directed edges indicate undirected edges in (c); (d) shows the graph estimated by performing intervention at node 2.	164
D.2	Phase plot of the Wiener coefficients, $W_i[j]$, computed using equation (5.5); data with 10,000 samples is generated using (D.1). Red plots show the phase plot, $\angle W_i[j]$, between the strict spouses i and j . σ is the standard deviation of the phase along the frequency. The spurious links $(i, j) = (2, 4)$ and $(i, j) = (3, 4)$ have $\sigma < 0.9$ whereas the remaining pair of nodes have $\sigma > 1.6$	164

D.3	Estimated conditional distribution when intervened at node 2 for two interventions, $do(X_2 = y_1)$ and $do(X_2 = y_2)$, where y_1 is all zero sequence and y_2 is a repeating sequence of 31 zeros and 33 twos. The first row shows the real part of the mean, the middle row is the imaginary part of the mean, and the bottom row is the standard deviation of the estimated distribution. The plots show that the mean of X_i , especially the mean of the imaginary part, follows the pattern of X_2 when there is a directed path to i from 2 ($i = 3, 5$). The mean of X_1 and X_4 are random, and does not follow the mean of X_2 , implying that there is no directed path to 1 and 4 from 2.	165
D.4	Phase plot of the Wiener coefficients, $W_i[j]$, computed using equation (5.5) for the restart and record data; data with 10,000 trajectories is generated for the AR in (D.1). Red plots show the phase plot, $\angle W_i[j]$, between the strict spouses i and j . σ is the standard deviation of the phase along the frequency. The spurious links $(i, j) = (2, 4)$ and $(i, j) = (3, 4)$ have $\sigma < 0.1$ whereas the remaining pairs of nodes have $\sigma > 1.5$	166
D.5	Estimated conditional distribution when intervened at node 2 for two interventions, $do(X_2 = y_1)$ and $do(X_2 = y_2)$, where y_1 is all zero sequence and y_2 is a repeating sequence of 31 zeros and 33 twos. The plots show that the mean of X_i , especially the mean of the imaginary part follows the pattern of X_2 when there is a directed path to i from 2 ($i = 3, 5$). The mean of X_1 and X_4 are random, and does not follow the mean of X_2 , implying that there is no directed path to 1 and 4 from 2.	167
D.6	Example LDIM with $W_{12,3}(\omega) = \beta(\omega), \forall \omega \in \Omega$	182
D.7	SEM 1	183
D.8	Graph 1	183

Chapter 1

Introduction

Modeling and representing complex systems by a network of interacting agents provide cost-effective and efficient ways to simulate, design, and test the solutions to real-world problems in science and engineering. Applications such as finance, neuroscience, climate science, power grids, etc. involve modeling the agents and states that interact and evolve dynamically and can contain feedback loops. Learning the interdependency (topology) and cause-and-effect structure among the agents is therefore of importance in the graphical network representation and effective modeling of the dynamical systems. Learning the unknown interaction structure of a network of agents from time-series measurements can be categorized as being active [5] or passive [6]. Active techniques require intervention in the normal operation of the system by injection of external signals and/or altering the network structure, by removing or adding agents to the network. Many critical systems such as the power grid, the financial markets, and the meteorological system either do not allow for active interventions or it is not possible to affect the system. On the contrary, passive techniques infer topology from time-series measurements, without affecting the underlying grid network.

Learning the topology of a network from time-series data is an active area of research, with considerable emphasis from the machine learning and probabilistic graphical model communities (see [5]-[10] and the references therein). However, here, most works assume that the nodes are random variables which fail to capture the dynamics of the interaction and thus are improper for the applications with dynamic dependencies that are common, for example, in the power grid [11] and its application to smart grid networks [12], climate science [13] and finance [14]. Moreover, in many scenarios, topology identification is the first step in system identification[15, 16].

1.1 Network Reconstruction and Prior Works

Filtering based topology reconstruction has gained considerable attention recently for unveiling the topology of dynamically related agents. In [17], the moral graph of a directed network is reconstructed using the magnitude response of multivariate Wiener filters. [18] provided a method that showed that the spurious links present in the moral graph can be removed by checking the phase response of the Wiener filter between the links. [17] and [19] provided algorithms for exact network reconstruction where all parent-child relations are uncovered but the results are restricted using Granger causality to systems with strictly causal dynamical dependencies. The aforementioned works, [17]-[19], assumed full network observability. Several works, [21]-[26], have studied topology identification in the presence of hidden nodes, but, restricted to radial tree topologies—characterized by undirected tree topology in [21], for polyforest networks in [22], and polytree networks in [23]-[25]. Network reconstruction with corrupted data streams was explored in [29]-[31].

In [32]-[38], the authors considered the problem of estimating conditional dependency structure of autoregressive (AR) Gaussian stochastic processes in the presence of latent nodes, with an emphasis on finiteness of the time-series available. Here, the problem is formulated in terms of sparse plus low-rank decomposition of the inverse of the power spectral density matrix (IPSDM). The articles [32]-[38] provided interesting optimization frameworks and theoretical guarantees to identify the conditional dependency from estimated IPSDM. Here, the graphical representations of conditional dependencies reconstructed from the IPSDM retrieves the moral graph [17]. As shown in [17, 39], moral graphs can admit many spurious edges.

In practice, observing time-series measurement at every node is not plausible, wherein it becomes important to learn the topology of the network when only a subset of the nodes are observed. Further, the moral graph relations can admit many spurious connections not present in the original topology. The thesis, especially, Chapter 2 demonstrate that the rank-sparsity patterns induced by the network topology on the imaginary part has properties that can be exploited toward the *exact* reconstruction of network topology. Here, under assumptions applicable for a large class of systems, all spurious links in the moral graph formed by observed nodes can be identified. Moreover, assuming conditions on hidden (latent) nodes, which follow from identifiability conditions, links between the hidden and observed nodes can be reconstructed resulting in the retrieval of the exact topology of the network. This chapter also serves as an important bridge

between the works presented in [40] and the works related to network structure reconstruction [17]-[28]. Furthermore, results here are applicable to many classes of directed graphs without self loops, not restricted to directed acyclic graphs or bi-directed graphs, unlike [18, 21].

One major caveat of the aforementioned literature on graphical models is that the results fail if the exogenous noise sources are spatially correlated, i.e. if the noise is correlated across the agents/variables. Prior works have studied the systems with spatially correlated noise sources [41, 27, 16]; however, these studies assume the knowledge of topology of the network. In a related work, [42] studied topology estimation under spatially correlated noise sources and used this estimated topology in local module identification. Chapter 3 studies the problem of topology identification of the LDIMs that allow spatially *correlated* noise sources, similar to the problem in [42]. However, Chapter 3 provides an alternate treatment, where the noise correlations are transformed to latent nodes. This transformation enables one to gain additional insights and apply techniques from topology identification with latent nodes to solve the problem.

All the aforementioned literature discussed correlation structure reconstruction. It is well known that correlation is not causation. Causal structure recovery is an important problem with applications in neuroscience [43], finance [44], meteorology [13], etc. The structure of directed interactions in a network of agents is conveniently represented using directed graphs, with the agents as nodes and the directed interactions as directed edges. If the underlying graph doesn't have cycles, it is called a Directed Acyclic Graph (DAG). In general, it is not possible to reconstruct the exact structure with the direction of different edges. Instead, in many networks it is possible to retrieve only the Markov equivalence graphs, the set of graphs satisfying the same conditional dependence property, from data without any intervention [45]. In applications such as finance [44], climate science [13] etc., the agent states, instead of being temporally independent, can evolve over time due to past directed interactions. Such temporal evolution can be represented by a linear dynamical system (LDS). In LDS, the interaction between agent states is captured by a linear time-invariant function. Chapter 4 studies the identifiability and present the sample complexity results for learning a DAG of LDS, which we term as *Dynamical* DAG or **DDAG**. This is distinguished from *static* DAG, where the agent states are temporally independent and the DAG does not correspond to temporal dynamics. Further, chapter 4 presents an information-theoretically optimal sample complexity analysis for learning a dynamical DAG (DDAG) excited by wide-sense stationary (WSS, i.e., non i.i.d) noise of equal power spectral density, using samples

of state trajectories of the underlying linear dynamical system (LDS).

The causal structure recovery results in Chapter 4 required the identifiability assumptions that either 1) the excitation noise sources are identical, or 2) the second order statistics of the excitation noise sources are known. Chapter 5 extends the causal structure recovery to more general set up without the two aforementioned assumptions. Considerable prior art and algorithms exist for recovering the causal graph of entities modeled as random variables, when interventions on the system are possible [46]. In a related problem of estimating the causal effect given the causal graph, the goal is to estimate the effect of a variable on another. Here, a powerful framework of do-calculus can be employed to ascertain the set of variables that need to be used to determine the total and direct effect of a variable on another which result in criteria such as the single-door, back-door and front-door criteria [47, 48].

As mentioned before, causal inference is more challenging in the presence of dynamical (across time) dependencies [49, 50, 51]. One of the established techniques for identifying the causation is Granger causality [52], which utilizes how well a set of variables can be employed to estimate a given variable while utilizing the time-arrow. Granger causality based approaches necessitate presence of delays in the dynamic dependencies. Such an assumption is rendered problematic when the data is collected at slower rates than the time-constants at which the dynamics evolve, thus precluding a number of practical scenarios. Another approach in handling dependencies across times is to consider the series at every time instant as a random variable thereby mapping the problem to a static version; the difficulties with such an approach stem from the lack of information on the size of the horizon in the past and future to be considered and the combinatorial explosion of the number of variables that entail [53, 54]. In a class of approaches for unveiling dynamic dependencies, models that include vector autoregressive (VAR) models [50, 51], additive noise models (ANM) [49, 51] and neural network based models (NNM) [55], of how the data is being generated is assumed and the causal graph structure is recovered via an estimation of model parameters. In another recent line of work, the main gist of the approach is to exploit the asymmetry in the relationships of a dependency and the inverse of the dependency. These asymmetries can result from nonlinear maps or from linear filters with inverses that have discernable differences that, for example, are characterizable via power spectral densities [56, 57]. Here, it is to be noted that works such as [56, 57] try to infer cause and effect relationship between two variables; however, the causal graph structure recovery entails determining how the influence flows from a variable to another with the possibility of intermediate

variables between the cause and the effected variables. We emphasize that effective methodologies for causal inference for systems with dynamic interactions remain much less developed than their static dependency counterparts. Several recent works have bridged the causality literature ([46]) with multivariate projections of dynamically related time-series data [49, 58, 27]. [27] extended the single-door criterion [48] to include cycles and proposed a similar criterion called revolving-door criterion. It was shown in [17] that the Wiener projections can be employed to recover the moral graph of entities interacting via linear time-invariant dynamics. Chapter 5 develops a framework for causal graph reconstruction for a system with dynamical interactions using a frequency domain approach, using both active (with intervention) and passive (from time series) approaches. Further, two probabilistic notions of independence for stochastic processes are provided, which is applied to extend single-,back-, and front-door criteria to networks with linear dynamical interactions.

1.2 Summary of Contributions

This thesis combines the graduate work that have appeared in various journals and conferences, and is organized as follows.

■ **Chapter 2** provides a novel algorithm for identifying the network topology in the presence of latent nodes. The algorithm reconstructs the moral graph of observed nodes and the Markov Blanket of latent nodes from the matrix decomposition of the Inverse of the Power Spectral Density Matrix (IPSDM); accounting for network identifiability issues associated with the latent nodes. The reconstruction algorithm relies on the sparse plus low-rank decomposition of the IPSDM into the unique sparse and low-rank parts. The chapter also provides necessary and sufficient conditions to obtain the unique sparse plus low-rank matrix decomposition of a given matrix, in the space of skew symmetric matrices. It is established that the unique decomposition of the imaginary part of IPSDM, which is skew symmetric, is sufficient to recover the exact topology of the entire network, in a large class of systems. For the more practical scenarios where we have access only to finite samples of time-series at each node, the chapter provides a concentration bound for estimation error of IPSDM.

■ **Chapter 3** provide approaches to reconstruct the topology of a network of agents excited by exogenous stochastic sources that are possibly correlated across the agents, from time-series measurements alone. The first major result of this chapter is to provide a transformation that converts an LDIM without latent nodes, but excited by spatially

correlated exogenous noise sources, to LDIMs with latent nodes. Here, the latent nodes are characterized by the maximal cliques in the correlation graph, the undirected graph that represents the spatial correlation structure. It is shown that, under affine correlation assumption, that is, when the correlated noise sources are related in affine way (Assumption 3.4.1), there exist transformed LDIMs with latent nodes, where all the nodes are excited by spatially uncorrelated noise sources. A key feature of the transformation is that the correlations are completely captured using the latent nodes, while the original topology remains unaltered. The original moral-graph/topology is the same as the moral-graph/topology among the observed nodes in the altered graph. Thus, this transformed problem is shown to be equivalent to topology identification of networks with latent nodes. Consequently, any of the aforementioned techniques for the networks with latent node can be applied on the transformed LDIM to reconstruct the original moral graph/topology.

Next, relaxing the affine correlation assumption, polynomial correlation is considered (Assumption 3.5.1); here, the focus is on noise sources with distributions that are symmetric around the mean. It is shown that, in this scenario, the transformed dynamical model can be excited by clusters of noise sources, where the clusters are uncorrelated. However, the noise sources inside the clusters can be correlated. Using the sparse+low-rank decomposition of IPSDM from [1], the true topology of the network along with the correlation structure is reconstructed, from the IPSDM of the original LDIM, without any additional information, if the network satisfies a necessary and sufficient condition. Notice that, the results discussed here are applicable to the networks with static random variables and AR models also, when the exogenous noise sources are spatially correlated.

■ **Chapter 4** provides an information-theoretically optimal sample complexity analysis for learning a dynamical DAG (DDAG) excited by wide-sense stationary (WSS, i.e., non i.i.d) noise of equal power spectral density, using samples of state trajectories of the underlying linear dynamical system (LDS). The chapter considers learning under two sampling scenarios, viz; 1) restart and record, 2) continuous sampling. While the former pertains to samples collected from disjoint (independent) trajectories of state evolution, the latter includes samples from a single but longer state trajectory (see Fig. 4.3). Surprisingly, the results in this article show that the estimation errors are not influenced by the sampling strategy (restart and record, or continuous) as long as the number of collected samples are over a determined threshold given by $n = O(q \log(p/q))$, where p is the number of nodes and q is the maximum number of parents per node. The chapter

also provides a matching information-theoretic lower-bound, $\max\left(\frac{\log p}{2\beta^2 + \beta^4}, \frac{q \log(p/q)}{M^2 - 1}\right)$, where β and M are system parameters (see Definition 4.2.1); thus obtaining an order optimal bound $n = \Theta(q \log(p/q))$.

■ **Chapter 5** develops a framework for a system with dynamical interactions using a frequency domain approach. The first part of the chapter applies Wiener filtering in the frequency domain. Here, the main steps taken for reconstruction of the moral graph (which recovers the kin structure) are: transform the time-series of each entity (modeled as a stochastic process) to the frequency domain using Fast Fourier Transform (FFTs), perform multivariate projections in the *frequency domain*, and construct an undirected graph using the non-zero entries of the obtained multivariate projection filter. Under the assumption that the data generative model is linear and time-invariant it is shown that the above process recovers the moral graph. With the standard assumption of faithfulness, a process akin to the Peter-Clark (PC) algorithm is employed to recover the essential graph with better computational efficiency, which exploits the phase properties of the complex valued estimates. When interventions are possible, it is shown that the causal graph can be recovered. It is demonstrated that the computational complexity for recovering the causal structure using the projections in frequency domain is $O(Tn^3 \log N)$ compared to the conventional projections using time-domain which are shown to have a complexity of $O(Tn^3 N^2)$; n , T and N are the number of nodes, number of samples, and the largest time-lag in the dependency between entities, respectively. In the second part of the chapter, two probabilistic notions of independence for stochastic processes are developed with guarantees on the convergence of the FFT model to the continuous frequency model provided. These notions help generalize and streamline the framework based on joint probability distributions in the static case to the dynamic case with the mechanics of do-calculus made possible in the dynamics case. For estimating Wiener coefficients from time-series data, a non-asymptotic concentration bound and a sample complexity analysis are also provided. Furthermore, for causal-effect estimation for LTI dependencies, back-door, single-door (with cycles) and front door criteria are obtained with proofs that are rendered straightforward in the frequency domain.

1.3 Publications

A large share of this thesis is based on the following publications:

Journal Articles:

- [1] Veedu, Mishfad Shaikh, Harish Doddi, and Murti V. Salapaka. "Topology learning of linear dynamical systems with latent nodes using matrix decomposition." *IEEE Transactions on Automatic Control* 67, no. 11 (2021): 5746-5761.
- [2] Veedu, Mishfad Shaikh, and Murti V. Salapaka. "Topology identification under spatially correlated noise." *Automatica* 156 (2023): 111182.

Conference Articles:

- [1] Veedu, M.S., Deka, D. and Salapaka, M., 2024, April. Information Theoretically Optimal Sample Complexity of Learning Dynamical Directed Acyclic Graphs. In *International Conference on Artificial Intelligence and Statistics* (pp. 4636-4644). PMLR.

Preprints:

- [1] Veedu, M.S., Melbourne, J. and Salapaka, M.V., 2023. Causal Structure Recovery of Linear Dynamical Systems: An FFT based Approach. *arXiv preprint arXiv:2309.02571*.

Chapter 2

Topology Learning of Linear Dynamical Systems with Latent Nodes using Matrix Decomposition

In this chapter, we present a novel approach to reconstruct the topology of networked linear dynamical systems with latent nodes. The network is allowed to have directed loops and bi-directed edges. The main approach relies on the unique decomposition of the inverse of power spectral density matrix (IPSDM) obtained from observed nodes as a sum of sparse and low-rank matrices. We provide conditions and methods for decomposing the IPSDM of the observed nodes into sparse and low-rank components. The sparse component yields the moral graph associated with the observed nodes, and the low-rank component retrieves parents, children and spouses (the Markov Blanket) of the hidden nodes. The chapter provides necessary and sufficient conditions for the unique decomposition of a given skew symmetric matrix into sum of a sparse skew symmetric and a low-rank skew symmetric matrices. It is shown that for a large class of systems, the unique decomposition of imaginary part of the IPSDM of observed nodes, a skew symmetric matrix, into the sparse and the low-rank components is sufficient to identify the moral graph of the observed nodes as well as the Markov Blanket of latent nodes. For a large class of systems, all spurious links in the moral graph formed by the observed nodes can be identified. Assuming conditions on hidden nodes required

for identifiability, links between the hidden and observed nodes can be reconstructed, resulting in the retrieval of the exact topology of the network from the availability of IPSDM. Moreover, for finite number of data samples, we provide concentration bounds on the entry-wise distance between the true IPSDM and the estimated IPSDM.

2.1 Introduction

Networks provide convenient representation of large scale complex systems utilized in diverse areas such as power grids, biology, finance, and neuroscience. Reconstructing the underlying network topology or the influence structure of the interaction from measurements is useful in predicting, and steering the behavior of the system towards a desired state. Learning the unknown interaction structure of a network of agents from time-series measurements can be categorized as being active [5] or passive [6]. Active techniques require intervention in the normal operation of the system by injection of external signals and/or altering the network structure, by removing or adding agents to the network. Many critical systems such as the power grid, the financial markets, and the meteorological system either do not allow for active interventions or it is not possible to affect the system. On the contrary, passive techniques infer topology from time-series measurements, without affecting the underlying grid network. Here, in practice, observing time-series measurement at every node is not plausible, wherein it becomes important to learn the topology of the network when only a subset of the nodes are observed.

Learning the topology of a network from time-series data is an active area of research, with considerable emphasis from the machine learning and probabilistic graphical model communities (see [5]-[10]). However, here, most works assume that the nodes are random variables which fail to capture the dynamics of the interaction and thus are improper for the applications with dynamic dependencies that are common, for example, in the power grid [11] and its application to smart grid networks [12], climate science [13] and finance [14]. Moreover, in many scenarios, topology identification is the first step in system identification [15, 16].

Filtering based topology reconstruction has gained considerable attention recently for unveiling the topology of dynamically related agents. In [17], the moral graph of a directed network is reconstructed using the magnitude response of multivariate Wiener filters. [18] provided a method that showed that the spurious links present in the moral graph can be removed by checking the phase response of the Wiener filter between

the links. [17] and [19] provided algorithms for exact network reconstruction where all parent-child relations are uncovered but the results are restricted using Granger causality to systems with strictly causal dynamical dependencies. The aforementioned works, [17]-[19], assumed full network observability. Several works, [21]-[26], have studied topology identification in the presence of hidden nodes, but, restricted to radial tree topologies—characterized by undirected tree topology in [21], for polyforest networks in [22], and polytree networks in [23]-[25]. Network reconstruction with corrupted data streams was explored in [29]-[31].

In [32]-[38], the authors considered the problem of estimating conditional dependency structure of autoregressive (AR) Gaussian stochastic processes in the presence of latent nodes, with an emphasis on finiteness of the time-series available. Here, the problem is formulated in terms of sparse plus low-rank decomposition of the inverse of the power spectral density matrix (IPSDM). The articles [32]-[38] provided interesting optimization frameworks and theoretical guarantees to identify the conditional dependency from estimated IPSDM. Here, the graphical representations of conditional dependencies reconstructed from the IPSDM retrieves the moral graph [17]. As shown in [17, 39], moral graphs can admit many spurious edges.

In this chapter, we approach the problem of reconstructing the topology in the presence of latent nodes. Similar to [32]-[38] a perspective of sparse plus low-rank decomposition of IPSDM associated with the observed nodes is taken.

Extending the results in [40], this chapter establishes conditions for a skew symmetric to admit a unique decomposition as a sparse and low-rank matrix. Towards decomposing a skew symmetric matrix, this chapter provides a characterization for tangent manifolds of skew symmetric matrices with a fixed rank and with a given sparsity pattern. Furthermore, the chapter provides additional theoretical insights into an empirical procedure presented in [40], which provides a sufficient condition that can be tractably assessed for a unique decomposition. Though not emphasized in the chapter, the methodology developed can be used to realize similar results for unique decomposition of Hermitian matrices as well.

Based on the extensions of the results in [40] established here, it is possible to obtain a decomposition of the IPSDM of the observed nodes, where the sparse part can be leveraged to realize the moral graph formed by the observed nodes. We establish identifiability conditions under which the low-rank component of the observed nodes' IPSDM yields the parents, children and spouses (all the two hop neighbors that form the Markov Blanket) of the latent (hidden) nodes. The IPSDM, being complex

and frequency dependent, has real and imaginary parts. We further emphasize the imaginary part of the IPSDM in the chapter as it has considerable structure applicable for a large class of problems. Here too the sparse component of the imaginary part of the IPSDM matrix, which is skew symmetric, is shown to yield the moral graph of networks governed by a linear dynamical model (defined later) that encompass a wide class of systems. The above approach can be employed toward the retrieval of the moral graph of networks of AR models and is applicable to networked systems addressed in [32] albeit, here we do not emphasize the finite data aspects.

The moral graph relations can admit many spurious connections not present in the original topology. We demonstrate that the rank-sparsity patterns induced by the network topology on the imaginary part has properties that can be exploited toward the *exact* reconstruction of network topology. Here, under assumptions applicable for a large class of systems, all spurious links in the moral graph formed by observed nodes can be identified. Moreover, assuming conditions on hidden (latent) nodes, which follow from identifiability conditions, links between the hidden and observed nodes can be reconstructed resulting in the retrieval of the exact topology of the network.

This chapter also serves as an important bridge between the works presented in [40] and the works related to network structure reconstruction [17]-[28]. Furthermore, results here are applicable to many classes of directed graphs without self loops, not restricted to directed acyclic graphs or bi-directed graphs, unlike [18, 21].

We summarize below the major contributions of the chapter.

- Provides non-trivial generalizations of results of [40] to skew symmetric matrices, with exact characterization of tangent manifolds of skew symmetric matrices with a given rank and skew symmetric sparse matrices. We also provide a sufficient condition that enables a practical way to select penalty factor for the convex optimization formulation that yields the unique, sparse plus low-rank matrix decomposition, restricted to skew symmetric matrices. This contribution is applicable to general skew symmetric matrix independent of its application to topology identification.
- Reconstructs the moral graph of observed nodes and the Markov Blanket of latent nodes from the matrix decomposition of IPSDM; accounting for network identifiability issues associated with the latent nodes.

- Establishes that the decomposition of imaginary part of IPSDM is sufficient to recover moral graph of observed nodes and the Markov Blanket of the latent nodes. Conditions and methods for unique decomposition as a sum of low rank and sparse matrices of the imaginary part are provided.
- For large class of systems, the exact topology of the entire network is reconstructed by the decomposition of imaginary part of the IPSDM.
- For the more practical scenarios where we have access only to finite samples of time series at each node, we provide a concentration bound for estimation error of IPSDM.

Organization of the chapter: Section 2.2 presents the unique decomposition of a skew symmetric matrix into its sparse and low-rank components. Section 2.3 discusses linear dynamical systems and its graphical representation. Section 2.4 addresses moral graph reconstruction, identifiability of latent nodes, and Markov Blanket reconstruction of the latent nodes. Exact topology reconstruction is studied in Section 2.5. Section 2.6 discusses IPSDM estimation from finite time series. Section 2.7 provides simulation results and Section 2.8 concludes the chapter.

Notations: Bold capital letters denote matrices and bold small letters denote vectors. $\mathbb{S}^n$ denotes the set of all $n \times n$ skew symmetric matrices with real entries. For a matrix $\mathbf{M}$, $[\mathbf{M}]_{ij}$, $\mathbf{M}_{ij}$, or $\mathbf{M}(i, j)$ denote the (i, j) -th element of $\mathbf{M}$, $\|\mathbf{M}\|_1$ is defined as $\sum_{i,j} |\mathbf{M}_{ij}|$, $\|\mathbf{M}\|_\infty$ is defined as $\max_{i,j} |\mathbf{M}_{ij}|$, and $\|\mathbf{M}\|_0$ denotes the number of non-zero entries in $\mathbf{M}$. $\|\mathbf{M}\|_*$ denotes the nuclear norm, which is the sum of singular values of $\mathbf{M}$, and $\|\mathbf{M}\|_2$ denotes spectral-norm, which is defined as the largest singular value. $\text{support}(\mathbf{M})$ is defined as $\{(i, j) : \mathbf{M}_{ij} \neq 0\}$. $\Im\{\mathbf{M}\}$ denotes imaginary part of $\mathbf{M}$. For a vector, $\|\mathbf{x}\|_2$ denotes euclidean-norm, defined as $\sqrt{\sum_i x_i^2}$. $\sigma_i(\mathbf{M})$ denotes i^{th} largest singular value of $\mathbf{M}$. For time-series $(\tilde{x}_i(t))_{t \in \mathbb{Z}}$, $\mathbf{x}(z) = \mathcal{Z}[\tilde{\mathbf{x}}(k)]$ denotes bilateral z-transform of $\tilde{\mathbf{x}}$. For a set $\mathcal{S}$, $|\mathcal{S}|$ denotes cardinality of the set. We use $j = \sqrt{-1}$. For a transfer function $h(z)$, $z \in \mathbb{C}$, $h = 0$ means h is identically zero, i.e., $h(z) = 0$ for every $|z| = 1$. $h \neq 0$ means h is not identically zero. Almost always or almost surely is defined for a probability measure that is absolutely continuous with respect to the Lebesgue measure (e.g. any continuous distribution).

2.2 Sparse plus Low-Rank Matrix Decomposition for Skew Symmetric Matrices

In this section, we discuss the following problem: suppose we are given a real skew symmetric matrix $\mathbf{C} \in \mathbb{S}^n$ that is obtained by adding a sparse matrix $\tilde{\mathbf{S}} \in \mathbb{S}^n$ and a low-rank matrix $\tilde{\mathbf{L}} \in \mathbb{S}^n$; when can we decompose the matrix and retrieve the component matrices? The material presented in this section provides the needed preliminaries and extensions of results from [40], which do not incorporate constraints of skew symmetry. We remark that the results discussed in this section for the space of skew symmetric matrices can be extended to the space of complex matrices. However, we do not discuss it here.

2.2.1 Optimization for Sparse plus Low-rank Decomposition

Consider the following optimization problem.

$$\begin{aligned} (\hat{\mathbf{S}}_\gamma, \hat{\mathbf{L}}_\gamma) &= \arg \min_{\mathbf{S}, \mathbf{L}} \gamma \|\mathbf{S}\|_0 + \text{rank}(\mathbf{L}) \\ &\text{subject to } \mathbf{S} + \mathbf{L} = \mathbf{C}, \\ &\quad \mathbf{S} + \mathbf{S}^T = \mathbf{0}, \mathbf{L} + \mathbf{L}^T = \mathbf{0}, \end{aligned} \tag{2.1}$$

where γ is a fixed penalty, selected a priori. (2.1) is a combinatorial optimization problem and is NP-hard [59]. ℓ_1 norm is often employed as a surrogate for ℓ_0 norm [60], with nuclear norm being a proxy for rank [61]. Thus a more tractable convex relaxation associated with (2.1) is:

$$\begin{aligned} (\hat{\mathbf{S}}_\gamma, \hat{\mathbf{L}}_\gamma) &= \arg \min_{\mathbf{S}, \mathbf{L}} \gamma \|\mathbf{S}\|_1 + \|\mathbf{L}\|_* \\ &\text{subject to } \mathbf{S} + \mathbf{L} = \mathbf{C}, \\ &\quad \mathbf{S} + \mathbf{S}^T = \mathbf{0}, \mathbf{L} + \mathbf{L}^T = \mathbf{0}. \end{aligned} \tag{2.2}$$

Remark 2.2.1. Given $\mathbf{C} = -\mathbf{C}^T$, imposing the constraint $\mathbf{S} + \mathbf{S}^T = \mathbf{0}$ renders $\mathbf{L} + \mathbf{L}^T$ superfluous.

In this chapter, the convex formulation in (2.2) is applied for retrieving the sparse and low-rank components from the given $\mathbf{C}$.

2.2.2 Affine Varieties and Tangent Spaces

In the seminal work [40], the convex optimization problem (2.2) without the constraints $\mathbf{S}^T = -\mathbf{S}$ and $\mathbf{L}^T = -\mathbf{L}$ is considered, which provided sufficient conditions to retrieve $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{L}}$ exactly. The results in [40] established results for general square matrices with a real field. In this section, we extend the results to skew symmetric matrices in real field, $\mathbb{S}^n$. In order to address the decomposition, we consider the sparse matrix sets as an affine variety and low-rank matrix sets as a manifold. We characterize the necessary and sufficient conditions required for the unique decomposition in terms of the tangent space to the affine variety of support constrained skew symmetric matrices at $\tilde{\mathbf{S}}$ —the original sparse matrix—and the tangent space to the manifold of rank constrained skew symmetric matrices at $\tilde{\mathbf{L}}$ —the original low-rank matrix. Note that an affine variety is defined as the zero set of a system of polynomial equations [62].

Remark 2.2.2. *The skew symmetric matrices with real entries have the property that all the non-zero eigenvalues are pure imaginary and they exist in conjugate pairs. Therefore, the rank r of every skew symmetric matrix must be even, and the multiplicity of the singular values must be a positive multiple of two.*

Next, we provide definitions of tangent spaces, specific to skew symmetric matrices in real field, as our major focus in this chapter is on imaginary part of Hermitian matrices. The affine variety of skew symmetric matrices constrained by support size m is defined as:

$$\mathcal{S}(m) := \{\mathbf{M} \in \mathbb{S}^n : |\text{support}(\mathbf{M})| \leq m\}. \quad (2.3)$$

Notice that $\mathcal{S}(m)$ is defined over the space of all skew symmetric matrices, $\mathbb{S}^n$, i.e., the set of all matrices $\mathbf{M} \in \mathbb{R}^{n \times n}$ with $\mathbf{M}^T = -\mathbf{M}$. We establish the following result on tangent spaces of sparse real skew symmetric matrices.

Lemma 2.2.1. *For any skew symmetric matrix $\mathbf{M} \in \mathbb{S}^n$, the tangent space $\Omega(\mathbf{M})$ with respect to $\mathcal{S}(|\text{support}(\mathbf{M})|)$ at $\mathbf{M}$ is:*

$$\Omega(\mathbf{M}) := \{\mathbf{N} \in \mathbb{S}^n : \text{support}(\mathbf{N}) \subseteq \text{support}(\mathbf{M})\}. \quad (2.4)$$

Proof. See Appendix A.1. □

The dimension of this tangent space is $|\text{support}(\mathbf{M})|/2$ owing to the skew symmetric property.

We define set of skew symmetric matrices of rank r as:

$$\mathcal{R}(r) := \{\mathbf{M} \in \mathbb{S}^n : \text{rank}(\mathbf{M}) = r\}. \quad (2.5)$$

It is shown in [63] that $\mathcal{R}(r)$ is a differential manifold, whose dimension is $nr - \frac{r^2-r}{2}$.

Lemma 2.2.2. *For any skew symmetric matrix $\mathbf{M} \in \mathbb{S}^n$, the tangent space $T(\mathbf{M})$ with respect to $\mathcal{R}(\text{rank}(\mathbf{M}))$ at $\mathbf{M}$ is:*

$$T(\mathbf{M}) := \{\mathbf{U}\mathbf{X}^T - \mathbf{X}\mathbf{U}^T : \mathbf{X} \in \mathbb{R}^{n \times r}\}, \quad (2.6)$$

where $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ is the compact singular value decomposition (CSVD) of $\mathbf{M}$.

Proof. See Appendix A.2. □

The dimension of $T(\mathbf{M})$ is $nr - \frac{r^2}{2}$. The following lemma is obtained based on Remark 2.2.2.

Lemma 2.2.3. *Let $\mathbf{M} \in \mathbb{S}^n$ be a skew symmetric matrix and let $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be CSVD of $\mathbf{M}$, $\mathbf{D} := \text{diag}(\sigma_1, \dots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$, where r is the rank of $\mathbf{M}$. Then, the projection matrices $\mathbf{U}\mathbf{U}^T$ and $\mathbf{V}\mathbf{V}^T$ of the given skew symmetric matrix $\mathbf{M}$ are equal.*

Proof. See Appendix A.3. □

Suppose we have prior information about $\Omega(\tilde{\mathbf{S}})$ and $T(\tilde{\mathbf{L}})$, in addition to being given $\mathbf{C} = \tilde{\mathbf{S}} + \tilde{\mathbf{L}}$. Then, it can be shown that a necessary and sufficient condition for unique identifiability of $(\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ in terms of the tangent spaces is

$$\Omega(\tilde{\mathbf{S}}) \cap T(\tilde{\mathbf{L}}) = \{\mathbf{0}\}, \quad (2.7)$$

i.e., the tangent spaces $\Omega(\tilde{\mathbf{S}})$ and $T(\tilde{\mathbf{L}})$ has trivial intersection. In other words, if the tangent spaces intersect only at origin, then we can retrieve the component matrices $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{L}}$, if we have access to $\Omega(\tilde{\mathbf{S}})$ and $T(\tilde{\mathbf{L}})$. Given exact characterization of tangent spaces for real skew symmetric matrices in Lemma 2.2.1 and Lemma 2.2.2, it is possible to test for the necessary and sufficient trivial intersection condition for a given matrix $\mathbf{M}$. Now, analogous to development in [40], we obtain the sparse and low-rank decomposition using convex optimization.

2.2.3 Sparse Plus Low-Rank Decomposition using Optimization

In general, it is not possible to recover the original sparse and low-rank matrices by solving (2.2). To begin with, the solution of the optimization problem depends intricately on the penalty factor γ . In Proposition 2.2.3, we prove that for γ close to zero the optimal solution $(\hat{\mathbf{S}}_\gamma, \hat{\mathbf{L}}_\gamma)$ returned by (2.2) is $(\mathbf{C}, \mathbf{0})$, whereas for γ sufficiently large $(\hat{\mathbf{S}}_\gamma, \hat{\mathbf{L}}_\gamma) = (\mathbf{0}, \mathbf{C})$.

Another issue in decomposing the given matrix $\mathbf{C}$ is if either $\tilde{\mathbf{S}}$ is low-rank, or $\tilde{\mathbf{L}}$ is sparse. For example, suppose that the low-rank matrix $\tilde{\mathbf{L}}$ is such that $\tilde{\mathbf{L}}_{11} \neq 0$ with every other entry zero, and $\tilde{\mathbf{S}}$ be any sparse matrix with $\tilde{\mathbf{S}}_{11} \neq 0$. Then, the optimization may return $(\mathbf{C}, \mathbf{0})$ or $(\mathbf{0}, \mathbf{C})$ as the solution depends on the rank of $\tilde{\mathbf{S}}$. Another example where the unique decomposition is not possible is when $\tilde{\mathbf{S}}$ has support restricted to the first column and the first column of $\tilde{\mathbf{L}}$ negates all the entries of $\mathbf{S}$. Then, a reasonable solution is $(\mathbf{0}, \mathbf{C})$.

Next, we characterize the optimal regions of (2.2) and provide sufficient conditions under which it obtain the unique decomposition, i.e., returns the true sparse and low-rank matrices. The following proposition provides a sufficient condition for (2.2) to return the optimum solution $(\hat{\mathbf{S}}_\gamma, \hat{\mathbf{L}}_\gamma) = (\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$.

Proposition 2.2.1. *Suppose that $\mathbf{C} = \tilde{\mathbf{S}} + \tilde{\mathbf{L}}$, $\mathbf{C} \neq \mathbf{0}$ where $\tilde{\mathbf{S}}, \tilde{\mathbf{L}} \in \mathbb{S}^n$ is given. Then, $(\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ is the unique optimizer of (2.2) if the following conditions are satisfied:*

1. $\Omega(\tilde{\mathbf{S}}) \cap T(\tilde{\mathbf{L}}) = \{\mathbf{0}\}$.
2. There exist duals $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathbb{R}^{n \times n}$ such that
 - (a) $P_{\Omega(\tilde{\mathbf{S}})}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T) = \gamma \text{sign}(\tilde{\mathbf{S}})$,
 - (b) $P_{T(\tilde{\mathbf{L}})}(\mathbf{Q}_1) = \mathbf{U}\mathbf{V}^T$,
 - (c) $\|P_{\Omega(\tilde{\mathbf{S}})^c}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T)\|_\infty < \gamma$,
 - (d) $\|P_{T^\perp}(\mathbf{Q}_1)\|_2 < 1$,

where $P_{\Omega(\tilde{\mathbf{S}})}(\mathbf{M})$ is obtained by setting entries of $\mathbf{M}$ outside the support of $\tilde{\mathbf{S}}$ to zero and projecting it to the space of skew symmetric matrices, and $P_{T(\tilde{\mathbf{L}})}(\mathbf{M}) := P_{\mathbf{U}}\mathbf{M} + \mathbf{M}P_{\mathbf{U}} - P_{\mathbf{U}}\mathbf{M}P_{\mathbf{U}}$; $P_{\mathbf{U}} = \mathbf{U}\mathbf{U}^T$, $\tilde{\mathbf{L}} = \mathbf{U}\Sigma\mathbf{V}^T$.

Proof. See Appendix A.4. □

2.2.4 Sufficient Conditions to Retrieve the True Matrices

Here, we provide some sufficient conditions that guarantee the existence of the duals $\mathbf{Q}_1$ and $\mathbf{Q}_2$ discussed in Proposition 2.2.1. The definitions $\mu(\tilde{\mathbf{S}}) := \max_{\mathbf{N} \in \Omega(\tilde{\mathbf{S}}): \|\mathbf{N}\|_\infty \leq 1} \|\mathbf{N}\|_2$ and $\xi(\tilde{\mathbf{L}}) := \max_{\mathbf{N} \in T(\tilde{\mathbf{L}}): \|\mathbf{N}\|_2 \leq 1} \|\mathbf{N}\|_\infty$ are used to characterize the properties of the tangent spaces.

Remark 2.2.3. Our definitions of $\mu(\tilde{\mathbf{S}})$ and $\xi(\tilde{\mathbf{L}})$ are different from the respective definitions in [40]. In fact, the values of our μ and ξ are less than or equal to the respective values in [40].

The following proposition provides a sufficient condition to obtain the unique decomposition.

Proposition 2.2.2. Suppose that $\mathbf{C} = \tilde{\mathbf{S}} + \tilde{\mathbf{L}}$ is given. Suppose that $\mu(\tilde{\mathbf{S}})\xi(\tilde{\mathbf{L}}) < \frac{1}{6}$. Then, the unique optimum for (2.2) is $(\hat{\mathbf{S}}_\gamma, \hat{\mathbf{L}}_\gamma) = (\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ if

$$\gamma \in \left(\frac{\xi(\tilde{\mathbf{L}})}{1 - 4\mu(\tilde{\mathbf{S}})\xi(\tilde{\mathbf{L}})}, \frac{1 - 3\mu(\tilde{\mathbf{S}})\xi(\tilde{\mathbf{L}})}{\mu(\tilde{\mathbf{S}})} \right). \quad (2.8)$$

Proof. See Appendix A.5. □

Remark 2.2.4. The range of values in (2.8) is a superset of the range specified in [40]. In the worst case, (2.8) would return the same interval specified in [40], due to difference in definitions of the tangent spaces, $\Omega(\tilde{\mathbf{S}})$ and $T(\tilde{\mathbf{L}})$.

Next, we define $\deg_{\max}$ of a matrix $\mathbf{M}$ as

$$\deg_{\max}(\mathbf{M}) := \max \left(\max_{1 \leq i \leq n} \left(\sum_{j=1}^n \mathbb{1}_{\{\mathbf{M}_{ij} \neq 0\}} \right), \max_{1 \leq j \leq n} \left(\sum_{i=1}^n \mathbb{1}_{\{\mathbf{M}_{ij} \neq 0\}} \right) \right),$$

where $\mathbb{1}_{\{x \neq 0\}} := 1$ if $x \neq 0$ and $\mathbb{1}_{\{x \neq 0\}} := 0$ if $x = 0$ denotes the indicator function. We define the *maximum incoherence* of the row/column space of the real skew symmetric matrix $\mathbf{M}$ as

$$\text{inc}(\mathbf{M}) := \max_k \|\mathbf{U}\mathbf{U}^T e_k\|_2,$$

where $\mathbf{U}\Sigma\mathbf{V}^T$ is the CSVD of $\mathbf{M}$. This definition is different from the one in [40] due to Lemma 2.2.3.

The following lemma extends the sufficient condition in Proposition 2.2.2 in terms of $\deg_{\max}$ and maximum incoherence.

Lemma 2.2.4. *Let $\mathbf{C} = \tilde{\mathbf{S}} + \tilde{\mathbf{L}}$ with $\deg_{\max}(\tilde{\mathbf{S}})$ and $\text{inc}(\tilde{\mathbf{L}})$ as defined above. If $\deg_{\max}(\tilde{\mathbf{S}})\text{inc}(\tilde{\mathbf{L}}) < \frac{1}{12}$, then the unique optimum of the convex program (2.2) is $(\hat{\mathbf{S}}_\gamma, \hat{\mathbf{L}}_\gamma) = (\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ for a range of values of γ given by:*

$$\gamma \in \left(\frac{2\text{inc}(\tilde{\mathbf{L}})}{1 - 8\deg_{\max}(\tilde{\mathbf{S}})\text{inc}(\tilde{\mathbf{L}})}, \frac{1 - 6\deg_{\max}(\tilde{\mathbf{S}})\text{inc}(\tilde{\mathbf{L}})}{\deg_{\max}(\tilde{\mathbf{S}})} \right). \quad (2.9)$$

Proof. The proof is similar to Corollary 3 in [40], and is skipped due to space constraint. $\square$

Thus, by picking a proper γ the convex optimization (2.2) returns the unique decomposition $(\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ without the need to determine $\Omega(\tilde{\mathbf{S}})$ and $T(\tilde{\mathbf{L}})$.

Remark 2.2.5. *Lemma 2.2.4 provides a conservative sufficient condition and hence covers only a subclass of uniquely decomposable matrices. In Section 2.7 we provide an example that does not satisfy the sufficient conditions, but still is uniquely decomposable using Algorithm 1, i.e., the network satisfies the transverse intersection (2.7), but not the sufficient condition in Lemma 2.2.4.*

It can be observed from the structure of singular vectors in the projection matrix $\mathbf{U}\mathbf{U}^T$ that n must be large for $\text{inc}(\tilde{\mathbf{L}})$ to be small enough to satisfy the condition in Lemma 2.2.4.

Moreover, using the results from [64], it can be shown that the number of non-zero entries in $\tilde{\mathbf{S}}$ must be at most $O(n)$ for the sufficient conditions to hold.

The convex program (2.2) is equivalent to the following formulation with the mapping $t = \frac{\gamma}{1+\gamma}$, where $t \in [0, 1]$:

$$\begin{aligned} (\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t) &= \arg \min_{\mathbf{S}, \mathbf{L}} t\|\mathbf{S}\|_1 + (1-t)\|\mathbf{L}\|_* \\ &\text{subject to } \mathbf{S} + \mathbf{L} = \mathbf{C}, \\ &\quad \mathbf{S}^T = -\mathbf{S}, \mathbf{L}^T = -\mathbf{L}. \end{aligned} \quad (2.10)$$

The following definitions are used to measure the closeness of the estimated matrices with the true matrices.

$$\text{tol}_t := \frac{\|\hat{\mathbf{S}}_t - \tilde{\mathbf{S}}\|_F}{\|\tilde{\mathbf{S}}\|_F} + \frac{\|\hat{\mathbf{L}}_t - \tilde{\mathbf{L}}\|_F}{\|\tilde{\mathbf{L}}\|_F}, \quad (2.11)$$

$$\text{diff}_t := (\|\hat{\mathbf{S}}_{t-\epsilon} - \hat{\mathbf{S}}_t\|_F) + (\|\hat{\mathbf{L}}_{t-\epsilon} - \hat{\mathbf{L}}_t\|_F), \quad (2.12)$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\epsilon > 0$ is a sufficiently small fixed constant. Note that tol_t requires the knowledge of the true matrices $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{L}}$, whereas $diff_t$ does not require any such prior information. Moreover, the mapping between t and γ is one-to-one.

In practice, we may not have access to any extra information other than $\mathbf{C}$; thus determining t required for the unique decomposition from Lemma 2.2.4 and tol_t becomes difficult. Here we provide guidance on which t (and this γ) to be employed. The following proposition provides a systematic approach to identify a proper penalty factor t for the unique decomposition.

Proposition 2.2.3. *Suppose we are given a matrix $\mathbf{C}$, which is obtained by summing $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{L}}$, where $\tilde{\mathbf{S}}$ is a sparse matrix and $\tilde{\mathbf{L}}$ is a low-rank matrix. If $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{L}}$ satisfies $\deg_{\max}(\tilde{\mathbf{S}})inc(\tilde{\mathbf{L}}) < 1/12$, then there exist at least three regions where $diff_t = 0$. In particular, there exists an interval $[t_1, t_2] \subset [0, 1]$ with $0 < t_1 < t_2 < 1$ such that $(\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t) = (\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ for any $t \in [t_1, t_2]$*

Proof. See Appendix A.6. □

Corollary 2.2.1. *By solving (2.10) and calculating $diff_t$ for every $t \in \{\epsilon, 2\epsilon, \dots, 1\}$ we obtain $\hat{\mathbf{S}}_t$ and $\hat{\mathbf{L}}_t$ and the zero regions specified in Proposition 2.2.3, specifically $[t_1, t_2]$, $0 < t_1 < t_2 < 1$, where $diff_t = 0$. For a given $t \in [t_1, t_2]$, if $\deg_{\max}(\hat{\mathbf{S}}_t)inc(\hat{\mathbf{L}}_t) < 1/12$, then the decomposition is exact, that is, $tol_t = 0$ and $(\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t) = (\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$.*

Remark 2.2.6. *Proposition 2.2.3 and Corollary 2.2.1 are applicable in any general sparse plus low-rank matrix decomposition, and is not restricted to skew symmetric matrix decomposition.*

Remark 2.2.7. *Conversely, if there are only two zero regions, t close to zero and t close to 1, then we can assert that it may not be possible to obtain a unique decomposition with this approach. Simulations show that $t \in [.26, .4]$ is a good region to look for t .*

Based on Proposition 2.2.3 and Corollary 2.2.1, we propose Algorithm 1 to obtain the unique decomposition, which returns the estimated sparse matrix $\hat{\mathbf{S}}$ and estimated low-rank matrix $\hat{\mathbf{L}}$.

Remark 2.2.8. *The aforementioned results can be extended to the space of complex Hermitian matrices also; however, we do not discuss them here.*

In the following section, we discuss some preliminaries of linear dynamical systems that are useful in understanding the rest of the chapter. In Section 2.4 and Section 2.5, we discuss how the matrix decomposition is extremely useful to reconstruct the moral graph/topology of a given linear dynamical model.

2.3 Linear Dynamical Systems

Consider a linear dynamical system with n interacting agents, each equipped with time-series measurements $(\tilde{x}_i(t))_{t \in \mathbb{Z}}$, $i \in \{1, \dots, n\}$, governed by the following linear dynamical model (LDM):

$$\tilde{\mathbf{x}}(k) = \sum_{l=-\infty}^{\infty} \tilde{\mathbf{H}}(l) \tilde{\mathbf{x}}(k-l) + \tilde{\mathbf{e}}(k), \quad (2.13)$$

where $\tilde{\mathbf{x}}(k) = [\tilde{x}_1(k), \dots, \tilde{x}_n(k)]^T$, $\tilde{\mathbf{e}}(k) = [\tilde{e}_1(k), \dots, \tilde{e}_n(k)]^T$, and for $i, j \in \{1, 2, \dots, n\}$, $i \neq j$, $\tilde{e}_i(k)$ is a zero mean wide sense stationary (WSS) process uncorrelated with $\tilde{e}_j(k)$. Additionally, the processes $\{\tilde{x}_i(k), \tilde{e}_i(k)\}_{i=1}^n$ are jointly WSS. Let $\mathbf{H}(z) = \mathcal{Z}[\tilde{\mathbf{H}}(k)]$. Then, $\tilde{\mathbf{H}}(l) \in \mathbb{R}^{n \times n}$ denotes the weighted adjacency matrix with diagonal entries $\tilde{\mathbf{H}}_{ii}(l) = 0$, $1 \leq i \leq n$, $l \in \mathbb{Z}$, such that $\mathbf{H}$ is well posed, i.e., every entry of $(\mathbf{I} - \mathbf{H}(z))^{-1}$ is analytic on the unit circle, $|z| = 1$, $z \in \mathbb{C}$. An LDM is said to be topologically detectable if $\Phi_{\mathbf{e}}(z)$ is positive definite for every $|z| = 1$. The above model can be represented using the following Transfer Function Model (TFM),

$$\mathbf{x}(z) = \mathbf{H}(z) \mathbf{x}(z) + \mathbf{e}(z), \quad z \in \mathbb{C}, \quad (2.14)$$

where $\mathbf{x}(z) = \mathcal{Z}[\tilde{\mathbf{x}}(k)]$ and $\mathbf{e}(z) = \mathcal{Z}[\tilde{\mathbf{e}}(k)]$. In general, there may exist nodes whose observations are not available and remain hidden. These nodes that are not accessible are called hidden/latent/unobservable nodes. $\mathcal{V}_o$ denotes the set of observable nodes with cardinality n_o and $\mathcal{V}_h$ is the set of latent nodes with cardinality n_h .

2.3.1 Graphical Representation

The Linear Dynamic Graph (LDG) associated with the LDM (2.14) is defined as the directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \dots, n\}$ and $\mathcal{E} = \{(i, j) | \mathbf{H}_{ji} \neq 0\}$. Thus, there exists a directed edge (i, j) from node i to node j in the LDG if and only if $\mathbf{H}_{ji} \neq 0$. For a directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, parent set of node j is $\mathcal{P}(j) := \{i | (i, j) \in \mathcal{E}\}$, child set of node

j is $\mathcal{C}(j) := \{i | (j, i) \in \mathcal{E}\}$ and spouse set of node j is $\mathcal{S}(j) := \{i | i \in \mathcal{P}(\mathcal{C}(j))\}$. Nodes i and j are strict spouses if $i \in \mathcal{S}(j)$, $i \notin \mathcal{C}(j) \cup \mathcal{P}(j)$. The Markov Blanket of node i , denoted $\text{kin}(i) := \mathcal{C}(i) \cup \mathcal{P}(i) \cup \mathcal{S}(i)$. The moral or the kin graph, $\text{kin}(\mathcal{G}) := \{(i, j) | i \in \text{kin}(j), i, j \in \mathcal{V}\}$, where (i, j) denotes an unordered pair. The topology of $\mathcal{G}$ is defined as an undirected graph $\text{top}(\mathcal{G}) := \{(i, j) | i \in \mathcal{P}(j) \cup \mathcal{C}(j), i, j \in \mathcal{V}\}$.

Similarly, an LDG obtained by restricting the vertex set to the observed nodes is defined by $\mathcal{G}_o(\mathcal{V}_o, \mathcal{E}_o)$, where $\mathcal{E}_o := \{(i, j) | i, j \in \mathcal{V}_o \text{ and } \mathbf{H}_{ji} \neq 0\}$. The topology among the observable nodes, $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_o) := \text{top}(\mathcal{G}_o(\mathcal{V}_o, \mathcal{E}_o))$. We define an undirected edge set $\bar{\mathcal{E}}_o := \{(i, j) | (i, j) \in \mathcal{E}_o \text{ or } (j, i) \in \mathcal{E}_o, i < j\}$. Similarly, the moral graph among the observable nodes is the undirected graph $\text{kin}(\mathcal{G}_o) := \{(i, j) | i \in \text{kin}(j), i, j \in \mathcal{V}_o\}$. We define a path between nodes i and j in an undirected graph as a set of nodes $\{i, x_0, \dots, x_k, j\}$ where $\{(i, x_0), (x_0, x_1), \dots, (x_{k-1}, x_k)\} \subseteq \mathcal{E}$. We define a path between nodes i and j in an undirected graph as a set of nodes $\{i, x_0, \dots, x_k, j\}$ where $\{(i, x_0), (x_0, x_1), \dots, (x_{k-1}, x_k)\} \subseteq \mathcal{E}$. A directed path in a directed graph is a path between nodes i and j with the constraint that all the edges are directed from i towards j . $d_{\text{hop}}(i, j)$ is defined for undirected graphs as the number of links between nodes i and j on the shortest path connecting i and j . It can be shown that $d_{\text{hop}}(i, j)$ is a metric for undirected graphs. For a node $i \in \mathcal{V}$, the degree of the node is $\text{deg}(i) := |\{j \in \mathcal{V} : (i, j) \in \mathcal{E}\}|$. Note that this definition is for undirected edges.

In the next section, we study properties of the IPSDM of a given LDM that are useful in topology/moral graph reconstruction.

2.4 Exact Reconstruction of Moral Graph of Observed Nodes and Markov Blanket of Latent Nodes

2.4.1 Moral Graph Reconstruction under Complete Observability

In this part of the section we present some important preliminaries for reconstruction of moral graphs from power spectral density matrices and methods on how to reconstruct moral graph under full observability of the network.

For the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, the power spectral density matrix (PSDM), $\Phi_{\mathbf{x}}(z) \in \mathbb{C}^{n \times n}$ is given by

$$\Phi_{\mathbf{x}}(z) := \sum_{k=-\infty}^{\infty} \mathbb{E}\{\tilde{\mathbf{x}}(k)\tilde{\mathbf{x}}^T(0)\}z^{-k}, \quad z \in \mathbb{C}, |z| = 1. \quad (2.15)$$

The PSDM and the IPSDM of the dynamical system governed by (2.14) can be respectively written as (see [17]) $\Phi_{\mathbf{x}}(z) = (\mathbf{I} - \mathbf{H}(z))^{-1} \Phi_{\mathbf{e}}(z) (\mathbf{I} - \mathbf{H}^*(z))^{-1}$ and

$$\Phi_{\mathbf{x}}^{-1}(z) = (\mathbf{I} - \mathbf{H}^*(z)) \Phi_{\mathbf{e}}^{-1}(z) (\mathbf{I} - \mathbf{H}(z)). \quad (2.16)$$

Note that $\Phi_{\mathbf{e}}^{-1}(z)$ is a diagonal matrix, since $\tilde{e}_i(k)$ is uncorrelated with $\tilde{e}_\ell(k)$ for $i \neq \ell$.

The following lemma (Theorem 27 in [17]) provides a sufficient condition to estimate the moral graph of $\mathcal{G}(\mathcal{V}, \mathcal{E})$ from the IPSDM.

Lemma 2.4.1. *Consider a well posed and topologically detectable LDM $(\mathbf{H}, \mathbf{e})$ with the associated LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$, described by (2.14), having full node observability. Let IPSDM of $\mathbf{x}$ be given by (2.16). Then, $[\Phi_{\mathbf{x}}^{-1}]_{i,j} \neq 0$, $i \neq j$ implies $i \in \text{kin}(j)$. Moreover, the converse holds almost always.*

Remark 2.4.1. *In [32] and the related works, $[\Phi_{\mathbf{x}}^{-1}]_{i,j} \neq 0$ is considered to be equivalent to nodes i and j being conditionally dependent given the rest of the observations. That is, retrieving conditional dependence is equivalent to reconstructing moral graph.*

Remark 2.4.2. *The results of [32] can be extended to AR models with WSS noise by considering conditional correlation instead of conditional dependence. In this case, retrieving conditional correlation is equivalent to moral graph reconstruction.*

2.4.2 Structure of the IPSDM with Latent Nodes

In the previous subsection we studied the properties of IPSDM under the assumption that all the nodes are observable. However, topology identification becomes complicated in the presence of latent nodes, often leading to lack of identifiability (see 2.4.3). Here, we discuss some of the special properties of the IPSDM in the presence of latent nodes that are exploited in this chapter.

By separating observable nodes and latent nodes, we represent $\mathbf{x}(z) = \begin{bmatrix} \mathbf{x}_o(z) \\ \mathbf{x}_h(z) \end{bmatrix}$ and $\mathbf{e}(z) = \begin{bmatrix} \mathbf{e}_o(z) \\ \mathbf{e}_h(z) \end{bmatrix}$, where $\mathbf{x}_o(z) = \mathcal{Z}[\tilde{\mathbf{x}}_o(k)]$, $\mathbf{x}_h(z) = \mathcal{Z}[\tilde{\mathbf{x}}_h(k)]$, $\mathbf{e}_o(z) = \mathcal{Z}[\tilde{\mathbf{e}}_o(k)]$, and $\mathbf{e}_h(z) = \mathcal{Z}[\tilde{\mathbf{e}}_h(k)]$. The TFM in (2.14) can be expressed as follows:

$$\begin{bmatrix} \mathbf{x}_o(z) \\ \mathbf{x}_h(z) \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{oo}(z) & \mathbf{H}_{oh}(z) \\ \mathbf{H}_{ho}(z) & \mathbf{H}_{hh}(z) \end{bmatrix} \begin{bmatrix} \mathbf{x}_o(z) \\ \mathbf{x}_h(z) \end{bmatrix} + \begin{bmatrix} \mathbf{e}_o(z) \\ \mathbf{e}_h(z) \end{bmatrix}. \quad (2.17)$$

Letting

$$\Phi_{\mathbf{x}}(z) = \begin{bmatrix} \Phi_{oo}(z) & \Phi_{oh}(z) \\ \Phi_{ho}(z) & \Phi_{hh}(z) \end{bmatrix}$$

and

$$\Phi_{\mathbf{x}}^{-1}(z) = \begin{bmatrix} \mathbf{K}_{oo}(z) & \mathbf{K}_{oh}(z) \\ \mathbf{K}_{ho}(z) & \mathbf{K}_{hh}(z) \end{bmatrix}$$

we have that (by ignoring the index z)

$$\Phi_{oo}^{-1} = \mathbf{K}_{oo} - \mathbf{K}_{oh}\mathbf{K}_{hh}^{-1}\mathbf{K}_{ho},$$

which follows by applying block matrix inversion formula and using Schur's complement representation [65]. Furthermore, using (2.16), the IPSDM corresponding to the observed variables can be written as:

$$\Phi_{oo}^{-1} = \mathbf{S} + \mathbf{L}, \text{ where} \quad (2.18)$$

$$\mathbf{S} = (\mathbf{I}_o - \mathbf{H}_{oo}^*)\Phi_{e_o}^{-1}(\mathbf{I}_o - \mathbf{H}_{oo}), \quad (2.19)$$

$$\mathbf{L} = \mathbf{H}_{ho}^*\Phi_{e_h}^{-1}\mathbf{H}_{ho} - \Psi^*\Lambda^{-1}\Psi, \quad (2.20)$$

$$\Psi = \mathbf{H}_{oh}^*\Phi_{e_o}^{-1}(\mathbf{I} - \mathbf{H}_{oo}) + (\mathbf{I} - \mathbf{H}_{hh}^*)\Phi_{e_h}^{-1}\mathbf{H}_{ho}, \text{ and}$$

$$\Lambda = \mathbf{H}_{oh}^*\Phi_{e_o}^{-1}\mathbf{H}_{oh} + (\mathbf{I} - \mathbf{H}_{hh}^*)\Phi_{e_h}^{-1}(\mathbf{I} - \mathbf{H}_{hh}).$$

The following proposition shows that $\text{support}(\mathbf{S})$ can retrieve the moral graph among the observable nodes, $\text{kin}(\mathcal{G}_o)$.

Proposition 2.4.1. *Consider a well-posed and topologically detectable LDM, $(\mathbf{H}, \mathbf{e})$, described by (2.14), with the associated graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Let $\mathbf{S}$ be given by (2.19). Then, the set $\hat{\mathcal{E}}_o := \{(i, j) | \mathbf{S}_{ij} \neq 0, i < j\}$ reconstructs $\text{kin}(\mathcal{G}_o)$ almost always.*

Proof. Notice that (2.16) and (2.19) are exactly the same, except $\mathbf{H}$, $\Phi_{\mathbf{e}}$, and $\mathbf{I}$ in (2.16) are replaced with $\mathbf{H}_{oo}$, Φ_{e_o} , and $\mathbf{I}_o$ respectively in (2.19). Thus, by applying Lemma 2.4.1 on $\mathbf{S}$, $\hat{\mathcal{E}}_o$ retrieves $\text{kin}(\mathcal{G}_o)$ almost always. $\square$

The following theorem shows that $\mathbf{S}$ is sparse if the moral graph, $\text{kin}(\mathcal{G}_o)$ is sparse, while $\mathbf{L}$ is a low-rank matrix if $n_h < n_o$. This particular structure aids in decomposing Φ_{oo}^{-1} into $\mathbf{S}$ and $\mathbf{L}$ (or more precisely $\Im\{\Phi_{oo}^{-1}\}$ into $\Im\{\mathbf{S}\}$ and $\Im\{\mathbf{L}\}$ using the results from Section 2.2). Notice that the index z is omitted from the notations. The results hold for every $|z| = 1$ uniformly.

Theorem 2.4.1. *Consider the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$ described by (2.14). Let $\mathcal{G}_o(\mathcal{V}_o, \mathcal{E}_o)$ be the LDG, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, restricted to the observed nodes. The following holds:*

$$|\text{support}(\mathbf{S})| \leq 2|\mathcal{E}_o| + 2|\mathcal{E}_o^{ss}| + n, \text{ rank}(\mathbf{L}) \leq 2n_h, \quad (2.21)$$

where $\mathcal{E}_o^{ss}$ denotes the set of undirected edges between the observable strict spouses with a common observable child.

Proof. See Appendix A.7. □

Remark 2.4.3. *Here, we are interested in the scenarios where $2|\mathcal{E}_o| + 2|\mathcal{E}_o^{ss}| = O(n)$; note that the maximum number of interconnections in a graph of n nodes is n^2 . In this sense, $\mathbf{S}$ is considered sparse. If we can uniquely decompose $\Phi_{oo}^{-1}(z)$ into the sparse matrix $\mathbf{S}$ and the low-rank matrix $\mathbf{L}$, then one can obtain $\text{kin}(\mathcal{G}_o)$ from $\mathbf{S}$ (see Proposition 2.4.1) and the Markov Blanket of the hidden nodes from $\mathbf{L}$ (see Section 2.4.4). Further, in Section 2.5.2, we reconstruct the exact topology of the entire network, including that of hidden nodes, under some assumptions. However, there are certain identifiability issues related to hidden nodes, inherent in network topology, which make the detection of hidden nodes from $\mathbf{L}$ difficult, even impossible in some cases. We address them next.*

2.4.3 Identifiability of Latent Nodes

Here, we discuss identifiability of the hidden nodes inherent to the graph structure and not limited to any specific reconstruction method. We illustrate the non-identifiability via examples.

Figure 2.1a and Figure 2.1d represent LDGs with LDMs given by $\tilde{x}_1 = \tilde{e}_1$, $\tilde{x}_2 = \tilde{x}_1 + \sum_{i=4}^n h_{2i} * \tilde{x}_i(k) + \tilde{e}_2$, $\tilde{x}_3 = h_{32} * \tilde{x}_2(k) + \tilde{e}_3$, where node 1 is latent, and $\tilde{x}_2 = \sum_{i=4}^n h_{2i} * \tilde{x}_i(k) + \hat{e}_2$, $\tilde{x}_3 = h_{32} * \tilde{x}_2(k) + \tilde{e}_3$ respectively. Let $\hat{e}_2 = \tilde{e}_2 + \tilde{e}_1$. Then the observed node time-series obtained from the two LDMs are identical and the two models are indistinguishable from Φ_{oo}^{-1} .

Similarly, consider the LDGs shown in Figure 2.1b and Figure 2.1e with LDMs given by $\tilde{x}_1 = \tilde{e}_1$, $\tilde{x}_2 = \tilde{x}_1 + \sum_{i=4}^n h_{2i} * \tilde{x}_i(k) + \tilde{e}_2$, $\tilde{x}_3 = h_{32} * \tilde{x}_2(k) + \tilde{e}_3$, $\tilde{x}_1 = \tilde{e}_1$, $\tilde{x}_2 = \tilde{x}_1 + \sum_{i=4}^n h_{2i} * \tilde{x}_i(k) + \tilde{e}_2$ respectively, where node 3 is latent. Again, the time-series among the observable nodes obtained from both the LDGs are the same and hence the two models are indistinguishable.

Based on the aforementioned discussion, we make the following assumptions for identifiability of a hidden node.

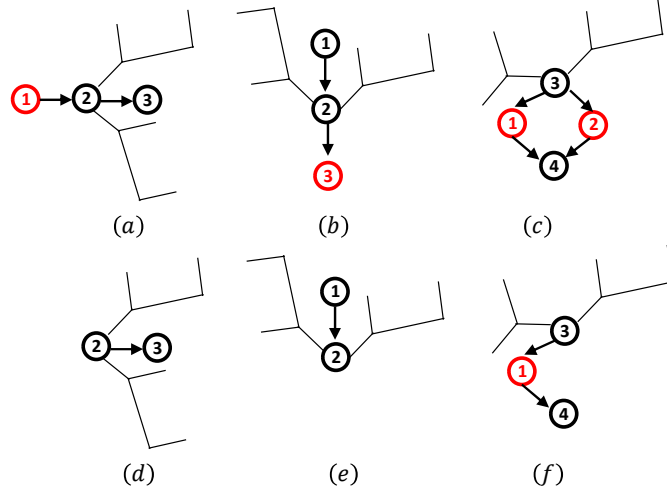


Figure 2.1: Non-identifiability of hidden nodes (red colored nodes): (a) hidden node is terminal and strict parent; (b) hidden node is terminal and strict child; (c) Markov Blankets of hidden nodes 1 and 2 overlapped.

Assumption 2.4.1. *Any hidden node k_h in $\mathcal{G}(\mathcal{V}, \mathcal{E})$ has at least one observed-child $c \in \mathcal{V}_o$. Further, k_h is a parent or child of another observable node $j \in \mathcal{V}_o \setminus \{c\}$.*

We next illustrate that non-identifiability issues arise when the Markov Blankets of hidden nodes overlap. Consider the LDGs shown in Figure 2.1c and Figure 2.1f with LDMs given by $\tilde{x}_3 = \sum_{i=5}^n h_{3i} * \tilde{x}_i(k) + \tilde{e}_3$, $\tilde{x}_1 = \tilde{x}_3 + \tilde{e}_1$, $\tilde{x}_2 = \tilde{x}_3 + \tilde{e}_2$, $\tilde{x}_4 = \tilde{x}_1 + \tilde{x}_2 + \tilde{e}_4$, and $\tilde{x}_3 = \sum_{i=5}^n h_{3i} * \tilde{x}_i(k) + \tilde{e}_3$, $\tilde{x}_1 = \tilde{x}_3 + \hat{e}_1$, $\tilde{x}_4 = \tilde{x}_1 + \tilde{e}_4$, where $\hat{e}_1 = \tilde{e}_1 + \tilde{e}_2$, respectively, where nodes 1 and 2 are latent. Both LDMs result in the same observed time-series, which leads to non-identifiability of hidden node 2. Hence, we make the following assumption about the spatial distribution of hidden nodes in the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$.

Assumption 2.4.2. *For every distinct $k_h, k'_h \in \mathcal{V}_h$, $d_{hop}(k_h, k'_h) > 4$.*

Assumption 2.4.2 is sufficient to ensure that Markov Blankets of any two distinct hidden nodes k_h and k'_h do not overlap. When the intersection of the Markov Blankets of k_h and k'_h contain more than one observable node, then one of the hidden nodes k_h or k'_h is non-identifiable (see the illustration associated with Figure 2.1c). If the intersection has at most one node, then both the hidden nodes might be identifiable. We make a slightly more conservative assumption that the Markov Blankets of the hidden nodes are non-overlapping. Moreover, the implication of Assumption 2.4.2 is that the Markov Blanket of a hidden node $h \in \mathcal{V}_h$ is $\{\mathcal{P}(h) \cup \mathcal{C}(h) \cup \mathcal{S}(h)\} \subseteq \mathcal{V}_o$.

Remark 2.4.4. Note that Assumption 2.4.2 might seem stronger than Assumption 2 in [21]. However, [21] restricted attention to radial topologies associated with bi-directed LDGs and assumed that the hidden nodes are at least three hops away from the leaf nodes. On the contrary, algorithms in Section 2.5 can reconstruct more general linear dynamical networks, including loopy networks.

2.4.4 Markov Blanket Reconstruction of the Latent nodes

Here, we provide the following definition and theorems which helps in learning the Markov Blanket of hidden nodes.

Definition 2.4.1. For $i, j \in \mathcal{V}_o$ and hidden node $h \in \mathcal{V}_h$, define $DE_h(i, j) := \{g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8, g_9\}$, $DE_h^{ss}(i, j) := \{g_{10}, g_{11}\}$, and $DM_h(i, j) := DE_h(i, j) \cup DE_h^{ss}(i, j)$, where $i \rightarrow h$ is denoted as (i, h) with $g_1 = \{(h, i), (h, j)\}$, $g_2 = \{(h, i), (j, h)\}$, $g_3 = \{(h, i), (h, k), (j, k)\}$, $\forall k \in \mathcal{V}_o \setminus \{i, j\}$, $g_4 = \{(i, h), (h, j)\}$, $g_5 = \{(i, h), (h, k), (j, k)\}$, $\forall k \in \mathcal{V}_o \setminus \{i, j\}$, $g_6 = \{(i, k), (h, k), (j, h)\}$, $\forall k \in \mathcal{V}_o \setminus \{i, j\}$, $g_7 = \{(i, k), (h, k), (h, j)\}$, $\forall k \in \mathcal{V}_o \setminus \{i, j\}$, $g_8 = \{(h, i), (j, i)\}$, $g_9 = \{(h, j), (i, j)\}$, $g_{10} = \{(i, h), (j, h)\}$, and $g_{11} = \{(i, k_1), (h, k_1), (h, k_2), (k_2, j)\}$, $\forall k_1, k_2 \in \mathcal{V}_o \setminus \{i, j\}$.

$DM_h(i, j)$ enumerates all the possible paths between two observable nodes i, j present in the Markov Blanket of hidden node h . We now present a result which infers the Markov Blanket of a hidden node h in $\mathcal{G}(\mathcal{V}, \mathcal{E})$ from $\text{support}(\mathbf{L})$.

Theorem 2.4.2. Suppose the LDM in (2.17) satisfies assumptions 2.4.1, 2.4.2. Let $\mathbf{L}$ be given by (2.20) and let $i, j \in \mathcal{V}_o$, $i \neq j$. Then, the following statements hold:

- (a) If $\mathbf{L}_{ij} \neq 0$, then there exists $g \in DM_h(i, j)$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$, for some $h \in \mathcal{V}_h$. Further, $d_{hop}(i, h) \leq 2$ and $d_{hop}(j, h) \leq 2$.
- (b) Given $\mathbf{L}_{ij} \neq 0$, suppose there exist $g_1 \in DM_{h_1}(i, j)$ and $g_2 \in DM_{h_2}(i, j)$ connecting i and j such that $g_1, g_2 \in \mathcal{G}(\mathcal{V}, \mathcal{E})$, for some $h_1, h_2 \in \mathcal{V}_h$. Then $h_1 = h_2$.

Proof. See Appendix A.8. □

Remark 2.4.5. We note that, for a set of system parameters, noise statistics can be construed such that a $g \in DM_h(i, j)$ is present in the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$, with $\mathbf{L}_{ij} = 0$. We remark that such cases are pathological; we will assume that the converse of Theorem 2.4.2(a) holds almost everywhere.

Thus, based on the locations of non-zero entries in $\mathbf{L}$, we construct $V_H = \{i | i \in \mathcal{V}_o, \exists k \in \mathcal{V}_o \setminus \{i\} \text{ s.t. } \mathbf{L}_{ik} \neq 0\}$ and $E_H = \{(i, j) | i, j \in V_H, \mathbf{L}_{ij} \neq 0\}$. The following result shows that (V_H, E_H) is a disjoint collection of connected undirected subgraphs (see Figure 2.4b for example). Moreover, the number of connected undirected subgraphs in E_H is equal to number of hidden nodes n_h in the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$.

Theorem 2.4.3. *Suppose the LDM in (2.17) satisfies assumptions 2.4.1, 2.4.2. Let $\mathbf{L}$ be given by (2.20). From $\mathbf{L}$, construct the following: $V_H = \{i | i \in \mathcal{V}_o, \exists k \in \mathcal{V}_o \setminus \{i\} \text{ s.t. } \mathbf{L}_{ik} \neq 0\}$, $E_H = \{(i, j) | i, j \in V_H, \mathbf{L}_{ij} \neq 0\}$. For every hidden node $l \in \mathcal{V}_h$, let $M_l = \mathcal{P}(l) \cup \mathcal{C}(l) \cup \mathcal{S}(l)$, $Q_l = \{(i, j) | i, j \in M_l, \mathbf{L}_{ij} \neq 0\}$. Then*

$$(a) \quad M_{l_1} \cap M_{l_2} = \emptyset, \text{ for all } l_1, l_2 \in \mathcal{V}_h, l_1 \neq l_2.$$

$$(b) \quad Q_{l_1} \cap Q_{l_2} = \emptyset, \text{ for all } l_1, l_2 \in \mathcal{V}_h, l_1 \neq l_2.$$

$$(c) \quad V_H = \bigcup_{l=1}^{n_h} M_l.$$

$$(d) \quad E_H = \bigcup_{l=1}^{n_h} Q_l.$$

Proof. See Appendix A.9. □

Remark 2.4.6. *The above theorem estimates the number of hidden nodes n_h as the number of connected undirected subgraphs in E_H . Each connected component (M_l, Q_l) is due to a hidden node $l \in \mathcal{V}_h$, that is, M_l is the Markov Blanket of l in $\mathcal{G}(\mathcal{V}, \mathcal{E})$, while Q_l is an undirected edge set that contain edges among any two distinct nodes in M_l .*

2.4.5 Moral Graph Reconstruction of Observable Nodes and Localization of Hidden Nodes from IPSDM

As shown in Theorem 2.4.1, $\mathbf{S}$ is sparse and $\mathbf{L}$ is low-rank. Then, one can retrieve $\mathbf{S}$ and $\mathbf{L}$ from Φ_{oo}^{-1} by employing the following optimization.

$$\begin{aligned} (\hat{\mathbf{S}}_\gamma, \hat{\mathbf{L}}_\gamma) &= \arg \min_{\mathbf{S}, \mathbf{L} \in \mathbb{C}^{n \times n}} \gamma \|\mathbf{S}\|_1 + \|\mathbf{L}\|_* \\ &\text{subject to } \mathbf{S} + \mathbf{L} = \Phi_{oo}^{-1}, \\ &\mathbf{S} - \mathbf{S}^* = \mathbf{0}, \mathbf{L} - \mathbf{L}^* = \mathbf{0}. \end{aligned} \tag{2.22}$$

Section 2.2 has provided certain sufficient conditions for the unique decomposition, in the space of skew symmetric matrices. As mentioned in Remark 2.2.8, one can extend the results to the space of complex Hermitian matrices also, which can be applied to solve (2.22). Then, $\text{support}(\widehat{\mathbf{S}}_\gamma)$ recovers the exact moral graph among the observable nodes, $\text{kin}(\mathcal{G}_o)$, and $\text{rank}(\widehat{\mathbf{L}}_\gamma)$ provides a lower bound on the number of hidden nodes. Additionally, as shown in Theorem 2.4.3, $\text{support}(\widehat{\mathbf{L}}_\gamma)$ retrieves Markov Blankets of all the hidden nodes.

2.4.6 Reconstruction based on the Imaginary Part of IPSDM

In Theorem 2.4.1, it was shown that $\mathbf{S}$ is sparse and $\mathbf{L}$ is low-rank, which implies that the same applies respectively to $\Im\{\mathbf{S}\}$ and $\Im\{\mathbf{L}\}$. Based on Theorem 2.4.1, one can show that $|\text{support}(\Im\{\mathbf{S}\})| \leq 2|\mathcal{E}_o| + 2|\mathcal{E}_o^{ss}|$ and $\text{rank}(\Im\{\mathbf{L}\}) \leq 4n_h$. Therefore, $\Im\{\mathbf{S}\}$ is sparse and $\Im\{\mathbf{L}\}$ is low-rank. As Φ_{oo}^{-1} is Hermitian, $\Im\{\Phi_{oo}^{-1}\}$ is skew symmetric, and hence the results from Section II are applicable here. Then, by applying the convex optimization (2.10) with $\mathbf{C} = \Im\{\Phi_{oo}^{-1}(z)\}$, one can retrieve the ground truth $\Im\{\mathbf{S}(z)\}$ and $\Im\{\mathbf{L}(z)\}$, for all $|z| = 1$, and appropriately selected γ as shown in Section 2.2. This procedure is provided in Algorithm 1. The caveat of decomposing $\Im\{\Phi_{oo}^{-1}\}$ is that the following assumption is required for consistent moral graph/topology reconstruction from $\Im\{\mathbf{S}\}$ and $\Im\{\mathbf{L}\}$.

Assumption 2.4.3. *For any $1 \leq i, k \leq n$, $i \neq k$, if $\mathbf{H}_{ik}(z) \neq 0$, then $\Im\{\mathbf{H}_{ik}(z)\} \neq 0$, for all z , $|z| = 1$.*

Remark 2.4.7. *Assumption 3 is necessary to reconstruct $\text{kin}(\mathcal{G}_o)$ and the Markov Blankets of the hidden nodes from the decomposition of $\Im\{\Phi_{oo}^{-1}\}$, instead of decomposing Φ_{oo}^{-1} directly. From (2.19), (2.20), it follows that $\Im\{\mathbf{S}\}$ and $\Im\{\mathbf{L}\}$ depend on elements of $\Im\{\mathbf{H}\}$. Hence, when Assumption 2.4.3 holds, the Lemma 2.4.1, Proposition 2.4.1, Theorem 2.4.1, 2.4.2, and 2.4.3 hold by replacing Φ_x^{-1} , Φ_{oo}^{-1} , $\mathbf{S}$, and $\mathbf{L}$ with $\Im\{\Phi_x^{-1}\}$, $\Im\{\Phi_{oo}^{-1}\}$, $\Im\{\mathbf{S}\}$, and $\Im\{\mathbf{L}\}$ respectively. In other words, $\text{kin}(\mathcal{G}_o)$ and the Markov Blanket of hidden nodes can be obtained by decomposing $\Im\{\Phi_{oo}^{-1}\}$ instead of Φ_{oo}^{-1} . We focus on the decomposition of $\Im\{\Phi_{oo}^{-1}\}$ in this chapter.*

Remark 2.4.8. *$\text{kin}(\mathcal{G}_o)$, obtained from $\Im\{\mathbf{S}\}$ using Proposition 2.4.1, would contain $\text{top}(\mathcal{G}_o)$ as well as additional edges due to strict spouses in $\mathcal{G}_o$. Likewise, (M_l, Q_l) obtained from $\text{support}(\Im\{\mathbf{L}\})$ using Theorem 2.4.3 may contain edges apart from $P(l) \cup C(l)$ of a hidden node l . Such spurious edges maybe many; examples include bi-directed LDGs. There is a need to eliminate them so that exact recovery of $\text{top}(\mathcal{G})$ is possible.*

In the next section, we develop techniques to eliminate the spurious edges and reconstruct the exact topology, for a wide class of networks.

2.5 Exact Topology Reconstruction

In this section, we develop methods for exact recovery of $top(\mathcal{G})$, under Assumption 2.5.1, which is applicable to wide class of applications.

2.5.1 Elimination of Strict Spouse Edges

Here, we show that, in certain LDMs, the strict spouse edges satisfy properties that can be exploited toward exact topology reconstruction. Notice that for *some* of the results (viz Theorem 2.5.1 and Theorem 2.5.2), we restrict our interest to the models that satisfy the following assumption. For the networks that satisfy Assumption 2.5.1, Theorem 2.5.1 will show that $support(\Im\{\mathbf{S}\})$ would reconstruct the exact topology, $top(\mathcal{G}_o)$ (not $kin(\mathcal{G}_o)$).

Assumption 2.5.1. *For the LDM in (2.14), and $i, k, l \in \mathcal{V}$, if $\mathbf{H}_{ki}(z) \neq 0$ and $\mathbf{H}_{kl}(z) \neq 0$, then $\angle \mathbf{H}_{ki}(z) = \angle \mathbf{H}_{kl}(z)$.*

Remark 2.5.1. *Assumption 2.5.1 is satisfied by a large class of engineering systems. For instance, Section 2 of [18] provides engineering systems that satisfy Assumption 2.5.1. Also, other examples include linearized chemical reaction ODEs [66]. For example, see (10.2.2) in [66].*

The following lemma (Theorem 3 in [18]) is useful in proving the subsequent results and helps in exploiting additional structure enjoyed by the IPSDM in the networks satisfying Assumption 2.5.1.

Lemma 2.5.1. *Consider a well-posed and topologically detectable LDM $(\mathbf{H}, \mathbf{e})$ with the associated graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, described by (2.14), having full node observability and satisfying Assumption 2.5.1. If i and j are strict spouses, then $[\Im\{\Phi_{\mathbf{x}}^{-1}\}]_{ij} = 0$.*

Lemma 2.5.1 eliminates the spurious edges formed due to strict spouse connections by observing the entries of $\Im\{\Phi_x^{-1}\}$. The proof follows from the expansion of (2.16). Here, $[\Im\{\Phi_x^{-1}\}]_{ij} = 0$ if $\mathbf{H}_{ij} = 0$, $\mathbf{H}_{ji} = 0$, and $\angle \mathbf{H}_{ki} = \angle \mathbf{H}_{kj}$ for all $k \in \mathcal{C}(i) \cap \mathcal{C}(j)$. This can be employed to separate the true parent-child connection from the strict spouse edges in the moral graph obtained from Lemma 2.4.1. Combining Lemma 2.4.1

and Lemma 2.5.1 with Assumption 2.4.3, one can conclude that, the undirected graph constructed from $\Im\{\Phi_x^{-1}\}$ is equal to $top(\mathcal{G})$, if Assumption 2.5.1 holds; else it would be $kin(\mathcal{G})$. We use this fact in the following theorem to reconstruct $top(\mathcal{G}_o)$.

Theorem 2.5.1. *Consider a well-posed and topologically detectable LDM $(\mathbf{H}, \mathbf{e})$ with the associated graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, described by (2.14) and satisfying Assumption 2.5.1. Let $\mathbf{S}$ be given by (2.19) and let $\hat{\mathcal{E}}_o := \{(i, j) : \Im\{\mathbf{S}_{ij}\} \neq 0, i < j\}$. Then, $\hat{\mathcal{E}}_o \subseteq \bar{\mathcal{E}}_o$. Additionally, if the LDM satisfies Assumption 2.4.3, then $\hat{\mathcal{E}}_o = \bar{\mathcal{E}}_o$ almost always.*

Proof. See Appendix A.10. □

In the following, we discuss the exact topology reconstruction from $\Im\{\Phi_{oo}^{-1}\}$ based on Theorems 2.4.3 and 2.5.1.

Algorithm 1 Matrix decomposition

Input: $\Phi_{oo}^{-1}(z)$: IPSDM among $\mathcal{V}_o, \varepsilon, z = e^{j\omega}, \omega \in (-\pi, \pi]$

Output: Matrices $\Im(\mathbf{S}(z))$ and $\Im(\mathbf{L}(z))$

- 1: Set $\mathbf{C} = \Im\{\Phi_{oo}^{-1}(z)\}$
 - 2: Initialize $(\hat{\mathbf{S}}_0, \hat{\mathbf{L}}_0) = (\mathbf{C}, \mathbf{0})$
 - 3: **for all** $t \in \{\varepsilon, 2\varepsilon, \dots, 1\}$ **do**
 - 4: Solve the convex optimization (2.10) and calculate $diff_t$ in (2.12)
 - 5: **end for**
 - 6: Identify the three regions where $diff_t$ is zero and denote the middle region as $[t_1, t_2]$.
 Pick a $t_0 \in [t_1, t_2]$ and the corresponding pair $(\hat{\mathbf{S}}_{t_0}, \hat{\mathbf{L}}_{t_0})$.
 - 7: **if** $deg_{max}(\hat{\mathbf{S}}_{t_0})inc(\hat{\mathbf{L}}_{t_0}) < \frac{1}{12}$ **then**
 - 8: $(\hat{\mathbf{S}}(z), \hat{\mathbf{L}}(z)) = (\hat{\mathbf{S}}_{t_0}, \hat{\mathbf{L}}_{t_0})$
 - 9: Return $(\hat{\mathbf{S}}(z), \hat{\mathbf{L}}(z))$
 - 10: **end if**
-

2.5.2 Reconstruction of Topology:

The topology, $\mathcal{T}(\mathcal{V}, \mathcal{E})$, of the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$ will be reconstructed in three steps: (a) recover the topology restricted to observed nodes given by $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_o)$, (b) determine the number of hidden nodes n_h in the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and (c) reconstruct the topology associated with each hidden node. We denote the reconstructed topology as $\mathcal{T}(\mathcal{V}_R, \mathcal{E}_R)$. The reconstruction is said to be exact when $\mathcal{T}(\mathcal{V}, \mathcal{E}) = \mathcal{T}(\mathcal{V}_R, \mathcal{E}_R)$.

The first step in topology reconstruction is to obtain $\hat{\mathbf{S}} = \Im\{\mathbf{S}(z)\}$ and $\hat{\mathbf{L}} = \Im\{\mathbf{L}(z)\}$ from Algorithm 1. Then, from $\hat{\mathbf{S}}$, we reconstruct the topology among observed nodes

$\mathcal{T}(\mathcal{V}_o, \mathcal{E}_o)$ as shown in Theorem 2.5.1. This is performed in Algorithm 2, which reconstructs $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_R)$, where $\mathcal{V}_o$ is the set of observable nodes and $\mathcal{E}_R$ is the undirected edge set reconstructed from $\text{support}(\widehat{\mathbf{S}})$. Apart from pathological cases, $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_R)$ is identical to $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_o)$ under Assumption 2.5.1.

Remark 2.5.2. *Algorithms 1 and 2 are applicable for any LDM with Assumption 2.4.3 and need not satisfy Assumption 2.5.1. However, in that case, $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_R)$ would return $\text{kin}(\mathcal{G}_o)$.*

Remark 2.5.3. *As mentioned in Section 2.4.5, one can decompose the complex Φ_{oo}^{-1} to obtain $\mathbf{S}$ and $\mathbf{L}$, instead of their respective imaginary parts, thus avoiding the need for Assumption 2.4.3. $\mathbf{S}$ and $\mathbf{L}$ can then retrieve moral graphs among the observable nodes and Markov Blanket of hidden nodes.*

Algorithm 2 Topology reconstruction of observable nodes

Input: $\widehat{\mathbf{S}}(z)$ from Algorithm 1, threshold τ

Output: Reconstructed topology among observable nodes $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_R)$

```

1: Edge sets  $\mathcal{E}_R \leftarrow \{\}$ 
2: for all  $(i, j) \in \{1, 2, \dots, n\} \times \{1, 2, \dots, n\}$  do
3:   if  $|\widehat{\mathbf{S}}_{ij}(z)| > \tau$  then
4:      $\mathcal{E}_R \leftarrow \mathcal{E}_R \cup \{(i, j)\}$ 
5:   end if
6: end for
7:  $\mathcal{V}_o \leftarrow \{1, 2, \dots, n\}$ 
8: Return  $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_R)$ 

```

We now proceed with estimating number of hidden nodes and reconstructing the topology associated with hidden nodes using $\Im\{\mathbf{L}\}$.

We emphasize that the undirected edges in Q_l are not necessarily present in the true topology $\mathcal{T}(\mathcal{V}, \mathcal{E})$. The only task remaining in constructing the topology $\mathcal{T}(\mathcal{V}, \mathcal{E})$ is finding the true parents or children of the hidden nodes. For this purpose, we consider each connected component (M_l, Q_l) , and reconstruct the topology associated with the hidden node l .

The following result is useful in reconstructing the exact topology associated with hidden node l from *degree* of nodes in (M_l, Q_l) . It shows that the information about the degree of each node is *sufficient* to reconstruct the topology associated with the hidden node l .

Theorem 2.5.2. *Suppose the LDM in (2.17) satisfies Assumptions 2.4.1-2.5.1. Consider a hidden node $l \in \mathcal{V}_h$ and its associated undirected graph (M_l, Q_l) as defined in Theorem 2.4.3 for $\mathfrak{S}\{\mathbf{L}\}$. Define $\alpha_l := \max_{j \in M_l} \deg_{M_l}(j)$, where $\deg_{M_l}(i) := |\{j \in M_l \setminus i : (i, j) \in Q_l\}|$. The following holds: $\exists k \in M_l$ such that $\deg_{M_l}(k) < \alpha_l$ if and only if $|\mathcal{P}(l) \setminus (\mathcal{C}(l) \cup \mathcal{S}(l))| \geq 2$ or $|\mathcal{S}(l) \setminus (\mathcal{C}(l) \cup \mathcal{P}(l))| \geq 2$ in $\mathcal{G}(\mathcal{V}, \mathcal{E})$.*

Proof. See Appendix A.11. □

Remark 2.5.4. *From the proof of the above theorem, the following holds:*

- (a) *Consider a node $a_1 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$. Then, for every $a_2 \in (\mathcal{C}(l) \cup \mathcal{P}(l) \cup \mathcal{S}(l)) \setminus \{a_1\}$, $(a_1, a_2) \in Q_l$.*
- (b) *Consider a strict spouse, $s_1 \in \mathcal{S}(l) \setminus (\mathcal{C}(l) \cup \mathcal{P}(l))$. Then, for any $a_1 \in \mathcal{C}(l) \cup \mathcal{P}(l)$, $(s_1, a_1) \in Q_l$.*
- (c) *Consider a strict parent, $p_1 \in \mathcal{P}(l) \setminus (\mathcal{C}(l) \cup \mathcal{S}(l))$. Then, for any $a_1 \in \mathcal{C}(l) \cup \mathcal{S}(l)$, $(p_1, a_1) \in Q_l$.*

Regardless of the number of strict spouses and strict parents in the LDG, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, for $M_l = \mathcal{C}(l) \cup \mathcal{P}(l) \cup \mathcal{S}(l)$, the following holds from (a), (b) and (c): for $i \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$ and for $k \in M_l \setminus i$, $(i, k) \in Q_l$. That is, a node from $\mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$ is connected to every other node from M_l in Q_l . Hence, $\deg_{M_l}(i) = |M_l| - 1$ and it is the node with maximum degree. Therefore, $\mathcal{C}(l) \cup \{\mathcal{P}(l) \cap \mathcal{S}(l)\} = \{i | i \in M_l, \deg_{M_l}(i) = |M_l| - 1\}$. The nodes in M_l with $\deg_{M_l}(i) < |M_l| - 1$ are either strict parents or strict spouses of hidden node l .

We provide the following assumption which is needed for exact reconstruction of topology associated with hidden nodes. When the below assumption is violated, that is, if the number of strict spouse for a hidden node h is one, then there will be a single false edge present in the reconstructed topology associated with the hidden node. Rephrasing, the strict spouse of a hidden node h will be considered as a neighbor in the reconstructed topology associated with the hidden node h . The number of such false edges reconstructed for a network is limited to a maximum of one per hidden node. Nevertheless, to avoid these false edges associated with a hidden node, we make the following assumption.

Assumption 2.5.2. *In the LDG, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, for any hidden node, $h \in \mathcal{V}_h$, if there exists a strict spouse, then there must exist at least one more strict spouse associated with h .*

Based on the above theorem and assumption, we propose Algorithm 3, which outputs the reconstructed topology $\mathcal{T}(\mathcal{V}_R, \mathcal{E}_R)$ that is identical to the true topology $\mathcal{T}(\mathcal{V}, \mathcal{E})$ of the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$. The Algorithm 3 consists of three parts as outlined below.

- (a) From support of $\mathfrak{S}\{\mathbf{L}\}$, determine V_H and E_H . The graph (V_H, E_H) will be a disjoint collection of connected subgraphs $\bigcup_{l=1}^{n_h} (M_l, Q_l)$. The number of hidden nodes, n_h , is given by the number of connected subgraphs. This is done in steps 1 – 10. $\mathcal{V}_H$ and $\mathcal{E}_H$ are both initialized with $\{\}$ (steps 11 – 12).
- (b) For each (M_l, Q_l) , we create a hidden node h_l and add this to $\mathcal{V}_H$ (steps 15 – 16). Next, we construct the topology associated with h_l . For this, we compute the degree of each node in (M_l, Q_l) and calculate its maximum, α_l (step 17). We check if there is a node in M_l with degree smaller than α_l . If there is no node in M_l with degree smaller than α_l , then add undirected edge (h_l, i) to $\mathcal{E}_H$ for all $i \in M_l$ (steps 18 – 28). Otherwise, we collect the nodes in M_l with degree α_l in the set d_{h_l} and the nodes with degree smaller than α_l in the set $\widetilde{M}_l$ (steps 29 – 38). The vertex set $d_{h_l} = \mathcal{C}(h_l) \cup \mathcal{P}(h_l) \cup \mathcal{S}(h_l)$. Thus, add (h_l, i) to $\mathcal{E}_H$ for all $i \in d_{h_l}$ (steps 39 – 41). The nodes in $\widetilde{M}_l$ are either strict parents of h_l or strict spouses of h_l in $\mathcal{G}(\mathcal{V}, \mathcal{E})$. We find the strict parents from $\widetilde{M}_l$ and add their edges with h_l to $\mathcal{E}_H$ (steps 42 – 47).
- (c) Repeat (b) for all $l = \{1, 2, \dots, n_h\}$ (step 13). Assign $\mathcal{V}_R$ as $\mathcal{V}_H \cup \mathcal{V}_o$ and $\mathcal{E}_R$ as $\mathcal{E}_H \cup \mathcal{E}_o$ (steps 49 – 50). The reconstructed topology of the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is $\mathcal{T}(\mathcal{V}_R, \mathcal{E}_R)$ (step 51).

The reconstructed topology $\mathcal{T}(\mathcal{V}_R, \mathcal{E}_R)$ is identical to the true topology $\mathcal{T}(\mathcal{V}, \mathcal{E})$ of the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$.

Till now, we have investigated topology identification under the assumption that the perfect PSDM is available. However, in practice, we have access only to finite number of observations of the time-series. In the next section, we show that if the number of observations, N , is large enough, then each entry of the estimated IPSDM can be brought ϵ -close to the actual IPSDM. Note that we do not assume the presence of any latent node here. The analysis in this section is applicable to networks with or without latent nodes.

Algorithm 3 Full Topology reconstruction with hidden nodes

Input: $\widehat{\mathbf{L}}(z)$ from Algorithm 1, threshold τ and $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_R)$ from Algorithm 2

Output: $\mathcal{T}(\mathcal{V}_R, \mathcal{E}_R)$, reconstructed topology of the LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$

```

1:  $V_H \leftarrow \{\}$ 
2:  $E_H \leftarrow \{\}$ 
3: for all  $(i, j) \in \{1, 2, \dots, n\} \times \{1, 2, \dots, n\}$  do
4:   if  $|\widehat{\mathbf{L}}_{ij}(z)| > \tau$  then
5:      $E_H \leftarrow E_H \cup \{(i, j)\}$ 
6:      $V_H \leftarrow V_H \cup \{i, j\}$ 
7:   end if
8: end for
9:  $(V_H, E_H) = \bigcup_{l=1}^{n_h} (M_l, Q_l)$  such that  $M_{l_1} \cap M_{l_2} = Q_{l_1} \cap Q_{l_2} = \emptyset \forall l_1, l_2 \in \{1, 2, \dots, n_h\}, l_1 \neq l_2$ .
10:  $n_h$  : number of disjoint connected subgraphs in  $E_H$ 
11: Vertex set  $\mathcal{V}_H \leftarrow \{\}$ 
12: Edge set  $\mathcal{E}_H \leftarrow \{\}$ 
13: for all  $l \in \{1, 2, \dots, n_h\}$ 
14:   flag = 0
15:   add a hidden node  $h_l$ 
16:    $\mathcal{V}_H \leftarrow \mathcal{V}_H \cup \{h_l\}$ 
17:    $\alpha_l := \max_{j \in M_l} \deg_{M_l}(j)$ 
18:   for all  $i \in M_l$  do
19:     if  $\deg_{M_l}(i) < \alpha_l$  then
20:       flag = 1
21:       break
22:     end if
23:   end for
24:   if flag == 0 then
25:     for all  $i \in M_l$  do
26:        $\mathcal{E}_H \leftarrow \mathcal{E}_H \cup \{(i, h_l)\}$ 
27:     end for
28:   end if
29:   If flag == 1
30:      $d_{h_l} \leftarrow \{\}, \widehat{M}_l \leftarrow \{\}$ 
31:     for all  $i \in M_l$  do
32:       if  $\deg_{M_l}(i) == \alpha_l$  then
33:          $d_{h_l} \leftarrow \{d_{h_l}\} \cup \{i\}$ 
34:       end if
35:       if  $\deg_{M_l}(i) < \alpha_l$  then
36:          $\widehat{M}_l \leftarrow \widehat{M}_l \cup \{i\}$ 
37:       end if
38:     end for

```

```

1: for all  $k \in d_{h_l}$  do
2:    $\mathcal{E}_H \leftarrow \mathcal{E}_H \cup \{(k, h_l)\}$ 
3: end for
4: for all  $k \in \widetilde{M}_l$  do
5:   if  $(k, d) \notin \mathcal{E}_R$  for all  $d \in d_{h_l}$  then
6:      $\mathcal{E}_H \leftarrow \mathcal{E}_H \cup \{(k, h_l)\}$ 
7:   end if
8: end for
9: end for
10:  $\mathcal{V}_R \leftarrow \mathcal{V}_H \cup \mathcal{V}_o$ 
11:  $\mathcal{E}_R \leftarrow \mathcal{E}_H \cup \mathcal{E}_R$ 
12: Return  $\mathcal{V}_R, \mathcal{E}_R$ 

```

2.6 IPSD Estimation from Finite time-series

In this section, we investigate the effect of finite time-series on the estimation of the IPSDM. Suppose that we are given the time-series $\{\mathbf{x}(t) \in \mathbb{R}^n\}_{t=1}^N$. We show that, with high probability, the entries of an estimated IPSDM are close to the original IPSDM, if the number of samples, N is large enough. To begin with, recall that (2.13) can be rewritten as $\widetilde{\mathbf{x}}(t) = \sum_{k=-\infty}^{\infty} \widetilde{\xi}(k) \widetilde{\mathbf{e}}(t-k)$, $t \in \mathbb{Z}$, where $\xi(z) := (\mathbf{I} - \mathbf{H}(z))^{-1}$. Similar to [67], for the convenience of analysis, we define:

$$\mathcal{H}_\rho(l, L) := \{(\mathbf{x}(t), t \in \mathbb{Z}) : 0 < l \leq |\lambda(\xi(z))| \leq L, \text{ for } |z| \leq \rho\}, \quad (2.23)$$

where $\lambda(\xi(z))$ denotes eigenvalues of $\xi(z)$, and restrict ourselves to the family $\mathcal{H}_\rho(l, L)$.

2.6.1 Estimation Error

Here, we characterize the difference between the actual PSDM, $\Phi_{\mathbf{x}}(z)$, in (2.15) and a PSDM estimated from finite time-series. We assume that the true correlation function $R_{\mathbf{x}}(k) = \mathbb{E}[\widetilde{\mathbf{x}}(0)\widetilde{\mathbf{x}}^T(k)]$ follows the relation

$$\|R_{\mathbf{x}}(k)\|_\infty \leq C_1 \rho^k, \quad (2.24)$$

for some ρ such that $|\rho| < 1$ (recall that $\widetilde{\mathbf{x}}$ is WSS). Note that this assumption is common in scalar Gaussian processes and is an example of strongly mixing process (see [67] and the references therein for details).

Let $\bar{\Phi}_{\mathbf{x}}(z)$ denote the truncated version of the PSDM, $\Phi_{\mathbf{x}}(z)$, and let $\hat{\Phi}_{\mathbf{x}}(z)$ be the PSDM estimated from data. That is, $\bar{\Phi}_{\mathbf{x}}(z) = \sum_{k=-p}^p R_{\mathbf{x}}(k)z^{-k}$, and $\hat{\Phi}_{\mathbf{x}}(z) = \sum_{k=-p}^p \hat{R}_{\mathbf{x}}(k)z^{-k}$, where $\hat{R}_{\mathbf{x}}(k) = \frac{1}{N-k} \sum_{l=1}^{N-k} \mathbf{x}(l)\mathbf{x}^T(l+k)$ is the estimated correlation matrix. We employ a concentration bound for $\|\Phi_{\mathbf{x}}(z) - \hat{\Phi}_{\mathbf{x}}(z)\|_{\infty}$ to bound $\|\Phi_{\mathbf{x}}^{-1}(z) - \hat{\Phi}_{\mathbf{x}}^{-1}(z)\|_{\infty}$, for any $|z| = 1$. By triangle inequality,

$$\|\Phi_{\mathbf{x}}(z) - \hat{\Phi}_{\mathbf{x}}(z)\|_{\infty} \leq \|\Phi_{\mathbf{x}}(z) - \bar{\Phi}_{\mathbf{x}}(z)\|_{\infty} + \|\bar{\Phi}_{\mathbf{x}}(z) - \hat{\Phi}_{\mathbf{x}}(z)\|_{\infty}. \quad (2.25)$$

$\|\Phi_{\mathbf{x}}(z) - \bar{\Phi}_{\mathbf{x}}(z)\|_{\infty}$ is the truncation error, which is the error in truncating PSDM to order p , and $\|\bar{\Phi}_{\mathbf{x}}(z) - \hat{\Phi}_{\mathbf{x}}(z)\|_{\infty}$ is the estimation error, which denotes the error in estimating the p -th order truncated PSDM.

The following proposition provides a bound on element-wise distance between $\bar{\Phi}_{\mathbf{x}}$ and $\Phi_{\mathbf{x}}$.

Proposition 2.6.1. *Consider a linear dynamic system governed by (2.13). Suppose that the autocorrelation function $R_{\mathbf{x}}(k)$ satisfies (2.24). For any $\varepsilon > 0$, if $p \geq \log_{\rho} \left(\frac{(1-\rho)\varepsilon}{2C_1} \right) - 1$, then the truncation error $\|\Phi_{\mathbf{x}}(z) - \bar{\Phi}_{\mathbf{x}}(z)\|_{\infty} \leq \varepsilon$.*

Proof. See Appendix A.12. □

Note that,

$$\begin{aligned} \|\bar{\Phi}_{\mathbf{x}}(z) - \hat{\Phi}_{\mathbf{x}}(z)\|_{\infty} &= \left\| \sum_{k=-p}^p [R_{\mathbf{x}}(k) - \hat{R}_{\mathbf{x}}(k)] z^{-k} \right\|_{\infty} \\ &\leq \sum_{k=-p}^p \|R_{\mathbf{x}}(k) - \hat{R}_{\mathbf{x}}(k)\|_{\infty}. \end{aligned} \quad (2.26)$$

Thus, we can upper bound $\|\bar{\Phi}_{\mathbf{x}}(z) - \hat{\Phi}_{\mathbf{x}}(z)\|_{\infty}$ by bounding on $\|R_{\mathbf{x}}(k) - \hat{R}_{\mathbf{x}}(k)\|_{\infty}$.

Next, we provide probably approximately correct (PAC) bounds for estimating auto-correlation matrices, which in turn is used in obtaining PAC bounds on PSDMs. The following proposition bounds deviation of each individual elements of auto-correlation matrix.

Proposition 2.6.2. *For every delay index $k \leq N - n$, $k \in \{0, \dots, p\}$, and all $\varepsilon > 0$, we have*

$$\mathbb{P} \left(\|\hat{R}_{\mathbf{x}}(k) - R_{\mathbf{x}}(k)\|_{\infty} > \varepsilon \right) \leq n^2 \exp \left(-(N - k) \min \left\{ \frac{\varepsilon^2}{32n^2C_1^2}, \frac{\varepsilon}{8nC_1} \right\} \right). \quad (2.27)$$

Proof. See Appendix A.13. $\square$

Now, we are ready to obtain the following lemma, which bounds the estimation error in (2.25) using (2.26).

Lemma 2.6.1. *For every $z = e^{j\omega}$, $\omega \in (-\pi, \pi]$, the error in estimating truncated PSDM $\hat{\Phi}_{\mathbf{x}}(z)$ is $\|\bar{\Phi}_{\mathbf{x}}(z) - \hat{\Phi}_{\mathbf{x}}(z)\|_{\infty} \leq \varepsilon$ with probability at least*

$$1 - n^2 \exp \left(-(N - p) \min \left\{ \frac{\varepsilon^2}{32(2p + 1)^2 n^2 C_1^2}, \frac{\varepsilon}{8(2p + 1)n C_1} \right\} \right).$$

Proof. See Appendix A.14. $\square$

By combining the above lemmas, we obtain the following theorem.

Theorem 2.6.1. *Consider a linear dynamic system governed by (2.13). Suppose that the autocorrelation function $R_{\mathbf{x}}(k)$ satisfies (2.24). Let $0 < \varepsilon_1 < \varepsilon$. Then, for any p such that $p \geq \log_{\rho} \left(\frac{(1-\rho)(\varepsilon-\varepsilon_1)}{2C_1} \right)$, we have $\|\Phi_{\mathbf{x}}(z) - \hat{\Phi}_{\mathbf{x}}(z)\|_{\infty} \leq \varepsilon$, with probability at least*

$$1 - n^2 \exp \left(-(N - p) \min \left\{ \frac{\varepsilon_1^2}{32(2p + 1)^2 n^2 C_1^2}, \frac{\varepsilon_1}{8(2p + 1)n C_1} \right\} \right).$$

Proof. The proof follows by combining Lemma 2.6.1 and Proposition 2.6.1. $\square$

2.6.2 IPSDM Estimation Error

Next, we provide a bound for the difference between the original and the estimated IPSDMs in terms of the difference between corresponding PSDMs.

Theorem 2.6.2. *Consider a linear dynamic system governed by (2.13) satisfying the Assumption (2.23). Suppose that the autocorrelation function $R_{\mathbf{x}}(k)$ satisfies (2.24). Let $0 < \varepsilon_1 < \varepsilon$. Then, for any p such that $p \geq \log_{\rho} \left(\frac{(1-\rho)(\varepsilon-\varepsilon_1)}{2C_1} \right) - 1$, we have*

$$\|\Phi_{\mathbf{x}}^{-1} - \hat{\Phi}_{\mathbf{x}}^{-1}\|_{\infty} \leq \frac{\sqrt{n} L^2 \sigma_{e[1]}^2}{l^4 \sigma_{e[n_o]}^4} \left(\frac{\sqrt{n} \varepsilon}{l^2 \sigma_{e[n_o]}^2 - \sqrt{n} \varepsilon} \right) \quad (2.28)$$

with probability at least

$$1 - n^2 \exp \left(-(N - p) \min \left\{ \frac{\varepsilon_1^2}{32(2p + 1)^2 n^2 C_1^2}, \frac{\varepsilon_1}{8(2p + 1)n C_1} \right\} \right),$$

where $\sigma_{e_{[1]}}$ and $\sigma_{e_{[n_o]}}$ denote the largest and the smallest eigenvalues of $\Phi_e(z)$ respectively.

Proof. See Appendix A.15. □

It follows from Theorem 2.6.2 that the element-wise distance between $\Phi_{\mathbf{x}}^{-1}$ and $\hat{\Phi}_{\mathbf{x}}^{-1}$ can be made arbitrarily small with high probability, by picking N large enough.

2.7 Simulation results

In this section, we validate our algorithms with simulations. All the simulations are performed in Matlab; to solve the optimization problem for matrix decomposition, we use YALMIP [68] with SDPT3 [69] solver.

For the simulations, we assume that we have access to the perfect IPSDM, Φ_{oo}^{-1} . From $\Im\{\Phi_{oo}\}$ alone, we aim to reconstruct the topology of the LDM which involves (a) recover the topology restricted to observed nodes given by $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_o)$, (b) determine the number of hidden nodes n_h in LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$, and (c) reconstruct topology associated with each hidden node.

Remark 2.7.1. *The objective of our simulation is to demonstrate the working of Algorithm 1 to decompose $\Im\{\Phi_{oo}^{-1}\}$ of this network into $\hat{\mathbf{S}} = \Im\{\mathbf{S}\}$ and $\hat{\mathbf{L}} = \Im\{\mathbf{L}\}$. Then, by applying algorithms 2 and 3 on $\hat{\mathbf{S}}$ and $\hat{\mathbf{L}}$, we can retrieve the original topology.*

2.7.1 LDM

For the simulation, we assume that Φ_{oo} of a topologically detectable LDM in (2.17) with 32 nodes, having the LDG as shown in Figure 2.2 is given. The nodes colored in red (30, 31, 32) are hidden and the rest are observed. Note that the network in Figure 2.2 and the corresponding LDM satisfies the assumptions 2.4.1-2.5.2.

2.7.2 Sparse plus low-rank decomposition of the Imaginary of IPSDM

We applied Algorithm 1 for matrix decomposition of $\Im\{\Phi_{oo}^{-1}(z)\}$ to obtain $\Im\{\mathbf{S}(z)\} = \hat{\mathbf{S}}$ and $\Im\{\mathbf{L}(z)\} = \hat{\mathbf{L}}$. Here, we took $z = e^{j\frac{3\pi}{8}}$ and $\epsilon = 0.01$ for running Algorithm 1.

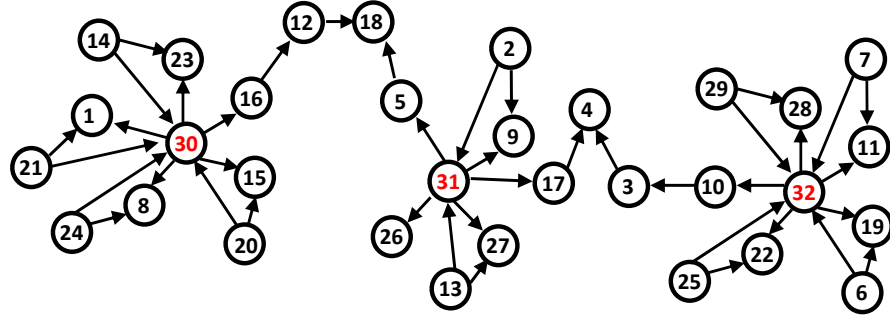


Figure 2.2: LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with hidden nodes shown in red.

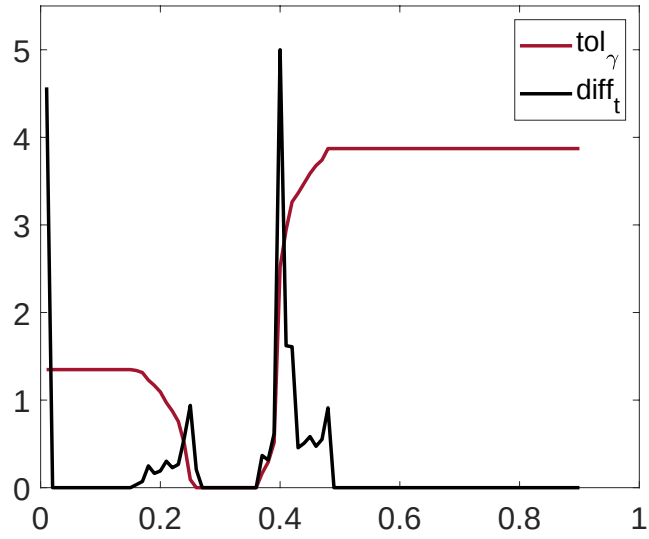
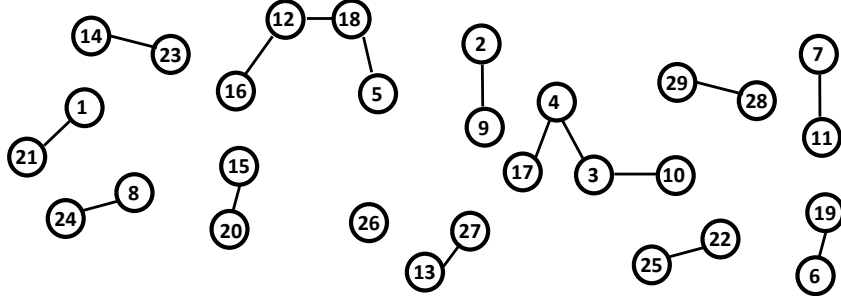
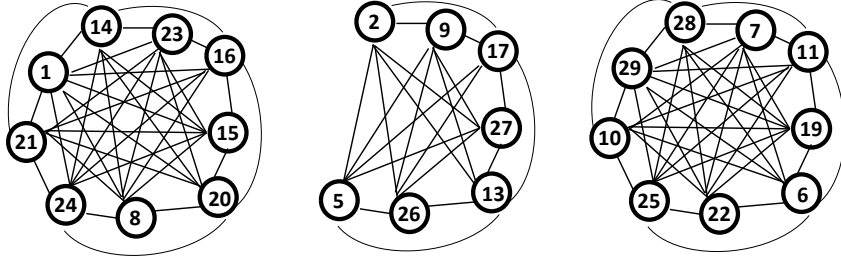


Figure 2.3: Comparison between tol_t and $diff_t$ for decomposition of $\Im\{\Phi_{oo}^{-1}\}$ obtained by the LDG in Figure 2.2



(a) Reconstructed topology among observable nodes, $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_o)$ obtained from $\mathfrak{S}\{\mathbf{S}\}$ using Algorithm 2.



(b) (V_H, E_H) obtained from $\mathfrak{S}\{\mathbf{L}\}$ (step 9 in Algorithm 3).

Figure 2.4: Applying Algorithms 2 on $\mathfrak{S}\{\mathbf{S}\}$ and Algorithm 3 on $\mathfrak{S}\{\mathbf{L}\}$.

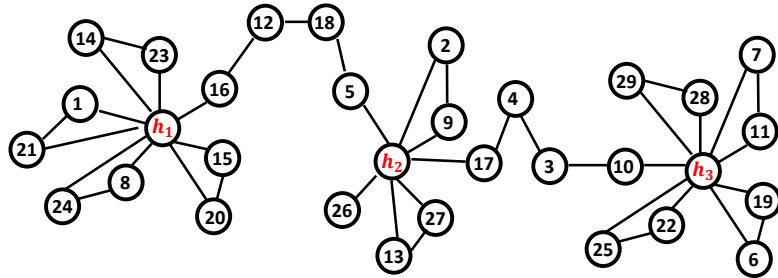


Figure 2.5: $\mathcal{T}(\mathcal{V}_R, \mathcal{E}_R)$: Final reconstructed topology with hidden nodes.

Figure 2.3 shows the values of tol_t and $diff_t$ corresponding to the decomposition of $\Im\{\Phi_{oo}^{-1}(z)\}$ for various values of t . From the plots, we see that the $diff_t$ plot shows three zero regions, similar to what was proposed in Proposition 2.2.3. We would like to stress that the network does not satisfy the sufficient condition of Lemma 2.2.4. Nevertheless, the plot can retrieve the true sparse and low-rank matrices by picking a t in the middle zero region, as described by Algorithm 2. This network belongs to the subclass of the networks that does not satisfy the sufficient condition of Lemma 2.2.4, but satisfy the necessary and sufficient condition of (2.7).

We make three important observations in Figure 2.3: (a) for small t , we observe a beginning zero region where $diff_t$ is zero. This verifies Proposition 2.2.3: when the value of t is very small (smaller than 0.15), the optimal objective value is obtained by setting zero to $\Im\{\mathbf{L}\}$ and the objective achieves the minimum value of zero at $(\Im\{\mathbf{S}\} + \Im\{\mathbf{L}\}, 0)$. As t is increased but less than 0.15, we still obtain the optimal solution at $(\Im\{\mathbf{S}\} + \Im\{\mathbf{L}\}, 0)$, and hence, $diff_t$ is zero for t very small. (b) The opposite behavior is observed when t is very large (greater than 0.5). This is the end zero region, which corresponds to the optimal solution at $(0, \Im\{\mathbf{S}\} + \Im\{\mathbf{L}\})$. Nevertheless, the value of tol_t is quite high for both of these scenarios, since neither of the solutions are correct. (c) In the middle region ($t \in [0.24, 0.35]$), it is observed that both the tol_t and $diff_t$ are zero; the range specified by Proposition 2.2.2 (See Corollary 2.2.1).

Note that $diff_t$ need not be zero in the middle region. However, it is observed from simulations that if we pick a t corresponding to minimum value of $diff_t$ in the middle region between (0.3, 0.4), it is still possible to reconstruct topology. This is because the finite value of $diff_t$ in middle region is due to element wise mismatch of non-zero entries. However, the structural pattern of sparse and low-rank matrices is almost preserved. For topology reconstruction, structural pattern is more important than the exactness of the element-wise values.

2.7.3 Reconstruction of Topology

Topology reconstruction of the LDM involves (a) recover the topology restricted to observed nodes given by $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_o)$, (b) determine the number of hidden nodes n_h in LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and (c) reconstruct the topology associated with each hidden node.

From Figure 2.3, it is evident that $\Im\{\Phi_{oo}^{-1}(z)\}$ was decomposed uniquely into $\Im\{\mathbf{S}\}$ and $\Im\{\mathbf{L}\}$ when $t \in [0.3, 0.4]$, since tol_t is zero. Indeed, due to perfect retrieval, the optimal solution is $(\Im\{\mathbf{S}\}, \Im\{\mathbf{L}\})$. We now use the decomposed matrices $\Im\{\mathbf{S}\}$ and

$\Im\{\mathbf{L}\}$ with $\tau = 10^{-6}$ to obtain the following:

- (a) From $\Im\{\mathbf{S}\}$, we apply Algorithm 2 to obtain the topology of the subgraph restricted to the observable nodes. The reconstructed topology among observed nodes is shown in Figure 2.4a, which matches perfectly with topology of the LDG in Figure 2.2 restricted to observed nodes $\mathcal{T}(\mathcal{V}_o, \mathcal{E}_o)$.
- (b) From $\Im\{\mathbf{L}\}$, we apply Algorithm 3 and construct (V_H, E_H) , which is a union of three disjoint connected subgraphs, i.e., $(V_H, E_H) = \bigcup_{l=1}^3 (M_l, Q_l)$ (refer Figure 2.4b). Therefore, the number of hidden nodes in the LDG are three hidden nodes present. n_h is 3.
- (c) In Figure 2.4b, each connected component (M_l, Q_l) is a clique. For $l = \{1, 2, 3\}$, Algorithm 3 considers l^{th} clique (M_l, Q_l) , places a single hidden node h_l in the clique and reconstructs the topology associated with hidden node h_l . Since (M_l, Q_l) is a clique, $M_l = \mathcal{C}(h_l) \cup \{\mathcal{P}(h_l) \cap \mathcal{S}(h_l)\}$.

From (a), (b), and (c), Algorithm 3 reconstructs the complete topology including latent nodes (see Figure 2.5), which matches exactly with the topology of the LDG. To summarize, Algorithm 1 decomposed $\Im\{\Phi_{oo}^{-1}(z)\}$ into sum of sparse $\Im\{\mathbf{S}\}$ and low-rank $\Im\{\mathbf{L}\}$. Using $\Im\{\mathbf{S}\}$, Algorithm 2 reconstructs the topology reconstructed to observable nodes. Using $\Im\{\mathbf{L}\}$, Algorithm 3 estimates the number of hidden nodes, reconstructs the topology associated with each hidden node, and the full topology of the LDG in Figure 2.2 is reconstructed as shown in Figure 2.5, which matched exactly with the true topology of the LDG.

2.8 Conclusions

In this chapter, a novel approach to reconstruct the topology of networked linear dynamical systems with latent agents, from IPSDM of the observed nodes, is presented. The network was allowed to have directed loops and bi-directed edges. It was shown that the IPSDM can be uniquely decomposed into sparse and low-rank matrices. The sparse component unveiled the moral graph of the observed nodes, and the low-rank component retrieved the Markov Blanket associated with the latent nodes. Necessary and sufficient conditions for unique sparse plus low-rank decomposition of a skew-symmetric matrix were established, along with an optimization-based algorithm that decomposes

the skew-symmetric matrix to yield the sparse component $\mathbf{S}$ and the low-rank component $\mathbf{L}$. For a large class of systems, the unique decomposition of the imaginary part of the IPSDM is sufficient to achieve the moral graph of the observed nodes including the Markov Blanket of latent nodes. Under some assumptions, it was shown that the imaginary part of the IPSDM can be employed to reconstruct the exact topology of the network. Furthermore, a concentration bound on IPSDM estimation from a finite time-series is provided.

Chapter 3

Topology Identification under Spatially Correlated Noise

This chapter addresses the problem of reconstructing the topology of a network of agents interacting via linear dynamics, while being excited by exogenous stochastic sources that are possibly correlated across the agents, from time-series measurements alone. It is shown, under the assumption that the correlations are affine in nature, such network of nodal interactions is equivalent to a network with added agents. The added agents are represented by nodes that are latent, where no corresponding time-series measurements are available; however, here all the exogenous excitements are spatially (that is, across agents) uncorrelated. Generalizing affine correlations, it is shown that, under polynomial correlations, the latent nodes in the expanded networks can be excited by clusters of noise sources, where the clusters are uncorrelated with each other. The clusters can be replaced with a single noise source if the latent nodes are allowed to have non-linear interactions. Finally, using the sparse plus low-rank matrix decomposition of the imaginary part of the inverse power spectral density matrix (IPSDM) of the time-series data, the topology of the network is reconstructed. Under non conservative assumptions, the correlation graph of the noise sources is retrieved.

3.1 Introduction

Networks and graphical models provide convenient tools for effective representations of complex high dimensional multi-agent systems. Such a representation is useful in applications including power grids [12, 11], meteorology [13], neuroscience [43], and finance

[70]. Knowledge of the network interaction structure (also known as network topology) helps in understanding, predicting, and in many applications, controlling/identifying the system behavior [71, 72, 27, 16, 73, 15]. In applications such as power grid, finance, and meteorological system, it is either difficult or impossible to intervene and affect the system. Hence, inferring the network properties by passive means such as time-series measurements is of great interest in these applications. An example is unveiling the correlation structure between the stocks in the stock market from daily share prices [70], which is useful in predicting the market behavior.

Learning the conditional independence relation between variables from time-series measurements is an active research field among machine learning (ML), probabilistic graphical model (PGM), and statistics communities [8, 74, 75, 76, 77, 9], where the system modules are considered as random variables. However, such studies fail to capture dynamic dependencies between the entities in a system, which are prevalent in most of the aforementioned applications. For dynamical systems, autoregressive (AR) models that are excited by exogenous Gaussian noise sources, which are independent across time and variables, are explored in [35, 78, 79, 80]. Here, the graph topology captured the sparsity pattern of the regressor coefficient matrices, which characterized the conditional independence between the variables. It was shown that the sparsity pattern of the inverse power spectral density matrix (IPSDM), also known as concentration matrix, identifies the conditional independence relation between the variables (see [79, 80]). As shown in [17], the conditional independence structure between the variables is equivalent to the moral-graph structure of the underlying directed graph. For multi-agent systems with linear dynamical interaction, excited by wide sense stationary (WSS) noise sources that are mutually uncorrelated across the agents, moral graph reconstruction using Wiener projection has gained popularity in the last decade [81, 82, 17]. Here, the moral-graph of the underlying linear dynamic influence model (LDIM) is recovered from the magnitude response of Wiener coefficients, obtained from time-series measurements. For a wide range of applications, the spurious connections in the moral-graph can be identified—returning the true topology—by observing the phase response of the Wiener coefficients [18]. Furthermore, for systems with strictly causal dynamical dependencies, Granger causality based algorithms can unveil the exact cause-effect nature of the interactions; thus recovering the exact parent-child relationship of the underlying graph [19, 83, 17].

In many networks, time-series at only a subset of nodes is available. The nodes where the time-series measurements are not available form the latent/confound/hidden

nodes. Topology reconstruction is more challenging in the presence of latent nodes as additional spurious connections due to confounded effects are formed when applying the aforementioned techniques. AR model identification of the independent Gaussian time-series discussed above, in the presence of latent nodes, was studied in [38, 36, 37, 84, 85, 33, 32, 86]. Here, the primary goal is to eliminate the spurious connections due to latent nodes and retrieve the original conditional independence structure from the observed time-series. In applications such as power grid, there is a need to retrieve the complete topology of the network, including that of the latent nodes. Such problems are studied for bidirected tree [21], poly-forest [22], and poly-tree [25, 24] networks excited by WSS noise sources that are uncorrelated across the agents. Recently, an approach to reconstruct the complete topology of a general linear dynamical network with WSS noise was provided in [1]. The work in [1] can be considered as a generalization of AR model identification with latent nodes, in asymptotic time-series regime. However, one major caveat of the aforementioned literature on graphical models is that the results fail if the exogenous noise sources are spatially correlated, i.e. if the noise is correlated across the agents/variables. Prior works have studied the systems with spatially correlated noise sources [41, 27, 16]; however, these studies assume the knowledge of topology of the network. In a related work, [42] studied topology estimation under spatially correlated noise sources and used this estimated topology in local module identification.

This chapter studies the problem of topology identification of the LDIMs that allow spatially *correlated* noise sources, similar to the problem in [42]. However, this chapter provides an alternate treatment, where the noise correlations are transformed to latent nodes. This transformation enables one to gain additional insights and apply techniques from topology identification with latent nodes to solve the problem.

The first major result of this chapter is to provide a transformation that converts an LDIM without latent nodes, but excited by spatially correlated exogenous noise sources, to LDIMs with latent nodes. Here, the latent nodes are characterized by the maximal cliques in the correlation graph, the undirected graph that represents the spatial correlation structure. It is shown that, under affine correlation assumption, that is, when the correlated noise sources are related in affine way (Assumption 3.4.1), there exist transformed LDIMs with latent nodes, where all the nodes are excited by spatially uncorrelated noise sources. A key feature of the transformation is that the correlations are completely captured using the latent nodes, while the original topology remains unaltered. The original moral-graph/topology is the same as the moral-graph/topology among the observed nodes in the altered graph. Thus, this transformed problem is shown

to be equivalent to topology identification of networks with latent nodes. Consequently, any of the aforementioned techniques for the networks with latent node can be applied on the transformed LDIM to reconstruct the original moral graph/topology.

Next, relaxing the affine correlation assumption, polynomial correlation is considered (Assumption 3.5.1); here, the focus is on noise sources with distributions that are symmetric around the mean. It is shown that, in this scenario, the transformed dynamical model can be excited by clusters of noise sources, where the clusters are uncorrelated. However, the noise sources inside the clusters can be correlated. Using the sparse+low-rank decomposition of IPSDM from [1], the true topology of the network along with the correlation structure is reconstructed, from the IPSDM of the original LDIM, without any additional information, if the network satisfies a necessary and sufficient condition. Notice that, the results discussed here are applicable to the networks with static random variables and AR models also, when the exogenous noise sources are spatially correlated.

The chapter is organized as follows. Section 3.2 introduces the system model, including LDIM, and the essential definitions. Section 3.3 discusses IPSDM based topology reconstruction. Section 3.4 describes the correlation graph to latent nodes transformation and some major results. Section 3.5 discusses transformation of LDIMs with polynomial correlation to LDIMs with latent nodes. Section 3.6 explains the sparse+low-rank decomposition technique and how the topology can be reconstructed without the knowledge of correlation graph. Simulation results are provided in Section 3.7. Finally, Section 2.8 concludes the chapter.

Notations: Bold capital letters, $\mathbf{A}$, denotes matrices; A_{ij} and $(\mathbf{A})_{ij}$ represents $(i, j)^{\text{th}}$ entry of $\mathbf{A}$; Bold small letters, $\mathbf{v}$, denotes vectors; v_i indicates i^{th} entry of $\mathbf{v}$; $[n] := \{1, \dots, n\}$; The subscript o denotes the observed nodes index set $[n]$ and the subscript h indicates the latent nodes index set $\{n+1, \dots, n+L\}$, For example, $\mathbf{x}_o := \{x_1, \dots, x_n\}$ and $\mathbf{x}_h := \{x_{n+1}, \dots, x_{n+L}\}$; $\Phi_{\mathbf{x}}(z) := [(\Phi_{\mathbf{x}}(z))_{ij}]$, $z \in \mathbb{C}$, $|z| = 1$ denote power spectral density matrix, where $(\Phi_{\mathbf{x}})_{ij}$ is cross power spectral density between the time-series at nodes i and j ; $\mathcal{F}$ denotes the set of real rational single input single output transfer functions that are analytic on unit circle; $\mathcal{A}$ denotes the set of rationally related zero mean jointly wide sense stationary (JWSS) scalar stochastic process; $\mathcal{Z}(\cdot)$ denotes bilateral z -transform; $\mathbb{S}^n$ denotes the space of all skew symmetric $n \times n$ matrices; $\mathbb{N}$ denotes the set of natural numbers, $\{0, 1, 2, \dots\}$; For a set $\mathcal{S}$, $|\mathcal{S}|$ denotes the cardinality of the set. $\|\mathbf{A}\|_*$ denotes the nuclear norm and $\|\mathbf{A}\|_1$ denote sum of absolute values of entries of $\mathbf{A}$. $\angle a$ denotes the phase of the complex number a .

3.2 System model

Consider a network of n interconnected nodes, where node i is equipped with time-series measurements, $(\tilde{x}_i(t))_{t \in \mathbb{Z}}$, $i \in [n]$. The network interaction is described by,

$$\mathbf{x}(z) = \mathbf{H}(z)\mathbf{x}(z) + \mathbf{e}(z), \quad (3.1)$$

where $\mathbf{x}(z) = [x_1(z), x_2(z), \dots, x_n(z)]$, $\mathbf{x}_i(z) = \mathcal{Z}[\tilde{\mathbf{x}}_i]$. $\mathbf{e}(z) = [e_1(z), e_2(z), \dots, e_n(z)] \in \mathcal{A}^n$ are the exogenous noise sources with $\Phi_e(z)$ is non-singular, but possibly non-diagonal. $\mathbf{H}(z)$ is the weighted adjacency matrix with $H_{ii} = 0$, for all $i \in [n]$. An LDIM is defined as the pair $(\mathbf{H}, \mathbf{e})$, whose output process is given by (3.1). The LDIM is well-posed if every entry of $(\mathbf{I}_n - \mathbf{H}(z))^{-1}$ is analytic on the unit circle, $|z| = 1$ and topologically detectable if Φ_e is positive definite.

Every LDIM has two associated graphs, viz: **1**) a directed graph, the linear dynamic influence graph (LDIG), $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = [n]$ and $\mathcal{E} := \{(i, j) : H_{ji} \neq 0\}$ and **2**) an undirected graph, the correlation graph, $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$, where $\mathcal{E}_c := \{(i, j) : (\Phi_e)_{ij} \neq 0, i \neq j\}$. Notice that, in a directed graph, if $(i, j) \in \mathcal{E}$ then there is a directed arrow from i to j in the graphical representation of $\mathcal{E}$ (see Figure 3.1(a) for example). Concisely, “LDIG $(\mathbf{H}, \mathbf{e})$ ” is used to denote the LDIG $\mathcal{G}(\mathcal{V}, \mathcal{E})$ corresponding to the LDIM $(\mathbf{H}, \mathbf{e})$. For a directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, the parent, the children, and the spouse sets of node i are $Pa(i) := \{j : (j, i) \in \mathcal{E}\}$, $Ch(i) := \{j : (i, j) \in \mathcal{E}\}$, and $Sp(i) := \{j : j \in Pa(Ch(i))\}$ respectively. The topology of a directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, denoted $\mathcal{T}(\mathcal{V}, \mathcal{E})$, is the undirected graph obtained by removing the directions from every edge $(i, j) \in \mathcal{E}$. Moral-graph (also called kin-graph) of $\mathcal{G}(\mathcal{V}, \mathcal{E})$, denoted $kin(\mathcal{G})$ is an undirected graph, $kin(\mathcal{G}) := \mathcal{G}(\mathcal{V}, \mathcal{E}')$, where $\mathcal{E}' := \{(i, j) : i \neq j, i \in (Sp(j) \cup Pa(j) \cup Ch(j))\}$. A clique is a sub-graph of a given undirected graph where every pair of nodes in the sub-graph are adjacent. A maximal clique is a clique that is not a subset of a larger clique. For a correlation graph, $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$, q is the number of maximal cliques with clique size > 1 . The problem we address is described as follows.

Problem 1. (P1) Consider a well-posed and topologically detectable LDIM, $(\mathbf{H}, \mathbf{e})$, where Φ_e is allowed to be non-diagonal and its associated graphs $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ are unknown. Given the power spectral density matrix $\Phi_{\mathbf{x}}$, where $\mathbf{x}$ is given by (3.1), reconstruct the topology of $\mathcal{G}$.

3.3 IPSDM based topology reconstruction

In this section, the IPSDM based topology reconstruction is presented. In [17], the authors showed that, for any LDIM characterized by (3.1), the availability of IPSDM, which can be written as

$$\Phi_{\mathbf{x}}^{-1} = (\mathbf{I}_n - \mathbf{H})^* \Phi_e^{-1} (\mathbf{I}_n - \mathbf{H}), \quad (3.2)$$

is sufficient for reconstructing the moral-graph of $(\mathbf{H}, \mathbf{e})$. That is, if $(\Phi_{\mathbf{x}}^{-1})_{ij} \neq 0$, then i and j are kins. However, an important assumption for the result to hold true is that Φ_e^{-1} is diagonal. If Φ_e^{-1} is non-diagonal, then the result does not hold in general. For $i \neq j$,

$$\begin{aligned} (\Phi_{\mathbf{x}}^{-1})_{ij} &= (\Phi_e^{-1})_{ij} - (\mathbf{H}^* \Phi_e^{-1})_{ij} - (\Phi_e^{-1} \mathbf{H})_{ij} + (\mathbf{H}^* \Phi_e^{-1} \mathbf{H})_{ij} \\ &= (\Phi_e^{-1})_{ij} - \sum_{k=1}^n (\mathbf{H}^*)_{ik} (\Phi_e^{-1})_{kj} - \sum_{k=1}^n (\Phi_e^{-1})_{ik} \mathbf{H}_{kj} \\ &\quad + \sum_{k=1}^n \sum_{l=1}^n (\mathbf{H}^*)_{ik} (\Phi_e^{-1})_{kl} \mathbf{H}_{lj}, \end{aligned}$$

and any of the four terms can cause $(\Phi_{\mathbf{x}}^{-1})_{ij} \neq 0$, depending on Φ_e^{-1} . Hence, this technique cannot be applied directly to solve Problem 1. For example, consider the network $\mathcal{G}$ given in Figure 3.1(a) with $(\Phi_e^{-1})_{13} \neq 0$. Then, $(\Phi_{\mathbf{x}}^{-1})_{23} \neq 0$ since $\mathbf{H}_{21}^* (\Phi_e^{-1})_{13} \neq 0$, which implies that the estimated topology has $(2, 3)$ present, while $(2, 3) \notin \text{kin}(\mathcal{G})$.

In the next section, the correlation graph, $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$, is scrutinized and properties of $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ are evaluated as the first step in unveiling the true topology when noise sources admit correlations.

3.4 Spatial Correlation to Latent Node Transformation

In this section, a transformation of the LDIM, $(\mathbf{H}, \mathbf{e})$, to an LDIM with latent nodes by exploiting the structural properties of the noise correlation is obtained. The transformation converts an LDIM without latent nodes and driven by spatially *correlated* exogenous noise sources to an LDIM with latent nodes that are excited by spatially *uncorrelated* exogenous noise sources. Assuming perfect knowledge of the noise correlation

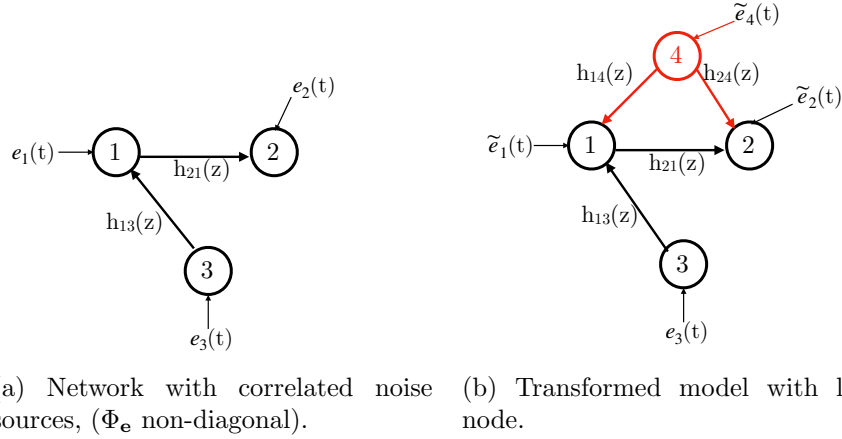


Figure 3.1: Transformation of an LDIM with correlated noise sources to a network with uncorrelated noise sources in the presence of a latent node.

structure, the latent nodes and their children in the transformed LDIM are characterized. It is shown that although the transformation is not unique, the topology of the transformation is unique under affine correlation (Assumption 3.4.1).

For $\tilde{\mathbf{e}} = [\tilde{\mathbf{e}}_o; \tilde{\mathbf{e}}_h]$ and $\tilde{\mathbf{H}} := \begin{bmatrix} \mathbf{H} & \mathbf{F} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, $\mathbf{H} \in \mathcal{F}^{n \times n}$, $\mathbf{F} \in \mathcal{F}^{n \times L}$, $L \in \mathbb{N}$, $(\tilde{\mathbf{H}}, \tilde{\mathbf{e}})$ (or with a slight abuse of notation, $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}})$) denotes an LDIM with L latent nodes, where $\mathbf{F}_{ik}$ denotes the directed edge weight from latent node k to observed node i and $\mathbf{H}_{ik}$ denotes the directed edge weight from observed node k to observed node i . Notice that the latent nodes considered in this chapter are strict parents and they do not have incoming edges.

3.4.1 Relation between Spatial Correlation and Latent Nodes

Here, it is demonstrated that the spatially correlated exogenous noise sources can be viewed as the children of a latent node that is a common parent of the correlated sources. The idea is explained in the motivating example below. Towards this, let us define the following notion of equivalent networks.

Definition 3.4.1. Let $(\mathbf{H}^{(1)}, \mathbf{e}^{(1)})$ and $(\mathbf{H}^{(2)}, \mathbf{e}^{(2)})$ be two LDIMs and let $\mathbf{x}^{(1)} = (\mathbf{I}_n - \mathbf{H}^{(1)})^{-1} \mathbf{e}^{(1)}$ and $\mathbf{x}^{(2)} = (\mathbf{I}_n - \mathbf{H}^{(2)})^{-1} \mathbf{e}^{(2)}$ respectively. Then, the LDIMs $(\mathbf{H}^{(1)}, \mathbf{e}^{(1)})$ and $(\mathbf{H}^{(2)}, \mathbf{e}^{(2)})$ are said to be equivalent, denoted $(\mathbf{H}^{(1)}, \mathbf{e}^{(1)}) \equiv (\mathbf{H}^{(2)}, \mathbf{e}^{(2)})$ if and only if $\Phi_{\mathbf{x}^{(1)}} = \Phi_{\mathbf{x}^{(2)}}$.

Consider the network of three nodes in Figure 3.1(a). Suppose the noise sources e_1 and e_2 are correlated, that is, $\Phi_{e_1 e_2} \neq 0$. Consider the network in Figure 3.1(b), where $\tilde{e}_1, \tilde{e}_2, \tilde{e}_4$, and e_3 are jointly uncorrelated and node 4 is a latent node. Note that e_3 is same in both the LDIMs while $\tilde{e}_1, \tilde{e}_2$ are different from e_1, e_2 . From the LDIM of Figure 3.1(a), it follows that

$$x_1 = e_1 + h_{13}x_3 \text{ \& } x_2 = e_2 + h_{21}x_1. \quad (3.3)$$

From the LDIM of Figure 3.1(b), it follows that

$$\begin{aligned} x_1 &= \tilde{e}_1 + h_{14}x_4 + h_{13}x_3, \\ x_2 &= \tilde{e}_2 + h_{24}x_4 + h_{21}x_1, \text{ \& } x_4 = \tilde{e}_4. \end{aligned} \quad (3.4)$$

Comparing (3.3) and (3.4), it is evident that if $\tilde{e}_1, \tilde{e}_2$, and $\tilde{e}_4$ are such that $e_1 = \tilde{e}_1 + h_{14}\tilde{e}_4$ and $e_2 = \tilde{e}_2 + h_{24}\tilde{e}_4$, then the time-series obtained from both the LDIMs are identical. That is, the LDIMs are equivalent.

For a given LDIM, $(\mathbf{H}, \mathbf{e})$, with correlated exogenous noise sources, one can define a space of equivalent LDIMs with uncorrelated noise sources that provide the same time-series. The following definition captures this space of “LDIMs with latent nodes and uncorrelated noise sources” that are equivalent to the original LDIM with correlated noise sources.

Assumption 3.4.1. *In LDIM $(\mathbf{H}, \mathbf{e})$, defined by (3.1), the exogenous noise processes $e_i, e_j \in \mathcal{A}$ are correlated only via affine interactions, i.e., $\Phi_{e_i e_j} \neq 0$ if and only if there exists an affine transform, $f(x) = a + bx$, $a \in \mathcal{A}$, $b \in \mathcal{F}$, $b \neq 0$ such that either $e_i = f(e_j)$ or $e_j = f(e_i)$.*

Definition 3.4.2. *For any LDIM, $(\mathbf{H}, \mathbf{e})$, with $\Phi_{\mathbf{e}}$ non-diagonal, and satisfying Assumption 3.4.1,*

$$\begin{aligned} \mathcal{L}(\mathbf{H}, \mathbf{e}) &:= \{(\tilde{\mathbf{H}}, \tilde{\mathbf{e}}) : \mathbf{e} = \tilde{\mathbf{e}}_o + \mathbf{F}\tilde{\mathbf{e}}_h, \tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H} & \mathbf{F} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tilde{\mathbf{e}} = [\tilde{\mathbf{e}}_o, \tilde{\mathbf{e}}_h]^T, \\ &\quad \tilde{\mathbf{e}} \in \mathcal{A}^{n+L}, \mathbf{F} \in \mathcal{F}^{n \times L}, L \in \mathbb{N}, \Phi_{\tilde{\mathbf{e}}} \text{ diagonal}\}, \end{aligned} \quad (3.5)$$

is the space of all $\mathcal{L}$ -transformations for a given $(\mathbf{H}, \mathbf{e})$.

Remark 3.4.1. *In (3.4.2), the number of latent nodes, L , is not fixed a priori.*

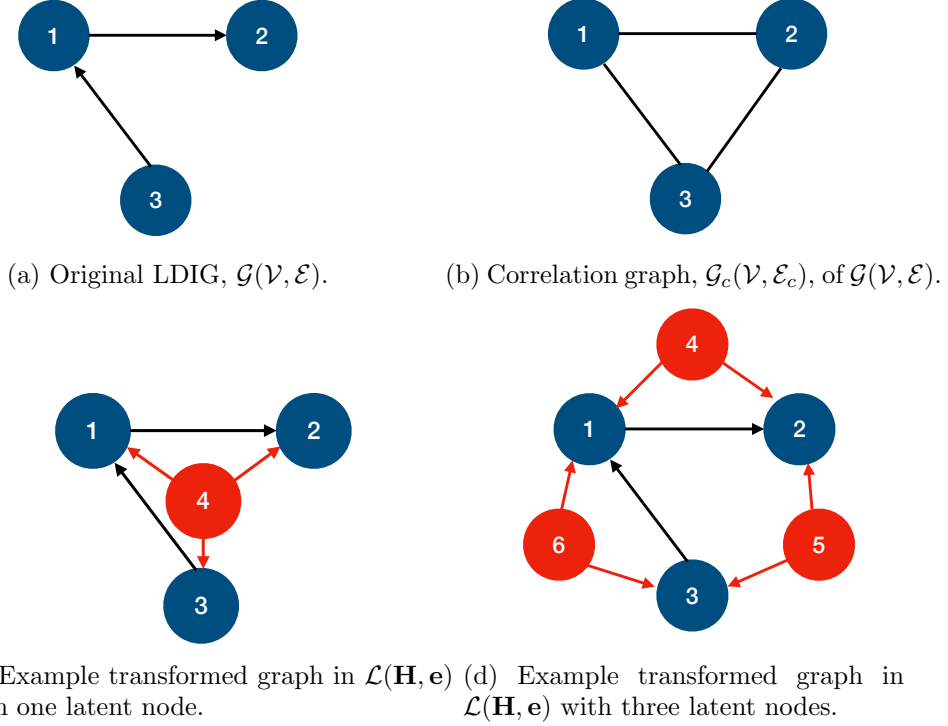


Figure 3.2: Demonstration of $\mathcal{L}(\mathbf{H}, \mathbf{e})$; noise sources are not shown. The exogenous noises are uncorrelated in Figure 3.2(c) and 3.2(d).

Based on the definition of $\mathcal{L}(\mathbf{H}, \mathbf{e})$ in (3.5), LDIM for $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}})$ can be rewritten as

$$\mathbf{x}(z) = \mathbf{H}(z)\mathbf{x}(z) + \mathbf{\Gamma}(z)\tilde{\mathbf{e}}(z), \quad (3.6)$$

where $\tilde{\mathbf{e}}(z) = [\tilde{e}_1(z), \dots, \tilde{e}_n(z), \dots, \tilde{e}_{n+L}(z)]$ are mutually uncorrelated and $\mathbf{\Gamma} = [\mathbf{I}_n, \mathbf{F}]$.

An LDIM, $(\tilde{\mathbf{H}}, \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e})$, is called a *transformed* LDIM of $(\mathbf{H}, \mathbf{e})$ in this chapter. Intuitively, $\mathcal{L}$ -transformation completely assigns the spatial correlation component present in the original noise source, $\mathbf{e}$, to latent nodes, without altering the LDIG, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, similar to the discussion on the aforementioned motivating example. For the LDIM in Figure 3.2(a) with correlation graph in Figure 3.2(b), Figure 3.2(c), and Figure 3.2(d) show some examples of the LDIGs that belong to $\mathcal{L}(\mathbf{H}, \mathbf{e})$. Thus, $\mathcal{L}$ -transformation returns a larger network with latent nodes. Notice that the latent nodes present in an LDIG of $(\tilde{\mathbf{H}}, \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e})$, are *strict parents*, whose interaction with the true nodes are characterized by $\mathbf{F}$.

The following theorem formalizes the relation between the correlation graph and the latent nodes in the transformed higher dimensional LDIMs. In particular, it shows that two noise sources e_i and e_j are correlated if and only if there is a latent node, which is a common parent of both the nodes i and j in the transformed LDIM.

Lemma 3.4.1. *Let $(\mathbf{H}, \mathbf{e})$ be an LDIM defined by (3.1) that satisfies Assumption 3.4.1, and let $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ be the correlation graph of $\{e_i\}_{i=1}^n$. Then, for every distinct $i, j \in [n]$, $(i, j) \in \mathcal{E}_c$ if and only if every LDIG $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e})$ contains a latent node h such that $h \in Pa(i) \cap Pa(j)$.*

Proof. Refer Appendix B.1 □

Thus, e_i and e_j in the original LDIM are correlated if and only if for every $\mathcal{L}$ -transformed LDIM there exists a k such that $F_{ik} \neq 0$ and $F_{jk} \neq 0$.

The following lemma shows that the number of latent nodes, L , present in any $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e})$ is at least q , the number of maximal cliques with clique size greater than one in $\mathcal{G}_c$.

Lemma 3.4.2. *Let $(\mathbf{H}, \mathbf{e})$ be an LDIM and let $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ be the correlation graph of the exogenous noise sources, $\mathbf{e}$. Then, the number, L , of latent nodes present in any $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e})$ satisfies $L \geq q$, where q is the number of maximal cliques with clique size > 1 in $\mathcal{G}_c$.*

Proof. Refer Appendix B.2. □

Remark 3.4.2. *Let $\mathcal{G}_c(\mathcal{V}_c, \mathcal{E}_c)$ be the correlation graph of the exogenous noise sources. Then, for any given maximal clique $\mathcal{G}^\ell(\mathcal{V}^\ell, \mathcal{E}^\ell) \subseteq \mathcal{G}_c(\mathcal{V}_c, \mathcal{E}_c)$ and for any $k_\ell \geq 1$, there exists a transformed LDIG $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e})$ with k_ℓ number of latent nodes such that the set $\mathcal{V}^\ell$ is equal to the children of the latent nodes. For $|\mathcal{V}^\ell| = 3$, Figure 3.2 shows examples with $k_\ell = 1, 3$ (see proof of Lemma 3.4.2 for details).*

As shown in Lemma 3.4.2 and Figure 3.2, the LDIMs in the characterizing space $\mathcal{L}(\mathbf{H}, \mathbf{e})$ of the equivalent LDIMs can have an arbitrary number of latent nodes, which leads to multiple transformed representations with varying number of latent nodes. The following definition restricts the number of latent nodes present in the equivalent transformed LDIGs and provides a minimal transformed representation—representation with minimal number of nodes—of the true LDIG.

Definition 3.4.3. *Define $\mathcal{L}_q(\mathbf{H}, \mathbf{e})$ to be the set of all LDIMs in $\mathcal{L}(\mathbf{H}, \mathbf{e})$ with the number of latent nodes equal to the number of maximal cliques with clique size > 1 present in $\mathcal{G}_c$.*

The following lemma shows that there exist a unique latent node in every transformed LDIG $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}_q(\mathbf{H}, \mathbf{e})$, corresponding to each maximal clique.

Theorem 3.4.1. *Let $(\mathbf{H}, \mathbf{e})$ be an LDIM that satisfies Assumption 3.4.1 and let $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ be the correlation graph of the exogenous noise sources, e. Suppose $\mathcal{G}_c$ has q maximal cliques $C_1, \dots, C_q$. Consider any LDIG $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}_q(\mathbf{H}, \mathbf{e})$ with q latent nodes, $h_1, \dots, h_q$. Then, for a maximal clique C_i in $\mathcal{G}_c$, there exists a unique latent node $h \in \{h_1, \dots, h_q\}$ such that $Ch(h) = C_i$.*

Proof. Refer Appendix B.3. □

Remark 3.4.3. *Notice that the existence of the unique latent node is true only for the space $\mathcal{L}_q(\mathbf{H}, \mathbf{e})$. Such a unique node might not exist for the transformed representations in $\mathcal{L}(\mathbf{H}, \mathbf{e})$. In other words, Theorem 4 identifies the minimal set of latent nodes required to explain the data.*

3.4.2 Uniqueness of the topology

Here, the LDIMs in $\mathcal{L}_q(\mathbf{H}, \mathbf{e})$ from the previous section is studied more carefully. In topology reconstruction, only the support structure of the transfer functions matter. The following proposition shows that the support structure is unique.

Proposition 3.4.1. *Let $(\mathbf{H}, \mathbf{e})$ be an LDIM that satisfies Assumption 3.4.1. The topology of every LDIG $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}_q(\mathbf{H}, \mathbf{e})$ is the same.*

Proof. Refer Appendix B.4. □

Based on the above results, the following transformation of Problem (P1) is formulated.

Problem 2. (P2) *Consider an LDIM $(\mathbf{H}, \mathbf{e})$ defined by (3.1) that satisfies Assumption 3.4.1, and $\Phi_{\mathbf{e}}$ allowed to be non-diagonal. Let $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}_q(\mathbf{H}, \mathbf{e})$ be an LDIM with LDIG given by $\mathcal{G}_t(\mathcal{V}_t, \mathcal{E}_t)$. Suppose that the time-series data among the observed nodes of $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}})$ is given. Then, reconstruct the topology among the observed nodes of $\mathcal{G}_t$.*

Remark 3.4.4. *The time series data among the observed nodes of the transformed LDIM is the same as the time series obtained from the original LDIM.*

Remark 3.4.5. *The problems (P1) and (P2) are equivalent. That is, the topology reconstructed in both the problems are the same, because topology among the observed nodes of any element of $\mathcal{L}(\mathbf{H}, \mathbf{e})$ is same as the topology of $(\mathbf{H}, \mathbf{e})$ (see Definition 3.4.2 and proof of Proposition 3.4.1).*

Therefore, instead of reconstructing the topology of $(\mathbf{H}, \mathbf{e})$ with Φ_e non-diagonal, it is sufficient to reconstruct the topology among the observed nodes for one of the LDIMs $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}_q(\mathbf{H}, \mathbf{e})$, which has $\Phi_{\tilde{\mathbf{e}}}$ diagonal.

3.5 Polynomial Correlation

The results in the previous section assumed that the correlations between the exogenous noise sources are affine in nature. In this section, a generalization of the affine correlation is addressed. It is shown that the noise sources that are correlated via non-affine, but a polynomial, interaction can be characterized using latent nodes with non-linear interaction dynamics. By lifting the processes to a higher dimension, the non-linearity is converted to linear interactions.

The following definitions are useful for the presentation in this section.

Definition 3.5.1. [62] Let $x = (x_1, \dots, x_m)$. A monomial in $x_1, \dots, x_m$ is a product of the form $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \dots x_m^{\alpha_m}$, where $\alpha \in \mathbb{N}^m$. A polynomial P in x with coefficient in a field (or a commutative ring), $\mathbb{F}$, is a finite linear combination of monomials, i.e.,

$$P(x) := \sum_{\alpha} a_{\alpha} x^{\alpha}, \quad a_{\alpha} \in \mathbb{F}.$$

- (i) The degree of a monomial x^{α} is $|\alpha| := \sum_{i=1}^m \alpha_i$.
- (ii) The total degree of $P \neq 0$ is the maximum of $|\alpha|$ such that $a_{\alpha} \neq 0$.

Definition 3.5.2. For $\mathbf{v} \in \mathcal{A}^m$, define the list of monomials of total degree at most p ,

$$\mathcal{M}(\mathbf{v}, p) := [f_0(\mathbf{v}), f_1(\mathbf{v}), \dots, f_p(\mathbf{v})]^T,$$

where $f_k(\mathbf{v})$ lists all the k -degree monomials, $f_k(\mathbf{v}) := \{\mathbf{v}_1^{\alpha_1} \mathbf{v}_2^{\alpha_2} \dots \mathbf{v}_m^{\alpha_m} \mid \sum_{i=1}^m \alpha_i = k\}$. For example, when $m = 2$, $f_0(\mathbf{v}) = 1$, $f_1(\mathbf{v}) = \{\mathbf{v}_1, \mathbf{v}_2\}$, and $f_2(\mathbf{v}) = \{\mathbf{v}_1^2, \mathbf{v}_1 \mathbf{v}_2, \mathbf{v}_2^2\}$. In general, the total number of monomials having degree k is given by $\binom{m+k-1}{k}$, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. The total number of monomials up to total degree p is

$$M = \sum_{k=0}^p \binom{m+k-1}{k}.$$

For $m = 2$ and $p = 3$,

$$\mathcal{M}(\mathbf{v}, 3) = [1, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_1^2, \mathbf{v}_1\mathbf{v}_2, \mathbf{v}_2^2, \mathbf{v}_1^3, \mathbf{v}_1^2\mathbf{v}_2, \mathbf{v}_1\mathbf{v}_2^2, \mathbf{v}_2^3]^T.$$

3.5.1 Characterization of Polynomial Correlations

Suppose $e_1 = \sum_{|\alpha| \leq M} a_{\alpha,1} \mathbf{v}^\alpha$ and $e_2 = \sum_{|\alpha| \leq M} a_{\alpha,2} \mathbf{v}^\alpha$, where $a_{\alpha,1}, a_{\alpha,2} \in \mathcal{F}$ and $\mathbf{v} \in \mathcal{A}^m$, $m, M \in \mathbb{N}$, with $\Phi_{\mathbf{v}}$ diagonal. Let $\mathbf{y} = \mathcal{M}(\mathbf{v}, p)$ be the vector obtained by concatenating $\mathbf{v}^\alpha$, $|\alpha| \leq p$. Letting $e_1 = \mathbf{A}_1 \mathbf{y}$ and $e_2 = \mathbf{A}_2 \mathbf{y}$, we have $\Phi_{e_1 e_2} = \mathbf{A}_1 \Phi_{\mathbf{y}} \mathbf{A}_2^*$, where $\mathbf{A}_i$ is the vector obtained by concatenating $a_{\alpha,i}$.

Notice that $\mathbf{y}$ is lifting of $\mathbf{v}$ into a higher dimensional space of polynomials. In the following, a discussion on the structure of $\Phi_{\mathbf{y}}$ is provided, with the help of examples.

3.5.2 Example: Lifting of a zero mean Gaussian Process

To illustrate lifting of noise processes to higher dimension, consider independent and identically distributed (IID) Gaussian process (GP). It is shown that, under lifting, the power spectral density matrix (PSDM) is block diagonal.

Consider an IID, GP $\{\mathbf{v}(k), k \in \mathbb{Z} \mid \mathbf{v}(k) \sim N(0, \sigma^2 \mathbf{I}_m)\}$. Then, $\mathbb{E}\{\mathbf{v}_i \mathbf{v}_j\} = \mathbb{E}\{\mathbf{v}_i\} \mathbb{E}\{\mathbf{v}_j\} = 0$. It can be shown that [87]

$$\mathbb{E}\{\mathbf{v}_i^p\} = \begin{cases} 0 & \text{if } p \text{ is odd,} \\ \sigma^p (p-1)!! & \text{if } p \text{ is even,} \end{cases} \quad (3.7)$$

where $p!! := (p-1)(p-3) \dots 3 \cdot 1$ denotes the double factorial. Consider $m = 2$ and let $\mathbf{y} = [y_1, \dots, y_{10}] := \mathcal{M}(\mathbf{v}, 3) = [1, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_1^2, \mathbf{v}_1\mathbf{v}_2, \mathbf{v}_2^2, \mathbf{v}_1^3, \mathbf{v}_1^2\mathbf{v}_2, \mathbf{v}_1\mathbf{v}_2^2, \mathbf{v}_2^3]^T$. That is, $\mathbf{y}$ lists all the monomials of $\mathbf{v}_1, \mathbf{v}_2$ with degree ≤ 3 .

Notice that $\mathbb{E}\{\mathbf{v}_1^i \mathbf{v}_2^j\} = \mathbb{E}\{\mathbf{v}_1^i\} \mathbb{E}\{\mathbf{v}_2^j\} \neq 0$ if and only if both i and j are even. Then, $\mathbb{E}\{y_2 y_5\} = \mathbb{E}\{\mathbf{v}_1^2 \mathbf{v}_2\} = 0$. Straight forward computation shows that $\mathbb{E}\{y_2 y_k\} \neq 0$ only for $k = 2, 7, 9$. Similarly, $\mathbb{E}\{y_7 y_k\} \neq 0$ if and only if $k = 7, 9, 2$, and $\mathbb{E}\{y_9 y_k\} \neq 0$ if and only if $k = 7, 9, 2$. Notice that $k = 2, 7, 9$ corresponds to the terms with odd power on x_1 and even power on x_2 . Repeating the same for every $\mathbb{E}\{y_i y_k\}$, $1 \leq i, k \leq 10$, one can show that, after appropriate rearrangement of rows and columns, the covariance matrix and the PSDM of $\mathbf{y}$ forms a block diagonal matrix, given by (3.8). Here, $\tilde{\mathbf{y}} =$

$$[y_1, y_4, y_6, y_2, y_7, y_9, y_3, y_8, y_{10}, y_5].$$

$$\Phi_{\tilde{\mathbf{y}}} = \begin{bmatrix} \Phi_{11} & \Phi_{14} & \Phi_{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \Phi_{41} & \Phi_{44} & \Phi_{46} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \Phi_{61} & \Phi_{46} & \Phi_{66} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \Phi_{22} & \Phi_{27} & \Phi_{29} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \Phi_{72} & \Phi_{77} & \Phi_{79} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \Phi_{92} & \Phi_{97} & \Phi_{99} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \Phi_{55} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \Phi_{33} & \Phi_{38} & \Phi_{310} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \Phi_{83} & \Phi_{88} & \Phi_{810} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \Phi_{103} & \Phi_{108} & \Phi_{1010} \end{bmatrix} \quad (3.8)$$

It is worth mentioning that since the Gaussian process is zero mean IID, the auto-correlation function $R_{\mathbf{v}_i \mathbf{v}_j}(k) = R_{\mathbf{v}_i \mathbf{v}_j}(0)\delta(k) = \mathbb{E}\{\mathbf{v}_i(0)\mathbf{v}_j(0)\}\delta(k)$, where δ is Kronecker delta and thus $\Phi_{\tilde{\mathbf{y}}}(z) = R_{\tilde{\mathbf{y}}}(0)$, $\forall |z| = 1$ is white (same for all the frequencies). The following proposition shows this for general m, p .

Proposition 3.5.1. *Consider the Gaussian IID process $\{\mathbf{v}(k), k \in \mathbb{Z} \mid \mathbf{v}(k) \sim N(0, \sigma^2 \mathbf{I}_m)\}$. Let $\mathbf{y} = \mathcal{M}(\mathbf{v}, p)$. Then, there exists a $\tilde{\mathbf{y}}$, a permutation of $\mathbf{y}$, such that $\Phi_{\tilde{\mathbf{y}}}$ is block diagonal with 2^m non-zero blocks.*

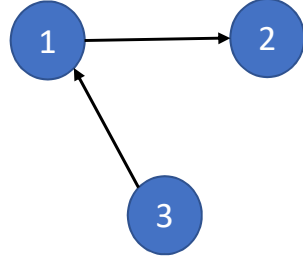
Proof: Refer Appendix B.5. □

Based on this example, one can extend the result to symmetric zero mean WSS processes. Symmetric distributions are the distributions with probability density functions symmetric around mean, for example, Gaussian distribution.

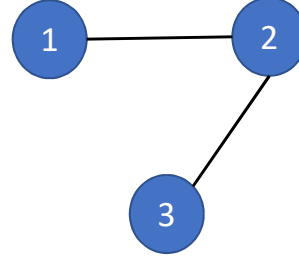
Definition 3.5.3. *A probability distribution is said to be symmetric around mean if and only if its density function, f , satisfies $f(\mu - x) = f(\mu + x)$ for every $x \in \mathbb{R}$, where μ is the mean of the distribution.*

Proposition 3.5.2. *Consider a zero mean WSS process, $\mathbf{v} \in \mathcal{A}^m$ with symmetric distribution and $\Phi_{\mathbf{v}}$ diagonal. Let $\mathbf{y} = \mathcal{M}(\mathbf{v}, p)$. Then, there exists a $\tilde{\mathbf{y}}$, a permutation of $\mathbf{y}$ such that $\Phi_{\tilde{\mathbf{y}}}$ is block diagonal with 2^m non-zero blocks.*

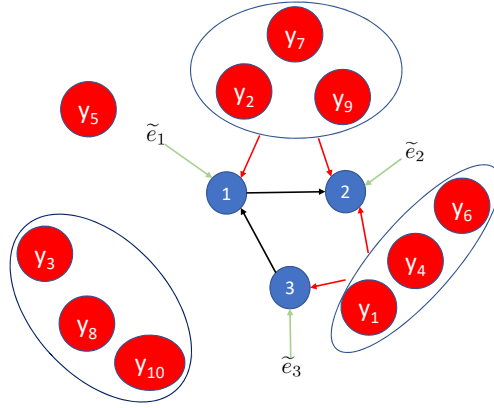
Proof. Proof is similar to the IID GP case. For symmetric distributions, odd moments are zero, since odd moments are the integral of odd functions. □



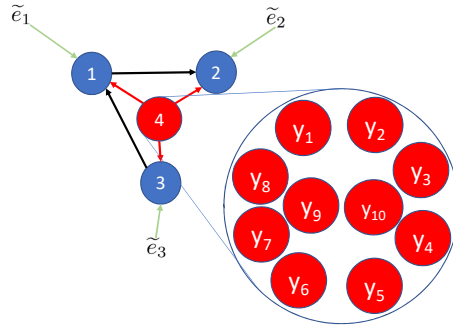
(a) Original LDIG, $\mathcal{G}(\mathcal{V}, \mathcal{E})$



(b) A correlation graph of $\mathcal{G}(\mathcal{V}, \mathcal{E})$



(c) An example transformed LDIG of sub figures above with $m = 2$, $p = 3$, and $n = 3$ with $2^m = 4$ clusters. Here, $e_1 = \tilde{e}_1 + F_{12}y_2 + F_{17}y_7 + F_{19}y_9$, $e_2 = \tilde{e}_2 + F_{22}y_2 + F_{27}y_7 + F_{29}y_9 + F_{21}y_1 + F_{24}y_4 + F_{26}y_6$, and $e_3 = \tilde{e}_3 + F_{31}y_1 + F_{34}y_4 + F_{36}y_6$.



(d) Transformed graph of the above LDIG when all the three nodes are correlated. The latent node 4 will capture the interactions due to the nodes $y_1, \dots, y_{10}$.

Figure 3.3: Transformation of an LDIM with correlated noise sources to a network with uncorrelated noise sources in the presence of a latent node.

As shown in the proof of Proposition 3.5.1, the monomial nodes corresponding to a given block diagonal (i.e., the same odd-even pattern) can be grouped into a cluster. Figure 3.3(c) shows such a clustering with $m = 2$ and $p = 3$. The red nodes are the lifted processes in the higher dimension (the monomial nodes $y_1, \dots, y_{10}$). The nodes inside a given cluster (shown by blue ellipse) are correlated with each other and the nodes from different clusters are not correlated with each other; this is a result of $\Phi_{\tilde{\mathbf{y}}}$ being block-diagonal. Here, e_1 and e_2 are correlated via the cluster of y_2, y_7 , and y_9 , whereas e_2 and e_3 are correlated via the cluster of y_1, y_4 , and y_6 . Notice the following important distinction from the LDIMs in Figure 3.1. Here, it is sufficient for the observed nodes 1, 2 to be connected to one of the latent nodes in a given cluster. It is not necessary for the nodes to have a common parent like the case in Figure 3.1. The common parent property from Section 3.4 is replaced with a common ancestral cluster here.

3.5.3 Transformation of Polynomial Correlation to Latent Nodes

Based on the aforementioned discussion, relaxation of Assumption 3.4.1 and extension of the LDIM transformation results from Section 3.4 is presented here. It is shown that in order to relax Assumption 3.4.1, non-linear interactions are required between the latent nodes and the observed nodes. The following assumption is a generalization of Assumption 3.4.1.

Assumption 3.5.1. *In LDIM $(\mathbf{H}, \mathbf{e})$, defined by (3.1), the noise processes $e_i, e_j \in \mathcal{A}$ are correlated if and only if there exist polynomials P_1, P_2 and $\tilde{\mathbf{e}} = [\mathbf{v}^T, \tilde{e}_1, \tilde{e}_2]^T$, $\mathbf{v} \in \mathcal{A}^m$, $e_1, e_2 \in \mathcal{A}$, with $\Phi_{\tilde{\mathbf{e}}}$ diagonal, $m \in \mathbb{N}$, such that $e_1 = \tilde{e}_1 + P_1(\mathbf{v})$ and $e_2 = \tilde{e}_2 + P_2(\mathbf{v})$, where $P_i(\mathbf{v}) = \sum_{|\alpha| \leq M} a_{\alpha,i} \mathbf{v}^\alpha$, $a_{\alpha,i} \in \mathcal{F}$, for some $M \in \mathbb{N}$.*

Remark 3.5.1. *To be precise, the extension of Assumption 3.4.1 is given by $y = P(x)$, where $x \in \mathcal{A}$ and $P(x) = \sum_{\alpha} a_{\alpha} x^{\alpha}$, $a_{\alpha} \in \mathcal{F}$ is a polynomial of degree less than or equal to M . Any $x \in \mathcal{A}$ can be written as $x = \tilde{e}_x + bv$, where $\tilde{e}_x, v \in \mathcal{A}$, $b \in \mathcal{F}$, and $\tilde{e}_x$ uncorrelated with v . Then, $y = P(\tilde{e}_x + bv)$, which is a special case of Assumption 3.5.1.*

Definition 3.5.4. For any LDIM, $(\mathbf{H}, \mathbf{e})$, with $\Phi_{\mathbf{e}}$ non-diagonal, and satisfying Assumption 3.5.1, and , $p > 1$

$$\begin{aligned} \mathcal{L}^{(p)}(\mathbf{H}, \mathbf{e}) &:= \{(\tilde{\mathbf{H}}, \mathbf{y}) : \mathbf{e} = \tilde{\mathbf{e}}_o + \mathbf{F}\mathbf{y}, \tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H} & \mathbf{F} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{F} \in \mathcal{F}^{n \times M}, \\ \mathbf{y} &= \mathcal{M}(\mathbf{v}, p), \tilde{\mathbf{e}}_o \in \mathcal{A}^n, \mathbf{v} \in \mathcal{A}^m, m \in \mathbb{N}, \tilde{\mathbf{e}} = [\tilde{\mathbf{e}}_o, \mathbf{v}]^T, \\ M &= \sum_{k=0}^p \binom{m+k-1}{k}, \Phi_{\tilde{\mathbf{e}}} \text{ diagonal} \} \end{aligned} \quad (3.9)$$

is the space of all $\mathcal{L}_p$ -transformations for a given $(\mathbf{H}, \mathbf{e})$. The matrix $\mathbf{F}$ is obtained by concatenating $\mathbf{F}_\alpha \in \mathcal{F}^{n \times 1}$. This is done by first listing all the $\mathbf{F}_\alpha$ corresponding to “degree one” monomials in lexicographic order, then “degree 2” in lexicographic order, etc. That is, $\mathbf{F} := [\mathbf{F}_{b_1}, \dots, \mathbf{F}_{b_L}, \mathbf{F}_{b_{11}}, \mathbf{F}_{b_{12}}, \dots, \mathbf{F}_{b_{1L}}, \mathbf{F}_{b_{21}}, \mathbf{F}_{b_{22}}, \dots]$, where $\mathbf{F}_{b_{i_1 \dots i_k}}$ denotes the column vector of $\mathbf{F}$ corresponding to $\alpha = b_{i_1} + \dots + b_{i_k}$, b_i : canonical basis. For $m = 2$, $p = 3$ from Section 3.5.2, $\mathbf{F} = [\mathbf{F}_{b_1}, \mathbf{F}_{b_2}, \mathbf{F}_{b_{11}}, \mathbf{F}_{b_{12}}, \mathbf{F}_{b_{22}}, \mathbf{F}_{b_{111}}, \mathbf{F}_{b_{112}}, \mathbf{F}_{b_{122}}, \mathbf{F}_{b_{222}}]$. Here, $\mathbf{F}_{b_2} = \mathbf{F}_{(0,1)}$, $\mathbf{F}_{b_{12}} = \mathbf{F}_{(1,1)}$, $\mathbf{F}_{b_{222}} = \mathbf{F}_{(0,3)}$.

With this new “polynomial lifting” definition, one can transform a given LDIM with correlated noise sources to a transformed LDIM with latent nodes. As shown in the following lemma, transformation of correlations to uncorrelated latent nodes from Section 3.4 is replaced with uncorrelated latent clusters here. For any cluster c , $Ch(c) := \{Ch(i) : i \in c\}$, that is, $Ch(c)$ denotes the union of the children of the nodes present in cluster c .

Lemma 3.5.1. Let $(\mathbf{H}, \mathbf{e})$ be an LDIM that satisfies Assumption 3.5.1, and let $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ be the correlation graph of the exogenous noise sources, $\{e_i\}_{i=1}^n$.

Then, for every distinct $i, j \in [n]$, $(i, j) \in \mathcal{E}_c$ if and only if for every LDIG $(\tilde{\mathbf{H}}, \mathbf{y}) \in \mathcal{L}^{(p)}(\mathbf{H}, \mathbf{e})$, there exists a cluster c such that $i, j \in Ch(c)$.

Proof. Refer Appendix B.6. □

The following theorem shows that a subgraph, $\mathcal{G}^\ell(\mathcal{V}^\ell, \mathcal{E}^\ell)$, of the correlation graph, $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$, forms a maximal clique in $\mathcal{G}_c$ if and only if for any transformed LDIG in $\mathcal{L}^{(p)}(\mathbf{H}, \mathbf{e})$, the set of nodes in $\mathcal{V}^\ell$ is equal to the set of the children of some latent cluster.

Theorem 3.5.1. Let $(\mathbf{H}, \mathbf{e})$ be an LDIM defined by (3.1) which satisfies Assumption 3.5.1, and let $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ be the correlation graph of the exogenous noise sources, $\mathbf{e}$. Suppose that $\mathcal{G}^\ell(\mathcal{V}^\ell, \mathcal{E}^\ell) \subseteq \mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ is a maximal clique with $|\mathcal{V}^\ell| > 1$. Then, for any LDIM

$(\tilde{\mathbf{H}}, \mathbf{y}) \in \mathcal{L}^{(p)}(\mathbf{H}, \mathbf{e})$, there exist latent clusters $C_1^\ell, \dots, C_{k_\ell}^\ell$ in the LDIG of $(\tilde{\mathbf{H}}, \mathbf{y})$ such that

$$\mathcal{V}^\ell = \bigcup_{i=1}^{k_\ell} Ch(C_i^\ell) \text{ and } \mathcal{E}^\ell = \bigcup_{i=1}^{k_\ell} \mathcal{E}_{\ell,i}, \quad (3.10)$$

where $\mathcal{E}_{\ell,i} := \{(k, j) : k, j \in Ch(c_i^\ell)\}$. In particular, for any latent cluster C in the LDIG of $(\tilde{\mathbf{H}}, \mathbf{y})$, $Ch(C)$ forms a clique in $\mathcal{G}_c$.

Proof. See Appendix B.7. □

Consider the LDIG shown in Figure 3.3(a) with correlation graph given by Figure 3.3(b). As shown in Figure 3.3(d), if all the three nodes are correlated, one can transform this LDIM to an LDIM with latent node 4. However, here node 4 should capture the interactions from the latent nodes $y_1 - y_{10}$. Therefore, during reconstruction, one should accommodate non-linear interaction between the latent node and the observed nodes.

In the next section, we describe a way to perform the reconstruction, when $\mathcal{G}_c$ is unavailable.

3.6 Moral-Graph Reconstruction by Matrix Decomposition

In the previous sections, frameworks that convert a network with spatially correlated noise to a network with latent nodes were studied. In this section, a technique is provided to reconstruct the topology of the transformed LDIMs using the sparse plus low-rank matrix decomposition of the IPSDM obtained from the observed time-series. Note that this result does not require any extra information other than the IPSDM obtained from the true LDIM.

3.6.1 Topology Reconstruction under Affine Correlation

Here, reconstruction under affine correlation is discussed. Recall from Definition 3.4.2 that $\mathbf{e} = \tilde{\mathbf{e}}_o + \mathbf{F}\tilde{\mathbf{e}}_h$. Then, PSDM of $\mathbf{e}$, Φ_e , can be written as [17]: $\Phi_e = \Phi_{\tilde{\mathbf{e}}_o} + \Phi_{\tilde{\mathbf{e}}_o\tilde{\mathbf{e}}_h}\mathbf{F}^* + \mathbf{F}\Phi_{\tilde{\mathbf{e}}_h\tilde{\mathbf{e}}_o} + \mathbf{F}\Phi_{\tilde{\mathbf{e}}_h}\mathbf{F}^* = \Phi_{\tilde{\mathbf{e}}_o} + \mathbf{F}\Phi_{\tilde{\mathbf{e}}_h}\mathbf{F}^*$, where the second equality follows because $\tilde{\mathbf{e}}_h$ and $\tilde{\mathbf{e}}_o$ are uncorrelated with mean zero. Then, IPSDM of the observed nodes in

$$([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e}),$$

$$\begin{aligned} \Phi_o^{-1} &= (\mathbf{I}_n - \mathbf{H})^* (\Phi_{\tilde{e}_o} + \mathbf{F} \Phi_{\tilde{e}_h} \mathbf{F}^*)^{-1} (\mathbf{I}_n - \mathbf{H}) \\ &\stackrel{(a)}{=} \mathbf{A} - \mathbf{B}, \end{aligned} \tag{3.11}$$

where $\mathbf{A} = (\mathbf{I}_n - \mathbf{H})^* \Phi_{\tilde{e}_o}^{-1} (\mathbf{I}_n - \mathbf{H})$ and $\mathbf{B} = (\mathbf{I}_n - \mathbf{H})^* \Phi_{\tilde{e}_o}^{-1} \mathbf{F} (\Phi_{\tilde{e}_h}^{-1} + \mathbf{F}^* \Phi_{\tilde{e}_o}^{-1} \mathbf{F})^{-1} \mathbf{F}^* \Phi_{\tilde{e}_o}^{-1} (\mathbf{I}_n - \mathbf{H})$. Equality (a) follows from the matrix inversion lemma [65]. (3.11) can then be rewritten as:

$$\Phi_o^{-1} = \mathbf{S} + \mathbf{L}, \text{ where} \tag{3.12}$$

$$\mathbf{S} = (\mathbf{I}_n - \mathbf{H}^*) \Phi_{\tilde{e}_o}^{-1} (\mathbf{I}_n - \mathbf{H}), \tag{3.13}$$

$$\mathbf{L} = -\Psi^* \Lambda^{-1} \Psi, \tag{3.14}$$

$\Psi = \mathbf{F}^* \Phi_{\tilde{e}_o}^{-1} (\mathbf{I}_n - \mathbf{H})$, and $\Lambda = \mathbf{F}^* \Phi_{\tilde{e}_o}^{-1} \mathbf{F} + \Phi_{\tilde{e}_h}^{-1}$, which is similar to the model in [1]. If the moral-graph of the original LDIG is sparse and $M \ll n$, then $\mathbf{S}$ is sparse and $\mathbf{L}$ is low-rank. It was shown in [17] that the support of $\mathbf{S}$ retrieves the moral graph of $\mathcal{G}$. Furthermore, as shown in [1], under some assumptions that is applicable to a large class of problems, (i, j) -th entry of $\mathbf{S}$ is strictly real if and only if the edge $i - j$ is a strict spouse edge. Thus, it can be shown that, in a large class of problems, support of $\Im\{\mathbf{S}\}$ retrieves the exact topology of $\mathcal{G}$. Following the approach in [1], we reconstruct the network topology from the sparse+low-rank decomposition of $\Im\{\Phi_o^{-1}(z)\}$, which is a skew-symmetric matrix. For completeness, the essential theories and a relevant algorithm from [1] is provided below in Section 3.6.4. The idea here is to decompose a given skew-symmetric matrix, i.e. $\Im\{\Phi_o^{-1}(z)\}$, into the sparse and low-rank components ($\Im\{\mathbf{S}(z)\}$ and $\Im\{\mathbf{L}(z)\}$ respectively), and then to reconstruct the moral graph/topology from $\Im\{\mathbf{S}(z)\}$, for some $|z| = 1$.

The next subsection shows how the sparse low-rank decomposition is applicable in polynomial correlation setting also.

3.6.2 Topology reconstruction under polynomial correlation

Under polynomial correlation, recall from Definition 3.5.4 that $\mathbf{e}(z) = \tilde{\mathbf{e}}_o(z) + \mathbf{F} \tilde{\mathbf{y}}$. Let $\mathbf{x}_h = \tilde{\mathbf{e}}_h = \tilde{\mathbf{y}}$ and let $\mathbf{x} = [\mathbf{x}_o^T, \mathbf{x}_h^T]^T$. Then one can write

$$\begin{bmatrix} \mathbf{x}_o(z) \\ \mathbf{x}_h(z) \end{bmatrix} = \begin{bmatrix} \mathbf{H}(z) & \mathbf{F}(z) \\ \mathbf{0}_{M \times n} & \mathbf{0}_{M \times M} \end{bmatrix} \begin{bmatrix} \mathbf{x}_o(z) \\ \mathbf{x}_h(z) \end{bmatrix} + \begin{bmatrix} \tilde{\mathbf{e}}_o(z) \\ \tilde{\mathbf{e}}_h(z) \end{bmatrix}, \quad (3.15)$$

where, $\mathbf{F}_{ij}$ captures the influence of $\tilde{\mathbf{y}}_j$ on x_i and $\Phi_{\tilde{\mathbf{y}}}$ is block diagonal.

The topology of the sub-graph restricted to the observed nodes in the above LDIM and the true topology of the network are equivalent. Moreover, similar to (3.11) (see [1] for details), one can obtain (3.12) to (3.14) exactly.

Here $\mathbf{S}(z), \mathbf{L}(z) \in \mathbb{C}^{n \times n}$, $\Psi \in \mathbb{C}^{M \times n}$, and $\Lambda \in \mathbb{C}^{M \times M}$. If $M \ll n$, then $\mathbf{L}$ is low-rank. Let $\mathcal{J} := \{j \in [M] \mid \exists i \in [n] \text{ with } \mathbf{F}_{ij} \neq 0\}$ be the set of monomials that has non zero contribution to e_i , $i \in [n]$ and let $L = |\mathcal{J}|$. Under this scenario, $\mathbf{L}$ is low-rank if $L \ll n$. Section 3.7.1 demonstrates an example on application of the sparse plus low-rank decomposition to reconstruct the topology under polynomial correlation.

3.6.3 Low-rank plus Sparse Matrix Decomposition

Here, the following problem is considered: given a matrix $\mathbf{C}$ that is known to be sum of a sparse skew-symmetric matrix $\mathbf{S}$ and a low-rank skew-symmetric matrix $\mathbf{L}$, retrieve the sparse and low-rank components. The following optimization program modified from [40] is used to obtain the sparse low-rank decomposition, where $0 \leq t \leq 1$ is a pre-selected penalty factor [1].

$$\begin{aligned} (\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t) &= \arg \min_{\hat{\mathbf{S}}, \hat{\mathbf{L}}} t \|\hat{\mathbf{S}}\|_1 + (1-t) \|\hat{\mathbf{L}}\|_* \\ &\text{subject to } \hat{\mathbf{S}} + \hat{\mathbf{L}} = \mathbf{C}, \\ &\hat{\mathbf{S}}^T = -\hat{\mathbf{S}}, \hat{\mathbf{L}}^T = -\hat{\mathbf{L}}, \end{aligned} \quad (3.16)$$

where $\mathbf{C} = \Im\{\Phi_o^{-1}(z)\}$, for some $z \in \mathbb{C}$, $|z| = 1$.

In the next subsection, a sufficient condition and an algorithm for the exact recovery of the sparse and the low-rank components from $\mathbf{C}$ using (3.16) are provided.

3.6.4 Sufficient Condition for Sparse Low-rank Matrix Decomposition

In this subsection, a sufficient condition (proved in Chapter 2) is provided to uniquely decompose a matrix as a sum of the sparse skew-symmetric and the low-rank skew-symmetric components. Furthermore, an algorithm is provided that utilizes this sufficient condition to retrieve the sparse and low-rank components.

The following definitions are used in the subsequent results.

$$deg_{max}(\mathbf{M}) := \max \left(\max_{1 \leq i \leq n} \left(\sum_{j=1}^n \mathbb{1}_{\{\mathbf{M}_{ij} \neq 0\}} \right), \max_{1 \leq j \leq n} \left(\sum_{i=1}^n \mathbb{1}_{\{\mathbf{M}_{ij} \neq 0\}} \right) \right), \quad (3.17)$$

$$inc(\mathbf{M}) := \max_k \|\mathbf{U}\mathbf{U}^T e_k\|_2, \quad (3.18)$$

where $\mathbf{U}\Sigma\mathbf{V}^T$ is the compact singular value decomposition of $\mathbf{M}$ and $\|\cdot\|_2$ is the Euclidean norm of a vector.

The following is a sufficient condition that guarantees the unique decomposition of $\mathbf{C}$ (see Chapter 2 for details).

Lemma 3.6.1. *Suppose that we are given a matrix $\mathbf{C}$, which is the sum of a sparse matrix $\tilde{\mathbf{S}} \in \mathbb{S}^n$ and a low-rank matrix $\tilde{\mathbf{L}} \in \mathbb{S}^n$. If $(\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ satisfies*

$$deg_{max}(\tilde{\mathbf{S}})inc(\tilde{\mathbf{L}}) < \frac{1}{12}, \quad (3.19)$$

then there exists a penalty factor $t \in [0, 1]$ such that (3.16) returns $(\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t) = (\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$.

Remark 3.6.1. *The results in [40, 1] are proved for the optimization problem with the objective function $\gamma\|\hat{\mathbf{S}}\|_1 + \|\hat{\mathbf{L}}\|_*$, where $\gamma \geq 0$. However, the results hold for (3.16) as well since the problems are equivalent (via the map $t = \gamma/(1 + \gamma)$)*

The sufficient condition (3.19) roughly translates to $\tilde{\mathbf{S}}$ being sparse and the number of maximal cliques, M , being small, with clique sizes not too small (see [1] for more details).

The following metrics are used to measure the accuracy of the estimates $(\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t)$ in

the optimization (3.16).

$$\begin{aligned} tol_t &:= \frac{\|\widehat{\mathbf{S}}_t - \tilde{\mathbf{S}}\|_F}{\|\tilde{\mathbf{S}}\|_F} + \frac{\|\widehat{\mathbf{L}}_t - \tilde{\mathbf{L}}\|_F}{\|\tilde{\mathbf{L}}\|_F}, \\ diff_t &:= (\|\widehat{\mathbf{S}}_{t-\epsilon} - \widehat{\mathbf{S}}_t\|_F) + (\|\widehat{\mathbf{L}}_{t-\epsilon} - \widehat{\mathbf{L}}_t\|_F), \end{aligned} \quad (3.20)$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\epsilon > 0$ is a sufficiently small fixed constant. Note that tol_t requires the knowledge of the true matrices $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{L}}$, whereas $diff_t$ does not.

The following Lemma is proved in [1] and is applied in Algorithm 4 later to retrieve the sparse and low-rank components.

Lemma 3.6.2. *Suppose we are given a matrix $\mathbf{C}$, which is obtained by summing $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{L}}$, where $\tilde{\mathbf{S}}$ is a sparse skew-symmetric matrix and $\tilde{\mathbf{L}}$ is a low-rank skew-symmetric matrix. If $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{L}}$ satisfies $\deg_{\max}(\tilde{\mathbf{S}})inc(\tilde{\mathbf{L}}) < 1/12$, then there exist at least three regions where $diff_t = 0$. In particular, there exists an interval $[t_1, t_2] \subset [0, 1]$ with $0 < t_1 < t_2 < 1$ such that the solution of (3.16) is $(\widehat{\mathbf{S}}_t, \widehat{\mathbf{L}}_t) = (\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ for any $t \in [t_1, t_2]$.*

Following the procedure in [1], moral-graph/topology reconstruction from the imaginary part of the IPSDM is considered here. The following are some assumptions from [1], assumed for the exact recovery of the topology. The details can be found in [1], and are skipped here. In the absence of Assumption 4, the reconstruction algorithm will detect some false positive edges, but none of the true edges are missed.

Assumption 3.6.1. *For any $i \in [n]$, $k \in [n]$ ($k' \in [L]$), if $\mathbf{H}_{ik}(z) \neq 0$, ($F_{ik'} \neq 0$) then $\Im\{\mathbf{H}_{ik}(z)\} \neq 0$ ($\Im\{\mathbf{F}_{ik'}(z)\} \neq 0$), for all $z \in \mathbb{C}$, $|z| = 1$.*

Assumption 3.6.2. *For the LDIM in (3.1), and $i, k, l \in \mathcal{V}$, if $\mathbf{H}_{ki}(z) \neq 0$ and $\mathbf{H}_{kl}(z) \neq 0$, then $\angle \mathbf{H}_{ki}(z) = \angle \mathbf{H}_{kl}(z)$.*

The following Lemma from [1] reconstructs the exact topology of the LDIM from $\mathbf{S}$ in (3.13).

Lemma 3.6.3. *Consider a well-posed and topologically detectable LDM $(\mathbf{H}, \mathbf{e})$ described by (3.1) with the associated graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, and satisfying Assumption 3.6.2. Let $\mathbf{S}(z)$ be given by (3.13) and let $\widehat{\mathcal{E}}_o := \{(i, j) : \Im\{\mathbf{S}_{ij}(z)\} \neq 0, i < j\}$ and $\overline{\mathcal{E}}_o := \{(i, j) \mid (i, j) \in \mathcal{E} \text{ or } (j, i) \in \mathcal{E}, i < j\}$, for some $z \in \mathbb{C}$, $|z| = 1$. Then, $\widehat{\mathcal{E}}_o \subseteq \overline{\mathcal{E}}_o$. Additionally, if the LDM satisfies Assumption 3.6.1, then $\widehat{\mathcal{E}}_o = \overline{\mathcal{E}}_o$ almost always.*

The following is the main result of this section, which follows from Lemma 3.6.1 and Lemma 3.6.3.

Theorem 3.6.1. *Let $(\mathbf{H}, \mathbf{e})$ be an LDIM with Φ_e non-diagonal that satisfy Assumption 3.6.1-Assumption 3.6.2. Suppose there exists an $(\mathbf{H}, \mathbf{F}, \tilde{\mathbf{e}}) \in \mathcal{L}_q(\mathbf{H}, \mathbf{e})$ such that for $\mathbf{S}$ and $\mathbf{L}$ in (3.12), $\deg_{\max}(\Im\{\mathbf{S}\})\text{inc}(\Im\{\mathbf{L}\}) < 1/12$. Then, the true topology of $(\mathbf{H}, \mathbf{e})$ can be reconstructed by solving the optimization problem (3.16) with $\mathbf{C} = \Im\{\Phi_o^{-1}(z)\}$ for some $z = e^{i\omega}$, $\omega \in (0, 2\pi]$. In particular, $\mathcal{T}(\mathcal{V}, \mathcal{E}) = \text{supp}(\hat{\mathbf{S}}_t)$, for appropriately selected t .*

Proof. Refer Appendix B.8. □

Additionally, applying the algorithms in [1], the correlation graph also can be reconstructed, if the LDIM satisfies the following assumption, as shown in simulation results.

Assumption 3.6.3. *For every distinct latent nodes k_h, k'_h in the transformed dynamic graph, the distance between k_h and k'_h is at least four hops.*

Algorithm 4 Matrix decomposition

Input: $\Phi_{oo}^{-1}(z)$: IPSDM among $\mathcal{V}_o, \varepsilon, z = e^{j\omega}$, $\omega \in (-\pi, \pi]$

Output: Matrices $\Im(\mathbf{S}(z))$ and $\Im(\mathbf{L}(z))$

- 1: Set $\mathbf{C} = \Im\{\Phi_{oo}^{-1}(z)\}$
 - 2: Initialize $(\hat{\mathbf{S}}_0, \hat{\mathbf{L}}_0) = (\mathbf{C}, \mathbf{0})$
 - 3: **for all** $t \in \{\varepsilon, 2\varepsilon, \dots, 1\}$ **do**
 - 4: Solve the convex optimization (3.16) and calculate diff_t in (3.20)
 - 5: **end for**
 - 6: Identify the three regions where diff_t is zero and denote the middle region as $[t_1, t_2]$.
 Pick a $t_0 \in [t_1, t_2]$ and the corresponding pair $(\hat{\mathbf{S}}_{t_0}, \hat{\mathbf{L}}_{t_0})$.
 - 7: **if** $\deg_{\max}(\hat{\mathbf{S}}_{t_0})\text{inc}(\hat{\mathbf{L}}_{t_0}) < \frac{1}{12}$ **then**
 - 8: $(\hat{\mathbf{S}}(z), \hat{\mathbf{L}}(z)) = (\hat{\mathbf{S}}_{t_0}, \hat{\mathbf{L}}_{t_0})$
 - 9: Return $(\hat{\mathbf{S}}(z), \hat{\mathbf{L}}(z))$
 - 10: **end if**
-

3.7 Simulation results

In this section, we demonstrate topology reconstruction of an LDIM $(\mathbf{H}, \mathbf{e})$ with Φ_e non-diagonal, from $\Phi_{\mathbf{x}}^{-1}$, using the sparse+low-rank decomposition technique discussed in

Section 3.6 for an affinely correlated network. Figure 3.4(a)-3.4(c) respectively depicts $\mathcal{G}(\mathcal{V}, \mathcal{E})$, $\mathcal{T}(\mathcal{V}, \mathcal{E})$, and $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$, and 3.5(a)-(b) shows $\mathcal{G}(\mathcal{V}_t, \mathcal{E}_t \setminus \mathcal{E})$ and $\mathcal{G}_t(\mathcal{V}_t, \mathcal{E}_t)$ described in Section 3.4. Simulations are done in Matlab. Yalmip [68] with SDP solver [69] is used to solve the convex program (3.16).

For the simulation, we assume we have access to the true PSDM, $\Phi_{\mathbf{x}}$, of the LDIM of Figure 3.4(a). Here, $\Phi_{\mathbf{e}}$ is non-diagonal with the (unkown) correlation structure as shown in Figure 3.4(c).

For the reconstruction, the imaginary part of the IPSDM, $\mathbf{C} = \Im\{\Phi_{\mathbf{x}}^{-1}(z)\}$ is employed in the convex optimization (3.16) for $z = e^{2\pi/8}$. Optimization (3.16) is solved for all the values of $t \in [\epsilon, 1]$, with the interval $\epsilon = 0.01$. Notice that for $t = 0$ $(\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t) = (\mathbf{C}, 0)$. Figure 3.6 shows the comparison of tol_t and $diff_t$ versus t . $(\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t)$ for $t = 0.36$ is picked, which belongs to the middle zero region of $diff_t$ as described in [1].

$\mathbb{I}_{\{\hat{\mathbf{S}}_t \neq \mathbf{0}\}}$ returned the exact topology of Figure 3.4(b). From $\mathbb{I}_{\{\hat{\mathbf{L}}_t \neq \mathbf{0}\}}$, by following Algorithms 2 and 3 in [1], $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ also is reconstructed, which matches Figure 3.4(c) exactly.

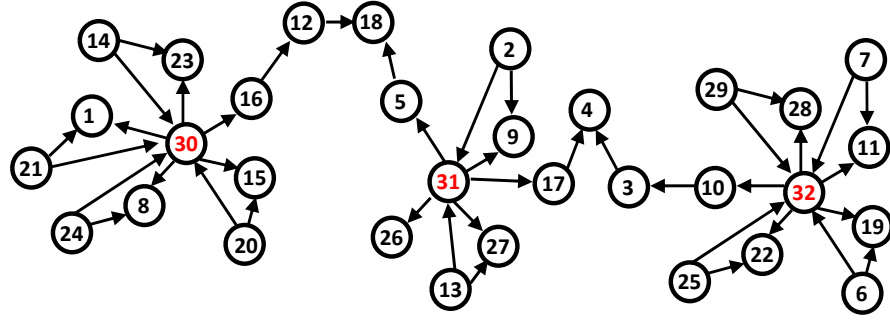
3.7.1 Polynomial correlation

Here, the topology reconstruction of the network when the noise processes are polynomially correlated, is shown. For the simulation, consider the LDIG shown in Figure 3.4(a) and correlation graph of 3.4(b). The noise processes are IID GP as shown in Section 3.5.2 with $m = 2$ and $p = 3$. The entries of $\mathbf{F}$ are such that only the co-efficients corresponding to y_2, y_5 , and y_9 are non-zero, and the coefficients $F_{ij} = 0$, for every $1 \leq i \leq 29, 1 \leq j \leq 10, j \notin \{2, 5, 9\}$.

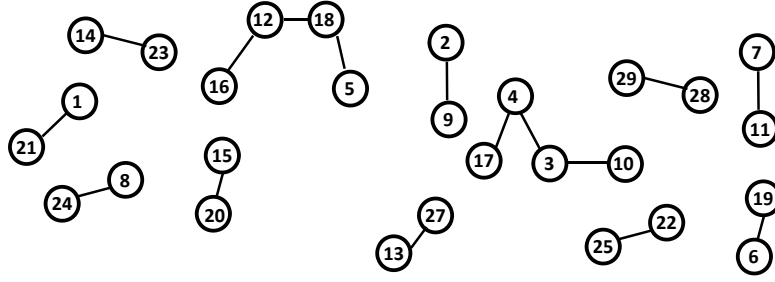
Figure 3.7 shows the $diff_t$ and tol_t plot of applying Algorithm 4 with $\epsilon = 0.01$. As shown in the plots, for $t \in [.28, .38]$, tol_t is zero, which corresponds to the exact decomposition. Additionally, as mentioned in Lemma 3.6.2, $diff_t$ is zero in this interval. The support of $\hat{\mathbf{S}}_t$ for some $t \in [.28, .38]$ reconstructs the exact topology, which validates Theorem 3.6.1.

3.7.2 Finite data simulation

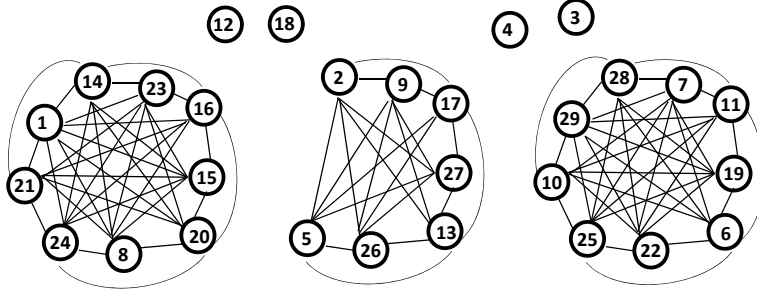
In this section, to evaluate the effect of finite data size, simulations are run on a synthetic data set based on the network shown in Figure 3.5(b). For the PSD estimation, Welch method [88] is used. Notice that the accuracy and the sample complexity of the estimation can be improved by employing advanced IPSDM estimation techniques



(a) The true LDIG, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, of $(\mathbf{H}, \mathbf{e})$.

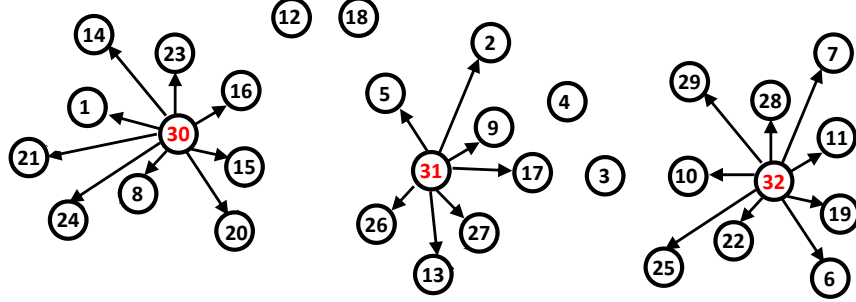


(b) The true topology, $\mathcal{T}(\mathcal{V}, \mathcal{E})$ of $(\mathbf{H}, \mathbf{e})$.

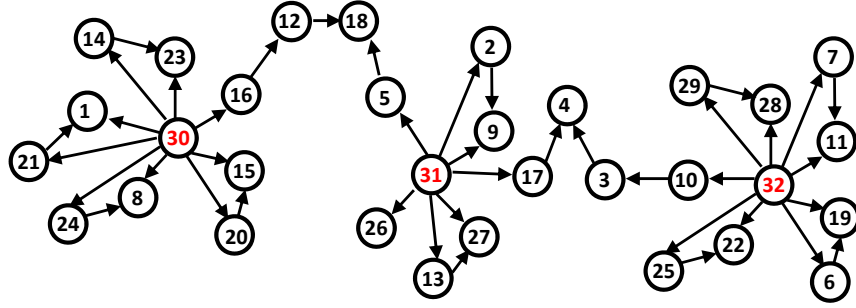


(c) Correlation graph, $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$, of $(\mathbf{H}, \mathbf{e}); M = 3$

Figure 3.4: Network structure



(a) Transformed correlation structure with latent nodes



(b) Transformed graph, $\mathcal{G}_t(\mathcal{V}_t, \mathcal{E}_t)$, of $(\tilde{\mathbf{H}}, \tilde{\mathbf{e}}) \in \mathcal{L}_q(\mathbf{H}, \mathbf{e})$.

Figure 3.5: Transformed network structure

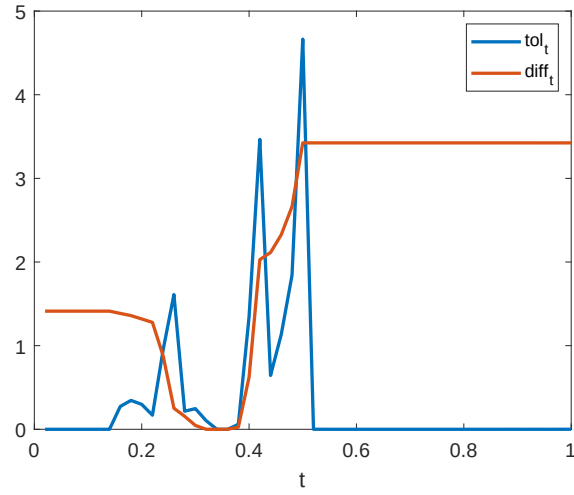


Figure 3.6: tol_t and diff_t plots for the network of Figure 3.4(a) under affine correlation

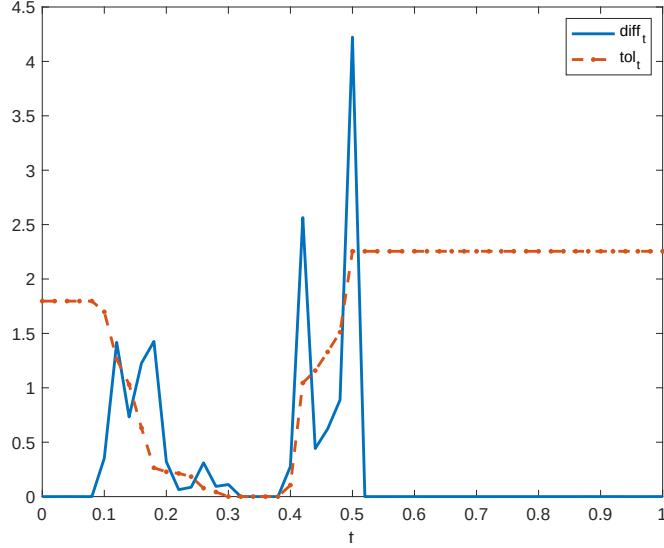


Figure 3.7: tol_t and diff_t plots for the network of Figure 3.4(a) under polynomial correlation

from literature, for example see [78, 32, 86]. Figure 3.8 shows the estimated results from a sample size of 10^6 ; Figure 3.8(a) shows the true and the estimated IPSDM matrices. The fourth matrix of Figure 3.8(a) shows the topology estimated directly from $\Im\{\hat{\Phi}_o^{-1}(e^{j2\pi/5})\}$, without decomposition. The estimation is done by hard thresholding. That is, the edge (i, j) is detected if $[\Im\{\hat{\Phi}_o^{-1}(e^{j2\pi/5})\}]_{i,j}$ is greater than a threshold, and not detected otherwise. The detection threshold is selected to obtain the minimum number of errors. 14 out of 16 edges are detected but 6 false positive edges are also detected, with the total error of 50% (8 out of 16 edges). This shows that estimating the topology directly from $\Im\{\hat{\Phi}_o^{-1}(e^{j2\pi/5})\}$ returns an undesirable number of errors.

Towards the exact topology retrieval, the optimization (3.16) is performed on $\Im\{\hat{\Phi}_o^{-1}(e^{j2\pi/5})\}$ to obtain sparse plus low-rank decomposition. Figure 3.8(b) shows the sparse part retrieved from the decomposition of $\Im\{\hat{\Phi}_o^{-1}(e^{j2\pi/5})\}$ for various t from 0 to 0.5. $\hat{\mathbf{S}}_t$ at $t = 0.34$ is selected for estimating the topology. Thus, as illustrated by the above example, with the approach proposed in the chapter it is possible to choose a threshold that yields 100% detection without sacrificing the false alarm performance.

In order to demonstrate that the techniques proposed in this chapter do not degrade drastically with lesser number of samples, a simulation is run on 6000 samples. Employing the detection from $\Im\{\hat{\Phi}_o^{-1}(e^{j2\pi/5})\}$ returned 48 false edges and missed one edge,

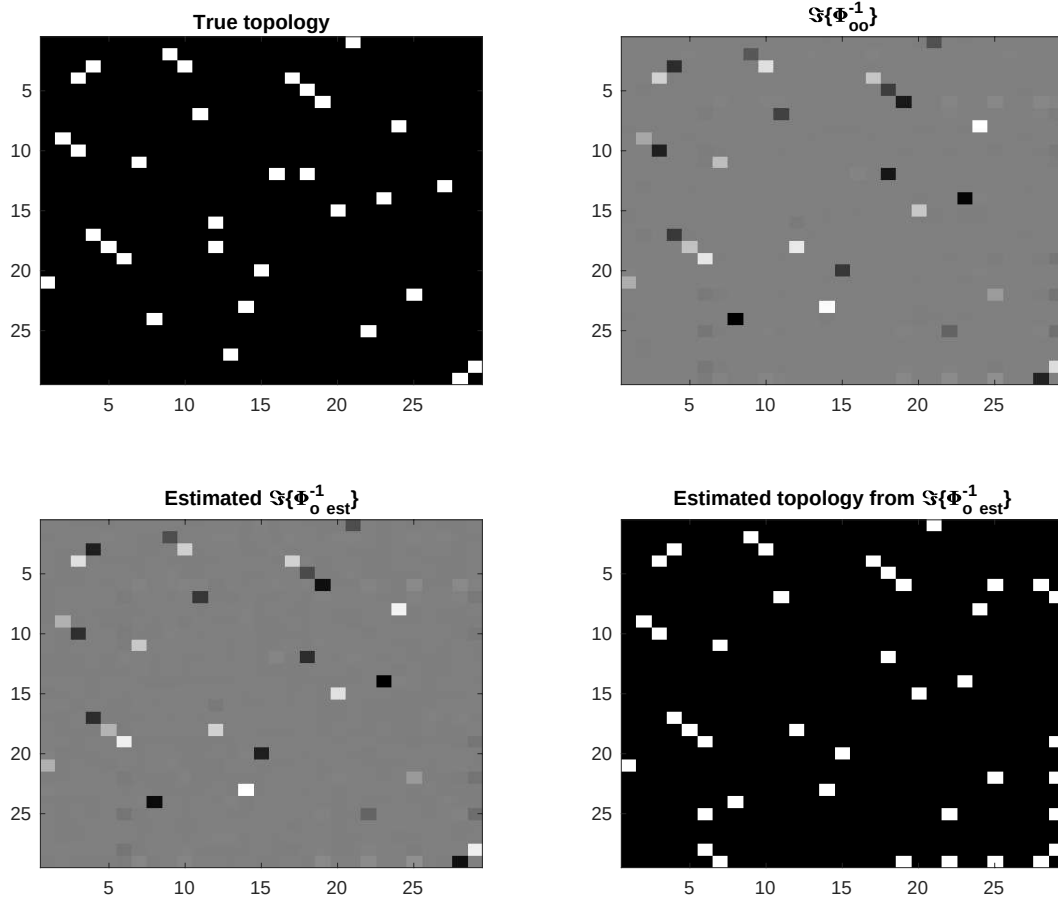


Figure 3.8: The true topology and the topology estimated directly from $\hat{\Phi}_o^{-1}(e^{j2\pi/5})$ for the network shown in Figure 3.4(a) under affine correlation with $N = 10^6$ data samples. The figures show the entries of matrices of size 29×29 . In the figures of topology, black denotes zero (edge absent) and white denotes 1 (edge present). In the gray scale images, gray denotes zero.

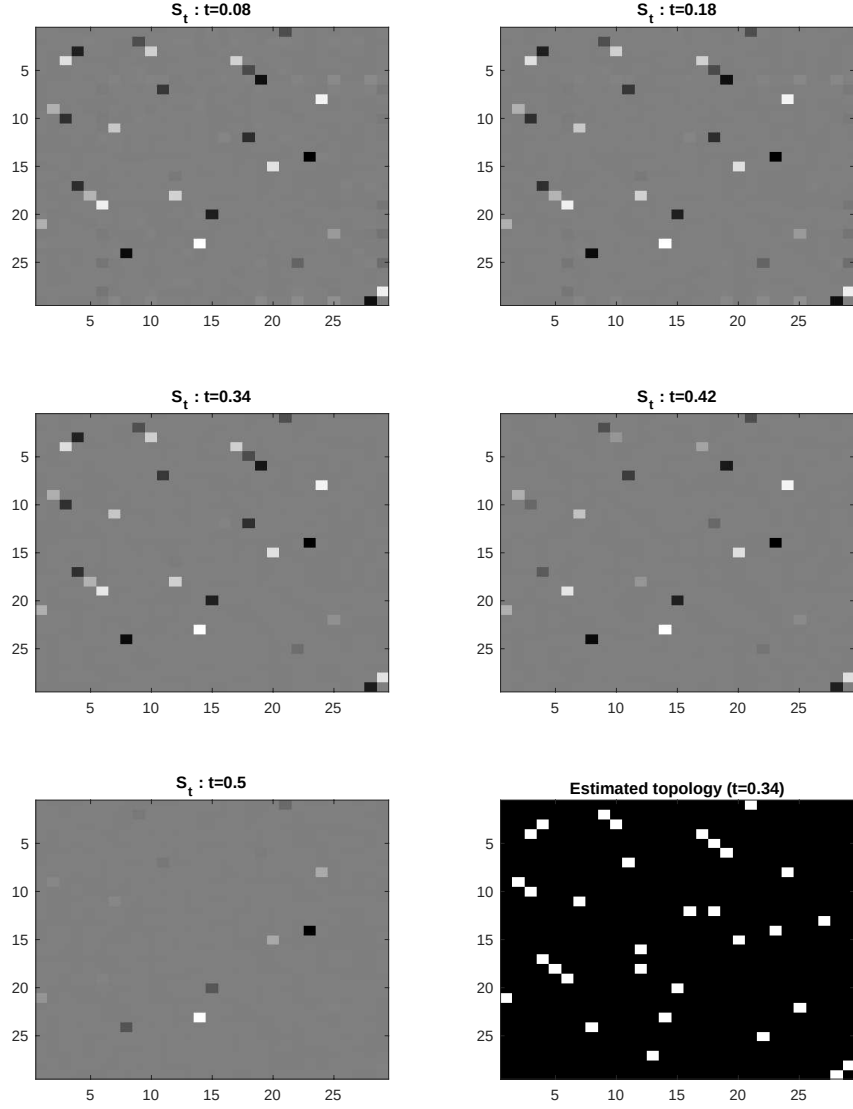


Figure 3.9: $\hat{\mathbf{S}}_t$ obtained by solving (3.16) for $\mathbf{C} = \Im\{\hat{\Phi}_o^{-1}(e^{j2\pi/5})\}$. Topology is estimated from $\hat{\mathbf{S}}_t$ for $t = 0.34$. In the figures of topology, black denotes zero (edge absent) and white denotes 1 (edge present). In the gray scale images, gray denotes zero.

thus a total of 49 errors. However, with the decomposition, detection from the support of $\widehat{\mathbf{S}}_t$ at $t = .34$ detected 14 out of 16 edges and missed none, thus giving an error rate of 87.5%. This confirms that the decomposition based reconstruction proposed by the chapter yield substantial advantages. The sample complexity analysis of the chapter's methods is open for future research.

3.8 Conclusion

In this chapter, the problem of reconstructing the topology of an LDIM with spatially correlated noise sources was studied. First, assuming affine correlation and the knowledge of correlation graph, the given LDIM was transformed into an LDIM with latent nodes, where the latent nodes were characterized using the correlation graph and all the nodes were excited by uncorrelated noises. For polynomial correlation, a generalization of the affine correlation, the latent nodes in the transformed LDIMs were excited using clusters of noise, where the noise clusters were uncorrelated with each other. Finally, using a sparse low-rank matrix decomposition technique, the exact topology of the LDIM was reconstructed, solely from the IPSDM of the true LDIM, when the network satisfied a sufficient condition required for the matrix decomposition. Simulation results are provided that verify the theoretical results.

Chapter 4

Information Theoretically Optimal Sample Complexity of Learning *Dynamical* Directed Acyclic Graphs

In this chapter, the optimal sample complexity of learning the underlying interaction/dependencies of a Linear Dynamical System (LDS) over a Directed Acyclic Graph (DAG) is studied. The sample complexity of learning a DAG's structure is well-studied for static systems, where the samples of nodal states are independent and identically distributed (i.i.d.). However, such a study is less explored for DAGs with dynamical systems, where the nodal states are temporally correlated. We call such a DAG underlying an LDS as *dynamical* DAG (DDAG). In particular, we consider a DDAG where the nodal dynamics are driven by unobserved exogenous noise sources that are wide-sense stationary (WSS) in time but are mutually uncorrelated, and have the same power spectral density (PSD). Inspired by the static settings, a metric and an algorithm based on the PSD matrix of the observed time series are proposed to reconstruct the DDAG. The equal noise PSD assumption can be relaxed such that identifiability conditions for DDAG reconstruction are not violated. For the LDS with WSS (sub) Gaussian exogenous noise sources, it is shown that the optimal sample complexity (or length of state trajectory) needed to learn the DDAG is $n = \Theta(q \log(p/q))$, where p is the number of

nodes and q is the maximum number of parents per node. To prove the sample complexity upper bound, a concentration bound for the PSD estimation is derived, under two different sampling strategies. A matching min-max lower bound using generalized Fano’s inequality also is provided, thus showing the order optimality of the proposed algorithm.

4.1 Introduction

Learning the interdependency structure in a network of agents, from passive time series observations, is a salient problem with applications in neuroscience [43], finance [44], meteorology [13], etc. Reconstructing the exact structure with the dependency/causation directions has a wide range of applications. For example, the identification of causation structure among the shares helps in obtaining robust portfolio management in the stock market [44]. Similarly, causal graphs are useful in understanding dynamics and identifying contributing factors of a public epidemic emergency situation [89].

The structure of directed interactions in a network of agents is conveniently represented using directed graphs, with the agents as nodes and the directed interactions as directed edges. If the underlying graph doesn’t have cycles, it is called a Directed Acyclic Graph (DAG). In general, it is not possible to reconstruct the exact structure with the direction of different edges. Instead, in many networks it is possible to retrieve only the Markov equivalence graphs, the set of graphs satisfying the same conditional dependence property, from data without any intervention [45]. In applications such as finance [44], climate science [13] etc., the agent states, instead of being temporally independent, can evolve over time due to past directed interactions. Such temporal evolution can be represented by a linear dynamical system (LDS). In LDS, the interaction between agent states is captured by a linear time-invariant function. In this chapter, we study the identifiability and present the first sample complexity results for learning a DAG of LDS, which we term as *Dynamical* DAG or **DDAG**. This is distinguished from *static* DAG, where the agent states are temporally independent and the DAG does not correspond to temporal dynamics.

4.1.1 Related Work

Static DAG Learning: The problem of obtaining an upper bound on the sample complexity of learning static DAGs goes back twenty-five years [90], [91]. However, tight

characterization of optimal rates for DAG learning is a harder problem compared to undirected networks [92], primarily due to the order identification step. Identifiability conditions for learning static DAGs with linear interactions and excited by equal variance Gaussian noise were given in [93]. Several polynomial time algorithms have been proposed for static DAG reconstruction using samples of states at the graph nodes; see [94, 95, 45, 96, 92, 97, 98], [99] [100] [101] [102], and the reference therein. An information-theoretic lower bound on structure estimation was studied in [94]. In [92], it was shown that the order optimal sample complexity for static Gaussian graphical model with equal variance is $n = \Theta(q \log(p/q))$, where p is the number of nodes and q is the maximum number of parents. The authors showed that the algorithm given in [96] provides an upper bound that matches a min-max lower bound for the number of samples. However, similar results for DDAGs, with underlying temporal directed interaction between agent states, have not been studied, to the best of our knowledge.

LDS Learning: Graph reconstruction, in general, is challenging in a network of LDS (undirected, or bi-directed), as it involves time-dependencies between collected samples of nodal states. Learning the conditional independence structure in LDS with independent and identically distributed (white) excitation was explored in [103], [104], [105], [106], [107], and the references therein. However, the methods in the cited papers do not extend to LDS that is excited by WSS (wide-sense stationary) noise, which makes correlations in state samples more pronounced. For LDS with WSS noise, [108], [109], [110], and [111], estimated the conditional correlation structure, which contains true edges in the network and extra edges between some two-hop neighbors [17]. A consistent algorithm for the recovery of exact topology in LDS with WSS noise was provided in [18], with the corresponding sample-complexity analysis developed in [112], using a neighborhood-based regression framework. However, the developed algorithms and related sample complexity results do not extend to directed graphs and hence exclude DDAG reconstruction. DDAG reconstruction from the time-series data has been explored using the framework of directed mutual information in [77] but without rate characterization for learning from finite samples. Recently, extending the results in [96], [113] studied the graph reconstruction for Discrete Time Fourier Transform (DTFT) models, a special case of LDS with i.i.d. noise, but again without rate characterization for learning from finite samples.

Contribution: This chapter presents an information-theoretically optimal sample complexity analysis for learning a dynamical DAG (DDAG) excited by wide-sense stationary

(WSS, i.e., non i.i.d) noise of equal power spectral density, using samples of state trajectories of the underlying linear dynamical system (LDS). To the best of our knowledge, this is the first work to study and prove sample complexity analysis for DDAGs. We consider learning under two sampling scenarios, viz; 1) restart and record, 2) continuous sampling. While the former pertains to samples collected from disjoint (independent) trajectories of state evolution, the latter includes samples from a single but longer state trajectory (see Fig. 4.3). Surprisingly, the results in this chapter show that the estimation errors are not influenced by the sampling strategy (restart and record, or continuous) as long as the number of collected samples are over a determined threshold given by $n = O(q \log(p/q))$, where p is the number of nodes and q is the maximum number of parents per node. We also provide a matching information-theoretic lower-bound, $\max\left(\frac{\log p}{2\beta^2 + \beta^4}, \frac{q \log(p/q)}{M^2 - 1}\right)$, where β and M are system parameters (see Definition 4.2.1); thus obtaining an order optimal bound $n = \Theta(q \log(p/q))$.

Our learning algorithm relies on first deriving rules for DDAG estimation using the Power Spectral Density Matrix (PSDM) of nodal states, inspired by the estimator for static DAGs based on covariance matrices [96]. Subsequently, the sample complexity associated with learning is derived by obtaining concentration bounds for the PSDM. In this regard, characterization of non-asymptotic bounds of PSDMs for a few spectral estimators have been obtained in [114, 1, 115] previously. A unified framework of concentration bounds for a general class of PSDM estimators was recently presented in [116]. Our concentration bounds of the PSDM are reached using different proof steps, based on Rademacher random variables and symmetrization argument [117].

Notations: Bold faced small letters, $\mathbf{x}$ denote vectors; Bold faced capital letters, $\mathbf{A}$ denote matrices; For a time-series, $\mathbf{x}, \check{\mathbf{x}}(t)$ denotes the value of $\mathbf{x}$ at time t , $\mathbf{x}(\omega)$ denotes the discrete time Fourier transform of $\mathbf{x}$, $\mathbf{x}(\omega) := \sum_{k=-\infty}^{\infty} \check{\mathbf{x}}(k) e^{-i\omega k}$, $\omega \in \Omega = [0, 2\pi]$; $\text{diag}(v_1, \dots, v_p)$ operator creates a diagonal matrix with diagonal entries $v_1, \dots, v_p$; $\Phi_{\mathbf{x}AB}$ or Φ_{AB} denotes the matrix obtained by selecting rows A and columns B in $\Phi_{\mathbf{x}}$; $\mathbf{A}^*$ denotes conjugate transpose of $\mathbf{A}$. $\|\mathbf{A}\|$ is the spectral norm of $\mathbf{A}$ and $\|\mathbf{v}\|_2$ is the Euclidean norm of vector $\mathbf{v}$.

4.2 System Model of DDAG

We describe different aspects of DDAG (DAG underlying an LDS) and the sampling strategies considered. We begin with some necessary DAG terminologies.

DDAG terminology: The *dynamical* directed acyclic graph (DDAG), is given by

$G := (V, E)$, where $V = \{1, \dots, p\}$, and E is edge set of directed edge $i \rightarrow j$. A directed path from i to j is a path of the form $i = v_0 \rightarrow v_1 \rightarrow \dots \rightarrow v_\ell \rightarrow v_{\ell+1} = j$, where $v_k \in V$ and $(v_k, v_{k+1}) \in E$ for every $k = 0, \dots, \ell$. A cycle is a directed path from i to i , which does not exist in DDAG G . For G , $pa(i) := \{j \in V : (j, i) \in E\}$ denotes the parents set and $desc(i)$ denotes the descendants of i , the nodes that have a directed path from i . The set $nd(i) := V \setminus desc(i)$ denotes the non-descendants set and the set $an(i) \subset nd(i)$ denotes the ancestors set, the nodes that have a directed path to i . At set $C \subseteq V$ is called ancestral if for every $i \in C$, $pa(i) \subseteq C$. Figure 4.1 shows an example DDAG with the node definitions. A node set $C \subseteq V$ is said to be a topological ordering on G if for every $i, j \in C$, $i \in desc(j)$ in G implies $i > j$. $\mathcal{G}_{p,q}$ denotes the family of DDAGs with p nodes and at most q parents per node.

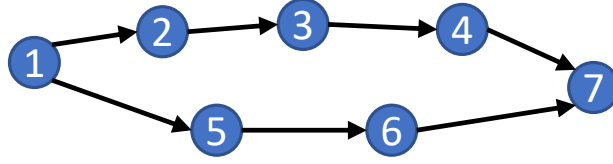


Figure 4.1: An example DDAG. Node 1 is an ancestor and node 7 is a descendant of every node in the graph. The set $\{1, 2, 5\}$ is an ancestral set but $\{2, 5\}$ is not. $an(3) = \{1, 2\}$, $desc(3) = \{4, 7\}$, $nd(3) = \{1, 2, 5, 6\}$.

Without a loss of generalizability, we use the same terminology for the DDAG and the underlying DAG.

LDS Model excited by equal PSD WSS noise: For the DDAG $G = (V, E) \in \mathcal{G}_{p,q}$, we consider a linear dynamical system (LDS) with p scalar state variables, corresponding to nodes in V . Node i is equipped with time series measurements, $\{\check{x}_i(k)\}_{k \in \mathbb{Z}}$, $1 \leq i \leq p$. The LDS evolves according to the linear time-invariant model,

$$\check{x}_i(k) = \sum_{(i,j) \in E, j \neq i}^p (\check{h}_{ij} \star \check{x}_j)(k) + \check{e}_i(k), \quad k \in \mathbb{Z}, \quad (4.1)$$

where transfer function $\check{h}_{ij} \neq 0$ when directed edge $(i, j) \in E$. The exogenous noise $\{\check{e}_i(k)\}_{k \in \mathbb{Z}}$, $1 \leq i \leq p$, are zero mean wide sense stationary Gaussian processes, uncorrelated across nodes. Taking the discrete-time Fourier transform (DTFT) of (4.1)

provides the frequency representation for every $\omega \in \Omega = [0, 2\pi]$,

$$\mathbf{x}_i(\omega) = \sum_{(i,j) \in E, j \neq i}^p \mathbf{H}_{ij}(\omega) \mathbf{x}_j(\omega) + \mathbf{e}_i(\omega), \quad 1 \leq i \leq p, \quad (4.2)$$

where $\mathbf{x}_i(\omega) = \mathcal{F}\{\check{x}_i\} := \sum_{k=-\infty}^{\infty} \check{x}_i(k) e^{-i\omega k}$, $\mathbf{e}_i(\omega) = \mathcal{F}\{\check{e}_i\}$, and $\mathbf{H}_{ij}(\omega) = \mathcal{F}\{\check{h}_{ij}\}$. The model in (4.2) can be represented in the matrix form to obtain the following LDS,

$$\mathbf{x}(\omega) = \mathbf{H}(\omega) \mathbf{x}(\omega) + \mathbf{e}(\omega), \quad \forall \omega \in \Omega, \quad (4.3)$$

where $\mathbf{e}(\omega)$ is the WSS noise. In this chapter, we are interested in the LDS with $\Phi_{\mathbf{e}}(\omega) = \sigma(\omega) \text{diag}(\alpha_1, \dots, \alpha_p)$, where α_i are known and can be a function of ω . For the simplicity of analysis, henceforth it is assumed that $\Phi_{\mathbf{e}}(\omega) = \sigma(\omega) \mathbf{I}$.

Remark 4.2.1. *The assumption $\Phi_{\mathbf{e}}(\omega) = \sigma(\omega) \mathbf{I}$ is a restrictive assumption. However, we would like to remark that some form of restriction is required due to the impossibility of DAG reconstruction in a general setup due to identifiability issues [101]. However, the assumption can be relaxed to incorporate the identifiability conditions on Φ_{e_i} and $\mathbf{H}$ to retrieve the topological ordering, similar to [45]. Furthermore, our results on DDAG reconstruction require the equal PSD to hold only at some known $\omega \in \Omega$ which is less restrictive.*

The power spectral density matrix (PSDM) of the time-series $\mathbf{x}$ at the angular frequency $\omega \in \Omega$ is given by

$$\Phi_{\mathbf{x}}(\omega) = \mathcal{F}\{R_{\mathbf{x}}(t)\} = \sum_{k=-\infty}^{\infty} R_{\mathbf{x}}(k) e^{-i\omega k}, \quad (4.4)$$

where $R_{\mathbf{x}}(k) := \mathbb{E}[\check{\mathbf{x}}(k) \check{\mathbf{x}}^T(0)]$ is the auto-correlation matrix of the time-series $\mathbf{x}$ at lag k . The (i, j) -th entry of $\Phi_{\mathbf{x}}$ is denoted by Φ_{ij} . For the LDS (4.3), the PSDM is given by

$$\Phi_{\mathbf{x}}(\omega) = (\mathbf{I} - \mathbf{H}(\omega))^{-1} \Phi_{\mathbf{e}}(\omega) ((\mathbf{I} - \mathbf{H}(\omega))^{-1})^*. \quad (4.5)$$

Consider the following additional non-restrictive assumptions on the power spectral density and correlation matrix of the LDS states.

Assumption 4.2.1. *There exists an $M \in \mathbb{R}$ such that $\frac{1}{M} \leq \lambda_{\min}(\Phi_{\mathbf{x}}) \leq \lambda_{\max}(\Phi_{\mathbf{x}}) \leq M$, where $\lambda_{\min}$ and $\lambda_{\max}$ respectively denote minimum and maximum eigenvalues.*

Assumption 4.2.2. *The auto-correlation matrix of the time-series $\mathbf{x}$ at lag k , $R_{\mathbf{x}}(k) := \mathbb{E}[\check{\mathbf{x}}(k)\check{\mathbf{x}}^T(0)]$ satisfies $\|R_{\mathbf{x}}(k)\| \leq C\rho^{-|k|}$, for some positive constants $C, \rho \in \mathbb{R}$, $\rho > 1$.*

In the remaining chapter, following these assumptions, our interest will be limited to the following family of DDAGs and corresponding LDS.

Definition 4.2.1. $\mathcal{H}_{p,q}(\beta, \sigma, M)$ denotes the family of LDS given by (4.3) such that the corresponding DDAG, $G(V, E) \in \mathcal{G}_{p,q}$ (p nodes with each node having a maximum q parents), with $|\mathbf{H}_{ij}(\omega)| \geq \beta$, $\forall (i, j) \in E, \omega$, $\Phi_{\mathbf{e}}(\omega) = \sigma(\omega)\mathbf{I}$, and $M^{-1} \leq \lambda_{\min}(\Phi_{\mathbf{x}}(\omega)) \leq \lambda_{\max}(\Phi_{\mathbf{x}}(\omega)) \leq M$, $\forall \omega \in \Omega$

Sampling Strategy for LDS states: We consider two sampling settings (see Fig. 4.2 for details):

- i) **restart and record:** The sampling is performed as follows: start recording and stop it after N measurements. For the next trajectory, the procedure is restarted with an independent realization and record for another epoch of N samples; repeat the process another $n - 2$ times providing n i.i.d trajectories of N samples each.
- ii) **continuous sampling:** Here, a single trajectory of length $n \times N$ is taken consecutively. Then, the observations are divided into n segments, each having N consecutive samples.

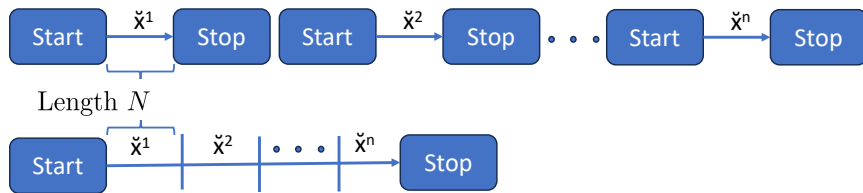


Figure 4.2: (a) shows restart and record sampling. (b) shows continuous sampling.

The data collected using either strategy is thus grouped into n trajectories of N samples each. For the finite length r^{th} trajectory, $\{\check{x}^r(t)\}_{t=0}^{N-1}$, define for each $\omega \in \Omega$,

$$\mathbf{x}^r(\omega) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \check{x}^r(k) e^{-i\omega k}, r = 1, \dots, n. \quad (4.6)$$

Note that $\mathbf{x}^r(\omega)$, if exists, is a zero mean random variable with the covariance matrix given by

$$\begin{aligned}\tilde{\Phi}_{\mathbf{x}}(\omega) &:= \mathbb{E} \{ \mathbf{x}^r(\omega) [\mathbf{x}^r(\omega)]^* \} \\ &= \frac{1}{N} \sum_{k=-(N-1)}^{N-1} (N - |k|) R_{\mathbf{x}}(k) e^{-i\omega k},\end{aligned}\tag{4.7}$$

$\forall \omega \in \Omega$. For the restart and record setting (unlike for continuous sampling), $\{\mathbf{x}^r(\omega)\}_{r=1}^n$ are i.i.d. Further, as $N \rightarrow \infty$, $\tilde{\Phi}_{\mathbf{x}}(\omega) \rightarrow \Phi_{\mathbf{x}}(\omega)$ uniformly in Ω [88].

4.3 Reconstructing DDAGs from PSDM

In this section, we discuss some results on how the PSDM, $\Phi_{\mathbf{x}}$, can be employed to completely reconstruct the DDAG, G when the time series is generated according to (4.3). Applying these results, inspired from the algorithm for static setting [96], we propose Algorithm 5 for reconstructing the DDAG, G . First, we prove that conditional PSD (CPSD) of i conditioned on C , defined as

$$f(i, C, \omega) := \Phi_{ii}(\omega) - \Phi_{iC}(\omega) \Phi_{CC}^{-1}(\omega) \Phi_{Ci}(\omega),\tag{4.8}$$

is a metric sufficient to obtain a topological ordering on the DDAG, G , which aids in the reconstruction of G . Notice that unlike [96]'s static setting, our algorithm will require the use of CPSD to unveil the dependencies in the DDAG and further affect the sample complexity of learning.

4.3.1 CPSD and Topological Ordering

Here, we will show that CPSD is the minimum for the nodes that have all the parents included in the conditioning set. We start by proving the result for the source nodes.

Lemma 4.3.1. *Consider the LDS described by (4.3). For any $\omega \in \Omega$, let $\alpha^* := \min_{k \in V} \Phi_{kk}(\omega)$. Then $\Phi_{ii}(\omega) = \alpha^*$ if and only if i is a source node.*

Proof. Let $\mathbf{T}(\omega) = (\mathbf{I} - \mathbf{H}(\omega))^{-1}$ and $\Phi_{\mathbf{e}}(\omega) = \sigma(\omega) \mathbf{I}_n$. Then, $\Phi_{\mathbf{x}}(\omega) = \sigma(\omega) \mathbf{T}(\omega) \mathbf{T}^*(\omega)$. By Cayley-Hamilton theorem, there exists constants $a_0(\omega), \dots, a_n(\omega)$ such that $(\mathbf{I} - \mathbf{H}(\omega))^{-1} = \sum_{k=0}^{n-1} a_k(\omega) (\mathbf{I} - \mathbf{H}(\omega))^k$. Using induction, it can be shown that non-diagonal entries of $(\mathbf{I} - \mathbf{H}(\omega))^k$ are zero if and only if there no k -hop path between i and j

(almost always). Similarly, (i, i) th entry is 1 if and only if there is no k -hop path between them (almost always).

Then (ignoring (ω)), $\Phi_{ii} = (\Phi_{\mathbf{x}})_{ii} = \sigma \sum_{k=1}^n \mathbf{T}_{ik} \mathbf{T}_{ik} = \sigma(\mathbf{T}_{ii}^2 + \sum_{k \neq i} \mathbf{T}_{ik}^2)$. If i is a source node, then $\mathbf{T}_{ik} = 0$ for every $k \neq i$ and $\mathbf{T}_{ii} = 1$, which implies $\Phi_{ii} = \sigma \mathbf{T}_{ii}^2$. For non-source nodes, $\mathbf{T}_{ik} \neq 0$ for some k , which gives $\Phi_{ii} > \sigma$. $\square$

Conditional PSD deficit

The following is the definition of conditional PSD deficit, which is helpful in proving the subsequent results and in retrieving the DDAG.

Definition 4.3.1 (The CPSD deficit).

$$\Delta := \min_{\omega \in [0, 2\pi]} \min_{j \in V} \min_{\substack{C \subseteq \text{nd}(j), \\ \text{Pa}(j) \setminus C \neq \emptyset}} f(j, C, \omega) - \sigma(\omega) \quad (4.9)$$

The following lemma shows that $f(i, C, \omega)$ from Eq. 4.8 can be used as a metric to obtain a topological ordering of G .

Lemma 4.3.2. *Consider the DDAG, G , governed by (4.3). Let $j \in V$ and let $C \subseteq V \setminus \{j\}$ be an ancestral set. Then for every $\omega \in \Omega$,*

$$\begin{aligned} f(j, C, \omega) &= \sigma(\omega) & : \text{ if } \text{Pa}(j) \subseteq C, \\ f(j, C, \omega) &\geq \Delta + \sigma(\omega) > \sigma(\omega) & : \text{ if } \text{Pa}(j) \not\subseteq C. \end{aligned}$$

The detailed proof is given in Appendix C.1. The proof uses the fact that conditioning on a superset of all parents makes node j conditionally independent of other ancestors and leaves only the exogenous noise at the node as the only component of its CPSD.

Corollary 4.3.1. *If $\text{Pa}(i) \subseteq C$ and $\text{Pa}(i) \not\subseteq D$, then $f(i, C, \omega) - f(i, D, \omega) \geq \Delta$.*

Lemma 4.3.3. $\Delta \geq \beta^2 \sigma$.

Proof. The proof follows from Lemma 4.3.2. Please see Appendix C.1 for details. $\square$

To determine the DDAG G 's structure, first determine a topological ordering of

nodes as follows: Beginning with an empty set $\mathcal{S}$, we iteratively add node i in $\mathcal{S}$ where

$$(i, C_i^*) \in \arg \min_{\substack{C \subseteq \mathcal{S}, |C| \leq q \\ 1 \leq j \leq p, j \notin \mathcal{S}}} f(j, C, \omega), \quad (4.10)$$

and f comes from (4.8). The following Lemma shows that $\mathcal{S}$ is a valid topological ordering w.r.t. G .

Lemma 4.3.4. *$\mathcal{S}$ is a valid topological ordering with respect to G .*

Proof. In the first step, $C = \emptyset$ and $f(i, C, \omega) = \Phi_{ii}$. By Lemma 4.3.1 and Lemma 4.3.2, $\Phi_{ii} = \sigma$ if i is a source node, where as $\Phi_{ii} \geq \sigma + \Delta$ if i is not a source node. Thus, the first node in the ordered set $\mathcal{S}$, $\mathcal{S}_1$, is always a source node. Induction assumption: Nodes $\mathcal{S}_1$ to $\mathcal{S}_n$ in $\mathcal{S}$ follow topological order. For the $n+1$, by Lemma 4.3.2, for every $C \subseteq \mathcal{S}$, $f(k, C, \omega)$ is minimum for $k \in V \setminus \mathcal{S}$ if and only if $Pa(k) \subseteq C$. Thus nodes $\mathcal{S}_1$ to $\mathcal{S}_{n+1}$ follow a topological order, which proves the result. $\square$

Identification of the parents: Parents of a node are identified from the ordered set by applying Corollary 4.3.1. Let $D = C \setminus \{k\}$. As shown in Corollary 4.3.1, if $Pa(i) \subseteq C$ and $k \in Pa(i)$, then $f(i, C, \omega) - f(i, D, \omega) \geq \Delta$. Thus, from the set $\mathcal{S}$, for every node $\mathcal{S}_i$, one can eliminate nodes $\mathcal{S}_1, \dots, \mathcal{S}_{i-1}$ by checking if the difference is greater than Δ . If the difference is greater than Δ for some $\mathcal{S}_k$, then $\mathcal{S}_k$ is a parent of $\mathcal{S}_i$. That is,

Lemma 4.3.5. *Let (i, C_i^*) be a solution of (4.10) and let*

$$P_i := \{j \in C_i^* \mid |f(i, C_i^*, \omega) - f(i, C_i^* \setminus j, \omega)| \geq \Delta\}.$$

Then, $Pa(i) = P_i$.

Applying the above procedure and Lemma 4.3.2-Lemma 4.3.5, one can formulate Algorithm 5 to obtain the ordering of the DDAG, G and eventually reconstruct the DDAG exactly. In Algorithm 5, $\hat{f}(i, C, \omega) := \hat{\Phi}_{ii}(\omega) - \hat{\Phi}_{iC}(\omega)\hat{\Phi}_{CC}^{-1}(\omega)\hat{\Phi}_{Ci}(\omega)$, an empirical estimate of $f(i, C, \omega)$, is employed instead of f and $\gamma = \Delta/2$. The following lemma proves that if the empirical estimate $\hat{f}(\cdot)$ is close enough to the original $f(\cdot)$, then Algorithm 5 reconstructs the DDAG exactly.

Lemma 4.3.6. *If $|\hat{f}(i, C, \omega) - f(i, C, \omega)| < \Delta/4$, for every i, ω, C , then Algorithm 5 reconstructs the DAG, G successfully. That is, $G = \hat{G}$.*

Algorithm 5 Reconstruction algorithm

Input: Estimated PSDM, $\widehat{\Phi}(\omega)$: Δ , $\omega \in [0, 2\pi]$

Output: $\widehat{G}$

1. Initialize the ordering, $\mathcal{S} \leftarrow ()$
 2. Repeat p times
 - (a) Compute $(j^*, C_j^*) \in \arg \min_{\substack{C \subseteq \mathcal{S}, |C| \leq q \\ 1 \leq j \leq p, j \notin \mathcal{S}}} \widehat{f}(j, C, \omega)$
 - (b) $\mathcal{S} \leftarrow (\mathcal{S}, j^*)$
 3. $\widehat{G} = (V, \widehat{E})$, $V \leftarrow \{1, \dots, p\}$, $\widehat{E} \leftarrow \emptyset$
 4. For $i = 1, \dots, p$
 - (a) $P_i := \left\{ j \in C_i^* \mid |\widehat{f}(\mathcal{S}_i, C_i^*, \omega) - \widehat{f}(\mathcal{S}_i, C_i^* \setminus j, \omega)| \geq \gamma \right\}$
 - (b) $\forall k \in P_i$, do $\widehat{E} \leftarrow \widehat{E} \cup \{(k, i)\}$
 5. Return $\widehat{G}$
-

Proof. See Appendix C.2 □

Therefore, it suffices to derive the conditions under which $|f(\cdot) - \widehat{f}(\cdot)| < \Delta/4$. In the following section, we derive a concentration bound to guarantee a small error, which is applied in obtaining the upper bound on the sample complexity of estimating G .

4.4 Finite Sample Analysis of Reconstructing DDAGs

In Lemma 4.3.6, it was shown that the DDAG, G , can be reconstructed exactly if the error in estimating $f(i, C, \omega)$ (given by (4.8)) is small enough. In this section, a concentration bound on the error in estimating $\Phi_{\mathbf{x}}$ from finite data is obtained, which is used later to obtain a concentration bound on the error in estimating the metric f .

Recall that we consider n state trajectories (see Fig. 4.3 for the two sampling strategies) with each trajectory being of length N samples, i.e. $\left\{ \{\check{\mathbf{x}}^r(k)\}_{k=0}^{N-1} \right\}_{r=1}^n$. For each trajectory, $\mathbf{x}^r(\omega)$, defined in (4.6), is a complex Gaussian with mean zero and covariance matrix, $\widetilde{\Phi}_x(\omega)$, i.e., for every $r = 1, \dots, n$, $\mathbf{x}^r(\omega) \sim \mathcal{N}(0, \widetilde{\Phi}_x(\omega))$, as given in Eq. 4.7. Increasing N ensures that $\widetilde{\Phi}_x$ is close to Φ_x . To estimate the PSDM, we thus rely on

the spectrogram method and estimate $\tilde{\Phi}_x(\omega)$ using finite n samples for $\mathbf{x}^r$.

4.4.1 Non-Asymptotic Estimation Error in Spectrogram Method

Let $\hat{\Phi}_x(\omega) := \frac{1}{n} \sum_{r=1}^n \mathbf{x}^r(\omega) [\mathbf{x}^r(\omega)]^*$. Let the estimation error in estimating $\Phi_{\mathbf{x}}$ by $\hat{\Phi}_x(\omega)$ be $Q := \hat{\Phi}_x(\omega) - \Phi_x(\omega)$.

Applying the triangle inequality, $\|Q\| \leq \|Q_{approx}\| + \|Q_2\|$, where $Q_{approx} := \tilde{\Phi}_x(\omega) - \Phi_x(\omega)$ and $Q_2 := \hat{\Phi}_x(\omega) - \tilde{\Phi}_x(\omega)$. Note that $\tilde{\Phi}_x(\omega)$ is the covariance of the DFT of each trajectory. To bound the estimation error, we bound both Q_{approx} and Q_2 .

The following Lemma shows that Q_{approx} is small if N (length of each trajectory) is large.

Lemma 4.4.1 (Lemma 5.1, [112]). *Consider an LDS given by (4.3) that satisfies Assumption 4.2.2. Let $Q_{approx} = \tilde{\Phi}_x(\omega) - \Phi_x(\omega)$ where $\tilde{\Phi}_x(\omega)$ is given in Eq. 4.7. Then $\|Q_{approx}\| < \varepsilon_1$ if $N > \frac{2C\rho^{-1}}{(1-\rho^{-1})^2\varepsilon_1}$.*

The next step is to characterize Q_2 , the error in estimating $\tilde{\Phi}_{\mathbf{x}}$ using $\hat{\Phi}_x(\omega)$. Since $\mathbb{E}(\mathbf{x}^r(\omega) [\mathbf{x}^r(\omega)]^*) = \tilde{\Phi}_x(\omega)$, $\hat{\Phi}_x(\omega)$ is an unbiased estimator of $\tilde{\Phi}_x(\omega)$. The following theorem provides a concentration bound on $\|Q_2\|$. The concentration bound is applicable under two sampling scenarios; the restart and record setting and continuous sampling setting, as shown in Fig. 4.3.

Theorem 4.4.1. *Suppose $\{\tilde{x}^r(k)\}_{k=0}^{N-1}$, $1 \leq r \leq n$ be the time series measurements obtained from an LDS governed by (4.3), satisfying Assumption 4.2.1. Then $\forall \omega \in \Omega$,*

$$\mathcal{P} \left(\left\| \hat{\Phi}_x(\omega) - \tilde{\Phi}_x(\omega) \right\| \geq \epsilon \right) \leq \exp \left(-\frac{\epsilon^2 n}{128M^2} + 6p \right). \quad (4.11)$$

The detailed proof is provided in Appendix C.3. We provide a concise proof sketch below.

Proof sketch of Theorem 4.4.1: The first step is to upper bound the spectral norm using a finite δ -cover of unit sphere $\mathbb{S}^p$, that is,

$$\|Q\| := \sup_{v \in \mathbb{S}^p} v^* Q v \leq 2 \max_{j=1, \dots, m} |(v^j)^* Q v^j|,$$

where v^j denotes the center of the j th sphere in the finite spherical cover. We then upper bound the moment-generating function, $\mathbb{E} [e^{\lambda \|Q\|}] \leq \sum_{j=1}^m \mathbb{E} [e^{2\lambda (v^j)^* Q v^j}] + \mathbb{E} [e^{-2\lambda (v^j)^* Q v^j}]$. The upper bound on $\mathbb{E} [e^{-2\lambda (v^j)^* Q v^j}]$ is divided into two parts. We first show the proof

for the restart and record sampling by exploiting the properties of i.i.d. nature of the DFT of the samples. Next, we generalize the results to the continuous sampling settings. The proofs involve using the Rademacher random variable based symmetrization argument followed by a series of standard inequalities. Finally, Chernoff bounding technique is applied to the upper bound of the moment-generating function upper bound. $\square$

By combining Lemma 4.4.1 and Theorem 4.4.1, the following corollary is obtained, which gives a concentration bound on the estimation error, $\|Q\|$.

Corollary 4.4.1. *Consider an LDS governed by (4.3) that satisfies Assumptions 4.2.1 and 4.2.2. Let $\{\check{x}^r(k)\}_{k=0}^{N-1}$, $1 \leq r \leq n$ be the time series measurements obtained for the LDS. Suppose that $N > \frac{2C\rho^{-1}}{(1-\rho^{-1})^2\varepsilon_1}$, where $0 < \varepsilon_1$. Let $0 < \varepsilon_1, \varepsilon_2 < \varepsilon$ be such that $\varepsilon_2 = \varepsilon - \varepsilon_1$. Then $\forall \omega \in \Omega$,*

$$\mathcal{P}\left(\left\|\Phi_x(\omega) - \hat{\Phi}_x(\omega)\right\| \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon_2^2 n}{128M^2} + 6p\right). \quad (4.12)$$

4.4.2 Sample Complexity Bounds: Upper Bound

In the previous subsection, concentration bounds on the estimation errors in PSDM were obtained. Here, a concentration bound on the error in estimating f is obtained, which is used to obtain a concentration bound in reconstructing the DDAG, G . The following result provides a concentration bound on $|f - \hat{f}|$.

Lemma 4.4.2. *Consider an LDS governed by (4.3) that satisfies Assumptions 4.2.1 and 4.2.2. Let $\{\check{x}^r(k)\}_{k=0}^{N-1}$, $1 \leq r \leq n$ be the time series measurements obtained for the LDS. Suppose that $N > \frac{2C\rho^{-1}}{(1-\rho^{-1})^2\varepsilon_1}$, where $0 < \varepsilon_1$. Let $0 < \varepsilon_1, \varepsilon_2 < \varepsilon$ be such that $\varepsilon_2 = \varepsilon - \varepsilon_1$. Then there exists a $c_0 \in \mathbb{R}$ such that, for any $\omega \in \Omega$,*

$$\mathcal{P}\left(|f(i, C, \omega) - \hat{f}(i, C, \omega)| \geq \varepsilon\right) \leq c_0 e^{\left(-\frac{\varepsilon_2^2 n}{10368M^6} + 6(q+1)\right)},$$

where q is the maximum number of parents any node has in G .

The proof, provided in Appendix C.4, uses Corollary 4.4.1 for different entries in the formula for $f(i, C, \omega)$ as defined in (4.8). Based on Lemma 4.4.2, the following upper bound on the probability of error in estimating G can be obtained.

Theorem 4.4.2. *Suppose $q \leq p/2$. Consider an LDS that belongs to $\mathcal{H}_{p,q}(\beta, \sigma, M)$ (Definition 4.2.1) that satisfies Assumptions 4.2.1 and 4.2.2. Let $\{\check{x}^r(k)\}_{k=0}^{N-1}$, $1 \leq r \leq n$*

be the time series measurements of the LDS and let $\hat{G}$ be the DDAG reconstructed by Algorithm 5. Suppose that $N > \frac{2C\rho^{-1}}{(1-\rho^{-1})^2\varepsilon_1}$, where $0 < \varepsilon_1 < \Delta/4$. Let $0 < \varepsilon_1, \varepsilon_2 < \varepsilon < \Delta/4$ be such that $\varepsilon_2 = \varepsilon - \varepsilon_1$. Then $\mathcal{P}\left(G(\omega) \neq \hat{G}(\omega)\right) \leq \delta$ if

$$n \gtrsim \frac{M^6 (q \log(p/q) - \log \delta)}{\epsilon_2^2}.$$

Proof. Applying the bound in Lemma 4.4.2,

$$\begin{aligned} \mathcal{P}\left(G(\omega) \neq \hat{G}(\omega)\right) &= \mathcal{P}\left(\bigcup_{k \in V, C \subseteq V \setminus \{k\}, |C| \leq q} \left\{|f(i, C, \omega) - \hat{f}(i, C, \omega)| > \epsilon\right\}\right) \\ &\stackrel{(a)}{\leq} \sum_{k \in V, C \subseteq V \setminus \{k\}, |C| \leq q} \mathcal{P}\left(\left\{|f(i, C, \omega) - \hat{f}(i, C, \omega)| > \epsilon\right\}\right) \\ &\stackrel{(b)}{\leq} p \left[\binom{p-1}{1} + \dots + \binom{p-1}{q} \right] c_0 e^{\left(-\frac{\epsilon^2 n}{10368 M^6} + 6(q+1)\right)} \\ &\stackrel{(c)}{\lesssim} c_1 p \times q \times (p/q)^q e^{\left(-\frac{\epsilon^2 n}{10368 M^6} + 6(q+1)\right)} \\ &\approx \exp\left(\log p + \log q + q \log\left(\frac{p}{q}\right) + \left(-\frac{\epsilon^2 n}{10368 M^6} + 6(q+1)\right)\right) \\ &\approx \exp\left(\log p + \log q + q \log\left(\frac{p}{q}\right) + 6q - \frac{\epsilon^2 n}{M^6}\right) \\ &\lesssim \exp\left(q \log\left(\frac{p}{q}\right) - \frac{\epsilon^2 n}{M^6}\right) < \delta, \end{aligned}$$

where (a) follows by union bound, (b) follows since $|V| = p$ and there are $\binom{p-1}{k}$ number of combinations with $|C| = k$. (c) follows by applying Stirling's approximation, $\binom{n}{k} \leq (ne/k)^k$. Thus, $n \gtrsim \frac{M^6(q \log(p/q) - \log \delta)}{\epsilon^2}$. By selecting the threshold γ in the algorithm appropriately, we can get a sample complexity $n \gtrsim \frac{M^6 q \log(p/q)}{\Delta^2}$. $\square$

In the following section, a matching lower bound is derived.

4.5 Sample Complexity Bounds: Lower Bound

The lower bound for reconstructing DDAG is derived using information-theoretic techniques, in particular Fano's inequality, and by restricting the interested family of graphs to a finite set. Notice that, except for a couple of non-trivial facts in dynamic setup, this

is a direct extension of the lower bound for the static case provided in [92]. The approach is to construct restricted ensembles of graphical models and then to lower bound the probability of error using Generalized Fano's inequality. Theorem 4.5.1 below provides the lower bound.

Theorem 4.5.1. *Suppose $q \leq p/2$. If*

$$n \leq (1 - 2\delta) \max \left(\frac{\log p}{2\beta^2 + \beta^4}, \frac{q \log(p/q)}{M^2 - 1} \right),$$

then

$$\inf_{\hat{G}} \max_{H \in \mathcal{H}_{p,q}(\beta, \sigma, M)} \mathcal{P}\{(G(H) \neq \hat{G})\} \geq \delta,$$

Proof. The proof is provided in Appendix C.5 □

That is, if $n \leq (1 - 2\delta) \max \left(\frac{\log p}{2\beta^2 + \beta^4}, \frac{q \log(p/q)}{M^2 - 1} \right)$, then any given estimator fails to reconstruct the DDAG with probability greater than δ . The lower-bound provides the fundamental limit on reconstructing the DDAG from finite samples.

Theorem 4.5.1 provides the lower bound $\Omega \left(\frac{\log p}{2\beta^2 + \beta^4} \vee \frac{q \log(p/q)}{M^2 - 1} \right)$. Notice that the upper bound in Theorem 4.4.2 is $O(q \log \frac{p}{q})$. Thus we obtain a matching order bound when the lower bound is dominated by the second term.

4.6 Simulation Experiments

The effectiveness of Algorithm 5 is demonstrated with the help of a simulation study here. In the simulations, we fix the number of nodes, p , and the number of parents, q as in [92]. The exogenous noise PSD is fixed to be the same for all the nodes. To generate the random DAGs, first obtain a random permutation of $[p]$ nodes to obtain an ordering τ . Then for each $j \in [p]$, randomly select $\min(q, j - 1)$ number of parents from $\tau_{[1:j-1]}$. Then for each edge (i, j) , generate $B_{ij} \sim \epsilon \times \text{Unif}(.5, 1)$, where ϵ is the Rademacher random variable. The matrix B is then scaled by $\frac{1}{1.002\|B\|}$ to satisfy Assumption 4.2.2. The data is generated according to the model

$$x_i(t) = \sum_{j \neq i} B_{ij} x_j(t - 1) + e_i(t), \forall i \in [p], t \in [T]. \quad (4.13)$$

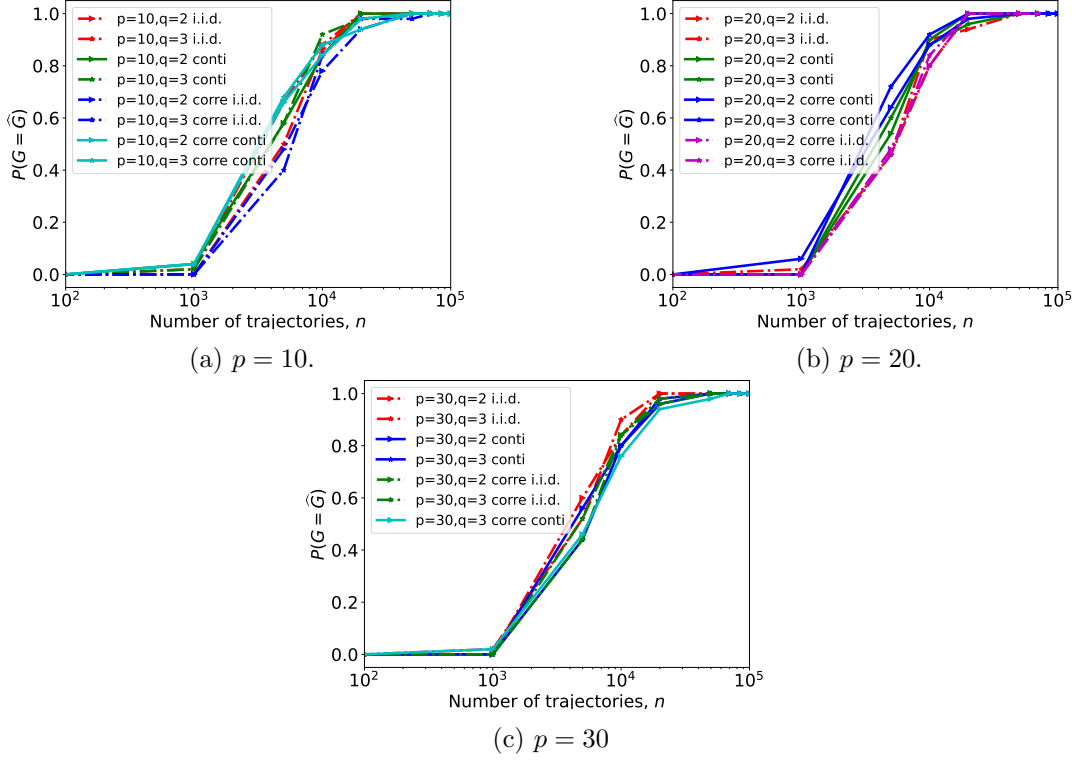


Figure 4.3: Probability of error with number of trajectories, for different networks under restart and record (RR) and continuous (conti) sampling strategies, when system is excited by either i.i.d. or WSS noise. p and q refer to the number of nodes and degree of each node in the network, respectively. The trajectory length $N = 64$ is fixed for each trajectory, and $\omega = \frac{17}{64}$.

We consider two scenarios for $e_i(t)$, viz 1) zero mean i.i.d. Gaussian noise with variance $\sigma = .5$, and 2) temporally correlated WSS Gaussian noise. The correlated WSS noise is generated according to the model $\mathbf{e}(t) = \alpha \mathbf{e}(t-1) + \mathbf{w}(t)$, for every $t \in [T]$, with $\alpha = 0.5$ and $\mathbf{w}(t)$ i.i.d. Gaussian noise with covariance matrix $\sigma \mathbf{I}$. The total number of samples, $T = N \times n$, where N is the length of each trajectory and n is the number of trajectories. We perform both restart and record and continuous sampling, as detailed in Fig. 4.3.

For each p and q , we generated 50 random networks and time series data according to the model above. For each set of parameters, n , T and N , we applied Algorithm 5 along with the estimator from Section 4.4 for $\omega = \frac{17}{64}$ to reconstruct the graph. Then, $\mathcal{P}(G = \hat{G})$ is empirically given by $\#(G = \hat{G})/50$. Fig. 4.3 show the comparison for

various test parameters. The simulations are performed using Python in a Macbook with M1 pro chip and 16 GB RAM.

Conclusion

In this chapter, we characterized the optimal sample complexity for structure identification with directions in linear dynamical networks. Inspired by the static setting, a metric, and an algorithm were proposed based on the power spectral density matrix to exactly reconstruct the DAG. It is shown that the optimal sample complexity is $n = \Theta(q \log(p/q))$. For the upper bound characterization, we obtained a tight concentration bound for the power spectral density matrix. An information-theoretic min-max lower bound also was provided for (sub) Gaussian linear dynamical systems. It was shown that the upper bound is order optimal with respect to the lower bound.

Chapter 5

Causal Structure Recovery of Linear Dynamical Systems: An FFT based Approach

Learning causal effects from data is a fundamental and well-studied problem across science, especially when the cause-effect relationship is static in nature. However, causal effect is less explored when there are dynamical dependencies, i.e., when dependencies exist between entities across time. The conventional causal structure recovery algorithms employ the tests such as the Fischer-z test and chi-square test to assess the conditional independence (CI). We show that employing Wiener Filter (WF) to assess CI provides better performance when compared to the static CI tests. The conventional approach of WF computation is computationally expensive. Exploiting the computation speed of FFT, we report a computationally efficient algorithm to compute the WF in the frequency domain. Applying the WF based CI test, we provide a modified version of the classical PC algorithm, called Wiener-PC, to reconstruct the Markov Equivalence Graphs (MEG) in the time-series with dynamical interactions. Moreover, for a large class of problems, we propose a novel computationally efficient algorithm for MEG reconstruction, called Wiener-Phase, which exploits phase properties of the complex frequency dependent WF. Concentration bounds are utilized to obtain non-asymptotic estimates of how well the Wiener coefficients can be determined, delineating the effect of data-size and accuracy of the estimates.

5.1 Introduction and literature survey

Causal identification from data is an active and important research area relevant to multiple domains including climate science [118], economics [70], neuroscience [119], and biology [120]. There is a considerable prior art on causation especially when the interactions are static in nature (see [48, 121, 46] and the references therein). In the case when entities can be modeled as random variables with an underlying joint probability distribution, without the possibility of actively intervening in the system, it is not possible in general to reconstruct the causal graph from observational data alone. Indeed, Markov equivalent graphs that capture the same set of conditional independence (CI) relations can be determined from data [46]. Some recent works on recovering a unique causal graph when the underlying data generative system is assumed to have more structure is presented in [51, 101]. Considerable prior art and algorithms exist for recovering the causal graph of entities modeled as random variables, when interventions on the system are possible [46]. In a related problem of estimating the causal effect given the causal graph, the goal is to estimate the effect of a variable on another. Here, a powerful framework of do-calculus can be employed to ascertain the set of variables that need to be used to determine the total and direct effect of a variable on another which result in criteria such as the single-door, back-door and front-door criteria [47, 48]

Causal inference is more challenging in the presence of dynamical (across time) dependencies [49, 50, 51]. One of the established techniques for identifying the causation is Granger causality [52], which utilizes how well a set of variables can be employed to estimate a given variable while utilizing the time-arrow. Granger causality based approaches necessitate presence of delays in the dynamic dependencies. Such an assumption is rendered problematic when the data is collected at slower rates than the time-constants at which the dynamics evolve, thus precluding a number of practical scenarios. Another approach in handling dependencies across times is to consider the series at every time instant as a random variable thereby mapping the problem to a static version; the difficulties with such an approach stem from the lack of information on the size of the horizon in the past and future to be considered and the combinatorial explosion of the number of variables that entail [53, 54]. In a class of approaches for unveiling dynamic dependencies, models that include vector auto regressive (VAR) models [50, 51], additive noise models (ANM) [49, 51] and neural network based models (NNM) [55], of how the data is being generated is assumed and the causal graph structure is recovered via an estimation of model parameters. In another recent line of work, the main

gist of the approach is to exploit the asymmetry in the relationships of a dependency and the inverse of the dependency. These asymmetries can result from nonlinear maps or from linear filters with inverses that have discernable differences that, for example, are characterizable via power spectral densities [56, 57]. Here, it is to be noted that works such as [56, 57] try to infer cause and effect relationship between two variables; however, the causal graph structure recovery entails determining how the influence flows from a variable to another with the possibility of intermediate variables between the cause and the effected variables. We emphasize that effective methodologies for causal inference for systems with dynamic interactions remain much less developed than their static dependency counterparts.

Several recent works have bridged the causality literature ([46]) with multivariate projections of dynamically related time-series data [49, 58, 27]. [27] extended the single-door criterion [48] to include cycles and proposed a similar criterion called revolving-door criterion. It was shown in [17] that the Wiener projections can be employed to recover the moral graph of entities interacting via linear time-invariant dynamics.

The main contributions of the chapter are summarized below. This chapter develops a framework for a system with dynamical interactions using a frequency domain approach. Here, the main steps taken for reconstruction of the moral graph (which recovers the kin structure) are: transform the time-series of each entity (modeled as a stochastic process) to the frequency domain using Fast Fourier Transform (FFTs), perform multivariate projections in the *frequency domain*, and construct an undirected graph using the non-zero entries of the obtained multivariate projection filter. Under the assumption that the data generative model is linear and time-invariant it is shown that the above process recovers the moral graph. With the standard assumption of faithfulness [46], a process akin to the Peter-Clark (PC) algorithm is employed to recover the essential graph with better computational efficiency, which exploits the phase properties of the complex valued estimates. When interventions are possible, it is shown that the causal graph can be recovered. It is demonstrated that the computational complexity for recovering the causal structure using the projections in frequency domain is $O(Tn^3 \log N)$ compared to the conventional projections using time-domain which are shown to have a complexity of $O(Tn^3 N^2)$; n , T and N are the number of nodes, number of samples, and the largest time-lag in the dependency between entities, respectively. Two notions of independence for stochastic processes are developed with guarantees on the convergence of the FFT model to the continuous frequency model provided. These notions help generalize and streamline the framework based on joint

probability distributions in the static case to the dynamic case with the mechanics of do-calculus made possible in the dynamics case. For estimating Wiener coefficients from time-series data, a non-asymptotic concentration bound and a sample complexity analysis are also provided. Furthermore, for causal-effect estimation for LTI dependencies, back-door, single-door (with cycles) and front door criteria are obtained with proofs that are rendered straightforward in the frequency domain.

Notations: Capital letters, X , denote either random variables or sets, the usage will be clear from the context; bold letters denote vectors or matrices. For any time signal, $x := \{x(t) : t \in \mathbb{Z}\}$, $\hat{x}$ denotes the Fourier transform of x , $\hat{x}(e^{j\omega}) := \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}$; for stochastic processes, $\mathbf{X}$, $\Phi_{\mathbf{X}}(e^{j\omega})$ denotes the power spectral density (PSD) of $\mathbf{X}$; $\mathbb{R}, \mathbb{C}, \mathbb{Z}$ denote the set of real numbers, complex numbers and integers respectively; $[N] := \{0, \dots, N-1\}$; for a complex number, $x = x_R + ix_I$, $\angle x$ denotes $\arctan(x_I/x_R)$; Ω denotes the interval $[0, 2\pi]$ and Ω_N denotes $\{0, 2\pi/N, \dots, 2\pi(N-1)/N\}$; for a transfer function, $\hat{x}$, $\hat{x} \neq 0$ means $\hat{x}$ is not identically zero; $\|\hat{x}\|_{\ell_{\infty}(\Omega_N)}$ denotes $\sup_{\omega \in \Omega_N} \hat{x}(\omega)$; in a given directed graph, $Pa(i)$, $Ch(i)$, $Sp(i)$ denotes the set of parents, children, and spouses of node i respectively. ℓ^1 denotes the space of absolutely summable sequences.

5.2 Preliminaries

5.2.1 Linear Dynamical Influence Models (LDIM)

Consider a network with n nodes, each node $i \in \{1, \dots, n\}$ having the time-series measurements, x_i , governed by the linear dynamical influence model (LDIM),

$$\hat{\mathbf{X}}(e^{j\omega}) = \hat{\mathbf{H}}(e^{j\omega})\hat{\mathbf{X}}(e^{j\omega}) + \hat{\mathbf{E}}(e^{j\omega}), \quad \forall \omega \in [0, 2\pi], \quad (5.1)$$

where $\hat{\mathbf{X}} = [\hat{X}_1, \dots, \hat{X}_n]^T$, and $\hat{\mathbf{E}} = [\hat{E}_1, \dots, \hat{E}_n]^T$ are the exogenous noise source with E_i, E_j jointly wide sense stationary for $i \neq j$. $\hat{\mathbf{H}}$ is a well posed transfer function, i.e., every submatrix of $(\mathbf{I} - \hat{\mathbf{H}}(e^{j\omega}))$ is invertible as well as every entry of $(\mathbf{I} - \hat{\mathbf{H}}(e^{j\omega}))^{-1}$ is analytic, and $\hat{\mathbf{H}}_{ii}(e^{j\omega}) = 0$ for every $\omega \in \Omega$. $\Phi_{\mathbf{E}}(e^{j\omega})$ is positive definite and diagonal for every ω . For notational convenience, $\hat{x}(e^{j\omega})$ is replaced with $\hat{x}(\omega)$ henceforth. Notice that the LDIM (5.1) can be represented in time-domain using the following linear time invariant model,

$$\mathbf{X}(k) = \sum_{l=-\infty}^{\infty} \mathbf{H}(l)\mathbf{X}(k-l) + \mathbf{E}(k), \quad (5.2)$$

where $\mathbf{E}(k) = [E_1(k), \dots, E_n(k)]^T$, and $\mathbf{X}(k) = [X_1(k), \dots, X_n(k)]^T$. The LDIM (5.1) can be represented using a graph $G = (V, \vec{\mathcal{E}})$, where $V = \{1, \dots, n\}$ and $\vec{\mathcal{E}} = \{(u, v) : H_{vu} \neq 0\}$, that is, there exists an edge $u \rightarrow v$ in $\vec{\mathcal{E}}$ if $\widehat{H}_{vu}(\omega) \neq 0$ for some $\omega \in [0, 2\pi]$. u is said to be a kin of v in G , denoted, $u \in \text{kin}_G(v)$, if at least one of the following exist in $\vec{\mathcal{E}}$: $u \rightarrow v$, $v \rightarrow u$, $u \rightarrow i \leftarrow v$ for some $i \in V$. $\text{kin}_G(j)$ is the set of kins of j in G .

5.2.2 Frequency Discretization

The LDIM in (5.1) can be sampled at frequencies ω in $\Omega_N = \{\omega_0, \dots, \omega_{N-1}\}$, where $\omega_k := \frac{2\pi k}{N}$, thus converting the continuous frequency model to practical model involving finite set of frequencies. In time-domain, the sampled set of relations of (5.1) is equivalent to the finite impulse response model described by,

$$\mathbf{X}(n) = \sum_{l=0}^{N-1} \mathbf{H}(l) \mathbf{X}(n-l)_{\text{mod } N} + \mathbf{E}(n), \quad n = 0, \dots, N-1, \quad (5.3)$$

where $\mathbf{X}(n)$ and $\mathbf{E}(n)$ are periodic with period N and $\mathbf{H}(l)$ is non-zero for at most $N-1$ lags. Let $\widehat{\mathbf{X}}(e^{\frac{j2\pi k}{N}}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \mathbf{X}(n) e^{-2\pi kn/N}$, $k = 0, \dots, N-1$ which provides the Discrete Fourier Transform (DFT) of the sequence $\mathbf{X}(n)$. Similar expressions hold for $\mathbf{H}(k)$ and $\mathbf{E}(k)$ and their corresponding DFTs. The directed graph associated with (5.3) is, $\check{G} = (V, \check{\mathcal{E}})$ is defined such that the edge $u \rightarrow v$ exists in $\check{\mathcal{E}}$ if $\widehat{H}_{vu}(e^{\frac{j2\pi k}{N}}) \neq 0$ for some k . A pertinent question here is: when and under what conditions (5.3) converges to (5.2). It is well known that the Fourier coefficients can be approximated via discretization. The following result, with proof included for completeness in the supplementary material, gives a uniform and explicit convergence rate for signals with bounded variation.

Theorem 5.2.1. *For a function $\widehat{f}$ of bounded variation V , on $[0, 2\pi]$ and $N \geq 1$, the estimation of $f(n) = \int_0^{2\pi} \widehat{f}(\omega) e^{j\omega n} d\omega$ given by $f^{(N)}(n) = \frac{1}{N} \sum_{k=0}^{N-1} e^{j2\pi nk/N} f(2\pi k/N)$ for $|n| \leq \sqrt{N}$ and zero otherwise, satisfies $\|f - \widehat{f}^{(N)}\|_{\ell_\infty(\Omega)} \leq C/\sqrt{N}$.*

Proof. See Appendix D.3 □

5.2.3 Graph Definitions

Consider a directed graph $G = (V, \vec{\mathcal{E}})$, where $V = \{1, \dots, n\}$. A *chain* from node i to node j is an ordered sequence of edges in $\vec{\mathcal{E}}$, $((\ell_0, \ell_1), (\ell_1, \ell_2), \dots, (\ell_{n-1}, \ell_n))$, where $\ell_0 = i$, $\ell_n = j$, and $(\ell_i, \ell_{i+1}) \in \vec{\mathcal{E}}$. Topology, $\text{top}(G)$, of a directed graph, $G = (V, \vec{\mathcal{E}})$, is an undirected graph, $\text{top}(G) = (V, \mathcal{E})$ where an edge $(u, v) \in \mathcal{E}$ if either $(u, v) \in \vec{\mathcal{E}}$ or

$(v, u) \in \vec{\mathcal{E}}$. A *path* from node i to node j in the directed graph G is an ordered set of edges $((\ell_0, \ell_1), (\ell_1, \ell_2), \dots, (\ell_{n-1}, \ell_n))$ with $\ell_0 = i$, $\ell_n = j$ in its topology, $(V, \mathcal{E})$, where $\{\ell_i, \ell_{i+1}\} \in \mathcal{E}$.

A path of the form $((\ell_0, \ell_1), \dots, (\ell_{n-1}, \ell_n))$ has a collider at ℓ_k if $\ell_k \in Ch(\ell_{k-1}) \cap Ch(\ell_{k+1})$. That is, there exists directed edges $\ell_{k-1} \rightarrow \ell_k \leftarrow \ell_{k+1}$ in G . Consider disjoint sets $X, Y, Z \subset V$ in a directed graph $G = (V, \vec{\mathcal{E}})$. Then, X and Y are ***d-separated*** given Z in G , denoted $d\text{-sep}_G(X, Z, Y)$ if and only if every path between $x \in X$ and $y \in Y$ satisfies at least one of the following: 1) The path contains a non-collider node $z \in Z$. 2) The path contains a collider node w such that neither w nor the descendants of w are present in Z . In-nodes of i is defined as $\{j : (i, j) \in \vec{\mathcal{E}}\}$.

5.2.4 Graph/Topology Learning

We first present a method for estimating the i^{th} time-series from another set of time-series. It is well known that the optimal estimate, $\tilde{X}_i$, (in the minimum mean square sense) of the process X_i from the rest of the time-series in $\bar{i} = V \setminus \{i\}$, satisfies $\hat{\tilde{X}}_i(\omega) = W_{i, \bar{i}}(\omega) \hat{X}_{\bar{i}}(\omega)$ where $W_{i, \bar{i}}(\omega) = \Phi_{\bar{i}, \bar{i}}^{-1}(\omega) \Phi_{i, \bar{i}}(\omega)$, with $\Phi_{i, j}(\omega) := \sum_{k=-\infty}^{\infty} R_{ij}(k) e^{-j\omega k}$, $R_{ij}(k) := \mathbb{E}[X_i(k) X_j^T(0)]$. Here $W_{i, C}(\omega)$ is multivariate Wiener filter obtained by projecting X_i onto the set of processes in C [122]. The entry of $W_{i, C}$ corresponding to j^{th} node in the set C , is denoted by $W_{i, C}[j]$. For notational convenience, the Wiener coefficient (vector of size $|\bar{i}|$) in the projection of X_i to $X_{\bar{i}}$ is denoted by W_i .

Lemma 5.2.1. *Consider a well-posed LDIM given by (5.1). Let $W_i := W_{i, \bar{i}}$ be the Wiener projection of $\mathbf{X}_i$ to $\mathbf{X}_{\bar{i}}$. Then, $W_i[j](\omega) \neq 0$ if and only if $i \in \text{kin}_G(j)$.*

Proof. See Appendix D.4

Assumption 5.2.1. *If a node k has multiple incoming edges in G , then for every pair of in-nodes i, j of k , $\angle \hat{H}_{ki} = \angle \hat{H}_{kj}$, where $i, j, k \in V$.*

Assumption 5.2.2. *If $\hat{H}_{ij}(\omega) \neq 0$ then $\Im\{\hat{H}_{ij}(\omega)\} \neq 0$, for every $i, j \in V$ and $\omega \in \Omega$.*

Lemma 5.2.2. [18] *Consider a well-posed LDIM given by (5.1) and satisfying Assumptions 5.2.1 and 5.2.2. Then for any $\omega \in \Omega$, $\Im\{W_i[j](\omega)\} \neq 0$ if and only if $(i, j) \in \vec{\mathcal{E}}$ or $(j, i) \in \vec{\mathcal{E}}$.*

We first briefly outline the *time-domain based approach* in determining the multivariate Wiener filter coefficients. Define $m = |C|$, $\mathbf{X}_i := [\mathbf{X}_i(T), \mathbf{X}_i(T-1), \dots, \mathbf{X}_i(N)]^T$ and $\mathbf{Y}_C := [\mathbf{X}_C(T), \mathbf{X}_C(T-1), \dots, \mathbf{X}_C(N)]^T \in \mathbb{R}^{(T-N+1) \times |C|N}$, where $\mathbf{X}_C(k) :=$

$[\mathbf{X}_{c_1}(k), \mathbf{X}_{c_1}(k-1), \dots, \mathbf{X}_{c_1}(k-N+1), \mathbf{X}_{c_2}(k), \mathbf{X}_{c_2}(k-1), \dots, \mathbf{X}_{c_2}(k-N+1), \dots, \mathbf{X}_{c_m}(k), \mathbf{X}_{c_m}(k-1), \dots, \mathbf{X}_{c_m}(k-N+1)] \in \mathbb{R}^{|C|N}$. Consider the least square formulation described by $\beta_T := \arg \min_{\beta \in \mathbb{R}^{|C|}} \frac{1}{T} \|\mathbf{X}_i - \mathbf{Y}_C \beta\|_2^2$ obtains the Wiener filter coefficients β_T for a finite time horizon of length T . Let

$$\widetilde{\mathbf{W}}_{i,C} := \arg \min_{\beta \in \mathbb{R}^{|C|}} \lim_{T \rightarrow \infty} \frac{1}{T} \|\mathbf{X}_i - \mathbf{Y}_C \beta\|_2^2. \quad (5.4)$$

Then, it can be shown that the Wiener filter $W_{i,C}$ coincides with $\widetilde{\mathbf{W}}_{i,C}$. Here β_T for sufficiently large T is employed as a representative for the Wiener filter $W_{i,C}$. Non-asymptotic concentration bounds and sample complexity results for estimating Wiener coefficient W_i from time-series data is provided in Section 5.4.2.

We now outline a process of determining multivariate Wiener filter by first transforming every time-series to its frequency domain representation followed by *projections in the frequency domain*. Consider the i^{th} time-series X_i which is partitioned in R segments with the r^{th} segment denoted by X_i^r . Each segment consists of N samples, for example, $X_i^r = \{X_i^r(0), \dots, X_i^r(N-1)\}$. Thus, the time series X_i is given by, $\{(X_i^r(t))_{t=0}^{N-1}\}_{r=1}^R$. Using the r^{th} segment of the X_i trajectory, given by $X_i^r(0), \dots, X_i^r(N-1)$, the FFT, $\widehat{\mathbf{X}}_i^r(e^{\frac{j2\pi k}{N}}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \mathbf{X}_i^r(n) e^{-2\pi k n/N}$ is computed. Let $C = \{c_1, c_2, \dots, c_m\} \subseteq V$ where V is the set of time-series. Let $\widehat{\mathbf{X}}_C^r$ be the vector $\begin{bmatrix} \widehat{\mathbf{X}}_{c_1}^r & \widehat{\mathbf{X}}_{c_2}^r \dots & \widehat{\mathbf{X}}_{c_m}^r \end{bmatrix}^T$ which is the vector obtained by stacking the Fourier coefficients obtained from the r^{th} segments of the time-series in the set C . Let $\mathcal{X}_C := [\widehat{\mathbf{X}}_C^1, \dots, \widehat{\mathbf{X}}_C^R]^T$, where $C \subseteq V$.

It can be shown that (see [112])

$$W_{i,C}(\omega) := \arg \min_{\beta \in \mathbb{C}^{|C|}} \lim_{N, R \rightarrow \infty} \frac{1}{2R} \|\mathcal{X}_i(\omega) - \mathcal{X}_C(\omega) \beta\|_2^2, \quad \omega \in [0, 2\pi]. \quad (5.5)$$

Here $\beta_{RN} := \arg \min_{\beta \in \mathbb{C}^{|C|}} \frac{1}{2R} \|\mathcal{X}_i(\omega) - \mathcal{X}_C(\omega) \beta\|_2^2$, $\omega \in [0, 2\pi]$ for sufficiently large RN is employed to represent the Wiener filter $W_{i,C}$.

5.2.5 Probabilistic Conditional Independence on a Set of Stochastic Processes and Factorization According to Directed Graphs

A framework for do-calculus for stochastic processes (SPs) requires a notion of independence in SPs and factorization according to conditional independence (CI). Towards this, we provide two notions of independence for SPs, the first one, motivated by the Kolmogorov extension theorem [123] is based on the finite dimensional distribution

(FDD) in time-domain. The second notion is defined in the Fourier domain. Using each of the two CI notions, one can define factorization of SPs according to graphs, perform do-calculus and derive the single-door, front-door and back-door methods for direct and total effect identification. The FDD based CI notions are developed in the supplementary material due to space constraint. Here, in the main chapter, we focus on the independence in frequency-domain, which requires existence of Fourier transform.

Conditional Independence on Stochastic Processes Interacting via LDIMs

An SP $X := \{X(t) : t \in I\}$, time indexed by $I = \mathbb{Z}$ (or $I = [N]$) on the probability space $(\Theta, \mathcal{F}, \mathcal{P})$ is a map from $\Theta \times I \rightarrow \mathbb{R}^I$. Here, $\mathbf{X}_i$ is used to denote the SP $\{X_i(t) : t \in I\}$. For any $\theta \in \Theta$, $t \in I$, $X(\theta, t)$ denotes an instance of X at time t . The set $X_\theta := \{X(\theta, t)\}_{t \in I}$ is the **sample path** corresponding to the event $\theta \in \Theta$. We assume that $X_\theta \in \ell^1$. For every $\theta \in \Theta$, the discrete time Fourier transform of the sample path, $\hat{X}_\theta : (\Theta, \mathcal{M}) \mapsto \mathbb{C}^\Omega$ is defined as $\hat{X}_\theta(\omega) := \sum_{t \in I} X(\theta, t)e^{-i\omega t}$, $\forall \omega \in \Omega$. Then, $\hat{X}(\omega) : \Theta \rightarrow \mathbb{C}$ is a random variable and it is possible to define the independence in the frequency-domain as follows.

A set of stochastic processes $\{X_i\}_{i=1}^m$ on $(\Theta, \mathcal{F}, \mathcal{P})$ are independent at $\omega \in \Omega$ if and only if

$$\mathcal{P} \left(\bigcap_{i=1}^m \left\{ \hat{X}_i(\omega) \in B_i \right\} \right) = \prod_{i=1}^m \mathcal{P} \left(\hat{X}_i(\omega) \in B_i \right).$$

Definition 5.2.1 (Conditional Independence). *Consider the stochastic processes X_1 , X_2 , and X_3 . Then X_1 is said to be conditionally independent of X_2 given X_3 at $\omega \in \Omega$ (denoted $X_1 \perp\!\!\!\perp^{(\omega)} X_2 \mid X_3$) if and only if*

$$\mathcal{P} \left(\bigcap_{i=1}^2 \left\{ \hat{X}_i(\omega) \in B_i \right\} \mid \hat{X}_3(\omega) \in B_3 \right) = \prod_{i=1}^2 \mathcal{P} \left(\hat{X}_i(\omega) \in B_i \mid \hat{X}_3(\omega) \in B_3 \right),$$

where $B_i \in \mathcal{B}^2$; $\mathcal{B}^2$ is the Borel sigma algebra.

In the LDIM (5.1), $\hat{\mathbf{X}}$, if it exists, is a linear transform of $\mathbf{X}$. Since FT-IFT pair is unique, the probability density functions (pdfs) are well-defined, if the pdf exists in time-domain. Then, $\hat{\mathbf{X}}(\omega)$ are random vectors with an associated probability measure $\mathbb{P}$ and a family of density functions $\{f^{(\omega)}\}_{\omega \in \Omega}$ on the set of nodes V .

We now define factorization at a specific frequency.

Definition 5.2.2 (Factorization of SPs according to a directed graph at a specific frequency). *A set V of stochastic processes is said to **factorize** according to a **directed** graph $G^{(\omega)} = (V, \vec{\mathcal{E}}^{(\omega)})$ at $\omega \in [0, 2\pi]$ if and only if the probability distribution at ω can be factorized according to the graph $G^{(\omega)}$, i.e.,*

$$\mathcal{P} \left(\hat{X}_1(\omega), \dots, \hat{X}_n(\omega) \right) = \prod_{i \in V} \mathcal{P} \left(\hat{X}_i(\omega) \mid \hat{X}_{Pa_i}(\omega) \right), \quad (5.6)$$

where Pa_i denotes the parents of i in $G^{(\omega)}$. If the pdf exists, then

$$f^{(\omega)}(\hat{x}_1(\omega), \dots, \hat{x}_n(\omega)) = \prod_{i=1}^n f^{(\omega)}(\hat{x}_i(\omega) \mid \hat{x}_{Pa_i}(\omega)). \quad (5.7)$$

Then, we can define a graph $G = (V, \vec{\mathcal{E}})$, where $V = \{1, \dots, n\}$ and $\vec{\mathcal{E}}$ is such that the edge $(i, j) \in \vec{\mathcal{E}}$ if and only if $(i, j) \in \vec{\mathcal{E}}^{(\omega)}$ for some $\omega \in \Omega$.

We now provide a definition of factorization of a set of stochastic processes.

Definition 5.2.3 (Conditional Independence–Frequency Domain (CIFD)). *Consider SPs $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$. Then $\mathbf{X}_1$ is said to be conditionally independent of $\mathbf{X}_2$ given $\mathbf{X}_3$ if and only if for almost all $\omega \in \Omega$,*

$$\mathcal{P} \left(\bigcap_{i=1}^2 \left\{ \hat{X}_i(\omega) \in B_i \right\} \mid \hat{X}_3(\omega) \in B_3 \right) = \prod_{i=1}^2 \mathcal{P} \left(\hat{X}_i(\omega) \in B_i \mid \hat{X}_3(\omega) \in B_3 \right),$$

where $B_i \in \mathcal{B}^2$.

Using this definition, one can define factorization of SPs according to graphs, which helps in performing do-calculus and causal identification.

Definition 5.2.4 (Factorization of stochastic process). *A set V of stochastic processes on $(\Theta, \mathcal{F})$ is said to factorize according to $G = (V, \vec{\mathcal{E}})$ if and only if*

$$\mathcal{P} \left(\hat{X}_1(\omega), \dots, \hat{X}_n(\omega) \right) = \prod_{i \in V} \mathcal{P} \left(\hat{X}_i(\omega) \mid \hat{X}_{Pa_i}(\omega) \right) \quad \text{for almost all } \omega \in \Omega, \quad (5.8)$$

where Pa_i denotes the parents of i in G .

5.3 Main Results

5.3.1 Computational Advantages of FFT based Wiener filter

The first contribution of the chapter delineates the computational complexity when the multivariate Wiener filters are obtained via time-domain projections in comparison to the computational complexity when the Wiener filter coefficients are determined via projection in the frequency domain. In Section 5.4 we show that the asymptotic computational complexity of determining the Wiener filter coefficient, W_i , using the time-domain approach (see equation (5.4)) is $O(Tn^2N^2)$ whereas when the Wiener filter is obtained via projections in the frequency domain (see equation (5.5)) it is $O(Tn^2 \log N)$; here each time series has T samples, N is the largest time lag, and n is the number of time-series.

5.3.2 Realizing the Essential Graph of SPs; using Phase Properties of Frequency Dependent Wiener Filtering

Mirroring the approach in the PC algorithm [124], we can apply the following approach to reconstruct the essential graph of dynamically related SPs. In the vanilla PC algorithm for reconstructing $G = (V, \vec{\mathcal{E}})$, the first step is to find the topology (also called skeleton) of the graph, $top(G)$ by performing, for every pair $i, j \in V$, conditional independence test on i and j conditioned on $C \subseteq V \setminus \{i, j\}$. If i and j are conditionally independent given C_{ij} for some $C_{ij} \subset V \setminus \{i, j\}$, then $(i, j) \notin top(G)$. Notice that C_{ij} is initialized by setting $C_{ij} = \{\}$, followed by the nodes with $|C_{ij}| = 1$, followed by $|C_{ij}| = 2$, and so on in the increasing order of size. Once the skeleton is retrieved, check for every $(i, j) \in top(G)$ and every k that is a common neighbor of i and j , if $k \in C_{ij}$. If $k \notin C_{ij}$, then k is a collider of the form $i \rightarrow k \leftarrow j$. By repeating this process for every nonadjacent pairs i, j and their common neighbors k , we can identify all the colliders in G . As shown in [110], multivariate Wiener filter can be employed to determine the CI structure in LDIMs. Let $G = (V, \vec{\mathcal{E}})$ be the graph representation of an LDIM (5.1). Then, $(i, j) \in top(G)$ if and only if $W_{i[j, C]}[j] = 0$. That is, if it holds that, in the projection of X_i on $X_{[j, C]}$ by (5.5), the coefficient of X_j is zero (i and j are Wiener separated by C). Then, we can employ the same steps in the vanilla PC algorithm, where conditional independence is tested using Wiener separation, to reconstruct the essential graph of G . The algorithm is called Wiener-PC (W-PC) algorithm here. We show in Section 5.4 that the computational complexity of W-PC is $O(RNn^{q+3} \log N)$,

where q is the maximum neighborhood size.

As shown in Lemma 5.2.2, in a large class of applications [1], support of the imaginary part of the frequency dependent Wiener filter retrieves $\text{top}(G)$ exactly. Thus, by analyzing the real and imaginary part of W_i separately, one can extract the strict spouse edges. This information can in turn be employed to identify the colliders in G in an efficient way (see Section 5.4). We call this algorithm Wiener-Phase here. The worst case asymptotic computational complexity of Wiener-Phase algorithm is $O(n^3(RN \log N + q^2))$ (see Section 5.4), which is advantageous, especially in highly connected graphs (with large q). We demonstrate the Wiener-phase algorithm in the simulations section.

5.3.3 Causal Inferences on Set of Stochastic Processes via Analysis of Dependencies at Specified Frequencies

Consider a set of SPs $\mathbf{X}$ on $(\Theta, \mathcal{F}, \mathcal{P})$. Let $A, B, C \subset V$ be disjoint sets. Then, as shown in Definition 5.2.3, A is independent of B given C , denoted $A \perp\!\!\!\perp B \mid C$, if and only if $A \perp\!\!\!\perp^{(\omega)} B \mid C$ for every $\omega \in \Omega$. The following lemma follows immediately.

Lemma 5.3.1. *Consider a set V of stochastic process $\mathbf{X}$ that factorizes according to a directed graph $G = (V, \vec{\mathcal{E}})$. Let $A, B, C \subset V$ be disjoint sets. If $\text{d-sep}_{G^{(\omega)}}(A, C, B)$ then $A \perp\!\!\!\perp^{(\omega)} B \mid C$ and if $\text{d-sep}_G(A, C, B)$, then $A \perp\!\!\!\perp B \mid C$. Further, if the distribution is faithful, then the converse also holds.*

The following lemma establishes a criterion which allows inference of conditional independence of stochastic processes $\{X_i\}_{i=1}^n$ from conditional independence of $\{\hat{X}_i(\omega)\}_{i=1}^n$ obtained by stochastic processes component at a specific frequency ω .

Lemma 5.3.2. *Consider a set V of stochastic process that factorizes according $G^{(\omega)} = (V, \vec{\mathcal{E}}^{(\omega)})$ at $\omega \in \Omega$. Let $G = (V, \vec{\mathcal{E}})$ be such that $(i, j) \in \vec{\mathcal{E}}$ if and only if $(i, j) \in \vec{\mathcal{E}}^{(\omega)}$ for some $\omega \in \Omega$. Let $X, Y, Z \subset V$ be disjoint sets. If Z d -separates X and Y in G , then Z d -separates X and Y in $G^{(\omega)}$.*

Proof. See Appendix D.4.

Much stronger results follow for LDIMs. The following result for LDIMs establishes the equivalence of d – separation on the graph G according to which the stochastic processes factor (according to Definition 5.2.4) and d – separation on the graph $G^{(\omega)}$ which factors the LDIM at a specific frequency $\omega \in \Omega$.

Lemma 5.3.3. *Consider a set V of stochastic process that is described by the LDIM (5.1) (here $\mathbf{E}_i(k) \perp\!\!\!\perp \mathbf{E}_j(l)$ for $i \neq j$ and $k, l \in \mathbb{Z}$). Moreover, suppose the set V of stochastic processes factorizes according to directed graph $G = (V, \vec{\mathcal{E}})$ and at a specific frequency $\omega \in \Omega$ factorizes according to a directed graph $G^{(\omega)} = (V, \vec{\mathcal{E}}^{(\omega)})$. Let $X, Y, Z \subset V$ be disjoint sets. Then Z d-separates X and Y in G if and only if Z d-separates X and Y in $G^{(\omega)}$, for almost all $\omega \in \Omega$.*

Proof. See Appendix D.4.3.

Lemma 5.3.4. *Consider a set V of stochastic process that is described by the LDIM (5.1) (here $\mathbf{E}_i(k) \perp\!\!\!\perp \mathbf{E}_j(l)$ for $i \neq j$ and $k, l \in \mathbb{Z}$). Moreover, suppose the set V of stochastic processes factorizes according to directed graph $G = (V, \vec{\mathcal{E}})$ and at a specific frequency $\omega \in \Omega$ factorizes according to a directed graph $G^{(\omega)} = (V, \vec{\mathcal{E}}^{(\omega)})$. Let $X, Y, Z \subset V$ be disjoint sets. Then $X \perp\!\!\!\perp^{(\omega)} Y \mid Z$ if and only if $X \perp\!\!\!\perp Y \mid Z$, for almost all $\omega \in \Omega$.*

Proof. See Appendix D.4.4

The CI and graph factorization machinery described here can be used in identifying the causal structure, and in estimating direct and total effect. We provide development of single-door, back-door and front-door criteria for causal effect identification in the supplementary material.

5.4 Complexity Analysis of Wiener filter

5.4.1 Computational Complexity of Estimating Wiener Filter

In the time-domain, consider projection of T samples $\{X_i(t)\}_{t=1}^T$ to past N values (including present) of $X_j(t)$, for every $t = 1, \dots, T$. The regression coefficients in the projection of $X_i(t)$ on the rest are $\{h_{ij}(t) : 0 \leq k \leq N-1, 1 \leq j \leq n, i \neq j\}$, i.e., $(n-1)N$ variables. The complexity of computing the least square for node i using (5.4) is $O(TN^2(n-1)^2) \approx O(TN^2n^2)$. Repeating this process n times, the total complexity for the projection of n time-series is $O(TN^2n^3)$.

If we compute FFT for a window size of N samples without overlap, then $N \log N$ computations are required to obtain FFT. Computing the Wiener co-efficient for a given ω (here the number of samples is T/N since we computed DFT with N samples) using the regression (5.5) takes $O(\frac{Tn^2}{N})$ computations, thus a total of $O(Tn^2 \log N)$ computations are required to compute the Wiener coefficient for a single frequency, in

contrast to $O(Tn^2N^2)$ in time-domain. Repeating this process n times give an overall complexity of $O(Tn^3 \log N)$ for our approach.

Wiener meets Peter-Clark

As shown in [110], Wiener filtering can identify the CI structure in linear dynamical models. Thus, combining it with PC algorithm gives W-PC algorithm. As shown before, computing W_i using FFT takes a complexity of $O(RNn^3 \log N)$. As shown in [124], PC algorithm requires $O(n^q)$ tests, thus giving an overall complexity of $O(RNn^{q+3} \log N)$, where q is the neighborhood size.

As shown in Section 5.3, Lemma 5.2.1 retrieves the Markov blanket structure in G and Lemma 5.2.2 returns the skeleton of G . Combining both, we propose Wiener-phase algorithm in the supplementary material to identify the colliders and thus the equivalence class of the graphs efficiently. $\mathcal{K}$ is the set of kin edges, $\mathcal{S}$ is the skeleton, and $\mathcal{SP}$ is the strict spouse edges. $\mathcal{C}_{ij} := \{k : \{i, k\} \in \mathcal{S}, \{j, k\} \in \mathcal{S}, \text{ and } \{i, j\} \in \mathcal{SP}\}$ is the set of potential colliders formed by i and j . For any $\{i, j\} \in \mathcal{SP}$, if $|\mathcal{C}_{ij}| = 1$, then $\mathcal{C}_{ij}$ is a collider. If $|\mathcal{C}_{ij}| > 1$, then for every $c \in \mathcal{C}_{ij}$, we can compute $W_{i, [j, c]}$. If $W_{i, [j, c]} \neq 0$, then c is a collider since $\{i, j\} \notin \mathcal{S}$. The complexity of computing $\mathcal{K}$ and $\mathcal{S}$ is $O(n^3RN \log N)$. Step 4 is repeated $O(nq)$ times and step 4 (a) which checks for potential colliders among the common neighbors of i and j takes $O(n^2q)$ operations. 4(b) and 4(c) take $O(n)$. Thus the complexity in computing step 4 is $O(n^3q^2)$, giving the overall complexity of $O(n^3(RN \log N + q^2))$. Notice that it is sufficient to compute Algorithm 6 for $O(1)$ number of ω . Thus, the final complexity remains the same.

5.4.2 Sample Complexity Analysis

In practice, due to finite time effects, the Wiener coefficients estimated can be different. Using the concentration bounds from [1] one can obtain a bound on the estimation error of $W_{ij.Z}(\omega) = \mathbf{W}_{i, [j, Z]}(\omega) := \Phi_{X_i \mathbf{X}_{[j, z]}}(\omega) \Phi_{\mathbf{X}_{[j, z]}}^{-1}(\omega)$.

Theorem 5.4.1. *Consider a linear dynamical system governed by (5.1). Suppose that the auto-correlation function $R_x(k) := \mathbb{E}\{X(n)X^T(n+k)\}$ satisfies exponential decay, $\|R_x(k)\|_2 \leq C\delta^{-|k|}$ and that there exists M such that $\frac{1}{M} \leq \lambda_{\min}(\Phi_{\mathbf{X}}) \leq \lambda_{\max}(\Phi_{\mathbf{X}}) \leq M$. Then for any $0 < \epsilon$ and $L \geq \log_\delta \left(\frac{(1-\delta)\epsilon}{2C} \right)$,*

$$\mathcal{P} \left(\|W_i - \widehat{W}_i\| > \epsilon \right) \leq n^2 \exp \left(-(T - L) \min \left\{ \frac{81\epsilon^2}{3200c_1^2 M^{16}(2L + 1)^2 n^2 C^2}, \frac{9\epsilon}{80c_1 M^4(2L + 1)nC} \right\} \right) \quad (5.9)$$

The proof of Theorem 5.4.1 is provided in the supplementary material. For small enough ϵ , the first term is the dominant one. Then $\mathcal{P} \left(\|W_i - \widehat{W}_i\| > \epsilon \right) < \delta$ if the number of samples, $T \gtrsim O \left(\frac{M^{16} L^2 n^2}{\epsilon^2} \ln \left(\frac{n^2}{\delta} \right) \right)$.

5.5 Limitations and Future Works

Frequency-domain independence are defined primarily for linear time-invariant models. It is unclear whether these notions of independence can explain non-linear models. FDD based notion fails for infinite convolution, that is, when the support of $\{H_{ij}(t) : t \in I\}$ is not finite. Sample complexity of dynamical causal inference is generally higher than the static counterpart. It is not clear if the sample complexity of frequency-domain approach is higher than the time-domain counterparts in dynamical setup; this is a future study. Employing efficient loss functions instead of vanilla least square loss in FFT domain might improve the computational and the sample complexities, and therefore require further scrutiny.

5.6 Simulation Results

The simulations are performed on a synthetic dataset. MacbookPro with M1-pro chip having 8 performance cores (up to 3.2GHz) and 2 efficiency cores (up to 2.6GHz) is used for the simulation. The data is generated according to an LDIM 5.2 with the causal graph shown in Figure 5.1(a). The total number of samples in the time-series data is $T = 10,000$ and $N = 64$ point FFT is computed. Figure 5.1(b) shows the estimated kin-graph from the magnitude of $\|W_i[j]\|_{\ell_\infty(\Omega_N)}$, where W_i is computed using (5.5).

We applied the Wiener-Phase Algorithm to compute the CPDAG in Fig 5.1(c). Figure 5.2 shows the phase response of Wiener filter, $W_i[j]$ for $i = 1, 2$ and $j = 4, 5$ versus ω . The phase response plot of $W_2[4]$ is almost constant at π (average = 3.1 and standard deviation = 0.1), whereas the phase response of the remaining pairs in the figure shows a larger variation in mean and variance. Thus, the link $X_2 - X_4$ is deemed spurious. In order to compute the rest of the directions, we perform intervention at node 2 with two

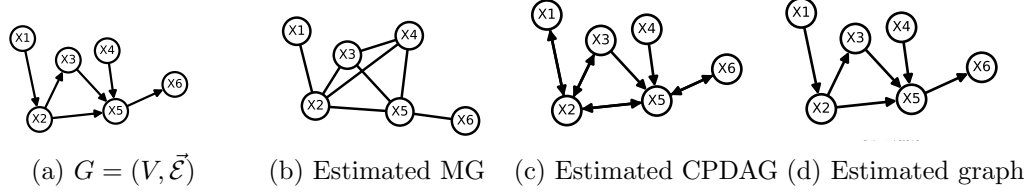


Figure 5.1: (a) shows the true causal generative graph; (b) shows the estimated kin-graph from $T = 10,000$ samples with $N = 64$ using $\ell_\infty(\Omega_N)$ norm on $W_i[j]$; (c) is the CPDAG estimated using Algorithm 6. The bi-directed edges indicate undirected edges in (c); (d) shows the graph estimated by performing the intervention at node 2.

different values $X_2 = y_1$ and $X_2 = y_2$, where $y_2 = 2y_1$. y_1 is a repeating sequence with 32 ones and 32 zeros; this particular pattern is used to obtain a nice 64-point FFT (sync function) for y_1 in the frequency-domain. Using this intervention, the time-series data generated by the LDIM was used to obtain the corresponding Fourier transform using FFT. The plots show the variation of mean of real and imaginary parts of $\hat{X}_i(\omega)$ versus ω under interventions on X_2 . It is evident from the estimated conditional distributions plots that $\mathcal{P}(\hat{X}_i(\omega)|do(X_2 = y_2))$ is different from $\mathcal{P}(\hat{X}_i(\omega)|do(X_2 = y_1))$ for $i = 3$ and 5, and very similar for $i = 4$. It can also be observed from the plots of the imaginary part that large values of mean for ω_k near $k = 0$ and $k = 64$. Detailed and expanded delineation of the results can be found in the supplementary material.

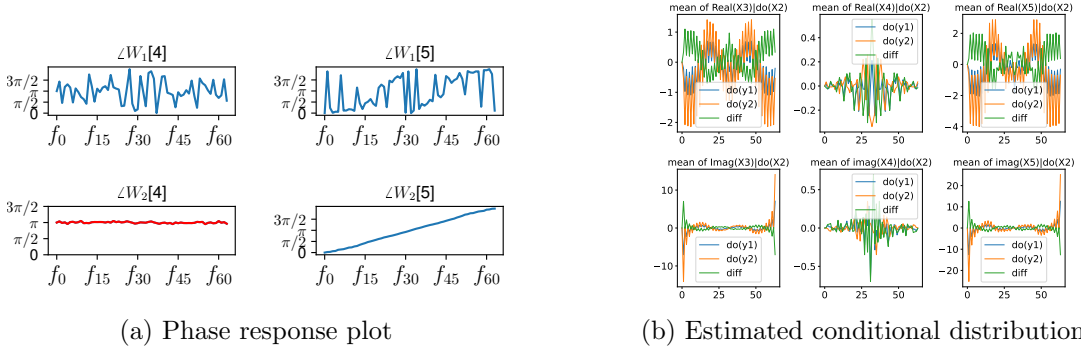


Figure 5.2: (a) shows the phase response of $W_i[j]$ for $i = 1, 2$ and $j = 4, 5$. (b) shows the plot of real and imaginary parts of the mean vs ω of the estimated conditional distribution, when conditioned on $do(X_2 = y_1)$ and $do(X_2 = y_2)$.

Conclusions

In this chapter, we studied the causal identification of linear dynamical systems from the frequency-domain perspective. It was shown that, if the Wiener filter is employed in the FFT domain, the computational complexity of identifying the correlation between the time-series can be reduced significantly. Exploiting the phase properties of the Wiener filter, an algorithm was proposed to efficiently identify the skeleton and the colliders in a graph from data, which can be applied in a large class of applications. Two notions of conditional independence were proposed to realize graph factorization and perform intervention-calculus in stochastic process. Back-door and front-door criteria were derived for stochastic processes using the proposed notions. Non-asymptotic concentration bounds were provide for the error in estimating the Wiener coefficients from data. Finally, theoretical results were verified with simulations.

Chapter 6

Conclusions and Future Research

6.1 Summary of Completed Works

This thesis makes several contributions to reconstruct the correlation and causation structure, and causal/parameter identification in linear dynamical systems, solely from time-series data.

In Chapter 2 we presented a novel approach to reconstruct the topology of networked linear dynamical systems with latent agents, from IPSDM of the observed nodes. The network was allowed to have directed loops and bi-directed edges. It was shown that the IPSDM can be uniquely decomposed into a sparse and a low-rank matrices. The sparse component unveiled the moral graph of the observed nodes, and the low-rank component retrieved the Markov Blanket associated with the latent nodes. Necessary and sufficient conditions for unique sparse plus low-rank decomposition of a skew symmetric matrix was established, along with an optimization based algorithm that decompose the skew symmetric matrix to yield the sparse component $\mathbf{S}$ and the low-rank component $\mathbf{L}$. For a large class of systems, the unique decomposition of imaginary part of the IPSDM is sufficient to achieve the moral graph of the observed nodes including the Markov Blanket of latent nodes. Under some assumptions, it was shown that the imaginary part of the IPSDM can be employed to reconstruct the exact topology of the network. Furthermore, concentration bounds on IPSDM estimation from finite time-series is provided.

In Chapter 3, we addressed the problem of reconstructing the topology of an LDIM with spatially correlated noise sources. Assuming affine correlation between the noise sources, the given LDIM was transformed into an LDIM with latent nodes, where the latent nodes were characterized using the correlation graph and all the nodes were

excited by uncorrelated noises. For polynomial correlation, a generalization of the affine correlation, the latent nodes in the transformed LDIMs were excited using clusters of noise, where the noise clusters were uncorrelated with each other. Finally, using a sparse low-rank matrix decomposition technique, the exact topology of the LDIM was reconstructed, solely from the IPSDM of the true LDIM, when the network satisfied a sufficient condition required for the matrix decomposition. Simulation results are provided that verify the theoretical results.

In Chapter 4, we characterized the optimal sample complexity for structure identification with directions in linear dynamical networks. Inspired by the static setting, a metric, and an algorithm were proposed based on the power spectral density matrix to exactly reconstruct the DAG. It was shown that the optimal sample complexity is $n = \Theta(q \log(p/q))$. For the upper bound characterization, we obtained a tight concentration bound for the power spectral density matrix. An information-theoretic min-max lower bound also was provided for (sub) Gaussian linear dynamical systems. It was shown that the upper bound is order optimal with respect to the lower bound.

In Chapter 5, we studied the causal identification of linear dynamical systems from the frequency-domain perspective. It was shown that, if the Wiener filter is employed in the FFT domain, the computational complexity of identifying the correlation between the time-series can be reduced significantly. Exploiting the phase properties of the Wiener filter, an algorithm was proposed to efficiently identify the skeleton and the colliders in a graph from data, which can be applied in a large class of applications. Two notions of conditional independence were proposed to realize graph factorization and perform intervention-calculus in stochastic process. Back-door and front-door criteria were derived for stochastic processes using the proposed notions. Non-asymptotic concentration bounds were provided for the error in estimating the Wiener coefficients from data. Finally, theoretical results were verified with simulations.

6.2 Future Works

A summary of open research topics that warrant further investigation is provided below:

- An important direction of future study is to study the network topology identification with latent nodes in the presence of non-stationary exogenous noise sources. In some of the real world applications the underlying assumption that the time-series measurements are wide sense stationary might fail. However, in applications

such as ocean temperature the time-series measurements show periodic behavior that can be modeled as cyclostationary processes. Extending the sparse low-rank decomposition framework in Chapter 2, it is possible to obtain a framework that involve block sparse plus low-rank decomposition of the IPSDM. Here, instead of the support structure of the IPSDM, block support structure of the IPSDM retrieves the correlation structure of the underlying time-series. Further, the necessary and sufficient conditions provided in Chapter 2 for the matrix decomposition can be extended to incorporate block-sparse low-rank decomposition. Another problem is the application of Wiener filter based topology identification in the presence of latent nodes. This problem is currently under investigation. Conventional Wiener filter based topology/moral-graph reconstruction approaches assume that all the agents are measurable. However, extending this approach to include latent nodes is a challenging problem, which is under investigation.

- Extending the results in this thesis, which focussed primarily on linear dynamical systems to non-linear systems is an important extension with several applications. Although the work on topology identification with polynomially correlated exogenous noise considered a special form of non-linearity, extension to a general form of non-linear interaction is an open problem. Application of the results here to cyber-physical systems with measurement noise and faulty/malignant agents is another extension of our results. This is pertinent in applications such as power-grids.
- The results in chapter 4 were applicable under a strong assumption that the noise sources are identical or when the noise PSDs were known. Instead of the top-down approach where the root nodes are identified first, one can implement a bottom-up approach using IPSDM instead of PSDM. In the bottom-up approach the leaf nodes are estimated first. Using the do calculus discussed in Chapter 5, one can implement a sequential approach with intervention to estimate the PSDs in LDIMs. This intervention based approach is applicable to more general dynamical systems, including static systems with unequal noise variances.

In conclusion, there are many challenging open questions on learning the correlation and causation structure from time-series measurements. We hope that the research presented in this thesis address some of these questions and serve as a base for a large variety of challenging future research directions.

References

- [1] Mishfad Shaikh Veedu, Harish Doddi, and Murti V Salapaka. Topology learning of linear dynamical systems with latent nodes using matrix decomposition. *IEEE Transactions on Automatic Control*, 67(11):5746–5761, 2021.
- [2] Mishfad Shaikh Veedu and Murti V Salapaka. Topology identification under spatially correlated noise. *Automatica*, 156:111182, 2023.
- [3] Mishfad Shaikh Veedu, Deepjyoti Deka, and Murti Salapaka. Information theoretically optimal sample complexity of learning dynamical directed acyclic graphs. *International Conference on Artificial Intelligence and Statistics*, pages 4636–4644, 2024.
- [4] Mishfad Shaikh Veedu, James Melbourne, and Murti V Salapaka. Causal structure recovery of linear dynamical systems: An fft based approach. *arXiv preprint arXiv:2309.02571*, 2023.
- [5] Y.-B. He and Z. Geng. Active learning of causal networks with intervention experiments and optimal designs. In *Journal of Machine Learning Research*, volume 9, pages 1096–1101, Nov 2008.
- [6] W. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2):195–210, April 1996.
- [7] G. Cavraro and V. Kekatos. Inverter probing for power distribution network topology processing. *IEEE Transactions on Control of Network Systems*, 6(3):980–992, Sep. 2019.

- [8] Elena Ceci, Yanning Shen, Georgios B. Giannakis, and Sergio Barbarossa. Graph-based learning under perturbations via total least-squares. *IEEE Transactions on Signal Processing*, 68:2870–2882, 2020.
- [9] Y. Shen, X. Fu, G. B. Giannakis, and N. D. Sidiropoulos. Topology identification of directed graphs via joint diagonalization of correlation matrices. *IEEE Transactions on Signal and Information Processing over Networks*, 6:271–283, 2020.
- [10] M. Jordan. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.
- [11] Allen J Wood, Bruce F Wollenberg, and Gerald B Sheblé. *Power generation, operation, and control*. John Wiley & Sons, 2013.
- [12] Sourav Patel, Sandeep Attree, Saurav Talukdar, Mangal Prakash, and Murti V Salapaka. Distributed apportioning in a power network for providing demand response services. In *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 38–44. IEEE, 2017.
- [13] M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou. Advanced spectral methods for climatic time series. *Reviews of Geophysics*, 40(1):3–1–3–41, 2002, <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2000RG000092>.
- [14] K. Deng, P. Barooah, P. G. Mehta, and S. P. Meyn. Building thermal model reduction via aggregation of states. In *Proceedings of the 2010 American Control Conference*, pages 5118–5123, June 2010.
- [15] Paul M.J. Van den Hof, Arne Dankers, Peter S.C. Heuberger, and Xavier Bombois. Identification of dynamic models in complex networks with prediction error methods—basic methods for consistent module estimates. *Automatica*, 49(10):2994–3006, 2013.
- [16] Karthik R. Ramaswamy and Paul M.J. Vandenhof. A local direct method for module identification in dynamic networks with correlated noise. *IEEE Transactions on Automatic Control*, pages 1–1, 2020.
- [17] D. Materassi and M. V. Salapaka. On the problem of reconstructing an unknown topology via locality properties of the wiener filter. *IEEE Transactions on Automatic Control*, 57(7):1765–1777, July 2012.

- [18] Saurav Talukdar, Deepjyoti Deka, Harish Doddi, Donatello Materassi, Michael Chertkov, and Murti V. Salapaka. Physics informed topology learning in networks of linear dynamical systems. *Automatica*, 112:108705, 2020.
- [19] M. Dimovska and D. Materassi. Granger-causality meets causal inference in graphical models: Learning networks via non-invasive observations. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 5268–5273, Dec 2017.
- [20] S. Talukdar, D. Deka, S. Attree, D. Materassi, and M. Salapaka. Learning the exact topology of undirected consensus networks. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 5784–5789, Dec 2017.
- [21] S. Talukdar, D. Deka, M. Chertkov, and M. Salapaka. Topology learning of radial dynamical systems with latent nodes. In *2018 Annual American Control Conference (ACC)*, pages 1096–1101, June 2018.
- [22] F. Sepehr and D. Materassi. Blind learning of tree network topologies in the presence of hidden nodes. *IEEE Transactions on Automatic Control*, 65(3):1014–1028, March 2020.
- [23] F. Sepehr and D. Materassi. Inferring the structure of polytree networks of dynamic systems with hidden nodes. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 4618–4623, Dec 2016.
- [24] Firoozeh Sepehr and Donatello Materassi. An algorithm to learn polytree networks with hidden nodes. In *Advances in Neural Information Processing Systems 32*, pages 15110–15119. Curran Associates, Inc., 2019.
- [25] D. Materassi and M. V. Salapaka. Network reconstruction of dynamical polytrees with unobserved nodes. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 4629–4634, 2012.
- [26] D. Materassi and M. V. Salapaka. Identification of network components in presence of unobserved nodes. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 1563–1568, 2015.
- [27] D. Materassi and M. V. Salapaka. Signal selection for estimation and identification in networks of dynamic systems: A graphical model approach. *IEEE Transactions on Automatic Control*, pages 1–1, 2019.

- [28] D. Materassi, G. Innocenti, L. Giarre, and Murli V. Salapaka. Model identification of a network as compressing sensing. *Systems & Control Letters*, 62:664–672, 2013.
- [29] Venkat Ram Subramanian, Andrew Lamperski, and Murli V Salapaka. Network topology identification from corrupt data streams. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1695–1700. IEEE, 2017.
- [30] Venkat Ram Subramanian, Andrew Lamperski, and Murli V Salapaka. Inferring directed graphs for networks from corrupt data-streams. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 4493–4498. IEEE, 2018.
- [31] Venkat Ram Subramanian, Andrew Lamperski, and Murli V Salapaka. Corruption detection in networks of bi-directional dynamical systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 4545–4550. IEEE, 2019.
- [32] M. Zorzi and R. Sepulchre. Ar identification of latent-variable graphical models. *IEEE Transactions on Automatic Control*, 61(9):2327–2340, Sep. 2016.
- [33] R. Liégeois, B. Mishra, M. Zorzi, and R. Sepulchre. Sparse plus low-rank autoregressive identification in neuroimaging time series. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 3965–3970, Dec 2015.
- [34] E. Nozari, Y. Zhao, and J. Cortés. Network identification with latent nodes via autoregressive models. *IEEE Transactions on Control of Network Systems*, 5(2):722–736, June 2018.
- [35] Daniele Alpagó, Mattia Zorzi, and Augusto Ferrante. Identification of sparse reciprocal graphical models. *IEEE Control Systems Letters*, 2(4):659–664, 2018.
- [36] Valentina Ciccone, Augusto Ferrante, and Mattia Zorzi. Factor models with real data: A robust estimation of the number of factors. *IEEE Transactions on Automatic Control*, 64(6):2412–2425, 2019.
- [37] Valentina Ciccone, Augusto Ferrante, and Mattia Zorzi. Learning latent variable dynamic graphical models by confidence sets selection. *IEEE Transactions on Automatic Control*, 65(12):5130–5143, 2020.
- [38] V. Ciccone, A. Ferrante, and M. Zorzi. Robust identification of “sparse plus low-rank” graphical models: An optimization approach. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2241–2246, Dec 2018.

- [39] Saurav Talukdar, Deepjyoti Deka, Donatello Materassi, and Murli Salapaka. Exact topology reconstruction of radial dynamical systems with applications to distribution system of the power grid. In *2017 American Control Conference (ACC)*, pages 813–818, 2017.
- [40] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.
- [41] Stefanie JM Fonken, Karthik Raghavan Ramaswamy, and Paul MJ Van den Hof. A scalable multi-step least squares method for network identification with unknown disturbance topology. *Automatica*, 141:110295, 2022.
- [42] Venkatakrishnan C. Rajagopal, Karthik R. Ramaswamy, and Paul M. J. Van Den Hof. Learning local modules in dynamic networks without prior topology information. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 840–845, 2021.
- [43] James M Bower and David Beeman. *The book of GENESIS: exploring realistic neural models with the GEneral NEural SIMulation System*. Springer Science & Business Media, 2012.
- [44] Young Shin Kim, Svetlozar T Rachev, Michele Leonardo Bianchi, Ivan Mitov, and Frank J Fabozzi. Time series analysis for financial market meltdowns. *Journal of Banking & Finance*, 35(8):1879–1891, 2011.
- [45] Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475. PMLR, 09–11 Apr 2018.
- [46] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.
- [47] Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37(6A), 2009.

- [48] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.
- [49] John AWB Costanzo and Osman Yağın. Data-driven i/o structure learning with contemporaneous causality. *IEEE Transactions on Control of Network Systems*, 7(4):1929–1939, 2020.
- [50] Joshin Krishnan, Rohan Money, Baltasar Beferull-Lozano, and Elvin Isufi. Simplicial vector autoregressive model for streaming edge flows. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [51] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.
- [52] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.
- [53] Zoubin Ghahramani. Learning dynamic bayesian networks. *Adaptive Processing of Sequences and Data Structures: International Summer School on Neural Networks “ER Caianiello” Vietri sul Mare, Salerno, Italy September 6–13, 1997 Tutorial Lectures*, pages 168–197, 2006.
- [54] Gabriele Lohmann, Kerstin Erfurth, Karsten Müller, and Robert Turner. Critical comments on dynamic causal modelling. *Neuroimage*, 59(3):2322–2329, 2012.
- [55] Raha Moraffah, Paras Sheth, Mansoor Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, 63:3041–3085, 2021.
- [56] Michel Besserve, Naji Shajarisales, Dominik Janzing, and Bernhard Schölkopf. Cause-effect inference through spectral independence in linear dynamical systems: theoretical foundations. In *Conference on Causal Learning and Reasoning*, pages 110–143. PMLR, 2022.

- [57] Naji Shajarisales, Dominik Janzing, Bernhard Schölkopf, and Michel Besserve. Telling cause from effect in deterministic linear dynamical systems. In *International Conference on Machine Learning*, pages 285–294. PMLR, 2015.
- [58] Donatello Materassi and Murti V Salapaka. Graphoid-based methodologies in modeling, analysis, identification and control of networks of dynamic systems. In *2016 American Control Conference (ACC)*, pages 4661–4675. IEEE, 2016.
- [59] Ravindran Kannan and Clyde L. Monma. On the computational complexity of integer programming problems. In Rudolf Henn, Bernhard Korte, and Werner Oettli, editors, *Optimization and Operations Research*, pages 161–172, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg.
- [60] Geethu Joseph. *Linear Dynamical systems with sparsity constraints: theory and algorithms*. PhD thesis, Dept. Elect. Comm. Eng., Indian Institute of Science, Bangalore, India, 2019.
- [61] M. Fazel. *Matrix Rank Minimization with Applications*. Ph. D. thesis, Department of Electrical Engineering, Stanford University, Palo Alto, CA, 2002.
- [62] David A. Cox, John Little, and Donal O’Shea. *Ideals, Varieties, and Algorithms—An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer, 2007.
- [63] Khazhgali Kozhasov. On minimality of determinantal varieties, arXiv preprint arXiv:2003.01049., 2003.01049.
- [64] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, Nov 2011.
- [65] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, USA, 2nd edition, 2012.
- [66] Linearizing ODEs, 3 2021. [Online; accessed 2021-06-26].
- [67] Alexander Goldenshluger and Assaf Zeevi. Nonasymptotic bounds for autoregressive time series modeling. *Ann. Statist.*, 29(2):417–444, 04 2001.

- [68] J. Lofberg. Yalmip : a toolbox for modeling and optimization in matlab. In *2004 IEEE International Conference on Robotics and Automation (IEEE Cat. No.04CH37508)*, pages 284–289, Sep. 2004.
- [69] R. H. Tütüncü, K. C. Toh, and M. J. Todd. Sdpt3—a matlab software package for semidefinite-quadratic-linear programming, version 3.0. *Web page <http://www.math.nus.edu.sg/mattohkc/sdpt3.html>*, 2001.
- [70] David Carfi and Giovanni Caristi. Financial dynamical systems. *Differential Geometry–Dynamical Systems*, 2008.
- [71] HJ Dreef, MCF Donkers, and Paul MJ Van den Hof. Identifiability of linear dynamic networks through switching modules. *IFAC-PapersOnLine*, 54(7):37–42, 2021.
- [72] Eduardo Mapurunga and Alexandre Sanfelici Bazanella. Optimal allocation of excitation and measurement for identification of dynamic networks. *arXiv preprint arXiv:2007.09263*, 2020.
- [73] Shengling Shi, Xiaodong Cheng, and Paul M. J. Van den Hof. Single module identifiability in linear dynamic networks with partial excitation and measurement. *arXiv preprint arXiv:2012.11414*, 2020.
- [74] Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.
- [75] Rohan Money, Joshin Krishnan, and Baltasar Beferull-Lozano. Online non-linear topology identification from graph-connected time series. In *2021 IEEE Data Science and Learning Workshop (DSLW)*, pages 1–6, 2021.
- [76] Rohan Money, Joshin Krishnan, and Baltasar Beferull-Lozano. Random feature approximation for online nonlinear graph topology identification. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2021.
- [77] Christopher J. Quinn, Negar Kiyavash, and Todd P. Coleman. Directed information graphs. *IEEE Transactions on Information Theory*, 61(12):6887–6909, 2015.

- [78] Daniele Alpago, Mattia Zorzi, and Augusto Ferrante. A scalable strategy for the identification of latent-variable graphical models. *IEEE Transactions on Automatic Control*, 67(7):3349–3362, 2022.
- [79] Enrico Avventi, Anders G. Lindquist, and Bo Wahlberg. Arma identification of graphical models. *IEEE Transactions on Automatic Control*, 58(5):1167–1178, 2013.
- [80] Jitkomut Songsiri and Lieven Vandenberghe. Topology selection in graphical models of autoregressive processes. *Journal of Machine Learning Research*, 11(91):2671–2705, 2010.
- [81] Giacomo Innocenti and Donatello Materassi. Modeling the topology of a dynamical network via wiener filtering approach. *Automatica*, 48(5):936–946, 2012.
- [82] D. Materassi and G. Innocenti. Topological identification in networks of dynamical systems. *IEEE Transactions on Automatic Control*, 55(8):1860–1871, 2010.
- [83] Mihaela Dimovska and Donatello Materassi. A control theoretic look at granger causality: extending topology reconstruction to networks with direct feedthroughs. *IEEE Transactions on Automatic Control*, Early Access:1–1, 2020.
- [84] Francesca Crescente, Lucia Falconi, Federica Rozzi, Augusto Ferrante, and Mattia Zorzi. Learning ar factor models. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 274–279, 2020.
- [85] Lucia Falconi, Augusto Ferrante, and Mattia Zorzi. A robust approach to arma factor modeling. *arXiv preprint arXiv:2107.03873*, 2021.
- [86] Mattia Zorzi and Alessandro Chiuso. Sparse plus low rank network identification: A nonparametric approach. *Automatica*, 76:355–366, 2017.
- [87] Athanasios Papoulis. *Probability and Statistics*. Prentice-Hall, Inc., USA, 1990.
- [88] Petre Stoica, Randolph L Moses, et al. *Spectral analysis of signals*, volume 452. Pearson Prentice Hall Upper Saddle River, NJ, 2005.
- [89] Xiaofei Yang, Tun Xu, Peng Jia, Han Xia, Li Guo, Lei Zhang, and Kai Ye. Transportation, germs, culture: a dynamic graph model of covid-19 outbreak. *Quantitative Biology*, 8:238–244, 2020.

- [90] Nir Friedman and Zohar Yakhini. On the sample complexity of learning bayesian networks. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 274–282, 1996.
- [91] Or Zuk, Shiri Margel, and Eytan Domany. On the number of samples needed to learn the correct structure of a bayesian network. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 560–567, 2006.
- [92] Ming Gao, Wai Ming Tai, and Bryon Aragam. Optimal estimation of gaussian dag models. In *International Conference on Artificial Intelligence and Statistics*, pages 8738–8757. PMLR, 2022.
- [93] Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.
- [94] Asish Ghoshal and Jean Honorio. Information-theoretic limits of bayesian network structure learning. In *Artificial Intelligence and Statistics*, pages 767–775. PMLR, 2017.
- [95] Asish Ghoshal and Jean Honorio. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. *Advances in Neural Information Processing Systems*, 30, 2017.
- [96] Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.
- [97] Gunwoong Park. Identifiability of additive noise models using conditional variances. *The Journal of Machine Learning Research*, 21(1):2896–2929, 2020.
- [98] Gunwoong Park and Garvesh Raskutti. Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods). *J. Mach. Learn. Res.*, 18(224):1–44, 2017.
- [99] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

- [100] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.
- [101] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.
- [102] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen, 2016.
- [103] Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, pages 1535–1567, 2015.
- [104] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in neural information processing systems*, 24, 2011.
- [105] Jitkomut Songsiri and Lieven Vandenberghe. Topology selection in graphical models of autoregressive processes. *The Journal of Machine Learning Research*, 11:2671–2705, 2010.
- [106] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.
- [107] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.
- [108] Alex Tank, Nicholas J Foti, and Emily B Fox. Bayesian structure learning for stationary time series. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 872–881, 2015.
- [109] Rainer Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51:157–172, 2000.

- [110] Donatello Materassi and Murti V Salapaka. Reconstruction of directed acyclic networks of dynamical systems. In *2013 American Control Conference*, pages 4687–4692. IEEE, 2013.
- [111] Donatello Materassi and Giacomo Innocenti. Unveiling the connectivity structure of financial networks via high-frequency analysis. *Physica A: Statistical Mechanics and its Applications*, 388(18):3866–3878, 2009.
- [112] Harish Doddi, Deepjyoti Deka, Saurav Talukdar, and Murti Salapaka. Efficient and passive learning of networked dynamical systems driven by non-white exogenous inputs. In *International Conference on Artificial Intelligence and Statistics*, pages 9982–9997. PMLR, 2022.
- [113] Aramayis Dallakyan. On learning time series summary dags: A frequency domain approach. *arXiv preprint arXiv:2304.08482*, 2023.
- [114] Mark Fiecas, Chenlei Leng, Weidong Liu, and Yi Yu. Spectral analysis of high-dimensional time series. *Electronic Journal of Statistics*, 13(2):4079–4101, 2019.
- [115] Danna Zhang and Wei Biao Wu. Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *The Annals of Statistics*, 49(1):233–254, 2021.
- [116] Andrew Lamperski. Non-asymptotic pointwise and worst-case bounds for classical spectrum estimators. *arXiv preprint arXiv:2303.11908*, 2023.
- [117] Martin J. Wainwright. *Random matrices and covariance estimation*, page 159–193. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.
- [118] Adrián Pérez-Suay and Gustau Camps-Valls. Causal inference in geoscience and remote sensing from observational data. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1502–1513, 2018.
- [119] Juan F Ramirez-Villegas, Michel Besserve, Yusuke Murayama, Henry C Evrard, Axel Oeltermann, and Nikos K Logothetis. Coupling of hippocampal theta and ripples with pontogeniculooccipital waves. *Nature*, 589(7840):96–102, 2021.
- [120] Pengfei Hu, Rong Jiao, Li Jin, and Momiao Xiong. Application of causal inference to genomic analysis: advances in methodology. *Frontiers in Genetics*, 9:238, 2018.

- [121] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- [122] Thomas Kailath, Ali H Sayed, and Babak Hassibi. *Linear estimation*. Number BOOK. Prentice Hall, 2000.
- [123] Cosma Shalizi and Aryeh Kontorovich. A course on random processes, for students of measure-theoretic probability, with a view to applications in dynamics and statistics. <http://www.stat.cmu.edu/cshalizi/almost-none/v0.1.1/almost-none.pdf>.
- [124] Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.
- [125] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [126] Bruce Lee and Andrew Lamperski. Non-asymptotic closed-loop system identification using autoregressive processes and hankel model reduction. *arXiv preprint arXiv:1909.02192*, 2019.
- [127] Wei Cui, Xu Zhang, and Yulong Liu. Covariance matrix estimation from linearly-correlated gaussian samples. *IEEE Transactions on Signal Processing*, 67(8):2187–2195, 2019.
- [128] Loukas Grafakos. *Classical fourier analysis*, volume 2. Springer, 2008.

Appendix A

Appendix of Chapter 2

A.1 Proof of Lemma 2.2.1

We show the result for an $n = 3$ skew symmetric matrix. It is straight forward to extend this to a general n . Let $\mathbf{A} = \begin{bmatrix} 0 & a & b \\ -a & 0 & 0 \\ -b & 0 & 0 \end{bmatrix}$ be a skew symmetric matrix. $\mathbf{A}$ can be defined as a point on the variety of skew symmetric matrices with support at most 4, i.e., $\mathcal{S}(4) = \mathbb{V}(\mathbf{X}_{11}, \mathbf{X}_{22}, \mathbf{X}_{33}, \mathbf{X}_{12}, \mathbf{X}_{12} + \mathbf{X}_{21}, \mathbf{X}_{13} + \mathbf{X}_{31}, \mathbf{X}_{23} + \mathbf{X}_{32}) \cup \mathbb{V}(\mathbf{X}_{11}, \mathbf{X}_{22}, \mathbf{X}_{33}, \mathbf{X}_{13}, \mathbf{X}_{12} + \mathbf{X}_{21}, \mathbf{X}_{13} + \mathbf{X}_{31}, \mathbf{X}_{23} + \mathbf{X}_{32}) \cup \mathbb{V}(\mathbf{X}_{11}, \mathbf{X}_{22}, \mathbf{X}_{33}, \mathbf{X}_{23}, \mathbf{X}_{12} + \mathbf{X}_{21}, \mathbf{X}_{13} + \mathbf{X}_{31}, \mathbf{X}_{23} + \mathbf{X}_{32})$.

Define $f_{ii} := \mathbf{X}_{ii}, \forall 1 \leq i \leq n$, $f_{ij} := \mathbf{X}_{ij} + \mathbf{X}_{ji}, \forall 1 \leq i < j \leq n$, and $g_{ij} := \mathbf{X}_{ij}, \forall 1 \leq i, j \leq n$. Note that $\mathbf{A}$ is a point on the variety defined by $f_{ii} = 0, 1 \leq i \leq n$, $f_{12} = 0$, $f_{13} = 0$, $f_{23} = 0$, and $g_{23} = 0$, and is non-singular point (with respect to the variety). The tangent space of $\mathcal{S}(4)$ at $\mathbf{A}$, $T(\mathbf{A})$, is (see Proposition 9.6.2 in [62]):

$$T(\mathbf{A}) = \mathbb{V}((d_{\mathbf{A}}(f_{ii}))_{i=1}^3, d_{\mathbf{A}}(f_{12}), d_{\mathbf{A}}(f_{13}), d_{\mathbf{A}}(f_{23}), d_{\mathbf{A}}(g_{23})), \quad (\text{A.1})$$

where $d_{\mathbf{A}}(f) := \sum_{i,j=1}^3 \frac{\partial f}{\partial x_{ij}}(\mathbf{A})(\mathbf{X}_{ij} - \mathbf{A}_{ij})$. It can be shown that $d_{\mathbf{A}}(f_{ii}) = \mathbf{X}_{ii}, \forall 1 \leq i \leq 3$, $d_{\mathbf{A}}(f_{ij}) = \mathbf{X}_{ij} + \mathbf{X}_{ji}, \forall 1 \leq i < j \leq 3$, and $d_{\mathbf{A}}(g_{23}) = \mathbf{X}_{23}$. Plugging this in (A.1), we obtain

$$T(\mathbf{A}) = \left\{ \begin{bmatrix} 0 & \mathbf{X}_{12} & \mathbf{X}_{13} \\ -\mathbf{X}_{12} & 0 & 0 \\ -\mathbf{X}_{13} & 0 & 0 \end{bmatrix} : \mathbf{X}_{12}, \mathbf{X}_{13} \in \mathbb{R} \right\}. \quad (\text{A.2})$$

Clearly, this is the space of all skew-symmetric matrices with support subset of support

of $\mathbf{A}$. Extending this analysis to general n at any point $\mathbf{A}$ gives us the result.

A.2 Proof of Lemma 2.2.2

The following lemma is useful in proving this.

Lemma A.2.1. *If $\mathbf{M}$ is a skew symmetric $n \times n$ matrix of rank r (r even or $n > r$ odd), then there exists a non-singular matrix $\mathbf{P}$ such that $\mathbf{M} = \mathbf{PQP}^T$, where $\mathbf{Q} = \begin{bmatrix} \mathbf{J} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$*

and $\mathbf{J} = \text{diag} \left(\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right)$.

Let $\mathcal{R}(r) := \{\mathbf{X} \in \mathbb{R}^{n \times n} : \text{rank}(\mathbf{X}) = r, \mathbf{X} + \mathbf{X}^T = 0\}$ be the space of skew symmetric matrices with rank r . By Youla decomposition (Lemma A.2.1), this is equivalent to $\mathcal{R}(r) = \{\mathbf{PQP}^T : \mathbf{P} \in \mathbb{R}^{n \times n} \text{ invertible}\}$.

As shown in [63], $\mathcal{R}(r)$ is a smooth manifold. Let $\mathbf{M}$ be an element in $\mathcal{R}(r)$. Let $\gamma : (-1, 1) \rightarrow \mathcal{R}(r)$ be a smooth map such that $\gamma(t) = \mathbf{P}(t)\mathbf{Q}\mathbf{P}^T(t)$ with $\gamma(0) = \mathbf{P}(0)\mathbf{Q}\mathbf{P}^T(0) = \mathbf{M}$, where $\mathbf{P}(t)$ is invertible. Then, $\frac{d\gamma(t)}{dt} = \frac{d\mathbf{P}(t)}{dt}\mathbf{Q}\mathbf{P}^T(t) + \mathbf{P}(t)\mathbf{Q}\frac{d\mathbf{P}^T(t)}{dt}$. For $t = 0$, this can be written as $\frac{d\gamma(t)}{dt}\Big|_{t=0} = \Delta\mathbf{M} + \mathbf{M}\Delta^T$, where $\Delta = \frac{d\mathbf{P}}{dt}(0)\mathbf{P}^{-1}(0)$. Define $W := \{\Delta\mathbf{M} + \mathbf{M}\Delta^T : \Delta \in \mathbb{R}^{n \times n}\}$ and let $\Gamma_M := \{\gamma'(0) : \gamma \text{ is a smooth map from } (-1, 1) \text{ to } \mathcal{R}(r)\}$ be the tangent space with respect to $\mathcal{R}(r)$ at $\mathbf{M}$. We claim that $\Gamma_M = W$. $\Gamma_M \subseteq W$ is obvious. To show the converse, let $w \in W$. Let Δ be such that $\Delta\mathbf{M} + \mathbf{M}\Delta^T = w$, and choose $\mathbf{P}(t) := \mathbf{P}(0) + t\Delta\mathbf{P}(0)$. Invertibility of $\mathbf{P}(t)$ follows from invertibility of $\mathbf{P}(0)$ and continuity of determinant. We need to show that this map belongs to the tangent space Γ_M . Note that $\frac{d\mathbf{P}(t)}{dt}\Big|_{t=0} = \Delta\mathbf{P}(0)$ and $\frac{d\gamma(t)}{dt}\Big|_{t=0} = \Delta\mathbf{M} + \mathbf{M}\Delta^T$, which proves the claim. Plugging in $\mathbf{M}^T = -\mathbf{M}$ and CSVD, $\mathbf{M} = \mathbf{U}\Sigma\mathbf{Q}\mathbf{U}^T$, in the definition of W , and substituting $\mathbf{X} = -\Delta\mathbf{U}\Sigma\mathbf{Q}$ completes the proof.

A.3 Proof of Lemma 2.2.3 :

SVD of a matrix $\mathbf{M}$ is given by $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T$, where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times r}$ are left and right singular matrices and $\Sigma = \text{diag}(\lambda_1, \lambda_1, \dots, \lambda_{r/2}, \lambda_{r/2})$. By Youla decomposition, any skew symmetric matrix $\mathbf{M}$ of rank r can be written in the form $\mathbf{M} = \mathbf{WQW}^T$, where

$$\mathbf{Q} = \text{diag} \left(\begin{bmatrix} 0 & \lambda_1 \\ \lambda_1 & 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 & \lambda_{r/2} \\ \lambda_{r/2} & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \dots \right)$$

and $\mathbf{W}$ is an orthogonal matrix. Note that this is equivalent to writing $\mathbf{M} = \mathbf{W}_r \Sigma \tilde{\mathbf{Q}} \mathbf{W}_r^T$, where $\mathbf{W}_r$ is obtained by picking first r columns of $\mathbf{W}$, $\Sigma = \text{diag}(\lambda_1, \lambda_1, \dots, \lambda_{r/2}, \lambda_{r/2})$, and $\tilde{\mathbf{Q}} = \text{diag}(\underbrace{\mathbf{L}, \dots, \mathbf{L}}_{r \text{ times}})$ with $\mathbf{L} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. That is, the svd, $\mathbf{M} = \mathbf{U} \Sigma \mathbf{V}^T$, can be rewritten with $\mathbf{V} = \mathbf{U} \tilde{\mathbf{Q}}^T$, where $\tilde{\mathbf{Q}}^T = \tilde{\mathbf{Q}}^{-1}$.

Thus, $\mathbf{V} \mathbf{V}^T = \mathbf{U} \tilde{\mathbf{Q}}^T \tilde{\mathbf{Q}} \mathbf{U}^T = \mathbf{U} \tilde{\mathbf{Q}}^{-1} \tilde{\mathbf{Q}} \mathbf{U}^T = \mathbf{U} \mathbf{U}^T$. This concludes the proof.

A.4 Proof of Proposition 2.2.1:

Subdifferential of a convex function f at a point $\mathbf{Y}$ in the domain $\mathcal{X}$ is defined as [125]:

$$\partial f(\mathbf{Y}) = \{g : f(\mathbf{X}) \geq f(\mathbf{Y}) + \langle g, \mathbf{X} - \mathbf{Y} \rangle, \forall \mathbf{X} \in \mathcal{X}\}, \quad (\text{A.3})$$

where $\langle \mathbf{A}, \mathbf{B} \rangle := \text{trace}\{\mathbf{A}^T \mathbf{B}\} = \sum_{i,j} \mathbf{A}_{ij} \mathbf{B}_{ij}$.

The lagrangian of (13), J , can be written as

$$J(\mathbf{S}, \mathbf{L}, \mathbf{Q}_1, \mathbf{Q}_2) = \gamma \|\mathbf{S}\|_1 + \|\mathbf{L}\|_* + \langle \mathbf{Q}_1, \mathbf{C} - \mathbf{S} - \mathbf{L} \rangle + \langle \mathbf{Q}_2, \mathbf{S} + \mathbf{S}^T \rangle \quad (\text{A.4})$$

Remark A.4.1. We can add another dual variable corresponding to the constraint $\mathbf{L} + \mathbf{L}^T = \mathbf{0}$. However, this would be redundant with $\mathbf{Q}_2$ as $\mathbf{C}$ is always a skew symmetric matrix.

The optimality conditions are then given by

$$\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T \in \gamma \partial \|\tilde{\mathbf{S}}\|_1, \text{ and } \mathbf{Q}_1 \in \gamma \partial \|\tilde{\mathbf{L}}\|_*. \quad (\text{A.5})$$

From the characterization of subdifferential of $\|\tilde{\mathbf{S}}\|_1$, it follows that

$$P_{\Omega(\tilde{\mathbf{S}})}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T) = \gamma \text{sign}(\tilde{\mathbf{S}}) \text{ and } \|P_{\Omega(\tilde{\mathbf{S}})^c}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T)\|_\infty \leq \gamma, \quad (\text{A.6})$$

where $P_{\Omega(\tilde{\mathbf{S}})}(\mathbf{A})$ is obtained by setting entries of $\mathbf{A}$ outside the support of $\tilde{\mathbf{S}}$ to zero and projecting it to the space of skew symmetric matrices.

Similarly, from the characterization of $\partial \|\tilde{\mathbf{L}}\|_*$ [40], we have

$$P_{T(\tilde{\mathbf{L}})}(\mathbf{Q}_1) = \mathbf{U} \mathbf{V}^T \text{ and } \|P_{T^\perp}(\mathbf{Q}_1)\|_2 \leq 1, \quad (\text{A.7})$$

where $P_{T(\tilde{\mathbf{L}})}(\mathbf{A}) := P_{\mathbf{U}} \mathbf{A} + \mathbf{A} P_{\mathbf{U}} - P_{\mathbf{U}} \mathbf{A} P_{\mathbf{U}}$; $P_{\mathbf{U}} = \mathbf{U} \mathbf{U}^T$, $\tilde{\mathbf{L}} = \mathbf{U} \Sigma \mathbf{V}^T$.

From the subgradient optimality conditions, we have that $(\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ is an optimal solution if there exist duals $\mathbf{Q}_1, \mathbf{Q}_2$ such that

$$\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T \in \gamma \partial \|\tilde{\mathbf{S}}\|_1, \text{ and } \mathbf{Q}_1 \in \gamma \partial \|\tilde{\mathbf{L}}\|_*. \quad (\text{A.8})$$

The second condition in the proposition statement guarantees the existence of such a dual since they satisfy (A.6) and (A.7).

Next, we show the uniqueness of the solution. We prove this by contradiction. Let $(\mathbf{Q}_S, \mathbf{Q}_L)$ be any subdifferential at $(\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ and let $(\tilde{\mathbf{S}} + \mathbf{N}_S, \tilde{\mathbf{L}} + \mathbf{N}_L)$ be another optimal solution. It follows, from the constraint $C = \tilde{\mathbf{S}} + \mathbf{N}_S + \tilde{\mathbf{L}} + \mathbf{N}_L$, that $\mathbf{N}_S + \mathbf{N}_L = \mathbf{0}$. Then,

$$\gamma \|\tilde{\mathbf{S}} + \mathbf{N}_S\|_1 + \|\tilde{\mathbf{L}} + \mathbf{N}_L\|_* \geq \gamma \|\tilde{\mathbf{S}}\|_1 + \|\tilde{\mathbf{L}}\|_* + \langle \mathbf{Q}_S, \mathbf{N}_S \rangle + \langle \mathbf{Q}_L, \mathbf{N}_L \rangle.$$

For notational simplicity, we write $\Omega(\tilde{\mathbf{S}})$, $\Omega^C(\tilde{\mathbf{S}})$, $T(\tilde{\mathbf{L}})$, and $T^\perp(\tilde{\mathbf{L}})$ as Ω , Ω^C , T and $T^\perp$ respectively.

$$\begin{aligned} \langle \mathbf{Q}_S, \mathbf{N}_S \rangle &= \langle \gamma \text{sign}(\tilde{\mathbf{S}}) + P_{\Omega^C}(\mathbf{Q}_S), \mathbf{N}_S \rangle \\ &= \langle \mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T - P_{\Omega^C}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T) + P_{\Omega^C}(\mathbf{Q}_S), \mathbf{N}_S \rangle \\ &= \langle \mathbf{Q}_1, \mathbf{N}_S \rangle - \langle \mathbf{Q}_2 + \mathbf{Q}_2^T, \mathbf{N}_S \rangle + \langle -P_{\Omega^C}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T) + P_{\Omega^C}(\mathbf{Q}_S), \mathbf{N}_S \rangle \\ &\stackrel{(i)}{=} \langle \mathbf{Q}_1, \mathbf{N}_S \rangle + \langle -P_{\Omega^C}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T) + P_{\Omega^C}(\mathbf{Q}_S), \mathbf{N}_S \rangle. \end{aligned}$$

Here, (i) follows since $\mathbf{Q}_2 + \mathbf{Q}_2^T$ is a symmetric matrix and $\mathbf{N}_S$ is skew symmetric matrix. Then, $\langle \mathbf{Q}_2 + \mathbf{Q}_2^T, \mathbf{N}_S \rangle = 0$. Similarly,

$$\begin{aligned} \langle \mathbf{Q}_L, \mathbf{N}_L \rangle &= \langle \mathbf{U}\mathbf{V}^T + P_{T^\perp}(\mathbf{Q}_L), \mathbf{N}_L \rangle \\ &\stackrel{(ii)}{=} \langle \mathbf{Q}_1 - P_{T^\perp}(\mathbf{Q}_1) + P_{T^\perp}(\mathbf{Q}_L), \mathbf{N}_L \rangle \\ &= \langle \mathbf{Q}_1, \mathbf{N}_L \rangle + \langle P_{T^\perp}(\mathbf{Q}_L) - P_{T^\perp}(\mathbf{Q}_1), \mathbf{N}_L \rangle, \end{aligned}$$

where (ii) follows since $\mathbf{Q}_1 = \mathbf{U}\mathbf{V}^T + P_{T^\perp}(\mathbf{Q}_1)$.

Then,

$$\begin{aligned}
\langle \mathbf{Q}_L, \mathbf{N}_L \rangle + \langle \mathbf{Q}_S, \mathbf{N}_S \rangle &= \langle \mathbf{Q}_1, \mathbf{N}_L \rangle + \langle P_{T^\perp}(\mathbf{Q}_L) - P_{T^\perp}(\mathbf{Q}_1), \mathbf{N}_L \rangle \\
&+ \langle \mathbf{Q}_1, \mathbf{N}_S \rangle + \langle P_{\Omega^C}(\mathbf{Q}_S) - P_{\Omega^C}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T), \mathbf{N}_S \rangle \\
&\stackrel{(iii)}{=} \langle P_{T^\perp}(\mathbf{Q}_L) - P_{T^\perp}(\mathbf{Q}_1), \mathbf{N}_L \rangle + \langle P_{\Omega^C}(\mathbf{Q}_S) - P_{\Omega^C}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T), \mathbf{N}_S \rangle \\
&= \langle P_{T^\perp}(\mathbf{Q}_L) - P_{T^\perp}(\mathbf{Q}_1), P_{T^\perp}(\mathbf{N}_L) \rangle \\
&+ \langle P_{\Omega^C}(\mathbf{Q}_S) - P_{\Omega^C}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T), P_{\Omega^C}(\mathbf{N}_S) \rangle.
\end{aligned}$$

Here, (iii) follows since $\mathbf{N}_S + \mathbf{N}_L = \mathbf{0}$. Now, choose $\mathbf{Q}_S = \gamma \text{sign}(P_{\Omega^C}(\mathbf{N}_S))$ and $\mathbf{Q}_L = \hat{\mathbf{U}}\hat{\mathbf{V}}^T$, where $P_{T^\perp}(\mathbf{N}_L) = \hat{\mathbf{U}}\hat{\Sigma}\hat{\mathbf{V}}^T$. Applying this substitution along with Holder's inequality [65], $\langle \mathbf{Q}_L, \mathbf{N}_L \rangle + \langle \mathbf{Q}_S, \mathbf{N}_S \rangle$ can be written as

$$\begin{aligned}
\langle \mathbf{Q}_L, \mathbf{N}_L \rangle + \langle \mathbf{Q}_S, \mathbf{N}_S \rangle &\geq (1 - \|P_{T^\perp}(\mathbf{Q}_1)\|_2) \|P_{T^\perp}(\mathbf{N}_L)\|_* \\
&+ (\gamma - \|P_{\Omega^C}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T)\|_\infty) \|P_{\Omega^C}(\mathbf{N}_S)\|_1.
\end{aligned}$$

Since $\|P_{\Omega^C}(\mathbf{Q}_1 - \mathbf{Q}_2 - \mathbf{Q}_2^T)\|_\infty < \gamma$ and $\|P_{T^\perp}(\mathbf{Q}_1)\|_2 < 1$, $\langle \mathbf{Q}_L, \mathbf{N}_L \rangle + \langle \mathbf{Q}_S, \mathbf{N}_S \rangle = 0$ if and only if both $P_{\Omega^C}(\mathbf{N}_S) = P_{T^\perp}(\mathbf{N}_L) = \mathbf{0}$, which implies that $P_\Omega(\mathbf{N}_S) + P_T(\mathbf{N}_L) = \mathbf{0}$ as $\mathbf{N}_S + \mathbf{N}_L = \mathbf{0}$. Since $\Omega(\tilde{\mathbf{S}}) \cap T(\tilde{\mathbf{L}}) = \{\mathbf{0}\}$, this would mean that $\mathbf{N}_S = \mathbf{N}_L = \mathbf{0}$.

A.5 Proof of Proposition 2.2.2:

It can be checked that if $\mu(\tilde{\mathbf{S}})\xi(\tilde{\mathbf{L}}) < \frac{1}{6}$, then $\frac{\xi(\tilde{\mathbf{L}})}{1-4\mu(\tilde{\mathbf{S}})\xi(\tilde{\mathbf{L}})} < \frac{1-3\mu(\tilde{\mathbf{S}})\xi(\tilde{\mathbf{L}})}{\mu(\tilde{\mathbf{S}})}$ and the range in (22) is non-empty.

Here, we construct the duals $\hat{\mathbf{Q}}_1, \hat{\mathbf{Q}}_2$ by picking both from the direct sum $\Omega \oplus T$, that is, $\hat{\mathbf{Q}}_1 = \hat{\mathbf{Q}}_{1,\Omega} + \hat{\mathbf{Q}}_{1,T}$ and $\hat{\mathbf{Q}}_2 = \hat{\mathbf{Q}}_{2,\Omega} + \hat{\mathbf{Q}}_{2,T}$, where $\hat{\mathbf{Q}}_{1,\Omega}, \hat{\mathbf{Q}}_{2,\Omega} \in \Omega$ and $\hat{\mathbf{Q}}_{1,T}, \hat{\mathbf{Q}}_{2,T} \in T$. We will show that, for any γ taken from (22), the duals satisfy the sufficient conditions specified by Proposition 1. This would then guarantee that (13) returns $(\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$ as the unique optimum solution.

Let $\hat{\mathbf{Q}} = \hat{\mathbf{Q}}_1 - \hat{\mathbf{Q}}_2 - \hat{\mathbf{Q}}_2^T$. $\hat{\mathbf{Q}}_2$ is a skew symmetric matrix by construction, and hence, $\hat{\mathbf{Q}}_2 + \hat{\mathbf{Q}}_2^T = \mathbf{0}$, which implies $\hat{\mathbf{Q}} = \hat{\mathbf{Q}}_1$. By following the similar arguments from Proposition 1 in [40], it can be shown that $\Omega \cap T = \{\mathbf{0}\}$ if $\mu(\tilde{\mathbf{S}})\xi(\tilde{\mathbf{L}}) < 1/6$. It follows that we can find a unique $\hat{\mathbf{Q}}_1$ such that $P_\Omega(\hat{\mathbf{Q}}_1) = \gamma \text{sign}(\tilde{\mathbf{S}})$ and $P_T(\hat{\mathbf{Q}}_1) = \mathbf{U}\mathbf{V}^T$. We will show that this particular $\hat{\mathbf{Q}}_1$ satisfies the remaining conditions in Proposition 1.

Let $\hat{\mathbf{Q}}_{1,\Omega} = \gamma \text{sign}(\tilde{\mathbf{S}}) + \epsilon_\Omega$ and $\hat{\mathbf{Q}}_{1,T} = \mathbf{UV}^T + \epsilon_T$. Then,

$$P_\Omega(\hat{\mathbf{Q}}) = P_\Omega(\hat{\mathbf{Q}}_{1,\Omega}) + P_\Omega(\hat{\mathbf{Q}}_{1,T}) \quad (\text{A.9})$$

$$= \gamma \text{sign}(\tilde{\mathbf{S}}) + \epsilon_\Omega + P_\Omega(\mathbf{UV}^T + \epsilon_T). \quad (\text{A.10})$$

From (A.6), it follows that $P_\Omega(\hat{\mathbf{Q}}) = \gamma \text{sign}(\tilde{\mathbf{S}})$, which indicates that $\epsilon_\Omega = -P_\Omega(\mathbf{UV}^T + \epsilon_T)$. Similarly, from the definition of $P_{T^\perp}(\hat{\mathbf{Q}}_1)$ and (A.7), we obtain $\epsilon_T = -P_T(\gamma \text{sign}(\tilde{\mathbf{S}}) + \epsilon_\Omega)$.

Next, following along the lines of Theorem 2 in [40], we can show that $\|P_{\Omega^c}(\hat{\mathbf{Q}})\|_\infty < 1$ if $\gamma > \frac{\xi(\tilde{\mathbf{L}})}{1-4\mu(\tilde{\mathbf{S}})\xi(\tilde{\mathbf{L}})}$ and $\|P_{T^\perp}(\hat{\mathbf{Q}}_1)\|_2 < 1$ if $\gamma < \frac{1-3\mu(\tilde{\mathbf{S}})\xi(\tilde{\mathbf{L}})}{\mu(\tilde{\mathbf{S}})}$. The proposition statement follows.

A.6 Proof of Proposition 2.2.3

For any matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\frac{1}{k_1}\|\mathbf{A}\|_1 \leq \|\mathbf{A}\|_* \leq k_2\|\mathbf{A}\|_1$, where k_1, k_2 are real numbers. Note that k_1 and k_2 can be functions of n . The tightest bound available is $k_1 = n$ and $k_2 = 1$ [65].

Define $f(\mathbf{S}, \mathbf{L}, t) := t\|\mathbf{S}\|_1 + (1-t)\|\mathbf{L}\|_*$. Clearly, the minimum value at $t = 0$ is $f(\hat{\mathbf{S}}_0, \hat{\mathbf{L}}_0, \mathbf{0}) = 0$, and is obtained at the point $(\hat{\mathbf{S}}, \hat{\mathbf{L}}) = (\mathbf{C}, \mathbf{0})$. Next, consider $t = \epsilon$ for some ϵ close to zero. The objective value at $(\mathbf{S}, \mathbf{L}) = (\mathbf{C} - \mathbf{N}, \mathbf{N})$ for $t = \epsilon$, where $\mathbf{N} \neq \mathbf{0}$ is

$$f(\mathbf{S}, \mathbf{L}, \epsilon) = \epsilon\|\mathbf{C} - \mathbf{N}\|_1 + (1-\epsilon)\|\mathbf{N}\|_*. \quad (\text{A.11})$$

Let the objective value for $t = \epsilon$ at the point $(\mathbf{C}, \mathbf{0})$ be $g(\epsilon) := f(\mathbf{C}, \mathbf{0}, \epsilon) = \epsilon\|\mathbf{C}\|_1$. This can be rewritten as

$$\begin{aligned} g(\epsilon) &= \epsilon\|\mathbf{C} - \mathbf{N} + \mathbf{N}\|_1 \leq \epsilon\|\mathbf{C} - \mathbf{N}\|_1 + \epsilon\|\mathbf{N}\|_1 \\ &\leq \epsilon\|\mathbf{C} - \mathbf{N}\|_1 + \epsilon k_1\|\mathbf{N}\|_*, \end{aligned} \quad (\text{A.12})$$

where the first inequality follows from the triangle inequality, and the second inequality follows since $\|\mathbf{N}\|_1 \leq k_1\|\mathbf{N}\|_*$.

Then, $f(\mathbf{S}, \mathbf{L}, \epsilon) - g(\epsilon) \geq (1 - (k_1 + 1)\epsilon)\|\mathbf{N}\|_*$. For $0 < \epsilon < \epsilon_L$, where $\epsilon_L = \frac{1}{k_1 + 1}$, we have that $f(\mathbf{S}, \mathbf{L}, \epsilon) - g(\epsilon) > 0$, since $\|\mathbf{N}\|_* > 0$. Thus, we have shown that for $t = \epsilon$ the value of the objective function at $(\mathbf{C} - \mathbf{N}, \mathbf{N})$ is strictly greater than that at $(\mathbf{C}, \mathbf{0})$.

Moreover, this is true for all $\mathbf{N} \neq \mathbf{0}$. Therefore, we can conclude that for $0 \leq t < \epsilon_L$, the optimal solution is $(\mathbf{C}, \mathbf{0})$.

Similarly, it is easy to see that the minimum value at $t = 1$, $f(\hat{\mathbf{S}}_1, \hat{\mathbf{L}}_1, 1) = 0$, and is obtained at the point $(\mathbf{S}, \mathbf{L}) = (\mathbf{0}, \mathbf{C})$. Proceeding similar to the above analysis for $t = 1$, with $(\mathbf{S}, \mathbf{L}) = (\mathbf{N}, \mathbf{C} - \mathbf{N})$ and $g(\epsilon) := f(\mathbf{0}, \mathbf{C}, \epsilon) = (1 - \epsilon)\|\mathbf{C}\|_*$, we can conclude that $f(\mathbf{S}, \mathbf{L}, \epsilon) - g(\epsilon) \geq (1 - (k_2 + 1)\epsilon)\|\mathbf{N}\|_1$. For $\epsilon_U < t \leq 1$, where $\epsilon_U = \frac{k_2}{k_2 + 1}$, we have that $f(\mathbf{S}, \mathbf{L}, \epsilon) - g(\epsilon) > 0$, for every $\mathbf{N} \neq \mathbf{0}$. Thus, for $t > \epsilon_U$, the optimal solution is $(\mathbf{0}, \mathbf{C})$.

Proposition 2.2.2 and Lemma 2.2.4 showed that (with the change of variable $t = \gamma/(1 + \gamma)$), for every t within the specified range, the optimal solution is $(\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t) = (\tilde{\mathbf{S}}, \tilde{\mathbf{L}})$. That is, if $\mu\xi < 1/6$ or $\deg_{\max}(\tilde{\mathbf{S}})inc(\tilde{\mathbf{L}}) < 1/12$, then we can find an interval $(t_1, t_2) \subset [0, 1]$ with $0 < t_1 < t_2 < 1$ such that $(\tilde{\mathbf{S}}, \tilde{\mathbf{L}}) = (\hat{\mathbf{S}}_t, \hat{\mathbf{L}}_t)$ for any $t \in (t_1, t_2)$. This implies that $diff_t = 0$ for $t \in [t_1 + \varepsilon, t_2 - \varepsilon]$ and sufficiently small interval ε . Therefore, there exists at least three zero regions for the plot of $diff_t$ versus t if any of the sufficient conditions mentioned above is satisfied and ε is small enough.

A.7 Proof of Theorem 2.4.1

To prove the first part, let $\bar{\mathcal{E}}_o$ be the set of undirected edges obtained by removing direction from $\mathcal{E}_o$, without repetition. Then, by Proposition 2.4.1, $\bar{\mathcal{E}}_o \cup \mathcal{E}_o^{ss} \supseteq \{(i, j) : \mathbf{S}_{ij} \neq 0, i < j\}$, i.e., $|support(\mathbf{S})| \leq 2|\bar{\mathcal{E}}_o| + 2|\mathcal{E}_o^{ss}| + n \leq 2|\mathcal{E}_o| + 2|\mathcal{E}_o^{ss}| + n$, where the additional n is due to the diagonal entries of $\mathbf{S}$.

To show the upper bound on rank, let $\mathbf{A} = \Psi^* \Lambda^{-1} \Psi$. Notice that $rank(\mathbf{A}) \leq \min\{rank(\Psi), rank(\Lambda)\} \leq n_h$. Similarly, $rank(\mathbf{H}_{ho}^* \Phi_{e_h}^{-1} \mathbf{H}_{ho}) \leq n_h$. Then, $rank(L) \leq 2n_h$ from (2.20).

A.8 Proof of Theorem 2.4.2

Recall expression (2.20) for $\mathbf{L}$. By expanding each term in (2.20) we obtain that if $\mathbf{H}_{hh} = 0$, then Λ is diagonal, and $\Psi(k_h, j)$ can be written as

$$\Psi(k_h, j) = \mathbf{H}_{oh}^*(j, k_h) \Phi_{e_o_j}^{-1} + \mathbf{H}_{ho}(k_h, j) \Phi_{e_h k_h}^{-1} - \sum_{k=1}^{n_o} \Phi_{e_o_k}^{-1} \mathbf{H}_{oh}^*(k, k_h) \mathbf{H}_{oo}(k, j)$$

Then, we can conclude the following from the aforementioned equations.

- If the hidden nodes are at least two hops away, then Λ is real and diagonal.
- If any of $k_h \rightarrow j$ or $j \rightarrow k_h$ or $k_h \rightarrow k \leftarrow j$ exists for some $k \in \mathcal{V}_o$, then $\Psi_{k_h, j} \neq 0$
- For $i \neq j$, $\mathbf{L}_{ij}$ is given by the expression (A.13) below.

$$\begin{aligned} \mathbf{L}_{i,j} = & \sum_{l=1}^{n_h} [\mathbf{H}_{ho}^*(l, i)] \Phi_{e_h, l}^{-1} \mathbf{H}_{ho}(l, j) - \sum_{l=1}^{n_h} d_l \\ & \left[\mathbf{H}_{oh}(i, l) \Phi_{e_h, l}^{-1} + [\mathbf{H}_{ho}^*(l, i)] \Phi_{e_h, l}^{-1} - \sum_{k=1}^{n_o} \Phi_{e_o, k}^{-1} \mathbf{H}_{oh}(k, l) [\mathbf{H}_{oo}^*(k, i)] \right] \\ & \left[\mathbf{H}_{oh}^*(j, l) \Phi_{e_h, l}^{-1} + [\mathbf{H}_{ho}(l, j)] \Phi_{e_h, l}^{-1} - \sum_{k=1}^{n_o} \Phi_{e_o, k}^{-1} \mathbf{H}_{oh}^*(k, l) [\mathbf{H}_{oo}(k, j)] \right] \end{aligned} \quad (\text{A.13})$$

Let a_1 , a_2 , and a_3 respectively denote first, second, and third terms in the first bracket and let b_1 , b_2 , and b_3 respectively denote first, second, and third terms in the second bracket. Then, (A.13) can be rewritten as,

$$\begin{aligned} \mathbf{L}(i, j) = & \sum_{l=1}^{n_h} [\mathbf{H}_{ho}(l, i)]^* \Phi_{e_h, l}^{-1} \mathbf{H}_{ho}(l, j) - \sum_{l=1}^{n_h} d_l [a_1 + a_2 + a_3] [b_1 + b_2 + b_3] \\ = & \sum_{l=1}^{n_h} [\mathbf{H}_{ho}(l, i)]^* \Phi_{e_h, l}^{-1} \mathbf{H}_{ho}(l, j) - \sum_{l=1}^{n_h} d_l [a_1 b_1 + a_1 b_2 + a_1 b_3 + a_2 b_1 + a_2 b_2 + \\ & a_2 b_3 + a_3 b_1 + a_3 b_2 + a_3 b_3]. \end{aligned} \quad (\text{A.14})$$

- (a) We use contrapositive argument to prove this. Suppose that, for every hidden node $h \in \mathcal{V}_h$, there does not exist $g \in DM_h(i, j)$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$. Then, from (A.14), it follows that $\mathbf{L}_{ij} = 0$. Therefore, if $\mathbf{L}_{ij} \neq 0$, then there exists $g \in DM_h(i, j)$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$, for some $h \in \mathcal{V}_h$. Clearly, for any $g \in DM_h(i, j)$, we have $d_{hop}(i, h) \leq 2$ and $d_{hop}(j, h) \leq 2$.
- (b) Let $i, j \in \mathcal{V}_o$. Suppose that $\mathbf{L}_{ij} \neq 0$. Then, from part (a), there exists $g_1 \in DM_{h_1}(i, j)$ such that $g_1 \in \mathcal{G}(\mathcal{V}, \mathcal{E})$, for some $h_1 \in \mathcal{V}_h$. Suppose for contradiction that there exists $h_2 \in \mathcal{V}_h \setminus \{h_1\}$ such that there exists $g_2 \in DM_{h_2}(i, j)$ such that $g_2 \in \mathcal{G}(\mathcal{V}, \mathcal{E})$. Then, we have $d_{hop}(h_1, h_2) \leq d_{hop}(h_1, i) + d_{hop}(i, h_2) \leq 4$ from Theorem 2.4.2(a), which contradicts Assumption 2.4.2. Hence, $h_1 = h_2$.

A.9 Proof of Theorem 2.4.3

- (a) Suppose $M_{l_1} \cap M_{l_2} \neq \emptyset$ for some $l_1, l_2 \in \mathcal{V}_h, l_1 \neq l_2$. Then, there exists $a \in M_{l_1} \cap M_{l_2}$. Since $a \in M_{l_1}$, it follows from the definition of M_{l_1} that $d_{hop}(a, l_1) \leq 2$. Similarly, $a \in M_{l_2}$ and $d_{hop}(a, l_2) \leq 2$. Then, we have $d_{hop}(l_1, l_2) \leq d_{hop}(l_1, a) + d_{hop}(a, l_2) \leq 4$, which contradicts Assumption 2.4.2. Therefore, $M_{l_1} \cap M_{l_2} = \emptyset$ for all $l_1, l_2 \in \mathcal{V}_h, l_1 \neq l_2$.
- (b) Suppose $Q_{l_1} \cap Q_{l_2} \neq \emptyset$ for some $l_1, l_2 \in \mathcal{V}_h, l_1 \neq l_2$. Then, there exists $(i_0, j_0) \in Q_{l_1} \cap Q_{l_2}$. $(i_0, j_0) \in Q_{l_1}$, which implies that $i_0, j_0 \in M_{l_1}$. Similarly, $(i_0, j_0) \in Q_{l_2}$ implies that $i_0, j_0 \in M_{l_2}$. Thus, $i_0, j_0 \in M_{l_1} \cap M_{l_2}$, which contradicts part (a). Therefore, $Q_{l_1} \cap Q_{l_2} = \emptyset$.

- (c) We first show that $V_H \supseteq \bigcup_{l=1}^{n_h} M_l$. To show this, let $a \in \bigcup_{l=1}^{n_h} M_l$. Then, $a \in M_l$ for some $l \in \mathcal{V}_h$. It follows from the definition of M_l that $a \in \mathcal{P}(l) \cup \mathcal{C}(l) \cup \mathcal{S}(l)$. If $a \in \mathcal{P}(l) \cup \mathcal{C}(l)$, then by Assumption 2.4.1, there exists $c \in M_l \setminus a$. If $a \in \mathcal{S}(l) \setminus (\mathcal{C}(l) \cup \mathcal{P}(l))$, then there exists $k \in M_l \setminus a$, such that $k \in \mathcal{C}(l) \cap \mathcal{C}(a)$. Given the existence of node $k \in \mathcal{C}(l)$, by Assumption 2.4.1 there exists $c \in M_l \setminus \{a, k\}$. Clearly there is a node $c \in M_l \setminus a$ in $\mathcal{G}(\mathcal{V}, \mathcal{E})$, which is also present in $DM_l(a, c)$. From (A.14) and Remark 2.4.5, it follows that $\mathbf{L}_{ac} \neq 0$. Because $\mathbf{L}_{ac} \neq 0$, it follows from the definition V_H , that $a \in V_H$.

Now, we show that $V_H \subseteq \bigcup_{l=1}^{n_h} M_l$. Let $a \in V_H$, then from the definition of V_H , there exists $c \in \mathcal{V}_o \setminus a$, such that $\mathbf{L}_{ac} \neq 0$. From Theorem 2.4.2 parts (a) and (b), there exists a unique hidden node $l \in \mathcal{V}_h$ such that $g \in DM_l(a, c)$ exists in $\mathcal{G}(\mathcal{V}, \mathcal{E})$. From the definition of $DM_l(a, c)$, it follows that $a \in M_l$. Hence, $a \in \bigcup_{l=1}^{n_h} M_l$, which concludes the proof.

We first show $E_H \supseteq \bigcup_{l=1}^{n_h} Q_l$. Let $(i_0, j_0) \in \bigcup_{l=1}^{n_h} Q_l$, then there exists a $l \in \mathcal{V}_h$, such that $(i_0, j_0) \in Q_l$. From the definition of $Q_l, \{i_0, j_0\} \in M_l \subset V_H$ and $\mathbf{L}_{i_0 j_0} \neq 0$. Thus $(i_0, j_0) \in E_H$.

To show the converse $E_H \subseteq \bigcup_{l=1}^{n_h} Q_l$, let $(i_0, j_0) \in E_H$, then $\mathbf{L}_{i_0 j_0} \neq 0$. From Theorem 2.4.2(b), there exists a unique hidden node $l \in \mathcal{V}_h$, such that a $g \in DM_l(i_0, j_0)$ exists in $\mathcal{G}(\mathcal{V}, \mathcal{E})$. It follows from the definition of $DM_l(i_0, j_0)$ that $\{i_0, j_0\} \in M_l$. $\{i_0, j_0\} \in M_l$ and $\mathbf{L}_{i_0 j_0} \neq 0$, thus $(i_0, j_0) \in Q_l \subset \bigcup_{l=1}^{n_h} Q_l$.

A.10 Proof of Theorem 2.5.1

From (2.19), observe that $\mathbf{S}$ is a Hermitian matrix, which is true if and only if $\Im\{\mathbf{S}\}$ is anti-symmetric, and thus the diagonal entries of $\Im\{\mathbf{S}\}$ are zeros. Lemma 2.4.1 showed that if $\mathbf{S}_{ij} \neq 0$, then $(i, j) \in \text{kin}(\mathcal{G}_o)$. Combining this with Lemma 2.5.1, we get if $\Im(\mathbf{S}_{ij}) \neq 0$ then $i \in \mathcal{C}(j)$ or $i \in \mathcal{P}(j)$. That is, $\widehat{\mathcal{E}}_o \subseteq \overline{\mathcal{E}}_o$.

To prove the equality, we need to show that $\widehat{\mathcal{E}}_o \supseteq \overline{\mathcal{E}}_o$. Let $i \neq j$ and suppose $[\mathbf{H}_{oo}]_{ij} \neq 0$. By expanding (2.19), $\mathbf{S}_{ij} = (\Phi_{e_o}^{-1})_{ij} - \sum_{k=1}^n (\mathbf{H}_{oo}^*)_{ik} (\Phi_{e_o}^{-1})_{kj} - \sum_{k=1}^n (\Phi_{e_o}^{-1})_{ik} (\mathbf{H}_{oo})_{kj} + \sum_{k=1}^n \sum_{l=1}^n (\mathbf{H}^*)_{ik} (\Phi_{e_o}^{-1})_{kl} (\mathbf{H}_{oo})_{lj}$. Clearly $\mathbf{S}_{ij} \neq 0$ since $[\mathbf{H}_{oo}]_{ij} \neq 0$, except for a few pathological cases that occur with Lebesgue measure zero. It follows by Assumption 2.4.3 that $\Im\{\mathbf{S}_{ij}\} \neq 0$, except over a set of Lebesgue measure zero.

A.11 Proof of Theorem 2.5.2:

First, notice that the following holds in (35) based on Lemma 6.

1. if $a_3 \neq 0$, then $a_3 \in \mathbb{R}$: the term a_3 corresponds to $l \rightarrow k \leftarrow i$, which implies that l and i are spouses.
2. if $b_3 \neq 0$, then $b_3 \in \mathbb{R}$: the term b_3 corresponds to $l \rightarrow k \leftarrow j$, which implies that l and j are spouses.
3. if $a_2 b_2 \neq 0$, then $a_2 b_2 \in \mathbb{R}$: the term $a_2 b_2$ corresponds to $i \rightarrow l \leftarrow j$, which implies that i and j are spouses. By similar argument, $\Im\{[\mathbf{H}_{ho}(l, i)]^* \Phi_{e_{h,t}}^{-1} \mathbf{H}_{ho}(l, j)\} = 0$

Hence, $\Im(a_2 b_2)$, $\Im(a_3)$ and $\Im(b_3)$ are zero. Then,

$$\Im(\mathbf{L})_{ij} = - \sum_{l=1}^{n_h} d_l [\Im(a_1 b_1) + \Im(a_1 b_2) + \Im(a_1 b_3) + \Im(a_2 b_1) + \Im(a_2 b_3) + \Im(a_3 b_1) + \Im(a_3 b_2)]$$

Remark A.11.1. Under Assumption 3, 4, for $i, j \in M_l$, such that (a) i, j are not strict parents or (b) i, j are not strict spouses of hidden node in $\mathcal{V}_h$, then from (A.15) it follows that $\Im\{\mathbf{L}_{ij}\} \neq 0$ almost always.

Remark A.11.2. Under Assumption 4, we apply Theorems 2 and 3, but by replacing $\mathbf{L}_{ij}$ in the definitions with $\Im\{\mathbf{L}_{ij}\}$ and $DM_h(i, j)$ with $DE_h(i, j)$. In other words, if $\Im\{\mathbf{L}_{ij}\}$ is not equal to zero, then there exists a $g \in DE_h(i, j)$ in $\mathcal{G}(\mathcal{V}, \mathcal{E})$.

($\Rightarrow$) From Assumption 2, it implies that $\mathcal{C}(l) \cup \mathcal{P}(l) \cup \mathcal{S}(l) \in \mathcal{V}_o$. We use contrapositive argument to show this, i.e., we will show that if $|\mathcal{P}(l) \setminus \{\mathcal{C}(l) \cup \mathcal{S}(l)\}| \leq 1$ and $|\mathcal{S}(l) \setminus \{\mathcal{C}(l) \cup \mathcal{P}(l)\}| \leq 1$, then $\nexists k \in M_l$ such that $\deg_{M_l}(k) < \alpha_l$. Here, we will show that for any node $a \in M_l$, $\deg_{M_l}(a) = |M_l| - 1$.

Case 1: $|\mathcal{P}(l) \setminus (\mathcal{C}(l) \cup \mathcal{S}(l))| = 0, |\mathcal{S}(l) \setminus \{\mathcal{C}(l) \cup \mathcal{P}(l)\}| = 0$:

Here, hidden node l does not have either strict parent and strict spouse, which implies that $M_l = \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$. Consider an observable node $a_1 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$; existence of a_1 is guaranteed from Assumption 1. Suppose there exists $a_2 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$, $a_2 \neq a_1$. If $a_1 \in \mathcal{C}(l)$, then clearly there exists a $g \in DE_l(a_1, a_2)$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$. Suppose $a_1 \in (\mathcal{P}(l) \cap \mathcal{S}(l)) \setminus \mathcal{C}(l)$. Then, again, from the definition of $DE_l(a_1, a_2)$, $\exists g \in DE_l(a_1, a_2)$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$. It follows from Remark A.11.1 that $\Im\{\mathbf{L}\}_{a_1 a_2} \neq 0$ and $\Im\{\mathbf{L}\}_{a_2 a_1} \neq 0$ almost everywhere, which implies $(a_1, a_2) \in Q_l$. Essentially, we have shown that for every $a_1 \in M_l$ and for every $k \in M_l \setminus \{a_1\}$, $(a_1, k) \in Q_l$, i.e., $\deg_{M_l}(a_1) = |M_l \setminus \{a_1\}| = |M_l| - 1$. The maximum degree is $\alpha_l = |M_l| - 1$, and thus, $\nexists k \in M_l$ such that $\deg_{M_l}(k) < \alpha_l$. We use the following conclusion in proof of Case 2, 3 and 4: Consider a node $a_1 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$. Then, for every $a_2 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l)) \setminus \{a_1\}$, $(a_1, a_2) \in Q_l$.

Case 2: $|\mathcal{P}(l) \setminus \{\mathcal{C}(l) \cup \mathcal{S}(l)\}| = 0, |\mathcal{S}(l) \setminus \{\mathcal{C}(l) \cup \mathcal{P}(l)\}| = 1$:

Here, there is no strict parent. One strict spouse s_1 is present, which implies that $M_l = \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l)) \cup s_1$. Consider any observable node $a_1 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$. Existence of a_1 is guaranteed by Assumption 1. It follows from Case 1 that $(a_1, a_2) \in Q_l, \forall a_2 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l)) \setminus a_1$. We next show that $(a_1, s_1) \in Q_l$. Since s_1 is strict spouse, there exists a $g \in DE_l(a_1, s_1)$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$. Then, by Remark A.11.1, $\Im\{\mathbf{L}\}_{a_1 s_1} \neq 0$ and $\Im\{\mathbf{L}\}_{s_1 a_1} \neq 0$ almost everywhere, which implies that $(a_1, s_1) \in Q_l$. Since $M_l = \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l)) \cup s_1$, we can conclude that for any $a_1 \in M_l \setminus s_1$, $(a_1, s_1) \in Q_l$. Moreover, $\deg_{M_l}(a_1) = |M_l| - 1$ and $\deg_{M_l}(s_1) = |M_l| - 1$. Note that, here, $\alpha_l = |M_l| - 1$. Thus, $\nexists k \in M_l$ such that $\deg_{M_l}(k) < \alpha_l$. We use the following conclusion in Case 4: Consider a node $s_1 \in \mathcal{S}(l) \setminus (\mathcal{C}(l) \cup \mathcal{P}(l))$, then for any $a_1 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$, $(s_1, a_1) \in Q_l$.

Case 3: $|\mathcal{P}(l) \setminus \{\mathcal{C}(l) \cup \mathcal{S}(l)\}| = 1, |\mathcal{S}(l) \setminus \{\mathcal{C}(l) \cup \mathcal{P}(l)\}| = 0$: Here, there is no strict spouse but one strict parent p_1 is present, which implies that $M_l = \mathcal{C}(l) \cup \{\mathcal{P}(l) \cap \mathcal{S}(l)\} \cup p_1$. Consider any observable node $a_1 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l))$. Existence of a_1 is guaranteed from the Assumption 1. From the conclusion of Case 1, it follows that for all $a_2 \in \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l)) \setminus \{a_1\}$, $(a_1, a_2) \in Q_l$. We next show that a_1 is also connected to p_1 . Since p_1 is strict parent, there exists a $g \in DE_l(a_1, p_1)$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$.

By Remark A.11.1, $\Im\{\mathbf{L}\}_{a_1 p_1} \neq 0$ and $\Im\{\mathbf{L}\}_{p_1 a_1} \neq 0$, which implies that $(a_1, p_1) \in Q_l$. Since $M_l = \mathcal{C}(l) \cup (\mathcal{P}(l) \cap \mathcal{S}(l)) \cup p_1$, we can conclude that $\deg_{M_l}(a_1) = |M_l| - 1$. Moreover, for any $a_1 \in M_l \setminus p_1$, $(a_1, p_1) \in Q_l$. Therefore, $\deg_{M_l}(p_1) = |M_l| - 1$. Here, $\alpha_l = |M_l| - 1$. Thus, $\nexists k \in M_l$ such that $\deg_{M_l}(k) < \alpha_l$.

Case 4 : $|\mathcal{P}(l) \setminus \{\mathcal{C}(l) \cup \mathcal{S}(l)\}| = 1, |\mathcal{S}(l) \setminus \{\mathcal{C}(l) \cup \mathcal{P}(l)\}| = 1$: Here, there is one strict parent p_1 , one strict spouse s_1 , which implies that $M_l = \mathcal{C}(l) \cup \{\mathcal{P}(l) \cap \mathcal{S}(l)\} \cup \{p_1, s_1\}$. Consider $a_1 \in \mathcal{C}(l) \cup \{\mathcal{P}(l) \cap \mathcal{S}(l)\}$. It follows from conclusion of Case 2 and 3 that $(a_1, s_1) \in Q_l$ and $(a_1, p_1) \in Q_l$. Thus, $\deg_{M_l}(a_1) = |M_l| - 1$. Moreover, $\exists g \in DE_l(p_1, s_1)$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$. It follows that $(p_1, s_1) \in Q_l$ and thus $\deg_{M_l}(p_1) = \deg_{M_l}(s_1) = |M_l| - 1$. Here, $\alpha_l = |M_l| - 1$ and hence, $\nexists k \in M_l$ such that $\deg_{M_l}(k) < \alpha_l$. This concludes the proof.

($\Leftarrow$) For the converse, we have from Assumption 1 that there exists an $i \in \mathcal{C}(l)$. Then, it follows that $\deg_{M_l}(i) = |M_l| - 1$. Thus, it suffices to show in each of the following cases that there exists a node $i \in M_l$ with $\deg_{M_l}(i) < |M_l| - 1$.

Suppose that $|\mathcal{P}(l) \setminus (\mathcal{C}(l) \cup \mathcal{S}(l))| \geq 2$. Then, there exist distinct $p_1, p_2 \in \mathcal{P}(l) \setminus (\mathcal{C}(l) \cup \mathcal{S}(l))$. Since p_1 and p_2 are strict parents in $\mathcal{G}(\mathcal{V}, \mathcal{E})$, there does not exist $g \in DE_h(p_1, p_2)$ for any $h \in \mathcal{V}_h$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$. Hence, by Remark A.11.1, $[\Im\{\mathbf{L}\}]_{p_1 p_2} = 0$, and $(p_1, p_2) \notin Q_l$. Thus, $\deg_{M_l}(p_1) < |M_l| - 1$ and $\deg_{M_l}(p_2) < |M_l| - 1$.

The similar proof holds if $|\mathcal{S}(l) \setminus (\mathcal{C}(l) \cup \mathcal{P}(l))| \geq 2$. Suppose that $|\mathcal{S}(l) \setminus (\mathcal{C}(l) \cup \mathcal{P}(l))| \geq 2$. Then, there exist distinct $s_1, s_2 \in \mathcal{S}(l) \setminus (\mathcal{C}(l) \cup \mathcal{P}(l))$. Since s_1 and s_2 are strict spouses, there does not exist $g \in DE_h(s_1, s_2)$ for any $h \in \mathcal{V}_h$ such that $g \in \mathcal{G}(\mathcal{V}, \mathcal{E})$. Hence, by Remark A.11.1, $[\Im\{\mathbf{L}\}]_{s_1 s_2} = 0$, and $(s_1, s_2) \notin Q_l$. Thus, $\deg_{M_l}(s_1) < |M_l| - 1$ and $\deg_{M_l}(s_2) < |M_l| - 1$, which concludes the proof.

A.11.1 LDG example:

An example illustrating a LDG, its topology, moral graph, and the definitions of strict parents and strict spouses, is shown in Fig. A.1.

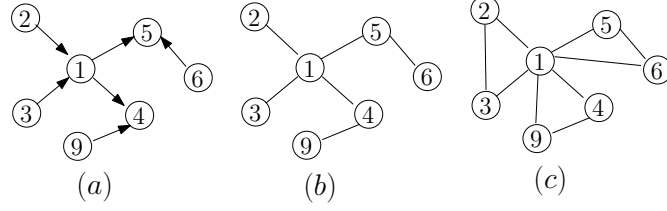


Figure A.1: (a) LDG $\mathcal{G}(\mathcal{V}, \mathcal{E})$, (b) $\text{top}(\mathcal{G})$ and (c) $\text{kin}(\mathcal{G})$. $\{2, 3\}$ are strict parents of 1. $\{6, 9\}$ are strict spouses of 1.

A.12 Proof of Proposition 2.6.1:

The truncation error in the estimation of the PSD matrix is given by,

$$\begin{aligned} \|\Phi_{\mathbf{x}}(z) - \bar{\Phi}_{\mathbf{x}}(z)\|_{\infty} &= \left\| \sum_{k=p+1}^{\infty} \left[R_{\mathbf{x}}(k)z^{-k} + R_{\mathbf{x}}(-k)z^k \right] \right\|_{\infty} \\ &\leq 2 \sum_{k=p+1}^{\infty} \|R_{\mathbf{x}}(k)\|_{\infty} \leq 2C_1 \frac{\rho^{p+1}}{1-\rho}, \end{aligned} \quad (\text{A.15})$$

where the inequality follows by triangle inequality, since $|z| = 1$. The lemma follows by letting $\varepsilon = 2C_1 \frac{\rho^{p+1}}{1-\rho}$.

A.13 Proof of Proposition 2.6.2:

Define $\mathbf{Z} = [\mathbf{x}_1^T, \dots, \mathbf{x}_N^T]$ and $R = \mathbb{E}\{\mathbf{Z}\mathbf{Z}^T\}$. Note that $\mathbf{Z}$ is obtained by stacking the observed time series along a column. This structure helps us in bounding the deviation of each individual elements.

The following lemmas are useful in the proof of Proposition 2.6.2.

Lemma A.13.1. *The following relation holds true for all matrices $A \in \mathbb{R}^{n \times n}$ [65]:*

$$\frac{1}{n} \|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{\infty} \leq \sqrt{n} \|\mathbf{A}\|_2. \quad (\text{A.16})$$

Lemma A.13.2. *For every symmetric matrix $\mathbf{S} \in \mathbb{R}^{nN \times nN}$ and for every $\varepsilon > 0$, the*

following bound holds for all $N \geq n$

$$\mathbb{P} \left(\frac{1}{N} \mathbf{Z}^T \mathbf{S} \mathbf{Z} > \frac{1}{N} \text{Tr}(\mathbf{R} \mathbf{S}) + \varepsilon \right) \leq \exp \left(-N \min \left\{ \frac{\varepsilon^2}{32 \|\mathbf{S}\|_2^2 n^2 C_1^2}, \frac{\varepsilon}{8 \|\mathbf{S}\|_2 n C_1} \right\} \right). \quad (\text{A.17})$$

Proof. The proof is similar to Lemma 4 in [126]. The lemma statement follows by plugging in the inequality $\|\mathbf{R}\| \leq n \|\mathbf{R}\|_\infty \leq n C_1$ from (A.16) and the bound on auto-correlation. $\square$

Lemma A.13.3. *For every $i, j \in \{1, \dots, n\}$, every $k \in \{0, \dots, p\}$, every $l \in \{1, \dots, N - k\}$, and every $N - k \geq n$, the following bound holds:*

$$\mathbb{P} \left(\left| [\widehat{R}_{\mathbf{x}}(k)]_{i,j} - [R_{\mathbf{x}}(k)]_{i,j} \right| > \varepsilon \right) \leq 2 \exp \left(-(N - k) \min \left\{ \frac{\varepsilon^2}{32 n^2 C_1^2}, \frac{\varepsilon}{8 n C_1} \right\} \right), \quad (\text{A.18})$$

where $[\widehat{R}_{\mathbf{x}}(k)]_{i,j} = \frac{1}{N-k} \sum_{l=1}^{N-k} x_i(l) x_j(l+k)$.

Proof. The idea behind the proof is to pick $\mathbf{S}$ such that $\mathbf{Z}^T \mathbf{S} \mathbf{Z} = \sum_{l=1}^{N-k} x_i(l) x_j(l+k)$. Designing $\mathbf{S}$ such that $\mathbf{S}_{lm} = 1$ if $\text{rem}(l, n) = i$ and $(\text{rem}(m - nk, n) = j) \& (m - nk > 0)$, where $\text{rem}(a, b)$ is the remainder of a/b will satisfy this condition. Then,

$$\text{Tr}\{\mathbf{S} \mathbf{R}\} = \text{Tr}\{\mathbf{S} \mathbb{E}\{\mathbf{Z} \mathbf{Z}^T\}\} = \mathbb{E}[\text{Tr}\{\mathbf{Z}^T \mathbf{S} \mathbf{Z}\}] = [R_{\mathbf{x}}(k)]_{ij}.$$

Lemma A.13.3 follows by applying Lemma A.13.2 and repeating the same for $-\mathbf{S}$ and for every i and j . $\square$

Remark A.13.1. *Note that in Lemma A.13.3 we have used $N - k$ instead of N . The bound still holds, since $N - K \leq N$.*

Now we can prove Proposition 2.6.2.

The result is obtained by applying union bound and Lemma A.13.3.

$$\begin{aligned} & \mathbb{P} \left(\|\widehat{R}_{\mathbf{x}}(k) - R_{\mathbf{x}}(k)\|_\infty > \varepsilon \right) \\ & \leq \mathbb{P} \left(\bigcup_{i,j=1}^n \left| [\widehat{R}_{\mathbf{x}}(k)]_{i,j} - [R_{\mathbf{x}}(k)]_{i,j} \right| > \varepsilon \right) \end{aligned} \quad (\text{A.19})$$

$$\leq \sum_{i,j=1}^n \mathbb{P} \left(\left| [\widehat{R}_{\mathbf{x}}(k)]_{i,j} - [R_{\mathbf{x}}(k)]_{i,j} \right| > \varepsilon \right), \quad (\text{A.20})$$

which gives the desired result by plugging in Lemma A.13.3.

A.14 Proof of Lemma 2.6.1:

From Proposition 2.6.2, we have

$$\mathbb{P}\left(\|\widehat{R}_{\mathbf{x}}(k) - R_{\mathbf{x}}(k)\|_{\infty} \leq \varepsilon\right) \geq 1 - n^2 \exp\left(-(N - k) \min\left\{\frac{\varepsilon^2}{32n^2C_1^2}, \frac{\varepsilon}{8nC_1}\right\}\right).$$

Then,

$$\sum_{k=-p}^p \|R_{\mathbf{x}}(k) - \widehat{R}_{\mathbf{x}}(k)\|_{\infty} \leq (2p + 1)\varepsilon \quad (\text{A.21})$$

with probability at least

$$1 - n^2 \exp\left(-(N - p) \min\left\{\frac{\varepsilon^2}{32n^2C_1^2}, \frac{\varepsilon}{8nC_1}\right\}\right).$$

Since ε is arbitrary, the lemma follows by picking appropriate epsilon and applying the inequality (26).

A.15 Proof of Theorem 2.6.2:

The following lemma from [65] is useful in deriving this.

Lemma A.15.1. *For any invertible matrices $\mathbf{A}$ and $\mathbf{B}$ with $\|\mathbf{A}^{-1}\|_2 \|\mathbf{B} - \mathbf{A}\|_2 < 1$, we have*

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_2 \leq \|\mathbf{A}^{-1}\|_2 \|\mathbf{A}\|_2^{-1} \|\mathbf{B} - \mathbf{A}\|_2 \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A}) \frac{\|\mathbf{B} - \mathbf{A}\|_2}{\|\mathbf{A}\|_2}}, \quad (\text{A.22})$$

where $\kappa(\mathbf{A})$ is the condition number of $\mathbf{A}$.

Let $\mathbf{A} = \Phi_{\mathbf{x}}(z)$ and $\mathbf{B} = \widehat{\Phi}_{\mathbf{x}}(z)$. Then, by applying Lemma A.15.1, we get

$$\|\Phi_{\mathbf{x}}^{-1} - \widehat{\Phi}_{\mathbf{x}}^{-1}\|_2 \leq \|\Phi_{\mathbf{x}}^{-1}\|_2 \|\Phi_{\mathbf{x}}\|_2^{-1} \|\widehat{\Phi}_{\mathbf{x}} - \Phi_{\mathbf{x}}\|_2 \frac{\kappa(\Phi_{\mathbf{x}})}{1 - \kappa(\Phi_{\mathbf{x}}) \frac{\|\widehat{\Phi}_{\mathbf{x}} - \Phi_{\mathbf{x}}\|_2}{\|\Phi_{\mathbf{x}}\|_2}}. \quad (\text{A.23})$$

From the definition (23), and by applying sub-multiplicative property of the spectral norm, $\|\Phi_{\mathbf{x}}^{-1}(z)\|_2 \leq \|(\mathbf{I} - H(z))^*\|_2 \|\Phi_e^{-1}(z)\|_2 \|(\mathbf{I} - H(z))\|_2 \leq \frac{1}{l^2 \sigma_{e[n_o]}^2}$, where $\sigma_{e[n_o]}$ is the smallest eigenvalue of $\Phi_e(z)$. Similarly, $\kappa(\Phi_{\mathbf{x}}) := \frac{\lambda_{max}(\Phi_{\mathbf{x}})}{\lambda_{min}(\Phi_{\mathbf{x}})} \leq \frac{L^2 \sigma_{e[1]}^2}{l^2 \sigma_{e[n_o]}^2}$.

Then,

$$\|\Phi_{\mathbf{x}}^{-1} - \hat{\Phi}_{\mathbf{x}}^{-1}\|_2 \leq \frac{L^2 \sigma_{e[1]}^2}{l^4 \sigma_{e[n_o]}^4} \left(\frac{\|\hat{\Phi}_{\mathbf{x}} - \Phi_{\mathbf{x}}\|_2}{l^2 \sigma_{e[n_o]}^2 - \|\hat{\Phi}_{\mathbf{x}} - \Phi_{\mathbf{x}}\|_2} \right). \quad (\text{A.24})$$

The theorem follows by applying (A.16).

Appendix B

Appendix of Chapter 3

B.1 Proof of Lemma 3.4.1

Let $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e})$ and let $i, j \in [n]$. By definition of $\mathcal{L}(\mathbf{H}, \mathbf{e})$, $e_i = \tilde{e}_i + \mathbf{F}_i \tilde{\mathbf{e}}_h$, where $\mathbf{F}_i$ denotes the i^{th} row of $\mathbf{F}$, for any $i \in [n]$. To prove the only if part, suppose that $\Phi_{e_i e_j} \neq 0$. By definition, $e_i e_j = (\tilde{e}_i + \mathbf{F}_i \tilde{\mathbf{e}}_h)(\tilde{e}_j + \mathbf{F}_j \tilde{\mathbf{e}}_h)$. Then, $\Phi_{e_i e_j} = \mathbf{F}_i \Phi_{\tilde{\mathbf{e}}_h} \mathbf{F}_j^* = \sum_k F_{ik} F_{jk}^* \Phi_{\tilde{e}_{h_k}}$, since $\tilde{e}_i$ and $\tilde{e}_j$ are uncorrelated for $i, j \in [n + L]$. Thus, $\sum_k F_{ik} F_{jk}^* \Phi_{\tilde{e}_{h_k}} \neq 0$, which implies that there exists a k such that $F_{ik} \neq 0$ and $F_{jk} \neq 0$. In other words, there exists a latent node h_k in the corresponding LDIG of $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}})$ such that $i, j \in Ch(h_k)$.

$\Leftarrow$ Let $i, j \in [n]$ such that $\Phi_{e_i e_j} = 0$. Then, from the proof of only if part, $0 = \mathbf{F}_i \Phi_{\tilde{\mathbf{e}}_h} \mathbf{F}_j^* = \sum_k F_{ik} F_{jk}^* \Phi_{\tilde{e}_{h_k}}$, which implies that $F_{ik} = 0$ or $F_{jk} = 0$, $\forall k$, except for a few pathological cases that occur with Lebesgue measure zero. We ignore the pathological cases here. Hence, there does not exist any latent node h such that both $i, j \in Ch(h)$. $\square$

The following result shows that if a subgraph, $\mathcal{G}^\ell(\mathcal{V}^\ell, \mathcal{E}^\ell)$, of the correlation graph, $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$, forms a maximal clique in $\mathcal{G}_c$, then for any transformed LDIG in $\mathcal{L}(\mathbf{H}, \mathbf{e})$, the set of nodes in $\mathcal{V}^\ell$ is equal to the set of children of some latent nodes in the LDIG.

B.2 Proof of Lemma 3.4.2

The following lemma is useful in proving Lemma 3.4.2.

Lemma B.2.1. *Let $(\mathbf{H}, \mathbf{e})$ be an LDIM defined by (3.1) which satisfies Assumption*

3.4.1, and let $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ be the correlation graph of the exogenous noise sources, $\mathbf{e}$. Suppose that $\mathcal{G}^\ell(\mathcal{V}^\ell, \mathcal{E}^\ell) \subseteq \mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$ is a maximal clique with $|\mathcal{V}^\ell| > 1$. Then, for every LDIG $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e})$, there exist latent nodes $h_1^\ell, \dots, h_{k_\ell}^\ell$ such that

$$\mathcal{V}^\ell = \bigcup_{i=1}^{k_\ell} Ch(h_i^\ell) \text{ and } \mathcal{E}^\ell = \bigcup_{i=1}^{k_\ell} \mathcal{E}_{\ell,i}, \quad (\text{B.1})$$

where $\mathcal{E}_{\ell,i} := \{(k, j) : k, j \in Ch(h_i^\ell)\}$.

In particular, for any latent node h in the LDIG $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}})$, $Ch(h)$ forms a clique (not necessarily maximal) that is restricted to $\mathcal{G}^\ell(\mathcal{V}^\ell, \mathcal{E}^\ell)$ in $\mathcal{G}_c$.

Proof. Let $\mathcal{V}^\ell \subseteq [n]$, $|\mathcal{V}^\ell| > 1$, such that $\mathcal{V}^\ell$ forms a clique in $\mathcal{G}_c$. Lemma 3.4.1 showed that, for any $i, j \in \mathcal{V}^\ell$, there exists a latent node h such that $i, j \in Ch(h)$ in the LDIG, $\mathcal{G}$, of $(\tilde{\mathbf{H}}, \tilde{\mathbf{e}})$. Since this is true for any pair $i, j \in \mathcal{V}^\ell$, there exists a minimal set of latent nodes $\mathcal{H}^\ell := \{h_i^l\}_{i=1}^{k_l}$, $k_l \in \mathbb{N} \setminus \{0\}$, s. t. for any $i, j \in \mathcal{V}^\ell$, we have $i, j \in Ch(h_p^l)$ for some $h_p^l \in \mathcal{H}$, in $\mathcal{G}$. Hence, $\mathcal{V}^\ell \subseteq \bigcup_{i=1}^{k_l} Ch(h_i^l)$. Similarly, $i, j \in Ch(h_p^l)$ implies $(i, j) \in \mathcal{E}_{l,p}$. Therefore, $\mathcal{E}^\ell \subseteq \bigcup_{i=1}^{k_l} \mathcal{E}_{l,i}$.

Next, we prove that all the children of a given latent node belong to a single (maximal) clique, which proves $\mathcal{V}^\ell \supseteq \bigcup_{i=1}^{k_l} Ch(h_i^l)$ and $\mathcal{E}^\ell \supseteq \bigcup_{i=1}^{k_l} \mathcal{E}_{l,i}$. Let $h \in \mathcal{H}^\ell$ be a latent node in the LDIG of $(\tilde{\mathbf{H}}, \tilde{\mathbf{e}})$ and suppose $i, j \in Ch(h)$. Then, from the definition of $(\tilde{\mathbf{H}}, \tilde{\mathbf{e}})$, there exists ℓ such that $F_{i\ell} \neq 0$ and $F_{j\ell} \neq 0$, and hence $\Phi_{e_{ie_j}} = \mathbf{F}_i \Phi_{\tilde{\mathbf{e}}_h} \mathbf{F}_j^* \neq 0$ a.e. Thus, $(i, j) \in \mathcal{E}_c$, excluding the pathological cases. Notice that this is true for any $i, j \in Ch(h)$. Therefore, $Ch(h) \subseteq \mathcal{V}^\ell$ forms a clique (not necessarily maximal) in $\mathcal{G}_c(\mathcal{V}, \mathcal{E}_c)$. Since this is true for every $h \in \mathcal{H}^\ell$, $\mathcal{V}^\ell \supseteq \bigcup_{i=1}^{k_l} Ch(h_i^l)$. The similar proof shows that $\mathcal{E}^\ell \supseteq \bigcup_{i=1}^{k_l} \mathcal{E}_{l,i}$. $\square$

The proof of Lemma 3.4.2 follows directly from (B.1) and by noting that $k_\ell \geq 1$ $\square$

B.3 Proof of Theorem 3.4.1

Lemma B.3.1. (*Pigeonhole principle*): The pigeonhole principle states that if n pigeons are put into m pigeonholes, with $n > m$, then at least one hole must contain more than one pigeon.

We use pigeonhole principle and Lemma B.2.1 to prove this via contrapositive argument. Recall that the number of latent nodes L is equal to the number of cliques M . Suppose there exists a clique $\mathcal{G}^l(\mathcal{V}^l, \mathcal{E}^l)$ such that, for some $i, j \in \mathcal{V}^l$, there does not exist

a latent node $h \in (\tilde{\mathbf{H}}, \tilde{\mathbf{e}})$ with $i, j \in Ch(h)$. By Lemma B.2.1, there exist latent nodes h, h' such that $i \in Ch(h)$ and $j \in Ch(h')$. By Lemma B.2.1 again, all the children of h are included in a single clique. That is, $Ch(h) \subseteq \mathcal{V}^l$ and $Ch(h') \subseteq \mathcal{V}^l$, since $i, j \in \mathcal{V}^l$. Then, excluding $\mathcal{V}^l$, h , and h' , we are left with $M - 1$ cliques and $L - 2 = M - 2$ latent nodes. Applying pigeonhole principle, with $M - 1$ pigeons (cliques) and $L - 2$ holes (latent nodes), there would exist at least one latent node k with $Ch(k)$ belonging to two different maximal cliques, which is a contradiction of Lemma B.2.1. $\square$

B.4 Proof of Proposition 3.4.1

Let $T_1, T_2 \in \{0, 1\}^{(n+L) \times (n+L)}$, $T_1 \neq T_2$, be the topologies of two distinct transformations $(\tilde{\mathbf{H}}^1, \tilde{\mathbf{e}}^1)$ and $(\tilde{\mathbf{H}}^2, \tilde{\mathbf{e}}^2)$ respectively. Without loss of generality, let (i, j) be such that $[T_1]_{ij} \neq 0$ and $[T_2]_{ij} = 0$. By the definition of $([\mathbf{H}, \mathbf{F}], \tilde{\mathbf{e}})$, if $i \leq n$ and $j \leq n$, then $[T_1]_{ij} = [T_2]_{ij} = \mathbb{1}_{\{H_{ij} \neq 0\} \cup \{H_{ji} \neq 0\}}$. If $i \leq n$ and $j \geq n$, then i is an observed node and j is a latent node. From Lemma 3.4.1, $F_{ij}^1 = F_{ij}^2 = 0$ if and only if $j \notin Pa(i)$. Thus, $[T_1]_{ij} = [T_2]_{ij}$, which is a contradiction, since both cannot be true. Similar contradiction holds if $i \geq n$ and $j \leq n$. If $i, j > n$, then $[T_1]_{ij} = [T_2]_{ij} = 0$ since $Pa(i) = Pa(j) = \emptyset$. Thus, the assumption leads to a contradiction, which implies that $T_1 = T_2$. $\square$

B.5 Proof of Proposition 3.5.1

Consider a pair of monomials y_k, y_l with $y_k = \mathbf{v}(0)^\alpha$ and $y_l = \mathbf{v}(0)^\beta$. For notational convenience, the index 0 is omitted. Then, $\mathbb{E}\{y_k y_l\} = \mathbb{E}\{\mathbf{v}^{\alpha+\beta}\} = \prod_{i=1}^m \mathbb{E}\{\mathbf{v}_i^{\alpha_i+\beta_i}\}$. By (3.7), $\mathbb{E}\{y_k y_l\} \neq 0$ if and only if $\alpha_i + \beta_i$ is even, $\forall i \in [m]$. Suppose α_i is odd. Then, β_i must be odd. Similarly, β_i must be even if α_i is even, $\forall i \in [m]$.

Define an element-wise boolean operator, $\mathcal{B} : \mathbb{N}^m \mapsto \{0, 1\}^m$ such that for $\mathbf{u} = \mathcal{B}(\alpha)$, $u_i = 0$ if α_i is odd and $u_i = 1$ if α_i is even. Then, for $y_k = \mathbf{v}^\alpha$ and $y_l = \mathbf{v}^\beta$, $\mathbb{E}\{y_k y_l\} \neq 0$ if and only if $\mathcal{B}(\alpha) = \mathcal{B}(\beta)$. Group the monomials with the the same odd-even pattern into one cluster. Since the total different values that $\mathcal{B}(\cdot)$ can take is 2^m , there are 2^m distinct clusters that are uncorrelated with each other.

Reorder $\mathbf{y}$ by grouping the monomials belonging to the same cluster together to obtain $\tilde{\mathbf{y}}$, similar to the $(n, p) = (2, 3)$ example in Section 3.5.2. Then, $\Phi_{\tilde{\mathbf{y}}}$ is a block diagonal matrix with 2^m blocks for m variable polynomials, where each diagonal block corresponds to one particular pattern of $\{Odd, Even\}^m$. $\square$

B.6 Proof of Lemma 3.5.1

Let $(\mathbf{H}, \mathbf{F}, \tilde{\mathbf{e}}) \in \mathcal{L}(\mathbf{H}, \mathbf{e})$ and let $i, j \in [n]$. By definition of $\mathcal{L}(\mathbf{H}, \mathbf{e})$, $e_i = \tilde{e}_i + \mathbf{F}_i \tilde{\mathbf{e}}_h$, where $\mathbf{F}_i$ denotes the i^{th} row of $\mathbf{F}$, for any $i \in [n]$. To prove the only if part, suppose that $\Phi_{e_i e_j} \neq 0$. By definition, $e_i e_j = (\tilde{e}_i + \mathbf{F}_i \mathbf{y})(\tilde{e}_j + \mathbf{F}_j \mathbf{y})$. Since $\tilde{\mathbf{e}}_i, \tilde{\mathbf{e}}_j$, and $\tilde{\mathbf{e}}_h$ are uncorrelated, and $\Phi_{\mathbf{y}}$ is block diagonal (Proposition 3.5.2),

$$\Phi_{e_i e_j} = \mathbf{F}_i \Phi_{\mathbf{y}} \mathbf{F}_j^* = \sum_{c=1}^{2^m} \sum_{k_1, k_2 \in \mathcal{C}_c} F_{ik_1} F_{jk_2}^* \Phi_{\mathbf{y}_{k_1} \mathbf{y}_{k_2}} \neq 0.$$

Thus, there exists a c such that $F_{ik_1} \neq 0$ and $F_{jk_2} \neq 0$ for some $k_1, k_2 \in \mathcal{C}_c$. That is, there exists a cluster c such that $i, j \in Ch(c)$.

$\Leftarrow$ Let $i, j \in [n]$ such that $\Phi_{e_i e_j} = 0$. Then, from the proof of only if part, $\sum_{c=1}^{2^m} \sum_{k_1, k_2 \in \mathcal{C}_c} F_{ik_1} F_{jk_2}^* \Phi_{\mathbf{y}_{k_1} \mathbf{y}_{k_2}} = 0$. That is, $F_{ik} = F_{jk} = 0$, $\forall k \in \mathcal{C}_c$, $\forall c \in [2^m]$, except for a few pathological cases that occur with Lebesgue measure zero. We ignore the pathological cases here. Hence, there does not exist any cluster c such that both $i, j \in Ch(c)$. $\square$

B.7 Proof of Theorem 3.5.1

The proof follows similar to the proof of Lemma B.2.1, by replacing latent nodes with latent cluster, as in the proof of Lemma 3.5.1. $\square$

B.8 Proof of Theorem 3.6.1

As shown in Lemma 3.6.1, if $\deg_{\max}(\Im\{\mathbf{S}\})\text{inc}(\Im\{\mathbf{L}\}) < 1/12$, then for the appropriately selected t , the convex program (2.10) retrieves $\Im\{\mathbf{S}\}$ and $\Im\{L\}$ when $\mathbf{C} = \Im\{\mathbf{S}\} + \Im\{\mathbf{L}\}$. If any one of the LDIMs $(\tilde{\mathbf{H}}, \mathbf{e}) \in \mathcal{L}_q(\mathbf{H}, \mathbf{e})$ satisfies this condition, then the imaginary part of $\hat{\mathbf{S}}_t = (\mathbf{I}_n - \mathbf{H}^*) \Phi_{\tilde{\mathbf{e}}_o}^{-1} (\mathbf{I}_n - \mathbf{H})$ returns the topology among the observed node of $(\tilde{\mathbf{H}}, \mathbf{e})$, by Lemma 3.6.3. The theorem follows from Remark 3.4.5 and Lemma B.2.1. $\square$

Appendix C

Appendix of Chapter 4

C.1 Proof of Lemma 4.3.2

Let $C \subseteq V \setminus \{j\}$ be an ancestral set and let $D = nd(j) \setminus C$. Then,

$$\mathbf{x}_j(\omega) = H_{jC}(\omega)X_C(\omega) + H_{jD}(\omega)X_D(\omega) + e_j(\omega).$$

Applying $\Phi_{e_j C} = \Phi_{e_j D} = 0$, we obtain $\Phi_{jC}(\omega) = H_{jC}(\omega)\Phi_{CC}(\omega) + H_{jD}(\omega)\Phi_{DC}(\omega)$ and

$$\begin{aligned} \Phi_j(\omega) &= H_{jC}(\omega)\Phi_C H_{jC}(\omega)^* + H_{jD}(\omega)\Phi_D H_{jD}(\omega)^* + H_{jC}(\omega)\Phi_{CD}(\omega)H_{jD}(\omega)^* \\ &\quad + H_{jD}(\omega)\Phi_{DC}(\omega)H_{jC}(\omega)^* + \Phi_{e_j e_j}(\omega). \end{aligned}$$

Then

$$f(j, C, \omega) = \Phi_j - \Phi_{jC}\Phi_C^{-1}\Phi_{Cj} = \Phi_{e_j e_j} + H_{jD}(\Phi_D - \Phi_{DC}\Phi_{CC}^{-1}\Phi_{CD})H_{jD}^*.$$

Notice that when $Pa(j) \subseteq C$, $H_{jD} = 0$, and $f(j, C, \omega) = \Phi_{e_j e_j}$, which shows the first part.

To prove the second part, suppose $Pa(j) \cap D \neq \emptyset$. We need to show that $H_{jD}(\Phi_D -$

$\Phi_{DC}\Phi_{CC}^{-1}\Phi_{CD})\mathbf{H}_{jD}^* > 0$. Let $A = nd(j) = C \cup D$ and $B = desc(j) \cup \{j\}$. From [21, 1],

$$\begin{aligned}\Phi_{AA}^{-1} &= \mathbf{S} + \mathbf{L}, \text{ where} \\ \mathbf{S} &= (\mathbf{I}_A - \mathbf{H}_{AA}^*)\Phi_{e_A}^{-1}(\mathbf{I}_A - \mathbf{H}_{AA}), \\ \mathbf{L} &= \mathbf{H}_{BA}\Phi_{e_B}^{-1}\mathbf{H}_{BA} - \Psi^*\Lambda^{-1}\Psi, \\ \Psi &= \mathbf{H}_{AB}^*\Phi_{e_A}^{-1}(\mathbf{I} - \mathbf{H}_{AA}) + (\mathbf{I} - \mathbf{H}_{BB}^*)\Phi_{e_B}^{-1}\mathbf{H}_{BA}, \text{ and} \\ \Lambda &= \mathbf{H}_{AB}^*\Phi_{e_A}^{-1}\mathbf{H}_{AB} + (\mathbf{I} - \mathbf{H}_{BB}^*)\Phi_{e_B}^{-1}(\mathbf{I} - \mathbf{H}_{BB}).\end{aligned}$$

Notice that since B is the set of descendants of j , $\mathbf{H}_{AB} = 0$, as cycles can be formed otherwise. Then, $\mathbf{L} = 0$ and $\Phi_{AA}^{-1} = (\mathbf{I}_A - \mathbf{H}_{AA}^*)\Phi_{e_A}^{-1}(\mathbf{I}_A - \mathbf{H}_{AA})$.

$$\Phi_{AA}^{-1} = \begin{bmatrix} \Phi_{DD} & \Phi_{DC} \\ \Phi_{CD} & \Phi_{CC} \end{bmatrix}^{-1} = \begin{bmatrix} K_{DD} & K_{DC} \\ K_{CD} & K_{CC} \end{bmatrix} = \frac{1}{\sigma} (\mathbf{I} - \mathbf{H}_{AA}^*) (\mathbf{I} - \mathbf{H}_{AA})$$

By Scur's complement, $(\Phi_D - \Phi_{DC}\Phi_{CC}^{-1}\Phi_{CD})^{-1} = K_{DD} = \frac{1}{\sigma}(\mathbf{I}_D - \mathbf{H}_{DD}^* - \mathbf{H}_{DD} + (\mathbf{H}_{AA}^*\mathbf{H}_{AA})_{D \times D})$. Moreover,

$$\mathbf{H}_{AA} = \begin{bmatrix} \mathbf{H}_{DD} & \mathbf{H}_{DC} \\ \mathbf{H}_{CD} & \mathbf{H}_{CC} \end{bmatrix} \text{ and } (\mathbf{H}_{AA}^*\mathbf{H}_{AA})_{D \times D} = \mathbf{H}_{DD}^*\mathbf{H}_{DD} + \mathbf{H}_{CD}^*\mathbf{H}_{CD}.$$

Since C is ancestral, $\mathbf{H}_{CD} = 0$ and

$$K_{DD} = \frac{1}{\sigma}(\mathbf{I}_D - \mathbf{H}_{DD})^*(\mathbf{I}_D - \mathbf{H}_{DD}).$$

Since G is a DAG, the rows and columns of $\mathbf{H}$ can be rearranged to obtain a lower triangular matrix with zeros on the diagonal. Thus eigenvalues of $(\mathbf{I}_D - \mathbf{H}_{DD})$ and its inverse are all ones. Hence minimum eigenvalue of K_{DD}^{-1} is greater than σ . Applying Rayleigh Ritz theorem on $\mathbf{H}_{jD}K_{DD}^{-1}\mathbf{H}_{jD}^*$, we have

$$\mathbf{H}_{jD}(\Phi_D - \Phi_{DC}\Phi_{CC}^{-1}\Phi_{CD})\mathbf{H}_{jD}^* = \mathbf{H}_{jD}K_{DD}^{-1}\mathbf{H}_{jD}^* \geq \sigma|D|\beta^2 \quad (\text{C.1})$$

which is strictly greater than zero if D is non-empty. $\square$

C.2 Proof of Lemma 4.3.6

The proof is done in two steps. First, we show that $\mathcal{S}$ in Algorithm 1 is a topological ordering. Then, we show that step (4) in Algorithm 1 can identify the parents of every node in G . The first step is shown via induction. Since $|\widehat{f}(i, C, \omega) - f(i, C, \omega)| < \Delta/4$ for empty set, $|\Phi_{ii} - \widehat{\Phi}_{ii}| < \Delta/4$ for every i . Recall from Lemma 3.3 that $\Phi_{ii} - \Phi_{jj} > \Delta$ if i is a source node and j is a non-source node. Then, $\widehat{\Phi}_{jj} \geq \Phi_{jj} - \Delta/4 \geq \Phi_{ii} + 3\Delta/4 \geq \widehat{\Phi}_{ii} + \Delta/2$. Thus, $i \in \arg \min_{1 \leq k \leq p} \widehat{\Phi}_{kk}$ if and only if $i \in \arg \min_{1 \leq k \leq p} \Phi_{kk}$ and thus $\mathcal{S}_1$ is always a source node.

For the induction step, assume that $\mathcal{S}_1, \dots, \mathcal{S}_n$ forms a correct topologically ordered set w.r.t. G . Let $C \subseteq \mathcal{S}(1 : n)$. If $Pa(i) \subseteq C$ and $Pa(j) \not\subseteq C$, then by applying Lemma 3.3, $\widehat{f}(j, C, \omega) > f(j, C, \omega) - \Delta/4 \geq \sigma + 3\Delta/4 = f(i, C, \omega) + 3\Delta/4 \geq \widehat{f}(i, C, \omega) + \Delta/2$. Thus, $i \in \arg \min_{k \in V \setminus \mathcal{S}} \widehat{f}(k, C, \omega)$ if and only if $i \in \arg \min_{k \in V \setminus \mathcal{S}} f(k, C, \omega)$ and thus $(\mathcal{S}, \mathcal{S}_{n+1})$ forms a topological order w.r.t. G , by Lemma 3.6.

To prove the second step, let $C \subseteq \mathcal{S}(1 : i)$. Since $\mathcal{S}(1 : i)$ is a valid topological ordering, $Pa(i) \subseteq \mathcal{S}(1 : i - 1)$. Let $k \in Pa(i)$ and let $D = C \setminus \{k\}$. Then, as shown in Corollary 3.4 $f(i, C, \omega) - f(i, D, \omega) \geq \Delta$, and

$$\begin{aligned} \Delta &\leq |f(i, C, \omega) - f(i, D, \omega)| \leq |f(i, C, \omega) - \widehat{f}(i, C, \omega)| + |\widehat{f}(i, C, \omega) - \widehat{f}(i, D, \omega)| \\ &\quad + |\widehat{f}(i, D, \omega) - f(i, D, \omega)| \\ &< \Delta/4 + \Delta/4 + |\widehat{f}(i, C, \omega) - \widehat{f}(i, D, \omega)| \\ \implies |\widehat{f}(i, C, \omega) - \widehat{f}(i, D, \omega)| &> \Delta/2. \end{aligned}$$

Suppose $k \notin Pa(i)$ but $k \in \mathcal{S}(1 : i)$. Then, for $D = C \setminus \{k\}$, $f(i, C, \omega) - f(i, D, \omega) = 0$. Repeating the same series of inequalities above by exchanging f and $\widehat{f}$, we obtain $|\widehat{f}(i, C, \omega) - \widehat{f}(i, D, \omega)| < \Delta/2$.

Thus, from the set $\mathcal{S}$, for every node $\mathcal{S}_i$, one can check nodes $\mathcal{S}_1, \dots, \mathcal{S}_{i-1}$ and verify if the difference of including and excluding the node is greater than $\Delta/2$. If the difference is greater than $\Delta/2$ for some k , then k is a parent of i , and if not, then the node is not a parent of i . That is, let $C_i = \{\mathcal{S}_1, \dots, \mathcal{S}_{i-1}\}$, $i > 1$, and let

$$\widehat{P}_i := \left\{ j \in C_i \mid |\widehat{f}(\mathcal{S}_i, C_i, \omega) - \widehat{f}(\mathcal{S}_i, C_i \setminus \{j\}, \omega)| > \Delta/2 \right\}.$$

Then, $Pa(i) = \widehat{P}_i$. □

C.3 Proof of Theorem 4.4.1

By the variational form of spectral norm [65],

$$\|Q\| = \sup_{v \in \mathbb{C}^p, \|v\|=1} |v^* Q v|,$$

where the max is taken over a p -dimensional unit complex sphere, $\mathbb{S}^p := \{v \in \mathbb{C}^p : \|v\|_2 = 1\}$. The first step here is to reduce supremum to finite maximization using finite covers of a unit ball, which is done using a δ cover. A δ -cover of a set $\mathcal{A}$ is a set $v^1, \dots, v^m$ such that for every $v \in \mathcal{A}$, there exists an $i \in 1, \dots, m$ such that $\|v^i - v\|_2 \leq \delta$. The following Lemma is obtained by extending example 5.8 in [117] to the complex field.

Lemma C.3.1. *Let $v^1, \dots, v^m$ be a δ -covering of the unit sphere $\mathbb{S}^p$. Then there exists such a covering with $m \leq (1 + 2/\delta)^{2p}$ vectors.*

Proof. The proof follows by extending (5.9) in [117], to the complex field. $\square$

Let $v \in \mathbb{S}^p$ and let v^j be such that $v = v^j + \Delta$, where $\|\Delta\| \leq \delta$. Then, $v^* Q v = (v^j)^* Q v^j + 2\Re\{\Delta^* Q v^j\} + \Delta^* Q \Delta$. Applying triangle inequality,

$$\begin{aligned} |v^* Q v| &\leq |(v^j)^* Q v^j| + 2\|\Delta\| \|Q\| \|v^j\| + \|\Delta\|^2 \|Q\| \\ &\leq |(v^j)^* Q v^j| + 2\delta \|Q\| + \delta^2 \|Q\| \\ &\leq |(v^j)^* Q v^j| + \frac{1}{2} \|Q\| \text{ for } \delta \leq 0.22474. \end{aligned}$$

Thus,

$$\begin{aligned} \|Q\| &= \max_{v \in \mathbb{S}^p} |v^* Q v| \leq \max_{j=1, \dots, m} |(v^j)^* Q v^j| + \frac{1}{2} \|Q\| \text{ and} \\ \|Q\| &\leq 2 \max_{j=1, \dots, m} |(v^j)^* Q v^j| \end{aligned}$$

Next, we find an upper bound for $\mathbb{E}[e^{\lambda \|Q\|}]$, which is treated with Chernoff-type bounding technique to obtain the desired result.

$$\begin{aligned} \mathbb{E}[e^{\lambda \|Q\|}] &\leq \mathbb{E}\left[\exp\left(2\lambda \max_{j=1, \dots, m} |(v^j)^* Q v^j|\right)\right] \\ &\leq \sum_{j=1}^m \mathbb{E}\left[e^{2\lambda (v^j)^* Q v^j}\right] + \mathbb{E}\left[e^{-2\lambda (v^j)^* Q v^j}\right] \end{aligned} \tag{C.2}$$

Next, we complete the proof for the restart and record sampling and the continuous sampling separately.

C.3.1 Restart and Record Sampling

Under the restart and record sampling settings, for any given $\omega \in \Omega$, $\{\mathbf{x}^r(\omega)\}_{r=1}^n$ is i.i.d. Thus

$$\begin{aligned}
\mathbb{E} [\exp (t(v^j)^* Q v^j)] &= \mathbb{E} \left[\exp \left(t(v^j)^* (\widehat{\Phi}_x(\omega) - \widetilde{\Phi}_x(\omega)) v^j \right) \right] \\
&= \mathbb{E} \left[\exp \left(\frac{t}{n} \sum_{r=1}^n (v^j)^* \mathbf{x}^r(\omega) [\mathbf{x}^r(\omega)]^* v^j - (v^j)^* \widetilde{\Phi}_x(\omega) v^j \right) \right] \\
&= \prod_{r=1}^n \mathbb{E} \left[\exp \left(\frac{t}{n} (v^j)^* \mathbf{x}^r(\omega) [\mathbf{x}^r(\omega)]^* v^j - (v^j)^* \widetilde{\Phi}_x(\omega) v^j \right) \right] \\
&= \left(\mathbb{E} \left[\exp \left(\frac{t}{n} (v^j)^* \mathbf{x}^1(\omega) [\mathbf{x}^1(\omega)]^* v^j - (v^j)^* \widetilde{\Phi}_x(\omega) v^j \right) \right] \right)^n \\
&= \left(\mathbb{E} \left[\exp \left(\frac{t}{n} |v^* \mathbf{x}^r(\omega)|^2 - v^* \widetilde{\Phi}_x(\omega) v \right) \right] \right)^n
\end{aligned}$$

Let $\varepsilon \in \{-1, +1\}$ be a Rademacher variable independent of $\mathbf{x}^r$. It can be shown that Proposition 4.11 in [117] will hold for complex numbers also. Then

$$\mathbb{E}_{\mathbf{x}^r(\omega)} \left[\exp \left(\frac{t}{n} |v^* \mathbf{x}^r(\omega)|^2 - v^* \widetilde{\Phi}_x(\omega) v \right) \right] \leq \mathbb{E}_{\mathbf{x}^r(\omega), \varepsilon} \left[\exp \left(\frac{2t\varepsilon}{n} |v^* \mathbf{x}^r(\omega)|^2 \right) \right] \quad (\text{C.3})$$

$$= \sum_{k=0}^{\infty} \frac{(2t/n)^{2k}}{2k!} \mathbb{E} \left[|v^* \mathbf{x}^r(\omega)|^{4k} \right] \quad (\text{C.4})$$

Recall that $\widetilde{\Phi}_{\mathbf{x}}$ is a positive definite matrix and $v^* \mathbf{x}^r \sim N(\mathbf{0}, \eta)$, where $\eta = v^* \widetilde{\Phi}_{\mathbf{x}} v \leq \lambda_{\max}(\widetilde{\Phi}_{\mathbf{x}}) \leq M$. The even moments of $y \sim N(0, \eta)$ is given by $\mathbb{E}\{y^{2k}\} = \eta^{2k} (2k-1)!! = \frac{(2k)!}{2^k k!} \eta^{2k}$. Then

$$\mathbb{E} \left[|v^* \mathbf{x}^r(\omega)|^{4k} \right] \leq \frac{(4k)!}{2^{2k} (2k)!} M^{2k}.$$

Therefore using the inequality $(4k)! \leq 2^{2k}[(2k)!]^2$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}^r(\omega)} \left[\exp \left(\frac{t}{n} |v^* \mathbf{x}^r(\omega)|^2 - v^* \tilde{\Phi}_x(\omega) v \right) \right] &\leq 1 + \sum_{k=1}^{\infty} \frac{(2t/n)^{2k}}{2k!} \frac{(4k)!}{2^{2k}(2k)!} M^{2k} \\
&\leq 1 + \sum_{k=1}^{\infty} \frac{(2t/n)^{2k}}{2k!} \frac{2^{2k}[(2k)!]^2}{2^{2k}(2k)!} M^{2k} \\
&= 1 + \sum_{k=1}^{\infty} \left(\frac{2Mt}{n} \right)^{2k} = \frac{1}{1 - \left(\frac{2Mt}{n} \right)^2} \\
&\leq \exp \left(\frac{8M^2 t^2}{n^2} \right)
\end{aligned}$$

whenever $\frac{2Mt}{n} < 3/4$, where the final inequality follows by applying $1 - x \geq e^{-2x}$ for $x \in [0, 3/4]$ (to be precise 0.77). Thus,

$$\mathbb{E} [\exp (t(v^j)^* Q v^j)] \leq \exp \left(\frac{8M^2 t^2}{n} \right), \quad \forall |t| \leq \frac{3n}{8M}.$$

Applying Lemma C.3.1 and the bound $2m \leq 2(1 + 2/0.22474)^{2p} \leq 2e^{4.6p} \leq e^{5p+0.693} \leq e^{6p}$,

$$\begin{aligned}
\text{From (C.2), } \mathbb{E} [e^{\lambda \|Q\|}] &\leq \mathbb{E} \left[\exp \left(2\lambda \max_{j=1, \dots, m} |(v^j)^* Q v^j| \right) \right] \\
&\leq \sum_{j=1}^m \mathbb{E} [e^{2\lambda (v^j)^* Q v^j}] + \mathbb{E} [e^{-2\lambda (v^j)^* Q v^j}] \\
&\leq 2m \exp \left(\frac{32M^2 \lambda^2}{n} \right) \\
&\leq \exp \left(\frac{32M^2 \lambda^2}{n} + 6p \right), \quad \forall |\lambda| \leq \frac{3n}{16M}.
\end{aligned}$$

Applying Chernoff-type bounding approach,

$$\mathcal{P} (\|Q\| \geq t) \leq e^{-\lambda t} \mathbb{E} [e^{\lambda \|Q\|}] \leq \exp \left(-\lambda t + \frac{32M^2 \lambda^2}{n} + 6p \right), \quad \forall |\lambda| \leq \frac{3n}{16M}.$$

The tightest bound is given by $g^*(t) := \inf_{|\lambda| \leq \frac{3n}{16M}} \left\{ -\lambda t + \frac{32M^2 \lambda^2}{n} + 6p \right\}$, where the objective is convex. Taking derivative w.r.t. λ and equating to zero, $\lambda^* = \frac{tn}{64M^2}$ and

$g^* = -\frac{t^2 n}{64M^2} + \frac{32M^2}{n} \frac{t^2 n^2}{64^2 M^4} + 6p = 6p - \frac{t^2 n}{128M^2}$, if t is such that $t \leq 12M$, which is reasonable as we can always pick $M \geq 1$.

Thus, $\mathcal{P}(\|\mathbf{Q}\| \geq t) \leq \exp\left(-\frac{t^2 n}{128M^2} + 6p\right)$ The theorem statement follows. $\square$

C.3.2 Continuous Sampling

In the continuous sampling setting, the samples $\check{\mathbf{x}}(0), \dots, \check{\mathbf{x}}(N-1), \check{\mathbf{x}}(N), \dots, \check{\mathbf{x}}(2N-1), \dots, \check{\mathbf{x}}((n-1)N), \dots, \check{\mathbf{x}}(nN-1)$ are sampled continuously and are correlated with each other. Thus, $\mathbf{x}^r(\omega)$ and $\mathbf{x}^s(\omega)$, $r \neq s$, $1 \leq r, s \leq n$, can be correlated, in contrast to the restart and record (RR) setting, where the $\mathbf{x}^r(\omega)$ and $\mathbf{x}^s(\omega)$, $r \neq s$ are i.i.d. For any given $\omega \in \Omega$, let $\mathbf{x}(\omega) := [[\mathbf{x}^1(\omega)]^T, [\mathbf{x}^2(\omega)]^T, \dots, [\mathbf{x}^n(\omega)]^T]^T \in \mathcal{C}^{pn \times 1}$ be the vectorized form of $\{\mathbf{x}^r(\omega)\}_{r=1}^n$ and let $\mathcal{C}(\omega) := \mathbb{E}\{\mathbf{x}(\omega)\mathbf{x}^*(\omega)\}$ be the covariance matrix of $\mathbf{x}(\omega)$. Under the RR setting, $\mathcal{C}(\omega) \in \mathcal{C}^{pn \times nn}$ will be a block-diagonal matrix (of block size $p \times p$), whereas in the continuous sampling, the non-block-diagonal entries of $\mathcal{C}(\omega)$ can be non-zero. However, the vector $\mathbf{x}(\omega)$, with correlated entries, can be written as a linear transformation of i.i.d. vector $\mathbf{w} \in \mathcal{C}^{pn \times 1}$ with unit variance, i.e. $\mathbf{x}(\omega) = \mathcal{C}^{1/2}(\omega)\mathbf{w}$, where $\mathcal{C}^{1/2}$ is the square-root of $\mathcal{C}$. As shown in [112], when $\{\check{\mathbf{e}}(k)\}_{k=1}^n$ in the linear time-invariant model (1) are Gaussian, $\mathbf{x}^r(\omega)$ and thus $\mathbf{x}(\omega)$ are Gaussian distributed. In this case, a candidate is $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{pn})$. It can be verified that $\mathbb{E}\{\mathbf{x}(\omega)\mathbf{x}^*(\omega)\} = \mathcal{C}^{1/2}(\omega)\mathbb{E}\{\mathbf{w}\mathbf{w}^*\}\mathcal{C}^{1/2}(\omega) = \mathcal{C}(\omega)$. Notice that the covariance matrix $\mathcal{C}(\omega)$ is a block matrix, defined as

$$\mathcal{C} = \begin{bmatrix} \mathcal{C}^{11} & \mathcal{C}^{12} & \dots & \mathcal{C}^{1n} \\ \mathcal{C}^{21} & \mathcal{C}^{22} & \dots & \mathcal{C}^{2n} \\ \vdots & & & \\ \mathcal{C}^{n1} & \mathcal{C}^{n2} & \dots & \mathcal{C}^{nn} \end{bmatrix}, \text{ where } \mathcal{C}^{rs}(\omega) \in \mathcal{C}^{p \times p}, 1 \leq r, s \leq n,$$

where the entries of $\mathcal{C}^{rs}(\omega)$ is given by $\mathbb{E}\{\mathbf{x}^r(\omega)[\mathbf{x}^s(\omega)]^*\}$. Recall that

$$\mathbf{x}^r(\omega) = \frac{1}{\sqrt{N}} \sum_{\ell=0}^{N-1} \check{\mathbf{x}}((r-1)N + \ell) e^{-i\omega\ell}.$$

Let $\mathbf{I}_r = [\mathbf{0} | \dots | \mathbf{0} | \mathbf{I}_{p \times p} | \dots | \mathbf{0}] \in \mathbb{R}^{p \times np}$ be such that r^{th} block is identity matrix.

Then $\mathbf{x}^r(\omega) = \mathbf{I}_r \mathbf{x}(\omega)$. The estimated PSDM is then given by

$$\hat{\Phi}_{\mathbf{x}}(\omega) = \frac{1}{n} \sum_{r=1}^n \mathbf{x}^r(\omega) [\mathbf{x}^r(\omega)]^* = \frac{1}{n} \sum_{r=1}^n \mathbf{I}_r \mathbf{x}(\omega) \mathbf{x}^*(\omega) \mathbf{I}_r^*.$$

Substituting $\mathbf{x}(\omega) = \mathcal{C}^{1/2}(\omega) \mathbf{w}$, and letting $\mathbf{B}(\omega) := (\mathcal{C}^{1/2})^*(\omega) \sum_{r=1}^n \mathbf{I}_r^* u u^* \mathbf{I}_r \mathcal{C}^{1/2}(\omega)$

$$\begin{aligned} \mathbb{E} [\exp(tu^* \mathbf{Q} u)] &= \mathbb{E} \left[\exp \left(\frac{t}{n} \sum_{r=1}^n u^* \mathbf{x}^r(\omega) [\mathbf{x}^r(\omega)]^* u - u^* \tilde{\Phi}_x(\omega) u \right) \right] \\ &= \mathbb{E} \left[\exp \left(\frac{t}{n} \sum_{r=1}^n [\mathbf{x}^*(\omega) \mathbf{I}_r^* u u^* \mathbf{I}_r \mathbf{x}(\omega) - \mathbb{E} \{ \mathbf{x}^*(\omega) \mathbf{I}_r^* u u^* \mathbf{I}_r \mathbf{x}(\omega) \}] \right) \right] \\ &= \mathbb{E} \left[\exp \left(\frac{t}{n} \sum_{r=1}^n \left[\mathbf{w}^* (\mathcal{C}^{1/2})^*(\omega) \mathbf{I}_r^* u u^* \mathbf{I}_r \mathcal{C}^{1/2}(\omega) \mathbf{w} \right. \right. \right. \\ &\quad \left. \left. - \mathbb{E} \left\{ \mathbf{w}^* (\mathcal{C}^{1/2})^*(\omega) \mathbf{I}_r^* u u^* \mathbf{I}_r \mathcal{C}^{1/2}(\omega) \mathbf{w} \right\} \right] \right) \right] \\ &= \mathbb{E} \left[\exp \left(\frac{t}{n} \left[\mathbf{w}^* (\mathcal{C}^{1/2})^*(\omega) \sum_{r=1}^n (\mathbf{I}_r^* u u^* \mathbf{I}_r) \mathcal{C}^{1/2}(\omega) \mathbf{w} \right. \right. \right. \\ &\quad \left. \left. - \mathbb{E} \left\{ \mathbf{w}^* (\mathcal{C}^{1/2})^*(\omega) \sum_{r=1}^n (\mathbf{I}_r^* u u^* \mathbf{I}_r) \mathcal{C}^{1/2}(\omega) \mathbf{w} \right\} \right] \right) \right] \\ &= \mathbb{E} \left[\exp \left(\frac{t}{n} [\mathbf{w}^* \mathbf{B}(\omega) \mathbf{w} - \mathbb{E} \{ \mathbf{w}^* \mathbf{B}(\omega) \mathbf{w} \}] \right) \right] \end{aligned}$$

Notice that $\mathbf{I}_r^* u = [0, \dots, 0, u^T, \dots, 0]^T$ is a column vector,

$$\mathbf{I}_r^* u u^* \mathbf{I}_r = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ & \dots & & \dots & \mathbf{0} \\ \vdots & \dots & \underbrace{u u^*}_{(r,r)^{th} \text{ block}} & \dots & \mathbf{0} \\ & & & \ddots & \mathbf{0} \\ \vdots & \vdots & \dots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \end{bmatrix} \text{ and } \sum_{r=1}^n \mathbf{I}_r^* u u^* \mathbf{I}_r = \begin{bmatrix} u u^* & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & u u^* & \dots & \mathbf{0} \\ \vdots & \ddots & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & u u^* \end{bmatrix},$$

i.e., $rank(\mathbf{I}_r^* u u^* \mathbf{I}_r) = 1$ and $rank(\mathbf{B}(\omega)) \leq n$. Let $\mathbf{B}(\omega) = \mathbf{U}(\omega) \Lambda(\omega) \mathbf{U}^*(\omega)$ be the eigen value decomposition of $\mathbf{B}(\omega)$, where $\Lambda = diag(\lambda_1, \dots, \lambda_n)$. Consequently, omitting ω

from the notations,

$$\begin{aligned}
\mathbb{E} [\exp (tu^* \mathbf{Q} u)] &= \mathbb{E} \left[\exp \left(\frac{t}{n} [\mathbf{w}^* \mathbf{B} \mathbf{w} - \mathbb{E} \{ \mathbf{w}^* \mathbf{B} \mathbf{w} \}] \right) \right] \\
&= \mathbb{E} \left[\exp \left(\frac{t}{n} [\mathbf{w}^* \mathbf{U} \mathbf{\Lambda} \mathbf{U}^* \mathbf{w} - \mathbb{E} \{ \mathbf{w}^* \mathbf{U} \mathbf{\Lambda} \mathbf{U}^* \mathbf{w} \}] \right) \right] \\
&\stackrel{(a)}{=} \mathbb{E} \left[\exp \left(\frac{t}{n} [\mathbf{w}^* \mathbf{\Lambda} \mathbf{w} - \mathbb{E} \{ \mathbf{w}^* \mathbf{\Lambda} \mathbf{w} \}] \right) \right] \\
&= \mathbb{E} \left[\exp \left(\frac{t}{n} \sum_{i=1}^n \lambda_i [w_i^2 - \mathbb{E} \{ w_i^2 \}] \right) \right] \\
&= \prod_{i=1}^n \mathbb{E} \left[\exp \left(\frac{t \lambda_i}{n} [w_i^2 - \mathbb{E} \{ w_i^2 \}] \right) \right],
\end{aligned}$$

where (a) follows because $\mathbf{w}$ is invariant under unitary transformations [127]. Let $\varepsilon \in \{+1, -1\}$ be a uniform random variable independent of $\mathbf{w}$. Similar to (C.3), we can now apply the Rademacher random variable trick.

$$\begin{aligned}
\mathbb{E}_{w_i} [\exp (\lambda [w_i^2 - \mathbb{E} \{ w_i^2 \}])] &\leq \mathbb{E}_{w_i, \varepsilon} [\exp (2 \lambda \varepsilon w_i^2)] \\
&= \sum_{k=0}^{\infty} \frac{(2 \lambda)^{2k}}{(2k)!} \mathbb{E} [w_i^{4k}] \\
&\leq \sum_{k=0}^{\infty} \frac{(2 \lambda)^{2k}}{2k!} \frac{(4k)!}{(2k)! 2^{2k}} \\
&\leq \sum_{k=0}^{\infty} (2 \lambda)^{2k} = \frac{1}{1 - 4 \lambda^2} \leq \exp(8 \lambda^2),
\end{aligned}$$

for every $|\lambda| < 3/8$. Thus, (with the substitution $\lambda = t \lambda_i / n$ and the upperbound

$$\lambda_i \leq \|\mathbf{B}\|).$$

$$\begin{aligned} \mathbb{E} [\exp (tu^* \mathbf{Q} u)] &= \prod_{i=1}^n \mathbb{E} \left[\exp \left(\frac{t\lambda_i}{n} [w_i^2 - \mathbb{E} \{w_i^2\}] \right) \right] \\ &\leq \prod_{i=1}^n \exp \left(\frac{8t^2\lambda_i^2}{n^2} \right) \\ &= \exp \left(\frac{8t^2}{n^2} \sum_{i=1}^n \lambda_i^2 \right) \\ &\leq \exp \left(\frac{8t^2}{n} \|\mathcal{C}\|^2 \right), \quad \forall |t| \leq \frac{3n}{8\|\mathcal{C}\|}, \end{aligned}$$

where we have used $\|B\| \leq \|\mathcal{C}\|$ in the final equality. Now, combining this with the δ -cover argument,

$$\begin{aligned} \mathbb{E} \left[e^{t\|Q\|} \right] &\leq \sum_{j=1}^m \mathbb{E} \left[e^{2t(v^j)^* Q v^j} \right] + \mathbb{E} \left[e^{-2t(v^j)^* Q v^j} \right] \\ &\leq 2m \exp \left(\frac{32\|\mathcal{C}\|^2 t^2}{n} \right) \\ &\leq \exp \left(\frac{32\|\mathcal{C}\|^2 t^2}{n} + 6p \right), \quad \forall |t| \leq \frac{3n}{16\|\mathcal{C}\|}. \end{aligned}$$

Finally, applying Chernoff bound, $\mathcal{P}(\|\mathbf{Q}\| \geq t) \leq \exp \left(-\frac{t^2 n}{128\|\mathcal{C}\|^2} + 6p \right).$

Tight upper bound for $\|\mathcal{C}\|$

An explicit expression for $\mathcal{C}$ is given as follows:

$$\begin{aligned}
\mathcal{C}^{rs}(\omega) &:= \mathbb{E} \{ \mathbf{x}^r(\omega) [\mathbf{x}^s(\omega)]^* \} = \frac{1}{N} \sum_{\ell=0}^{N-1} \sum_{k=0}^{N-1} \mathbb{E} \{ \check{x}((r-1)N + \ell) [\check{x}((s-1)N + k)]^T \} e^{-i\omega(\ell-k)} \\
&= \frac{1}{N} \sum_{\ell=0}^{N-1} \sum_{k=0}^{N-1} R_{\check{\mathbf{x}}}((r-s)N + \ell - k) e^{-i\omega(\ell-k)} \\
&= \frac{1}{N} \sum_{\tau=-N+1}^{N-1} (N - |\tau|) R_{\check{\mathbf{x}}}((r-s)N + \tau) e^{-i\omega\tau} \\
&= \sum_{\tau=-N+1}^{N-1} \left(1 - \frac{|\tau|}{N} \right) R_{\check{\mathbf{x}}}((r-s)N + \tau) e^{-i\omega\tau} \\
&= \sum_{\tau=-N+1}^{N-1} \left(1 - \frac{|\tau|}{N} \right) e^{-i\omega\tau} R_{\check{\mathbf{x}}}((r-s)N + \tau).
\end{aligned}$$

Let $\alpha_\tau = e^{-i\omega\tau} \left(1 - \frac{|\tau|}{N} \right)$. Then

$$\mathcal{C} = \sum_{\tau=-N+1}^{N-1} \alpha_\tau \begin{bmatrix} R_{\check{\mathbf{x}}}(\tau) & R_{\check{\mathbf{x}}}(-N + \tau) & \dots & R_{\check{\mathbf{x}}}((1-n)N + \tau) \\ R_{\check{\mathbf{x}}}(N + \tau) & R_{\check{\mathbf{x}}}(\tau) & \dots & R_{\check{\mathbf{x}}}((2-n)N + \tau) \\ \vdots & \vdots & \ddots & \\ R_{\check{\mathbf{x}}}((n-1)N + \tau) & R_{\check{\mathbf{x}}}((n-2)N + \tau) & \dots & R_{\check{\mathbf{x}}}(\tau) \end{bmatrix}.$$

Notice that $\check{g}(\tau) := 1 - |\tau|/N$ is a triangle function, and the Fourier transform of $\check{g}(\tau)$, $g(\omega)$ has the property that $|g(\omega)| \leq 1, \forall \omega \in \Omega$. Let $\mathbf{u} \in \mathcal{C}^{np}$ with $\|\mathbf{u}\|_2 \leq 1$ and $u(i) \in \mathcal{C}^p$ be the i th p -dimensional subvector in u . Further, $\|u(\omega)\|_2^2 = \left\| \sum_{k=1}^n e^{-i\omega k} u(k) \right\|_2^2 \leq$

$\sum_{k=1}^n |e^{-i\omega k}| \|u(k)\|_2^2 = \|\mathbf{u}\|_2^2 \leq 1$. Then

$$\begin{aligned}
|\mathbf{u}^* \mathcal{C} \mathbf{u}| &= \left| \sum_{i,j=1}^n u^*(i) \sum_{\tau=-N+1}^{N-1} \alpha_\tau R_{\check{\mathbf{x}}}((i-j)N + \tau) u(j) \right| \\
&= \left| \sum_{\tau=-N+1}^{N-1} e^{-i\omega\tau} \check{g}(\tau) \sum_{i,j=-n}^n u^*(i) R_{\check{\mathbf{x}}} \left(\left((i-j) + \frac{\tau}{N} \right) N \right) u(j) \right|, \because u(i) := 0, \forall i \leq 0 \\
&= \left| \sum_{\tau=-N+1}^{N-1} e^{-i\omega\tau} \check{g}(\tau) (v^* \star \mathbf{y} \star u) \left(\frac{\tau}{N} \right) \right|, \because v(i) = u(-i), \forall i, \text{ and } \mathbf{y}(\tau) = R_{\check{\mathbf{x}}}(N\tau) \\
&\stackrel{(a)}{=} \left| \frac{1}{2\pi} \int_{\Omega} g(\nu - \omega) v^*(N\omega) \mathbf{y}(N\omega) u(N\omega) d\omega \right| \\
&\leq \frac{1}{2\pi} \int_{\Omega} |v^*(N\omega) \mathbf{y}(N\omega) u(N\omega)| d\omega \quad (\because |g(\nu - \omega)| \leq 1) \\
&\leq \frac{M}{2\pi} \int_{\Omega} |v^*(N\omega) u(N\omega)| d\omega \quad (\text{by spectral norm definition and } \|\Phi_{\mathbf{x}}\| \leq M) \\
&\leq \frac{M}{2\pi} \int_{\Omega} \|v^*(N\omega)\|_2 \|u(N\omega)\|_2 d\omega \\
&\leq M \quad (\because \|u(\omega)\|_2 \leq 1)
\end{aligned}$$

Thus, $\mathcal{P}(\|\mathbf{Q}\| \geq t) \leq \exp\left(-\frac{t^2 n}{128M^2} + 6p\right)$, similar to the restart and record case.

C.4 Proof of Lemma 4.4.2

Notice that $\|A\mathbf{x}\|_2 \leq \|A\| \|\mathbf{x}\|_2$ for every matrix A and vector $\mathbf{x}$. Applying this identity with $\mathbf{x} = [1, 0, \dots, 0]$, $\|\Phi_{Ci}\|_2 \leq \|\Phi_{AA}\|$, where $A = [k, C]$. Then, applying CBS inequality for complex vectors, $|x^* Ay| \leq \|x\|_2 \|Ay\|_2 \leq \|x\|_2 \|A\| \|y\|_2$, the error can be

upper bounded as

$$\begin{aligned}
|f(i, C, \omega) - \hat{f}(i, C, \omega)| &= \\
&|(\Phi_{ii} - \Phi_{iC}\Phi_{CC}^{-1}\Phi_{Ci}) - (\hat{\Phi}_{ii} - \hat{\Phi}_{iC}\hat{\Phi}_{CC}^{-1}\hat{\Phi}_{Ci})| \\
&= |(\Phi_{ii} - \hat{\Phi}_{ii}) + (\hat{\Phi}_{iC}\hat{\Phi}_{CC}^{-1}\hat{\Phi}_{Ci} - \Phi_{iC}\Phi_{CC}^{-1}\Phi_{Ci})| \\
&\leq |\Phi_{ii} - \hat{\Phi}_{ii}| + |\hat{\Phi}_{iC}(\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1})\hat{\Phi}_{Ci}| \\
&\quad + |(\hat{\Phi}_{iC} - \Phi_{iC})\Phi_{CC}^{-1}\hat{\Phi}_{Ci}| + |\Phi_{iC}\Phi_{CC}^{-1}(\hat{\Phi}_{Ci} - \Phi_{Ci})| \\
&\leq |\Phi_{ii} - \hat{\Phi}_{ii}| + \|\hat{\Phi}_{iC}\|_2 \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \|\hat{\Phi}_{Ci}\|_2 \\
&\quad + \|\hat{\Phi}_{iC} - \Phi_{iC}\|_2 \|\Phi_{CC}^{-1}\| \|\hat{\Phi}_{Ci}\|_2 + \|\Phi_{iC}\|_2 \|\Phi_{CC}^{-1}\| \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2 \\
&\leq |\Phi_{ii} - \hat{\Phi}_{ii}| + \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \|\hat{\Phi}_{Ci}\|_2^2 \\
&\quad + M \|\hat{\Phi}_{iC} - \Phi_{iC}\|_2 \|\hat{\Phi}_{Ci}\|_2 + M^2 \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2 \\
&\leq |\Phi_{ii} - \hat{\Phi}_{ii}| + \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2^2 + \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \|\Phi_{Ci}\|_2^2 \\
&\quad + M \|\hat{\Phi}_{iC} - \Phi_{iC}\|_2 \left(\|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2 + \|\Phi_{Ci}\|_2 \right) + M^2 \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2 \\
&\leq |\Phi_{ii} - \hat{\Phi}_{ii}| + \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2^2 + \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| M^2 \\
&\quad + M \|\hat{\Phi}_{iC} - \Phi_{iC}\|_2 \left(\|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2 + M \right) + M^2 \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2 \\
&= |\Phi_{ii} - \hat{\Phi}_{ii}| + \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2^2 + \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| M^2 \\
&\quad + M \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2^2 + M^2 \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2 + M^2 \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2 \\
&\leq |\Phi_{ii} - \hat{\Phi}_{ii}| + \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2^2 + M^2 \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \\
&\quad + M \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2^2 + 2M^2 \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2.
\end{aligned}$$

The above expression can be bounded above if we can bound the three errors,

$\|\hat{\Phi}_{ii} - \Phi_{ii}\| = \epsilon_i$, $\|\hat{\Phi}_{AA} - \Phi_{AA}\| = \epsilon_A$, and $\|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| = \epsilon_{Cinv}$. Simplifying the above expression,

$$\begin{aligned}
|f(i, C, \omega) - \hat{f}(i, C, \omega)| &\leq |\Phi_{ii} - \hat{\Phi}_{ii}| + \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2^2 + M^2 \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \\
&\quad + M \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2^2 + 2M^2 \|\hat{\Phi}_{Ci} - \Phi_{Ci}\|_2 \\
&\leq \epsilon_i + \epsilon_{Cinv} \epsilon_A^2 + 2M^2 \epsilon_{Cinv} + M \epsilon_A^2 + 2M^2 \epsilon_A \\
&\leq \epsilon_i + \epsilon_{Cinv} (\epsilon_A^2 + 2M^2) + 3M^2 \epsilon_A \\
&\leq \epsilon_i + 3M^2 \epsilon_{Cinv} + 3M^2 \epsilon_A.
\end{aligned}$$

Pick $\epsilon_i = 3M^2\epsilon_{Cinv} = 3M^2\epsilon_A = \epsilon/3$. Then $|f(i, C, \omega) - \hat{f}(i, C, \omega)| < \epsilon$. From Section 5.8 in [65],

$$\begin{aligned} \|\Phi_{CC} - \hat{\Phi}_{CC}\| &\leq \|\Phi_{CC}\| \|\Phi_{CC}^{-1}\|^{-1} \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \frac{M^2}{1 - M^2 \frac{\|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\|}{\|\Phi_{CC}^{-1}\|}}, \\ &\leq \frac{M^4 \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\|}{1 - M \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\|} \leq \epsilon \implies \|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \leq \frac{\epsilon}{M^4 + M\epsilon} \leq \frac{\epsilon}{M^4}. \end{aligned}$$

Therefore, to guarantee that $\|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| < \epsilon$, it is sufficient to guarantee that $\|\hat{\Phi}_{CC} - \Phi_{CC}\| < \epsilon$ since $M \geq 1$. Rewriting Corollary 4.3,

$$\mathcal{P}\left(|\Phi_{ii} - \hat{\Phi}_{ii}| \geq \epsilon\right) \leq e^{-\frac{\epsilon^2 n}{128M^2} + 6}, \quad (\text{C.5})$$

$$\mathcal{P}\left(\|\Phi_{AA} - \hat{\Phi}_{AA}\| > \epsilon\right) \leq e^{-\frac{\epsilon^2 n}{128M^2} + 6(q+1)}, \text{ and} \quad (\text{C.6})$$

$$\mathcal{P}\left(\|\Phi_{CC} - \hat{\Phi}_{CC}\| > \epsilon\right) \leq e^{-\frac{\epsilon^2 n}{128M^2} + 6q}, \quad \forall \epsilon \geq 0. \quad (\text{C.7})$$

Plugging these bounds in the above expressions gives the concentration upper bound

$$\begin{aligned} \mathcal{P}\left(|f(i, C, \omega) - \hat{f}(i, C, \omega)| \geq \epsilon\right) &\leq \mathcal{P}\left(|\Phi_{ii} - \hat{\Phi}_{ii}| \geq \epsilon/3\right) + \mathcal{P}\left(\|\hat{\Phi}_{CC}^{-1} - \Phi_{CC}^{-1}\| \geq \epsilon/(9M^2)\right) \\ &\quad + \mathcal{P}\left(\|\hat{\Phi}_{AA} - \Phi_{AA}\| \geq \epsilon/(9M^2)\right) \\ &\leq \mathcal{P}\left(|\Phi_{ii} - \hat{\Phi}_{ii}| \geq \epsilon/3\right) + \mathcal{P}\left(\|\hat{\Phi}_{CC} - \Phi_{CC}\| \geq \epsilon M^2/(9)\right) \\ &\quad + \mathcal{P}\left(\|\hat{\Phi}_{AA} - \Phi_{AA}\| \geq \epsilon/(9M^2)\right) \\ &\leq e^{-\left(-\frac{\epsilon^2 n}{1152M^2} + 6\right)} + e^{-\left(-\frac{\epsilon^2 M^2 n}{10368} + 6q\right)} + e^{-\left(-\frac{\epsilon^2 n}{10368M^6} + 6(q+1)\right)} \\ &\leq c_0 e^{-\left(-\frac{\epsilon^2 n}{10368M^6} + 6(q+1)\right)}. \end{aligned}$$

□

C.5 Lower bound: Proof of Theorem 4.5.1

The proof is based on Generalized Fano's inequality.

Lemma C.5.1 (Generalized Fano's method). [92] *Consider a class of observational distribution $\mathcal{F}$ and a subclass $\mathcal{F}' = \{F_1, \dots, F_r\} \subseteq \mathcal{F}$ with r distributions and the*

estimators $\hat{\theta}$. Then

$$\inf_{\hat{\theta}} \max_{F \in \mathcal{F}} \mathbb{E}\{\mathbb{1}(\theta(F) \neq \hat{\theta})\} \geq \frac{\alpha_r}{2} \left(1 - \frac{n\beta_r + \log 2}{\log r}\right),$$

where n is the number of samples,

$$\begin{aligned} \alpha_r &:= \max_{k \neq j} \mathbb{1}(\theta(F_k) \neq \theta(F_j)), \\ \beta_r &:= \max_{k \neq j} KL(F_k || F_j), \end{aligned}$$

with $KL(P||Q) := \mathbb{E}_P \left[\log \frac{P}{Q} \right] = \mathbb{E}_P [\log P] - \mathbb{E}_P [\log Q]$ being the KL divergence.

Corollary C.5.1. Consider subclass of graphs $\mathcal{G}' = \{G_1, \dots, G_r\} \subseteq \mathcal{G}_{p,q}$, and let $\mathbf{H}^i$ be the distribution corresponding to a distinct $G_i \in \mathcal{G}'$. Then, any estimator $\hat{G} := \bigcup_{\omega \in \Omega} \hat{G}(\omega)$ of G_i is δ unreliable,

$$\inf_{\hat{G}} \sup_{G_i \in \mathcal{G}'} \mathcal{P}\{(G(\mathbf{H}^i) \neq \hat{G})\} \geq \delta,$$

if

$$n \leq \frac{(1 - 2\delta) \log r - \log 2}{\beta_r}$$

Therefore, building a lower bound involves finding a subclass that has 1) small β_r and 2) large r . First, we can find an upper bound of r by upper bounding the number of directed graphs possible with at most q parents. Overall, p^2 number of possible positions are there and at most pq many edges. The number of possible ways to choose k edges is $\binom{p^2}{k}$. Thus $r = \sum_{k=1}^{pq} \binom{p^2}{k} \leq pq \binom{p^2}{pq} \lesssim pq(p/q)^{pq}$. Therefore, $\log r \lesssim \log(pq) + pq \log(p/q) \lesssim pq \log(p/q)$. Similarly, it can be shown that $\log r \gtrsim pq \log(p/q)$ [92].

Consider the LDS (3), with $\check{e}(k) \sim N(0, \sigma_k \mathbf{I})$, i.i.d. across time. Then $\mathbf{e}(\omega) \sim N(0, \Phi_{\mathbf{e}}(\omega))$ and $\mathbf{x}(\omega) \sim N(0, \Phi_{\mathbf{x}})$, where $\Phi_{\mathbf{e}}(\omega) = \sum_{k \in \mathbb{Z}} \sigma_k \mathbf{I}$ and $\Phi_{\mathbf{x}}(\omega) = (I - \mathbf{H}(\omega))^{-1} \Phi_{\mathbf{e}}(\omega) ((I - \mathbf{H}(\omega))^{-1})^*$.

Ensemble A: Consider all possible DAGs in $\mathcal{G}'$ with i.i.d. Gaussian exogenous distribution such that $\Phi_{\mathbf{e}}(\omega)$ exists. For the two distributions, F_k and F_j such that for any

$\omega \in \Omega$, $F_k(\omega) \sim \mathcal{N}(\mathbf{0}, \Phi^{(k)}(\omega))$ and $F_j(\omega) \sim \mathcal{N}(\mathbf{0}, \Phi^{(j)}(\omega))$,

$$\begin{aligned}
KL(F_k(\omega) || F_j(\omega)) &= \frac{1}{2} \left(\mathbb{E}_{F_j(\omega)}[\mathbf{x}^*(\omega)[\Phi^{(k)}(\omega)]^{-1}\mathbf{x}(\omega)] - \mathbb{E}_{F_j}[\mathbf{x}^*(\omega)[\Phi^{(j)}(\omega)]^{-1}\mathbf{x}(\omega)] \right) \\
&= \frac{1}{2} \left(\mathbb{E}_{F_j(\omega)}[tr(\mathbf{x}(\omega)[\Phi^{(k)}(\omega)]^{-1}\mathbf{x}^*(\omega))] - p \right) \\
&= \frac{1}{2} \left(tr([\Phi^{(k)}(\omega)]^{-1}\Phi^{(j)}(\omega)) - p \right) \\
&\leq \frac{1}{2} \left(\sqrt{\|[\Phi^{(k)}(\omega)]^{-1}\|_F^2 \|\Phi^{(j)}(\omega)\|_F^2} - p \right) \\
&\leq \frac{1}{2} (pM^2 - p) \leq (M^2 - 1)p
\end{aligned}$$

Therefore one of the lower bounds is

$$\inf_{\hat{G}} \sup_{G \in \mathcal{G}'} \mathcal{P}\{(G \neq \hat{G})\} \geq \delta,$$

if

$$\begin{aligned}
n &\leq \frac{(1 - 2\delta)pq \log(p/q) - \log 2}{(M^2 - 1)p} \\
&\lesssim \frac{q \log(p/q)}{M^2 - 1}.
\end{aligned}$$

Ensemble B: Here, we consider graphs in $\mathcal{H}_{p,q}(\beta, \sigma, M)$ (recall Definition 2.4) with a single edge $u \rightarrow v$ with $H_{vu}(\omega) = \beta$ for every $\omega \in \Omega$, i.e. constant matrix. For LDS with i.i.d. Gaussian noise with PSD matrix $\Phi_{\mathbf{x}}$ that satisfies this condition, $\mathbf{H}$ is such that $H_{vu} \neq 0$ and $H_{ij} = 0$ otherwise. Here, the total number of graphs, $r = 2\binom{p}{2} = (p^2 - p) \approx p^2$

Notice that $[\Phi_{\mathbf{x}}^{-1}]_{ij} = (\mathbf{I}_{ij} - H_{ij} - H_{ji}^* + \sum_{k=1}^p H_{ki}^* H_{kj})/\sigma$. Thus (ignoring ω),

$[\Phi_{\mathbf{x}}^{-1}]_{uv} = \frac{-H_{vu}^* + \sum_{k=1}^p H_{ku}^* H_{kv}}{\sigma} = \frac{-H_{vu}^*}{\sigma} = -\beta/\sigma$ and $[\Phi_{\mathbf{x}}^{-1}]_{ij} = 0$ if $i, j \neq u, v$. Then,

$$\begin{aligned}
\mathbf{x}^* \Phi_{\mathbf{x}}^{-1} \mathbf{x} &= \sum_{ij} x_i^* [\Phi_{\mathbf{x}}^{-1}]_{ij} x_j \\
&= \frac{1}{\sigma} \left[\sum_{i=1}^p |x_i|^2 (1 + \sum_k |H_{ki}|^2) + x_u^* [\Phi_{\mathbf{x}}^{-1}]_{uv} x_v + x_v^* [\Phi_{\mathbf{x}}^{-1}]_{vu} x_u \right] \\
&= \frac{1}{\sigma} \left[\sum_{i \neq u} |x_i|^2 + (1 + |H_{vu}|^2) |x_u|^2 - 2\beta \Re\{x_u^* x_v\} \right] \\
&= \frac{1}{\sigma} \left[\sum_i |x_i|^2 + \beta^2 |x_u|^2 - 2\beta \Re\{x_u^* x_v\} \right] \\
&= \frac{1}{\sigma} \left[\sum_{i \neq v} |x_i|^2 + |x_v - \beta x_u|^2 \right]
\end{aligned}$$

Therefore,

$$\begin{aligned}
KL(F^{uv} || F^{jk}) &= \mathbb{E}_{F^{uv}} [\log F^{uv} - \log F^{jk}] \\
&= \frac{1}{2\sigma} \mathbb{E}_{F^{uv}} [|x_v|^2 + |x_v - \beta x_u|^2 - |x_k|^2 - |x_k - \beta x_j|^2] \\
&= \frac{1}{2\sigma} [\beta^2 \sigma + \mathbb{E}_{F^{uv}} (\beta^2 |x_j|^2 - 2\beta \Re\{x_j^* x_k\})]
\end{aligned}$$

Considering all the cases of (u, v) vs (j, k) it can be shown that $KL(F^{uv} || F^{jk}) \leq \beta^2 + \beta^4/2$ [92]. Thus, $n \gtrsim \frac{\log p}{\beta^2 + \beta^4/2}$ gives the second lower bound. The lower bound follows by combining ensembles A and B . $\square$

Appendix D

Appendix of Chapter 5

Table of Contents

- Wiener-Phase algorithm – Appendix D.1
- Simulation – Appendix D.2
- Theorem 5.2.1 (convergence of FFT) – Appendix D.3
- Proof of Lemma 5.3.2 (d-separation in G implies d-separation in $G^{(\omega)}$) – Appendix D.4
- Proof of Lemma 5.3.3 – Appendix D.4.3
- Proof of Lemma 5.3.4 – Appendix D.4.4
- Theorems on Single-door, back-door and front-door – Appendix D.6
- Proof of single-door – Appendix D.7
- Proof of Theorem D.8 (Back-door) – Appendix D.8
- Proof of Theorem 5.4.1 – Appendix D.9
- Primer on Wiener filter – Appendix D.10
- Primer on stochastic processes and FDD based conditional independence – Appendix D.11

D.1 Wiener-Phase Algorithm

In here, we provide the Wiener-Phase which is described in Algorithm 6. As shown in Lemma 5.2.2, in a large class of applications [1], support of the imaginary part of the frequency dependent Wiener filter retrieves $top(G)$ exactly. Further, as shown in Lemma 5.2.1, support of W_i retrieves the Markov blanket structure in G . Combining both, we can obtain the Wiener-phase algorithm below to identify the colliders and thus the equivalence class of the graphs efficiently. $\mathcal{K}$ is the set of kin edges obtained using Lemma 5.2.1 and $\mathcal{S}$ is the skeleton obtained from Lemma 5.2.2. Consider any edge (i, j) that belongs to $\mathcal{K}$ but does not belong to $\mathcal{S}$. Then i and j share a common child without a direct link between them (strict spouses) in G . Following this procedure, we can construct all the strict spouses. Let this set be $\mathcal{SP}$. The complexity of computing $\mathcal{K}$ and $\mathcal{S}$ is $O(n^3 RN \log N)$

Now consider the skeleton $\mathcal{S}$. In Step 4, we identify the common child between the strict spouses i, j as follows. For any $i, j \in \mathcal{SP}$, let $\mathcal{C}_{ij} := \{k : \{i, k\} \in \mathcal{S}, \{j, k\} \in \mathcal{S}\}$, which is a set that contains all nodes c which has a link to both i and j in the skeleton. Note that the set $\mathcal{C}_{ij}$ can contain non-colliders, which will be eliminated using the following steps. For any $\{i, j\} \in \mathcal{SP}$, if $|\mathcal{C}_{ij}| = 1$, then $c \in \mathcal{C}_{ij}$ is a collider (because the link between i and j in $\mathcal{K}$ is formed by at least one collider). If $|\mathcal{C}_{ij}| > 1$, then for every $c \in \mathcal{C}_{ij}$, we can compute $W_{i[j, c]}$. If $W_{i[j, c]}[j] \neq 0$, then c is a collider since $\{i, j\} \notin \mathcal{S}$. The complexity of computing $\mathcal{K}$ and $\mathcal{S}$ is $O(n^3 RN \log N)$. Step 4 is repeated $O(nq)$ times and Step 4(a) which checks for potential colliders among the common neighbors of i and j takes $O(n^2 q)$ operations. Steps 4(b) and 4(c) take $O(n)$. Thus the complexity in computing Step 4 is $O(n^3 q^2)$, giving the overall complexity of $O(n^3(RN \log N + q^2))$. Notice that it is sufficient to compute Algorithm 6 for $O(1)$ number of ω . Thus, the final complexity remains the same.

Algorithm 6 Wiener-Phase algorithm

Input: Data $\mathcal{X}(\omega)$, $\omega \in [0, 2\pi]$ **Output:** $\hat{G}$

1. Initialize the ordering, $\mathcal{S} \leftarrow ()$
 2. For $i = 1, \dots, n$
 - (a) Compute $W_i(\omega)$ using (5.5)
 - (b) for $j = 1, \dots, n$
 - i. If $|W_i[j]| > \tau$, then $\mathcal{K} \leftarrow \mathcal{K} \cup \{i, j\}$
 - ii. If $|\Im\{W_i[j]\}| > \tau$, then $\mathcal{S} \leftarrow \mathcal{S} \cup \{i, j\}$
 3. compute $\mathcal{SP} := \mathcal{K} \setminus \mathcal{S}$
 4. for $\{i, j\} \in \mathcal{SP}$
 - (a) for $k = 1, \dots, n$
 - i. if $\{i, k\} \in \mathcal{S}$ and $\{j, k\} \in \mathcal{S}$
 - $C_{ij} \leftarrow C_{ij} \cup k$
 - (b) If $|C_{ij}| = 1$: $Col \leftarrow Col \cup k$
 - (c) Else: for $c \in C_{ij}$ compute $W_{i \cdot [j, c]}$
 - i. If $|W_{i \cdot [j, c]}[j]| > \tau$: $Col \leftarrow Col \cup c$
 5. Return $\hat{G}$
-

D.2 Simulation

The synthetic data is generated using AR model, with the non-zero entries that respect the graph structure shown in Figure 5.1(a),

$$\mathbf{X}_i(t) + a_i(1)\mathbf{X}_i(t-1) + a_i(2)\mathbf{X}_i(t-2) + a_i(3)\mathbf{X}_i(t-3) = \sum_{j \neq i} b_{ij}\mathbf{X}_j(t-1) + \mathbf{E}_i(t), \quad (\text{D.1})$$

where $\mathbf{E}_i(t)$ is zero mean i.i.d. Gaussian noise with diagonal covariance matrix and $i = 1, \dots, 6$. The data is generated continuously according to the AR model in (D.1). That is, for $t = 1$ to 10,000, $\mathbf{X}(t)$ is computed based on (D.1), which satisfies Assumption 5.2.2. Hence Lemma 5.2.2 and Algorithm 6 can be applied to reconstruct the essential graph. Figure D.2 shows the phase response of the Wiener coefficients, computed using (5.5) (Step 2(a) in Algorithm 6) and Figure D.3 shows the estimated

conditional distributions on performing intervention at node 2.

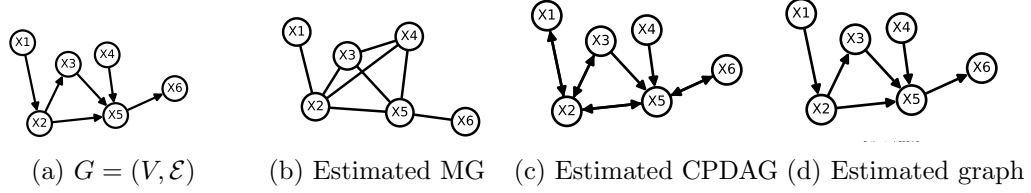


Figure D.1: (a) shows the true causal generative graph; (b) shows the estimated kin-graph from $T = 10,000$ samples with $N = 64$ using $\ell_\infty(\Omega_N)$ norm on $W_i[j]$; (c) is the CPDAG estimated using Algorithm 6. The bi-directed edges indicate undirected edges in (c); (d) shows the graph estimated by performing intervention at node 2.

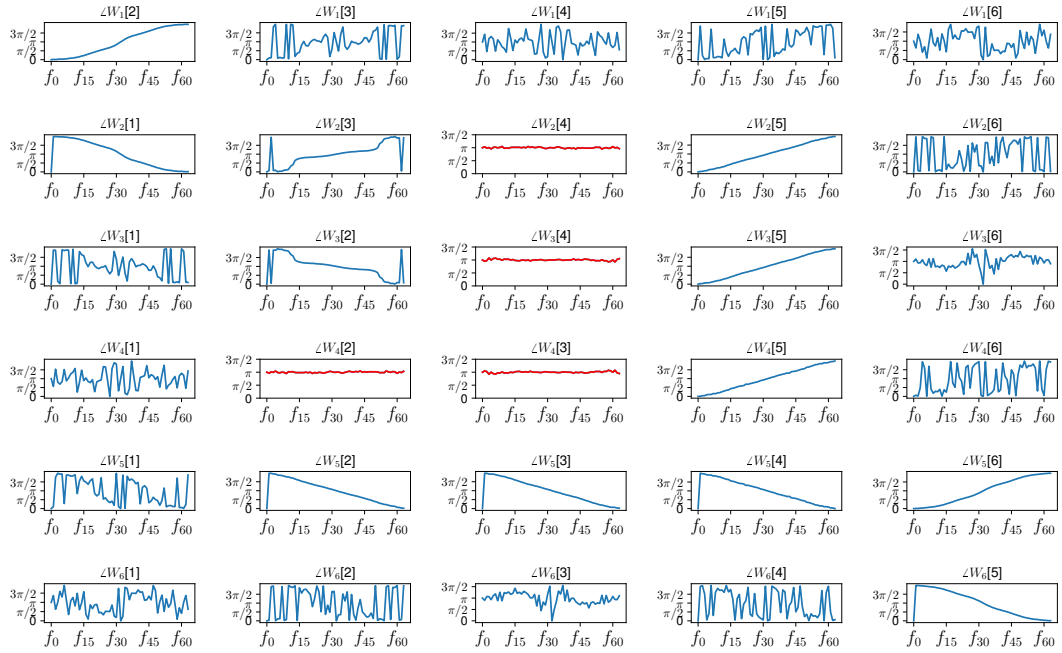


Figure D.2: Phase plot of the Wiener coefficients, $W_i[j]$, computed using equation (5.5); data with 10,000 samples is generated using (D.1). Red plots show the phase plot, $\angle W_i[j]$, between the strict spouses i and j . σ is the standard deviation of the phase along the frequency. The spurious links $(i, j) = (2, 4)$ and $(i, j) = (3, 4)$ have $\sigma < 0.9$ whereas the remaining pair of nodes have $\sigma > 1.6$.

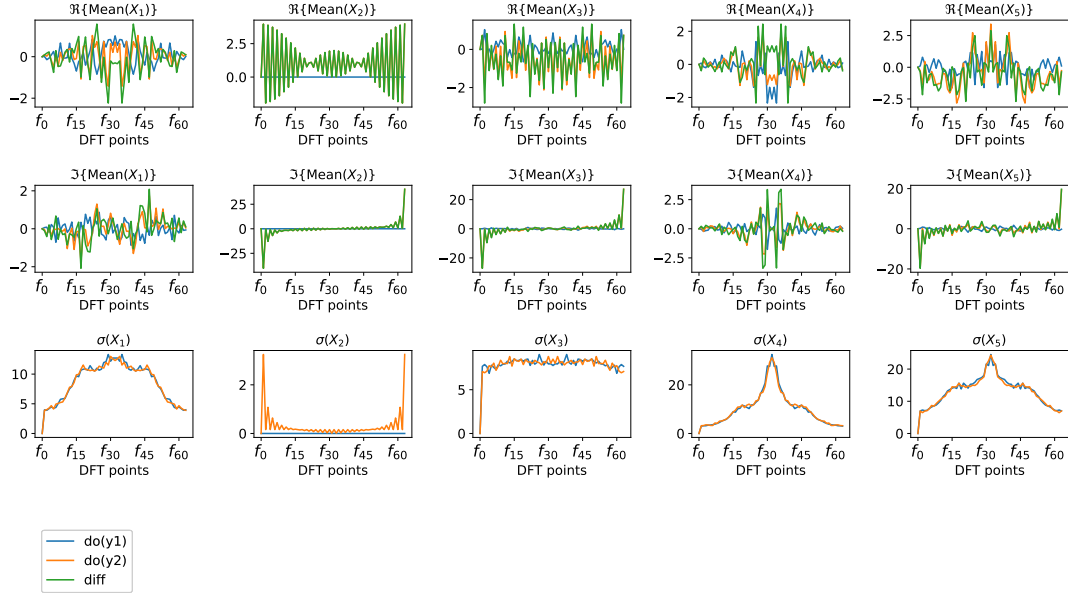


Figure D.3: Estimated conditional distribution when intervened at node 2 for two interventions, $do(X_2 = y_1)$ and $do(X_2 = y_2)$, where y_1 is all zero sequence and y_2 is a repeating sequence of 31 zeros and 33 twos. The first row shows the real part of the mean, the middle row is the imaginary part of the mean, and the bottom row is the standard deviation of the estimated distribution. The plots show that the mean of X_i , especially the mean of the imaginary part, follows the pattern of X_2 when there is a directed path to i from 2 ($i = 3, 5$). The mean of X_1 and X_4 are random, and does not follow the mean of X_2 , implying that there is no directed path to 1 and 4 from 2.

D.2.1 Restart and Record Sampling

In restart and record sampling [112], the i^{th} time-series X_i is partitioned into R segments with the r^{th} segment denoted by X_i^r . Each segment consists of N samples, for example, $X_i^r = \{X_i^r(0), \dots, X_i^r(N-1)\}$. The time series X_i is then given by $\{(X_i^r(t))_{t=0}^{N-1}\}_{r=1}^R$. In restart and record sampling, each of the time series is considered independent of the other. Here, following (D.1), the initial conditions of all $\mathbf{X}_i$ are set to zero. $\mathbf{E}_i(t)$ is generated using i.i.d. Gaussian distribution with mean zero and variance one. Samples of $\mathbf{X}_i(t)$, $i = 1, \dots, n$ are realized for $t = 0, \dots, N-1$, which are recorded. The procedure is restarted with different realizations for another epoch of $N-1$ samples. Using the r^{th} segment of the X_i trajectory, given by $X_i^r(0), \dots, X_i^r(N-1)$, the FFT, $\hat{\mathbf{X}}_i^r(e^{j\frac{2\pi k}{N}}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \mathbf{X}_i^r(n) e^{-2\pi kn/N}$ is computed. Figure D.4 and Figure D.5 show

the plots obtained for the restart and record data.

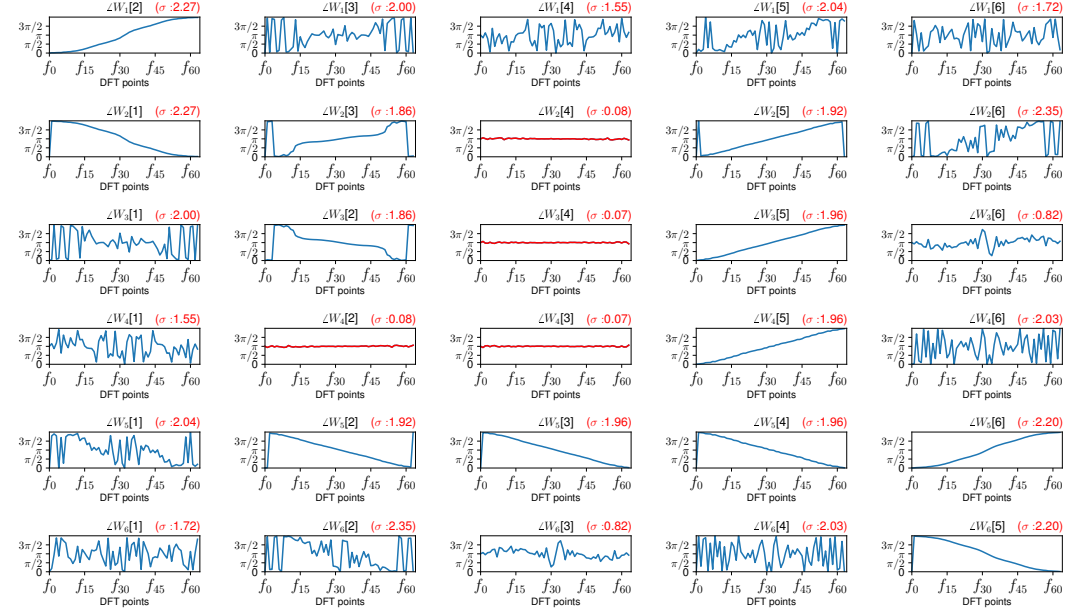


Figure D.4: Phase plot of the Wiener coefficients, $W_i[j]$, computed using equation (5.5) for the restart and record data; data with 10,000 trajectories is generated for the AR in (D.1). Red plots show the phase plot, $\angle W_i[j]$, between the strict spouses i and j . σ is the standard deviation of the phase along the frequency. The spurious links $(i, j) = (2, 4)$ and $(i, j) = (3, 4)$ have $\sigma < 0.1$ whereas the remaining pairs of nodes have $\sigma > 1.5$.

D.3 Convergence

Notation: For a continuous function f , $\hat{f}$ denotes the Fourier transform of f .

Definition D.3.1. [128] *The total variation of a continuous complex-valued function f , defined on $[a, b] \subset \mathbb{R}$ is the quantity, $V_a^b(f) := \sup_{p \in \mathcal{P}} \sum_{i=0}^{n_p-1} |f(x_{i+1}) - f(x_i)|$, where $P = \{x_0, \dots, x_{n_p-1}\}$ is a partition and $\mathcal{P}$ is the set of all partitions of $[a, b]$.*

A continuous complex-valued function f is said to be of bounded variation on a chosen interval $[a, b] \subset \mathbb{R}$ if its total variation is finite, i.e. if $V_a^b(f) < +\infty$.

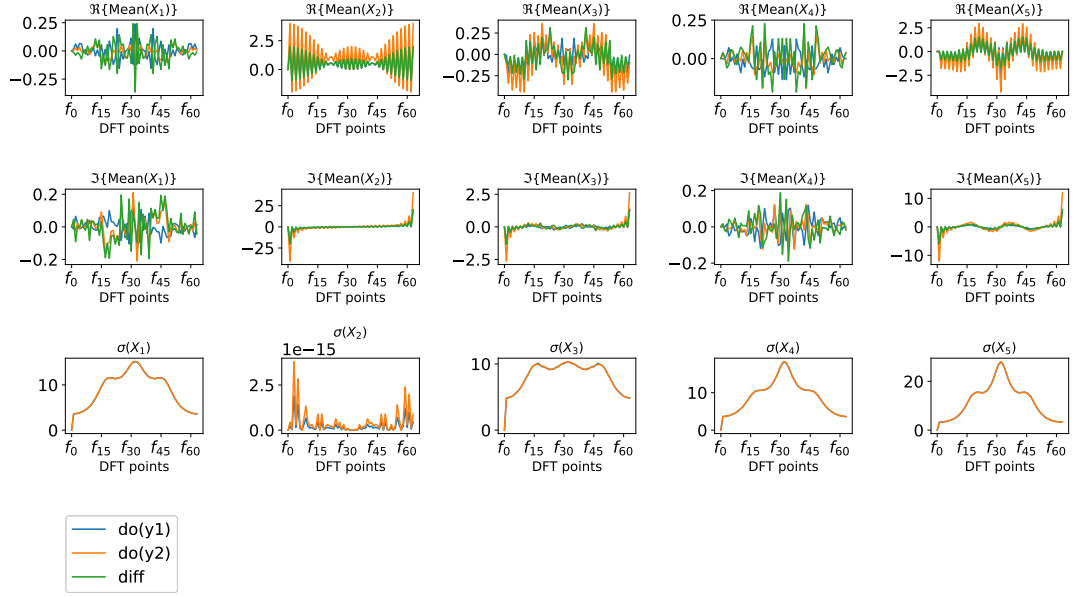


Figure D.5: Estimated conditional distribution when intervened at node 2 for two interventions, $do(X_2 = y_1)$ and $do(X_2 = y_2)$, where y_1 is all zero sequence and y_2 is a repeating sequence of 31 zeros and 33 twos. The plots show that the mean of X_i , especially the mean of the imaginary part follows the pattern of X_2 when there is a directed path to i from 2 ($i = 3, 5$). The mean of X_1 and X_4 are random, and does not follow the mean of X_2 , implying that there is no directed path to 1 and 4 from 2.

Theorem D.3.1 (Theorem 5.2.1). *For a function $\hat{f}$ of bounded variation V , on $[0, 2\pi]$ and $N \geq 1$, the estimation of $f(n) = \int_0^{2\pi} \hat{f}(\omega) e^{j\omega n} d\omega$ given by*

$$f^{(N)}(n) = \frac{1}{N} \sum_{k=0}^{N-1} e^{j2\pi nk/N} \hat{f}(2\pi k/N), \text{ for } |n| \leq \sqrt{N},$$

and zero otherwise, satisfies $\|f - f^{(N)}\|_{\ell_\infty(\Omega)} \leq C/\sqrt{N}$.

Proposition D.3.1. *For $\hat{f}$ with bounded variation, $V_a^b(\hat{f})$, on $[a, b]$ and $t \in [a, b]$*

$$\left| \int_a^b \hat{f}(x) dx - (b-a)\hat{f}(t) \right| \leq (b-a)V_a^b(\hat{f}).$$

Proof. Since $\hat{f}$ has bounded variation, $M := \sup_{x \in [a, b]} \hat{f}(x)$ and $m := \inf_{x \in [a, b]} \hat{f}(x)$ are

both finite. Thus,

$$\left| \int_a^b \widehat{f}(x) dx - (b-a)\widehat{f}(t) \right| \leq \int_a^b \left| \widehat{f}(x) - \widehat{f}(t) \right| dx \leq (M-m)(b-a) \leq (b-a)V_a^b(\widehat{f}).$$

Corollary D.3.1. *For $\widehat{f}$ a function of bounded variation on $[0, 1]$, given by $V_0^1(\widehat{f})$, the error in a Riemann sum approximation,*

$$\varepsilon_N := \left| \int_0^1 \widehat{f}(x) dx - \frac{1}{N} \sum_{n=0}^{N-1} \widehat{f}(n/N) \right| \leq V_0^1(\widehat{f})/N.$$

Proof. Note that

$$\varepsilon_N \leq \sum_{n=0}^{N-1} \left| \int_{n/N}^{(n+1)/N} \widehat{f}(x) dx - \frac{1}{N} \widehat{f}(n/N) \right| \leq \sum_{n=0}^{N-1} V_{n/N}^{(n+1)/N}(\widehat{f})/N \leq V_0^1(\widehat{f})/N,$$

where the second inequality follows from Proposition D.3.1.

Proposition D.3.2 (Proposition 3.2.14, [128]). *If $\widehat{f}$ is of bounded variation, then*

$$f(n) \leq \frac{V_0^1(\widehat{f})}{2\pi|n|}, \quad n \neq 0.$$

Corollary D.3.2. *For $n \in \mathbb{N}$, and $\varepsilon > 0$ fixed. the estimation of the Fourier series $f(n)$ given by*

$$f^{(N)}(n) = \begin{cases} \sum_{k=0}^{N-1} e^{j2\pi kn/N} \widehat{f}(k/N) & |n| \leq N^{1-\varepsilon} \\ 0 & \text{otherwise} \end{cases}$$

converges uniformly to $f(n)$.

Proof. For $|n| < N^{1-\varepsilon}$, it follows by Corollary D.3.2 that

$$\begin{aligned} \left| f(n) - f^{(N)}(n) \right| &\leq \sqrt{2} \frac{\|\widehat{f}\|_{\infty} 2\pi|n| + V_0^1(\widehat{f})}{N} \\ &\leq \frac{\sqrt{2}\|\widehat{f}\|_{\infty} 2\pi}{N^{\varepsilon}} + \frac{\sqrt{2}V_0^1(\widehat{f})}{N}. \end{aligned}$$

Similarly, for $n \geq N^{1-\varepsilon}$

$$\left| f(n) - f^{(N)}(n) \right| \leq \frac{V_0^1(\widehat{f})}{2\pi|n|} \leq \frac{V_0^1(\widehat{f})}{2\pi N^{1-\varepsilon}}.$$

The proof of Theorem D.3.1 follows by letting $\varepsilon = 1/2$. □

D.4 Proof of Lemmas

D.4.1 Proof sketch of Lemma 5.2.1

Recall that $\Phi_{\mathbf{X}}^{-1} = (\mathbf{I} - \mathbf{H})^* \Phi_{\mathbf{E}}^{-1} (\mathbf{I} - \mathbf{H})$. Then, for any $i \neq j$, $[\Phi_{\mathbf{X}}^{-1}]_{ij} = -\mathbf{H}_{ji}^* - \mathbf{H}_{ij} + \sum_{k=1}^n \mathbf{H}_{ki}^* \mathbf{H}_{kj}$. The Lemma follows. Detailed proof is provided in [17]. □

D.4.2 Proof of Lemma 5.3.2

Here, for convenience, we're restating Lemma 5.3.2

Lemma D.4.1 (Lemma 5.3.2). *Consider a set V of stochastic process that factorizes according $G^{(\omega)} = (V, \mathcal{E}^{(\omega)})$ at $\omega \in \Omega$. Let $G = (V, \mathcal{E})$ be such that $(i, j) \in \mathcal{E}$ if and only if $(i, j) \in \mathcal{E}^{(\omega)}$ for some $\omega \in \Omega$. Let $X, Y, Z \subset V$ be disjoint sets. If Z d-separates X and Y in G , then Z d-separates X and Y in $G^{(\omega)}$.*

Proof. We prove this by contrapositive argument. Suppose Z , d-connects X and Y in $G^{(\omega)}$. Then, there exists an $x \in X$ and $y \in Y$ and a path $\pi : x - s_1 - \dots - s_n - y$ such that π is a d-connected path between x and y given Z in $G^{(\omega)}$. Then, in G , since we are not removing any edge that is present in $G^{(\omega)}$, the d-connected path remains d-connected.

D.4.3 Proof of Lemma 5.3.3

Lemma D.4.2 (Lemma 5.3.3). *Consider a set V of stochastic process that is described by the LDIM (5.1) (here $\mathbf{E}_i(k) \perp\!\!\!\perp \mathbf{E}_j(l)$ for $i \neq j$ and $k, l \in \mathbb{Z}$). Moreover, suppose the set V of stochastic processes factorizes according to directed graph $G = (V, \mathcal{E})$ and at a specific frequency $\omega \in \Omega$ factorizes according to a directed graph $G^{(\omega)} = (V, \mathcal{E}^{(\omega)})$. Let $X, Y, Z \subset V$ be disjoint sets. Then Z d-separates X and Y in G if and only if Z d-separates X and Y in $G^{(\omega)}$, for almost all $\omega \in \Omega$.*

Proof. The proof follows from Lemma D.5.1. Recall that G is defined as $G := \bigcup_{\omega \in \Omega} G^{(\omega)}$. Applying Lemma D.5.1 immediately provides $\text{supp}(\mathbf{H}(\omega_1)) = \text{supp}(\mathbf{H}(\omega_2))$ and $\mathcal{G}^{(\omega_1)} = \mathcal{G}^{(\omega_2)}$ for any $\omega_1, \omega_2 \in \Omega$, almost always. Thus, $\mathcal{G}^{(\omega)} = \mathcal{G}$ almost always in LDIMs. The result follows.

D.4.4 Proof of Lemma 5.3.4

The proof follows by combining Lemma 5.3.2 and Lemma 5.3.3.

D.5 Which frequency to choose?

Most practical systems are finite dimensional with finite dimensional realizations. These admit rational transfer functions.

The following Lemma proves that, in identifying the CI, graph/moral graph, and topology, it is sufficient to work with one frequency selected arbitrarily, in order to get meaningful result, when we have access to large enough data.

Lemma D.5.1. *Consider any rational polynomial transfer function*

$$\mathbf{H}_{ij}(z) = C \frac{(z - a_1)(z - a_2) \dots (z - a_p)}{(z - b_1)(z - b_2) \dots (z - b_q)}, \quad p \leq q, \quad C \in \mathbb{C}.$$

For any distinct $\omega_1, \omega_2 \in \Omega$, $\mathbf{H}_{ij}(e^{j\omega_1}) \neq 0$ if and only if $\mathbf{H}_{ij}(e^{j\omega_2}) \neq 0$ almost surely w.r.t. a continuous probability measure.

Proof. By the fundamental theorem of algebra [62], there exists at most p complex zeros for $\mathbf{H}_{ij}(z)$, if $\mathbf{H}_{ij}(z)$ is not identically zero, which form a set of Lebesgue measure zero. Suppose $\mathbf{H}_{ij}(e^{j\omega_1}) \neq 0$, which implies that $\mathbf{H}_{ij}$ is not identically zero. Thus, $\mathbf{H}_{ij}(e^{j\omega_2}) \neq 0$ almost everywhere. $\square$

Remark D.5.1. *Since $\Phi_x^{-1}(e^{j\omega}) = (\mathbf{I} - \mathbf{H}(e^{j\omega}))^* \Phi_e^{-1}(e^{j\omega})(\mathbf{I} - \mathbf{H}(e^{j\omega}))$, the same can be said about $\Phi_{\mathbf{X}}(e^{j\omega_1})$ and $\Phi_{\mathbf{X}}(e^{j\omega_2})$ and Wiener coefficients also. Thus, it is sufficient to pick any $\omega \in \Omega$ and apply the algorithms for that frequency.*

D.6 Direct and Total Effect Identification in SPs

D.6.1 Direct Effect Identification

In the classical causal inference from IID data, some of the popular approaches on parameter identification involve single-door, back-door and front-door criterion [48]. In this work, we extend these results to dynamically related stochastic processes by realizing them in frequency-domain. For the single-door criterion (with cycles) in frequency-domain (for WSS processes), first proved in [27], we provide a simpler proof.

Theorem D.6.1 (Single-door criterion). *Consider a directed graph, $G = (V, \mathcal{E})$, that represents an LDIM, defined by (5.1). Let $U, Y \in V$ and $Z \subset V$ with $X_Y(\omega) = \alpha(\omega)X_U(\omega) + \sum_{p \in Pa(Y) \setminus U} H_{Yp}(\omega)X_p(\omega) + E_Y(\omega)$, for every $\omega \in \Omega$, where $\alpha(\omega)$ is the weight of the link $U \rightarrow Y$. Let $G' := G_{\overline{U \rightarrow Y}}$ be the graph obtained by deleting the edge $U \rightarrow Y$. Suppose that U, Y , and Z satisfy the single door criteria described by: (C1) Y has no descendants in Z ; (C2) Z d-separates U and Y in $G_{\overline{U \rightarrow Y}}$; (C3) There is no directed cycle involving Y . Then the coefficient of U in the projection of Y to $[U, Z]$, denoted, $W_{YU \cdot Z}(\omega) = \alpha(\omega)$.*

Proof. See Appendix D.7

Suppose there exists $Pa_i := Pa(i) \subset V$, $i \in V$ such that the factorization in (5.6) and (5.7) exist. Then one can perform atomic interventions and the do-calculus on the SPs by performing them at every $\omega \in \Omega$.

D.6.2 Total Effect Identification

One can define the back-door criterion on SPs using CIFD as follows. Notice that a similar result can be obtained using the FDD based independence notion also. The results in this section require the following definitions. A *directed path* from node i to node j in the directed graph G is an ordered set of edges $((\ell_0, \ell_1), (\ell_1, \ell_2), \dots, (\ell_{n-1}, \ell_n))$ with $\ell_0 = i$, $\ell_n = j$ in $G = (V, \mathcal{E})$, where $(\ell_i, \ell_{i+1}) \in \mathcal{E}$.

Definition D.6.1 (Back-door Criterion). *Given an ordered pair of SPs (W, Y) in a DAG $G^{(\omega)}$, a set of SPs Z satisfies the back-door criterion relative to (W, Y) if the following hold:*

- *No node in Z is a descendant of W and*
- *Z blocks (or d-separates) every path between W and Y that contains an arrow into W .*

Theorem D.6.2 (Back-door Adjustment). *Consider a set of stochastic processes $\{\mathbf{X}_V\} := \mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_n]$, where $V = \{1, \dots, n\}$. Suppose that $G^{(\omega)}$ is a DAG compatible with $\mathcal{P}$ with the pdf given by $f^{(\omega)}$. Further, suppose that the processes $W, Y, Z \subset V$ are such that W and Y are disjoint and Z satisfies the back-door criterion (Definition D.6.1) relative to (W, Y) . Then, the causal effect of W on Y (for notational simplicity W, Y, Z denotes $\mathbf{X}_W, \mathbf{X}_Y, \mathbf{X}_Z$) is identifiable and is given by $f^{(\omega)}(y(\omega) \mid do(W = w)) = \int f^{(\omega)}(y(\omega) \mid w(\omega), z(\omega)) f^{(\omega)}(z(\omega)) d\nu(z(\omega)) \omega \in \Omega$.*

Proof. See Appendix D.8

Definition D.6.2 (Front-door Criterion). *Given an ordered pair of sets $(W, Y) \subset V \times V$ in a DAG $G^{(\omega)}$, a set of SPs $Z \subset V$ satisfies the front-door criterion relative to (W, Y) if the following hold:*

- *Every directed path from W to Y has a node $z \in Z$.*
- *All back-door paths between Z and Y are blocked by W . That is, W d-separates every path between Z and Y which has an arrow into Z and*
- *There are no back-door paths activated between W and Z . That is, the empty set d-separates any path between W and Z with arrow into W .*

Theorem D.6.3 (Front-door adjustment). *Consider a set of stochastic processes $\{\mathbf{X}_V\} := \mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_n]$, where $V = \{1, \dots, n\}$. Suppose that $G^{(\omega)}$ is a DAG compatible with $\mathcal{P}$ with the pdf given by $f^{(\omega)}$. Further, suppose that the sets $W, Y, Z \subset V$ are such that W and Y are disjoint and Z satisfies the front-door criterion (Definition D.6.2) relative to (W, Y) . Then, the causal effect of W on Y (for notational simplicity W, Y, Z denotes $\mathbf{X}_W, \mathbf{X}_Y, \mathbf{X}_Z$) is identifiable and is given by:*

$$f^{(\omega)}(y(\omega) \mid do(W = w)) = \int f^{(\omega)}(z(\omega) \mid w(\omega)) \int f^{(\omega)}(y(\omega) \mid w'(\omega), z(\omega)) f^{(\omega)}(z(\omega)) d\nu(w'(\omega)) d\nu(z(\omega)).$$

Proof. The proof follows similar to the back-door criterion.

Remark D.6.1. *One can extend the rules of do-calculus also to the SPs in the same way and is skipped due to space constraints.*

D.7 Proof of Theorem D.6.1 (Single-door)

The following Proposition is useful in proving the succeeding results.

Lemma D.7.1 (Linearity of Wiener Coefficient). *Let $X, Y, Z_1, \dots, Z_n$ be stochastic processes, and let $\tilde{Z} := \text{span}(X, Z_1, \dots, Z_n)$ and $\mathbf{Z} = [Z_1, \dots, Z_n]$. Let $W_{YX.\mathbf{Z}}$ be the coefficient of X in projecting Y on to $\tilde{Z}$. Suppose $Y(\omega) = \sum_{k=1}^m \alpha_k(\omega) Y_k(\omega)$, where*

$\alpha_1, \dots, \alpha_m$ are transfer functions and $m \in \mathbf{N}$. Then,

$$W_{YX \cdot \mathbf{Z}}(\omega) = \sum_{i=1}^m \alpha_i(\omega) W_{Y_i X \cdot \mathbf{Z}}(\omega)$$

Proof. See Appendix D.10.4.

The following lemma, which follows from [27] shows the relation between d-separation and Wiener coefficients.

Lemma D.7.2. *Consider a directed graph $G = (V, \mathcal{E})$ that represents an LDIM, (5.1). Let $X, Y, Z \subset V$ be disjoint sets. Suppose that $\text{d-sep}_G(X, Z, Y)$. Then $W_{ij \cdot Z} = 0$, for every $i \in X$ and $j \in Y$.*

Proof. The result follows from Theorem 24 in [27].

Now we can prove the single-door criterion.

D.7.1 Proof: single-door

Notation: Here, $W_{pU \cdot Z}^{(L')}$ denotes the co-efficient of $\mathbf{X}_U$ in the projection of the time-series $\mathbf{X}_p$ to $\mathbf{X}_{[U, Z]}$.

Claim 1: If U and Y are d-separated by Z in graph G' , then U and any $p \in \text{Pa}(Y) \setminus (Z \cup \{U\})$ are also d-separated by Z in G' .

Proof. The claim can be proved by contrapositive argument. Suppose U and $p \in \text{Pa}(Y) \setminus \{U\}$ be d-connected by Z in G' . Then, there exists a path $PT : \{U, \pi_1, \dots, \pi_n, p\}$ in G' such that

1. If the path is a chain then $\pi_i, p \notin Z$ (since the path is disconnected otherwise).
2. If there exists a fork π_j in the path, then $\pi_j \notin Z$.
3. If there are colliders in $(\pi_i)_{i=1}^n$, then either the colliders or their descendants are present in Z .

Presence of the path PT implies that there exists a path $PT_Y := \{U, \pi_1, \dots, \pi_n, p \rightarrow Y\}$, which is obtained by adding $p \rightarrow Y$ to PT . In the path PT_Y , p can be a fork element or a chain element, which does not belong to Z . Thus, the path PT_Y remains d-connected given Z in G' , which contradicts the d-separation assumption.

Claim 2: For any $p \in \text{Pa}(Y) \setminus (Z \cup \{U\})$ $\text{d-sep}_{G'}(U, Z, p) \implies \text{d-sep}_G(U, Z, p)$

Suppose U and p are d-connected by Z in G . Then, there exists a path $PT : (U, p_1, \dots, p_n, p_{n+1} = p)$ in G such that

1. If PT is a chain, then $p_i, p \notin Z$.
2. If there exists a fork p_j , then $p_j \notin Z$.
3. If there are colliders in $(p_i)_{i=1}^n$, then for every collider, either the collider or its descendant is present in Z .

Notice that the only difference between G and G' is the absence of edge $U \rightarrow Y$ in G' .

We devide the proof into two cases

Case 1: Suppose the path $PT = U - \pi_1 - \pi_2 - \dots - \pi_n - p$ has no colliders in the graph G . Then it has atmost one fork.

1. Suppose the path PT has one fork at π_k . Thus the path $PT \equiv U \leftarrow \pi_1 \leftarrow \pi_2 \leftarrow \dots \leftarrow \pi_{k-1} \leftarrow \pi_k \rightarrow \pi_{k+1} \rightarrow \dots \rightarrow p$. Suppose the subpath $PT_2 \equiv \pi_k \rightarrow \pi_{k+1} \rightarrow \dots \rightarrow p$ has the link $U \rightarrow Y$. Then the path $\pi_k \rightarrow \pi_{k+1} \rightarrow \dots \rightarrow U \rightarrow Y \rightarrow \dots \rightarrow p \rightarrow Y$ exists in G which involves Y in a directed loop leading to a contradiction. Thus $U \rightarrow Y$ is not present in PT_2 . Now suppose the link $Y \leftarrow U$ is present in the subpath $PT_1 \equiv U \leftarrow \pi_1 \leftarrow \pi_2 \leftarrow \dots \leftarrow \pi_{k-1} \leftarrow \pi_k$. Suppose $PT_1 \equiv U \leftarrow \pi_1 \leftarrow \pi_2 \leftarrow Y = \pi_{l-1} \leftarrow U = \pi_l \leftarrow \dots \leftarrow \pi_{k-1} \leftarrow \pi_k$ which implies Y is in a directed loop leading to a contradiction. Thus the path PT does not have a link of the form $U \rightarrow Y$ and remains d-connected by Z in G' .
2. Suppose the path has no forks. As there are no colliders the path is a chain.
 - (a) Suppose the path, PT is chain from U to Y . Suppose there is the link $U \rightarrow Y$ as a subpath in PT . Then Y will be a part of a directed loop which leads to a contradiction.
 - (b) Suppose the path PT is a chain from p to U . Suppose there a link $Y \leftarrow U$ in PT . Again, Y will be part of a directed loop which is a contradiction. Thus the PT remains d-connected by Z in G' .

Case 2: Suppose the path PT has m colliders, $c_1 \dots, c_m$ with $PT \equiv U - \dots - \rightarrow c_1 \leftarrow \dots \leftarrow \rightarrow c_2 \leftarrow \dots \leftarrow \rightarrow c_m \leftarrow \dots \rightarrow p$. Consider the subpath $PT_i \equiv c_i \leftarrow \pi_1 - \dots - \pi_q \rightarrow c_{i+1}$ of PT between c_i and c_{i+1} which are colliders. As the path PT is d-connected by Z , c_i and c_{i+1} or their descendants are in Z . Without loss of generality we assume the path $z_\ell = \ell_k \leftarrow \ell_{k-1} \leftarrow \dots \leftarrow \ell_1 \leftarrow c_i \leftarrow \pi_1 - \dots - \pi_q \rightarrow c_{i+1} \rightarrow r_1 \rightarrow r_2 \dots \rightarrow r_s = z_r$ where z_r and z_ℓ are descendants of c_{i+1} and c_i in Z respectively exists (d-connecting the path). Note that none of the nodes

ℓ_j or r_j can be Y as no descendant of Y can be in Z . Thus the link $U \rightarrow Y$ cannot be present in the subpath $PT_{i\ell} \equiv z_\ell = \ell_k \leftarrow \ell_{k-1} \leftarrow \dots \leftarrow \ell_1 \leftarrow c_i$ or in $PT_{ir} \equiv c_{i+1} \rightarrow r_1 \rightarrow r_2 \dots \rightarrow r_s = z_r$.

The subpath PT_i has no colliders and is a path between colliders and thus it has one fork say at π_t so that the path $PT_i \equiv c_i \leftarrow \pi_1 - \dots - \leftarrow \pi_t \rightarrow - \dots \rightarrow c_{i+1}$. Without loss of generality assume Y appears in the subpath $c_i \leftarrow \pi_1 \leftarrow \dots \leftarrow \pi_t$. Then examining the path $PT_{i\ell}$ concatenated with $c_i \leftarrow \pi_1 \leftarrow \dots \leftarrow \pi_t$ results in Y having a descendant in Z which is a contradiction. Thus Y cannot be in $c_i \leftarrow \pi_1 - \dots - \leftarrow \pi_t$. Similarly Y cannot be in $\pi_t \rightarrow \dots \rightarrow c_{i+1}$. Thus the link $U \rightarrow Y$ cannot be in PT_i . In conclusion, the link $U \rightarrow Y$ cannot be in $z_\ell = \ell_k \leftarrow \ell_{k-1} \leftarrow \dots \leftarrow \ell_1 \leftarrow c_i \leftarrow \pi_1 - \dots - \pi_q \rightarrow c_{i+1} \rightarrow r_1 \rightarrow r_2 \dots \rightarrow r_s = z_r$ which implies the path PT_i remains d-connected by Z in G' .

Thus the (U, p) remain d-connected by Z in G' .

Thus, $\text{d-sep}_{G'}(U, Z, p) \implies \text{d-sep}_G(U, Z, p)$ and it follows from Lemma D.7.2 that $W_{pU \cdot Z}^{(L)} = 0$. Notice that if $p \in Z$, then $W_{pU \cdot Z}^{(L)} = 0$. From Linearity of W , Lemma D.7.1,

$$\begin{aligned} W_{Y, U \cdot Z}^{(L)} &= W_{(\alpha U + \sum_{p \in Pa(Y) \setminus U} a_{Yp} p + \varepsilon_Y), x \cdot Z}^{(L)} \\ &= \alpha W_{U, U \cdot Z}^{(L)} + \sum_{p \in Pa(Y) \setminus \{x\}} a_{Yp} W_{p, U \cdot Z}^{(L)} + W_{\varepsilon_Y, U \cdot Z}^{(L)} \\ &= \alpha + 0. \end{aligned}$$

If $Pa(Y) \setminus \{U\} = \emptyset$, then the above equations hold.

D.8 Proof of Theorem D.6.2

D.8.1 Atomic interventions

$$\mathcal{P}(X_{(1)}(\omega), \dots, X_{(n)}(\omega) \mid do(x_i)) = \begin{cases} \prod_{k=1, k \neq i}^n \mathcal{P}(X_k(\omega) \mid X_{(Pa_k)}(\omega)) & \text{if } X_i(\omega) = x_i(\omega) \\ 0 & \text{otherwise} \end{cases} \quad (\text{D.2})$$

Definition D.8.1 (The Causal Markov Condition). *Suppose that we want to condition on $X_i(\omega) = x_i(\omega)$. Then, the intervened distribution is obtained by removing the factor*

$\mathcal{P}(X_i(\omega) \mid X_{Pa_i}(\omega))$ from the product in (5.6). That is, the distribution satisfies (D.2). Alternately, $\mathcal{P}(\cdot)$ can be replaced with the pdf $f(\cdot)$ if f exists.

We first prove the following auxiliary theorem and lemma.

Theorem D.8.1 (Causal effect). *Let $V := \{1, \dots, n\}$, $A \subset V$ and let $i \in V$. The causal effect of X_i on a set of SPs X_A can be identified if $i \cup A \cup Pa_i$ are observable and $A \cap (i \cup Pa_i) = \emptyset$. The causal effect of X_i on X_A is given by*

$$f^{(\omega)}(x_A(\omega) \mid x_i(\omega)) = \int f^{(\omega)}(x_A(\omega) \mid x_i(\omega), x_{Pa_i}(\omega)) f^{(\omega)}(x_{Pa_i}(\omega)) d\nu(x_{Pa_i}(\omega)),$$

if $X_i(\omega) = x_i(\omega)$ and zero otherwise.

Proof. The proof follows by Baye's rule and the definition of conditional probability, $\mathcal{P}(Y \mid Z) = \frac{\mathcal{P}(Y, Z)}{\mathcal{P}(Z)}$. If $X_i = x_i$, applying (D.2),

$$\begin{aligned} \mathcal{P}(X_1(\omega), \dots, X_n(\omega) \mid X_i(\omega)) &= \prod_{k=1, k \neq i}^n \mathcal{P}(X_k(\omega) \mid X_{Pa_k}(\omega)) \\ &= \frac{\prod_{k=1}^n \mathcal{P}(X_k(\omega) \mid X_{(Pa_k)}(\omega))}{\mathcal{P}(X_i(\omega) \mid X_{Pa_i}(\omega))} \\ &= \frac{\mathcal{P}(X_1(\omega), \dots, X_n(\omega))}{\mathcal{P}(X_i(\omega) \mid X_{Pa_i}(\omega))} \\ &= \mathcal{P}(X_1(\omega), \dots, X_n(\omega) \mid X_i(\omega), X_{Pa_i}(\omega)) \mathcal{P}(X_{Pa_i}(\omega)). \end{aligned}$$

Let $U = (V \setminus (\{i\} \cup A))$ and $U_i := V \setminus (\{i\} \cup A \cup Pa_i)$. Let f be the joint density

function. Then,

$$\begin{aligned}
f(x_A(\omega) \mid x_i(\omega)) &\stackrel{(a)}{=} \int f(x_V(\omega) \mid x_i(\omega)) d\nu(x_U(\omega)) \\
&\stackrel{(b)}{=} \int f(x_V \mid x_i, x_{Pa_i}) f(x_{Pa_i}) d\nu(x_U) \\
&\stackrel{(c)}{=} \int \int f(x_V \mid x_i, x_{Pa_i}) f(x_{Pa_i}) d\nu(x_{U_i}) d\nu(x_{Pa_i}) \\
&= \int f(x_{Pa_i}) \int f(x_V \mid x_i, x_{Pa_i}) d\nu(x_{U_i}) d\nu(x_{Pa_i}) \\
&= \int f(x_{Pa_i}) \int \frac{f(x_V)}{f(x_i, x_{Pa_i})} d\nu(x_{U_i}) d\nu(x_{Pa_i}) \\
&\stackrel{(d)}{=} \int f(x_{Pa_i}) \frac{f(x_A, x_i, x_{Pa_i})}{f(x_i, x_{Pa_i})} d\nu(x_{Pa_i}) \\
&= \int f(x_A \mid x_i, x_{Pa_i}) f(x_{Pa_i}) d\nu(x_{Pa_i}),
\end{aligned}$$

where (a) follows by marginalization, (b) definition of conditional probability, (c) Fubini's theorem, and (d) follows by marginalization over U_i . Thus, $f(x_A \mid x_i)$ can be determined if X_A , X_i , and X_{Pa_i} are measured. $\square$

The following is an auxiliary Lemma useful in proving back-door criterion.

Lemma D.8.1. *Consider a set of stochastic processes $\mathbf{X} := \mathbf{X}_V = \{X_1, \dots, X_n\}$, where $V = \{1, \dots, n\}$. Suppose $G^{(\omega)}$ is a directed graph compatible with $\mathcal{P}(\cdot)$ at ω . Further, suppose that the processes $W, Y, Z \subset V$ are such that*

1. $W \perp\!\!\!\perp^{(\omega)} Z \mid Pa_W$,
2. $Y \perp\!\!\!\perp^{(\omega)} Pa_W \mid W, Z$

where $Pa_W \subset V$ is the set of parents of W . Then,

$$f^{(\omega)}(y(\omega) \mid do(W = w)) = \int f^{(\omega)}(y \mid w(\omega), z(\omega)) f^{(\omega)}(\omega) d\nu(z(\omega)).$$

Proof. Let $U = Pa_W$. From Theorem D.8.1 (ignoring the index ω),

$$\begin{aligned}
f(y \mid do(W = w)) &= \int f(y \mid w, u) f(u) d\nu(u) \\
&\stackrel{(a)}{=} \int \int f(y \mid w, u, z) f(z \mid u, w) f(u) d\nu(z) d\nu(u) \\
&\stackrel{(b)}{=} \int \int f(y \mid w, z) f(z \mid u, w) f(u) d\nu(z) d\nu(u) \\
&\stackrel{(c)}{=} \int \int f(y \mid w, z) f(z \mid u) f(u) d\nu(u) d\nu(z) \\
&= \int f(y \mid w, z) \left[\int f(z \mid u) f(u) d\nu(u) \right] d\nu(z) \\
&\stackrel{(d)}{=} \int f(y \mid w, z) f(z) d\nu(z).
\end{aligned}$$

where (a) follows by marginalization over Z , (b) $Y \perp\!\!\!\perp U \mid W, Z$, (c) $z \perp\!\!\!\perp w \mid u$, and (d) marginalization. $\square$

D.8.2 Back-door Criterion

Definition D.8.2 (Back-door Criterion). *Given an ordered pair of SPs (W, Y) in a DAG $G^{(\omega)}$, a set of SPs Z satisfies the back-door criterion relative to (W, Y) if the following hold:*

- No node in Z is a descendant of W and
- Z blocks (or d-separates) every path between W and Y that contains an arrow into W .

Proof of Theorem D.6.2: The proof follows from Lemma D.8.1. If Z does not contain descendants of W then $Z \perp\!\!\!\perp W \mid Pa_W$. If Z blocks every back-door path between Y and W , then $Y \perp\!\!\!\perp Pa_W \mid W, Z$, given that Z is not a descendent of W . The absence of a loop involving W , Y , and Z ensures that Z is not a descendent of W . The result follows. $\square$

D.9 Proof of Theorem 5.4.1

Lemma D.9.1 (Theorem 6, [1]). *Consider a linear dynamical system governed by (5.1). Suppose that the autocorrelation function $R_x(k)$ satisfies exponential decay,*

$\|R_x(k)\|_2 \leq C\delta^{-|k|}$. For any $0 < \epsilon_1 < \epsilon$ and $L \geq \log_\delta \left(\frac{(1-\delta)\epsilon}{2C} \right)$,

$$\mathcal{P} \left(\|\Phi_{\mathbf{X}} - \widehat{\Phi}_x\|_{max} > \epsilon \right) \leq p^2 \exp \left(-(N-L) \min \left\{ \frac{\epsilon_1^2}{32(2L+1)^2 n^2 C^2}, \frac{\epsilon_1}{8(2L+1)nC} \right\} \right). \quad (\text{D.3})$$

Remark D.9.1. In the above theorem, we can take $\epsilon_1 = 9\epsilon/10$, which gives a good looking result. The following Lemma uses this.

The following lemma bounds the IPSDM by applying the sufficient condition that $\|\Phi_{\mathbf{X}} - \widehat{\Phi}_x\|_{max} < \epsilon/M^4 \implies \|\Phi_{\mathbf{X}}^{-1} - \widehat{\Phi}_x^{-1}\|_{max} < \epsilon$ and applies $\epsilon_1 = 9\epsilon/10$.

Lemma D.9.2. Consider a linear dynamical system governed by (5.1). Suppose that the auto-correlation function $R_x(k)$ satisfies exponential decay, $\|R_x(k)\|_2 \leq C\delta^{-|k|}$ and that there exists M such that $\frac{1}{M} \leq \lambda_{min}(\Phi_{\mathbf{X}}) \leq \lambda_{max}(\Phi_{\mathbf{X}}) \leq M$. Then for any $0 < \epsilon$ and $L \geq \log_\delta \left(\frac{(1-\delta)\epsilon}{2C} \right)$,

$$\mathcal{P} \left(\|\Phi_x^{-1} - \widehat{\Phi}_x^{-1}\|_{max} > \epsilon \right) \leq n^2 \exp \left(-(N-L) \min \left\{ \frac{81\epsilon^2}{3200M^{16}(2L+1)^2 n^2 C^2}, \frac{9\epsilon}{80M^4(2L+1)nC} \right\} \right). \quad (\text{D.4})$$

Proof. The following lemma from [65] is useful in deriving this.

Lemma D.9.3. For any invertible matrices $\mathbf{A}$ and $\mathbf{B}$ with $\|\mathbf{A}^{-1}\|_2 \|\mathbf{B} - \mathbf{A}\|_2 < 1$, we have

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_2 \leq \|\mathbf{A}^{-1}\|_2 \|\mathbf{A}\|_2^{-1} \|\mathbf{B} - \mathbf{A}\|_2 \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A}) \frac{\|\mathbf{B} - \mathbf{A}\|_2}{\|\mathbf{A}\|_2}}, \quad (\text{D.5})$$

where $\kappa(\mathbf{A})$ is the condition number of $\mathbf{A}$.

Let $\mathbf{A} = \Phi_x(z)$ and $\mathbf{B} = \widehat{\Phi}_x(z)$. Notice that $\kappa(\Phi_x) \leq M^2$. Then, by applying Lemma D.9.3,

$$\begin{aligned} \|\Phi_x^{-1} - \widehat{\Phi}_x^{-1}\|_2 &\leq \|\Phi_x^{-1}\|_2 \|\Phi_x\|_2^{-1} \|\widehat{\Phi}_x - \Phi_x\|_2 \frac{M^2}{1 - M^2 \frac{\|\widehat{\Phi}_x - \Phi_x\|_2}{\|\Phi_x\|_2}}, \\ &\leq \frac{M^4 \|\widehat{\Phi}_x - \Phi_x\|_2}{1 - M \|\widehat{\Phi}_x - \Phi_x\|_2} < \epsilon, \end{aligned} \quad (\text{D.6})$$

if $\|\widehat{\Phi}_x - \Phi_x\|_2 < \epsilon/M^4$. Letting $\epsilon_1 = \frac{9\epsilon}{10M^4}$ and $\epsilon \rightarrow \epsilon/M^4$ gives the Lemma statement. For any matrix A and vector $\mathbf{x}$, $\|A\mathbf{x}\|_2 \leq \|A\|_2$. Also, $\|A\|_{\max} = \max_{i,j} e_i^T A e_j \leq \|A\|_2$. Similarly, there exists a $c_1 \in \mathbb{R}$ such that $\|A\|_2 \leq c_1 \|A\|_{\max}$. Then the above concentration bounds on the maximum value extends to the spectral norm also.

Let $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}$ be the estimated values of $\mathbf{A}, \mathbf{B}$. Then,

$$\begin{aligned} \|\mathbf{AB} - \widehat{\mathbf{A}}\widehat{\mathbf{B}}\|_2 &= \|\mathbf{AB} - \widehat{\mathbf{A}}(\widehat{\mathbf{B}} + \mathbf{B} - \mathbf{B})\|_2 \\ &\leq \|(\mathbf{A} - \widehat{\mathbf{A}})\mathbf{B}\|_2 + \|\widehat{\mathbf{A}}(\widehat{\mathbf{B}} - \mathbf{B})\|_2 \\ &\leq \|(\mathbf{A} - \widehat{\mathbf{A}})\mathbf{B}\|_2 + \|\mathbf{A}(\widehat{\mathbf{B}} - \mathbf{B})\|_2 + \|(\widehat{\mathbf{A}} - \mathbf{A})(\widehat{\mathbf{B}} - \mathbf{B})\|_2 \\ &\leq M\|\mathbf{A} - \widehat{\mathbf{A}}\|_2 + M\|\widehat{\mathbf{B}} - \mathbf{B}\|_2 + \|\widehat{\mathbf{A}} - \mathbf{A}\|_2\|\widehat{\mathbf{B}} - \mathbf{B}\|_2 \\ &< M\frac{\epsilon}{3M} + M\frac{\epsilon}{3M} + \frac{\sqrt{\epsilon}\sqrt{\epsilon}}{3}. \end{aligned}$$

For our Wiener coefficient estimation, $\mathbf{A}$ is $\Phi_{\mathbf{X}}$ and $\mathbf{B}$ is $\Phi_{\mathbf{X}}^{-1}$. This gives us a concentration bound on the estimation error of Wiener coefficients using Lemma D.9.1 and Lemma D.9.2. Notice that $\|\mathbf{A} - \widehat{\mathbf{A}}\| < \epsilon/(3M^5)$ implies that $\|\mathbf{AB} - \widehat{\mathbf{A}}\widehat{\mathbf{B}}\|_2 < \epsilon$. Therefore

$$\mathcal{P}(\|W_i - \widehat{W}_i\| > \epsilon) \leq n^2 \exp\left(-(N-L) \min\left\{\frac{81\epsilon^2}{3200c_1^2 M^{16}(2L+1)^2 n^2 C^2}, \frac{9\epsilon}{80c_1 M^4(2L+1)nC}\right\}\right). \quad (\text{D.7})$$

□

D.10 Primer on Wiener Projection of Time-Series and its Application in Parameter Identification

Here, we provide some results on Wiener projection and Wiener coefficients and shows how the Wiener coefficients are useful in system identification.

Suppose $X_1, \dots, X_n$ be stochastic processes. Consider the projection of X_1 to $X_{\bar{1}}$,

$$\Pi_{X_{\bar{1}}}(x_1) := \arg \min_{\beta \in \mathbb{C}^{n-1}} \|X_1 - \beta X_{\bar{1}}\|^2 = \mathbf{W}_{1,\bar{1}} \mathbf{X}_{\bar{1}}, \quad (\text{D.8})$$

where $\mathbf{W}_{1,\bar{1}}(\omega) := \Phi_{X_1 \mathbf{X}_{\bar{1}}}(\omega) \Phi_{\mathbf{X}_{\bar{1}}}^{-1}(\omega)$, $\mathbf{X}_{\bar{1}} := [X_2, \dots, X_n]^T$. Let the solution be

$$X_{1,2,3,\dots,n}(\omega) = \sum_{j=2}^n X_j(\omega) W_{1j \cdot 23 \dots (j-1)(j+1) \dots n}(\omega), \quad (\text{D.9})$$

where

$$W_{1j \cdot 23 \dots (j-1)(j+1) \dots n} = \Phi_{x_1 x_j} \left(\Phi_{X_{\bar{1}}}^{-1} \right)_{[j, :]}. \quad (\text{D.10})$$

The following result provides a closed form expression for Wiener coefficients in terms of co-factors of $\Phi_{\mathbf{X}}$.

Proposition D.10.1. *Let $W_{1\bar{1}}(\omega) = \begin{bmatrix} W_{12}(\omega) & \dots & W_{1n}(\omega) \end{bmatrix}$, $\Phi_{1\bar{1}}(\omega) = \begin{bmatrix} \Phi_{12}(\omega) & \Phi_{13}(\omega) & \dots & \Phi_{1n}(\omega) \end{bmatrix}$, and*

$$\Phi_{\bar{1}\bar{1}}(\omega) = \begin{bmatrix} \Phi_{22}(\omega) & \Phi_{23}(\omega) & \dots & \Phi_{2n}(\omega) \\ \Phi_{32}(\omega) & \Phi_{33}(\omega) & \dots & \Phi_{3n}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{n2}(\omega) & \Phi_{n3}(\omega) & \dots & \Phi_{nn}(\omega) \end{bmatrix}.$$

Let the co-factor of $\Phi_{ij}(\omega)$ be $C_{ij}(\omega)$. Then, for any $k \neq 1$,

$$W_{1k \cdot 2 \dots (k-1)(k+1) \dots n}(\omega) = -\frac{C_{k1}(\omega)}{C_{11}(\omega)}. \quad (\text{D.11})$$

Proof. See Appendix D.10.3.

Corollary D.10.1. *The coefficient of X_k in the projection of X_i on*

$$\mathbf{X}_{\bar{i}} := \text{span}\{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n\}$$

is

$$W_{ik \cdot 123 \dots (i-1)(i+1) \dots (k-1)(k+1) \dots n} = -\frac{C_{ki}}{C_{ii}}. \quad (\text{D.12})$$

Remark D.10.1. *For $n = 3$,*

$$W_{12 \cdot 3} = -\frac{C_{21}}{C_{11}} = \frac{\Phi_{12}\Phi_{33} - \Phi_{31}\Phi_{32}}{\Phi_{22}\Phi_{33} - \Phi_{23}\Phi_{32}}. \quad (\text{D.13})$$

The following lemma shows the linear/conjugate-linear property of the PSD, which is used later in proving single-door criterion.

Lemma D.10.1 (Linearity of PSD).

1. *Linear on the first term*

$$\Phi_{(\alpha X_1 + \beta X_2)X_3} = \alpha \Phi_{X_1 X_3} + \beta \Phi_{X_2 X_3}.$$

2. *Conjugate-linear on the second term*

$$\Phi_{X_3(\alpha X_1 + \beta X_2)} = \alpha^* \Phi_{X_3 X_1} + \beta^* \Phi_{X_3 X_2}.$$

D.10.1 Parameter Identification

To demonstrate the application of Wiener projections in system identification, consider the following LDIM corresponding to Figure D.6: $X_1 = \beta(\omega)X_2 + \varepsilon_1$, $X_2 = \alpha(\omega)X_3 + \varepsilon_2$, and $X_3 = \varepsilon_3$, where $\varepsilon_1, \varepsilon_2$, and ε_3 are mutually uncorrelated WSS processes and $X_1, \dots, X_3$ are the observed time series.

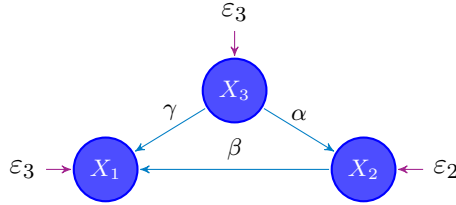


Figure D.6: Example LDIM with $W_{12.3}(\omega) = \beta(\omega), \forall \omega \in \Omega$.

Applying Lemma (D.10.1)(ω is omitted to avoid cluttering),

$$\begin{aligned} \Phi_{23} &= \Phi_{X_2 X_3} = \Phi_{(\alpha X_3 + \varepsilon_2) X_3} = \alpha \Phi_{X_3}, \\ \Phi_{13} &= \Phi_{(\beta X_2 + \gamma X_3 + \varepsilon_1) X_3} = \Phi_{(\beta(\alpha X_3 + \varepsilon_2) + \gamma X_3 + \varepsilon_1) X_3} \\ &= \beta \alpha \Phi_{X_3} + \gamma \Phi_{X_3}, \\ \Phi_{12} &= \Phi_{(\beta X_2 + \gamma X_3 + \varepsilon_1) X_2} = \beta \Phi_{X_2} + \gamma \Phi_{X_3 X_2} \\ &= \beta \Phi_{X_2} + \alpha \gamma \Phi_{X_3}. \end{aligned}$$

Consider the coefficient $W_{12.3}$ of X_2 , in the projection of X_1 to $\text{span}(X_2, X_3)$. From (D.13), $W_{12.3}(\omega) = \beta(\omega)$, for every $\omega \in \Omega$. Thus, we can estimate $\beta(\omega)$ from the partial Wiener coefficient $W_{12.3}(\omega)$, for every $\omega \in \Omega$. However, this doesn't work for all the LDIMs.

Consider the LDIM in Figure D.7, described by $X = a(\omega)Y + \varepsilon_X$, $Y = b(\omega)Z + \varepsilon_Y$, and $Z = \varepsilon_Z$. Similar to the above computation, the Wiener coefficient of projecting Y

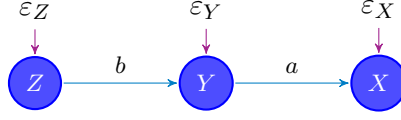


Figure D.7: SEM 1

to X and Z can be written as

$$\begin{aligned} W_{YZ \cdot X}(\omega) &= \frac{\Phi_{YZ}\Phi_{XX} - \Phi_{YX}\Phi_{ZX}}{\Phi_{ZZ}\Phi_{XX} - \Phi_{XZ}\Phi_{ZX}} \\ &= b(\omega) \frac{\Phi_Z(\omega)}{|a(\omega)|^2 \Phi_Y(\omega) + \Phi_Z(\omega)}. \end{aligned}$$

Thus, $W_{YZ \cdot X}(\omega) \neq b(\omega)$ if $a(\omega) \neq 0$. That is, Wiener coefficients do not always give the System parameter directly. However, this is possible if some additional properties are satisfied. For example, projecting Z to Y alone will return the parameter $b(\omega)$ here.

D.10.2 Motivating Example for Back-door criterion

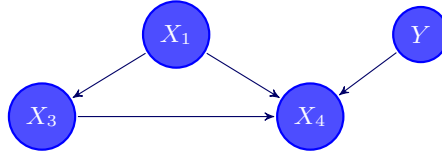


Figure D.8: Graph 1

Example D.10.1. Consider the graph in Fig. D.8. The pdf can be factorized as

$$f^{(\omega)}(x_1(\omega), x_2(\omega), x_3(\omega), x_4(\omega)) = f^{(\omega)}(x_1) f^{(\omega)}(x_2) f^{(\omega)}(x_3 | x_1) f^{(\omega)}(x_4 | x_1 x_2 x_3).$$

Using (D.2), the control distribution on intervention at node 3 is written as

$$f^{(\omega)}(x_1, \dots, x_4 | do(x_3)) = f^{(\omega)}(x_1(\omega)) f^{(\omega)}(x_2(\omega)) f^{(\omega)}(x_4(\omega) | x_1(\omega) x_2(\omega) do(x_3(\omega))).$$

Thus, the causal effect of $X_3 = x_3$ on $X_4 = x_4$ is given by (from Theorem D.8.1)

$$f^{(\omega)}(x_4 \mid do(x_3)) = \int f^{(\omega)}(x_1(\omega)) f^{(\omega)}(x_2(\omega)) f^{(\omega)}(x_4(\omega) \mid x_1(\omega)x_2(\omega) do(x_3(\omega))) d\nu(x_1(\omega) \times x_2(\omega)).$$

D.10.3 Proof of Proposition D.10.1

Let $\eta_{1,2,3,\dots,n} := x_1 - \sum_{k=2}^n W_{1k}x_k$. From projection theorem [17], it can be that $\langle \eta_{1,2,3,\dots,n}, x_j \rangle = 0$, for every $j = 2, \dots, n$. Expanding the inner-product,

$$\begin{aligned} \langle x_1 - \sum_{k=2}^n W_{1k}x_k, x_j \rangle &= 0 \text{ for } j = 2, \dots, n, \\ \int_{\Omega} \Phi_{x_1 x_j}(\omega) - \Phi_{(\sum_{k=2}^n W_{1k}x_k) x_j}(\omega) \frac{d\omega}{2\pi} &= 0, \text{ for } j = 2, \dots, n, \\ \int_{\Omega} \Phi_{x_1 x_j}(\omega) - \sum_{k=2}^n W_{1k}(\omega) \Phi_{x_k x_j}(\omega) \frac{d\omega}{2\pi} &= 0, \text{ for } j = 2, \dots, n. \end{aligned}$$

Solving this for every ω [17],

$$\sum_{k=2}^n W_{1k}(\omega) \Phi_{x_k x_j}(\omega) = \Phi_{x_1 x_j}(\omega) \text{ for } j = 2, \dots, n.$$

Ignoring the index ω , this can be represented as

$$\Phi_{1\bar{1}} = W_{1\bar{1}} \Phi_{\bar{1}\bar{1}}.$$

Assuming $\Phi_{\bar{1}\bar{1}}$ is invertible,

$$W_{1\bar{1}} = \Phi_{1\bar{1}} \Phi_{\bar{1}\bar{1}}^{-1} = \Phi_{1\bar{1}} \frac{1}{\det(\Phi_{\bar{1}\bar{1}})} \text{adj}(\Phi_{\bar{1}\bar{1}}).$$

By definition, cofactor of Φ_{ij} is $C_{ij} = (-1)^{i+j} \det(\Phi_{\bar{i}\bar{j}})$. Then, $C_{11} = \det(\Phi_{\bar{1}\bar{1}})$ and

$$W_{1\bar{1}} = \frac{1}{C_{11}} \Phi_{1\bar{1}} \text{adj}(\Phi_{\bar{1}\bar{1}}).$$

The result follows. □

D.10.4 Proof of Lemma D.7.1

We prove this for $m = 2$. General m follows by induction.

Suppose $y = \alpha_1 y_1 + \alpha_2 y_2$, $\alpha_1, \alpha_2 \in \mathcal{F}$ and let $\tilde{\mathbf{z}} := [x \ \mathbf{z}]$. By Wiener projection, $\Pi_{\tilde{\mathbf{z}}}(y) = \mathbf{W}_{y\tilde{\mathbf{z}}}\tilde{\mathbf{z}}$, where $\mathbf{W}_{y\tilde{\mathbf{z}}} = \Phi_{y\tilde{\mathbf{z}}}\Phi_{\tilde{\mathbf{z}}}^{-1}$. Recall that $\Phi_{y\tilde{\mathbf{z}}} = [\Phi_{yx} \ \Phi_{y\mathbf{z}}]$, and from (D.10), $W_{yx \cdot \mathbf{z}} = b_1^T \Phi_{y\tilde{\mathbf{z}}} (\Phi_{\tilde{\mathbf{z}}}^{-1})$. Moreover, by linearity (Lemma D.10.1), $\Phi_{y\tilde{\mathbf{z}}} = \alpha_1 \Phi_{y_1\tilde{\mathbf{z}}} + \alpha_2 \Phi_{y_2\tilde{\mathbf{z}}}$. Thus,

$$\begin{aligned} \Phi_{y\tilde{\mathbf{z}}}\Phi_{\tilde{\mathbf{z}}}^{-1}\tilde{\mathbf{z}} &= \alpha_1 \Phi_{y_1\tilde{\mathbf{z}}}\Phi_{\tilde{\mathbf{z}}}^{-1}\tilde{\mathbf{z}} + \alpha_2 \Phi_{y_2\tilde{\mathbf{z}}}\Phi_{\tilde{\mathbf{z}}}^{-1}\tilde{\mathbf{z}} \\ &= \alpha_1 [\Phi_{y_1x} \ \Phi_{y_1\mathbf{z}}] \Phi_{\tilde{\mathbf{z}}}^{-1}\tilde{\mathbf{z}} + \alpha_2 [\Phi_{y_2x} \ \Phi_{y_2\mathbf{z}}] \Phi_{\tilde{\mathbf{z}}}^{-1}\tilde{\mathbf{z}}. \end{aligned}$$

The proof follows since $W_{y_i x \cdot \mathbf{z}} = b_1^T \Phi_{y_i\tilde{\mathbf{z}}} (\Phi_{\tilde{\mathbf{z}}}^{-1})$ for $i = 1, 2$.

D.11 A Primer on Stochastic Processes and Conditional Independence Notion in Time Domain

Let $(\Theta, \mathcal{M}, \mathcal{P})$ be a probability space, where Θ is a sample space, $\mathcal{M}$ is a sigma field, and $\mathcal{P}$ is a probability measure. Let I be an index set and let (S, Σ) be a measurable space.

Definition D.11.1. A stochastic process (SP) $\mathbf{X} := \{X(t) : t \in I\}$ is a collection of random variables $X(t)$, indexed by I , where $X(t) : (\Theta, \mathcal{M}) \mapsto (S, \Sigma)$. That is, $\{X(t)\}$ takes values in a common space (S, Σ) . The expected value of $X(t)$ is defined as $\mathbb{E}[X(t)] := \int_{\Theta} X(t) d\mathcal{P}(\theta)$.

Remark D.11.1. When $S = \mathbb{R}^n$ and $\Sigma = \mathcal{B}_{\mathbb{R}^n}$, the Borel sigma field, $X(t)$ is a random vector and can be written as $X(t) = [X_1(t), \dots, X_n(t)]^T$, where $X_i(t) : (\Theta, \mathcal{M}) \mapsto (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

Definition D.11.2. Consider a stochastic process $\mathbf{X} : \Theta \mapsto S^I$, defined on the probability space $(\Theta, \mathcal{F}, \mathcal{P})$. The law or the distribution of $\mathbf{X}$, is defined as the push forward measure, $\mu = \mathcal{P} \circ \mathbf{X}^{-1}$. That is,

$$\mu(B) := \mathcal{P}(\{\theta \in \Theta : \mathbf{X}(\theta) \in B\}), \quad \forall B \in \Sigma^I,$$

where Σ^I denotes the product sigma algebra.

The distribution of any SP $\mathbf{X} = \{X(t) : t \in I\}$ with I non-finite is infinite-dimensional. So, finite-dimensional distributions (FDD) are used to characterize the SPs [123].

Definition D.11.3 (Finite Dimensional Distribution). *Let $\mathbf{X}$ be a stochastic process with distribution μ as above. The FDD for $t_1, \dots, t_k \in I$ (in arbitrary order), where $k \in \mathbf{N}$, is defined as $\mu_{t_1, \dots, t_k} := \mathcal{P} \circ (X(t_1), \dots, X(t_k))^{-1}$.*

Example D.11.1. Suppose $n = 1$ and let $B_k = \bigtimes_{i=1}^k (-\infty, x_{t_i}]$, where $x_{t_1}, \dots, x_{t_k} \in \mathbb{R}$. Then,

$$\mu_{t_1, \dots, t_k}(B_k) = \mathcal{P}(X(t_1) \leq x_{t_1}, \dots, X(t_k) \leq x_{t_k}).$$

Definition D.11.4. (Projection Operator, Coordinate Map): A projection operator π_t is a map from S^I to S such that $\pi_t \mathbf{X} = X(t)$. That is, projection operator provides the marginal distributions, which is obtained by integrating out the rest of the variables from the distribution of $\mathbf{X}$.

For $i = 1, \dots, n$, $\pi_t^{(i)} : S^I \mapsto S$, defined as $\pi_t^{(i)} \mathbf{X} = X_i(t)$, denote the i -th component of $X(t)$. For an index set $J \subset I$, $\mu_J^{(i)} := \mu \circ (\pi_J^{(i)})^{-1}$ is the i -th component of the projection of μ to J .

Definition D.11.5 (Projective Family of Distributions). A family of distributions μ_J , $J \subset I$, is projective or **consistent** when for every $J \subset K \subset I$, we have $\mu_J = \mu_K \circ (\pi_J^K)^{-1}$.

Lemma D.11.1. [123] The FDDs of a stochastic process always form a projective family. That is, for every $J \subset K \subset I$, we have $\mu_J = \mu_K \circ (\pi_J^K)^{-1}$.

Remark D.11.2. The FDDs form the set $\{\mu_J : J \in \text{Fin}(I)\}$, where $\text{Fin}(I)$ denotes the collection of all finite subset of I .

The existence of the unique probability measure is given by the following theorem.

Lemma D.11.2 (Daniel-Kolmogorov extension theorem). [123] Suppose that we are given, for every $J \subset I$, a consistent family of distributions μ_J on (S^J, Σ^J) . Then, there exists a unique probability measure $\mathcal{P}$ on $(\Theta, \mathcal{M})$ such that

$$\mu_J = \mathcal{P} \circ \pi_J^{-1}.$$

Remark D.11.3. *The application of Lemma D.11.2 is that it is sufficient to study the FDDs, which are easier to handle. Two SPs that have same FDDs have the same distribution [123].*

Next, the notion of independence and conditional independence are defined for SPs

D.11.1 FDD based Independence in Stochastic Process

An application of Lemma D.11.2 is that the independence can be defined on FDDs.

Definition D.11.6 (Independence-FDD). *Stochastic processes $\mathbf{X}_1, \dots, \mathbf{X}_m$ are independent if and only if for every $J_1, \dots, J_m \in \text{Fin}(I)$, and $B_{J_i} \in \Sigma^{J_i}$, $i = 1, \dots, m$,*

$$\mathcal{P} \left(\bigcap_{i=1}^m \{ \pi_{J_i} \mathbf{X}_i \in B_{J_i} \} \right) = \prod_{i=1}^m \mathcal{P} (\pi_{J_i} \mathbf{X}_i \in B_{J_i}).$$

Definition D.11.7 (Conditional independence-FDD). *Consider three stochastic processes $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ and the finite projection $\pi_{J_i} \mathbf{X}_i$, $i = 1, 2, 3$. $\mathbf{X}_1$ is said to be conditionally independent of $\mathbf{X}_2$ given $\mathbf{X}_3$ if and only if for every $J_1, J_2 \in \text{Fin}(I)$, there exists $J_3 \in \text{Fin}(I)$ such that*

$$\mathcal{P} \left(\bigcap_{i=1}^2 \{ \pi_{J_i} \mathbf{X}_i \in B_{J_i} \} \mid \pi_{J_3} \mathbf{X}_3 \in B_{J_3} \right) = \prod_{i=1}^2 \mathcal{P} \left(\pi_{J_i} \mathbf{X}_i \in B_{J_i} \mid \pi_{J_3} \mathbf{X}_3 \in B_{J_3} \right),$$

where $B_{J_i} \in \Sigma^{J_i}$, $i = 1, \dots, 3$.

Remark D.11.4. *The similar definition holds for conditional independence of set of stochastic processes also.*

Remark D.11.5. *This definition for DAG factorization with FDD will work with the AR processes and the time-series model with independent noise (TiMINo) with finite lags, provided in [51], which include non-linear models. However, FDD based conditional independence notion will fail for infinite convolution model, since a finite J_3 might not exist. However, the frequency domain independence will work in the infinite linear convolution model.*