Department of Applied Physics

Connectivity inference with asynchronously updated kinetic Ising models

Hong-Li Zeng

Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 117/2014

Connectivity inference with asynchronously updated kinetic Ising models

Hong-Li Zeng

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall M1 of the main building on 15 August 2014 at 12 noon.

Aalto University
School of Science
Department of Applied Physics
Complex Systems and Materials

Supervising professor
Professor Mikko Alava

Thesis advisor
Professor Erik Aurell

Preliminary examiners
Professor Manfred Opper
Professor Federico Ricci-Tersenghi

Opponent
Professor Reimer Kühn, King’s College London, United Kingdom

Aalto University publication series
DOCTORAL DISSERTATIONS 117/2014

ISBN 978-952-60-5802-3
ISBN 978-952-60-5803-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5803-0

Unigrafia Oy
Helsinki 2014

Finland

NORDIC ECOLABEL
441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Hong-Li Zeng

Name of the doctoral dissertation
Connectivity inference with asynchronously updated kinetic Ising model

Publisher School of Science

Unit Department of Applied Physics

Series Aalto University publication series DOCTORAL DISSERTATIONS 117/2014

Field of research Theoretical and Computational Physics

Manuscript submitted 2 May 2014 **Date of the defence** 15 August 2014

Permission to publish granted (date) 27 June 2014 **Language** English

□ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

This thesis focuses on the inference of network connections from statistical physics point of view. The reconstruction methods of the asynchronously updated kinetic Ising model with an asymmetric Sherrington-Kirkpatrick (SK) model is studied theoretically. Both approximate and exact learning rules for the couplings from the generated dynamical data are developed. The approximate formulae are based on naive mean field (nMF) and Thouless-Anderson-Palmer (TAP) equations respectively. The exact learning rules are derived for two cases: one in which both the spin history and the update times are known and one in which only the spin history. One can average over all possible choices of update times to obtain an averaged learning rule that depends only on spin correlations. We studied all the learning rules numerically. Good convergence is observed in accordance with the theoretical expectations.

The developed inference learning rules are applied to two data sets. One is spike trains recorded from 20 retinal ganglion cells and the other is generated by transactions of 100 highly traded stocks on the New York Stock Exchange (NYSE).

For the neuron data set, we compared the inferred asynchronous couplings with the equilibrium ones. The results show that the inferred couplings from these two models are very similar. This implies that real dynamical process of the neuron system satisfies the Gibbs equilibrium conditions and that the final distribution of states is the Gibbs stationary distribution.

For the financial data set, three inference methods are applied to reconstruct the coupling matrices between traded stocks. They are equilibrium, synchronous and asynchronous inference formula respectively. All of them are based on mean-field approximation. Synchronous and asynchronous Ising inference methods give results which are coherent with equilibrium case, but more detailed since the obtained interaction networks are directed.

Keywords network inference, asynchronous update, kinetic Ising model

ISBN (printed) 978-952-60-5802-3 **ISBN (pdf)** 978-952-60-5803-0

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 124 **urn** http://urn.fi/URN:ISBN:978-952-60-5803-0

Preface

The research work presented in this thesis has been performed in the Complex Systems and Materials (CSM) Group of the Department of Applied Physics at the Aalto University School of Science (*known* as the Helsinki University of Technology before 2011).

I am deeply indebted to my tutor as well as thesis instructor Prof. Erik Aurell for your support and help during my research and studies. Your "try it" encouraged me to enjoy the freedom in scientific research. Without your professional instructions and effective solutions to problems, this thesis would never have been finished. I am also extremely grateful to my supervisor Prof. Mikko Alava for your detailed instructions of my daily studies and work. Your "we are almost there" made me believe that our research is always hopeful to show good results.

I was privileged to work together with my co-authors Erik Aurell, Mikko Alava, Hamed Mahmoudi, John Hertz, Yasser Roudi, and Rémi Lemoy. I would like to thank Michael Berry of Princeton University for providing the neural spike trains from salamander retina and Matteo Marsili of ICTP for the financial data from traded stocks on the New York Stock Exchange (NYSE). The pre-examiners Manfred Opper and Federico Ricci-Tersenghi and the opponent Reimer Kuhn deserve thanks for evaluating the thesis.

I thank Arttu Lehtinen, Mikael Mohtaschemi, Shaomeng Qin and Xavier Illa as we were in the same working room before my maternity leave. The coffee flavor and pulla for publication celebration also deserve thanks as they will be sweet memories for ever. I would like to thank Antti Puisto, Amandine Miksic, Sanja Janicevic as my new roommates when I turned back. I thank the other group members for funny discussions during our free time.

I would like to thank the facilities provided by COMP, FICS, COIN as well as library facilities and computer facilities (Triton cluster and the workstation named as Boromir). I would also like to thank the hospitality of Nordita, Niels Bohr Institute and Norwegian University of Science and Technology (NTNU).

Finally, I would like to thank all my friends in both Finland and China for support during these years. I would like to thank my husband Yandong for his personal support and great patience at all times. My parents, brothers have given me their unequivocal support throughout, as always, for which my mere expression of thanks does not suffice. My little boy Jinran also deserves a "thanks" for speeding up the writing process.

I love you all, thank you!

Contents

List of publications

This thesis is based on the following papers, which will be henceforth referred to by the respective Roman numeral.

I Hong-Li Zeng, Erik Aurell, Mikko Alava, and Hamed Mahmoudi. *Network inference using asynchronously updated kinetic Ising model.* Physical Review E **83**, 041135 (2011).
DOI: 10.1103/PhysRevE.83.041135

II Hong-Li Zeng, Mikko Alava, Erik Aurell, John Hertz, and Yasser Roudi. *Maximum Likelihood Reconstruction for Ising Models with Asynchronous Updates.* Physical Review Letters **110**, 210601 (2013).
DOI: 10.1103/PhysRevLett.110.210601

III Hong-Li Zeng, John Hertz, and Yasser Roudi. L_1 *Regularization for Reconstruction of a Non-equilibrium Ising Model.* Accepted by Physica Scripta. arXiv: 1211.3671

IV Hong-Li Zeng, Rémi Lemoy, and Mikko Alava. *Financial interaction networks inferred from traded volumes.* Journal of Statistical Mechanics: Theory and Experiment P07008 (2014).
DOI: 10.1088/1742-5468/2014/07/P07008

Author's contribution

Publication I: Network inference using asynchronously updated kinetic Ising model
The author performed the numerical simulations, prepared all figures, analysis the results and wrote the manuscript.

Publication II: Maximum Likelihood Reconstruction for Ising Models with Asynchronous Updates
The author performed the numerical simulations, prepared the figure for the reconstruction error of algorithms, analysis the results and participated in writing the manuscript.

Publication III: L_1 Regularization for Reconstruction of a Non-equilibrium Ising Model
The author did the numerical simulations, prepared all figures, analysis the results and participated in writing the manuscript.

Publication IV: Financial interaction networks inferred from traded volumes
The author did the numerical simulations, prepared all figures except the network samples, participated in analyzing the results.

List of Symbols and Abbreviations

T_p: temperature
J_{ij}: couplings between i and j
θ_i: external field on spin i
$s_i(t)$: state of spin i at time t
$H_i(t) = \theta_i + \sum_j J_{ij} s_j(t)$: effect field on spin i at time t
$m_i = \langle s_i(t) \rangle_t$: mean value of spin i over time t
$C_{ij}(\tau) = \langle s_i(t) s_j(t) \rangle_t - m_i m_j$: connected correlation between spin i and j with a time delay of τ
L: data length
T: average number of updates for a spin
$C_{jk}^{(i)} = \langle (1 - \tanh^2 H_i(t)) s_j(t) s_k(t) \rangle_t$: Fisher information matrix for spin i
γ: inverse time scale
F_i: total firing number of neuron i
t_i^f: firing time of fth spike for neuron i
Δt: time bins over the time series of the finance data
χ: threshold parameter over volume
$V_i(t; \Delta t)$: volumes traded in time window $[t, t + \Delta t)$
V_i^{av}: average volume of stock i per second
$V_i^{th} = \chi V_i^{av} \Delta t$: threshold volume of stock i in the Δt window.

Chapter 1

Introduction

To deal with the massive data originating from high throughput technologies has been a fascinating challenge in recent years in several research fields. Examples are simultaneous recordings from large number of neurons (an illustration of neural data is shown in Figure 1.1) and stock trading recordings for various companies (the time series of stock trading volumes from Fannie Mae company is presented in Figure 1.2).

Network theory provides a feasible way to extract useful information from such data. The constituting elements of a system can be mapped to the vertices of a network, and one can then focus on finding the functional connections in the network under study. The hope is that such results can help us understand the dynamics of the system. For instance, several kinds of neurons are identified in the biological systems (place cells, retina cells, etc.). If one knows the couplings between them, one may get better understanding about how they interact with each other and how they produce the recorded experimental data[1].

We will be concerned with interaction when the behaviors of the elements in a system can be described by two states. For example, for the time-binned spike histories of neurons, if we assume that the firing rates are low enough that there is at most one neuron spiking during one time bin. The state of a neuron can then be described as firing / not firing in a certain time bin. Similarly, the activity of a stock trades can also be described as active / not active patterns. This representation of data makes itself to formulate a simple statistical physical model: Ising model. What we are interested is finding out the pair-wise interactions between the elements of a system which produces the given data in the framework of the Ising model.

In Chapter 2, both standard and kinetic Ising models are described. For an Ising model composed by N binary spins $\mathrm{s} = \{s_1, ..., s_N\}$, each spin in the system experiences an external field θ_i, and the coupling between pairwise spins i and

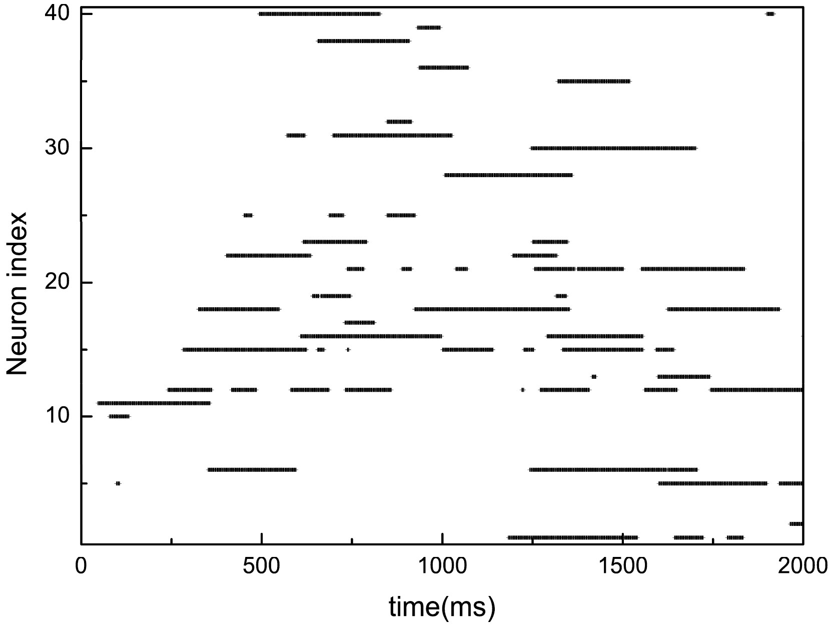

Fig. 1.1: A sample data for spike trains recorded from Salamander retina under visual stimulation by a repeated 26.5-second movie clip. The original data comes in the form of spike times. Here, the size of time bin is $\delta t = 1\text{ms}$. A $+1$ will put when there is a spike in a time bin. We continue setting $+1$ in to the following bins for a period drawn from an exponential distribution with mean 10ms (a typical time scale which corresponds to the inverse of the width of the auto correlation function for neural data) before setting a -1 for the following bins if no other spikes were recorded. Data is provided by Michael Berry of Princeton University through personal communication.

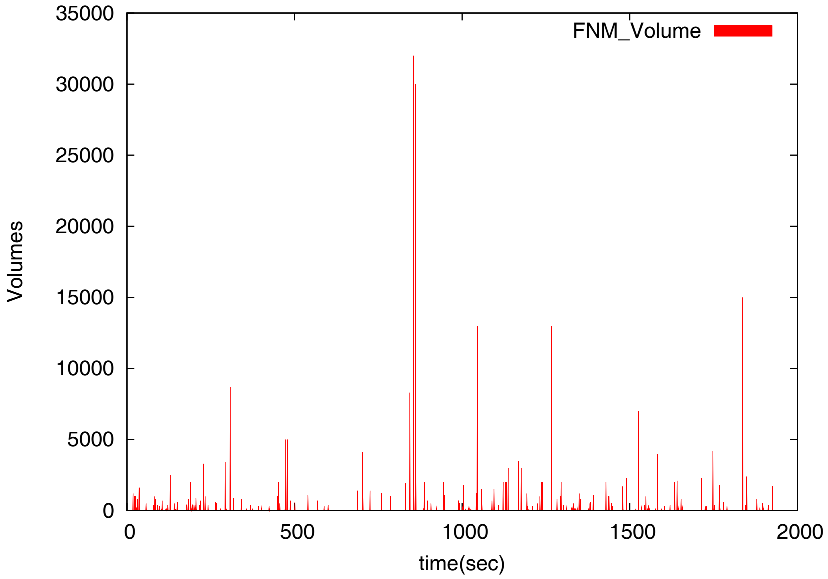

Fig. 1.2: A sample of time series of stock trading volumes for Fannie Mae (FNM), a company for secondary mortgage market, home loan and mortgage. Data is provided by Matteo Marsili of ICTP through personal communication.

j is J_{ij}, the task is to find out the first moment $\langle s_i \rangle$ (magnetization) and second moment $\langle s_i s_j \rangle$ (correlation) etc., which can be measured by experiments. The probability distribution of the Ising system to be in a certain configuration at temperature T_p follows the Gibbs-Boltzmann equilibrium distribution. This fact means that the standard Ising model has a close relation to the maximum entropy model [2]. A closed macroscopic system is said to be in the state of *thermal equilibrium* when the macroscopic physical quantities in any macroscopic subsystem are to a high degree of accuracy equal to their mean values. Physical systems in thermal equilibrium can be described by the Boltzmann distribution, which has the maximum possible entropy if the average energy of the system is given [3]. The relation between them will be further described in detail in this chapter.

The standard Ising model does not include dynamics and is described by the Gibbs-Boltzmann equilibrium distribution only. However, as the given data in, say, biological system, financial system etc., is always a noisy dynamical one, we will investigate a stochastic dynamical one. The Ising model can be given a dynamics following Glauber [4]. We refer the Ising model with dynamics as the kinetic Ising model. The Glauber dynamics is described by a master equation as

$$\frac{dp(\mathrm{s};t)}{dt} = \sum_i \omega_i(-s_i)p(s_1,...,-s_i,...,s_N;t) - \sum_i \omega_i(s_i)p(\mathrm{s};t),$$

where $\omega_i(s_i) = \frac{1}{2}[1 - s_i \tanh\beta(\theta_i + \sum_j J_{ij}s_j)]$ is the flipping rate that the ith spin flips from the value s_i to $-s_i$, while the others unchanged. The first term is a gain function and contributes to the probability distribution from the flipping of the opposite state and the second term is a loss function by flipping out of the same state. This dynamics can also be described in an alternative way as follows: each spin has a probability $\propto \gamma\delta t$ to be updated in the infinitesimal interval $[t, t+\delta t)$, where γ has an inverse time scale. For a spin that is updated, the total "field" is $H_i(t) = \theta_i + \sum_j J_{ij}s_j(t)$, where θ_i is the external field on spin i and J_{ij} the pairwise coupling between spin i and j. The spin i will take its new value for time step $t+\delta t$ with a probability of $p(s_i(t+\delta t)|\{s_j(t)\}) = \frac{1}{2}[1 + s_i(t+\delta t)\tanh H_i(t)]$, where the set of $s_j(t)$ is the nearest neighbors of i. If θ_i is independent of t and the matrix J is symmetric, then this model has a stationary distribution which is the Gibbs-Boltzmann distribution. However, when the θ_i are time dependent or the coupling matrix J is not symmetric, a stationary distribution may still exist, but in general, it may not have a simple description. Such a state is not described by the Gibbs-Boltzmann distribution.

For the numerical simulation of Glauber dynamics, we discretize the time t and the system is updated in equal time intervals. Each time interval is divided into N small time increments. During each time increment, only one spin is randomly selected to update. In this way, not every spin is guaranteed to be selected in a time interval consisting of N spin updates. Some spins will be updated more than once in the time interval. However, every spin is updated once per time interval on average. We refer this update way of spins as asynchronous update. On the other hand, if we update all spins in the system simultaneously at a time interval, then we call this updating approach as synchronous update. In our work, we are mainly focus on asynchronously updated case based on the following reasons: firstly, the asynchronous update will converge to the stationary state which is Gibbs-Boltzmann distribution for a symmetric model, while this may not true for synchronous case. Secondly, the possible applications are naturally asynchronous. For instance, the expression of gene is not a synchronous process, the transcription of DNA, the transport of enzymes, and degradation can vary widely from gene to gene and may take from milliseconds up to a few seconds [5]. More studies in [6, 7] expected that the biological systems do not have a completely synchronous dynamics.

In Chapter 3, different inference methods for network couplings by using asynchronously updated kinetic Ising model are presented. Both approximate methods and learning methods are derived for inferring the connections. The approximate methods are based on mean field equations, in which long time Monte Carlo runs are avoided. This part of work appears in paper I *Network inference using asynchronously updated kinetic Ising model*. One is based on Curie-Weiss mean-field equations which was first applied to magnetic systems and thus we call the inference method as naive Mean field (nMF) approximation inference method. The other is based on an improved equations by Onsager. They were then applied to spin glasses by Thouless, Anderson and Palmer [8], so nowadays, they are called TAP equations in statistical mechanics. The inference method based on this is then called TAP approximation method. TAP approximation adds simple corrections to the nMF approximation, taking into account the effect of the focused spin on itself via its influence on other neighboring spins. The input quantities of these two approximate inference methods are magnetization and correlations, which could be observed from the spin histories. For the inference effectiveness of these two mean-field based approaches, the performance of TAP is somewhat expected to be better than nMF. However, in most application scenarios, network inference using asynchronously updated kinetic Ising models should work well enough using nMF reconstruction and the further step to TAP reconstruction would not be needed.

When one further looks at the kinetic Ising model with asynchronous updates closer, one will find it can be described as a double stochastic process: both the spin history and the update times are stochastic variables. Two cases are considered when inferring the network couplings. One in which one knows both the spin history and the update times and one in which only the spin history. For the first case, one can average over all possible choices of update times to obtain a learning rule that depends only on spin correlations and can also be derived from the equations of motion for the correlations. For the second case, the same rule can be derived within a further decoupling approximation. Thus, the first algorithm needs the full spin and update history and its average version needs the spin correlations at and near $t = 0$. The second case needs the spin history only. The performance of these algorithms is promising in practical terms and agree with the theoretical expectations. In particular, their performances are better than the approximate methods based on mean-field found earlier. This part of work corresponds to paper II *Maximum Likelihood Reconstruction for Ising Models with Asynchronous Updates*.

In chapter 4, we introduce L_1 regularization to the inferred interactions which aims at eliminating the least important couplings in a system. This work refers to paper III *L_1 Regularization for Reconstruction of a non-equilibrium Ising Model*. The idea is to minimize a cost function with respect to couplings J

[9, 10, 11]

$$E = -\mathbf{L}_0 + \lambda \sum_{ij} |J_{ij}|,$$

where the first term is the negative log-likelihood of spin history and the additional term the L_1 norm with λ the strength of L_1 penalty. In order to see how L_1 regularization works in detail, we use a simple gradient descent algorithm which leads to an additional term $-\Lambda \mathrm{sgn}\left(J_{ij}\right)$ in the learning rule for couplings. By using this method, the L_1 term in the cost function is non-differentiable. We deal with this problem by setting $J_{ij} = 0$ by hand. The pruning process of connections is tracked by increasing the λ value from 0 to a large value. Some insight into how this happens was made possible by using an approximation scheme based on a quadratic expansion of the cost function around its minimum. Further approximation by neglecting the off-diagonals in the inverse Fisher matrices behaviors worse than the one with whole elements, which implies the off-diagonals play an important role in regularization. Both exact L_1 and approximate learning rules are performed on a simple and sparse network model, where the connectivity c for each node in the network is much smaller than the system size N. However, we hope the learning methods with/without approximations will be useful in analyzing data from complex systems with sparse property.

The inference methods are firstly tested to reconstruct asymmetric Sherrington-Kirkpatrick (aSK) theoretically, where the interactions are identically and independently Gaussian variables. However, in Chapter 5, two examples are shown for the applications of these inference methods to real data. The first one is: averaged version of the learning rules which originating from maximizing the log likelihood of the history are applied to spike trains from 20 retinal ganglion cells to obtain couplings J_{ij}s and fields θ_is. The raw spike trains are recorded from salamander retina under visual stimulation by a repeated 26.5-second movie clip. We considered spike trains of length of 3180 seconds for 20 neurons with the highest firing rates in the data set. We extracted data as follows: the whole data length are cut to fine time windows with a time bin of size $\delta t = 0.5$ms. A +1 is assigned in a time bin when there is a spike in the bin, we continued putting +1 into the following bins for a period drawn from an exponential distribution with mean 10ms before setting -1 for the subsequent bins if no other spikes were recorded. We apply the asynchronous learning rule mentioned above to the mapped data and obtain the functional connections. The resulting couplings are quite similar to the Gibbs equilibrium ones except that self-connections are missing for the latter inference method. This part of work corresponds to the real application of the average algorithm which is presented in paper II *Maximum Likelihood Reconstruction for Ising Models with Asynchronous Updates*.

The second application example is the nMF approximation methods (equilibrium, synchronous and asynchronous version) applied to financial data on 100 highly traded stocks on the New York Stock Exchange. In order to use the inference methods, we extract data by considering the information of the trade time and volumes. Here, the sliding time window of size Δt is used, with a shifting constant $\Delta s = 1$ second. We use a threshold method to map the data into binaries. For each stock i the sum of volumes $\sum_{t'=t-\Delta t}^{t'=t} V_i(t')$ traded in the time bin of length Δt ending at time t is compared with a given volume threshold $V_{th}^i = \chi V_{av}^i \Delta t$, where V_{av}^i is the average volume of the considered stock over the whole data length, and χ controls the volume threshold. If the sum of volumes in the time bin with length Δt is not less than the threshold, then a +1 will be assigned to that bin; otherwise, a -1 will be assigned. The resulting functional connections given by these two nMF inference methods are coherent with equilibrium ones, while more detailed as the obtained couplings in the inferred network are directed. We find that the volume information of the stocks transaction is enough to obtain the collective behaviors in the stock market which are usually observed through the price information. The details are shown in paper IV *Financial interaction networks inferred from traded volumes*

Chapter 6 will present the further developments and conclusion.

Chapter 2

Ising model

This chapter recalls the standard Ising model which is a mathematic model used to describe the ferromagnetism in statistical mechanics. It is an easy model to define while has wonderfully rich behaviors. Then we move to introducing the kinetic Ising model which is further developed by taking into account Glauber dynamics, where the states of spins in the system evolve with time t and the couplings can be either symmetric or asymmetric. Only the symmetric couplings lead to stationary distributions which are the same with the equilibrium Ising model.

2.1 The standard Ising model

2.1.1 Equilibrium Ising model

The conventional Ising model for ferromagnetism is consisted by N connected spins which are usually located on a lattice, especially the square-lattice. Each of them is connected to its nearest neighbors through an interaction matrix J_{ij}. The couplings J_{ij}s imply the influences from spin j to spin i. A positive interaction J_{ij} is called ferromagnetic, where the neighboring spin i of spin j tends to be with a same orientation while a negative J_{ij} is antiferromagnetic where the neighboring spins tend to have an opposite sign. Zero J_{ij} means there is no interactions between spin j and i. The spins in the system could be subjected to an external field θ_i which indicates how likely spin i tends to be "up" in the absence of the other spins. Positive field on spin i tends to drive it to be "up", while negative one tends to let spin i be "down". Each spin in the system will be in one of a binary states, +1 or -1. Thus there are 2^N possible configurations s : $\{s_i = \pm 1, i = 1, ..., N\}$ in total for an N spin system.

With the pairwise coupling matrix J between spins and external field θ_i on each individual spin, the Hamiltonian function or energy function of an N spin system in state s : $\{s_i = \pm 1, i = 1...N\}$ is defined as follows:

$$E(\mathbf{s}) = -\sum_i \theta_i s_i - \sum_{i<j} J_{ij} s_i s_j, \tag{2.1}$$

where the first term is contributed by the external field and the second sum is over each pairs of neighboring spins where every pair is counted only once.

The probability distribution of the system stays at configuration s at temperature T_p follows Gibbs-Boltzmann equilibrium distribution, which is:

$$p(\mathbf{s}) = \frac{1}{Z(T_p)} \exp\left(\frac{-E(\mathbf{s})}{k_B T_p}\right), \tag{2.2}$$

where k_B is the Boltzmann constant. To simplify, $\frac{1}{k_B T_p}$ is denoted as the inverse temperature β. Z, the partition function, which is a constant to make the above equation (2.2) as a probability measure, is then defined as following:

$$Z(T_p) = \sum_{\mathbf{s}} \exp\left(-\beta E(\mathbf{s})\right). \tag{2.3}$$

where the sum is over all the spin configurations in the system. The partition function Z is an important quantity. Say, if one has it, the negative free energy of the system is obtained by $\log Z$. However, this quantity is difficult to calculate when the system size grows large because the sum is over all spin configurations. For a system with N spins, the calculation is over 2^N terms, which becomes large and intractable when the system size $N > 20$.

The expectation values of the first and second moments under the distribution shown in equation (2.2) $\langle s_i \rangle$ and $\langle s_i s_j \rangle$ are called *magnetizations* and *correlations*, denoted by m_i, χ_{ij} respectively. These two moments are quantities which can be measured in experiments. They are what the direct standard ferromagnetic Ising model be concerned about. However, we focus on the *inverse problem*: we are given the measured magnetizations and correlations, which could be calculated from the data of spin states, and we want to find out the “bias” field θ_i and the couplings J_{ij} which can reproduce the observed values of $\langle s_i \rangle$ and $\langle s_i s_j \rangle$.

In standard Ising model, the temperature T_p plays an important role. However, for the inverse problem, T_p plays the role of setting a common scale of inferred J_{ij} and θ_i. So in some cases, we will take an assumption that $T_p = 1$ which will make the problem more clear.

2.1.2 Relation to maximum entropy principle

We will see the relation between standard Ising model and maximum entropy principle [2] in this section.

Suppose we are given a set of data, which originate from a same probability distribution. Then among all possible probability distributions that can reproduce the data set, the one which best represents the data set has the maximum entropy. This maximum entropy principle has a close connection to the Ising model in statistical physics as shown below:

Assume the probability of a system in the state s is $p(\mathrm{s})$, which is a discrete probability, then the entropy of the system as given by Shannon in 1948 is:

$$S = \sum_{\mathbf{s}} -p(\mathbf{s}) \log \left(p(\mathbf{s})\right). \tag{2.4}$$

With the knowledge about the measured magnetizations m_i and correlations χ_{ij} from the given data, Lagrange multipliers J_{ij}, θ_i and λ can be introduced to build a model for probability distribution $p(\mathrm{s})$. Then the following term need to be maximized under constraints:

$$\begin{aligned}\sum_{\mathbf{s}} -p(\mathbf{s}) \log \left(p(\mathbf{s})\right) + \sum_i \theta_i \left(\sum_{\mathbf{s}} s_i p(\mathbf{s}) - m_i\right)& \\ + \sum_{i,j} J_{ij} \left(\sum_{\mathbf{s}} s_i s_j p(\mathbf{s}) - \chi_{ij}\right)& \\ + \lambda \left(\sum_{\mathbf{s}} p(\mathbf{s}) - 1\right)&\end{aligned} \tag{2.5}$$

Use $p(\mathrm{s}) + \delta p(\mathrm{s})$ instead of $p(\mathrm{s})$ in equation (2.5), and expand the $\log$ function with respective to $\delta p(\mathrm{s})$ to the first order, then keep only the variational terms, we have the following equation:

$$\sum_{\mathbf{s}} \delta p(\mathbf{s}) \left\{ \log p(\mathbf{s}) - 1 + \sum_i \theta_i s_i + \sum_{ij} J_{ij} s_i s_j + \lambda \right\} + \mathcal{O}(\delta^2 p(\mathbf{s})) = 0 \tag{2.6}$$

Thus, the maximization of equation (2.5) corresponds the following term equals to zero.

$$-\log p(\mathbf{s}) - 1 + \sum_i \theta_i s_i + \sum_{ij} J_{ij} s_i s_j + \lambda = 0$$

Which leads to a probability distribution as:

$$p(\mathbf{s}) = \exp\left(-1 + \sum_i \theta_i s_i + \sum_{i,j} J_{ij} s_i s_j + \lambda\right). \tag{2.7}$$

which is the same with equation (2.2) if $\beta = 1$ in that equation. This implies that the Gibbs-Boltzmann equilibrium distribution has the maximum possible entropy given the measured m_i and χ_{ij}. The maximum entropy probability distribution defines an energy function for the system, and energy function relevant problem is an Ising model.

2.2 Kinetic Ising model

There is no dynamics in the original Ising model and thus the states of spins are independent of time. Although the standard Ising model plays an important and fundamental role in ferromagnetism systems, it is possible and natural to introduce dynamics to spins in the system with the aim to generalize it and apply to wider problems. We refer the Ising model as kinetic one if the states of spins are dependent of time and follow a certain dynamics. This indicates that the kinetic Ising model can be obtained by providing a transition rate to the Ising model which allowing the spin system to hop between different configurations. Both of equilibrium and kinetic Ising model have attracted long-term interest which is partly because of the simplicity of the models and the wide applications of them.

2.2.1 Reasons for moving to kinetic spins

For the sake of our further investigations to biological, finance systems, etc., the generalization of the standard Ising model is needed as both of the systems are usually stochastic and dynamic. There are few other reasons to move to kinetic Ising model:

- Gibbs-Boltzmann equilibrium distribution is unlikely to hold in, say, biological or finance systems. The systems are usually driven by the external field which could be time dependent $\theta_i(t)$. And the real given data are always over time. Kinetic models have a bigger relevance to such systems.
- When the problem is posed as inferring the parameters of an equilibrium Gibbs distribution (2.2), the partition function Z in the distribution will be difficult to calculate when the system size grows large. As for a system size of N, there are 2^N microstates in all. Under such case, only approximate methods are available.

- For real applications, the given data length L may be not as long as 2^N when N grows bigger. This implies that only part of the configuration are known for the inference of the model parameters.

- The interaction matrix J shown in distribution (2.2) are not necessarily symmetric outside the original equilibrium system. The influence between elements in a system could be unequal to each other in real systems, i.e., $J_{ij} \neq J_{ji}$. It is not satisfy detailed balance when matrix J is asymmetric. Applying equilibrium approaches to non-equilibrium cases, the inferred couplings may have no obvious relationship to the real ones.

2.2.2 Sherrington-Kirkpatrick (SK) model

Unlike in the original ferromagnetic Ising model, the spins are only coupled with the neighbors symmetrically in a short space, for kinetic version, spins could be coupled in different ways. In the following, we will introduce the Sherrington-Kirkpatrick (SK) model which meet this requirement. The standard SK model [12] is a system of N spins, which can be used to model N spins or agents with binary states ± 1. It is a fully connected graph, i.e., all elements have been coupled to the others in the system. The interactions J_{ij} between each pair of elements can be extended to the asymmetric forms as follow [13]:

$$J_{ij} = J^{s}_{ij} + kJ^{as}_{ij}, \qquad k \geq 0, \tag{2.8}$$

where J^{s}_{ij} and J^{as}_{ij} are symmetric and asymmetric interaction respectively:

$$\begin{aligned} J^{s}_{ij} &= J^{s}_{ji}, \\ J^{as}_{ij} &= -J^{as}_{ji} \end{aligned} \tag{2.9}$$

The parameter k in equation (2.8) measures the asymmetric degree of the interactions J_{ij}. With $k = 0$, J_{ij}'s are a fully symmetric model while $k \neq 0$ means the J_{ij} and J_{ji} are uncorrelated quantities.

Both symmetric and asymmetric couplings are identically and independently Gaussian distributed random variables with means zeros and variances as:

$$\langle {J^{s}_{ij}}^{2} \rangle = \langle J^{as2}_{ij} \rangle = \frac{g^2}{N}\frac{1}{1+k^2}. \tag{2.10}$$

This means the coupling matrix J follows a Gaussian distribution

$$p\left(J_{ij}\right) \propto \exp\left(-\frac{\left(J_{ij}-\mu\right)^2}{2\sigma^2}\right) \tag{2.11}$$

with means $\mu = 0$ and variance $\sigma^2 = g^2/N$.

2.2.3 Glauber dynamics

With the definition of the underlying SK network model, which determines the coupling approach between spins, we introduce dynamics to spins in the system. Then the states of spins will be followed as function of time. Thus the generalized model we consider here is a stochastic one. The behaviors of N spins are stochastic function of time $s_i(t) = \pm 1, i = 1, ..., N$. The state of a spin jumps between 1 or -1 randomly. The hops are influenced by the interactions of spins with an external field which is usually considered as a thermal bath. The transition probability of each spin is determined by the current values of its neighbors and the influence from the heat bath. Thus it is possible that the correlations could appear between spins because of the existence of interactions.

We start from introducing the master equation which describes the derivative of the joint probability distribution $p(s_1, ..., s_N; t)$ of spin states in system at time t as follows::

$$\frac{d}{dt}p(s_1, ..., s_N; t) = \sum_i \omega_i(-s_i)p(s_1, ..., -s_i, ..., s_N; t) - \sum_i \omega_i(s_i)p(\mathbf{s}; t), \tag{2.12}$$

where $\omega_i(s_i)$ is the flipping rate, i.e., the probability for the state of ith neuron changes from s_i to $-s_i$ per unit time while the other spins are momentarily unchanged. Equation (2.12) shows that the configuration $s_1, ..., s_N$ is destroyed by a flip of any spin s_i, but it can also be created by the flip from any configuration with the form $s_1, ... - s_i, ..., s_N$. The flipping rate of spin i is given as follows:

$$\begin{aligned} \omega_i(s_i) &= \frac{\gamma}{1 + \exp\left[2\beta s_i\left(\theta_i + \sum_j J_{ij}s_j(t)\right)\right]} \\ &= \frac{\gamma}{2}\left[1 - s_i \tanh\left(\beta\left(\theta_i + \sum_j J_{ij}s_j(t)\right)\right)\right] \end{aligned} \tag{2.13}$$

As mentioned above that the effective field on spin i is composed of the influence from neighboring spins and the reservoir, which can be written as follows for the sake of convenience:

$$H_i(t) = \sum_j J_{ij}s_j(t) + \theta_i. \tag{2.14}$$

If the couplings are symmetric (i.e., k in equation (2.8) equals 0), then the steady state of the dynamics given by equations (2.12) and (2.13) is:

$$p(s_1, ..., s_N) \propto \exp\left(\beta \sum_i s_i\theta_i + \beta \sum_{ij} s_i s_j J_{ij}\right). \tag{2.15}$$

which is the Gibbs-Boltzmann distribution as shown in equation (2.2). However, when the couplings J are not symmetric anymore, then equations (2.12) and (2.13) still have a steady state (under general condition), but this state does not have a simple description.

The inverse temperature β could be set to 1, cause any effects of changing of it in Glauber dynamics can be realized through changing the coupling strength g which appears in the variance of Js. The effect can also be realized by changing the field strengths θ_i, however, it is set as a fixed and time-independent value in the following work.

2.2.4 Numerical simulations

There is another way to describe the stochastic Glauber dynamics which is widely used for the numerical simulation. The basic idea is that we discretize the time t and the system is updated in equal time intervals. Each time interval is divided into N small time increments. During each time increment, only one spin is randomly selected to update. By this way, not every spin is guaranteed to be selected in a time interval consisting of N spin updates. Some spins will be updated more than once in the time interval. However, every spin is updated once per time interval on average. How to choose spin to update is critical for Glauber dynamics. We refer the update way of spins as asynchronous update if only one spin is randomly selected to update. On the other hand, if we update all spins in the system simultaneously at a time interval, then we call this updating approach as synchronous update.

Asynchronous update

The following two schemes gives detailed descriptions of the asynchronous update for Glauber dynamics casted on the original Ising model, this part of work appears in paper II *Maximum Likelihood Reconstruction for Ising Models with Asynchronous Updates*.

1. Consider a time discretization with a time step increment of δt. At each step, update a random chosen spin i with a probability $\gamma_i \delta t$, where γ_i are constants with dimension of inverse time. In our work, this parameter is assumed to be *a priori*. In order to make the case simple, we assume $\gamma_i = \gamma$ for all spins and we take $\gamma = 1$. For the update of spin i, it will take value $s_i(t + \delta t)$ as follows:

$$s_i(t+\delta t) = \begin{cases} +1 & \text{with probability} \quad 1/\{1+\exp[-2\beta H_i(t)]\} \\ -1 & \text{with probability} \quad 1/\{1+\exp[2\beta H_i(t)]\} \end{cases}$$

The new value of the updated spin i may be equal to the old one; updating does not necessarily mean flipping. Multiple spins can be updated in one time step, but for $\delta t \ll 1$ in most steps at most one spin is updated. When $\gamma\delta t = 1$, the model will be in the synchronous case. Thus, one can interpolate between the synchronous and asynchronous models by varying the parameter γ. In this formulation, the asynchronous model is double stochastic: the dynamics of one set of stochastic variables (the spins) are conditional on the dynamics of the other (the updates).

2. Start from the Glauber master equation (2.12). Then at each time step every spin is flipped with a probability $\gamma\delta t\frac{1}{2}\left[1 - s_i(t)\tanh H_i(t)\right]$. Same as in Scheme 1, multiple spins can flip in a single time step, but this happens with a probability of order $(\delta t)^2$. Thus, with $\delta t \ll 1$, in most time increment, at most only one spin is flipped.

The difference between these two schemes is that in scheme 1, two set of random variables, the update times (which is denoted as τ_i and the spin histories $s_i(t)$, while in scheme 2 contains only the $s_i(t)$. The update times τ_i in scheme 1 can be marginalizing out, which will lead the scheme 1 exactly to scheme 2, even if $\gamma\delta t$ is not small. Nevertheless, knowing the "the history of the system" (i.e., a realization of its stochastic evolution) means something different in the two schemes. In the first we know all the update times, while in the second we know only those at which the updated spins flipped.

Synchronous dynamics

For synchronous updates, all spins will update simultaneously instead of randomly asynchronous. By this way, the model will be easier to be applied to time-binned data, which are always the case for neuron spiking trains. Roudi and Hertz have made several contributions to the descriptions of synchronous updates and have applied to neuron spiking data [14, 15]. With synchronously updated Glauber dynamics, each spin has a probability $\propto \gamma\delta t$ to be updated in a infinitesimal time interval $[t, t + \delta t)$. For the sake of making model simpler, the time increment in the simulation is chosen as $\gamma\delta t = 1$. This means the spin updates are independent Poisson processes.

The time t is discretized also for synchronous case. The initial spin configuration is specified as as 1 or -1 randomly. Then at each discrete time step, the spins are assigned with a new value according to the following distribution (with $\gamma = 1$):

$$\forall i \in \{1, ..., N\} : p(s_i(t+1) = \pm 1) = \frac{1}{2}\left[1 \pm \tanh\left[\beta H_i(t)\right]\right] \tag{2.16}$$

where the instantaneously total field on spin i is:

$$H_i(t) = \theta_i + \sum_j J_{ij} s_j(t-1) \tag{2.17}$$

we take θ_i to be time-independent, however, it can be generalized to be time dependent.

This Markov chain in equation (2.16) can also be described in terms of the microscopic state probability $p_t(\mathbf{s})$, which indicates the probability of the spin system in state s at each time t [16]:

$$p_t(\mathbf{s}) = \sum_{\mathbf{s}'} W_t\left[\mathbf{s};\mathbf{s}'\right] p_{t-1}(\mathbf{s}') \tag{2.18}$$

With a transition probability W_t as:

$$\begin{aligned} W_t\left[\mathbf{s};\mathbf{s}'\right] &= \prod_i \frac{\exp\left(\beta s_i H_i(t-1)\right)}{2\cosh\left(\beta H_i(t-1)\right)} \\ &= \prod_i \frac{1}{2}\left[1 + s_i \tanh \beta H_i(t-1)\right] \end{aligned} \tag{2.19}$$

With a finite β and N, the process will evolves into a stationary distribution which is an equilibrium state if matrix J is symmetric. In the detailed balance case, the corresponding equilibrium probability has a Gibbs-Boltzmann form as shown in equation (2.2) [17].

It is notable that the inverse temperature β controls the stochasticity of the dynamics. $\beta = 0$ corresponds to a fully random case while for $\beta = \infty$, the process is “frozen” as a deterministic case where the configurations of spins will not change with time. As mentioned above, the effect of β can be realized by changing the coupling strength parameter g and the external field θ_i.

Observable from kinetic Ising model

Based on the above two kinds of updating ways applied on Ising model, the spin history $s_i(t)$ is obtained. With which we can naturally define the time dependent means and connected correlations with time delay as follows:

$$\begin{aligned} m_i &= \langle s_i(t) \rangle. \\ C_{ij}(\tau) &= \langle s_i(t+\tau) s_j(t) \rangle - m_i m_j. \end{aligned} \tag{2.20}$$

An example of numerical calculation of the pair-wise correlation functions with different time delays for the kinetic Ising model with asynchronous updates is

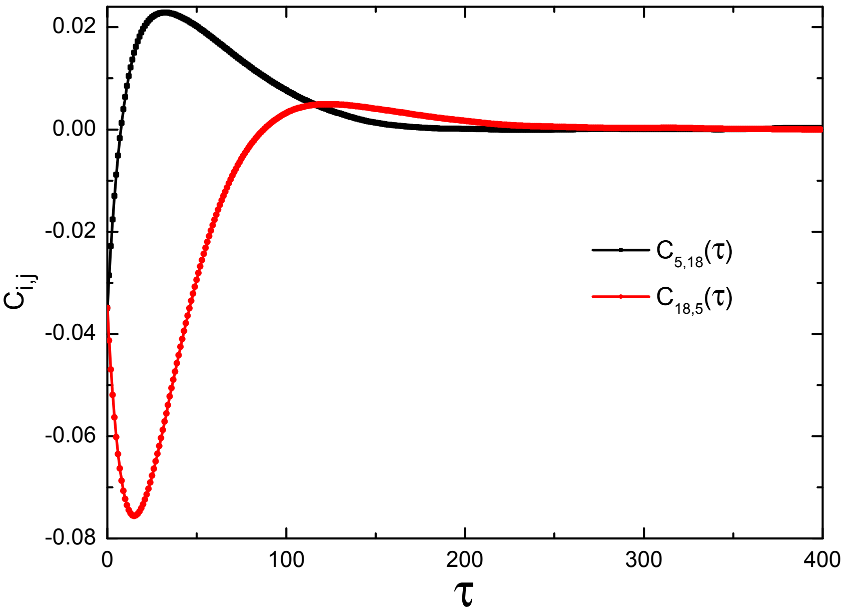

Fig. 2.1: Cross correlations over different time delays for spin pair (5,18) and (18,5). With asymmetric couplings, the correlations between them are not symmetric as $C_{5,18} \neq C_{5,18}$. With longer time delay, both $C_{5,18}$ and $C_{18,5}$ approach to zero.

shown in Figure 2.1. The correlations between spin with index 5 and 18 with a time delay denoted as τ are calculated in this figure. Fully asymmetric SK model is adopted for the simulation. The result shows that $C_{5,18}(\tau) \neq C_{18,5}(\tau)$. The present results are based a data length of $L = N \times 10^8$. The correlations $C_{ij}(\tau)$ approaches to zero with longer enough time delay.

Chapter 3

Inference

This chapter describes the development of inference methods that are used to find out the parameters of pairwise binary models first. Then we mainly focus on the algorithms that are used to reconstruct the couplings and fields of the asynchronously updated kinetic Ising model. The algorithms could be approximate which are based on mean-field equations or exact that start from maximizing the log likelihood of system histories. The performs of both approximate and exact learning rules are presented for different values of parameters. With the theoretical inference approaches, we will show how can they can be used to study the connections from real data.

3.1 Development of pair-wise inference methods

Gibbs Equilibrium Model

In principle, it has been known for some time how to do the inference for the couplings and fields with the given binary data, as been measured like, say, Ising model or similar two-state systems. The implied assumption is that the observed data could be sampled from a same probability distribution. With the measured magnetization m_is and connected correlations C_{ij}s, one could find that the possible distribution which has a maximum entropy will be familiar with the Gibbs-Boltzmann equilibrium distribution, as shown in equation (2.2) [1]. The parameters θ_i and J_{ij} one wants to infer are Lagrange multipliers used in the constrained maximization.

One can adjust the values of parameters J_{ij} and θ_i to maximize the probability distribution (2.2) with the given data. Then the learning rules are called Boltzmann machines in Neuroscience. The method is iterative but (when it converges) exact [18]:

$$\begin{aligned}\delta\theta_i &= \eta\left(\langle s_i\rangle_{Data} - \langle s_i\rangle_{Model}\right),\\ \delta J_{ij} &= \eta\left(\langle s_i s_j\rangle_{Data} - \langle s_i s_j\rangle_{Model}\right).\end{aligned} \tag{3.1}$$

η is a learning rate which need to be chosen small enough to get convergence. To estimate the second averages, one needs to perform Monte Carlo runs with the current values of J_{ij} and θ_i. For large N, these runs can be very time consuming. Besides, the learning will be slow when working for data with long recordings. One likes to estimate the averages from the model as good as that from the original data, then the length of Monte Carlo runs have to be equal to the provided data length. This may take many iterations to obtain stationary coupling and fields. As claimed by Hertz and his collaborators that the learning rules are impractical to try to work with $N > 100$.

Naive Mean Field approximation

To avoid long Monte Carlo runs, one can use mean field methods to get approximate algorithms. From the updating rules of Glauber dynamics, on knows the exact value of m_i which is conditioned on the neighboring s_j through the interaction matrix J is follows:

$$\begin{aligned}m_i &= 1\times p(s_i = 1|s_j) - 1\times p(s_i = -1|s_j)\\ &= \frac{e^{\theta_i + \sum_j J_{ij}s_j} - e^{-\theta_i - \sum_j J_{ij}s_j}}{e^{\theta_i + \sum_j J_{ij}s_j} - e^{-\theta_i - \sum_j J_{ij}s_j}}\\ &= \tanh(\theta_i + \sum_j J_{ij}s_j)\end{aligned} \tag{3.2}$$

The mean field approximation is obtained by replacing s_j inside the $\tanh$ function with its average m_j:

$$m_i = \tanh\left(\theta_i + \sum_j J_{ij}m_j\right). \tag{3.3}$$

which means spin i only takes into account the influences from its nearest neighbors. This is expected to be a good approximation when there are many spins directly connected to spin i and in which the interactions (the Js) are the same.

From equation (3.3), it is easy to write down the formula for field θ_i as:

$$\theta_i = \tanh^{-1} m_i - \sum_j J_{ij}m_j. \tag{3.4}$$

The derivative of θ_i with respect to m_j is the inverse susceptibility matrix which equals to the inverse connected correlation matrix in terms of equilibrium statistical mechanics.

$$\begin{aligned}(C^{-1})_{ij} &= \frac{\partial \theta_i}{\partial m_j} \\ &= \frac{\delta_{ij}}{1-m_i^2} - J_{ij}.\end{aligned} \tag{3.5}$$

Thus, if one knows the correlation matrix, one has an inference algorithm based on naive mean field approximation as ($i \neq j$):

$$J_{ij} = -(C^{-1})_{ij} \tag{3.6}$$

and θ_i could be calculated by equation (3.4).

TAP approximation

When one takes into account the Onsager term, which takes away the contribution to the neighbor magnetization m_j from the central unit s_i when estimating the field acting on s_i, the TAP equation can be written as [8, 16]:

$$m_i = \tanh\left[\theta_i + \sum_j J_{ij} m_j - m_i \sum_j J_{ij}^2 (1-m_j^2)\right] \tag{3.7}$$

TAP equations should be used in spin glasses where the couplings J_{ij} are random and with a very small mean as the Onsager term has the same order with the naive mean field [8]. In [19], Plefka pointed out that equation (3.3) and (3.7) are the first two terms in the sequence with better approximation which can be derived systematically.

With equation (3.7), it is easy to write down the equation for the fields in the following form:

$$\theta_i = \tanh^{-1} m_i - \sum_j J_{ij} m_j + m_i \sum_j J_{ij}^2 (1-m_j^2) \tag{3.8}$$

Then, perform similar derivative with respective to m_j as in naive mean field case, the inverse correlation matrix can be obtained as:

$$\begin{aligned}(C^{-1})_{ij} &= \frac{\partial \theta_i}{\partial m_j} \\ &= \frac{\delta_{ij}}{1-m_i^2} - J_{ij} - 2J_{ij}^2 m_i m_j.\end{aligned} \tag{3.9}$$

This means that if the correlation matrix is known, the couplings J_{ij} can be solved from above equation (3.9) with $i \neq j$ [15, 16, 20, 21]:

$$(C^{-1})_{ij} = -J_{ij} - 2J_{ij}^2 m_i m_j. \tag{3.10}$$

One may note that for TAP case, there are $N(N-1)/2$ independent quardratic equations to be solved for coupling matrix J. This makes the TAP case more complicated compared with the inference method with the naive mean field approximation.

For the above equilibrium inference methods, the self-couplings are equal to zero by convention.

3.2 Learning from general to details (kinetic models)

This section will cover the derivation of learning algorithms for parameters with the asynchronous updated kinetic Ising model. We first derive the inference equations for couplings and fields based on mean field equation and then move to derive them by maximizing the log likelihood of the spin history.

3.2.1 Mean field approximation

This part of work refers to the derivation of inference equations in paper I *Network inference using asynchronously updated kinetic Ising model.*

Similarly to the equilibrium case, we derive the inference formula with both naive mean field equation and TAP equation in the frame of kinetic Ising model with asynchronous updates.

With Glauber dynamics, the state of spin i is time dependent $s_i(t)$. With this, we can naturally define the time-dependent means and correlations as shown in equation (2.20). With $H_i(t) = \theta_i + \sum_j J_{ij} s_j(t)$, and from the master equation (2.12) and the flipping rate (2.13), we have the equations of motion for means and correlations as:

$$\begin{aligned} \frac{dm_i}{dt} &= -m_i + \langle \tanh\left[H_i(t)\right] \rangle. \\ \frac{d\langle s_i(t) s_j(t_0) \rangle}{dt} &= -\langle s_i(t) s_j(t_0) \rangle + \langle \tanh\left[H_i(t) s_j(t_0)\right] \rangle. \end{aligned} \tag{3.11}$$

For the second equation of equation (3.11), the term on the left-hand side and the first term on the right-hand side can be solved based on the empirical data which could be produced by the Glauber dynamics. However, the calculation of the second term involves all kinds of higher-order correlations and is therefore not easily expressed only in terms of means and pairwise correlations. In order to solve the second equation in (3.11), approximations of tanh function are obviously needed. We use the nMF and TAP approximation to deal with this problem.

Let

$$b_i = \theta_i + \sum_j J_{ij} m_j \tag{3.12}$$

and rewrite $H_i(t) = \theta_i + \sum_j J_{ij} s_j(t)$ as

$$H_i = b_i + \sum_j J_{ij}(s_j - m_j) \equiv b_i + \sum_j J_{ij} \delta s_j(t) \tag{3.13}$$

with

$$\delta s_i(t) = s_i(t) - m_i \tag{3.14}$$

Thus the terms in $\tanh$ function in the second equation of (3.11) are as follows:

$$H_i(t) s_j(t_0) = (b_i + \sum_k J_{ik} \delta s_k(t))(m_j + \delta s_j(t_0)) \tag{3.15}$$

Then expand the $\tanh$ function with respect to b_i:

$$\langle s_i(t) s_j(t_0) \rangle + \frac{d\langle s_i(t) s_j(t_0)\rangle}{dt} = m_i m_j + (1 - m_i^2)\left(\sum_j J_{ik} \langle \delta s_k(t) \delta s_j(t_0) \rangle\right) \tag{3.16}$$

with

$$C_{ij}(t, t_0) = \langle \delta s_i(t) \delta s_j(t_0) \rangle = \langle s_i(t) s_j(t_0) \rangle - m_i m_j \tag{3.17}$$

and defining

$$\begin{aligned} D_{ij}(t, t_0) &= C_{ij}(t, t_0) + \frac{dC_{ij}(t, t_0)}{dt} \\ &= (1 - m_i^2) \sum_k J_{ik} C_{kj}(t, t_0) \end{aligned} \tag{3.18}$$

In the limit $t \to t_0$, we have the equation that need to infer the network connections:

$$J = A^{-1} D C^{-1}, \tag{3.19}$$

where $A_{ij} = \delta_{ij}(1 - m_i^2)$.

Equation (3.19) is a linear matrix equation with respect to J_{ij}. We can solve it for J_{ij} directly. With the inferred J_{ij}, the fields θ_i can be solved by equation (3.3).

Next, we derive the inference formula with TAP equation.

Similarly to the derivation of the inference formula with nMF approximation, we start from the $H_i(t)$ term in the $\tanh$ function as shown in the equation of motion of correlations (3.11). It can be rewritten as follows:

$$H_i(t) = b_i \mp m_i \sum_{k \neq i} J_{ik}^2 (1 - m_k^2) + \sum_k J_{ik} \delta s_k(t). \tag{3.20}$$

and the TAP equation

$$m_i = \tanh \left[b_i - m_i \sum_k J_{ik}^2 (1 - m_k^2) \right],$$

we expand the $\tanh$ function in the second equation of (3.11) with respect to

$$b_i - m_i \sum_{k \neq i} J_{ik}^2 (1 - m_k^2)$$

to the third order and keep the terms only up to the third order of J. Then the corresponding TAP inference formula for coupling J_{ij} is obtained, which is formally the same as in the nMF approximation.

$$J = A^{-1} D C^{-1}. \tag{3.21}$$

However, matrix **A** in TAP formula is different

$$A_{ij} = \delta_{ij} (1 - m_i^2) \left[1 - (1 - m_i^2) \sum_j J_{ij}^2 (1 - m_j^2) \right]. \tag{3.22}$$

Equation (3.21) is a function of the couplings **J**, and therefore it is a nonlinear equation for matrix **J**.

We try to solve equation (3.21) for **J** though two approaches. One way is to solve it iteratively. We start from reasonable initial values J_{ij}^0 and insert them in the right hand side of the formula. The resulting J_{ij}^1 is the solution after one iteration. This can be again replaced in the right hand side to get the second iteration results and etcetera ...

$$J^{t+1} = A(J^t)^{-1} D C^{-1} \tag{3.23}$$

An alternative way is solving it directly, as done for the synchronous update model in [15], casting the inference formula to a set of cubic equations. For equation (3.22), we denote

$$F_i = (1 - m_i^2) \sum_j J_{ij}^2 (1 - m_j^2) \tag{3.24}$$

and plug it into equation (3.21), and then obtain the following equation for J_{ij}:

$$J_{ij}^{\text{TAP}} = \frac{V_{ij}}{(1 - m_i^2)(1 - F_i)} \tag{3.25}$$

where $V_{ij} = [DC^{-1}]_{ij}$. Inserting equation (3.25) into equation (3.24), we obtain the cubic equation for F_i as:

$$F_i(1 - F_i)^2 - \frac{\sum_j V_{ij}^2 (1 - m_j^2)}{1 - m_i^2} = 0. \tag{3.26}$$

With the obtained physical solution for F_i, we get the reconstructed couplings J^{TAP} as

$$J_{ij}^{\text{TAP}} = \frac{J_{ij}^{\text{nMF}}}{1 - F_i}. \tag{3.27}$$

It is worth mentioning that for the cubic equation (3.25), we have three solutions with possible imaginary parts. Here we study the real roots of the cubic equation and ignore those solutions with imaginary parts. When three solutions are all real ones, we take the smallest one.

3.2.2 Maximum likelihood reconstruction

This part corresponds to the derivation of inference formula derivation part in paper II *Maximum Likelihood Reconstruction for Ising Models with Asynchronous Updates*.

For kinetic Ising model with asynchronous updates, the updating process can be considered as a double stochastic process: the spin histories as well as the update times of spins. Doubly stochastic processes are in fact abundant in real life. An example is a securities market [22, 23] where traders place limit orders: conditional offers to buy securities if their market price falls below a threshold, or to sell if the market price rises above it. If offers are made, other traders may respond or not; if they do, transactions take place. Whether or not limit offers are placed define a first set of stochastic variables depending on which transactions may or may not occur, defining a second set.

We recall the Glauber dynamics as presented in chapter 2. For an N spins system $s_i = \pm 1, i = 1, ..., N$ with coupling matrix J_{ij} and field parameter θ_i. The dynamics can be described in either of the following two ways.

(1) With a time discretization of size δt, update spin i with probability $\gamma\delta t$. A new value $s_i(t+\delta t)$ with probability $\exp(s_i(t+\delta t)H_i(t))/2\cosh H_i(t)$. The new value, $s_i(t+\delta t)$ may be equal to the old one as updating does not necessarily mean flipping. Multiple spins can be updated in one time step, but for $\delta t \ll 1$ (the limit we consider) in most steps at most one spin is updated. In this formulation, the model is doubly stochastic: the dynamics of one set of stochastic variables (the spins) are conditional on the dynamics of the other (the updates). Here the temperature in this model equal to 1, because it can be absorbed into the definitions of the fields and couplings.

(2) Start from the Glauber master equation (2.12). Then at every step every spin is flipped with a probability $\gamma\delta t\,(1-s_i(t)\tanh H_i(t))\,/2$. As in scheme (1), multiple spins can flip in a single time step, but this happens with probability of order $(\delta t)^2$. Thus, $\delta t \ll 1$, in most time intervals at most one spin is flipped.

The difference between the schemes is that in scheme (1) we have two sets of random variables, the update times (which we denote by $\{\tau_i\}$) and the spin histories $\{s_i(t)\}$, while scheme (2) contains only the $\{s_i(t)\}$. One can easily show that marginalizing out the $\{\tau_i\}$ in scheme (1) leads exactly to scheme (2), even if $\gamma\delta t$ is not small. Thus, all averages over histories involving spins only (i.e., not involving the update times) will be the same in the two schemes. Nevertheless, knowing "the history of the system" (i.e., a realization of its stochastic evolution) means something different in the two schemes. In the first we know all the update times, while in the second we only know those at which the updated spins flipped.

Consider scheme (1) above. Suppose we are given a history of the system, i.e., the data $\mathrm{s} \equiv \{s_i(t)\}$ and $\tau \equiv \{\tau_i\}$, of length L steps, and we are asked to reconstruct the couplings and fields. We do this by maximizing the likelihood $P(s,\tau) = P(s|\tau)p(\tau)$ over these parameters. For each spin i, the τ_i are a (discretized) Poisson process, i.e., every t has probability $\gamma\delta t$ of being a member of the set τ. Thus the probability of the update history, $p(\tau)$, is independent of the model parameters, and we can take as objective function $\log P(\mathrm{s}|\tau)$, i.e.,

$$\mathcal{L}_1 = \sum_i \sum_{\tau_i} \left[s_i(\tau_i+\delta t)H_i(\tau_i) - \log 2\cosh H_i(\tau_i) \right].$$

This is just like the synchronous-update case except that the sum over times is only over the update times. It leads to a learning rule

$$\delta J_{ij} \propto \frac{\partial \mathcal{L}_1}{\partial J_{ij}} = \sum_{\tau_i} [s_i(\tau_i+\delta t) - \tanh(H_i(\tau_i))]s_j(\tau_i). \tag{3.28}$$

Defining $J_{i0} = \theta_i$, $s_0(t) = 1$, this equation also includes the learning rule for θ_i. We call this algorithm "spin- and update-history-based", or "**SUH**".

In scheme (2), we know only the spin history, not the update times. Since this scheme is equivalent to the first one with the τ_i marginalized out, we treat it by maximizing $P(s) = \sum_\tau P(S|\tau)p(\tau)$ [24], leading to

$$\mathcal{L}_2 = \sum_{i,t} \log \left[(1 - \gamma\delta t)\delta_{s_i(t+\delta t), s_i(t)} + \gamma\delta t \frac{e^{s_i(t+\delta t)H_i(t)}}{2\cosh H_i(t)} \right].$$

as objective function. Separating terms with and without spin flips, the resulting learning rules will be

$$\begin{aligned} \delta J_{ij} &\propto \frac{\partial \mathcal{L}_2}{\partial J_{ij}} \\ &= \sum_{\textit{flips}} [s_i(t+\delta t) - \tanh(H_i(t))] s_j(t) + \frac{\gamma\delta t}{2} \sum_{\textit{no flips}} q_i(t) s_i(t+\delta t) s_j(t), \end{aligned} \tag{3.29}$$

where $q_i(t) \equiv [1 - \tanh^2(H_i(t))]$, and it includes the rule for the θ_i with the convention $J_{i0} = \theta_i$, $s_0(t) = 1$. We call this the "spin-history-only" ("**SHO**") algorithm.

Reconstruction errors for both algorithms can be calculated by analyzing the Fisher information matrices. For SHO the Fisher matrix elements read

$$\begin{aligned} -\frac{\partial^2 \mathcal{L}_2}{\partial J_{ij} \partial J_{kl}} &= \delta_{ik} \sum_{flips} q_i(t) s_j(t) s_l(t) \\ &\quad + 2\delta_{ik}\gamma\delta t \sum_{\mathrm{no}\, flips} q_i(t) s_i(t+\delta t) \tanh(H_i(t)) s_j(t) s_l(t). \end{aligned} \tag{3.30}$$

In the weak coupling limit, this matrix has nonzero elements only for $j = l$, and the mean value of these non-zero elements yields the inverse of the mean square reconstruction error (MSE). Without external fields, the second term in equation (3.30) vanishes; thus, for a data set with length L, the MSE in this case is $2/(L\delta t\gamma)$, noting that the probability that a time step is a flip is $\gamma\delta t/2$. For SUH the calculation is analogous and for $\theta_i = 0$ and weak couplings, the MSE will be $(L\delta t\gamma)^{-1}$, i.e., a factor of two smaller than for SHO.

Next an algorithm by averaging the one for SUH in equation (3.28) over all updating histories. Denoting $C_{ij}(t) = \langle s_i(t + t_0) s_j(t_0) \rangle$, the time derivative of it can be written as

$$\dot{C}_{ij}(t) = \lim_{\delta t \to 0} \frac{\langle s_i(t+\delta t) s_j(t_0) \rangle - \langle s_i(t) s_j(t_0) \rangle}{\delta t}, \tag{3.31}$$

where $\langle ... \rangle$ means an average over all realizations of the stochastic dynamics. Separating time steps into those at which an update occurred and those at which no update occurred yields

$$\dot{C}_{ij}(t) = \lim_{\delta t \to 0} \left\{ \gamma \delta t \frac{\langle s_i(\tau_i + \delta t) s_j(t_0) \rangle_{\tau_i} - \langle s_i(\tau_i) s_j(t_0) \rangle_{\tau_i}}{\delta t} \right\}. \tag{3.32}$$

There is no contribution from steps with no flip because then $s_i(t + \delta t) = s_i(t)$ and the numerator would be zero. Thus we have expressed the average over all realizations of the first term in equation (3.28) in terms of spin correlation functions and their time derivatives:

$$\langle s_i(\tau_i + \delta t) s_j(\tau_i) \rangle_{\tau_i} = \frac{1}{\gamma} \dot{C}_{ij}(0) + C_{ij}(0). \tag{3.33}$$

In averaging the second term in equation (3.28), the average over $\{\tau_i\}$ can be replaced by an average over all times, since the quantity $\tanh H_i(t) s_j(t)$ is insensitive to whether an update is being made. Thus, averaging equation (3.28) over all possible histories yields

$$\delta J_{ij} \propto \gamma^{-1} \dot{C}_{ij}(0) + C_{ij}(0) - \langle \tanh(H_i(t)) s_j(t) \rangle. \tag{3.34}$$

We will refer to the update rule given by equation (3.34) as the averaged-SUH rule, or "**AVE**" . This rule has the same structure as the one for the synchronous-update model [15], with $\langle s_i(t+1) s_j(t) \rangle$ replaced by $C(0) + \gamma^{-1} \dot{C}(0)$.

AVE requires knowing the equal-time correlations, their derivatives at $t = 0$, and $\langle \tanh(H_i(t)) s_j(t) \rangle$. This latter quantity depends on the model parameters (through $H_i(t)$), so, in practice, estimating it at each learning step requires knowing the entire spin history, the same data as SHO learning needs.

For **SHO** learning, when one performs the average over spin flip times, an algorithm like equation (3.34) could be obtained. Denote the local fields at time t generated by the true model (the one that generated the data) by $\tilde{H}_i(t)$, and, as before, the local field calculated using the inferred parameters as $H_i(t)$. At each time step t, then, the probability of flipping spin i is $\gamma \delta t [1 - s(t) \tanh \tilde{H}_i(t)]/2$. When one allots a weight $\gamma \delta t [1 - s(t) \tanh \tilde{H}_i(t)]/2$ for the first term in equation (3.29) and the second a weight $1 - \gamma \delta t [1 - s(t) \tanh \tilde{H}_i(t)]/2 \approx 1$ getting

$$\begin{aligned} \delta J_{ij} &\propto \left\langle \frac{\partial \mathcal{L}_1}{\partial J_{ij}} \right\rangle_0 \\ &= \frac{\gamma}{2T} \int dt [\tanh \tilde{H}_i(t) - \tanh H_i(t)] \times [1 + s_i(t) \tanh H_i(t)] s_j(t). \end{aligned} \tag{3.35}$$

The learning thus converges when the discrepancy $\tanh(H(t)) - \tanh(\tilde{H}(t))$ is zero. Noting also that the arguments above leading to equation (3.33) yields $\langle \tanh \tilde{H}(t) s_j(t) \rangle_t = \gamma^{-1}\dot{C}(0) + C(0)$, we write equation (3.35) as

$$\begin{aligned}\delta J_{ij} \propto\, & \gamma^{-1}\dot{C}_{ij}(0) + C_{ij}(0) - \langle \tanh H_i(t) s_j(t) \rangle_t \\ & + \langle [\tanh \tilde{H}_i(t) - \tanh H_i(t)] s_i(t) \tanh H_i(t) s_j(t) \rangle_t\end{aligned} \tag{3.36}$$

The first line is identical to equation (3.34). We can obtain a learning rule heuristically by an *ad hoc* factorization of the average in the second line as

$$\begin{aligned}& \langle [\tanh \tilde{H}_i(t) - \tanh H_i(t)] s_i(t) \tanh H_i(t) s_j(t) \rangle_t \\ \approx\, & \langle \tanh \tilde{H}_i(t) - \tanh H_i(t) s_j(t) \rangle_t \langle s_i(t) \tanh H_i(t) \rangle_t\end{aligned}$$

yielding

$$\begin{aligned}\delta J_{ij} \quad \propto \quad & [\gamma^{-1}\dot{C}_{ij}(0) + C_{ij}(0) - \langle \tanh H_i(t) s_j(t) \rangle_t] \\ \times \quad & \langle [1 + s_i(t) \tanh H_i(t)] \rangle_t.\end{aligned} \tag{3.37}$$

This just amounts to varying the learning rate in equation (3.34) proportional to the time-averaged probability of not flipping according to the model. Thus we arrive by a different route at the **AVE** rule, equation (3.34).

In this subsection, we start from two likelihood functions for the data producing by Glauber dynamics, one in which update times are known, the other only the spin history, we derive two different learning rules. These learning rules have different precisions for inferring the couplings, and that they have a nontrivial relation to each other: averaging over possible update times, they both lead to a third one, but with different learning rates. Surprisingly, this third learning rule can also be derived from the forward equations of motion for the correlations of the asynchronous Ising model [4] and without appealing to a likelihood function. This relates two previously unrelated approaches of learning the couplings.

3.3 Performance of algorithms

In this section, the performance of these five algorithms for reconstructing the asynchronously updated kinetic Ising model are present. We compared the performance of the algorithms **SUH**, **SHO**, and **AVE** to each other and to the naive mean-field (**nMF**) and Thouless-Anderson-Palmer (**TAP**) approximations to **AVE** investigated in [25] for fully asymmetric SK models [12]. The couplings are zero-mean i.i.d. normal variables with variance g^2/N (J_{ij} is independent of J_{ji}). We study these for different values of g and θ, the system size N and the data length L.

As a performance measure, we use the mean square error (MSE) on the J_{ij}. The MSE is used to measure the difference between the reconstructed network structure and the original true ones, which is

$$\mathrm{MSE} = \frac{\sum_{i \neq j} (J_{ij}^{re} - J_{ij}^{true})^2}{N(N-1)}. \tag{3.38}$$

where J_{ij}^{true} represents the true network couplings and J_{ij}^{re} for the reconstructed ones.

Figure 3.1 shows the performance of the algorithms. As anticipated above, the error for **SUH** is half of that for **SHO** learning; see figure 3.1(a). The same panel also shows that **AVE** and **SHO** appear to perform equally well for large enough L. In retrospect, this is not surprising, since both algorithms effectively use the same data (the spin history). For small L, the averaging that yields **AVE** from **SHO** may be prone to fluctuations yielding the two learning rules behaving differently. Figure 3.1(b) shows that the MSE for the exact algorithms is insensitive to N, while the approximate algorithms improve as N becomes larger (note however the opposite trend in figure 3.1(a)); in these calculations, the average numbers of updates and flips per spin were kept constant, taking $L = 5 \times 10^5 N$.) Figure 3.1(c) shows that the performance of the three exact algorithms is also not sensitive at all to θ, while nMF and TAP work noticeably less well with a non-zero θ. Finally, the effects of (inverse)g are depicted in figure 3.1(d). For fixed L, all the algorithms do worse at strong couplings (large g). The nMF and TAP do so in a much more clear fashion at smaller g, growing approximately exponentially with g for g greater than ≈ 0.2. In the weak-coupling limit, all algorithms perform roughly similarly, except that **SUH** enjoys its factor-2 advantage (conferred by knowledge of the update times), as already seen in figure 3.1(a).

The approximate learning rules (nMF and TAP) are much faster in reconstructing the couplings while with worse accuracy compared with that of the exact iterative learning rules (AVE, SHO and SUH).

3.4 Learning from data

The above mentioned algorithms, for both the ones based on mean field approximations and the ones relying on likelihoods, are tested on synthetic data produced by Glauber dynamics with asymmetric SK model. However, the purpose for us to study the learning algorithms is try to apply them to the recorded data by experiments.

The essential part of the applications for these algorithms contains the mapping of data. As the algorithms are derived from an Ising binary model, we should

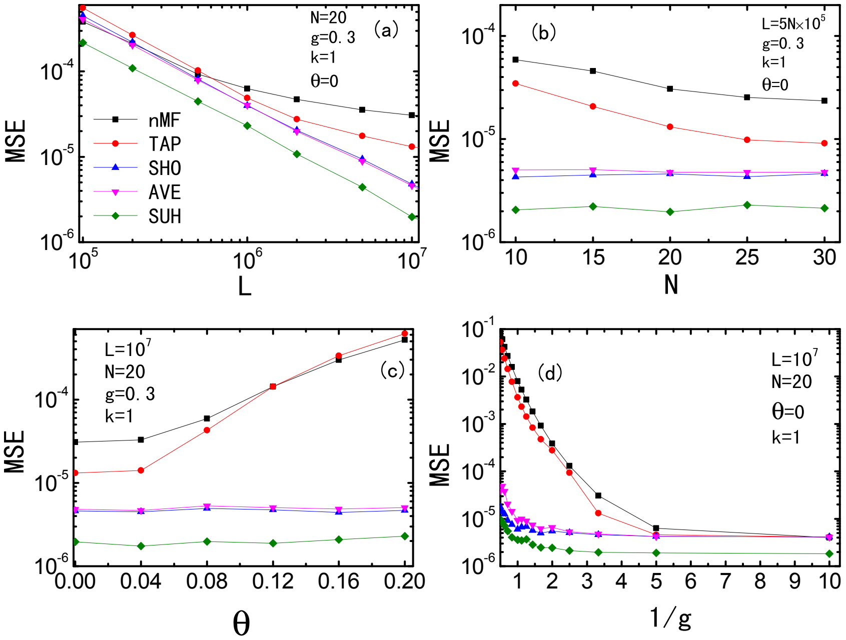

Fig. 3.1: Mean square error (MSE) versus (a) data length L, (b) system size N, (c) external field θ and (d) temperature $1/g$. Black squares show nMF, red circles, TAP, blue up triangle SHO, pink down triangle AVE and green diamond SUH, respectively. The parameters are $g = 0.3$, $N = 20$, $\theta = 0$, L=10^7 except when varied in a panel.

start from the notation of the data. Usually, the given data is dependent of time. Thus, we can bin the time of the data with an assumption that the active rates are low enough that there is at most only one event per time bin. By this way, time is discretized in the units of the bin size. Then, the state of element i at time t is denoted as $s_i(t)$, with $s_i(t) = +1$ if it is active and $s_i(t) = -1$ if it does not.

This representation of data leads itself to be described in terms of Ising model. And the algorithms we derived in the above sections could be applied to infer the functional connections between elements. We first apply the **AVE** algorithm to spike trains from 20 retinal ganglion cells. The inferred couplings are comparable with that obtained from the Gibbs equilibrium model. The second example is applying non-equilibrium nMF inference formula to New York Stock Exchange (NYSE) data based on the information of transaction time and volumes to get some sight of the connections between different stocks. The details about the applications to real data will be presented in the following chapter 5.

Chapter 4

L_1 Regularization

4.1 Introduction

This chapter refers to paper III *L_1 Regularization for Reconstruction of a Non-equilibrium Ising Model.*

The part of work is an extension of the derived learning rules for asynchronous kinetic Ising model. Be different from the recent work in which the L_1 regularizer has been taken into account [26, 27] to infer the couplings more efficiently, we focused on its application of non-equilibrium models as the equilibrium ones are not the ideal choice for network reconstruction in many practical applications. Several recent studies have moved to kinetic models, using exact and approximate learnings for reconstructing the couplings in non-equilibrium models [15, 25, 28, 29]. This part of work has not yet exploited the potential power of L_1 regularization in inferring the connections. Thus, we introduced L_1 regularization to infer the couplings in a sparse asymmetric, asynchronously updated kinetic Ising model as it tends to produce sparse models.

The idea of L_1 regularization is simply to minimize a cost function which is composed of negative log likelihood and L_1 norm. However, the L_1 norm is not differentiable with respective to the couplings. When they approach to zero, we need to deal with them by hand or taking approximations to avoid the problem. In our work, we do the calculations in several ways: (1) by iterative minimization of the cost function, which is referred as *full L_1 regularization*. (2) approximate scheme based on a quadratic expansion of the cost function around its minimum. (3) approximate method where the learning rule depends only on the diagonals of Fisher information matrix. (4) approximate method where the learning rule depends only on the initial slope of the inferred parameters without regularization.

We also tried to studying the consistency of L$_1$ regularization in logistic and linear regression problems [26, 30, 31]. Asymptotic analysis shows that increasing correlations between input variables in a regression problem has a negative effect on the performance of L$_1$ for these problems [30, 31]. However, in the asymmetric SK model, correlations between the spins are controlled by the magnitude of the couplings. Larger couplings are in general easier to identify and therefore, without L$_1$, as we show, these competing factors result in a decrease in reconstruction error when coupling strength is increased. We show that this continues to be the case for both full and approximate L$_1$ regularization.

4.2 Dynamics and underlying network

We consider a kinetic Ising model endowed with the asynchronously updated Glauber dynamics which has been introduced in section 2.2.3. The dynamics are performed on a diluted binary asymmetric SK model: J_{ij} is independent of J_{ji}, and the interactions vary only in sign, not in magnitude: each coupling has the distribution

$$p(J) = \frac{c}{2N}\delta\left(J - \frac{g}{\sqrt{c}}\right) + \frac{c}{2N}\delta\left(J + \frac{g}{\sqrt{c}}\right) + \left(1 - \frac{c}{N}\right)\delta(J). \tag{4.1}$$

where c is the average in-degree (and out-degree). We are interested in sparse networks, i.e., $c \ll N$. We use $N = 40$ and $c = 5$ in our computations. Furthermore, as mentioned above, we model asymmetrically coupled spins, taking each J_{ij} independent of J_{ji}. This model can have a stationary distribution (and does for the parameters we use here), but it is not of Gibbs-Boltzmann form, and no simple expression for it is known.

4.3 Exact learning with L$_1$ regularization

As derived in chapter 3, the negative log likelihood of the spin history and updating history can be written as follows:

$$\mathcal{L}_0 = \sum_i \sum_{\tau_i} \left[s_i(\tau_i + \delta t)H_i(\tau_i) - \log 2\cosh H_i(\tau_i)\right]. \tag{4.2}$$

We can maximize the log-likelihood by simple gradient descent with a learning rate η:

$$\delta J_{ij} = \eta\frac{\partial \mathcal{L}_0}{\partial J_{ij}} = \eta \sum_{\tau_i}[s_i(\tau_i + \delta t) - \tanh H_i(\tau_i)]s_j(\tau_i). \tag{4.3}$$

For the exact learning rule, we take the initial couplings input $J_{ij}^{(0)} = 0$ and iterate equation (4.3), obtaining the corrections $\delta J_{ij}^{(n+1)}$ using the nth estimate $J_{ij}^{(n)} = J_{ij}^{(n-1)} + \delta J_{ij}^{(n)}$ on the right-hand side. Inserting each $J^{(n)}$ into equation (4.2) gives $\mathcal{L}_0^{(n)}$ at each iteration step n. If we find an increase in likelihood $\mathcal{L}_0^{(n)} - \mathcal{L}_0^{(n-1)} < 10^{-5}$, we consider the iteration to be convergent and stop.

For finite data length L, this procedure will in general produce a fully connected network. To sparsify it, L_1 penalize term is introduced to the cost function as:

$$E = -\mathcal{L}_0 + \Lambda \sum_{ij} |J_{ij}|. \tag{4.4}$$

where the first term is the negative log-likelihood and the second term is the L_1 norm. The minimization of equation (4.4) leads to an additional term in the learning rule for couplings:

$$\delta J_{ij} = \eta \left\{ \sum_{\tau_i} \left[s_i(\tau_i + \delta t) - \tanh H_i(\tau_i) \right] s_j(\tau_i) - \Lambda \operatorname{sgn}(J_{ij}) \right\}. \tag{4.5}$$

The log-likelihood function $\mathcal{L}_0$ is smooth and convex as a function of Js, so the cost function is still concave but not smooth at the place where any $J_{ij} = 0$. This leads to complications in the minimization whenever a minimum of E is at $J_{ij} = 0$: We deal with this problem by setting $J_{ij} = 0$ whenever the change (4.5) would cause J_{ij} to change sign. Then, if the minimum of E truly lies at this $J_{ij} = 0$, the estimated J_{ij} will oscillate between zero and a small nonzero value (using $\operatorname{sgn}(0) = 0$). The size of these oscillations is proportional to the learning rate η, so a sufficiently small η ensures that these couplings can be pruned by a simple rounding procedure, with a negligible chance of removing couplings that are not truly zero at the minimum. In the case that J_{ij} is not zero at the minimum, its estimated value will continue to change and it will move toward its optimal value after the step where it was set to zero.

For this learning algorithm with L_1 regularization, we take as initial couplings the J_{ij}s obtained as described above without regularization. Then, for each value of Λ, we iterate equation (4.5) to obtain successive parameter estimates. At each step n, we computer the cost function $E^{(n)}$ using the current parameter estimates and stop the iteration process if $E^{(n-1)} - E^{(n)} < 10^{-5}$. The resulting Js are then taken as the initial couplings for the next value of Λ. This procedure is carried out for all the values of Λ for which we want to evaluate the cost function.

4.4 Approximate learning schemes

We can get some insight into how the learning rule works with regularization by expanding the cost function (4.4) to second order around its minimum $\mathbf{J^0}$ when $\Lambda = 0$. Up to a constant, we have

$$\frac{E}{T} = \frac{1}{2}\sum_{ijk} C^{(i)}_{jk} v_{ij} v_{ik} + \lambda \sum_{ij} |J^0_{ij} + v_{ij}| \tag{4.6}$$

where $v_{ij} = J_{ij} - J^0_{ij}$, $T = L/N$ is the average number of updates per spin, $\lambda = \Lambda/T$, and

$$C^{(i)}_{jk} = \frac{1}{T}\sum_{\tau_i}(1 - \tanh^2 H^0_i(\tau_i))s_j(\tau_i)s_k(\tau_i), \tag{4.7}$$

where $H^0_i(\tau_i)$ is $H_i(\tau_i)$ evaluated with $J_{ij} = J^0_{ij}$.

Since the quantities in the sum in (4.7) are insensitive to whether spin i is updated, the average over updates may safely be replace by an average over all times,

$$C^{(i)}_{jk} = \langle (1 - \tanh^2 H^0_i(t))s_j(t)s_k(t)\rangle_t \tag{4.8}$$

the Fisher information matrix for spin i, which is a more robust quantity.

Minimizing (4.6), we get, to first order in λ,

$$\sum_k C^{(i)}_{jk} v_{ik} = -\lambda \mathrm{sgn}(J^0_{ij} + v_{ij}) \approx -\lambda \mathrm{sgn}(J^0_{ij}). \tag{4.9}$$

Solving this equation for v_{ij},we obtain:

$$v_{ij} = -\lambda \sum_k \left[C^{(i)}\right]^{-1}_{jk} \mathrm{sgn}(J^0_{ik}). \tag{4.10}$$

This equation shows how the regularization term shrinks the magnitudes of the couplings.

In the weak coupling limit (small g) and with a uniform external field, J_{ij} are just shrunk in magnitude proportional to λ until they reach zero and are pruned. This is a trivial kind of regularization: couplings that survive the pruning procedure the longest are simply the ones with the biggest initial absolute values. However, at larger coupling this is not the case. Some J_{ij} will be shrunk more rapidly than others, depending on the size and signs of the terms in the sum in (4.10).

Based on the quadratic expansion (4.6), we can carry out the pruning in an approximate way. We start from J^0_{ij} and a small value of λ, calculate the shifts v_{ij}

by (4.10) and remove any J_{ij} that would go though zero. With the resulting new J_{ij}s (some of them now equal to zero), increase λ, recalculate the Fisher information matrix and calculate new shifts in the parameter values. Again remove any couplings that change sign, and continue until the desired degree of pruning has been achieved. This amounts to numerical integration of the differential equation, describing how the regularization works under increasing λ.

$$\frac{dJ_{ij}(\lambda)}{d\lambda} = -\sum_k \left[C^{(i)}(\lambda)\right]^{-1}_{jk} \mathrm{sgn}(J_{ik}(\lambda)). \tag{4.11}$$

Note that now the matrix $\mathsf{C}^{(i)}(\lambda)$ depends on λ, since it is computed in the absence of the bonds that were removed at previous steps.

4.5 Performance of Algorithms

The problem we consider now becomes to identify the positive and negative couplings in the network, i.e., correctly classifying every potential bond as $+$, $-$ or 0. With an interesting intermediate length of data as T of 200 realizes and $g = 1/\sqrt{2}$, the couplings J_{ij}s without regularization ($\lambda = 0$) are shown in Fig. 4.1. The partial histograms from the zero and nonzero-J classes are partly overlap to each other.

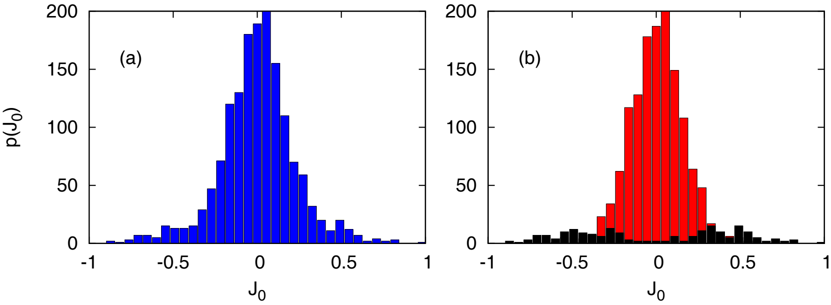

Fig. 4.1: Distribution of the inferred couplings without L$_1$ regularization, $g = 1/\sqrt{2}$ with data lengths $T = 200$. The left panel shows a histogram of the J_{ij} obtained, and the right panel shows these sorted according to whether the bond was present (black) or absent (red) in the network that generated the data.

Based on Js inferred with $\Lambda_0 = 0$ as shown in Fig. 4.1, four pruning methods were employed. Fig. 4.2 shows how the Js inferred by each method vary as the regularization coefficient λ is increased. Here, we only show positive J_0s. Bonds actually present in the model are plotted in black and red for absents.

Fig. 4.2a shows the Js inferred using exact learning with full L_1 regularization (4.5). It is apparent that the pruning process for the case shown here is not trivial: Some true (black) bonds, for which rather small values were inferred at $\lambda = 0$ because of insufficient data, are “rescued" (they fall off more slowly with λ than red ones with nearly the same initial inferred Js), and some spurious (red) bonds with high inferred values at $\lambda = 0$ are driven to zero faster than black ones with the same initial inferred Js. Thus, as λ increases the red and black lines tend to get separated, and one can do the pruning almost correctly just by turning λ up until the desired number of bonds have been removed.

We also studied three approximate methods. The first one infers Js using the quadratic expansion of the cost function and performs the regularization follows equation (4.11). We refer this as “**approximation 1**". The qualitative features of Fig. 4.2a are apparently reproduced by this approximation. In the second approximate learning, the off-diagonal elements of $\left[C^{(i)}(\lambda)\right]_{jk}^{-1}$ are ignored in (4.11), which we refer as “**approximation 2**". As shown in Fig. 4.2c, the separation of red and black curves is not as good that in Fig. 4.2b. It is evident that the slopes of the $J_{ij}(\lambda)$ curves vary rather slowly with λ in 4.2c. Therefore, we also tried a linear extrapolation based on the slopes of the curves in figure 4.2c at $\lambda = 0$. We denote this method as “**approximation 3**". One needs only to do the learning at $\lambda = 0$ (to get the $J_{ij}(\lambda)$) and calculate the Fisher matrices (to get the $dJ_{ij}/d\lambda$). Fig. 4.2d shows the result of this minimal algorithm. For approximation 3, the inferred Js that have been shrunk to zero are removed permanently. For the other three approaches, the inferred Js have a chance to be “resurrected” with increasing λ, though in the results presented here we haven’t observed this.

In order to quantify the performance of these algorithms, the empirical classification errors are computed. There are three kinds of bonds in the actual network, negative (-), positive (+) and zero. And the errors could be false positive (FP) (actual absent is predicted as present), true positive (TP) (actual present is identified as present). Then, the Receiver Operating Characteristic (ROC) curves are calculated for them. For a given λ, the false positive rate (FPR) and true positive rate (TPR) are defined as:

$$\begin{aligned} \text{TPR} &= \frac{N(\text{TPs})}{N(\text{actual presents})} \\ \text{FPR} &= \frac{N(\text{FPs})}{N(\text{actual zeros})} \end{aligned} \tag{4.12}$$

The ROC curve is a plot of TPR versus FPR. In Fig. 4.3, we plot the ROC curves for all of our methods. We further measure the performance of the different

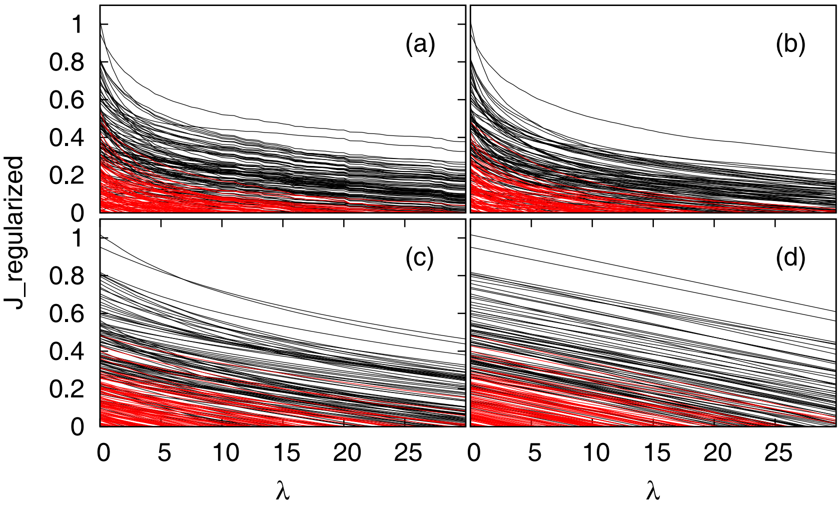

Fig. 4.2: Inferred couplings as functions of regularization coefficient λ for four methods: (a) full L_1 regularization using (4.5), (b) integration of (4.11), (c) integration of (4.11) with diagonal approximation of the inverse Fisher matrix, (d) linear extrapolation in λ of the curves in (c). Black lines represent bonds actually presents, while red lines represent ones equal to zero in the network used to generate the data. We show equal number of red and black ones.

methods quantitatively by defining an error measure, ϵ:

$$\epsilon = 1 - \text{area under ROC curve}. \tag{4.13}$$

The values of ϵ for full L_1 and approximations 1, 2, 3 are 0.03, 0.06, 0.08, 0.09 respectively. Which means full L_1 algorithm works best, followed by approximations 1, 2, 3.

We establish a baseline for the performance of the methods by performing a simple pruning procedure that does not require any L_1 regularization calculation. For a given cut value $\hat{J}$, the bonds whose Js lie in the range $[-\hat{J}, \hat{J}]$ as absent and those outside that interval as present. The black Js in Fig. 4.1b which lie within the interval are FNs and the red ones outside the interval are FPs. Varying $\hat{J}$, we obtain an ROC curve. We refer to this procedure as “J0-cut”. The curve with light blue squares in Figure 3 is calculated using this method. It gives the same value of ϵ (0.09) as Approximation 3, and the ROC curves nearly coincide.

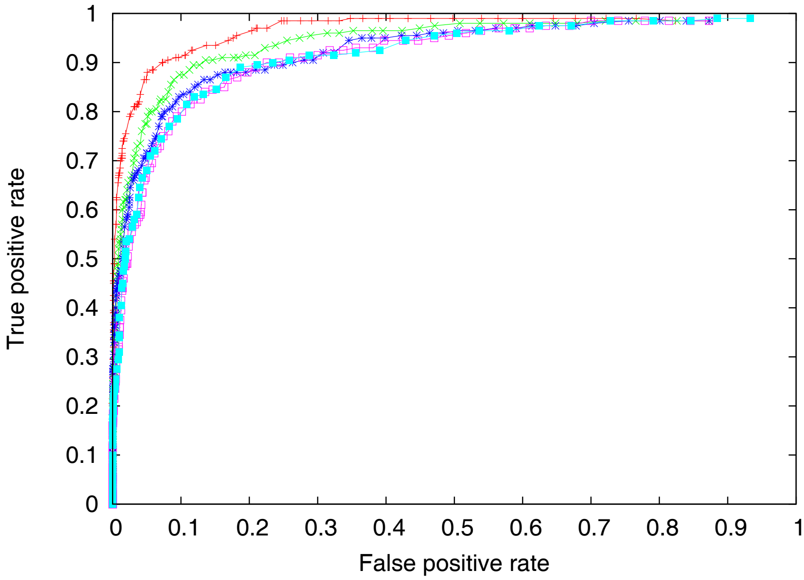

Fig. 4.3: ROC curves for full L_1 regularization, Approximations 1, 2, 3, and the J0-cut method are shown in red, green, blue, pink and light blue, respectively.

4.6 Effects of coupling strength g on L_1 regularization

A well known result about L_1 regularization is that the presence of correlations between covariates in a regression model will have a negative effect on the consistency of the reconstruction [30, 31]. Given the likelihood in equation (4.2), the inference of the connections in the kinetic Ising model can be considered as a regression problem, where the spin configurations at time t are the predictors of the values at time $t + 1$. One can therefore naively expect that increasing the strength of the connections, g, and thus the correlations between the covariates, will have a negative effect on the L_1-regularized inference. However, this is not true in our case.

To study the effect of the couplings and correlations, we thus calculate the ROC curves for full L_1 regularization and Approximation 1, respectively, for two other values of $g = 0.5, 1$.

Figure 4.4 shows how the ROC curves change as we change g for a fixed data

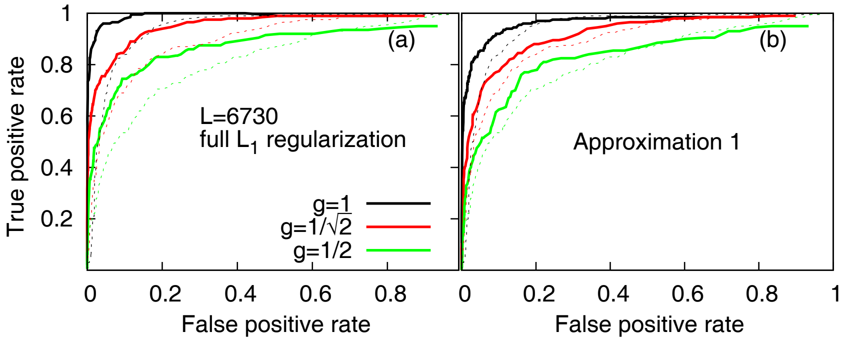

Fig. 4.4: ROC curves for full L_1 regularization (left, solid lines) and Approximations 1 (right, solid lines) with $g = 0.5,\ 1/\sqrt{2},\ 1$ respectively. The green lines for $g = 0.5$, red for $g = 1/\sqrt{2}$ and black for $g = 1$. Corresponding dashed lines are for J0-cut method of these gs. The length of the data is $L = 6730$.

length. The first observation is that for the J0-cut, shown by dashed lines, increasing g helps recovering the correct connections. We see the same trend for the exact L_1 regularization as well as Approximation 1, shown by solid curves. This can be understood intuitively in the following way. Increasing g has two negative effects. First, it increases the equal time correlations, that is, thinking about the problem as a regression problem, one would be dealing with more correlated covariates. Second, increasing g increases the correlation time and therefore different data points will be more correlated and less informative about the presence or absence of a connection. On the other hand, with larger g, the parameters to be inferred are bigger and can be identified more easily. It is the relative strength of these three factors that determine the net effect of increasing g on the inference, and, as we see, the last of these wins out over the other two for the coupling strengths we have studied. Thus, with stronger couplings, L_1 regularization is able to provide increasingly more accurate network reconstructions and more benefit over an un-regularized reconstruction.

4.7 Conclusion

We have studied the reconstruction of sparse asynchronously updated kinetic Ising networks with L_1 regularization. With smaller data length, simple maximization of the log likelihood of the system history will infer fully connected network where many bonds are actually not present. The histogram of the couplings can overlap strongly, and nontrivial methods are required to perform optimal pruning of the inferred coupling set. Here we used L_1 regularization to do this, minimizing a cost function that includes the L_1-norm of the parameter vector as a penalty term. We performed this minimization in four ways, one exact and the other three involving various degrees of approximation.

Calculations on a model network at intermediate coupling strength revealed that the full L_1 regularization classified the bonds significantly better than a naive method based on retaining the strongest bonds. Approximation 1 was somewhat worse than the exact algorithm, but still significantly better than the naive method. Our other two approximations, obtained by successive simplifications of Approximation 1, however, did not perform measurably better than the naive way, as measured by the areas under their ROC curves. These conclusions are general with respect to the coupling strengths we used. The regularization helps more with stronger coupling strengths.

This work is the first that we know of that takes a detailed look at how L_1 regularization works in a non-equilibrium model, by studying how bonds are removed successively as the regularization parameter Λ is increased. Some insight into how this happens was made possible by studying the quadratic expansion of the cost function about its minimum, which also led to the relatively successful Approximation 1. The process would have been more transparent if we could have made further simplifying approximations, as we did for Approximation 2, where we neglected off-diagonal elements of the inverse Fisher matrices. The fact that this approximation performed rather poorly (while Approximation 1 did quite well) indicates that the off-diagonal terms in (4.11) are necessary, and we lack generic insight about them.

For the kinetic SK model, increasing the couplings and thus the correlations helps the performance of L_1. This was true both for the exact L_1 solution and, for small data length, Approximation 1. Although at a first look this might sound inconsistent with the results of the regression studies with L_1 [30, 31], a closer look shows that this is not the case. In regression problems, correlations between the input covariates and the strength of the couplings between the inputs and the output are independent parameters. While for the model studied here in which these two effects covary in a way that is controlled by the magnitude of the couplings and have opposing effects on network reconstruction.

Chapter 5

Applications

This chapter refers to the applications of the derived algorithms to real recorded experimental data. Two data sets are investigated. One is neuronal spike trains and the other one is transaction data of stocks on financial market. The former one corresponds to the application part of paper II *Maximum Likelihood Reconstruction for Ising Models with Asynchronous Updates* and the latter one appears on paper IV *Financial interaction networks inferred from traded volumes*.

For the neuronal data, we use **AVE** learning rule in equation (3.34) to reconstruct the asynchronous connections of the neuron network as well as Boltzmann machine learning to infer the equilibrium couplings. The asynchronous couplings are comparable with the equilibrium ones. This implies that the dynamical process of this neuron system satisfies the Gibbs equilibrium condition and the parameters can be obtained by the maximum entropy model also. However, the asynchronous model allows the inference of self-couplings which are not presented in the equilibrium model. Furthermore, the equilibrium model needs Monte Carlo samples which makes the inference slower than the asynchronous model.

For the financial stock trades data, we use approximate inferring algorithms which are based on mean-field equations to infer a financial network composed by 100 traded stocks. By transforming the data of transaction times and volumes to binary strings, three inference methods are used to reconstruct the network. They are equilibrium, synchronous and asynchronous inference methods respectively. On one hand, the synchronous and asynchronous algorithms produce comparable results with that from equilibrium inference. On the other hand, the non-equilibrium models allow the inference of self-couplings (diagonal elements of the coupling matrix) and directed links which are not present in the equilibrium model.

5.1 Reconstruction of a neuron network

This section refers to the application part of paper II *Maximum Likelihood Reconstruction for Ising Models with Asynchronous Updates*. In this work, we extract information from the records of a neuron system. One way to decode the experimental data from neurons is to learn the network of which these neurons are part. With the model network, one can adjust the values of parameters of it to produce data close to the original spike trains as much as possible. Then, the couplings could explain the how neurons in the network produce the data and how they influence each other. Here, we choose the simplest network model as Ising model which deals with binary strings and infers the pairwise connections between neurons. Next, we show how to do it from the given data by the asynchronous model.

5.1.1 Data description and representation of data

We considered neuronal spike trains from salamander retina under stimulations by a repeated 26.5-second movie clip. The provided data set records the spiking times for the neuron and has a data length of 3180 seconds (120 repetitions of the movie clip). Here, we only focuses on $N = 20$ neurons with the highest firing rates in the data set. Data sets of this type have been studied previously using equilibrium Ising model. The data has been binned with time windows of 10 ms or more. The reason for choosing this window size is that they are larger than the typical temporal correlation width of the neurons (the typical time scale of the autocorrelation function of a neuron). Here, since we are using the kinetic model, we have the ability to study this data set using a much shorter time bin which can make low enough firing rates there is (almost) never more than one spike per bin. Then, the temporal correlations with time delays between neuron pairs as well as the self-correlations become important.

For the asynchronous Ising model, the time bins are $\delta t = 1/(\gamma N)$. For neural data, γ can be interpreted as the inverse of the time length of the autocorrelation function which is typically 10 ms. To generate the binary spin history from this spike train data set, we should therefore bin the spike trains into time bins of length $\gamma \delta t = 1/20$. Which means the size of time bins should be chosen as $\delta t = 1/(20\gamma) = 0.5$ ms. We can just simply transform the spin trains in to binaries in a common way as follows: a +1 is assigned to every time bin in which there is a spike and a -1 when there is no spikes. However, this translation will always end up with isolated instances of +1 while superfluous of -1s which is not the expected case for asynchronous Ising model. Thus, we introduce memory process for each neuron to the data set. It is a time period with an exponential distribution with mean of $1/\gamma$ in the data translation. Denote the total firing number of neuron i as F_i, and t_i^f as the firing time of fth spike for neuron i, where $i = 1, ..., N$ and $f = 1, ..., F_i - 1$, then the mapping of the spike history is follows:

$$\mathrm{s}_i(t) = \begin{cases} 1, & \text{if } t \in \left[t_i^f, \min\left(t_i^{f+1}, t_i^n + X\right)\right) \text{ with } X \sim \exp(\gamma^{-1}) \\ -1, & \text{otherwise} \end{cases} \tag{5.1}$$

where X is a period drawn from exponential distribution with mean 10 ms. By this way, we obtain the asynchronous type of data that are needed for the asynchronous model.

5.1.2 Inference methods for connections

We take the binary spin history that transformed from the original spiking trains to infer the couplings with the asynchronous Ising model by using the "AVE" learning rule in equation (3.34)

$$\delta J_{ij} = \eta\{\gamma^{-1}\dot{C}_{ij}(0) + C_{ij}(0) - \langle\tanh(H_i(t))s_j(t)\rangle\}.$$

With $J_{i0} = \theta_i, s_0(t) = 1$, the above equation also includes the learning rule for the field θ_i. Here, the learning rate is chosen as $\eta = 0.5$. The initial conditions are zero couplings and the external fields are $\theta_i = \tanh^{-1} m_i$ for the learning iterations.

We also used the same spin history to fit an equilibrium Ising model by using exact Boltzmann learning as shown in equation (3.1)

$$\begin{aligned} \delta\theta_i &= \eta\left(\langle s_i\rangle_{Data} - \langle s_i\rangle_{Model}\right), \\ \delta J_{ij} &= \eta\left(\langle s_i s_j\rangle_{Data} - \langle s_i s_j\rangle_{Model}\right). \end{aligned}$$

The learning rate η is 0.5 also and there are 100000 Monte Carlo steps per iteration for the second term in the Boltzmann learning rule. The initial conditions are same with that for asynchronous case.

5.1.3 Results

In the current inference of retina functional connections, the value of model parameters like window size δt, inverse time scale γ are set as *a priori* according to the previous studies on equilibrium Ising model. This avoids systematic studies over the value of parameters.

As presented in Fig. 5.1, the inferred couplings by Gibbs equilibrium and asynchronous kinetic Ising model are very close to each other. We also tested what

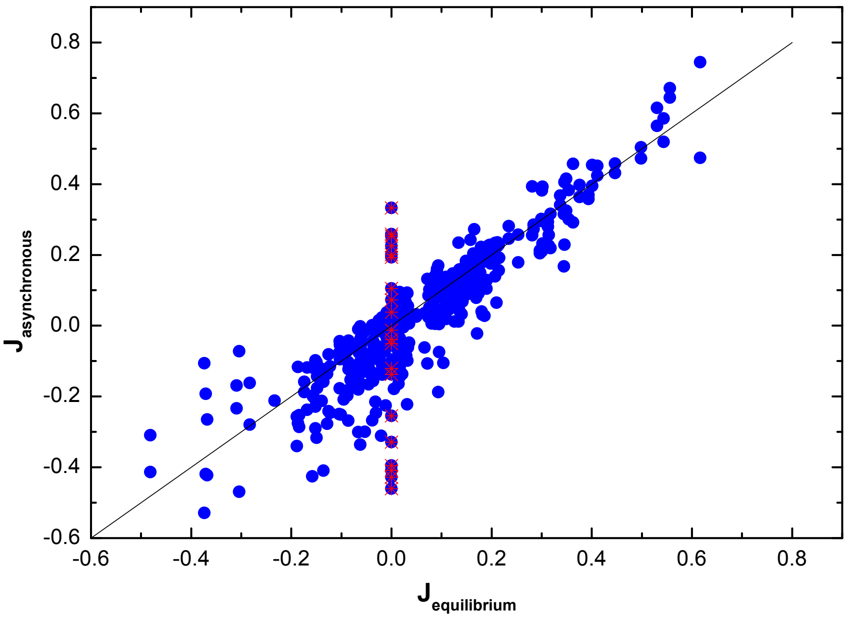

Fig. 5.1: Inferred asynchronous versus equilibrium couplings for retinal data. Red stars show the self-couplings which by convention are equal to zero for the equilibrium model.

happens to the couplings of the asynchronous model if during learning we symmetrized the couplings matrix at each iteration by adding its transpose to itself and dividing by two and also putting the self-couplings to zero. We find that the resulting asynchronous couplings get even closer to the equilibrium ones.

However, the asynchronous model allows the inference of self-couplings (diagonal elements of the coupling matrix) which are not present in the equilibrium model. As shown in Fig. 5.1, the diagonals from the equilibrium model equals to zeros by convention and denoted by the red stars. Furthermore, to be different from the symmetric couplings by the equilibrium model, the asynchronous model provides more details as the inferred couplings are directed and asymmetric.

This result provides a guide for the use of the maximum entropy equilibrium Ising model: if the asynchronous couplings were far away form the equilibrium ones, it would imply that the real dynamical process did not satisfy the Gibbs equilibrium conditions and that the final distribution of states is not the Gibbs equilibrium Ising model. Since inferring the equilibrium model is an exponentially difficult problem, requiring time consuming for Monte Carlo samplings while the asynchronous approach does not. The asynchronous learning rules thus allow the inference of functional connections that for the retinal data largely agree with the maximum entropy equilibrium model, but the inference is much faster.

5.2 Reconstruction of a finance network

This section refers to paper IV *Financial interaction networks inferred from traded volumes*. In which we use three approximate inference methods based on mean-field equation to infer a financial network from trade data of 100 stocks. They are equilibrium, synchronous and asynchronous (non-equilibrium) ones for Ising model respectively. The recorded data are transformed into binaries by local averaging and thresholding. This introduces additional parameters that have to be studied systematically to understand the behavior of the system. On one hand, the inferred couplings from synchronous and asynchronous methods are quite similar to the equilibrium ones. All produce network communities that have close industrial features. On the other hand, the non-equilibrium ones are more detailed as they are directed compared with that from the equilibrium ones.

5.2.1 Data description and representation

The recorded data was generated by transactions on the New York Stock Exchange (NYSE) over a few years, and each trade is characterized by a time, a volume traded, and a price. We only focus on the trades for 100 trading days between 02.01.2003 and 30.05.2003. However, we only use the information of trading time and volume.

We study the 10^4 central seconds of each day to avoid the opening and closing periods of the stock exchange, which is same with that in [32]. Two parameters are introduced to the data transform as the sliding window are adopted. One is the size of the sliding time window (denoted as Δt), the other one is the shifting constant (which is $\Delta s = 1$ second, the time resolution of the data). This means that the information contained in two mapped data points separated by a time less than Δt is partly redundant. However, it also means that no information from the original data is lost.

In the present work, only volume information of a trade is considered. For each stock i, we consider the sum of the volumes $V_i(t, \Delta t)$ traded in window $[t, t + \Delta t)$, and compare it to a given volume threshold $V_i^{th} = \chi V_i^{av} \Delta t$, where

V_i^{av} is the average (over the whole time series) volume of the considered stock traded per second, and χ a parameter governing our volume threshold:

$$\mathrm{s}_i(t) = \begin{cases} 1, & if \ \ V_i(t, \Delta t) \geq V_{th}^i \\ -1, & if \ \ V_i(t, \Delta t) < V_{th}^i \end{cases} \tag{5.2}$$

The parameters Δt and χ will be explored systematically for the inference with the goal that to find values of the parameters which yield inferred couplings containing interesting information.

5.2.2 Inference methods for connections

With the transformed binaries, it is natural to define magnetization m_i and correlations $C_{ij}(\tau)$ as shown in equation (2.20).

With them, we will use three different inference methods that are based on mean-field approximation. The inference formula for couplings are different for each method:

- Equilibrium inference ($i \neq j$), which only focuses on equal time correlations [21]
$$J_{ij} = -C(0)_{ij}^{-1}$$
- Synchronous inference is suitable for non-equilibrium inference, and considers also time-lagged correlations with a time lag τ in addition to equal time correlations [15], which can be rewritten as:
$$J_{ij} = \frac{1}{1 - m_i^2} \left(C(\tau) C(0)^{-1} \right)_{ij}$$
- Asynchronous inference [25], also modeling non-equilibrium processes, uses the derivative of the time-lagged correlations $\dot{C}_{ij}(\tau)$, as shown in equation (3.19) and be rewritten as:
$$J_{ij} = \frac{1}{1 - m_i^2} \left(\frac{dC(\tau)}{d\tau} |_{\tau=0} C(0)^{-1} \right)_{ij}$$

The inference formula for fields is same for these three inferences, which can be obtained from the mean-field approximation:

$$\theta_i = \operatorname{arctanh} m_i - \sum_{j \neq i} J_{ij} m_j.$$

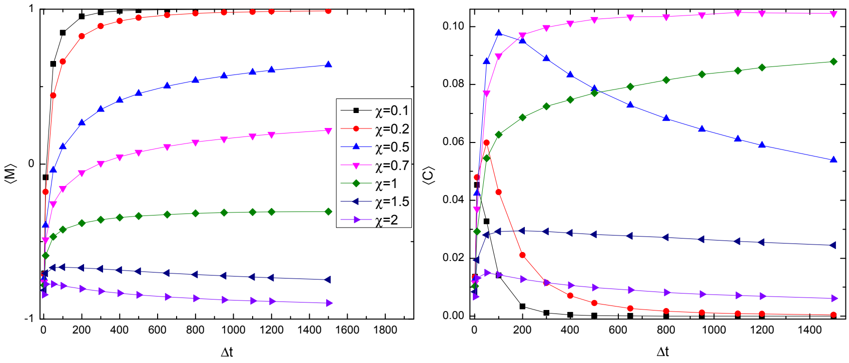

Fig. 5.2: Evolution with mapping parameters χ and Δt of the magnetization and the connected correlations, averaged over the whole dataset (all stocks or pairs of stocks, and whole time period).

The main difference between the two last methods is that synchronous inference assumes that all spins are proposed to update in parallel at a discrete time, while asynchronous inference does not have a such assumption: update times themselves are stochastic variables. The asynchronous method is supposedly more powerful, as it monitors the decay in time of all pair correlations.

It is noticeable that be different from equilibrium case, the synchronous and asynchronous inference ones have an additional parameter τ , which is the time-lag of correlations. For the asynchronous case, this time scale does not appear explicitly in the formula, but arises when the derivative is computed from the data.

5.2.3 Results

We show the average magnetization and connected correlations as a function of the window size Δt for several χs. As presented in Fig. 5.2, for short window size, correlations are small at any volume scale χ, which can be linked to the fact that the average magnetization tends to -1. Correlations are small for long window size with either small or big χ because magnetization tends to be 1 or -1 respectively.

The distributions of couplings for different values of the parameters are presented in Fig. 5.3. For asynchronous inference, the derivative of the time-lagged correlations $\dot{C}_{ij}(\tau)$ is computed through a linear fitting of this function $C_{ij}(\tau)$ using four points: $C(0)$, $C(\Delta t/5)$, $C(2\Delta t/5)$ and $C(3\Delta t/5)$. This explains

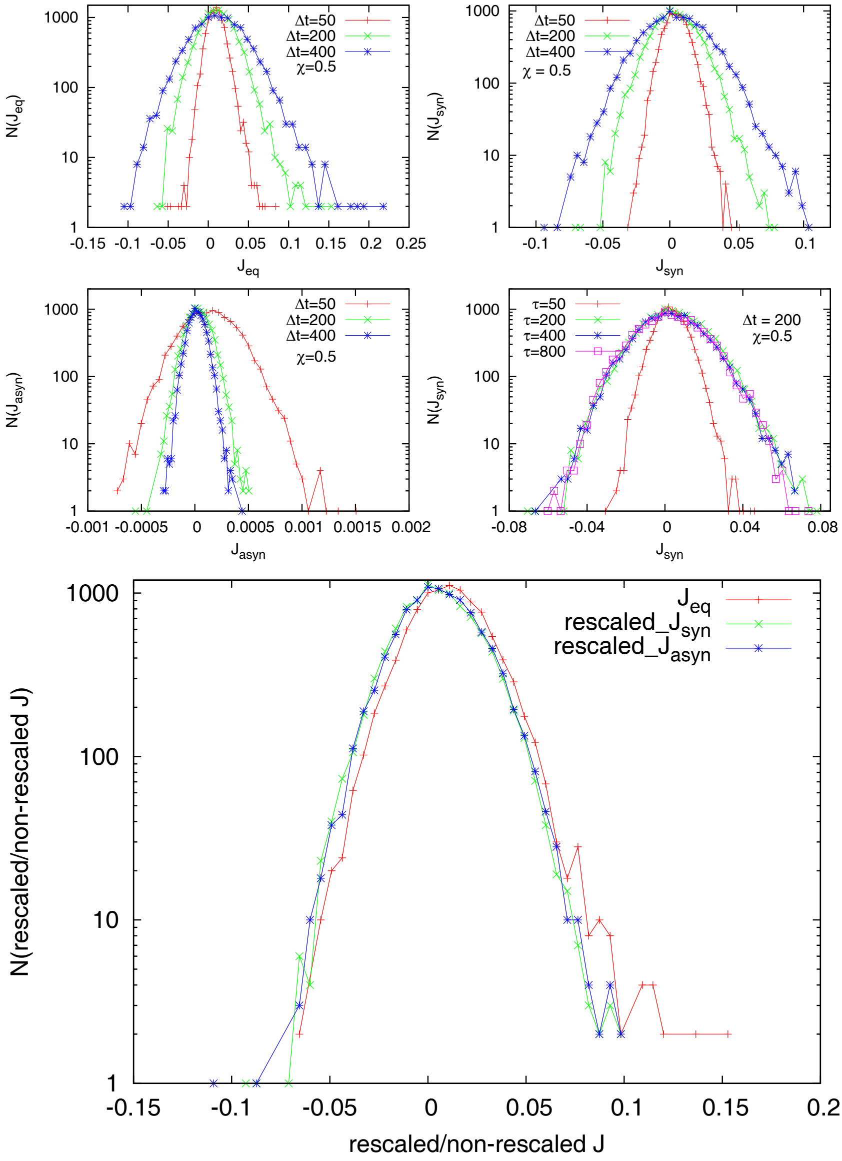

Fig. 5.3: Histograms $N(J)$ of inferred couplings. Upper panel has four log-lin subplots with: upper left one, histogram of J_{eq} with different time bins, upper right $N(J_{syn})$, using $\tau = \Delta t$; bottom left $N(J_{asyn})$ and bottom right $N(J_{syn})$ with different values of τ. Bottom panel: couplings obtained by the three inference methods. J_{syn} and J_{asyn} are rescaled to have the same standard deviation as $N(J_{eq})$. For the three versions, $\chi = 0.5$ and $\Delta t = 200$ seconds, and for synchronous inference $\tau = \Delta t$.

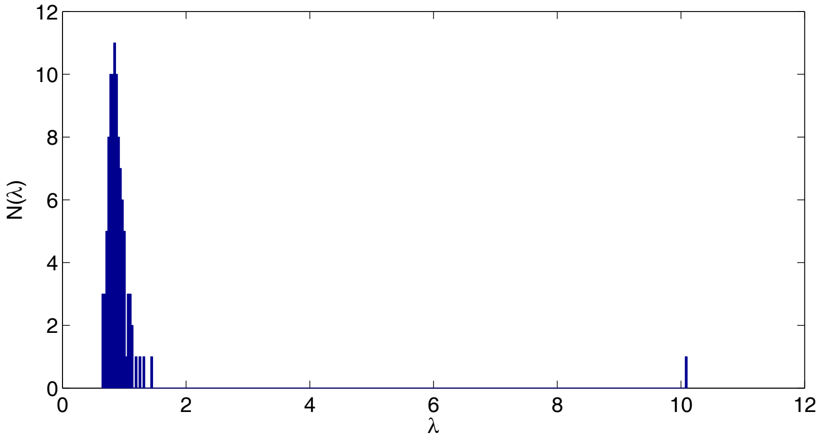

Fig. 5.4: Histograms of the eigenvalues of the equal time connected correlation matrix. Parameters: $\chi = 0.5$ and $\Delta t = 100$ seconds.

why the histogram of J_{asyn} becomes sharper when Δt is increased on the upper panel of Fig. 5.3, as this parameter is then in the denominator of the derivative.

The bottom panel of Fig. 5.3 shows that the three inference methods give similar distributions of couplings. For comparison, the distributions are rescaled on the bottom panel so as to have the same standard deviation. The upper panel shows how these distributions change with the parameters. It can be remarked that for small time scales, they have a strictly positive mean and a long positive tail. For higher time scales, the distributions are more centered around zero, but they keep an asymmetry and a longer positive tail than the negative one. This prevalence of positive couplings can intuitively be linked with the market mode phenomenon [33, 34, 35, 36]: a large eigenvalue appears, corresponding to a collective activity of all stocks, illustrated in Fig. 5.4.

With increasing values of Δt, the histograms of J_{eq} (and J_{syn} with $\tau = \Delta t$) become broader, which implies larger interactions between stocks appears. The last figure of the upper panel of Fig. 5.3 shows that the histogram of J_{syn} does not change much with τ for high values of this parameter, which indicates J_{syn} are insensitive to big values of τ.

To measure the similarity of interaction matrices J and J' which inferred from different inference methods, a similarity measurement $Q_{J,J'}$ is defined as:

$$Q_{J,J'} = \frac{\sum_{i,j} J_{ij} J'_{ij}}{\sum_{i,j} \max(J_{ij}, J'_{ij})^2} \tag{5.3}$$

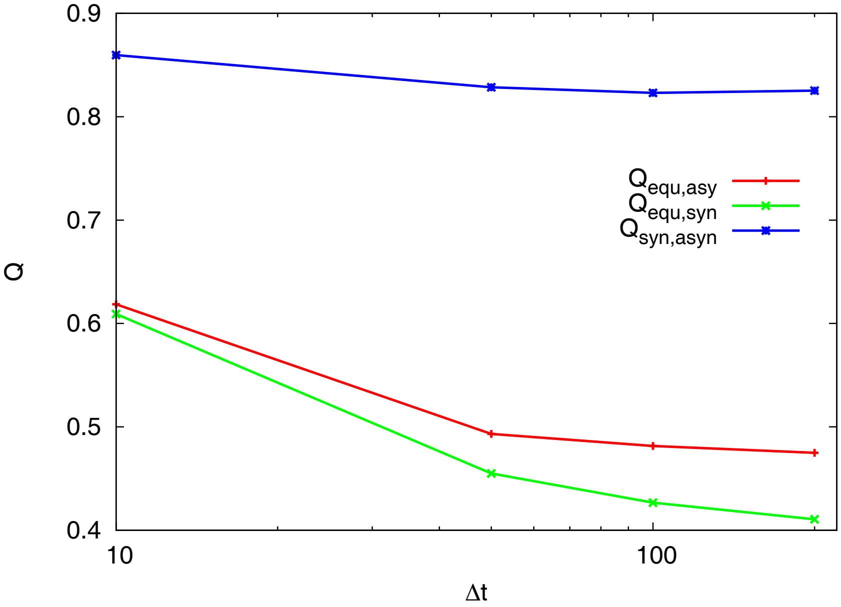

Fig. 5.5: Similarity $Q_{J,J'}$ between interaction matrices obtained with different inference methods, versus window length Δt. The couplings are rescaled to have the same mean. $\chi = 0.5, \tau = \Delta t$ for the synchronous inference, and the same fitting as for Fig. 5.3 is used for the asynchronous inference.

This measurement compares elements of two matrices one by one and gives a global similarity measure. It takes real values between 1 (when $J_{ij} = J'_{ij}$ for all i and j) and -1 ($J_{ij} = -J'_{ij}$ for all i and j), and values close to zero indicate uncorrelated couplings. The values of Q is smaller than 0.02 in absolute value when all elements of the vectors J_{ij} and J'_{ij} are drawn independently at random from a same Gaussian distribution, of mean 0, and for different values of the standard deviation of this distribution. However, Fig. 5.5 displays high similarities between couplings obtained from equilibrium, synchronous and asynchronous inference. Synchronous and asynchronous inference give especially close results, while equilibrium inference gives couplings which differ more from the other two methods. All similarities increase when Δt decreases, which is also consistent with the *Epps effect* (the phenomenon that the empirical correlation between the returns of two different stocks decreases as the sampling frequency of data increases [37]) and the fact that the system becomes less interacting on small time scales.

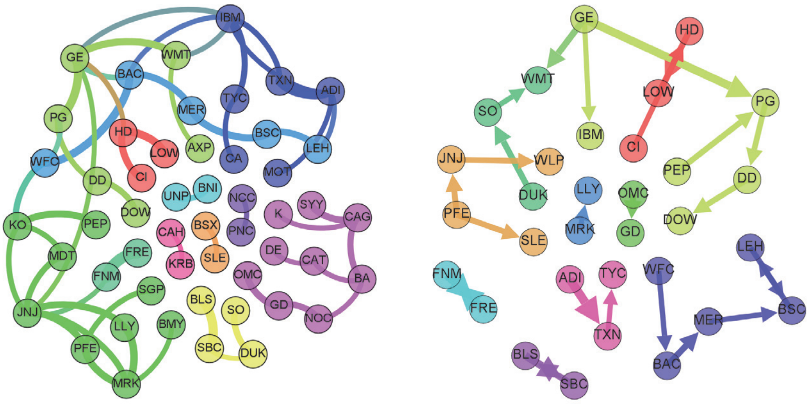

Fig. 5.6: Inferred financial networks, showing only the largest interaction strengths (proportional to the width of links and arrows). Colors are indicative, and chosen by a modularity-based community detection algorithm [16]. Parameters: $\chi = 0.5$ and $\Delta t = 100$ seconds. Left panel: equilibrium inference. Right panel: synchronous inference, with $\tau = 20$ seconds.

To show the statistical error of the inference methods, we perform the Glauber dynamics with the inferred couplings and fields obtained from the inference of the financial data. With the provided synthetic data, we perform inference again. The agreement between two sets of couplings is very good, with the MSE equals to 1.6×10^{-6}. The synchronous and asynchronous inference methods give similar low inference errors when performing corresponding test.

5.2.4 Examples of inferred finance networks

As the inferred finance networks are densely connected, we focus only on the largest couplings, which can be easily explained by closely related activities of the considered stocks. The left panel of Fig. 5.6 shows that with equilibrium inference, more than half the stocks in the data can be displayed on a network where almost all links have simple economical interpretations.

The right panel of Fig. 5.6 presents the results of synchronous inference in the same conditions. It shows that the results of equilibrium and synchronous inference are consistent, and that synchronous inference provides additional information, as it infers an directed network (all this is also true for asynchronous inference). For instance, GE is clearly a node which in influences others and is not strongly influenced itself at this level of interaction, and the financial sector is a directed chain.

From the network samples, we have the following two basic conclusions. First, they show market mode (most of the interaction strengths found are usually positive, which indicates that the financial market has a clear collective behavior) [34, 35] even only trade and volume information is considered. Stocks tend to be traded or not traded at the same time.

In addition, the strongest inferred interactions can be easily understood by similarities in the industrial activities of the considered stocks. This means that financial activity tends to concentrate on a certain activity sector at a certain time. For price dynamics this phenomenon is well-known [33, 36, 38], but it is more surprising that it appears also based on the information of traded volumes.

Symbol	**Name**	**Description**
ABT	Abbott Laboratories	Pharmaceutical
ADI	Analog Devices	Semiconductors
AFL	Aflac Incorporated	Insurance
AIG	American International Group	Insurance
ALL	Allstate Co.	Insurance
AVP	Avon Products	Personal care manufacturer
AXP	American Express Co.	Financial services
BAC	Bank of America Co.	Financial
BA	The Boeing Co.	Aerospace and defense
BAX	Baxter International	Pharmaceutical, health care
BBY	Best Buy Co. Inc.	Electronics retailer
BK	The Bank of New York Mellon Co.	Financial services
BLS	BellSouth	Telecommunications
BMY	Bristol-Myers Squibb Company	Pharmaceutical
BNI	Burlington Northern Santa Fe Co.	Railway, railroad
BSC	Bear Stearns	Investment bank
BSX	Boston Scientific Co.	Medical devices
BUD	Anheuser-Busch InBev	Beverages
CA	CA Technologies	Software corporation
CAG	ConAgra Foods Inc.	Packaged food
CAH	Cardinal Health Inc.	Pharmaceutical
CAT	Caterpillar Inc.	Machinery, financial
CCU	Compania Cervecerias Unidas S.A.	Beverages
CI	CIGNA Co.	Health care management
CL	Colgate-Palmolive Company	Health care
DD	E. I. du Pont de Nemours and Company	Chemical company
DE	Deere and Company	Heavy equipment
DHR	Danaher Co.	Conglomerate

Symbol	Name	Description
DIS	The Walt Disney Company	Mass media
DOW	The Dow Chemical Company	Chemicals
DUK	Duke energy co.	Energy
EMC	EMC Co.	Computer storage
EMR	Emerson Electric Co.	Electrical equipment
FDC	First Data Co.	Financial services
FNM	Fannie Mae	Home loan and mortgage
FON	Fiber Optic Network	Multimedia Communicator
FRE	Freddie Mac	Home loan and mortgage
GCI	Gannett Co., Inc	Media
G	Genpact	Management and Technology
GD	General Dynamics Co.	Aerospace
GDT	Guidant Co.	Designs and manufacture
GDW	Golden West Financial	Financial
GE	General Electric Company	Conglomerate
GIS	General Mills	Food
GM	General Motors Company	Automotive
GPS	The Gap, Inc.	Retail
HD	The Home Depot, Inc.	Retailing: home construction
HDI	Harley-Davidson Inc	Motorcycle manufacturers
IBM	International Business Machines Co.	IT services
IGT	International Game Technology	Gaming technology
IP	International Paper Company	Pulp and paper
ITW	Illinois Tool Works Inc.	Manufacturing
JNJ	Johnson and Johnson	Medical and pharmaceutical
K	Kellogg Company	Food
KMB	Kimberly-Clark Co.	Personal care
KO	The Coca-Cola Company	Carbonated soft drink
KRB	MBNA Co.	Banking
KR	The Kroger Co.	Retail
KSS	Kohl's Co.	Retail
LEH	Lehman Brothers Holdings Inc.	Investment services
LLY	Eli Lilly and Company	Bio-pharmacy
LOW	Lowe's Companies Inc.	Retailing
MCD	McDonald's Co.	Restaurants
MDT	Medtronic, Inc.	Medical equipment
MEL	mellon financial co.	Financials
MER	Merrill Lynch Wealth Management	Investment
MMC	Marsh-McLennan Companies, Inc.	Insurance brokers
MOT	Motorola, Inc.	Telecommunications
MRK	Merck and Co. Inc.	Bio-pharmacy

Symbol	Name	Description
NCC	National City Co.	Banks
NEM	Newmont Mining Co.	Metals and mining
NOC	Northrop Grumman Co.	Aerospace-defense
OMC	Omnicom Group Inc.	Communication
ONE	Higher One Holdings, Inc.	College business
OXY	Occidental Petroleum Co.	Oil and gas
PEP	Pepsico	Beverages
PFE	Pfizer Inc.	Pharmacy
PG	The Procter and Gamble Company	Consumer goods
PGR	Progressive Co.	Insurance
PNC	The PNC Financial Services Group, Inc.	Financial services
PPG	PPG Industries Inc.	Glass and Chemicals
RD	Royal Dutch Shell	gas and oil
SBC	SBC Communications Inc.	Telecommunication
SCH	Charles Schwab Co.	Brokerage and banking
S	Sprint Co.	Telecommunications
SGP	Schering-Plough Co.	pharmaceuticals
SLB	Schlumberger Limited	Oilfield services
SLE	Chicago-based Sara Lee Co..	Consumer-goods
SO	Southern Company	Electric utility
STI	SunTrust Banks, Inc.	Banking
SYY	Sysco Co.	Food
TRB	Tribune Company	Multimedia corporation
TXN	Texas Instruments Inc	Semiconductor
TYC	Tyco International Ltd.	Security
UNP	Union Pacific Co.	Railroad
UTX	United Technologies Co.	Conglomerate
WAG	Walgreen Co.	Retailing
WFC	Wells Fargo and Company	Banking, Financial
WLP	WellPoint Inc.	Managed health care
WMT	Wal-Mart Stores Inc	Retailing

Chapter 6

Conclusions

This thesis is composed by three parts: derivations of learning rules for asynchronous updated kinetic Ising model (papers I and II) in chapter 3, L_1 regularization (paper III) in chapter 4 and applications of learning rules to recorded experimental data (paper II and paper IV) in chapter 5.

6.1 Learning rules for asynchronous Ising model

Both approximate and exact learning rules of asynchronously updated kinetic Ising model have been derived in Chapter 3.

Two approximate learning rules are based on different levels of mean-field equations. One is based on the Curie-Weiss approximation applied to the magnetic systems, which we refer as naive mean-field method (nMF). The other one is on improved equations where the Onsager term has been considered. This one is denoted as TAP inference. Both of them are starting from the equation of motion for the correlations.

In addition, two exact learning rules are derived from maximizing two kinds of log likelihoods. One in which both spin history and updating history of spins are known. However, in the other case, only spin history is known. These two leaning rules are referred as **SUH** and **SHO** respectively. We also derived average version of **SUH** over the update times, which is denoted as **AVE** and surprisingly, it can also be derived from the equation of motion for the correlations. This indicate that the nMF inference rule is not just heuristic: it can be derived from the likelihood also.

It is expected to develop new inference methods based on the out-of-equilibrium properties of kinetic Ising model in the further research work. For instance, the generalized fluctuation-dissipation theorem (FDT) can be investigated for the specific kinetic Ising model from the non-equilibrium statistical mechanics point of view and then new inference method could be expected. These current derived inference methods provide promising performance in practical terms. The derivation of inference can be extended to different inverse statistical mechanical problems which maybe beyond the particular case of kinetic Ising model also.

6.2 L_1 regularization

L_1 regularization is applied to infer sparse asynchronous Ising model as it tends to produce sparse model in Chapter 4. With the purpose of monitoring the behavior of L_1 regularization, we only use simple gradient descent algorithm: iteratively minimize a cost function equal to minus log likelihood of data and plus an L_1 penalty norm. To heal the non-differentiability of the L_1 norm with respect to the couplings, we put them to zeros by hand when they change their signs during the minimization. We refer this as full L_1 regularization. We also perform the approximate calculation which is based on a quadratic expansion of the cost function around its minimum. The pruning of connections is tracked by increasing the strength of L_1 penalty from zero to large values.

We find that increasing the coupling strength improves the reconstruction of connections, which seems contrary to regression models which are typically studied in the context of L_1 regularization. However, it is not the case. In regression problems, correlations between the input covariates and the strength of the couplings between the inputs and the output are independent parameters. This is not the case for the model studied here or for many other kinetic models in which these two effects covary in a way that is controlled by the magnitude of the couplings and have opposing effects on network reconstruction.

The inferred couplings from the derived algorithms are fully connected. However, in real applications, some suspicious and spurious weak links need to be eliminated. In such case, L_1 regularization should be added as it allows to sort true small couplings from truly zero couplings. L_1 regularization is expected to be applied in the network reconstruction from real experimental data for biological system, financial system, etc.

6.3 Applications of learning rules

Two of the derived learning rules are applied to inferring interaction network from the recorded experimental data in Chapter 5. The first case is applying "**AVE**" learning rule to reconstruct the couplings between neurons from recorded spike trains. The data are firstly transformed to binary strings with asynchronous style by introducing memory effect to each neuron. On one hand, the reconstructed couplings by asynchronous kinetic Ising model present very similar results comparing with that from Gibbs equilibrium model, as they are close to each other on the scatter plot of them. On the other hand, the asynchronous model allows the inference of self-couplings (diagonal elements of the coupling matrix) which are not present in the equilibrium model.

The second case is to reconstruct a financial network from trades data of 100 stocks recorded from NYSE. With a sliding window of size Δt, we move the time window with a sliding constant $\Delta s = 1$ second. We map the raw data to binary strings by local averaging and thresholding. The interaction matrix of the financial network is obtained by three inference methods which are all based on mean-field equations. They are equilibrium, synchronous and asynchronous methods respectively. On one hand, coupling matrices inferred by equilibrium methods are quite close to that from synchronous and asynchronous methods, on the other hand, the latter provide more details as the inferred couplings are directed.

For the first data set, we are only focused the values of one set of parameters. It is possible to investigate them systematically. Furthermore, we can also try to infer the couplings by other asynchronous learning rule (say, **SHO**) or synchronous model. For the second data set, we can perform different ways of data mapping, in which the information of price can also be included. Additionally, the current mapping are based thresholding of the average, for which we can move to quantiles to avoid the heavy fluctuations in the system behaviors.

Bibliography

[1] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. *Weak pairwise correlations imply strongly correlated network states in a neural population*. Nature **440**, 1007 (2006).

[2] C. E. Shannon. *Mathematical theory of communication*. Univ. of Illinois Press, Bell System Technical Journal **207**, 379 (1948).

[3] L. Landau and E. Lifshitz. *Statistical physics*. Course of theoretical physics, Pergamon International Library of Science, Technology, Engineering and Social Studies, Oxford: Pergamon Press **1** (1980).

[4] R. J. Glauber. *Time-dependent statistics of the Ising model*. Journal of Mathematical Physics **4**, 294 (1963).

[5] M. Chaves, R. Albert, and E. D. Sontag. *Robustness and fragility of boolean models for genetic regulatory networks*. Journal of theoretical biology **235**, 431 (2005).

[6] F. Greil and B. Drossel. *Dynamics of critical Kauffman networks under asynchronous stochastic update*. Physical Review Letters **95**, 048701 (2005).

[7] K. Klemm, S. Bornholdt, and H. G. Schuster. *Beyond Hebb: Exclusive-or and biological learning*. Physical Review Letters **84**, 3013 (2000).

[8] D. Thouless, P. Anderson, and R. Palmer. *Solution of 'solvable model of a spin glass'*. Philosophical Magazine **35**, 593 (1977).

[9] R. Tibshirani. *Regression shrinkage and selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological) **58**, 267 (1996).

[10] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. *High-dimensional Ising model selection using L_1-regularized logistic regression*. The Annals of Statistics **38**, 1287 (2010).

[11] J. Friedman, T. Hastie, and R. Tibshirani. *Regularization paths for generalized linear models via coordinate descent*. Journal of Statistical Software **33**, 1 (2010).

[12] D. Sherrington and S. Kirkpatrick. *Solvable model of a spin-glass*. Physcal Review Letters **35**, 1792 (1975).

[13] A. Crisanti and H. Sompolinsky. *Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model*. Physical Review A **36**, 4922 (1987).

[14] Y. Roudi, J. Tyrcha, and J. Hertz. *Ising model for neural data: Model quality and approximate methods for extracting functional connectivity*. Physical Review E **79**, 051915 (2009).

[15] Y. Roudi and J. Hertz. *Mean field theory for nonequilibrium network reconstruction*. Physical Review Letters **106**, 048702 (2011a).

[16] Y. Roudi and J. Hertz. *Dynamical TAP equations for non-equilibrium Ising spin glasses*. Journal of Statistical Mechanics: Theory and Experiment P03031 (2011).

[17] P. Peretto. *Collective properties of neural networks: a statistical physics approach*. Biological cybernetics **50**, 51 (1984).

[18] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. *A learning algorithm for Boltzmann machines*. Cognitive Science **9**, 147 (1985).

[19] T. Plefka. *Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model*. Journal of Physics A: Mathematical and General **15**, 1971 (1982).

[20] T. Tanaka. *Mean-field theory of Boltzmann machine learning*. Physical Review E **58**, 2302 (1998).

[21] H. J. Kappen and F. Rodriguez. *Efficient learning in Boltzmann machines using linear response theory*. Neural Computation **10**, 1137 (1998).

[22] A. Ranaldo. *Order aggressiveness in limit order book markets*. Journal of Financial Markets **7**, 53 (2004).

[23] S. Maslov. *Simple model of a limit order-driven market*. Physica A: Statistical Mechanics and its Applications **278**, 571 (2000).

[24] C. Kipnis and C. Landim. *Scaling limits of interacting particle systems*, volume 320. Springer (1999).

[25] H.-L. Zeng, E. Aurell, M. Alava, and H. Mahmoudi. *Network inference using asynchronously updated kinetic Ising model*. Physical Review E **83**, 041135 (2011).

[26] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. *High-dimensional graphical model selection using L_1-regularized logistic regression*. Advances in Neural Information Processing Systems **19**, 1465 (2007).

[27] E. Aurell and M. Ekeberg. *Inverse Ising inference using all the data*. Physical Review Letters **108**, 090201 (2012).

[28] J. A. Hertz, Y. Roudi, A. Thorning, J. Tyrcha, E. Aurell, and H.-L. Zeng. *Inferring network connectivity using kinetic Ising models*. BMC Neuroscience **11**, P51 (2010).

[29] H.-L. Zeng, M. Alava, E. Aurell, J. Hertz, and Y. Roudi. *Maximum likelihood reconstruction for Ising models with asynchronous updates*. Physical Review Letters **110**, 210601 (2013).

[30] K. Knight and W. Fu. *Asymptotics for Lasso-type estimators*. Annals of Statistics 1356–1378 (2000).

[31] P. Zhao and B. Yu. *On model selection consistency of Lasso*. The Journal of Machine Learning Research **7**, 2541 (2006).

[32] I. Mastromatteo and M. Marsili. *On the criticality of inferred models*. Journal of Statistical Mechanics: Theory and Experiment P10012 (2011).

[33] T. Bury. *Statistical pairwise interaction model of stock market*. The European Physical Journal B **86**, 1 (2013).

[34] J.-P. Bouchaud and M. Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge university press (2003).

[35] R. N. Mantegna and H. E. Stanley. *An introduction to econophysics: correlations and complexity in finance* (2003).

[36] C. Biely and S. Thurner. *Random matrix ensembles of time-lagged correlation matrices: derivation of eigenvalue spectra and analysis of financial time-series*. Quantitative Finance **8**, 705 (2008).

[37] T. W. Epps. *Comovements in stock prices in the very short run*. Journal of the American Statistical Association **74**, 291 (1979).

[38] L. Kullmann, J. Kertész, and K. Kaski. *Time-dependent cross-correlations between different stock returns: A directed network of influence.* Physical Review E **66**, 026125 (2002).

Aalto-DD 117/2014

ISBN 978-952-60-5802-3
ISBN 978-952-60-5803-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Applied Physics
www.aalto.fi