Modelling financial lead-lag interactions with Kinetic Ising Models

Carlo Campajola

Scuola Normale Superiore

Classe di Scienze

SCUOLA
NORMALE
SUPERIORE

PhD Thesis

Corso di perfezionamento in Matematica per la Finanza

Advisors:

Prof. Dr. Fabrizio Lillo

Prof. Dr. Daniele Tantari

Contents

Acknowledgements

I want to thank my supervisors, Fabrizio Lillo and Daniele Tantari, who have given me guidance to turn ideas into science and always spurred me to the best of my capabilities. I am also grateful to the head of our PhD program, Stefano Marmi, for his dedication and joyful support.

A special thanks goes to all my colleagues at the Quantitative Finance group in Pisa, particularly to Clemente, Frèderic, Gael and Walter with whom I started this adventure, and to Giulia, Mateusz, Giuseppe and Danilo who were of great help academically and with their friendship.

I thank my coauthors Piero and Domenico who have contributed to the works featured in this thesis, as well as my collaborators in other works not included here. A particular thanks to Claudio J. Tessone, who hosted me for a visiting period and helped in my development as an aspiring researcher.

A huge thanks to my friends and family for always being of invaluable support, in particular to Francesco for his true friendship and to Irene and Andrea for hosting me in the difficult times of the COVID-19 pandemic.

Finally, my greatest luck has been having my girlfriend Carla by my side all this time, and I could never be grateful enough for this.

Pisa, 2020

Chapter 1

Introduction

Modern economic and financial systems are some of the most interesting and complex structures that can be studied. What makes them particularly fascinating is that, unlike other systems such as ecologies, materials or proteins, the ways in which they are formed and behave do not descend from physical or biological laws, but only from the imagination and inventive of their creators and of those who act within these structures, ultimately of human beings.

For this reason it has become growingly popular to study financial markets from the participants' perspective, rather than by only looking at the dynamics of asset prices, making it possible to account for the heterogeneity of strategies and behaviors that different individuals (or organizations) have in their toolbox. This has produced the vast literature on agent-based models (ABM), which aim at reproducing stylized price dynamics determined by the collective behavior of more or less rational traders sharing more or less information (Cont (2007); Farmer and Foley (2009); Alfi et al. (2009a,b); Leal et al. (2016); Fagiolo et al. (2019a)).

Some of the ABM literature focuses on the study of phenomena in fi-

nancial markets that have been broadly described as herd behavior, usually defined as the human tendency to mimic the actions of a social group with which the individual identifies, regardless of whether those actions are rational or irrational. In finance this reflects in the observation that large amounts of traders sometimes show remarkably similar behavior in the short term, often causing price movements that are not justified by fundamental information and increasing price volatility (Grinblatt et al. (1995); Bouchaud et al. (2009); Toth et al. (2015)).

One of the criticisms that have been moved to ABMs is that they are often prone to having large numbers of parameters that cannot be estimated or statistically validated on empirical data, which in recent years has led to more effort being put into the development of validation methods for these models (Alfarano et al. (2005); Barde (2016); Fagiolo et al. (2019b)).

As classical economic models used to borrow methods and theories from the natural sciences, in particular from Netwonian physics and Darwinian evolutionary theory (and still unconsciously do, see Montes (2003) for a philosophical critique), agent-based models borrow from modern developments of these disciplines, namely statistical physics and (evolutionary) game theory. Statistical physics in particular provides a modelling approach which is ideal to settings where a large number of interacting agents are taken into account, as it has been developed to describe extraordinarily large amounts of interacting particles whose collective behaviour is determined by their individual, extremely stylized properties. While naively translating physical models to financial markets is not necessarily a good idea, mainly because as mentioned financial markets are ultimately not physical systems, the insights that can be produced by adapting such models to financial systems are considerable (Sornette (2014); Challet et al.

(2013)).

Throughout this thesis we provide novel contributions to several streams of literature, ranging from statistical mechanics to agent-based modelling and financial econometrics, their connection being in the application to financial systems of the Kinetic Ising Model (KIM).

The KIM (Derrida et al. (1987); Crisanti and Sompolinsky (1988)), developed as a neural network model, describes the dynamics of a system of binary variables - named "spins" in the statistical physics literature - that have a lagged influence on each other. For a model with N variables $s(t) \in \{-1, 1\}^N$, the number of parameters is $N(N+1)$, where N parameters h_i, $i = 1, \ldots, N$, describe the tendency specific to spin i to have positive or negative value in a vacuum, and a matrix of N^2 parameters J_{ij} summarize the effect of the interaction between spin s_i and s_j. This interaction is lagged in time, which means that spin i's evolution is influenced by spin j's past value through the J_{ij} parameter while J_{ji} accounts for the reciprocal (and possibly asymmetric) effect. The nature of this interaction is considered irrelevant, the only thing that matters is its magnitude: the farther from 0 the value of J_{ij}, the stronger the effect that s_i experiences coming from s_j.

As mentioned, the KIM has been developed as a neural network model in the late 1980s: these were among the first prototypes that were proposed to understand the neurological functioning of animal brains and set some of the pillars of what has then become the field of artificial intelligence (Hopfield (1982); Coolen (2001a,b)). Much like brains, which are made of overall very simple units (neurons) but are able of extremely complex functions, neural networks are able to store and retrieve complex information presented to them by storing it in the interaction patterns among their nodes.

In particular recurrent neural networks such as the KIM are able to store and reproduce temporal patterns.

Given that the model is defined in such a way that the specific nature of the interaction between variables is irrelevant, it has been used to describe time series from neuron spike trains in animal brains (Hertz et al. (2010); Capone et al. (2015)) as well as to model interacting traders in financial markets (Bornholdt (2001); Bouchaud (2013)). This is the key difference between what we propose in this thesis and for instance an agent-based model: we do not focus on the mechanisms that make two variables interact and reduce all that information to a single number, the coupling J_{ij}.

This is both our main advantage and our main limitation: making this assumption allows to use a very simple and flexible model to study a variety of different systems without running into the hardships of calibrating many parameters, while still keeping track of some of the system's complexity; on the other hand it also limits the depth to which the model describes the peculiarities of the system. We believe that taking this approach is functional to evaluate the presence of lagged dependencies between binary variables in high-dimensional datasets, where it would be excessively complicated to model each interaction by its own specific properties, or in situations where the nature of the interaction is unknown to the modeller but the information about its existence can be used to make more informed decisions.

In our applications to financial systems we show how the KIM can be used to describe the relation between traders strategies as well as stocks volatility, allowing to quantify herding effects or to study abnormal collective price movements at high frequency with a simple and effective method that can be implemented for real time use.

Another perspective that can be considered is the one of financial econo-

metrics and time series analysis, where the KIM would be defined as a vector logistic autoregressive model of order 1. While in this thesis we do not focus on the use of our models for hypothesis testing, it is clear that our contributions take inspiration from that stream of literature too, particularly in the last part where we present the Score-Driven KIM. Furthermore a model very similar to the KIM, the Vector Discrete AutoRegressive model of Jacobs and Lewis (1978), has been used in several financial applications in recent years (Taranto et al. (2014); Mazzarisi et al. (2020a,b)): in an appendix to this thesis we expand on the relation between the two, finding they are equivalent with the restriction that the values of J_{ij} need to be positive. This puts two very different streams of literature into communication and it is our hope that it will open opportunities to improve the understanding and functionality of these models.

Having established the motivation for our work, in the following paragraphs we will briefly introduce the content of the remaining chapters of this thesis. After a review of relevant literature in Chapter 2, we begin in Chapter 3 by developing an inference method for the KIM from time series with significant amounts of missing values.

We tackle the problem of inferring a weighted causality network from multiple binary time series by using the KIM in datasets where a fraction of observations is missing. This is highly relevant in a number of real world situations which show up in social sciences, where even if the existence of an agent is known to the observer it is not always possible to measure its state, for reasons that go from observations being costly to intrinsic features of the system.

The literature on Restricted Boltzmann Machines[1] is not new to this

[1]RBMs are a family of neural network models widely used in machine learning for classification, dimensionality reduction and feature learning tasks (see Hinton (2012) and

sort of problem but typically considers a setting where a fraction of nodes is permanently hidden, rather than having all nodes being hidden for a fraction of their observations (Dunn and Roudi (2013); Decelle et al. (2016)). What we present in Chapter 3 then is a generalization to the existing methods by allowing the missing observations to show up for any variable at any time.

Our algorithm relies on the path integral method of Martin et al. (1973) to approximate the log-likelihood of the model, allowing to calculate it in polynomial time rather than exponential time, which can be easily paired with model selection techniques such as the LASSO (Tibshirani (1996)) or Decimation (Decelle and Zhang (2015)) to obtain a sparse network solution. The inference algorithm properly accounts for the presence of missing values by computing their posterior means, which are in turn used to improve the accuracy of the network inference in a sort of Expectation-Maximization procedure (Dempster et al. (1977)).

We test the performance of the algorithm on synthetic data and find interesting properties regarding the dependency on heterogeneity of the observation frequency of spins: in particular the more heterogeneous is the distribution of missing values across variables, the least efficient the method is at correctly reconstructing both the network and the missing values. We also find that some of the assumptions necessary to the analytical derivation of the approximated log-likelihood do not impact the quality of the estimation too much even when they are violated.

One possible application of this modelling approach is presented in Chapter 4, where we use it to infer lead-lag relationship networks between

references therein for a comprehensive review of methods and applications). They rely on a bipartite weighted network of visible and hidden units which is tuned to match a target probability distribution observed from data.

investors in the foreign exchange market (FX) and to reconstruct the aggregate state of supply and demand at all times. We analyze records from the electronic trading platform of a major dealer in the FX market, where clients of any sort can request a quote from the dealer to exchange some amount of one currency for another.

The architecture of the FX market relies on a centralized interdealer exchange with a continuous double-auction mechanism where only few market members (the dealers) are allowed to trade, mostly large banks and financial institutions. These dealers in turn offer their intermediation services to the public, by providing proprietary electronic trading platforms with other trading mechanisms, typically in the shape of on-demand over-the-counter (OTC) trading. This mechanism operates based on the dealer maintaining a balanced portfolio of currencies, named the *inventory*, such that a client can obtain immediate execution of her trades in exchange for a premium rewarding the dealer for taking the risk related to continuously holding large amounts of currencies with fluctuating market value.

Taking the perspective of the dealer, a significant part of its risk management is related to mitigating adverse selection risk, namely the risk of trading with a more informed counterpart (Kyle (1985); Glosten and Milgrom (1985)). In particular in the case of the dealer it is highly likely that significant fractions of its clients are trading based on more information, as they are probably specialized in that business, which results in the risk of accumulating significant amounts of "bad" inventory while trading away "good" inventory: if for example clients require US Dollars and pay in Euro, the dealer will accumulate Euro and be lacking US Dollars in its reserves and will thus need to trade Euros for US Dollars on the interdealer market; however if the clients were informed that the EUR/USD exchange rate

would go up in the near future, meaning Euros are worth less compared to US Dollars, the dealer will incur in a loss due to its uninformed trading. This is referred to as inventory risk (Ho and Stoll (1980)) and the dealer offsets it by imposing a premium to its clients in the form of a spread between exchange rates, which is asymmetric with respect to the interdealer spread to incentivize client's flow in a favorable direction and wide enough to let it safely trade on the interdealer market.

For this reason it is relevant to the dealer to understand how information propagates among its clients, identifying players that can forerun large order flows across the market and thus correctly price the risk they will bear.

We propose to approach the problem using the Kinetic Ising Model and analyze the trade records of one dealer's clients at the 5 minutes time scale on the EUR/USD spot exchange rate market. We assume that trades are observations of the opinion the trader holds about the rate, namely if she is buying USD she believes the EUR/USD rate will go up and viceversa. In particular we take the sign of net volume $V_i(t)$ of EUR acquired in exchange for USD by agent i in the 5 minutes window $(t - 5m, t]$: if it is positive (resp. negative) we assign a value of $+1$ (-1) to its opinion $s_i(t)$, which is considered as a spin of the KIM.

However most of the traders are not active every 5 minutes, even if it is reasonable to assume that they still hold an opinion and refrain from trading because of transaction costs, limited liquidity, risk-aversion or other causes, but they or other traders adopting similar strategies might be active on other venues: for this reason the time series contains a significant amount of missing values which, if correctly estimated, would provide a clearer picture of the state of supply and demand in the whole market, even the parts to which the dealer is not directly connected.

Thanks to the inference algorithm we present in Chapter 3 we are able to reconstruct a lead-lag network between traders unveiling the way in which their trades relate over time, as well as to reconstruct the unobserved opinions. We apply influencer detection techniques to the networks to identify leading players in the market and we define a new herding measure, based on both the observed and estimated traders opinions. We show that this herding measure has Granger Causality relations with the state of liquidity in the centralized interdealer market, thus linking to transaction costs the dealer pays when rebalancing its inventory. Overall our results show that the dealer efficiently propagates favorable states of liquidity to its clients while absorbing temporary flow imbalances, thus contributing to market efficiency and stability.

Up to this point we have considered stationary models, where parameters do not vary in time or they do so slowly enough that they can be considered constant in a subsample. However this stationarity assumption is not particularly realistic in a variety of situations, as events that warp the dynamics of financial variables happen all the time. One prominent example of time-varying parameter in financial literature is the volatility of price returns, towards which a huge modelling effort has been devoted since decades (Bollerslev (1986); Heston (1993); Cox (1996)). It is widely accepted that even something as simple as the variance of log-returns is not constant even throughout the same trading day, hinting that it is important to consider approaches that take into account time-varying parameters when modelling financial markets.

In the last research Chapter of this thesis, Chapter 5, we extend the Kinetic Ising Model to its Score-Driven formulation, a particularly flexible and interesting form of time-varying parameters modelling approach. Intro-

duced by Creal et al. (2013) and Harvey (2013), score-driven models are a particular class of observation-driven models, which differ from parameter-driven models by the deterministic evolution of the time-varying parameters as functions of the observations, avoiding the addition of further sources of stochasticity.

In score-driven models the evolution of the time-varying parameters at time t depends deterministically on the observations through the score, i.e. the gradient of the conditional log-likelihood with respect to the time-varying parameters. We will better introduce the details of the approach in Chapter 5, but the idea is that the model parameters evolve towards their local maximum likelihood value driven by a dynamics resembling the Newton method for optimization, where the previous value of the parameters is taken as the starting point for a step in the steepest descent direction.

The main advantage of taking this approach, compared to a more "classical" parameter-driven model, lies in the less challenging estimation: being the time-varying parameters fully determined by the observations, there is no need for Monte Carlo simulations when computing the value of the likelihood and thus even very complex models can be efficiently estimated with relatively low effort. This is true in principle for any observation-driven model, but the score-driven models have been shown to be optimal among their "relatives" in terms of information theory by Blasques et al. (2015). Another advantage is that the model can be used as a misspecified filter, where the time-varying parameters are estimated from the data without knowledge of their actual laws of motion: as long as it is meaningful to introduce a dynamical parameter, endowing it with a score-driven dynamics allows to effectively measure its behaviour without relying on additional assumptions, as has been shown by Koopman et al. (2016).

In this thesis we propose two specifications of the Score-Driven KIM: the Dynamical Noise KIM (DyNoKIM) and the Dynamic Endogeneity KIM (DyEKIM). The two differ by the number and kind of parameters which are considered to be time-varying: in the DyNoKIM we only have one dynamical parameter capturing the level of randomness in the observations, while in the DyEKIM we factor the parameters of the KIM in a way that different time-varying parameters account for the relative importance of one set of effects over the others, particularly focusing on distinguishing between endogenous and exogenous dynamics of the observations.

We show that the DyNoKIM, with its time-varying "inverse noise" parameter $\beta(t)$ inspired by the inverse temperature of statistical physics, is particularly useful to assess the reliability of forecasts made by the model, as exhibited by computing the theoretical form of the Area Under the ROC Curve (AUC) at different values of β. The AUC is a standard metric of performance for binary classifiers, which the KIM de facto is, and summarizes the specificity and sensitivity of the classifier (Bradley (1997)): we show that the AUC is an increasing function of the value of the inferred $\beta(t)$ (and thus decreasing in the estimated noise level), with a functional form depending on the other parameters of the model and the data distribution. Since β can be estimated in real time, and forecasts made at higher $\beta(t)$ values can be considered more "reliable" than the ones made at lower $\beta(t)$, this provides a useful tool to continuously monitor the forecast ability of the model and to decide how to account for its predictions in a more informed and data-driven fashion.

We apply the DyNoKIM to a dataset of US stock prices at the 5 seconds time scale, where we map times t where stock i changes price to positive values of the spin $s_i(t) = +1$, while if the price does not change we take

$s_i(t) = -1$. This quantity, referred to as stock activity in the literature (Rambaldi et al. (2015); Wheatley et al. (2019)), is taken as a proxy for high-frequency volatility, meaning that periods where the $s_i(t)$ are consistently more positive than negative are periods of higher volatility and viceversa. We show that the AUC measured empirically for one-step ahead forecasts matches the theoretical dependence on $\beta(t)$, further justifying this modelling approach for real world applications.

In the same Chapter 5 we also propose a more elaborate model, the DyEKIM, which we design to discern between endogenous and exogenous dynamics of the observations in the KIM framework. Intuitively, if the dynamics is endogenous it means that future realizations of the observations have a strong dependency on their past realizations, either in the form of auto-correlations or of lagged cross-correlations, while if the dynamics is exogenous it is driven by other factors, captured by external regressors or common trends. By defining a set of time-varying parameters each acting as a common factor for auto-correlations (the diagonal of the interaction matrix J), lagged cross-correlations (the off-diagonal terms of J), idiosyncratic and common trends (the bias vector h) and external regressors we are able to nicely separate these effects, thus gaining insight on the relative importance each of them has in determining the observations at a given point in time.

We provide two example applications for the Dynamic Endogeneity KIM, one applied to a similar dataset of US stock prices as the previous example and one applied to the traders activity dataset of Chapter 4. In the US stocks application we consider two events that caused some turmoil in the market, the Flash Crash of May 6, 2010 (SEC (2010); Kirilenko et al. (2017); Menkveld and Yueshen (2019)) and the Federal Open Market Com-

mittee meeting report announcement of July 31, 2019 (Powell (2019)). The ability to separate endogenous from exogenous effects granted by the time-varying parameters of the DyEKIM is useful to understand which mechanisms are in effect before, during, and after these events. The main difference between the two events is their predictability: while the Flash Crash happened completely unexpected and for initially obscure causes, FOMC announcements are scheduled events taking place periodically during the year, meaning that the market can "prepare" for the latter in the previous days and hours, thus reducing abnormal effects at the exact time of the announcement.

Indeed this difference is also highlighted by the patterns of our time-varying parameters, where we find evidence that the Flash Crash originated from an increase in exogenous volatility which then triggered a consistent amount of volatility spillovers across stocks, while the reaction to the FOMC scheduled announcement is much more contained and quickly absorbed by the market. Both effects are consistent with relevant literature on this sort of events, such as Kirilenko et al. (2017) and Hautsch et al. (2011).

What is actually interesting is that the FOMC report of July 31, 2019 was followed by a press conference by the FOMC Chairman Jerome H. Powell, which caused some turmoil due to a few unexpected statements by the Chairman during the press Q&A regarding future policy decisions. In that case we see an effect in our time-varying parameters which is much more similar to the one observed for the Flash Crash. Overall then this experiment shows that the DyEKIM can be used to assess the level of endogeneity in the dynamics and that its interpretation is consistent with other analyses of similar events.

The second example application for the DyEKIM regards the same

dataset analyzed in Chapter 4, where we now study the behaviour of the traders in the hours before and after a set of macroeconomic news announcements. These are scheduled announcements, such as unemployment rate reports or FOMC meetings, which are particularly relevant for one currency or another, e.g. a FOMC announcement will cause a re-evaluation of the value of the US Dollar relative to other currencies but should not affect the Euro or the Pound. We find that the traders diverge from their typical strategic behaviour in proximity of scheduled news, with their trading dynamics becoming less endogenous and less driven by prices but more driven by risk-aversion, as we show that they present a common trend to drop the currency affected by the news in the minutes leading to the announcement.

Finally, this thesis contains an Appendix where we show an interesting result regarding the equivalence between the Kinetic Ising Model and the Vector Discrete AutoRegressive model of order 1 (Jacobs and Lewis (1978)), provided that the elements of the J matrix of the KIM are all $J_{ij} \geq 0$. The two models have been designed and studied for over thirty years in two very different streams of literature, which we now put in connection by formalizing an equivalence theorem. We hope that this result will prove useful to both scientific communities, showing once more that the cross-fertilization between disciplines can be beneficial.

Chapter 2

From Statistical Mechanics to Finance

2.1 The Ising model and its successors

It has been a long and rich history the one that started in 1925 with the publication of Ernst Ising's doctoral thesis on a new model ideated by his supervisor Wilhelm Lenz (Ising (1925)). Originally meant to model the ferromagnetic phase transition in solid state materials, that is the empirical observation that below a certain *critical temperature* some materials develop an intrinsic magnetic dipole moment which is zero otherwise, Ising's first conclusion after solving the 1-dimensional version was that the model was not good for its purpose and had to be discarded, since he proved that the sought phase transition was not occurring.

It took 10 years before the then Manhattan Project scholar Rudolf Peierls, while working with Hans Bethe and Max Born, showed that the phase transition would occur on lattices of dimension greater or equal than two (Peierls (1936)) and 10 more before an exact solution of the 2D Ising

model was published by Lars Onsager (Onsager (1944)). Following Onsager's paper the interest in the Ising model was revamped and, while Ernst Ising himself (after surviving WWII as a German Jew and emigrating to the US) never published again in his career, the model carrying his name is one of the most celebrated and influential in statistical physics and its implications have reached far beyond the borders of physics itself.

In its original formulation, the Ising model describes a set of N interacting *spins* $s \in \{-1, 1\}^N$ described at equilibrium by the Hamiltonian

$$\mathcal{H} = -J \sum_{\langle i,j \rangle} s_i s_j$$

where J is a parameter characterizing the interaction and $\langle i, j \rangle$ indicates a set of neighbouring spins in a given space, as for example a lattice or any network. The Boltzmann probability distribution for the equilibrium states of this model at a given temperature T is then formulated as

$$P(s|J, \beta) = \frac{1}{Z} \exp\{\beta J \sum_{\langle i,j \rangle} s_i s_j\}$$

where Z is the normalizing partition function and $\beta = 1/k_B T$, with k_B the Boltzmann universal constant. By looking at this formula it appears clear that if $J > 0$, called the *ferromagnetic* case, spins will be more likely to be found in configurations where they are aligned (that is, with the same sign) with their neighbours, whereas if $J < 0$, called the *antiferromagnetic* case, spins will be favouring configurations in which they have opposite sign than their neighbours.

Countless variations on this theme have been produced through the decades, first exploring higher dimensionalities d of the space in which the system evolves (Onsager (1944), Ferrenberg and Landau (1991)), then in-

vestigating higher-dimensional spins (Fisher (1964), Stanley and Kaplan (1966), Kosterlitz (1974)) and all possibly imaginable combinations. Most of these variations share one common property, one that is very much appealing to physicists, which is that they are *equilibrium* models: in a nutshell, the model is analyzed in its equilibrium states, trying to predict the properties of materials in experiments that typically are not able to measure quantities on time scales where out-of-equilibrium behavior is visible.

Equilibrium statistical mechanics is the theory showing how macroscopic observable quantities such as temperature, pressure or the magnetic field emanating from a magnet are the result of the microscopic interaction of atomic particles, a macroscopic average of many microscopic states. It is often said to provide a conjunction between microscopic quantum mechanics and the macroscopic classical mechanics, and throughout the years a multitude of methods and models have been developed in order to efficiently study, simulate, and infer statistical mechanical models to describe real, macroscopic systems, giving birth to the growing field of *complex systems*.

The most celebrated of the descendents of the Ising model are probably the Edwards-Anderson (EA) (Edwards and Anderson (1975)) and the Sherrington-Kirkpatrick (SK) model (Kirkpatrick and Sherrington (1978)), where a completely new layer of complexity is added on top of Ising's original formulation: in the EA model the interaction term J becomes a symmetric matrix of random variables J_{ij}, so that any equilibrium property now depends not on the value of J but on its *distribution*. The Hamiltonian reads

$$\mathcal{H} = -\sum_{\langle i,j \rangle} J_{ij} s_i s_j - \sum_i h_i s_i$$

where $J_{ij} = J_{ji} \sim \mathcal{N}(J_0, J_1^2)$ and h is a vector of local fields, biasing the

spins towards one direction. This model is meant to characterize a class of systems called *spin glasses*, that are materials with non-trivial magnetic properties due to the presence of impurities in random sites of their crystalline structures, but it is far more general than that as we will show. The SK model is the solvable version of the EA model, allowing the structure of the coupling matrix J to not depend on the underlying space and running the sum over $i < j$ rather than $\langle i, j \rangle$. This simplification allowed to give a complete characterization of the equilibrium states of the model, which can be summarized in three categories (called *phases*) depending on the values of J_0 and J_1 with respect to the temperature T:

- a *ferromagnetic phase* (large J_1/T and large J_0/J_1 ratio) where spins align all together in one direction;
- a *paramagnetic phase* (small J_1/T and small J_0/J_1 ratio) where spins have random direction and don't show collective behaviour;
- a *spin glass phase* (large J_1/T and small J_0/J_1 ratio) where if one takes a single realization of the J_{ij}s the individual spins freeze in one of the two states, but do not show global ordering properties.

The spin glass phase is a peculiar phenomenon which arises as a consequence of the randomness of J, and its discovery has been hugely impactful in defining a whole new class of models for complex systems. Although the main topic of this thesis is not spin glass models, most of the concepts we will build upon in the next chapters are rooted in the spin glass and statistical physics literature, from which we will extensively borrow methods and ideas for the formulation and inference of models for financial time series.

2.2 Why Ising models

A legitimate question to ask is: why should Ising models be so interesting and widespread beyond the pretty narrow application of magnetic materials?

The answer was given in a beautiful and enlightening article by Edwin Jaynes (Jaynes (1957)), where the foundations were laid for the bridge between the information theory of Shannon (Shannon (1948); Cover and Thomas (2012)) and statistical mechanics, which proved crucial in explaining a simple yet powerful idea: the entropy defined by Shannon for communication systems and the entropy defined by Gibbs for physical systems not only share their formula, but also the concept behind them is the same.

The argument goes along these lines: take a random variable $x \in \{x_1, \ldots, x_i, \ldots, x_n\}$ with corresponding probabilities p_i, which are unknown. The only known quantity is the expected value of a function $f(x)$

$$\langle f(x) \rangle = \sum_{i=1}^{n} p_i f(x_i) \tag{2.1}$$

Is it possible from this information to determine the form of the probability distribution of x that requires the least possible arbitrary assumptions? Jaynes points out that this can be framed as a constrained optimization problem, where the cost function is provided by the Shannon (or Gibbs) entropy of the probability distribution p

$$H(p_1, \ldots, p_n) = -K \sum_i p_i \log p_i$$

where K is a positive constant. As Shannon proved, this quantity is the only one which is always positive, is increasing with increasing uncertainty about the random variable x and is additive for independent sources

of uncertainty. One then needs to maximize this function subject to two constraints: the first is that $\{p_i\}$, being a probability distribution, needs to be normalized, that is $\sum_i p_i = 1$; the second is the known information that we have, that is Eq. 2.1. Introducing Lagrangian multiplier constants λ, μ the problem is recast into

$$\{p_i\} = \arg\max_{\{p_i\}} \left[H\left(\{p_i\}\right) + \lambda \left(\sum_i p_i - 1 \right) + \mu \left(\sum_i p_i f(x_i) - \langle f(x) \rangle \right) \right] \tag{2.2}$$

which gives the result

$$\begin{aligned} p_i &= \exp\{-\lambda - \mu f(x_i)\} \\ \lambda &= \log Z(\mu) \\ Z(\mu) &= \sum_i \exp\{-\mu f(x_i)\} \end{aligned}$$

This can be easily generalized if a set of constraints is given, and it is particularly relevant to our case when such constraints are the averages of binary variables $\langle s_i \rangle$ and their correlations $\langle s_i s_j \rangle$. In this case it is easy to see that the solution to the optimization problem leads to find

$$\begin{aligned} p(\{s\}) &= \exp\{\sum_{i<j} J_{ij} s_i s_j + \sum_i h_i s_i - \log Z(J,h)\} \\ Z(J,h) &= \sum_{\{s\}} \exp\{\sum_{i<j} J_{ij} s_i s_j + \sum_i h_i s_i\} \end{aligned}$$

which coincides exactly with the Boltzmann distribution of the SK model for $\beta = 1$. This means that the SK model is intrinsically the optimal model (following this *Maximum Entropy* principle) for a system of binary random variables for which we only hold information about averages

and correlations, and is the main reason for which the family of models descending from Ising's initial formulation has been keeping scientists of many backgrounds interested for almost a century.

2.3 The Kinetic Ising Model

In this thesis our main focus is on the Kinetic Ising Model (KIM), an out-of-equilibrium version of the SK model (Derrida et al. (1987); Crisanti and Sompolinsky (1988)) developed a few years later and proposed as dynamical model for asymmetric neural networks. As we mentioned above, one of the main assumptions in spin glass models is that $J_{ij} = J_{ji}$, that is interactions are symmetric between spins. It is sufficient to break this assumption to have the model completely change its properties, as the asymmetry is at odds with the concept of correlation where $\langle s_i s_j \rangle = \langle s_j s_i \rangle$ by definition and might look contrasting to what was stated in the previous section.

However there is still one ingredient that has been missing in this discussion, one that is extremely relevant in the analysis of financial variables: time.

If we introduce dynamics into the equations, there is a whole new set of correlations that can be constrained via the Maximum Entropy principle, that are correlations at lag l $\langle s_i(t+l)s_j(t)\rangle$. It is straightforward to see that now these correlations are not invariant to permutation of i and j, hence if for the SK model the number of constraints was $N + N(N-1)/2$ it has now grown to $N + N(N-1)$. The result of the constrained optimization of Eq. 2.2 for $l = 1$ will then be

$$p(\{s_i(t+1)\}|\{s_i(t)\}, J, h) = \frac{1}{Z(t)} \exp\{\sum_{i,j} J_{ij} s_i(t+1) s_j(t) + \sum_i h_i s_i(t+1)\} \tag{2.3}$$

Typically, in the physics literature, the J elements are assumed to be *iid* Gaussian random variables, $J_{ij} \sim \mathcal{N}(J_0, J_1^2/N)$ and the properties of the model as data generating process are the object of analysis. This simple change in the structure of J has a huge impact on the behaviour of the model, which loses its spin glass phase and only preserves a dynamic phase transition between a paramagnetic and a ferromagnetic phase when the mean of the J elements, J_0, is greater than 1.

It goes beyond the scope of this thesis to characterize the model in its physical formulation, for which results can be found in the literature (see Crisanti and Sompolinsky (1988); Derrida et al. (1987); Coolen (2001a,b)). It is instead our goal to use the model in the context of financial time series analysis, and for this reason we need to tackle the problem of inferring the model parameters from data. In the following paragraphs we will summarize the state of the art regarding inference methods and model selection criteria for Kinetic Ising Models, which are the true foundations for this work.

2.4 Inference methods for the Kinetic Ising Model

Inferring a model from data is the process of computing the set of parameters that is the most likely given the observations, which is typically achieved by maximizing the *posterior* probability

$$\{J,h\} = \arg\max_{\{J,h\}} p(\{J,h\}|\{s(t)\})$$

By applying Bayes's formula to the posterior, we can recognize it splits in two probabilities: the *prior* and the *likelihood*:

$$p(\{J,h\}|\{s(t)\}) = \frac{\overbrace{p(\{J,h\})}^{\text{Prior}}\overbrace{p(\{s(t)\}|\{J,h\})}^{\text{Likelihood}}}{p(\{s(t)\})} \tag{2.4}$$

and $p(\{s(t)\})$ is just a normalizing factor. The prior can be arbitrarily chosen by the modeller, as it reflects any external knowledge about the model that is not dependent on the data, as for example information about how sparse the J matrix should be. For the moment being we will consider a uniform prior, so that the above relation simplifies to

$$p(\{J,h\}|\{s(t)\}) \propto p(\{s(t)\}|\{J,h\})$$

This relation states that, under the uniform prior assumption, any set of parameters $\{J,h\}$ that maximizes the likelihood is also the maximum posterior estimator for the model.

The likelihood for the Kinetic Ising Model reads

$$p(\{s(t)\}|\{J,h\}) = \prod_t \prod_i \frac{1}{Z(t)} \exp\left[s_i(t+1)\left(\sum_j J_{ij}s_j(t) + h_i\right)\right]$$

where $Z(t) = 2\cosh\left[h_i + \sum_j J_{ij}s_j(t)\right]$. The easiest and more straightforward way of maximizing this function is via Gradient Ascent methods (Nesterov (2008); Bottou (2010); Kingma and Ba (2014)), which basically rely on computing the gradient of the logarithm of the likelihood (the *log-likelihood*) and following the steepest path towards the maximum. Many

different implementations of this algorithm can be found in publicly available code libraries, each proposing some feature that is supposed to make the estimation faster and more reliable. We have reported in the bibliography the ones that we used throughout this thesis, but they are far from being the only available options.

While using Gradient Ascent methods is definitely the most intuitive way to tackle this problem, there are settings where, resorting to some mild assumptions, the inference of the parameters can be made even simpler or allow to extract more information from the data. As we will show in Chapter 3, when a fraction of the data is not observable (due to measurement errors, noise or cost) it is useful to construct a method to use the inferred model to estimate the missing data, and recursively use this information to improve the accuracy of the inferred parameters. In order to do so, a set of assumptions and approximations is necessary following what is called a *Mean Field Method* (Opper and Saad (2001)).

The first Mean Field Methods developed for the inference of the Kinetic Ising Model are the ones by Roudi and Hertz (2011a,b), where they derive the so-called Naive Mean Field (NMF) and Thouless-Anderson-Palmer (TAP) approximations. The baseline assumption they make is that, as in the original physical formulation, $J_{ij} \sim^{iid} \mathcal{N}(0, J^2/N)$. Calling $m_i = \langle s_i \rangle$ and $\delta s_i(t) = s_i(t) - m_i$ (averages are taken over time), they calculate from the data the lagged and synchronous correlations, $D_{ij} = \langle \delta s_i(t+1) \delta s_j(t) \rangle$ and $C_{ij} = \langle \delta s_i(t) \delta s_j(t) \rangle$.

The Naive Mean Field approximation is based on the assumption that each individual spin will "see" others as if they were represented by their average values, thus assuming that the fluctuations in the sum $\sum_j J_{ij} s_j(t)$ are negligible (or, in other words, that its variance tends to 0). This is

typically a reasonable approximation only when $N \to \infty$ and J is a dense matrix, but it is a standard starting point for analysis. This leads to the self-consistency relation

$$m_i = \tanh\left(h_i + \sum_j J_{ij}^{NMF} m_j\right)$$

which, after a series expansion of the tanh term leads to find

$$\langle \delta s_i(t+1)\delta s_j(t)\rangle = (1-m_i^2)\sum_k J_{ik}^{NMF}\langle \delta s_k(t)\delta s_j(t)\rangle$$

which can be rewritten to give a simple form for the J^{NMF} matrix, by calling $A_{ij}^{NMF} = (1-m_i^2)\delta_{ij}$ where δ_{ij} is the Kronecker delta symbol,

$$J^{NMF} = \left[A^{NMF}\right]^{-1} DC^{-1} \tag{2.5}$$

The slightly more complicated TAP approximation, which was formulated for the SK model by Thouless et al. (1977), orbits around the relation

$$m_i = \tanh\left[h_i + \sum_j J_{ij}^{TAP} m_j - m_i \sum_j \left[J^{TAP}\right]_{ij}^2 (1-m_j^2)\right]$$

shown to be valid for the Kinetic Ising Model by Roudi and Hertz (2011a). In this approximation they find that Eq. 2.5 is still valid once one modifies A into

$$A_{ij}^{TAP} = A_{ij}^{NMF}\left[1 - (1-m_i^2)\sum_l \left[J^{TAP}\right]_{il}^2 (1-m_l^2)\right]$$

which however needs to be solved iteratively, as now J^{TAP} appears on both sides of the Eq. 2.5.

One further development was proposed by Sakellariou (2013), who developed an exact inference method based only on the assumption of Gaussian

random couplings J_{ij} and of a large enough number of spins N. Given this, it is straightforward to see that the sum $\sum_j J_{ij}s_j(t)$ in the limit $N, T \to \infty$ is itself a Gaussian random variable with mean and variance

$$g_i = \sum_j J_{ij} m_j$$
$$\Delta_i = \sum_{j,k} J_{ij} J_{ik} \left[\langle s_j(t) s_k(t)\rangle - m_j m_k\right] = \sum_j J_{ij}^2 (1 - m_j^2)$$

where the last equality comes from the fact that the J_{ij}s are independent of each other and so $j \neq k$ terms vanish in the $N \to \infty$ limit. As a result, the time averages can now be replaced with Gaussian integrals

$$m_i = \int Dx \tanh\left[h_i + g_i + x\sqrt{\Delta_i}\right]$$

where $Dx = \frac{dx}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$ is a Gaussian integration measure.

The lagged and instantaneous correlations, following the same argument, are shown to be related by

$$D = AJC$$

where again $D_{ij} = \langle \delta s_i(t+1) \delta s_j(t)\rangle$, $C_{ij} = \langle \delta s_i(t) \delta s_j(t)\rangle$ and A is a diagonal matrix with entries

$$A_{ii} = \int Dx \left[1 - \tanh^2\left(h_i + g_i + x\sqrt{\Delta_i}\right)\right]$$

which again leads to find a recursive relation to determine J of the same form as Eq. 2.5. The important improvement for this method is that it does not rely on the typical assumption of Mean Field Methods of weak interaction, while both the NMF and TAP results are only valid in the limit

$J_{ij} \to 0 \; \forall \, i, j$, but on the other hand it requires the numerical solution of multiple integrals to be able to infer the parameters.

2.5 Model selection criteria

All of the above results implicitly assume that there is no restriction on the structure of J and h, meaning that if one were to draw the resulting model as a network where the links are non-zero elements of J they would be most likely facing a fully connected network, a problem that is common to Maximum Entropy models. However a model with too many parameters is almost as uninformative as one with none, since the goal of fitting a model is extracting few relevant features and mechanisms that can improve the understanding of the process that generated the data, or even help in forecasting future observations. As someone said, "the best model for a cat is the same cat, but what can you learn from it?"[1].

This is the reason why a significant stream of literature has focused in developing efficient model selection criteria, which help the modeller in selecting the most relevant parameters and discard the ones that don't contribute much to the description of the data. Some of them, like the Akaike and Bayes Information Criteria (Akaike (1974); Schwarz et al. (1978)), define a quantity that is the difference between some function of the number of parameters and the log-likelihood of the fitted model and then look at which model minimizes such quantity; another option is the likelihood ratio test, which tries to determine whether a model with an extra parameter is statistically "better" than without it through a statistical test that is asymptotically correct, but with the drawback that for a finite number of

[1] I thank Dr. Andrea Baronchelli for this quote.

parameters the test statistic could have unknown distribution.

Another stream of literature puts its focus on the choice of meaningful prior distributions to be put in Eq. 2.4, in order to either impose a structure on the J matrix or to penalize models with too many parameters. Probably the most relevant approach of this kind for our case is the LASSO regularization (Tibshirani (1996)), where a prior of the kind $p(\{J\}) \propto \exp\left[\sum_{ij} \lambda |J_{ij}|\right]$ is added to penalize the quantity of non-zero elements of J, with $\lambda > 0$ being a free parameter to be determined with out-of-sample validation. This has been used for Ising models as a standard technique to obtain sparse models (Ravikumar et al. (2010)), but it has been challenged by more model-specific techniques such as Decimation (Decelle and Zhang (2015); Decelle et al. (2016)). We will describe in more detail the Decimation approach in Chapter 3, but it is in principle a different take on the likelihood ratio test which, instead of producing a test statistic to compare models, defines a transformed log-likelihood function which compares the restricted model with the complete and empty ones. The model having a log-likelihood that is the farthest from a linear interpolation of the two extreme cases is selected. In Chapter 3 we show how this method outperforms the standard LASSO approach for our applications, consistent with the results of Decelle and Zhang (2015).

2.6 Handling missing data: deletion, direct estimation, imputation

An extremely interesting problem in the modelling of time series is how one can deal with missing observations, a rather common scenario in a variety of real world settings. Starting with Rubin (1976), the problem has been

tackled from statisticians in a more and more elaborate way by first studying how missing data can affect the model estimation and when to ignore the mechanism leading to missing observations and then progressively devising methods that allow to improve the model inference and predict what values would better fill the gaps in the data.

Buhi et al. (2008) provide an overview of commonly adopted methods, which typically belong to one of three categories: deletion, direct estimation or imputation. Deletion techniques are the most used: they discard partial observations, only taking into account for statistical analysis the samples where there is no missing data. This is typically the standard for computing correlations or inferring generalized linear models, with most statistical software using this as the default method. Clearly this approach can significantly reduce the sample size, leading to higher estimation errors and lower statistical power, and is very weak when the data are not missing at random, that is there is an underlying reason for which those data are missing.

Direct estimation techniques instead consider for statistical inference any piece of available data, removing from the analysis only the missing entries instead of whole samples. The advantage of these methods is that one does not reduce the sample size and, in the case of Bayesian inference methods, can even try to overcome any bias in the sampling by adding prior information to the model.

The last and most interesting family of methods is the one of imputation techniques: these not only do not discard any available data, but try to use available information and modelling assumptions to fill in (*impute*) the missing observations. Probably the most popular and celebrated approach of this kind is the Expectation-Maximization algorithm (Dempster et al.

(1977)), which alternates a log-likelihood maximization step and a missing data expectation step, filling gaps in the data by substituting them with their posterior expectations given the current set of model parameters. Iterating this procedure has been shown to produce consistent estimates for data that are missing at random (that is where the sampling is not biased), as shown in Little and Rubin (2019).

2.7 Taking averages without sampling: the generating functional

What the Expectation-Maximization approach typically does in the Expectation step is computing averages either by having an analytical solution to the expectation integral or by sampling from the model's probability distribution with a Monte Carlo algorithm when such analytical solutions are not available. However Monte Carlo sampling is extremely inefficient when the cardinality of the configuration space is exponentially large in the number of variables (as it is for most models) and it thus requires an exponentially large time to produce consistent averages.

One method to produce analytical expressions for posterior expectations in highly complex settings is the generating functional approach, originally proposed by Martin et al. (1973), which takes advantage of the concept of *path integral*. When dealing with a time-evolving quantity the expectation is to be carried over all possible paths it will follow during its dynamics. What Martin et al. (1973) first and Janssen (1976) and De Dominicis (1978) then realized is that the computational complexity of sampling from a very large configuration space can be shifted to the solution of a high dimensional integral over a set of auxiliary fields which allow to obtain the sought

averages as derivatives of the resulting functional, called indeed the *generating functional*. These high dimensional integrals have all the advantages that continuous mathematics provides for calculus, allowing to use exact and approximate techniques (as for instance the saddle point method) to obtain solutions.

To illustrate the principles behind this technique, which we will extensively use throughout this thesis, let's consider a simple setting where one variable $x(t)$ evolves in discrete time according to some stochastic law

$$x(t+1) = F(x(t)) + \eta(t) \tag{2.6}$$

where $\eta(t)$ is a discrete time martingale and F some generic function. Imagine we want to evaluate the average over all possible realizations of some quantity $\phi(t)$ which is a function of the underlying stochastic process, that is $\mathbb{E}[\phi(t)] = \mathbb{E}[\phi[x(t)]]$ with the expectation to be taken on the $P(\{x(t)\})$ measure.

As the value of ϕ is determined uniquely by the realization of the stochastic process x, the average operation is equivalent to a functional integration over all possible functions $x(t)$, taking into account the restriction that they should respect the law of Eq. 2.6, what is commonly known as a path integral. Such a restriction can be expressed by the use of a Dirac delta function, here given in its integral representation

$$\delta(x - x_0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp\{-i\hat{x}(x - x_0)\} d\hat{x}$$

Then the required average takes the form

$$\mathbb{E}[\phi[x(t)]] = \mathbb{E}\left[\int \mathcal{D}x \phi[x(t)] \prod_t \delta(x(t+1) - F(x(t)) - \eta(t)) \right]$$

where the expectation is taken over the noise term η. It is then easy to rewrite this last equation as

$$\mathbb{E}[\phi[x(t)]] = \int \mathcal{D}[x,\hat{x}]\phi[x(t)]e^{-i\sum_t \hat{x}(t)[x(t+1)-F(x(t))]}\mathbb{E}\left[e^{-i\sum_t \hat{x}(t)\eta(t)}\right] \quad (2.7)$$

which is generally much simpler to solve as the expectation now involves only the noise term and no other possibly very complicated function.

One important property of this approach is that Eq. 2.7 can be further manipulated to obtain a generating functional for any quantity ϕ. Defining an auxiliary variable $\psi(t)$ we can define the generating functional for the moments of ϕ as

$$\mathcal{G}[\psi] = \log \int \mathcal{D}[x,\hat{x}]e^{\sum_t \psi(t)\phi[x(t)]}e^{-i\sum_t \hat{x}(t)[x(t+1)-F(x(t))]}\mathbb{E}\left[e^{-i\sum_t \hat{x}(t)\eta(t)}\right] \quad (2.8)$$

and it is straightforward to find that the n-th moment of ϕ corresponds to the n-th derivative of $\mathcal{G}$ with respect to the auxiliary variable $\psi(t)$ in the limit of $\psi(t) \to 0$, or more explicitly

$$\mathbb{E}\left[\phi^n[x(t)]\right] = \lim_{\psi(t)\to 0} \frac{\partial^n \mathcal{G}}{\partial \psi(t)^n}$$

While Eq. 2.8 might look complicated at a first glance, it has to be noticed that the functional integration is now trivial in x thanks to the introduction of the δ functions, and the exponential form of the integrand allows, under mild assumptions on F being smooth, to apply integration techniques such as the saddle-point approximation to solve the integral in $\hat{x}$.

2.8 Ising models for finance

Ising-like models and their countless variations have been used throughout the last decades to describe data or model systems with the most diverse nature (Bury (2013); Bouchaud (2013); Tanaka and Scheraga (1977); Cocco et al. (2017); Kadirvelu et al. (2017)) and to increase our understanding of how natural, artificial, social and economic systems work.

On the one hand these models, studied in their original physical formulation, can be manipulated to generate a wide range of behaviours mimicking the features of these systems (Bouchaud (2013); Bornholdt (2001)), and use a deductive approach to explain the stylized properties of data we observe in the real world. On the other hand one can use these models in the fashion of descriptive and forecasting models (Bury (2013); Cocco et al. (2017); Ibuki et al. (2013); Kadirvelu et al. (2017)), by using Maximum Likelihood (ML) and Maximum A Posteriori (MAP) techniques to fit the model to the data, inductively working towards an explanation of the observations. This is typically referred to as the inverse formulation of the model, while the former is the direct formulation.

Both approaches have been taken in the context of financial data, mostly in a consistent stream of literature commonly known as *econophysics*, but more recently the impact of models of this family in the economic and econometric literatures has been growing thanks to the increasingly popular Agent-Based Models methodology as in Cont (2007), Farmer and Foley (2009) and Fagiolo et al. (2019b).

Perhaps the most general framework that has been proposed is the one of the Random Field Ising Model (RFIM) of Bouchaud (2013), which is essentially a Kinetic Ising Model with a random external driver, very similar to what we study throughout the thesis except for the fact that the

coupling coefficients J_{ij} are assumed to be positive. This formulation is by the way equivalent to the Discrete AutoRegressive model of order 1 (see Jacobs and Lewis (1978)) with an exogenous regressor, as we show in Appendix A. Again, the original purpose of the model was to explain hysteresis phenomena in magnetism, but it is clear that these kinds of mechanisms can be mapped more or less for free to social and economic dynamics. In particular hysteresis cycles are situations where a system globally switches non-linearly between two behaviours and where the change from behaviour A to behaviour B happens at a different point in the space of parameters than the change from B to A. The RFIM framework shows that if the interactions between the Ising spins J are large enough the system presents hysteresis when looking at the average behaviour, while it behaves quasi-linearly with weaker interactions (or higher noise levels). This kind of behaviour has been compared to extreme phenomena we observe in socio-economic systems, such as financial bubbles and economic crises (Roehner and Sornette (2000)) or major paradigm shifts in political and technological environments (Bikhchandani et al. (1992)).

Kaizoji (2000) constructs a model very similar to the ones we use in this thesis except for the time component, which is not included. He analyzes the equilibrium results for a standard Ising model in the Mean Field approximation and maps them to behaviour of investors in a stock market, having them react to two factors: the *investment environment*, namely the difference in yield between the risky asset and the safe asset, and the *bandwagon effect*, that is the chance the investor copies the trending behaviour of others, assuming both these effects are equal for all traders. The main result is the mapping of the Ising phase transition to the financial interpretation of a *bear* or *bull* market, namely time frames where the market

is (excessively) pessimistic or optimistic towards traded assets. The author then proceeds to fit the model to empirical monthly data from the infamous 1990-1992 Japan bear market, finding that the interplay between the two effects was highly predictive of the monthly TOPIX (TOkyo stock Price IndeX) variations in the 1987-1992 time frame.

In a further article, Kaizoji et al. (2002) extend the same Ising model to a setting with heterogeneous agents, namely where a fraction behaves as a *fundamentalist* ignoring others' behaviours and trade with a contrarian strategy, while others still interact between each other and take their strategic decisions based on knowledge of their neighbours' strategies and on whether they belong to the majority or minority group. The authors then proceed to model the price formation mechanism in their synthetic market, and run simulations where they reproduce volatility clustering, log-returns fat tails and the intermittency of bull and bear markets, as was also shown in Bornholdt (2001).

Harras et al. (2012) propose a model which is closely related to the one we present in Chapter 5. There they simulate a Kinetic Ising Model with dynamical interaction strength, showing that when interpreting spins as traders that choose whether to buy or sell an asset and with a Walrasian price formation the autocorrelation of volatility has long memory, which is not found for log-returns, again a sign of volatility clustering.

Another similar example of this sort of modelling can be found in Kristoufek and Vosvrda (2018), where the authors investigate the dynamics of the model when local interactions among participants are affected by a global interaction generating a minority game (Challet et al. (2013)). This interplay produces a dynamics of the price that can diverge from its efficient value when herding is either too weak or to strong.

Ising models have also been applied in economic settings to analyse competition and technology adoption, as for instance in Biely et al. (2009), to describe segregation between communities in cities (Schelling (1971)) and models that are essentially equivalent to the Ising model have been used to describe social choice (Brock and Durlauf (1999)) and opinion dynamics (Sood and Redner (2005); Castellano et al. (2009); De Vincenzo et al. (2017)).

2.9 Collective phenomena in finance: herding

One prominent example of behaviour not dictated by the rationality of a representative agent is what Nietzsche called "herd behaviour" in his Untimely Meditations (Nietzsche (1997)) and in later works (Nietzsche (1974); Nietzsche and Common (1950)). By this expression he meant the concept that most people tend to behave by aligning to the mass, suppressing their individuality in favour of identifying with the majority. While his feelings about this behaviour were of absolute abhorrence, to the point that his writings became the basis for despicable ideologies later on, proofs of the existence of such behaviours can be found in a multitude of settings, including financial markets where "herding" has had a growing stream of literature since the inception of behavioral finance (Shiller (2015)).

In synthesis, herding is a collective behaviour expressed by a group of individuals who coordinate without centralized direction. The coordination can come from several mechanisms, as for example social pressure and imitation (Sood and Redner (2005)) or the aggregation of purely selfish mechanisms as the "selfish herds" of preys running from predators (Hamil-

ton (1971)). In finance it is typically defined as a trading activity that does not reflect the private information held by market participants due to their knowledge that others are trading too (Bikhchandani and Sharma (2000)). This can lead to abnormal price movements on the short-to-medium term. The most commonly studied events of this kind are bubbles and crashes, where market prices take irrationally large deviations from fundamental values due to collective speculation and suddenly revert to more reasonable quantities, often destroying large amounts of wealth in the process (Lux and Sornette (2002)).

One of the first examples of modelling for herd behaviour can be found in Banerjee (1992). There the author studies a game where each player can hold a piece of private information about what is the correct choice to make and players sequentially make their choice, which is public information for players acting later. Under mild assumptions on the choice strategy the equilibrium is largely dictated by the first few players, with all others following and ignoring their private information in what is called an *information cascade*. Indeed if the private information held by players has probability β of being correct and they hold such private information with probability α, there is a finite probability p that *none* of the players makes the correct choice

$$p = \frac{(1-\alpha)(1-\beta)}{1-\alpha(1-\beta)}$$

As a corollary, the author shows that introducing a mechanism that hides the choice of the first few players to others will get rid of herding behaviour: this result in particular has shaped some market designs where orders can be hidden for a limited amount of time.

Recently agent-based models have been increasingly popular in the the-

oretical investigation of herding. An example is the work of Alfarano et al. (2005), where they propose an agent-based model for herding behaviour by assuming traders can pick one of two strategies, the fundamentalist and the noise trading, and switch between the two based on a probabilistic rule which accounts for mean field interaction with others. They then construct a model market with Walrasian auction to determine equilibrium prices, leading to a skewed and fat-tailed volatility distribution which is also found for daily financial data.

In Barde (2016) the methodology of Alfarano et al. (2005) is used to model price evolutions and compared to the performance of the more commonly adopted GARCH model (Bollerslev (1986)), showing that having a more detailed model for the dynamics generating the price allows to outperform the GARCH around extreme events.

In terms of empirical works, Lakonishok et al. (1992) ask the question whether institutional investors "herd" more than individual investors and whether their behaviour destabilizes asset prices, leading them away from their equilibrium values. They empirically evaluate trading patterns of actively managed pension funds based on quarterly data and find that institutional herding exists but is not destabilizing the prices, statistically defining herding as

$$H_t = \left| \frac{B_t}{B_t + S_t} - p_t \right| - AF_t$$

where B_t is the number of (net) buying funds on a given asset at time t, S_t is the number of (net) selling funds on the same asset, p is the expectation of $B/(B+S)$ under the null hypothesis that the fraction of buyers is equal for all stocks in the market and AF is an adjustment factor that rescales the measure around 0.

This sort of evidence is also found by Grinblatt et al. (1995), however there the authors introduce a more refined herding measure based on funds activity accounting for the sign of trades, which shows a more generalized herding behaviour of fund managers, particularly the ones that focus on momentum strategies. They do so by defining a herding measure for the single fund which is the interaction of the measure of Lakonishok et al. (1992) with a signed variable, which indicates whether the fund is trading "with the herd" or "against" it.

In a further study, Grinblatt and Keloharju (2000) analyze a detailed dataset of daily ownership of stocks in the Helsinki Stock Exchange. They categorize each market participant by its overall trading strategy, finding that most private household investors are contrarians and most foreign investors are momentum-driven, while Finnish institutional investors are mostly contrarians but in a less marked way with respect to households. However they don't find performance differences between the categories that match the strength of behavioral differences, meaning that institutional investors are not behaving "more rationally" than households by being more sophisticated than simply buying losers and selling winners.

Nofsinger and Sias (1999) investigate whether trading by institutional investors is more or less related to price returns than herding of individuals and if this is due to a common "feedback trading" strategy (i.e. traders identically react to a common signal which are price returns) or to other sorts of herding. They find that institutional trading has significantly higher correlation with returns, consistent with the hypothesis of institutions engaging in feedback trading more than individuals or that herding from institutions impacts the price more. The authors do not disentangle the two effects, which amounts to determining the direction of causality between the obser-

vations, for which they do not find sufficient evidence to make conclusive claims.

In a similar spirit, Lillo et al. (2008) study the trading behaviour of a set of investors on the Spanish stock market. There the authors categorize investing strategies based on an eigenvector decomposition of the correlation matrix of inventory variations, finding that the principal component of said matrix is closely related to returns. This result allows to define three investor categories as trending, reversing or neutral. They then check for herding within these categories, using herding measures similar to Lakonishok et al. (1992) and Grinblatt and Keloharju (2000), and find that traders in the reversing category herd significantly more than others, both on relatively short timescales (15 minute) and on the daily scale, mostly in a feedback trading fashion as testified by a Granger Causality analysis.

A recent study by Cai et al. (2019) finds that institutional traders are even more subject to herding in the corporate bond market, and that the effect of this behaviour is asymmetric depending on whether it is a "buy herding" or a "sell herding". In the first case they show that herding when buying fixed-income securities is beneficial to price discovery, generating a more permanent price impact, while on the other hand herding of sellers generates a large transitory price impact, signifying that it is distorting the price dynamics beyond equilibrium levels. A similar analysis is performed by Galariotis et al. (2016) on the European governments bond markets, finding that herding is exacerbated by macroeconomic announcements during the Euro crisis of 2010-2011.

A different take on herding measures can be found in Toth et al. (2015), where the authors look at the aggregate effect of herding by splitting the autocorrelation of the order flow at high frequency in two components, one

dictated by order-splitting (the common practice of diluting in time a large trade to optimize transaction costs) and the other dictated by herding. They propose two methodologies to perform the decomposition, one based on an Ising model itself, and they show that herding is not a factor in the determination of the autocorrelation of the order flow on short timescales.

Recent efforts have been directed to identifying herding phenomena based on the refinement of automated text analysis techniques. It is the case of Palmer et al. (2018), where the authors adopt topic modelling techniques to identify similarities between financial analysts reports. They find that these similarities increased during the financial crisis of 2008, but the degree to which this is due to herding rather than common fundamental information is still unclear.

2.10 Going micro to understand macro: investors networks

As the amount and quality of available market data increased in the last decade, a number of more and more refined methods to analyze the microstructure of the markets have been developed. What is typically known as market microstructure is the set of rules and actors that are the building blocks of modern financial markets and the study of how the design of a market affects the price formation process and trading strategies. One of the recently popular topics in this branch of finance has been the reconstruction of investors networks, that are networks where nodes represent traders and links some form of interaction between them.

Estimating such networks clearly requires an amount of detail in the data that is not easy to find, as no modelling of this kind is possible if the

scientist has no access to single traders actions, and this explains why this sort of analysis is relatively recent. Indeed with the advent of automated and digitalized markets it has been much easier to register huge quantities of transactions at a relatively low cost, making it possible to keep track of all the events that happen in the market up to the finest detail of single limit orders in an order book.

One of the first examples of investors networks is found in Tumminello et al. (2012), where the authors analyse through the methodology of Statistically Validated Networks (SVN, Tumminello et al. (2011)) the behaviour of a very heterogeneous set of traders in the Finnish stock market on a daily timescale. In their work they identify a network where nodes are traders (both institutional and retail) and links represent a tendency to trade together on a given stock which cannot be explained by the null hypothesis of trading on random days (keeping the trading frequency constant). They then provide a cluster analysis of the SVNs, for which they find that investors belonging to the same category (household, financial institution, foreign, government, non-profit, insurance) are indeed more likely to trade together. The same method was later adopted by Curme et al. (2015) for the investigation of lead-lag networks between financial assets at high frequency and by Musciotto et al. (2018) to study how the investors networks evolve in time, finding that in times of higher volatility these networks become more heterogeneous, that is the number of groups that use different strategies is larger, an observation that was predicted by Farmer (2002).

Another example of SVN applied for the inference of synchronous and lagged relationships among traders can be found in Challet et al. (2018), where first the authors apply the SVN methodology to identify clusters of investors in the Foreign Exchange market that show high synchronicity at

the 1 hour timescale and then construct a lead-lag SVN, this time treating the clusters as a representative agent. In this way they are able to analyze whether there are trading strategies that consistently anticipate others, and they show that the order flow can be predicted with good accuracy.

Departing from SVNs one recent work by Gutiérrez-Roig et al. (2019) uses Information Theoretic methods such as the Mutual Information and the Symbolic Transfer Entropy to quantify synchronicity and lead-lag relations among investors in the Spanish stock market on the daily timescale. Similarly to Challet et al. (2018) they resort to machine learning methods to nowcast and forecast traders activity based on their networks, finding that nowcasting is improved by the mutual information network for some activity patterns and that forecasting is marginally improved by the transfer entropy network.

In Chapter 4 we propose to use the Kinetic Ising Model as an alternative method to study investors networks, specifically targeted at identifying possible high-frequency herding phenomena that can impact the liquidity supply in the market.

Chapter 3

Kinetic Ising model and missing data

Almost all results in this chapter previously appeared in Campajola et al. (2019)

3.1 Introduction

As we have shown in Chapter 2, the Kinetic Ising Model (Derrida et al. (1987); Crisanti and Sompolinsky (1988)) is a Maximum Entropy model describing a set of binary units - named "spins" in the physics literature - that influence each other through time. The simplicity of the model makes it extremely flexible in the kinds of systems it can represent, ranging from networks of neurons in the brain (Capone et al. (2015)) all the way to traders in a financial market (Bornholdt (2001); Sornette (2014)). Recent work on the inference of the Kinetic Ising Model has led to the development of exact solutions (Sakellariou (2013)), cavity methods (Zhang (2012)) and Mean Field (Roudi and Hertz (2011a)) techniques for the inference of the

parameters, and the latter have been used to work with partially observed systems linking to the realm of (Semi-) Restricted Boltzmann Machines (Dunn and Roudi (2013)).

This latest stream of literature sparked our interest for the model applied to time series of financial data at high frequency, where we typically encounter problems related to the lack of homogeneously frequent and synchronized observations (Aït-Sahalia et al. (2010); Buccheri et al. (2020); Corsi et al. (2012)).

The literature on Kinetic Ising Model has previously considered mainly the inference problem in the presence of hidden nodes (Dunn and Roudi (2013)), i.e. part of the spins are *never* observed, but it is known that they exist and interact with the visible ones. This setting is of particular interest in neuroscience where an experiment typically monitors the firing activity of a subset of neurons. In other domains, such as in economics, finance, and social sciences, another type of missing data is often present, namely the case where even for the visible agents (nodes), observations are missing a significant fraction of the times. Moreover in these cases there is a strong heterogeneity of the frequency of observations, i.e. some nodes are frequently observed while other are rarely observed. There are different sources for this lack of data: in some cases, it might be due to the fact the observation is costly for the experimenter, whereas in other cases it is intrinsic to the given problem. Consider, for example, the problem of inferring the opinion of investors from their trading activity. When an investor buys (sells) it is reasonable to assume that she believes the price will increase (decrease), but in many circumstances the investor will not trade leading to missing observations for her belief. Using a suitable inference model, as the one proposed in this paper, it is possible to estimate her

belief from the inferred structure of interaction among investors and the observed state of the set of visible ones. We will also include external fields (for example the market price in the previous example) that can influence spins (investors' opinion).

Missing data is a common problem in many fields of science, and several techniques have been developed to overcome this issue. Starting with the historical paper by Rubin (1976), the interest for the problem has grown and different kinds of deletion (Buhi et al. (2008)), imputation (Rubin (2004)) and estimation (Dempster et al. (1977); Marlin and Zemel (2009); Mohan et al. (2013)) methods have been developed, each answering questions for specific classes of missing data problems. Our contribution fits in the family of Maximum Likelihood estimators and the Expectation-Maximization (EM) method, which has been proved to provide bias-free estimates as long as the data are missing at random by Little and Rubin (2019).

Taking inspiration from the work by Dunn and Roudi (2013), we extend the formulation of the inference procedure to cases where the missing observations are unevenly cross-sectionally distributed, meaning that time series are sampled at a constant rate and whenever no observations are found between two timestamps a missing value is recorded. The result is an algorithm closely related to an Expectation-Maximization (EM) method (Dempster et al. (1977)), iteratively alternating a step of log-likelihood gradient ascent (Nesterov (2008)) and the self-consistent resolution of TAP equations (Roudi and Hertz (2011a)), that gives as output both a coupling matrix and an approximated maximum-likelihood estimate of the missing values.

To evaluate the algorithm performance we devise a series of tests stressing on different characteristics of the input, simulating synthetic datasets

with several regimes of intrinsic noise, observation frequency, heterogeneity of variables and model misspecification. We thus define some performance standards that can be expected given the quality of data fed to the method, giving an overview of how flexible the approach is.

3.2 Solving the Inverse Problem with missing values

The Kinetic Ising Model (or non-equilibrium Ising Model) (Derrida et al. (1987)) is defined on a set of spins $y \in \{-1, +1\}^N$, whose dynamics is described by the transition probability mass function

$$p[y(t+1)|y(t)] = \\ = Z^{-1}(t) \exp\left[\sum_{\langle i,j \rangle} y_i(t+1) J_{ij} y_j(t) + \sum_i y_i(t+1) h_i\right] \tag{3.1}$$

where $\langle i, j \rangle$ is a sum over neighbouring pairs on an underlying network, J_{ij} are independent and identically distributed couplings, h is the vector of spin-specific fields and $Z(t)$ is a normalizing constant also known as the partition function.

In our treatment of the problem we will adopt a Mean Field (MF) approximation, which relies on the assumption that the dynamics of a spin i depends only on an effective field locally "sensed" by the spin rather than on the sum of the single specific interactions with others. The result of this picture is that the topology of the underlying network is considered irrelevant and assumed fully connected - although the goal of the inference would be the reconstruction of the network nonetheless - thus the sum on

neighbours is substituted by a sum on all the other spins. This recasts the transition probability into the following form

$$p[y(t+1)|y(t)] = Z^{-1}(t) \exp\left[\sum_{i=1}^{N} y_i(t+1)\tilde{g}_i(t)\right] \tag{3.2}$$

where $\tilde{g}_i(t) = \sum_{j=1}^{N} J_{ij} y_j(t) + h_i$ is the local effective field of spin i and J is now a square and fully asymmetric matrix with normally distributed entries $J_{ij} \sim \mathcal{N}(0, J_1^2/N)$, where the assumption on the distribution and the scaling of the variance with N^{-1} will be necessary in the forthcoming calculations.

Consider observing only a fraction $M(t)/N$ of spins at each time step, and define $G(t)$ as the $M(t) \times N$ matrix mapping the configuration $y(t)$ into the observed vector $s(t) \in \{-1, 1\}^{M(t)}$. Also define $F(t)$ as the $(N - M(t)) \times N$ matrix mapping $y(t)$ into the unobserved spins vector $\sigma(t) \in \{-1, 1\}^{N-M(t)}$. We require that both matrices are right-invertible at all t, thus they must have full rank, that implies that observations are not linear combinations of the underlying variables as our interest is in a partially observed system rather than a low-dimensional observation of a high-dimensional system. For the sake of simplicity we assume that the entries are either 0 or 1, meaning observation is not noisy or distorted and the right-inverse matrices will coincide with the transpose.

In the upcoming calculations we will use some simplifying custom notation in order to reduce what can be some cumbersome equations. We will thus denote $\sum'_i$ the sum over indices i at time $t+1$, while the regular $\sum_i$ indicates a sum over indices i at time t and $\sum^-_i$ a sum at time $t-1$. Accordingly, we will indicate with s_i spin i at time t, with s_i^- at time $t-1$ and with s_i' at time $t+1$, and the same applies for g, σ and any other variable. Also indices i, j, k, l are used for observed variables, whereas indices

a, b, c, d will identify unobserved variables.

In this notation, the probability mass function is rewritten as

$$p[\{s', \sigma'\}|\{s, \sigma\}] = Z^{-1} \exp\left[\sum_i{}' s'_i g'_i + \sum_a{}' \sigma'_a g'_a\right] \tag{3.3}$$

Defining the matrices $J^{oo}(t+1) = G(t+1)JG^T(t)$, $J^{oh}(t+1) = G(t+1)JF^T(t)$, $J^{ho}(t+1) = F(t+1)JG^T(t)$ and $J^{hh}(t+1) = F(t+1)JF^T(t)$ the local fields are

$$\begin{aligned} g_i &= \sum_j J^{oo}_{ij} s^-_j + \sum_b J^{oh}_{ib} \sigma^-_b + h_i \\ g_a &= \sum_j J^{ho}_{aj} s^-_j + \sum_b J^{hh}_{ab} \sigma^-_b + h_a \end{aligned} \tag{3.4}$$

and the partition function or normalization constant is

$$Z = \prod_{i,a}{}' 2\cosh(g'_i) 2\cosh(g'_a)$$

The ultimate purpose of this work is to devise an approximate method to obtain Maximum Likelihood Estimates (MLE) for the parameters J, h and the unobserved spins σ. The likelihood function is just the product through time of the independent transition probabilities expressed in Eq. 3.3, taking the trace over the missing values

$$p[\{s\}] = \mathrm{Tr}_\sigma \prod_t p[\{s', \sigma'\}|\{s, \sigma\}] \tag{3.5}$$

To solve the problem, our approach is closely related to the one developed by Dunn and Roudi (2013), where the authors investigate on a system where only a subset of spins is observable. The extension to our case is presented below.

The trace of Eq. 3.5 is computationally intractable for large systems with many hidden variables. However the path integral formulation first proposed by Martin et al. (1973) allows to decouple spins and perform the trace at the cost of computing a high dimensional integral. Define the functional

$$\mathcal{L}[\psi] = \log \mathrm{Tr}_\sigma \prod_t \exp\left[\sum_a \psi_a \sigma_a\right] p[\{s', \sigma'\}|\{s, \sigma\}] \tag{3.6}$$

Notice that this is equivalent to the log-likelihood if $\psi_a(t) = 0 \; \forall a, t$, thus the goal of the calculation will be to efficiently maximise $\mathcal{L}[\psi]$ in the J, h coordinates considering the limit when $\psi \to 0$. As will become clear in the next steps, the introduction of these so-called "auxiliary fields" is necessary to switch from the unknown values σ to their posterior expectations m, thus smoothing the log-likelihood function eliminating unknown binary variables from its formula. Call

$$\begin{aligned}
Q[s, \sigma] &= \sum_t \sum_i s_i g_i + \sum_t \sum_a \sigma_a g_a + \\
&\quad - \sum_t \sum_i \log 2\cosh(g_i) - \sum_t \sum_a \log 2\cosh(g_a) \\
\Delta &= \sum_t \sum_i i\hat{g}_i \left[g_i - \sum_j J^{oo}_{ij} s^-_j - \sum_b J^{oh}_{ib} \sigma^-_b - h_i \right] + \\
&\quad + \sum_t \sum_a i\hat{g}_a \left[g_a - \sum_j J^{ho}_{aj} s^-_j - \sum_b J^{hh}_{ab} \sigma^-_b - h_a \right]
\end{aligned}$$

where e^Δ, integrated over the $\hat{g}$s is the integral representation of the Dirac delta function. Then one obtains

$$\mathcal{L}[\psi] = \log \int \mathcal{D}\mathcal{G} \exp[\Phi] \tag{3.7}$$

where $\mathcal{G} = \{g_i, g_a, \hat{g}_i, \hat{g}_a\}_t$ and

$$\Phi = \log \mathrm{Tr}_\sigma \exp \left[Q + \Delta + \sum_t \sum_a \psi_a \sigma_a \right] \tag{3.8}$$

Now the trace can be easily computed since the introduction of the delta function has decoupled the σs by fixing the value of the local fields g, obtaining

$$\begin{aligned}
\Phi = \sum_t \Bigg[& \sum_i \left[s_i g_i - \log 2\cosh(g_i) \right] - \sum_a \log 2\cosh(g_a) + \\
& + \sum_i i\hat{g}_i \left[g_i - \sum_j J^{oo}_{ij} s^-_j - h_i \right] + \\
& + \sum_a i\hat{g}_a \left[g_a - \sum_j J^{ho}_{aj} s^-_j - h_a \right] + \\
& + \sum_a \log 2\cosh \left[g^-_a - \sum_i i\hat{g}_i J^{oh}_{ia} - \sum_b i\hat{g}_b J^{hh}_{ba} + \psi^-_a \right] \Bigg]
\end{aligned}$$

As mentioned, the cost is computing the integral of Eq. 3.7, which can be solved via the saddle-point approximation, where the saddle-point is obtained by the extremization of Φ with respect to the coordinates in $\mathcal{G}$. Setting $\nabla_{\mathcal{G}} \Phi = 0$ gives

$$\begin{aligned}
g^0_i &= h_i + \sum_j{}^- J^{oo}_{ij} s^-_j + \sum_a{}^- J^{oh}_{ia} m^-_a \\
g^0_a &= h_a + \sum_j{}^- J^{ho}_{aj} s^-_j + \sum_b{}^- J^{hh}_{ab} m^-_a \\
i\hat{g}^0_i &= \tanh(g_i) - s_i \\
i\hat{g}^0_a &= \tanh(g_a) - m_a
\end{aligned}$$

which, substituted in Φ, give the zero-order solution to the saddle-point integral.

The missing part of the puzzle is the posterior mean $\mathbb{E}[\sigma_a(t)]$, for which $\mathcal{L}$ acts as the generating functional

$$\mathbb{E}[\sigma_a(t)] = m_a(t) = \lim_{\psi_a(t)\to 0} \mu_a(t) = \lim_{\psi_a(t)\to 0} \frac{\partial \mathcal{L}}{\partial \psi_a(t)}$$

where the expectation is performed under the posterior measure $p[\{\sigma\}|\{s, J, h\}]$. Thus we find

$$\lim_{\psi_a\to 0} \frac{\partial \mathcal{L}}{\partial \psi_a} = m_a = \tanh\left[g_a^0 - {\sum_i}' i\hat{g}_i^{0\prime} J_{ia}^{oh} - {\sum_b}' i\hat{g}_b^{0\prime} J_{ba}^{hh} \right]$$

This zero-order approximation is rather rough, nonetheless the saddle-point method can be solved at higher orders of approximation. The second-order (*i.e.* Gaussian) correction to the saddle point solution of the integral in Eq. 3.7 is

$$\delta\mathcal{L} = -\frac{1}{2}\log\det[\nabla^2_{\mathcal{G}}\mathcal{L}]$$

where $\nabla^2_{\mathcal{G}}\mathcal{L}$ is the Hessian matrix in the $\mathcal{G}$ space of $\mathcal{L}$ evaluated at the saddle point. This is a forbidding task to tackle numerically, since the matrix has $(4NT)^2$ elements, but with a few algebraic manipulations the computations become feasible.

The Hessian matrix elements can be summarized in the following sub-matrices $A^{tt'}, ..., G^{tt'}$, given by

$$
\begin{aligned}
\frac{\partial^2\Phi}{\partial g_i(t)\partial g_j(t')} &= A_{ij}^{tt'} = -\delta_{ij}\delta_{tt'}(1-\tanh^2[g_i^0(t)]) \\
\frac{\partial^2\Phi}{\partial \hat{g}_i(t)\partial \hat{g}_j(t')} &= B_{ij}^{tt'} = -\delta_{tt'}\sum_a{}^{-} J_{ia}^{oh}(t)J_{ja}^{oh}(t)[1-\mu_a^2(t-1)] \\
\frac{\partial^2\Phi}{\partial g_a(t)\partial g_b(t')} &= C_{ab}^{tt'} = -\delta_{ab}\delta_{tt'}\left[\mu_a^2(t)-\tanh^2[g_a^0(t)]\right] \\
\frac{\partial^2\Phi}{\partial \hat{g}_a(t)\partial \hat{g}_b(t')} &= D_{ab}^{tt'} = -\delta_{tt'}\sum_c{}^{-} J_{ac}^{hh}(t)J_{bc}^{hh}(t)\left[1-\mu_c^2(t-1)\right] \\
\frac{\partial^2\Phi}{\partial \hat{g}_i(t)\partial \hat{g}_b(t')} &= E_{ib}^{tt'} = -\delta_{tt'}\sum_a{}^{-} J_{ia}^{oh}(t)J_{ba}^{hh}(t)\left[1-\mu_a^2(t-1)\right] \\
\frac{\partial^2\Phi}{\partial \hat{g}_i(t)\partial g_b(t')} &= F_{ib}^{tt'} = -i\delta_{t-1,t'}J_{ib}^{oh}(t)\left[1-\mu_b^2(t-1)\right] \\
\frac{\partial^2\Phi}{\partial g_a(t)\partial \hat{g}_b(t')} &= \delta_{ab}\delta_{tt'} + G_{ab}^{tt'} = \delta_{ab}\delta_{tt'} - i\delta_{t+1,t'}J_{ba}^{hh}(t+1)\left[1-\mu_a^2(t)\right] \\
\frac{\partial^2\Phi}{\partial g_i(t)\partial \hat{g}_j(t')} &= \delta_{ij}\delta_{tt'} \\
\frac{\partial^2\Phi}{\partial g_i(t)\partial g_b(t')} &= \frac{\partial^2\Phi}{\partial g_i(t)\partial \hat{g}_b(t')} = 0 \qquad \forall\, t,t',i,b
\end{aligned}
$$

and in matrix form it has the following almost block-diagonal form (we show the sub-matrix for times $t, t+1$)

$$
\left[\begin{array}{cccc|cccc}
A^{tt} & i\mathbb{I} & 0 & 0 & 0 & 0 & 0 & 0 \\
i\mathbb{I} & B^{tt} & 0 & E^{tt} & 0 & 0 & 0 & 0 \\
0 & 0 & C^{tt} & i\mathbb{I} & 0 & [F^{t+1,t}]^T & 0 & G^{t,t+1} \\
0 & [E^{tt}]^T & i\mathbb{I} & D^{tt} & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & A^{t+1,t+1} & i\mathbb{I} & 0 & 0 \\
0 & 0 & F^{t+1,t} & 0 & i\mathbb{I} & B^{t+1,t+1} & 0 & E^{t+1,t+1} \\
0 & 0 & 0 & 0 & 0 & 0 & C^{t+1,t+1} & i\mathbb{I} \\
0 & 0 & [G^{t,t+1}]^T & 0 & 0 & [E^{t+1,t+1}]^T & i\mathbb{I} & D^{t+1,t+1}
\end{array}\right]
$$

We are interested in the determinant, and in particular its logarithm. Dividing the Hessian in the matrices α containing block-diagonal elements and β containing the rest, we find

$$\log\det(\alpha+\beta) = \log\det(\alpha)+\log\det[\mathbb{I}+\alpha^{-1}\beta] = \log\det(\alpha)+\mathrm{Tr}\log[\mathbb{I}+\alpha^{-1}\beta] \approx$$
$$\approx \log\det(\alpha) + \mathrm{Tr}[\alpha^{-1}\beta] + \frac{1}{2}\mathrm{Tr}\{[\alpha^{-1}\beta]^2\} + ... \quad (3.9)$$

Given that α is block-diagonal, so will be α^{-1}, then $\mathrm{Tr}[\alpha^{-1}\beta] = 0$ and we ignore higher order terms assuming the off-diagonal part of the Hessian matrix is small compared to the diagonal one. In our initial assumption, the couplings J_{ij} are Gaussian random variables with mean of order $1/N$ and variance of order J_1^2/N, which means $\log\det(\alpha)$ is quadratic in J_1. The determinant now can be computed and a weak couplings expansion (i.e. $J_1 \to 0$) can be made to eliminate the logarithm, leading to the final approximate form of the correction

$$\delta\mathcal{L} \approx -\frac{1}{2}\sum_t{}' \sum_i{}' \left[\left(1-\tanh^2(g_i')\right)\sum_b \left[J_{ib}^{oh\prime}\right]^2 (1-\mu_b^2)\right] +$$
$$-\frac{1}{2}\sum_t{}' \sum_a{}' \left[\left(\mu_a'^2-\tanh^2(g_a')\right)\sum_b \left[J_{ab}^{hh\prime}\right]^2 (1-\mu_b^2)\right]$$

Given the new form of $\mathcal{L}_1 = \mathcal{L}_0 + \delta\mathcal{L}$, we need to recalculate the self-consistency relation for $m_a(t)$ and the learning rule for J. As for $m_a(t)$, we can easily see that it is going to coincide with $m_a(t) = \lim_{\psi_a(t)\to 0}\mu_a(t)+l_a(t)$, where

$$
\begin{aligned}
l_a(t) =& \frac{\partial(\delta\mathcal{L})}{\partial\psi_a(t)} = \\
=& \mu_a(1-\mu_a^2)\left[\sum_i{}' \left[\left(1-\tanh^2(g_i')\right)\left[J_{ia}^{oh\prime}\right]^2\right]\right] \\
&+\mu_a(1-\mu_a^2)\left[\sum_b{}^{-}\left[J_{ab}^{hh}\right]^2(1-\mu_b^{-2})+\right. \\
&\left.+\sum_b{}'\left(\mu_b'^2-\tanh^2(g_b')\right)\left[J_{ba}^{hh\prime}\right]^2\right] \qquad (3.10)
\end{aligned}
$$

Implementing the MSR method has introduced an explicit dependence of the $\mathcal{L}$ functional from the auxiliary fields $\hat{g}$ and ψ, which however make little sense in terms of the model itself. Now that we have solved the integral at the saddle-point and in its immediate neighbourhood the auxiliary fields can be absorbed back into the original variables by performing a Legendre transform of $\mathcal{L}$, exploiting the fact that $\mathcal{L}$ is convex and that we would rather have it depend on the conjugate field of ψ, that is μ. The transform is

$$
\Gamma[\mu] = \mathcal{L} - \sum_t\sum_a \psi_a(t)\mu_a(t) \; s.t. \; -\psi_a(t) = \frac{\partial\Gamma[\mu]}{\partial\mu_a(t)} \qquad (3.11)
$$

and so we can adopt Γ as the functional to be maximised in the learning process instead. At zero-order, this is easily found to be

$$
\begin{aligned}
\Gamma_0[\mu] = \sum_t\Bigg[&\sum_i{}'\left[s_i'g_i^{0\prime}-\log 2\cosh(g_i^{0\prime})\right]+ \\
&+\sum_a{}'\left[\mu_a'g_a^{0\prime}-\log 2\cosh(g_a^{0\prime})\right]+\sum_a S[\mu_a]\Bigg] \qquad (3.12)
\end{aligned}
$$

where $S[x] = -\frac{1+x}{2}\log(\frac{1+x}{2}) - \frac{1-x}{2}\log(\frac{1-x}{2})$ is the entropy of an uncoupled spin with magnetization x. It is relevant to mention that so far the

functional is expressed in terms of μ, while we have already highlighted that after the Gaussian correction a new term l is introduced in the formula for m. However, since we are restricting to second order in J, the terms containing l in Γ are all of superior order and are thus negligible in this approximation, then $\Gamma_0[m] \approx \Gamma_0[\mu]|_{\mu=m}$. Performing the exact same steps on the correction term $\delta\mathcal{L}$ one finds the corrected functional

$$\Gamma_1[m] = \Gamma_0[m] + \delta\mathcal{L}[m]$$

Γ_1 is the functional to be optimized through an Expectation-Maximization-like algorithm, recursively computing the self-consistent magnetizations m given J, h and then climbing the gradient $\nabla_{J,h}\Gamma_1$ to obtain a new J matrix and h vector.

The formulas necessary to the EM-like algorithm, namely the log-likelihood gradient and the self-consistent relations for the magnetizations, respectively read

$$\begin{aligned}
\frac{\partial\Gamma_1}{\partial J_{kl}} = \sum_t \Bigg[& \sum_i{}' \left[\frac{\partial g'_i}{\partial J_{kl}} \left(s'_i - \tanh(g'_i)\right)\right] + \\
& + \sum_a{}' \left[\frac{\partial g'_a}{\partial J_{kl}} \left(m'_a - \tanh(g'_a)\right)\right] + \\
& + \sum_i{}' \left[\frac{\tanh(g'_i)}{\cosh^2(g'_i)} \frac{\partial g'_i}{\partial J_{kl}} \sum_{bmn} G'_{im} J^2_{mn} F^T_{nb} (1 - m_b^2)\right] + \\
& + \sum_i{}' \left[-\left(1 - \tanh^2(g'_i)\right) \sum_b G'_{ik} J_{kl} F^T_{lb} (1 - m_b^2)\right] + \\
& + \sum_a{}' \left[\frac{\tanh(g'_a)}{\cosh^2(g'_a)} \frac{\partial g'_a}{\partial J_{kl}} \sum_{bmn} F'_{am} J^2_{mn} F^T_{nb} (1 - m_b^2)\right] + \\
& + \sum_a{}' \left[-\left(m_a^{2\prime} - \tanh^2(g'_a)\right) \sum_b F'_{ak} J_{kl} F^T_{lb} (1 - m_b^2)\right] \Bigg]
\end{aligned}$$

where the fields g and their derivatives are given by

$$
\begin{aligned}
g'_i &= \sum_j \sum_{kl} G'_{ik} J_{kl} G^T_{lj} s_j + \sum_b \sum_{kl} G'_{ik} J_{kl} F^T_{lb} m_b + h_i \\
g'_a &= \sum_j \sum_{kl} F'_{ak} J_{kl} G^T_{lj} s_j + \sum_b \sum_{kl} F'_{ak} J_{kl} F^T_{lb} m_b + h_a \\
\frac{\partial g'_i}{\partial J_{kl}} &= \sum_j G'_{ik} G^T_{lj} s_j + \sum_b G'_{ik} F^T_{lb} m_b \\
\frac{\partial g'_a}{\partial J_{kl}} &= \sum_j F'_{ak} G^T_{lj} s_j + \sum_b F'_{ak} F^T_{lb} m_b
\end{aligned}
$$

The self consistency equations for the magnetizations m are then obtained by imposing $\partial \Gamma_1 / \partial m_a(t) = 0$, finding

$$
\begin{aligned}
m_a = \tanh \Bigg[g_a + m_a \Bigg[& \sum_i{}' \left(1 - \tanh^2(g'_i)\right) \sum_{kl} G'_{ik} J^2_{kl} F^T_{la} + \\
& + \sum_b{}' \left(m_b^{2\prime} - \tanh^2(g'_b)\right) \sum_{kl} F'_{bk} J^2_{kl} F^T_{la} + \\
& - \sum_c{}^{-} \sum_{kl} F_{ak} J^2_{kl} F^{T\,-}_{lc} (1 - m_c^{2\,-}) \Bigg] + \\
& + \sum_i{}' \left(s'_i - \tanh(g'_i)\right) \sum_{kl} G'_{ik} J_{kl} F^T_{la} + \\
& + \sum_b{}' \left(m'_b - \tanh(g'_b)\right) \sum_{kl} F'_{bk} J_{kl} F^T_{la} + \\
& + \sum_i{}' \frac{\tanh(g'_i)}{\cosh^2(g'_i)} \sum_{oqb} G'_{io} J_{oq} F^T_{qb} (1 - m_b^{2\prime}) \sum_{kl} G'_{ik} J_{kl} F^T_{la} + \\
& + \sum_c{}' \frac{\tanh(g'_c)}{\cosh^2(g'_c)} \sum_{oqb} F'_{co} J_{oq} F^T_{qb} \left(1 - m_b^{2\prime}\right) \sum_{kl} F'_{ck} J_{kl} F^T_{la} \Bigg]
\end{aligned}
\tag{3.13}
$$

Once this approximate log-likelihood is maximized and the final iteration of the expectation part of the algorithm is finished, the result is an (approx-

imated) Maximum Likelihood Estimate of the couplings as well as a Maximum A Posteriori estimate of the hidden spins σ, given by $\hat{\sigma}(t) = \text{sign}(m_t)$.

Summarizing, the procedure is the following:

Algorithm

- Initialize J, h, $m(t)$
- Until convergence is reached
  - compute the self-consistent magnetizations $m(t)$
  - compute the gradient $\nabla_{J,h}\Gamma_1$
  - apply Gradient Ascent step, in our case Nesterov's II method proximal gradient ascent with backtracking line search
- Possibly involve LASSO ℓ_1-norm regularization or pruning techniques to obtain a sparse model.

3.3 Tests on synthetic data

We perform a series of tests on the algorithm in order to assess its performance in several diverse conditions of data availability. We particularly focus on how we select the observed spins and on the structure of the coupling matrix J in the data generating model. To construct the $G(t)$ and $F(t)$ matrices, we assign to each spin a probability p_i of being observed, meaning that $y_i(t)$ is observed with probability p_i for all t.

We explore how the performance of the inference depends on the following model specifications:

0. The average observation frequency, taking the Bernoulli probabilities

$p_i = p$, $\forall i = 1, \ldots, N$;

1. The heterogeneity of the Bernoulli probabilities p_i, which we choose to be distributed according to a Beta distribution $B(a(K), b(K))$ with given mean K and shape parameters a and b;

2. The scale J_1 of the J entries, which are distributed as $J_{ij} \sim \mathcal{N}(0, J_1^2/N)$;

3. The structure of the J matrix, specifically whether the underlying network is fully connected or an Erdős-Rényi random network of varying density, adopting either the LASSO ℓ_1 regularization (Tibshirani (1996)) or the decimation procedure of Decelle and Zhang (2015) to select the links;

4. The asymmetry of the J matrix. One of the key assumptions in the calculation is that $J_{ij} \neq J_{ji}$ and that they are independent and identically distributed, and we investigate how far one can violate it up to the case of a symmetric J matrix;

5. The dependency on the length of the time series relative to the number of units involved, T/N, to check the estimate asymptotic efficiency.

In Test 0 we study the performance of the algorithm in a very simple setting of missing information, where each variable has the same probability of being observed and the generating model is a fully-connected Kinetic Ising model. This is intended to study the effect the average amount of missing information in the sample has on the inference, without considering the possibility of having heterogeneous types of nodes. In this setting we also introduce a procedure we call Recursive E-M: by properly iterating the algorithm multiple times it allows to boost data artificially thus achieving

good performances even when the fraction of missing values is particularly high.

In Test 1 we explore the possibility that spins have heterogeneous observational properties. We sample the $\{p_i\}$ from a Beta distribution varying parameters to probe different levels of heterogeneity. The Beta distribution allows to range from a sharply peaked unimodal distribution to a sharply peaked bimodal distribution tuning the shape parameters α and β, while keeping the mean K constant: the former case is a situation of perfect homogeneity in the frequency of observations calling back to Test 0, while the latter is the extreme heterogeneity of having some units that are (almost) always hidden while the others are (almost) always observed. We select some intermediate cases to characterize how heterogeneity in observation frequency affects the identification of the model parameters.

Test 2 aims at assessing whether there is a minimal interaction strength to have the inferential process converging and how the approximations necessary to develop the method impact the accuracy of the inference. Indeed while J_1 in the physical model is proportional to the ratio between the strength of the magnetic coupling interaction and the temperature at which the system is observed, from a modelling perspective it is inversely proportional to the impact of the noise on the dynamics. Given the approximation of Eq. 3.9, if J_1 gets too large, the precision with which the parameters are identified should get worse. We thus expect to find an optimal region for the inference to be accurate, bounded from below by an identifiability threshold and from above by the limit of validity of the expansion.

In Test 3 we pursue the goal of making the methodology useful for real world scenarios, where it is highly unlikely that all spins interact among themselves and the underlying network is probably sparse. We compare

the performance of two well established techniques, the LASSO ℓ_1 regularization and the decimation procedure, and explore how these two methods perform paired with our algorithm by simulating data on a set of Erdős-Rényi random networks with different densities.

In a similar spirit, in Test 4 we study how the i.i.d. assumption made in Eq. 3.9 affects the performance in situations where coupling coefficients are pairwise correlated or even symmetric, a condition we envision to be more realistic in social and economic environments (Squartini et al. (2013)). We vary the correlation parameter $\mathrm{Cor}(J_{ij}, J_{ji}) = \rho$ for $i \neq j$ between 0 and 1, with the symmetric case being also of special interest because the model transforms into a dynamical form of the Sherrington-Kirkpatrick model, thus connecting to the extensive literature on the topic.

Finally, a sanity check is made in Test 5 by looking at the dependency of performance metrics on the ratio T/N, that is the ratio between the number of observations and the number of spins, to characterize the convergence rate of the estimator towards the true value and its consistency.

We test the algorithm and evaluate the performance using mainly two metrics, one relative to the reconstruction of the couplings and one to the reconstruction of missing values:

1. The Root Mean Square Error (RMSE) on the elements of the matrix J, $\mathrm{RMSE} = \sqrt{\langle (\hat{J}_{ij} - J_{ij})^2 \rangle_{ij}}$, suitably rescaled when comparing experiments with different J_1;

2. The "Reconstruction Efficiency" (RE), namely the fraction of spins that are correctly guessed among the hidden ones averaged throughout the time series, or $\mathrm{RE} = \langle \frac{1}{N - M(t)} \sum_a \delta_{\hat{\sigma}_a(t), \sigma_a(t)} \rangle_t$ where $\hat{\sigma}_a(t)$ is the sign of the self-consistent magnetization $m_a(t)$ calculated using the

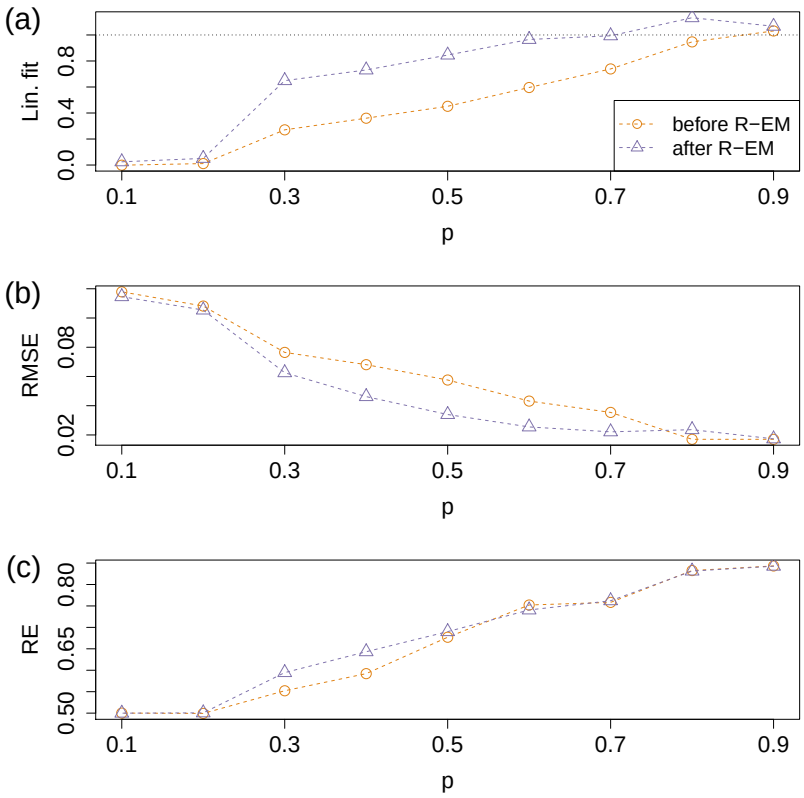

Figure 3.1: (a) Angular coefficient of the linear fit $\hat{J}_{ij} = aJ_{ij} + c$ before and after R-EM varying the average observation density p; (b) Root Mean Squared Error on the couplings; (c) Reconstruction Efficiency.

inferred coupling matrix $\hat{J}$.

Test 0: dependency on a homogeneous p_i

The algorithm is outstandingly resilient to cases with few observations available. We simulate a system of $N = 100$ spins, for $T = 10000$ time steps, with $J_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/N)$ lying on a fully connected network and we give a probability of observation to each variable $p_i = p$, with p ranging from 0.1 to 0.9. As can be seen from the top panel of Figure 3.1, showing the linear regression coefficient a of $\hat{J}_{ij} = aJ_{ij} + c$, with one iteration of the method we

get a very reliable result for the couplings for $p \geq 0.8$, although below this value the lack of data reduces the quality of the estimation and moves the estimates towards 0. To overcome this issue, we propose the aforementioned R-EM procedure as a further enhancement of our algorithm: once a maximum of the approximate likelihood has been reached, a fraction of hidden spins is substituted with their maximum likelihood estimates $\hat{\sigma}_a = \text{sign}(m_a)$ and the inference is run again on the new, artificially boosted data. Since m is proportional to the probability of the spin being up, we choose the missing values to be substituted at every t as the ones with the most polarized magnetization, *i.e.* for which m is closer to ± 1. This artificial boosting on the data shows promising results since with a few recursions the performance is noticeably better even in cases with severe lack of observations, as is also reflected in the middle and bottom panels of Figure 3.1. We defer a more rigorous treatment of this recursive method to future work, while still proposing it here as we find it surprisingly accurate.

The bottom panel of Figure 3.1 shows the Reconstruction Efficiency, which gets worse almost linearly as the number of observations decreases and on which the R-EM has a smaller effect, albeit still being a clear improvement. It is evident from all panels that when a large fraction of data is missing ($p \leq 0.2$) the inference fails to identify any of the parameters and the model is no better than a coin flip at reconstructing configurations.

In the following paragraphs we will always show results obtained with the R-EM procedure, as the performance is typically better or not significantly different from the single iteration method.

Test 1: heterogeneous p_i

In Test 1 we want to highlight how our model is a generalization of the one studied extensively by Dunn and Roudi (2013) and to characterize the impact of heterogeneity on the inference performance. To give a better comparison with the aforementioned paper, we realize simulations morphing from our initial specification of $p_i = p \ \forall i$, studied in Test 0, to a case very close to the one of Dunn et al. where $p_i \in \{0, 1\}$, that is some variables are always observed and some are always hidden. We choose to take the probabilities distributed according to a Beta distribution, $p_i \sim B(a(K), b(K))$, giving us the possibility of leaving the average number of observations constant while skewing the distribution between a fully bimodal (small $b(K)$) and a sharp quasi-delta function (large $b(K)$). We choose the parameters a and b such that the mean $\mathbb{E}[p_i] = K$ is constant, so that different tests can be compared and the role of heterogeneity is highlighted. This binds the values of a and b through $a = \frac{Kb}{1-K}$.

The results of Figure 3.2 clearly show that when the distribution is bimodal, that is when some variables are very rarely observed, the performance of the algorithm is worse. With a sample size of $T = 10^4$ and $N = 40$, the Dunn et al. model approximated by $B(a(K), 0.1)$ is identified with reasonable performance only when $K \geq 0.8$. This is extremely mitigated when the observations are more homogeneously distributed, particularly in the case of the coupling coefficients whose estimation seem to require a rather homogeneous distribution of observations among variables to be reliable. On the other hand, the reconstruction efficiency is far less demanding in terms of data quality and a reasonable performance is achieved even with sparse data and heterogeneous observations.

In Figure 3.3 we plot the Root Mean Square Error on couplings condi-

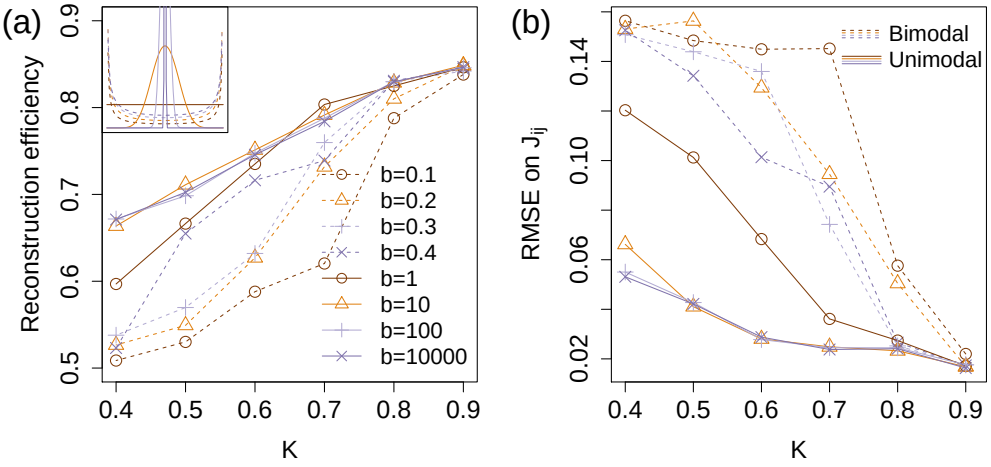

Figure 3.2: (a) Reconstruction efficiency as a function of K with different Beta parameters. Inset: the pdf of the adopted Beta distributions with $K = 0.5$ (color coding is the same as in the main panel) (b) Root Mean Square Error on the couplings as a function of K with different Beta parameters.

tional on the probability of observing subsequently the spins at their ends. This probability is simply given by $p_{ij} = p_i p_j$ since observations are independently sampled, and the RMSE is

$$\mathrm{RMSE}(p) = \sqrt{\langle (\hat{J}_{ij} - J_{ij})^2 \rangle_{p_{ij}=p}}$$

where the mean is taken on links that have (close to) the same joint observation probability. The plots highlight how for pairs with less frequent joint observations the precision of the fit is significantly worse, however it is also clear that the error grows for the more frequently observed couplings too. This is partially mitigated when one looks at the linear fit between the inferred Js and the true ones, meaning that the error is mostly affected by the variance component rather than the bias one.

The overall effect of heterogeneity is thus a decrease in the quality of the inference, with a stronger effect on couplings that are between the least

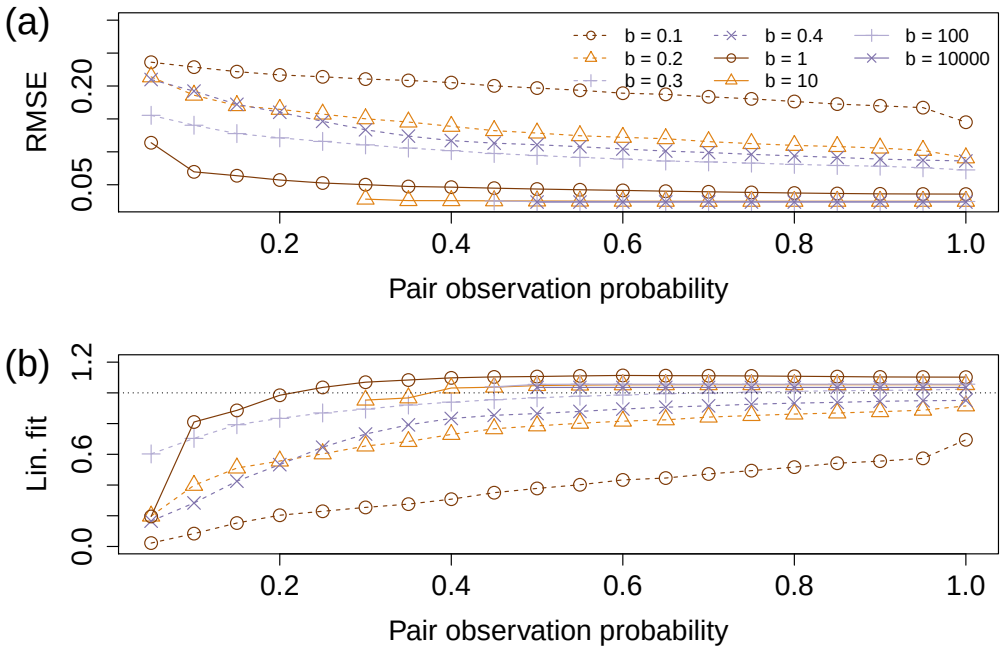

Figure 3.3: Quality of inference varying the probability of observing the end nodes at subsequent times. (a) RMSE for different values of the Beta b parameter with mean $K = 0.7$; (b) Linear fit coefficient for different values of the b parameter, $K = 0.7$.

observed pairs of spins and an important loss in accuracy, but with a bias component that is mitigated for the most frequently observed pairs.

Test 2: dependency on J_1

So far we have dealt with elements of J drawn i.i.d. from a $\mathcal{N}(0, 1/N)$ distribution. We want to relax this hypothesis and, while changing the mean value of the distribution would not be particularly meaningful in that it would just shift the correlation patterns between variables, it makes sense to investigate the behaviour as one changes the variance and thus the strength of the interactions. While there is no phase transition in the underlying model as long as the J_{ij} are i.i.d., we want to check how weak

can the couplings be in order to be correctly inferred and give a reliable reconstruction of the data. In other words, we are trying to identify a threshold in the interaction strength below which the algorithm is unable to converge.

We report results for an experiment with $N = 100$, $T = 10000$, $p_i = p = 0.8$ and J_1 ranging from 0.05 to 13. We see from Figure 3.4 that increasing the typical size of couplings positively affects the quality of the inference, as should be expected since the dynamics is less affected by randomness. In the top panel we plot the reconstruction efficiency which has a steady increase and saturates towards 1 after $J_1 \simeq 5$. The bottom panel shows the relative RMSE, that is RMSE$/J_1$, and we see that it drops below 5% for $J_1 > 0.5$. It is rather surprising to see how, regardless of the small couplings expansion we utilize in Eq. 3.9, the algorithm seems to work efficiently even in cases where the variance of the couplings J_1^2/N is of order 1, albeit a region of optimality for the inference of the couplings seems to lie within $0.5 \leq J_1 \leq 7$.

Test 3: impact of network structure

We test the algorithm performance on some more realistic network structure than the fully connected one. It is indeed known that real networks, and particularly social networks, are typically sparse and thus network models have to implement some pruning mechanism permitting to discriminate between noise, spurious correlations and actual causal relations. We generate our data simulating the Kinetic Ising model on one of the simplest random network models, the Erdős-Rényi model, with edges that have weights J_{ij} normally distributed with variance $1/N$, $N = 100$ and $T = 10000$ and with a probability of observing the variables of $p \in \{0.8, 0.6, 0.4\}$. One then

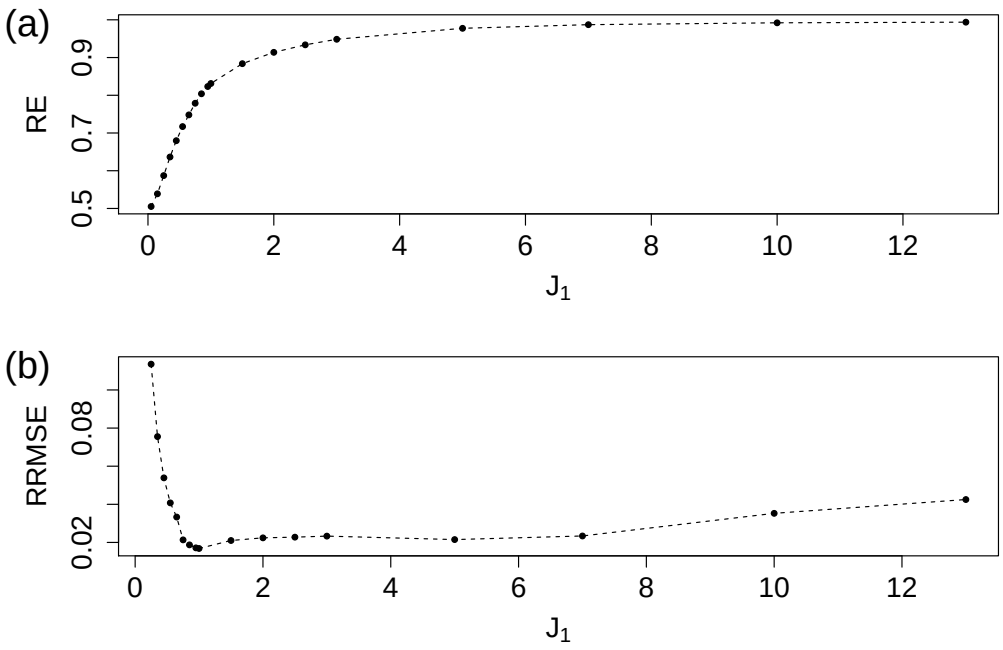

Figure 3.4: (a) Reconstruction Efficiency as a function of J_1. (b) Rescaled RMSE (by J_1) on the couplings as a function of J_1.

needs to adjust the algorithm to give sparse solutions, as the mean field approximation will tend to return fully connected J matrices. The adjustments we make are the LASSO regularization and the decimation procedure of Decelle and Zhang (2015). The first is the well known ℓ_1 norm regularization of the objective function, which projects the maximum likelihood fully connected solution on a symplex of dimensions determined by a free parameter λ (which has to be validated out of sample).

The second is a recently proposed technique that selects parameters starting to decimate them from the least significant ones and repeating the process until a so-called Tilted log-Likelihood function shows a discontinuity in the first derivative.

To briefly describe the procedure, call $\mathcal{L}_{max}$ the value of the log-likelihood provided by the maximum likelihood algorithm without any constraint and

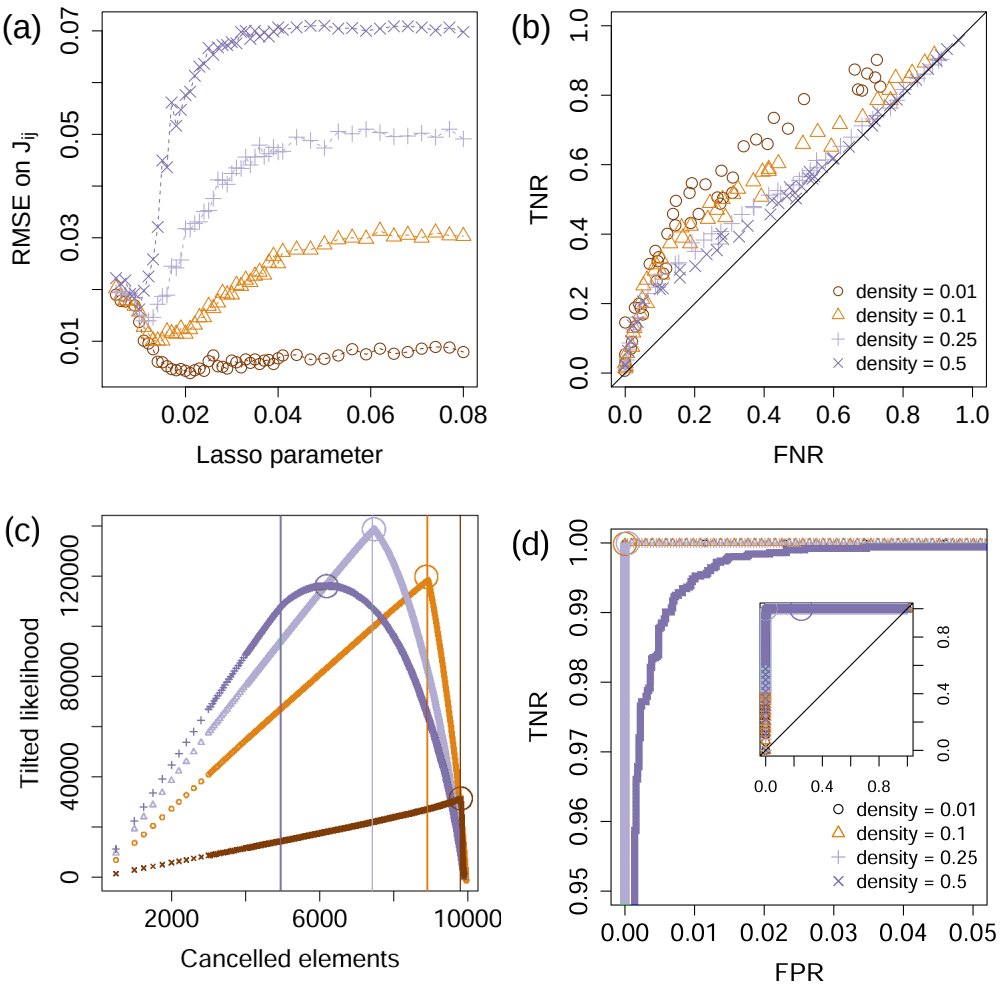

Figure 3.5: (a-b) Results from the LASSO with 80% observations: (a) RMSE on couplings as a function of the LASSO parameter; (b) ROC curves. (c-d) Results from the decimation procedure with 80% observations: (c) Tilted likelihood evolution through the decimation process, vertical lines show the correct number of null elements; (d) ROC curves through the decimation process with different network densities. The circle identifies the point at which the Tilted Likelihood is maximized.

then call x the fraction of parameters J_{ij} that are being set to 0. Finally call $\mathcal{L}(x)$ the log-likelihood of the model with the fraction x of decimated parameters and $\mathcal{L}_1$ the log-likelihood of a model with no couplings that is, in case $h_i = 0 \,\forall\, i$, $\mathcal{L}_1 = -\sum_t M(t) \log 2$. The Tilted log-Likelihood takes

the form

$$\mathcal{L}^{tilted}(x) = \mathcal{L}(x) - ((1-x)\mathcal{L}_{max} + x\mathcal{L}_1)$$

that is, the difference between a convex combination of the original log-likelihood with the log-likelihood of a system with no parameters and the log-likelihood of the decimated model. This function is strictly positive and is 0 only for $x = 0, 1$, since $\mathcal{L}(0) = \mathcal{L}_{max}$ and $\mathcal{L}(1) = \mathcal{L}_1$, thus there has to be a maximum. The decimation process thus consists in gradually increasing the fraction of pruned parameters x until the maximum of the Tilted log-Likelihood is found, giving the optimal set of parameters of the model.

We show in Figure 3.5 and 3.6 the results of the test. We observe how the ROC curves seem to lean strongly in favor of the decimation approach, which tends to score perfectly on the False Positives Ratio (FPR) - True Negatives Ratio (TNR) plane. However the maximum of the Tilted Likelihood does not always correspond to the optimal score in the ROC diagram, both in the case of a non-sparse network and when the data has a large number of missing values. While the former case is not particularly interesting in that a dense network model fitted on real data would be prone to overfitting and of disputable use, the latter is much more of a concern, albeit the process is still surprisingly efficient even when data is extremely sparse.

Even if the decimation procedure is consistently outperforming the LASSO, there is reason to still hold the ℓ_1 regularization as a viable option. Indeed when one introduces local fields h of non-negligible entity, the decimation procedure is not anymore reliable in that the Tilted Likelihood becomes non-convex as shown in Figure 3.6 and the maximum is not in the correct

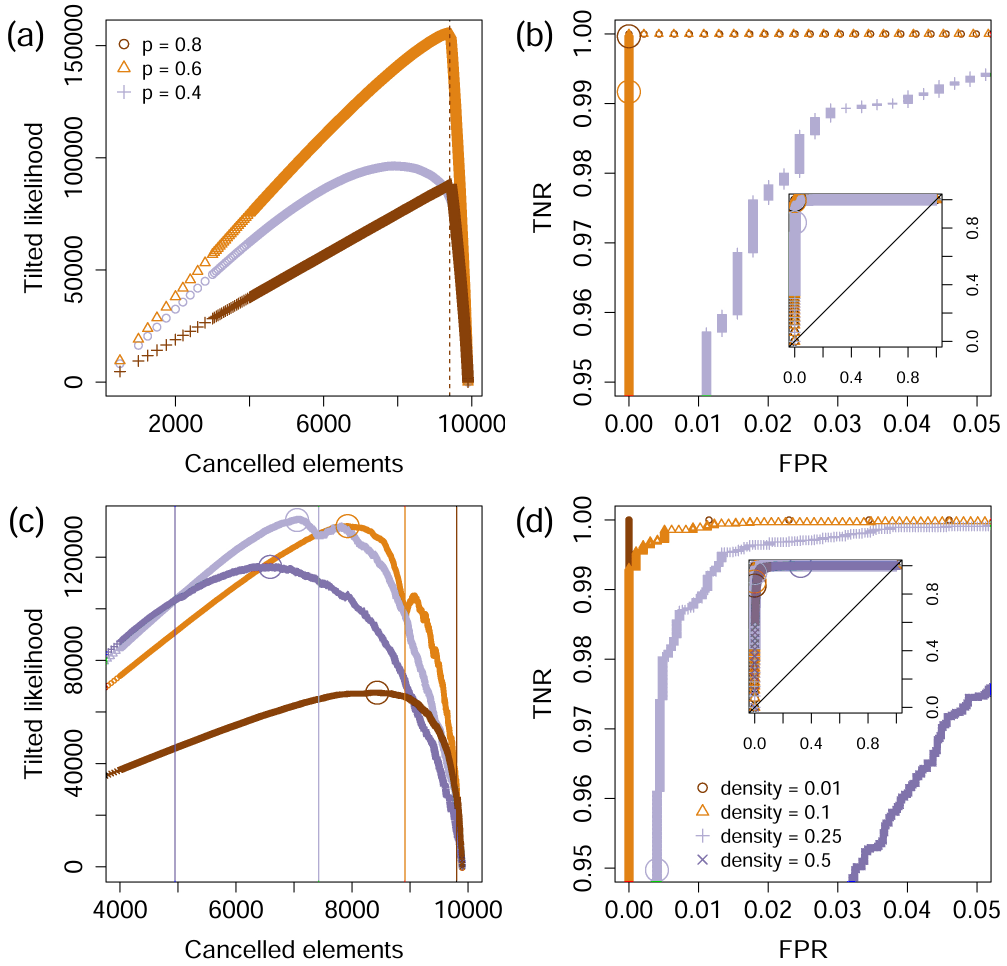

Figure 3.6: (a-b) Results from the decimation procedure with 80%, 60% and 40% observations available and a network density of 0.05: (a) Tilted Likelihood evolution through the decimation, vertical line shows the correct number of null elements; (b) ROC curves through the decimation process with different observation densities. (c-d) Results from the decimation introducing local fields h: (c) Tilted likelihood, vertical lines show the correct number of null elements; (d) ROC curves. The introduction of local fields makes the tilted likelihood non-convex and seriously affects the performance.

position. This is due to the underestimation of the h parameters during the log-likelihood maximization of the fully connected model, where part of the role of the local fields is absorbed in couplings that should be pruned.

However these couplings are still relevant to the model since they compensate for the underestimated h parameters, giving the Tilted likelihood a non-convex form and shifting its maximum towards a more dense network model. This situation does not occur with the LASSO regularization as the pruning is performed at the same time as the maximization, giving the LASSO the advantage of a much more reliable fit of the local fields albeit with an overall worse performance in the inference of the nonzero couplings.

Test 4: Impact of asymmetricity assumption

Another assumption we made to perform the calculations in Equation 3.9 was that the J_{ij} are iid Gaussian random variables. In the case of social networks and trade networks reciprocity, that is the correlation between J_{ij} and J_{ji}, is often found to be much higher than what would be expected in an iid setting (Squartini et al. (2013)). We ask ourselves how impactful is this assumption on the outcome of the inference and we test the algorithm on data generated from a model with $N = 100$, $T = 10000$, $p_i = p = 0.8$, $J_1 = 1$ and such that $\mathrm{Cor}(J_{ij}, J_{ji}) = \rho,\ i \neq j$. We show the results for this series of tests in Figure 3.7. What we find is that the ρ parameter barely affects the performance and even makes it easier to infer the hidden variables, albeit marginally. Indeed we only used the assumption to approximate the determinant of the Hessian in the second order correction to the saddle-point solution, and letting the couplings not be reciprocally independent should affect the approximation slightly by having some elements of J^2 that vanish slower than others in the sums. It is possible that having a large enough N facilitates the inference then, since the amount of those slowly vanishing terms grows with N while the number of entries of J grows with N^2.

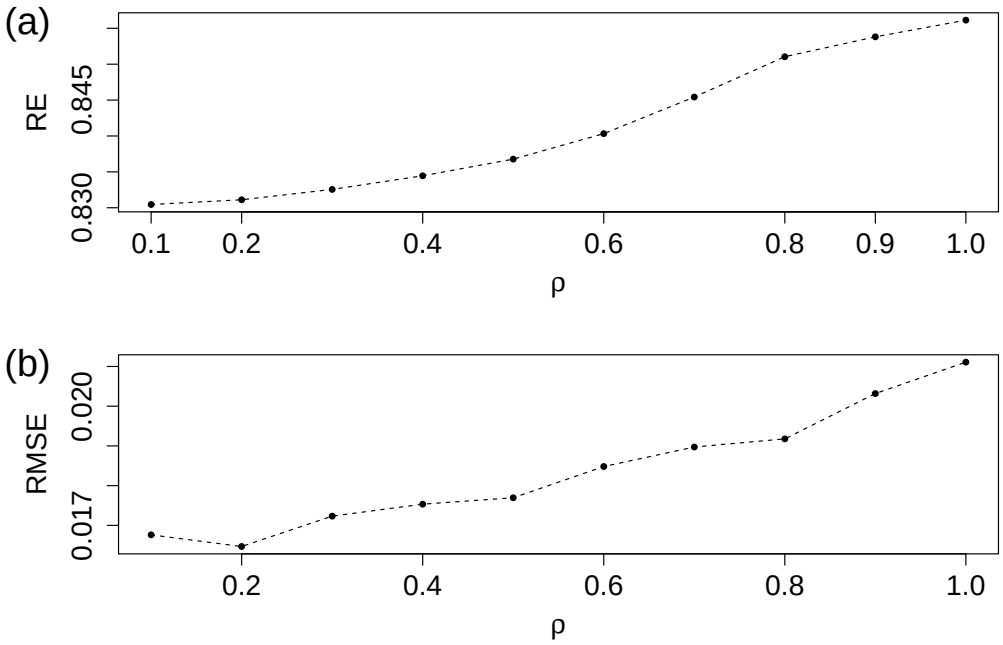

Figure 3.7: (a) Reconstruction Efficiency varying the correlation between symmetric elements of J; (b) RMSE on the couplings.

We then turn our attention to the extreme case of $\rho = 1$, corresponding to the well known Sherrington-Kirkpatrick (SK) model (Kirkpatrick and Sherrington (1978)), one of the first and most studied spin glass models in the literature. The SK model has the peculiarity of undergoing a phase transition at $J_1 = 2$ in our notation for the Hamiltonian (since we have not included a factor $1/2$ to remove double counting), where for $J_1 > 2$ the spin glass phase arises and multiple equilibrium states appear such that the model is not easy to infer anymore. It is thus interesting to see whether this affects the inference from dynamical configurations and how the identifiability transition is reached. We perform the experiment of varying J_1 in this framework and show the results in Fig. 3.8. We find the expected increase in rescaled error (that is, RMSE$/J_1$) marking the transition, surrounded by a finite-size scaling noisy region, while the reconstruction efficiency of the

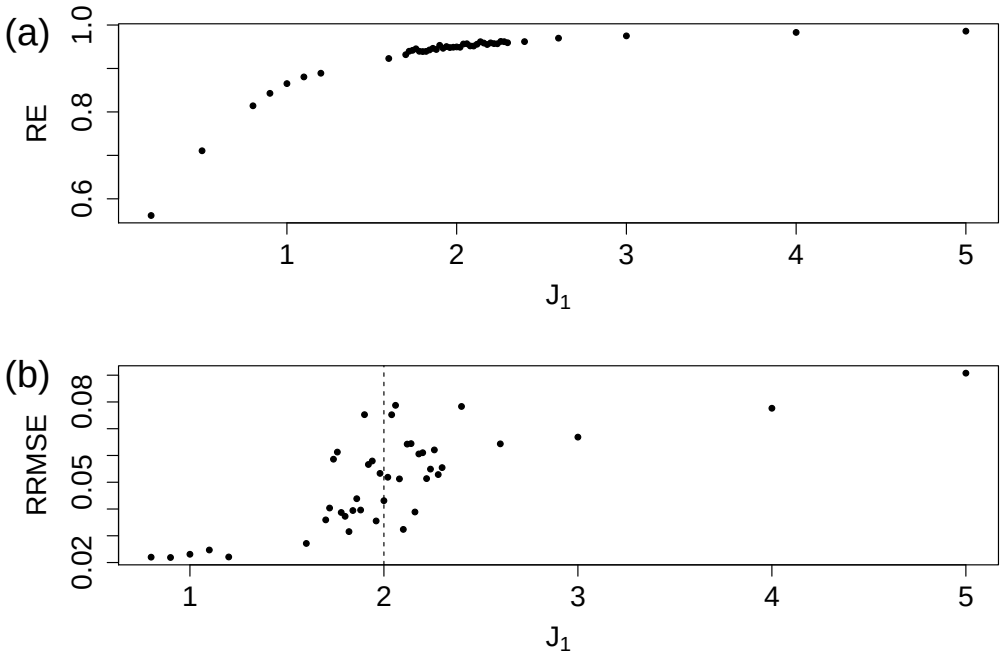

Figure 3.8: (a) Reconstruction Efficiency as a function of J_1 in the SK model; (b) Rescaled RMSE on couplings as a function of J_1.

configurations remains very good. This fits in the narrative of the phase transition of the SK model, since in the spin glass phase an equilibrium configuration of the model can be generated by multiple - and in principle indistinguishable - choices of parameters which we indeed struggle to identify with our methodology.

Test 5: sample size and convergence

We finally devolve our attention to the convergence properties of our estimator and how they are affected by finite sample sizes. The relevant parameter to be varied is the ratio between the length of the time series T and the number of units that are modelled, N. We run simulations with $N = 100$, $J_1 = 1$, $p_i = p = 0.8$ and varying T between 100 and 25000, and report the

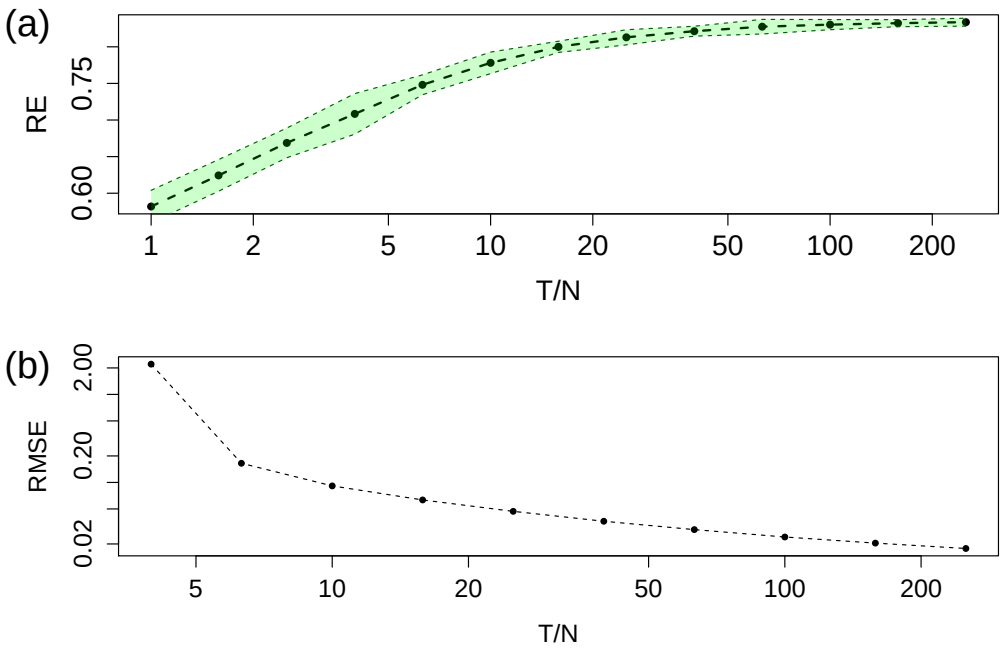

Figure 3.9: (a) Reconstruction Efficiency as a function of the T/N ratio; (b) RMSE as a function of the T/N ratio. Area in green is 1 standard deviation from the mean over 30 repetitions.

results in Figure 3.9. It can be seen that the RMSE on J_{ij} diminishes, after $T/N = 20$, with what might look like a power law behaviour with exponent close to 0.5, although we do not provide an exact law for the convergence. The RMSE is below 5% of J_1 when T/N is larger than 20 and is steadily converging towards 0. Regarding the reconstruction efficiency we see that it saturates quickly towards 90% and then it keeps increasing towards 100%. This evidence is a heuristic proof that the estimator is converging and is important to estimate how reliable a result might be given the T/N ratio of the data. Although a more rigorous law would be much more appealing for the task, it would require being able to write the posterior of J, σ given s, which to the best of our knowledge is not a feasible calculation in this setting.

Additional parameters: exogenous drivers

The model can be easily extended to a version in which an exogenous driver (or multiple ones), observed at all times, affects the dynamics of the variables. In a financial setting the first external driver would be given by the log-returns r_t and the associated parameter would be the typical reaction of a trader to price changes, typically categorized between contrarians and chartists whether they go "against" the flow (i.e. sell when the price rises and viceversa) or follow the trend. In the model, this is introduced by adding a set of linear parameters b in the local fields that couple the variables to the driver

$$g_k(t) = \sum_l y_l(t) + h_k + b_k r_t$$

The introduction of the parameter does not complicate the inference process at all and is particularly important if one wants to use the model to describe and possibly forecast order flows in financial markets. We omit the results for this section for the sake of space and because no significant dependency on the size of the b_k parameters is found for our performance metrics.

3.4 Conclusions

In this Chapter we developed a methodology to perform inference of Kinetic Ising Models on datasets with missing observations. We successfully adapt a known approximation from the Mean Field literature to the presence of missing values in the sample and devise several performance tests to characterize the algorithm and show its potential. We also propose a recursive methodology, R-EM, that gradually reconstructs the dataset with

inferred quantities and tries to refine the inference, and show its efficacy on synthetic data.

The main results are that it is indeed possible to infer Kinetic Ising Models from incomplete datasets and that our procedure is resilient to noise, heterogeneity in the nature of data and in the frequency of missing values, and overall quantity of missing data. We make the algorithm ready for real-world applications by implementing pruning techniques in the form of LASSO and decimation, and give a brief overview of what we think are the better uses for each.

The methodology lends itself to applications on many diverse datasets, but our main focus in the next chapters will be on opinion spreading in financial markets where transactions occur at high frequency, such as the FX or the cryptocurrency markets. We indeed envision our algorithm can identify significant structures of lagged correlations between traders, that in turn can be mapped to a network of lead-lag relations. Such a network would be particularly useful to get a quantitative picture of how possible speculative or irrational price movements can occur due to voluntary or involuntary coordination between traders and to devise appropriate strategies to counteract them.

Chapter 4

Traders networks and herding

Almost all results in this chapter previously appeared in Campajola et al. (2020)

4.1 Introduction

A significant part of risk management for financial intermediaries is related to the mitigation of the adverse selection risk (Kyle (1985); Glosten and Milgrom (1985)), namely the risk of trading with a counterpart that has access to better information on the traded asset. This risk is exacerbated in contexts where multiple counterparties might - consciously or not - co-ordinate their trades, introducing not only an adverse selection risk against a specific counterpart but against a group of traders, a situation typically referred to as inventory risk (Ho and Stoll (1980)). Understanding how information propagates in the market is crucial to identify key players that can forerun the order flow, a knowledge that an intermediary can exploit to better hedge against inventory risk.

Methods to detect lead-lag relationships between financial variables have

been extensively studied in the literature, starting with correlations between financial assets (Jegadeesh and Titman (1995)) and evolving towards more complex measurement methods such as cascade models (Lux et al. (2001)) or Vector AutoRegressive models (Barigozzi and Brownlees (2019)). In recent years there has been a rising interest in methods to cluster together traders based on their strategic and behavioral features as well as studying how they influence each other. A prominent example is the Statistically Validated Networks (SVN) methodology, first described in Tumminello et al. (2011) and then applied to financial data (Tumminello et al. (2012); Curme et al. (2015); Musciotto et al. (2018)), which has then been extended to the Statistically Validated Lead-Lag Networks methodology (Challet et al. (2018); Cordi et al. (2019)) to analyse how investors can be classified based on their strategic behaviour and which clusters correlate at different time-scales. Challet et al. (2016) proposed a Machine Learning method to construct lead-lag networks between clusters of investors and predict the order flow, while Gutiérrez-Roig et al. (2019) rely on information-based methods to achieve similar results.

We propose our approach as an alternative to the aforementioned methods, introducing the Kinetic Ising Model as an opinion spreading mechanism whose parameters can be inferred from the data, adopting the algorithm we showed in Chapter 3.

Our goal is to find significant lead-lag relationships between single market participants on the intra-day time-scale, as well as exploiting these lead-lag relations to estimate the current implicit state of supply and demand. There are two main innovations in our approach with respect to the above mentioned ones: on the one hand, treating the data as a whole in a multivariate model, instead of running multiple pairwise tests - as pre-

viously cited methods do - allows us to correctly identify correlations and causalities, whereas a pairwise approach is potentially prone to cases where spurious effects appear; on the other hand, we also have the ability to handle missing observations, which in the case of financial markets is an effect of the intrinsic asynchronicity of trade records, as shown by Aït-Sahalia et al. (2010) and Corsi et al. (2012).

The main purpose of financial markets is to aggregate the public opinion about a particular asset, determining the correct price as the optimal match between supply and demand. The opinion of a particular trader about the asset price is thus expressed when they perform a transaction: when they buy an asset at price p, they believe the correct price (the "value" of the asset) is $p' > p$, and vice-versa. Due to transaction costs, limited liquidity and other frictional effects, the traders incur in a cost whenever they want to express their opinion, inducing them to trade less than they would in an ideal situation. As a result, when looking at trade records on the intraday time-scale, it is very hard to aggregate time at a level such that every participant trades in every time slice. However it is reasonable to assume that, even if a trader has not traded in the last time interval, they still hold an opinion about the asset, which could be reflected in other trades they perform on other markets or could influence other traders in their future actions.

We choose to model this system through the Kinetic Ising Model, assuming traders' opinions can either be positive (belief that $p' > p$) or negative ($p' < p$) and thus be represented by binary spins that evolve in discrete time. Their coupling factors will then carry the information about lead-lag relationships in the spreading of opinions at the considered time-scale. As mentioned, the only observations available about such opinions are the

trades that investors make, meaning that the data will likely present a significant amount of missing values if one takes a reasonably short time step. A good reason to choose the Kinetic Ising Model then is the possibility to infer the model parameters efficiently even from incomplete data, thanks to the Expectation-Maximization-like algorithm developed in Campajola et al. (2019) and shown in Chapter 3, while also getting a Maximum Likelihood estimate of the unobserved opinions. Such estimations can then be used to make an informed guess about the hidden opinion, by simply taking their sign.

The case is particularly relevant for the foreign exchange (FX) market. The market has a multi-dealer organization, where a centralized double-auction exchange is accessible to few market members (the dealers) which in turn offer, through their proprietary platform, a trading service to their clients. The dealer then acts as a liquidity provider, while also absorbing temporary shocks in supply and demand through its inventory which is then rebalanced by trading with other dealers on the centralized platform. Optimal dealership (mostly known as optimal market making) is a vastly studied problem in finance (see Guéant (2016) for a comprehensive review), trying to devise how to optimally rebalance the inventory one accumulates when satisfying clients' requests and what is the fee the dealer has to charge clients in exchange for the immediacy of their transaction. One of the costs faced by dealers is the cost of liquidity on the inter-dealer market, which can be particularly high when all market participants experience the same kind of pressure from their clients. To predict what this cost will be it can be useful to understand what the aggregate opinion of traders is, even the ones the dealer doesn't observe due to lack of trading activity, either because they might influence other clients actions or because they are active

with other competing dealers and will eventually impact the cost of liquidity shortly afterwards.

Our modelling approach also allows to analyse the inferred lead-lag networks to identify key nodes in the opinion spreading process, whether the network changes over time as traders enter and exit the market, and to study how influential nodes are relevant for the prediction of the order flow and future liquidity.

The chapter is organized as follows: in Section 4.2 we describe the dataset and model we use, in Section 4.3 we show the results coming from multiple network analysis metrics, we analyze the performance of the model when trying to forecast the order flow and we define a herding measure from the inferred opinions, for which we test for Granger Causality (Granger (1969)) effects with several liquidity imbalance measures. Section 4.4 concludes the chapter.

4.2 Dataset

Our dataset consists of all the trades performed in the period going from January 2012 to December 2013 on the eFX platform of a major dealer in the EUR/USD spot exchange rate market, including an anonymized identifier of the market agent requesting the trade, the volume and sign of the transaction, the time of request, and the price in EUR/USD quote offered by the dealer.

We select trades occurring on working days between 8AM and 4PM GMT and we split the dataset by month, resulting in 24 time series of trades with information about time, volume, sign, and identity of the counterpart. We then aggregate trades performed by the same agent i within 5 minutes

time windows and take the sign of the aggregate volume $V_i(t)$ of EUR acquired in exchange for USD as the information on whether the agent has sold ($V_i(t) < 0$), bought ($V_i(t) > 0$) or has stayed idle ($V_i(t) = 0$) at time t. Finally, we call p_i the fraction of time intervals in which trader i was active - that is, the fraction of non-missing data - and for each month we remove traders that were active in a fraction $p_i \leq 0.3$ of the total number of samples.

The final dataset involves a total of 68 traders, with an average of 16 traders active each month, a minimum of 9 and a maximum of 29 and we report some statistics in Table 4.1. To better understand the heterogeneity in the activity of market agents involved, we compute the Gini coefficient on the monthly p_is and find the distribution of observations to be mostly homogeneous, typically having only one agent that is much more active than all the others.

	T	N	$\overline{s}$	$\overline{p_i}$
Minimum	679	9	0.02	0.45
Maximum	2231	29	0.13	0.55
Mean	2039.33	16.46	0.08	0.49
Stdev	308.57	4.59	0.03	0.02
	Trader p_i	p_i Gini	Trader ACF1	Flow ACF1
Minimum	0.30	0.16	-0.15	0.04
Maximum	0.99	0.23	0.57	0.19
Mean	0.49	0.19	0.07	0.12
Stdev	0.18	0.02	0.07	0.05

Table 4.1: Basic statistics of the dataset: (top) number of time steps T and of agents N of the monthly time series, monthly average sign of observed trades $\overline{s}$, monthly fraction of observations $\overline{p_i}$; (bottom) single trader monthly fraction of non-missing values p_i, monthly Gini coefficient of p_i, ACF at lag 1 of single traders and of the aggregate order flow.

The sign of the aggregate trade volume $s_i(t) = \text{sign}[V_i(t)]$ is intended as a proxy of the opinion $y_i(t)$ the trader has at that time on whether

the price should go up or down in the near future, while the zeros are intended as missing observations on their opinion. As shown in Table 4.1 the AutoCorrelation Function (ACF) at lag 1 on the aggregate order flow is typically higher than the average ACF of single traders, suggesting that traders act in coordination on a short lag, having their opinions diffuse gradually over a network of information spreading.

The model we use to describe our data is the Kinetic Ising Model with an external regressor for log-returns, described by the transition probability

$$p\left[y(t+1)|y(t)\right] = Z^{-1}(t)\exp\left[\sum_{\langle i,j\rangle} y_i(t+1)J_{ij}y_j(t) + \right. \\ \left. + \sum_i y_i(t+1)\left(h_i + b_i r(t)\right)\right] \tag{4.1}$$

where again the $y(t)$s are not always observed, thus we have the observed opinions $s(t)$ alongside the unobserved opinions $\sigma(t)$ at each time step and the usual reparametrization we showed in the previous chapter.

We infer the parameters of Eq. 4.1 on monthly subsets of data to account for non-stationarity and for traders that enter and exit the platform throughout the considered two year period. The outcome is a series of weighted and directed networks whose weighted adjacency matrix for month k is $A(k) = J^T(k)$ (transposing conforms the matrix to the standard definition of adjacency matrix, which has non-zero element a_{ij} if there is a link from node i to node j), where the nodes are traders and the links represent an influence relation between the sign of the opinion of the origin node at time t and the sign of the opinion of the end node at time $t+1$. The links can have either positive or negative weight: when it is positive it means that the follower tends to agree with the leader opinion, while when it is negative they tend to disagree.

Since we hypothesize that returns $r(t)$ can affect the trading behaviour of traders, we introduce the 5-minutes log-returns as a control variable in the model, with a trader-specific parameter b_i capturing their reaction to a price change in the previous time window. We use the mid-price in the order book of the EBS electronic inter-dealer exchange to which the dealer has access as a market member: although traders do not specifically trade at that price, it is the only price indicator that we can reliably use while not introducing trade-specific effects.

4.3 Results

The networks resulting from the model inference (as for example the one shown in Figure 4.1) are then analysed to find out whether there are traders that are more influential than others, how the network changes over time, and how accurate is the prediction of trade signs. We start by defining a characterization of the nodes as influencers and followers based on an adapted weighted version of the PageRank (Brin and Page (1998)) measure as proposed by Kiss and Bichler (2008). Then in the next subsection we compute the persistence of the neighbourhood of nodes as described by Nicosia et al. (2013) to quantify the local stability of the networks in time and try to disentangle degree-related effects from preferential attachment by comparing the results with the ones obtained by randomly rewiring the networks and reshuffling the time series. We then compute the out-of-sample accuracy of prediction of trade signs to evaluate model performance compared to a Logistic AutoRegressive (LAR) model of order 1, taking as input the previous trade sign of trader i (where available) and the last log-return. We also evaluate the forecasting and nowcasting performance of

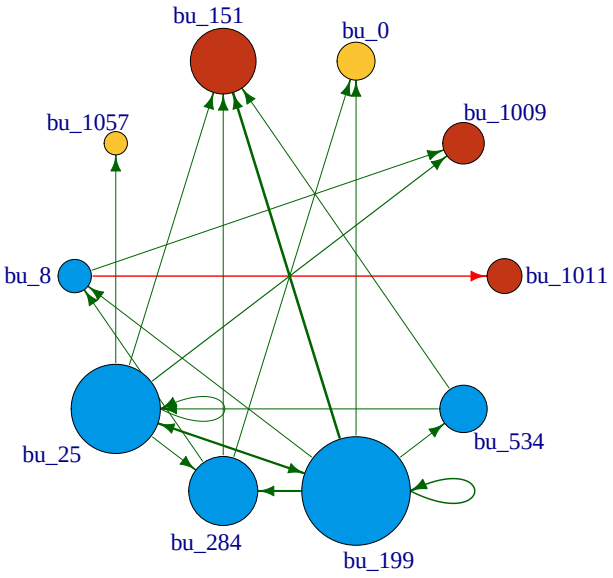

Figure 4.1: The inferred lead-lag network at month 13. Node coloring follows the PageRank influence categorization described in section 4.3, the size of both nodes and links is proportional to their strength and weight, respectively. The link color indicates whether it is positively (green) or negatively (red) weighted.

the model, utilizing parameters fitted on one month to predict trade signs in the next one, always comparing with the LAR benchmark. Finally we show a further interesting feature of our approach which allows us to define a micro-level herding measure. We take this measure and run a Granger Causality analysis between it and a set of liquidity imbalance measures computed on the order book of the EBS inter-dealer exchange to highlight the functioning of the multi-dealer market and emphasize the role of the dealer as a liquidity provider.

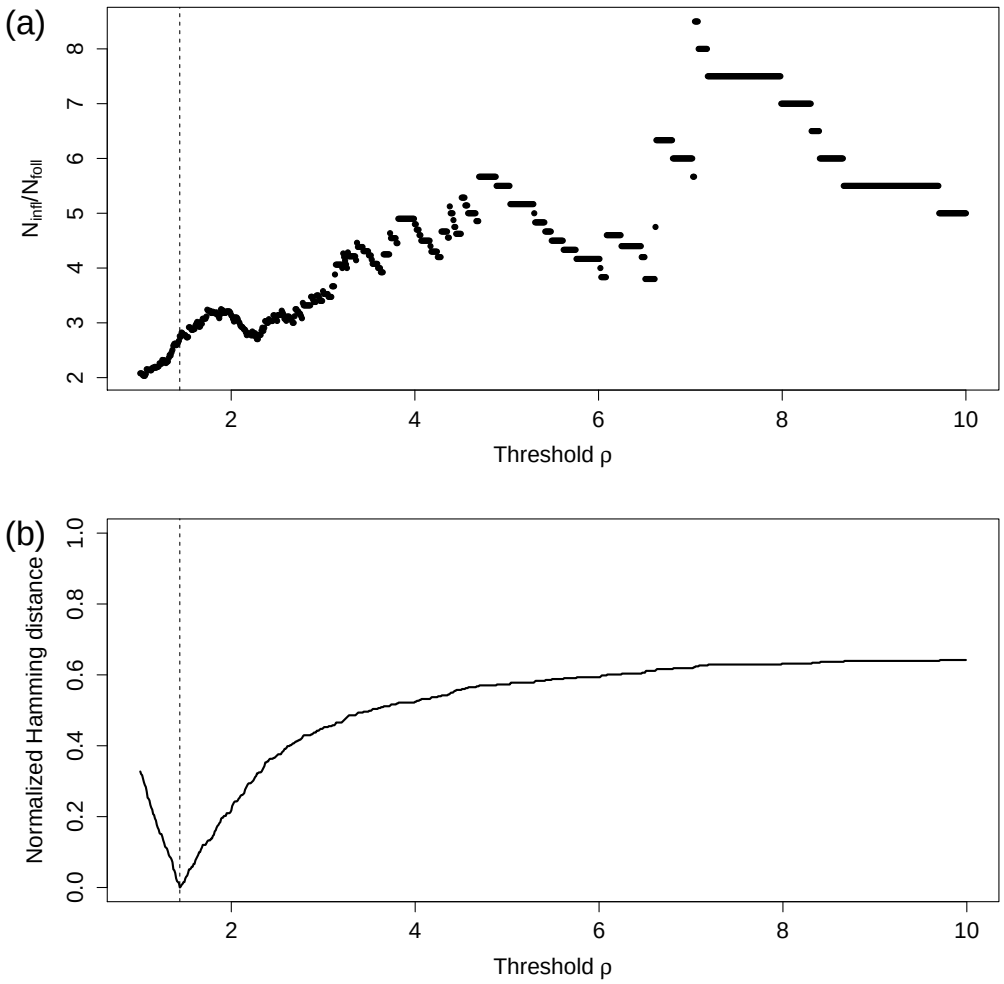

Figure 4.2: Stability of the PageRank ratio categorization. (a) Ratio between the number of identified influencers and the number of identified followers varying the threshold parameter ρ; (b) Normalized Hamming distance between the categorization chosen optimizing Eq. 4.3 and other categorizations varying ρ. Vertical dashed lines mark the value of ρ^*.

Influence network: key players and properties

It is our interest to identify key actors in the market that carry information about behavioral trends and that might "lead the pack", forerunning the order flow. To categorize traders in our network we adopt the measure

developed by Kiss and Bichler (2008) as a modification of the PageRank algorithm by Brin and Page (1998). The PageRank measure identifies important nodes based on how likely it is that a so-called *random surfer*, that is a random walker with some probability of restarting from a random node, ends up on some specific node of the network. In particular we want to label our nodes in 3 categories: influencers, followers, and neutrals. Kiss and Bichler (2008) define the Weighted PageRank (WPR) and the Weighted SenderRank (WSR) measures, where a node has higher WPR (WSR) the larger the relative strength[1] of incoming (outgoing) links it has from (towards) highly ranked nodes. Both these measures have a minimum value related to a parameter f called *damping factor*, representing the probability that the random surfer keeps walking instead of jumping to a random node, which we choose to be the literature standard 0.85 for both. The resulting measure for node i is then defined as

$$\mathrm{WSR}_i = (1-f) + f \sum_{j \in L_i} \frac{w_{ij}}{S_i} \mathrm{WSR}_j$$
$$\mathrm{WPR}_i = (1-f) + f \sum_{j \, s.t. \, i \in L_j} \frac{w_{ji}}{S_j} \mathrm{WPR}_j$$

where L_i is the set of nodes that have an incoming link from node i, w_{ij} is the weight of the link between i and j and $S_i = \sum_{L_i} w_{ij}$ is the out-strength of node i.

Notice that since the links can have negative weights we take the absolute value of the weight to account for negative influence as well. We then define the category C_i^t of trader i in month t based on the ratio between

[1]The strenght of a node is the sum of the weights of all links pointing at (in-strength) or departing from (out-strength) that node.

their WSR and WPR:

$$C_i^t = \begin{cases} \text{Influencer, } \mathcal{I} & \text{if } \mathrm{WSR}_i^t/\mathrm{WPR}_i^t > \rho \\ \text{Follower, } \mathcal{F} & \text{if } \mathrm{WSR}_i^t/\mathrm{WPR}_i^t < 1/\rho \\ \text{Neutral, } \mathcal{N} & \text{otherwise} \end{cases} \tag{4.2}$$

where ρ is a threshold ratio that can be arbitrarily decided. To make this decision less arbitrary, we try to find an optimal value of the ratio in order to maximize categorization diversity cross-sectionally while keeping it consistent through time. The idea is thus to minimize a measure of diversity for the single agent across months, while maximizing the same measure between different agents in the same month, and taking ρ as the optimal in terms of Euclidean distance from the ideal case of perfect trader consistency in time and perfect uniformity of categorization cross-sectionally.

Call $p_i^\rho(C) = 1/T \sum_t \delta(C_i^t, C)$ the empirical probability at which trader i is assigned to category C using threshold ρ: the measure of diversity we choose is the normalized *Total Variation Distance* $d(p_i^\rho)$ from the uniform distribution, namely

$$d(p_i^\rho) = \frac{3}{2} \sup_{C \in \{\mathcal{I},\mathcal{F},\mathcal{N}\}} \left| p_i^\rho(C) - \frac{1}{3} \right|$$

where p_i^ρ is compared to the uniform distribution which takes value $1/3$ for all categories, and the factor $3/2$ is making sure that $d(p_i^\rho)$ is normalized to 1 in the case of maximum homogeneity, while it is 0 at maximum diversity. Call $\check{C}_i^\rho = \arg\max p_i^\rho(C)$ the most frequent categorization of trader i at threshold ρ, and define the frequency of category $\check{C}$ among the $\check{C}_i$s as $f^\rho(\check{C}) = 1/N \sum_i \delta(\check{C}_i, \check{C})$. Finally, call $\zeta^\rho = d(f^\rho)$ the cross-sectional

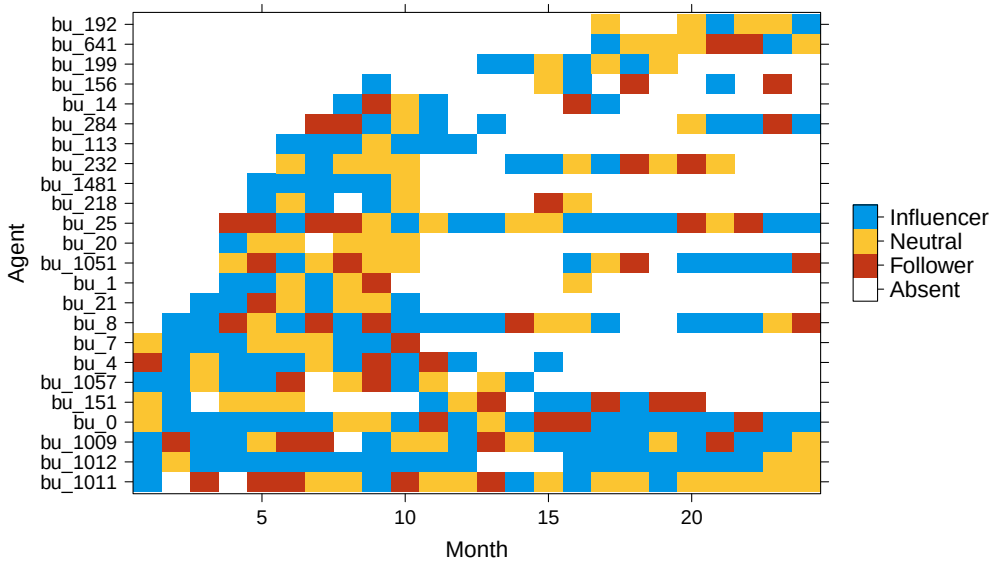

Figure 4.3: PageRank categorization of agents across months.

diversity between the most frequent categories the traders are assigned to. Then we optimize ρ as

$$\rho^* = \arg\min_{\rho} \left[(\zeta^\rho)^2 + (\mathbb{E}_i[d(p_i^\rho)] - 1)^2 \right] \tag{4.3}$$

that is we minimize the Euclidean distance from the ideal case of having each trader in the same category every month ($\mathbb{E}_i[d(p_i^\rho)] = 1$) and evenly spread categories across agents ($\zeta^\rho = 0$), obtaining a threshold value of 1.44.

We show how the selection of the threshold affects the categorization in Fig. 4.2, plotting the influencers/followers ratio and the Hamming distance between the chosen and all other categorizations. In the region surrounding the chosen threshold the ratio of influencers to followers is rather stable at around 3, and the normalized Hamming distance (that is the fraction of categories changing between two choices of ρ) between the chosen category and its neighbourhood is rather low and smoothly varying when moving

away from the chosen threshold, a sign that the categorization is stable enough to justify using this selection method.

The resulting categories for traders that exist in the data for more than 5 months are shown in Fig. 4.3. It is rather interesting to see how some traders show a consistent behaviour across the whole dataset being identified mostly as influencers (see for example trader #1012, #113 and #1481), while others have a more swinging nature.

Network persistence

To understand how variable the network is from month to month we compute the neighbourhood persistence measure proposed by Nicosia et al. (2013), defined as

$$D_i(t, t+1) = \frac{\sum_j a_{ij}(t) a_{ij}(t+1)}{\sqrt{\sum_j a_{ij}(t) \sum_k a_{ik}(t+1)}} \tag{4.4}$$

where a_{ij} are the elements of the network adjacency matrix. Since the network is directed we compute the measure on the three possible neighbourhoods - the in, out, and total neighbourhood - changing the summation indices appropriately: in particular, Eq. 4.4 refers to the out-neighborhood, while summing over rows instead of columns produces the measure for the in-neighborhood and the total is obtained by using the symmetrized adjacency matrix $A^T + A$. We compare it to the same measure averaged over $10,000$ order randomizations of the network time series to isolate the actual persistence in time from the average connectivity the trader has. In Figure 4.4a we plot the two quantities for the 10 nodes in the network that show the largest persistence and for all neighbourhood types. We see that these nodes tend to have abnormally persistent neighbourhoods, sometimes more

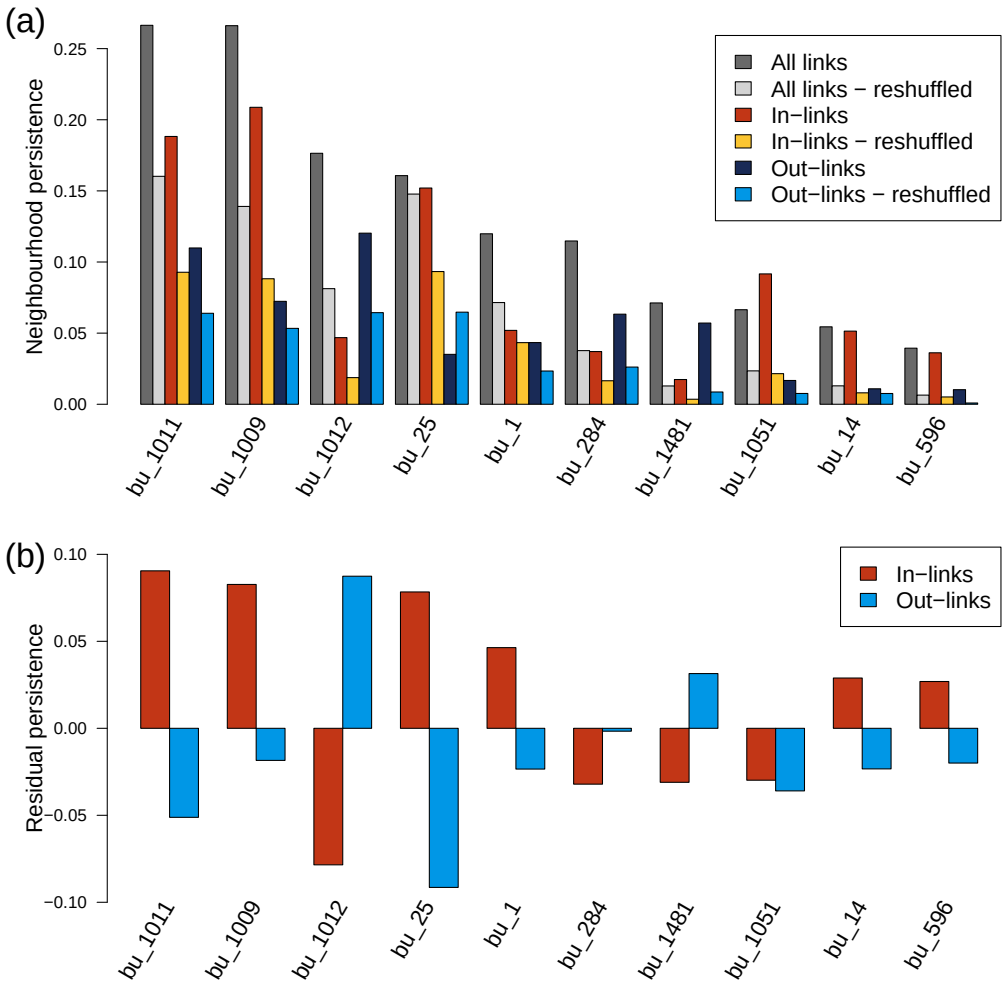

Figure 4.4: (a) Neighbourhood persistence measured before and after a random reshuffle of the network time series for a subset of nodes; (b) Residual in- and out-neighbourhood persistence after a degree-preserving rewiring of the network for the same subset of nodes.

in the in-neighbourhood and sometimes in the out-neighbourhood, a sign that some preferential relationships exist and replicate themselves in time. We also test the persistence by doing a random degree-preserving rewiring of the networks (while keeping the temporal structure) in order to remove the effect of the node degree, as a more connected node is more likely to

have a more persistent neighbourhood than a less connected one. Figure 4.4b shows the difference in the directed neighbourhood persistence between the original networks and the rewired ones. When this residual persistence is positive it means that the node has a persistence higher than in the null configuration model and viceversa.

The results show that there are indeed nodes that show a higher (or lower) persistence in their neighbourhoods even when ruling out the effect of the in- and out-degree, while this is typically not true for the undirected version (which roughly corresponds to the sum of the two). A node with a higher persistence of the out-neighbourhood is a node that attaches preferentially to some other nodes, meaning, in our convention, that it has influence over a persistent set of nodes, while the opposite is true for a node with higher persistence of the in-neighbourhood. For example, node #1011 has overly persistent incoming links and non-persistent outgoing links, meaning it is typically influenced by the same set of nodes, while its influenced neighbours are more randomly selected. The opposite happens for node #1012, which is indeed consistently recognized as an influencer by Weighted PageRank. Overall this analysis shows that, even if the network density is rather high and it is difficult to extract significant community structures, there is evidence of some preferential attachment mechanism at work in the directed network.

Out of sample validation and forecasting

In this subsection we perform out of sample validation and forecasting for the presented model. Specifically, we neglect some observed trades and we test whether our model is able to correctly guess them. In the forecasting exercise we instead train the model in a subperiod and test whether we are

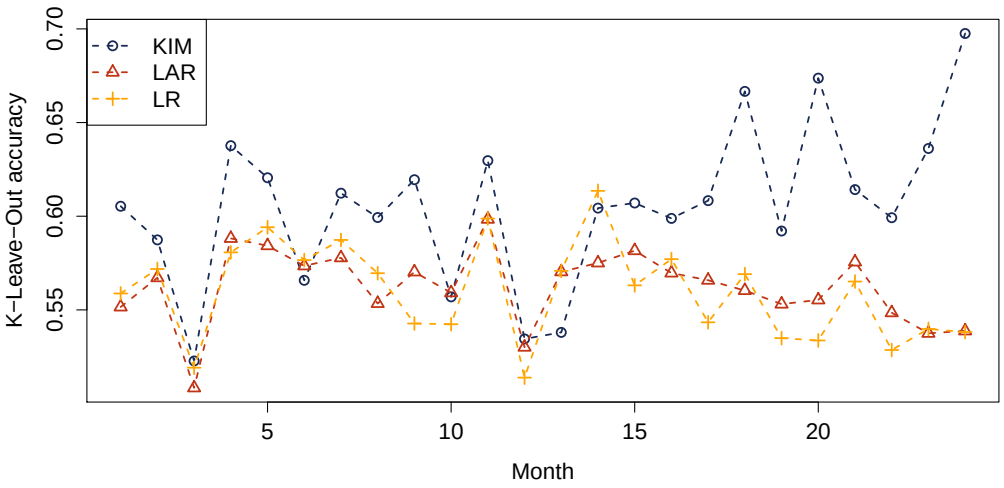

Figure 4.5: Out-of-sample K-leaveout accuracy of the KIM model compared to a Logistic AutoRegressive (LAR) model and to a Logistic Regression on log-returns (LR) for every month in the dataset.

able to predict the trading activity in the following subperiod. In all cases we consider models with and without 5-minutes log-returns as a control variable.

In Figure 4.5 we plot the performance one has predicting out-of-sample trade signs using the Kinetic Ising Model compared to the average of N logistic univariate logistic regressions, with the log-returns as an independent variable and, in the AutoRegressive version (LAR), the trade sign of the trader at the previous time interval (if available). The performance is measured by K-leaveout cross-validation, consisting of hiding $K = 5\%$ observations from the sample and then comparing the predicted trade sign with the actual one. The measure is then the fraction of correctly identified trade signs.

Overall the performance of our model is better than the benchmarks in a range from 5% to over 10% (excluding a couple of months where it does slightly worse), and in the best case the model predicts trade signs

with 70% accuracy, while on average it scores around 60%. While being nothing too extraordinary, this result shows that the model can provide a valid platform for descriptive and forecasting purposes.

The difference in performance between the KIM and the univariate logistic regression models is larger when the cross-correlation at lag 1 between the order flow and the log-returns is non-significant (not shown here). This tells us that, while the simpler models capture what is probably the most important interaction observed in the market (the reaction of traders to price changes), when this interaction is weaker they fail to capture any significant effect. However a significant amount of coordination persists regardless of whether it is caused by price movements or by other mechanisms, and it can be explained by our modelling approach.

We thus try to use the Kinetic Ising Model to forecast order signs: as a proof of concept, we take the result from one month and use the inferred parameters to produce the one-step-ahead forecasts in the next month.

Calling $\{J^M, h^M, b^M\}$ the set of inferred model parameters at month M and $s_i^M(t)$ the observed order sign of trader i at time t in month M, we forecast one step ahead using $\hat{s}_i^M(t+1) = \text{sign}(\check{m}_i(t+1))$, where

$$\check{m}_i^M(t+1) = \tanh\left[h_i^{M-1} + b_i^{M-1} r_t + \sum_{j \in \text{obs}(t)} J_{ij}^{M-1} s_j^M(t) + \sum_{b \notin \text{obs}(t)} J_{ib}^{M-1} m_b^M(t)\right]$$

and $\text{obs}(t)$ is the set of observed indices at time t. This quantity is then compared to the time $t+1$ observations and the average number of correct guesses is reported as the forecasting performance. Notice that every time an observation is added the $\check{m}^M(t)$ vector is updated through Eq. 3.13 to include the new information and keep the forecasting just to one step ahead of the observations.

We analyze the performance of the KIM when using all traders or only the influencers subset to predict future order signs compared to the same task performed with a LAR model. The results (not reported for the sake of space) show that there is no significant increase in performance by introducing the multivariate modeling, and restricting the prediction to using only traders that were identified as influencers in the previous month doesn't seem to change radically the forecasting accuracy.

We observe that the performance is marginally better ($\sim 55\%$) than a random guess and that is rather stable across time horizons (we also tested the one-step ahead forecasts using models several months after their inference without noticing significant changes). Our hypothesis is that both the LAR and the KIM methods, when used for forecasting, mostly rely on the log-returns to guess the next trade, which we believe is the reason why the accuracy of predictions is just a few percents higher than a coin flip and it does not vanish at longer time horizons. While this may seem at odds with the results shown in the previous sections, it has to be pointed out that the main objective of our model is to infer the state of investors when they do not trade, not forecasting, and that to do so we take advantage of future information in Eq. 3.13, something that is clearly not possible for one step ahead forecasts.

Predicting liquidity from inferred opinions

One possible use for our modelling approach is to produce a "herding" measure, given by the average opinion of traders at any point in time. Indeed a by-product of the model estimation is a maximum likelihood estimate of the unobserved opinions in the market, which we can use to generalize the buy-sell imbalance that trade signs show to an implied opinion imbal-

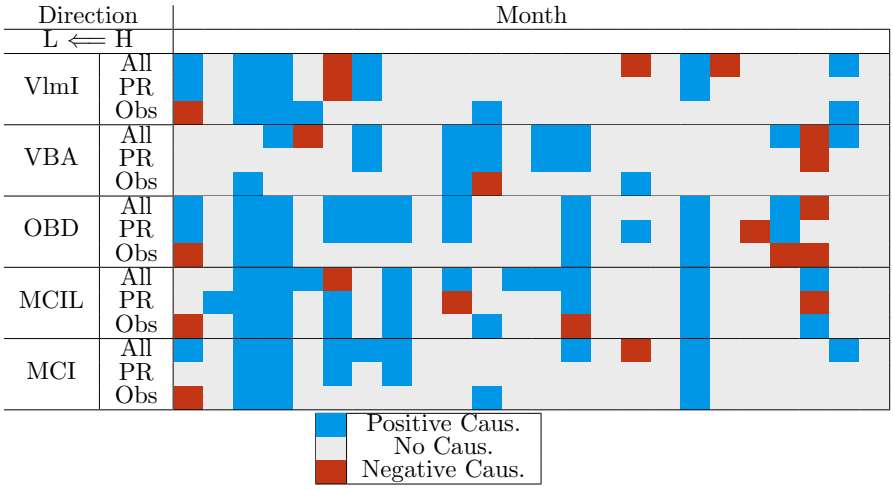

Figure 4.6: Causality relations from herding to liquidity. The model is based on 5 minutes lags with order up to 12, that is one hour, and the reported sign is the one of the coefficient is the one for the minimum order showing Granger Causality effects. The herding measure either accounts for all traders ("All"), only the ones belonging to the influencers group under the PR influence measure ("PR") or only the observed ones ("Obs").

ance. Typically herding is defined as an irrational behaviour that crowds show where a large fraction of agents co-ordinate based on social interaction rather than as a reaction to information, often resulting in unjustified macroscopic phenomena as, in the case of financial markets, price volatility jumps and dramatic liquidity imbalances. Herding has been documented in fund industry (Grinblatt et al. (1995)) as well as in institutional and individual investors (Nofsinger and Sias (1999); Grinblatt and Keloharju (2000)), and in market members (Lillo et al. (2008)).

The herding measure we define, as a simplification of the one already present in Lakonishok et al. (1992), is

$$H(t) = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i(t)$$

where $\hat{y}_i(t)$ is either the observed sign of the transaction $s_i(t)$ executed by

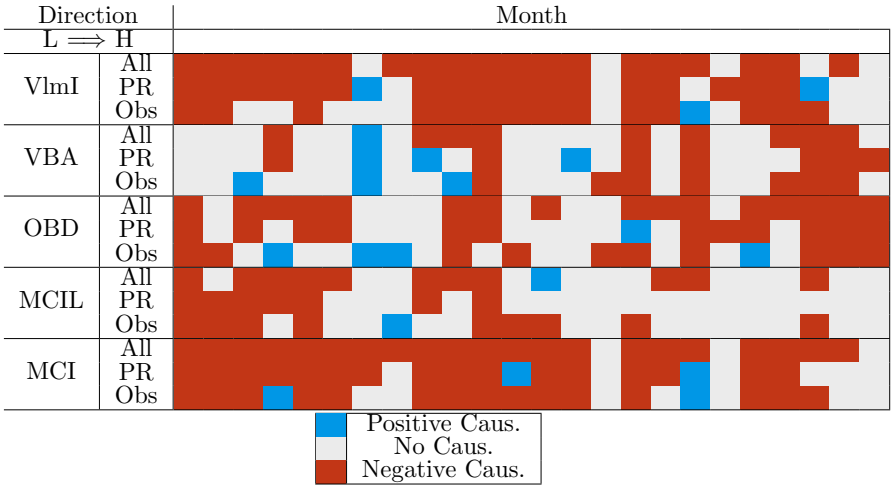

Figure 4.7: Causality relations from liquidity to herding. The model is based on 5 minutes lags with order up to 12, that is one hour, and the reported sign is the one of the coefficient is the one for the minimum order showing Granger Causality effects. The herding measure either accounts for all traders ("All"), only the ones belonging to the influencers group under the PR influence measure (and "PR") or only the observed ones ("Obs").

trader i at time t or the one inferred as sign$[m_i(t)]$. The main difference with the existing definitions is of course that we include also inferred states.

We want to show how this measure can be used to study a typical problem dealers are confronted with, that is facing poor liquidity conditions in the inter-dealer market when their inventory becomes unbalanced due to unexpected trading pressure from clients. To this end, we take into account a set of liquidity imbalance measures in the interdealer market:

- **VBA**: Dollar Volume at best Bid-Ask. It is the difference between the volume of limit orders at the best bid level and the volume of limit orders at the best ask, normalized by the total volume at those levels;

- **OBD**: Order Book Depth. It is the difference between the number

of levels that have to be explored to execute a buy market order of 10^7 units of currency (which is the typical imbalance that the dealer accumulates in a 5-minutes time window) and an equal sell market order size;

- **MCI imbalance**: It is the imbalance between the Marginal Cost of Immediacy between the ask and bid side. MCI, introduced by Cenesizoglu and Grass (2018), is defined as

$$\mathrm{MCI}_A = \frac{\mathrm{VWAPM}_A}{\mathrm{Vlm}_A}$$

$$\mathrm{VWAPM}_A = \log \frac{\frac{\mathrm{Vlm}_A}{\sum_{l=1}^{L} Q_{A,l}}}{0.5(P_{A,1} + P_{B,1})}$$

$$\mathrm{Vlm}_A = \sum_{l=1}^{L} P_{A,l} Q_{A,l}$$

where $P_{A,l}$ is the price at level l on the Ask side and $Q_{A,l}$ is the quantity available at level l on the Ask side. The same can be defined for the Bid side and the measure we use is $MCI_A - MCI_B$. The quantity is computed for $L = 10$ (MCI) and for $L = OBD$ (MCIL), in order to capture book-wide imbalances as opposed to typical transaction size imbalances.

- **VlmI**: Dollar Volume Imbalance. It is the normalized amount of dollars in orders on the bid side of the book minus the same quantity on the ask side;

All these measures are defined such that a positive imbalance means that liquidity is higher for the bid side of the order book, that is it is easier for a market participant to execute a sell market order (the asset is always considered to be EUR and the quotes are given in USD, as in the dealer platform data).

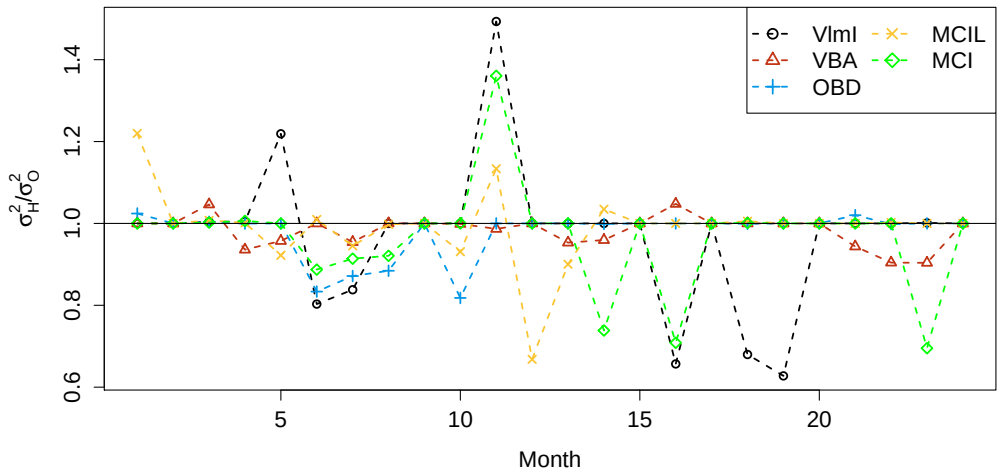

Figure 4.8: Residuals variance ratio between models using our herding measure H_t or the observed trade sign imbalance. The ratio is mostly less than 1, showing models have a better goodness of fit when using H_t.

To explore the relationship between these measures and $H(t)$ we run Granger Causality tests on pairwise Vector AutoRegressive (VAR) models for which we report results in Figures 4.6 and 4.7. We choose this method over alternatives such as Transfer Entropy (Bossomaier et al. (2016); Novelli et al. (2019)) for its simplicity and ease of interpretation, but it is possible that comparing with more elaborate techniques can produce more interesting results. The figures show the causality relations we find in both directions and specify whether the coefficient of the VAR model for which the causality is found is positive or negative. If a positive (negative) causality is found, it means that an increase in the first variable is causing an increase (decrease) in the other. To reduce the number of false positives we implement the false discovery rate (FDR) method of Benjamini and Hochberg (1995) for multiple hypothesis testing, setting the significance threshold at 0.05.

The results highlight the importance of the dealer in distributing liquid-

ity and absorbing temporary imbalance in the supply and demand: indeed most of the relations running in the direction $H \to L$, that is Herding to Liquidity, are positively signed, while the opposite is true when looking at the $L \to H$ direction. This means that when the herding measure is positive, and so the majority of traders on the eFX platform is buying EUR, the liquidity on the EBS market will make it harder for the dealer to quickly rebalance her inventory as the imbalances are typically positive, meaning it is easier to sell than to buy EUR. On the other hand, when the EBS market conditions are favorable for the dealer to sell (positive L), this is typically followed by a majority of traders selling EUR to the dealer (negative H), as it is likely that the dealer is offering better quotes given the ease she has in unloading excess inventory.

We also show how the herding relation to liquidity is typically unchanged whether one includes in its computation all traders or only the subset of influencers as identified by the Weighted PageRank measure, meaning that they are indeed among the most informative traders in this sense, while only using the observed trades and ignoring the opinions reconstructed through the Kinetic Ising Model one finds less and more incoherent causality relations. As a further argument in support, the quality of the Vector AutoRegressive model fit is generally better when considering our measure over the observed trades imbalance, as shown by Figure 4.8. There we compare the variance of the residuals on the liquidity side of the VAR model when using our herding measure $H(t)$ or the observed trades as the other model variable. We see that the ratio is typically less than 1, meaning the variance is smaller (and thus the fit better) with our measure.

To further investigate this relation, we apply the test of Granger Causality in tail originally proposed by Hong et al. (2009) between $H(t)$ and the

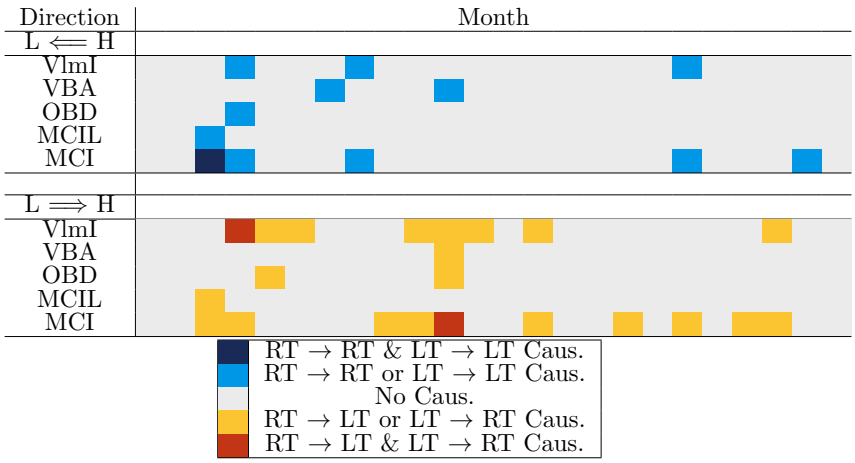

Figure 4.9: Tail Granger Causality relations identified by the test of Hong et al. (2009). We see how the results agree with the Granger Causality in mean, showing the strong connection between the markets also on extreme events.

liquidity measures. It is indeed interesting to see whether the Granger Causality only appears on average or it shows also between extreme events. The test is built to identify causality relations between binary time series, representing occurrences of extreme (tail) events with respect to recent history. Such events are identified as values of the liquidity or herding measure that exceed the 90% empirical conditional quantile (or are below the 10% quantile), measured as proposed by Davis (2016) on the past 2 hours of data at all points. The measurements above the 90% threshold are denominated as "right tail" (RT) events, while the events below the 10% one are "left tail" (LT) events. We show the results of the analysis in Figure 4.9, where we see that the picture given by the Granger Causality in mean is confirmed and the effects are particularly recurrent for the total volume imbalance (VlmI) and the MCI measure.

This last result highlights why it could be important to estimate the

unseen opinions of traders: given the multi-dealer structure of the spot rate FX market, a dealer only has a partial picture of what the supply and demand for the asset looks like at any given time, offered by the trades she sees from her clients. However these clients might have access to other dealer platforms and trade at their convenience with one or the other, thus hiding their opinion to the single dealer while still using market liquidity. This is then reflected in the order book of the inter-dealer market, where liquidity deteriorates whenever a shift in supply and demand occurs and makes it costly for the dealers to efficiently rebalance their inventories.

Simulation study

In conclusion to this chapter we present a brief simulation study aiming to provide further insight to the reader regarding the results that should be expected from this approach. We produce a synthetic dataset of opinions based on the basic statistics of our data, summarized in Table 4.1, by simulating the Kinetic Ising Model fixing N, T and the distribution of traders probability of observation p_i to closely resemble the ones that we observe in the data. We thus choose $N = 20$, $T = 2000$ and the distribution of p_i is assumed to be a Beta distribution, $p_i \sim B(\alpha, \beta)$, with parameters $\alpha = \beta \approx 4.01$. The value of the parameters is obtained following Pham-Gia and Turkkan (1992) in order to be consistent with the observed average Gini coefficient of p_i and the mean cross-sectional p_i of 0.5. The remaining free parameters are the ones directly related to what we aim to infer, that is the structure of the interaction matrix J and the magnitude of its elements.

We thus explore several degrees of sparsity of the underlying J matrix by sampling it as an Erdős-Rényi random graph with parameter $d_J \in [0, 1]$ describing the probability of a link, i.e. the density of the graph, and vary

the parameter J_1 which regulates the magnitude of the coupling coefficients assuming that for the existing links $J_{ij} \sim \mathcal{N}(0, J_1/\sqrt{N})$. The scaling with $\sqrt{N}$ is necessary to be able to compare parameters coming from models with different N, as it correctly normalizes the sum in Eq. 4.1.

We show these results in Figure 4.10, by plotting the Reconstruction Efficiency (RE) of hidden opinions, that is the fraction of hidden opinions that is correctly guessed, varying d_J and J_1 and showing the region we find empirically from our trading dataset. While no particular dependence of the RE is to be expected from the network density, as shown in Figure 4.10a, in Figure 4.10b we also see how it is instead strongly dependent on the magnitude of the couplings. This is also predictable from the theory, as we show with an hyperbolic tangent fit. Indeed the probability distribution of a hidden value σ_i given m_i is

$$p(\sigma_i(t) = \pm 1 | m_i) = \frac{1 \pm m_i}{2} \tag{4.5}$$

Here m_i depends from J_1 through Eq. 3.13 where J_1 is, given its definition and the Central Limit Theorem, the typical size of any sum of the kind $\sum_j J_{ij} s_j$ or $\sum_b J_{ib} m_b$, assuming all m_bs are estimated with no error and $N \to \infty$. Indeed the coefficients of the fit $\mathrm{RE} = a + b \tanh(J_1)$ are found to be $a = 0.49 \pm 0.03$ and $b = 0.43 \pm 0.03$ to 95% confidence for $N = 20$, $T = 2000$ and similar results are obtained for a larger system with $N = 100$, $T = 10000$. The small discrepancy between the theoretical value of $b = 0.5$ and the one we measure in simulations is most likely due to the presence of more than one hidden value, introducing uncertainty in the estimation of m itself.

We also plot the Root Mean Squared Error on J elements in relative terms to the magnitude of the parameters J_1, showing that in the region in

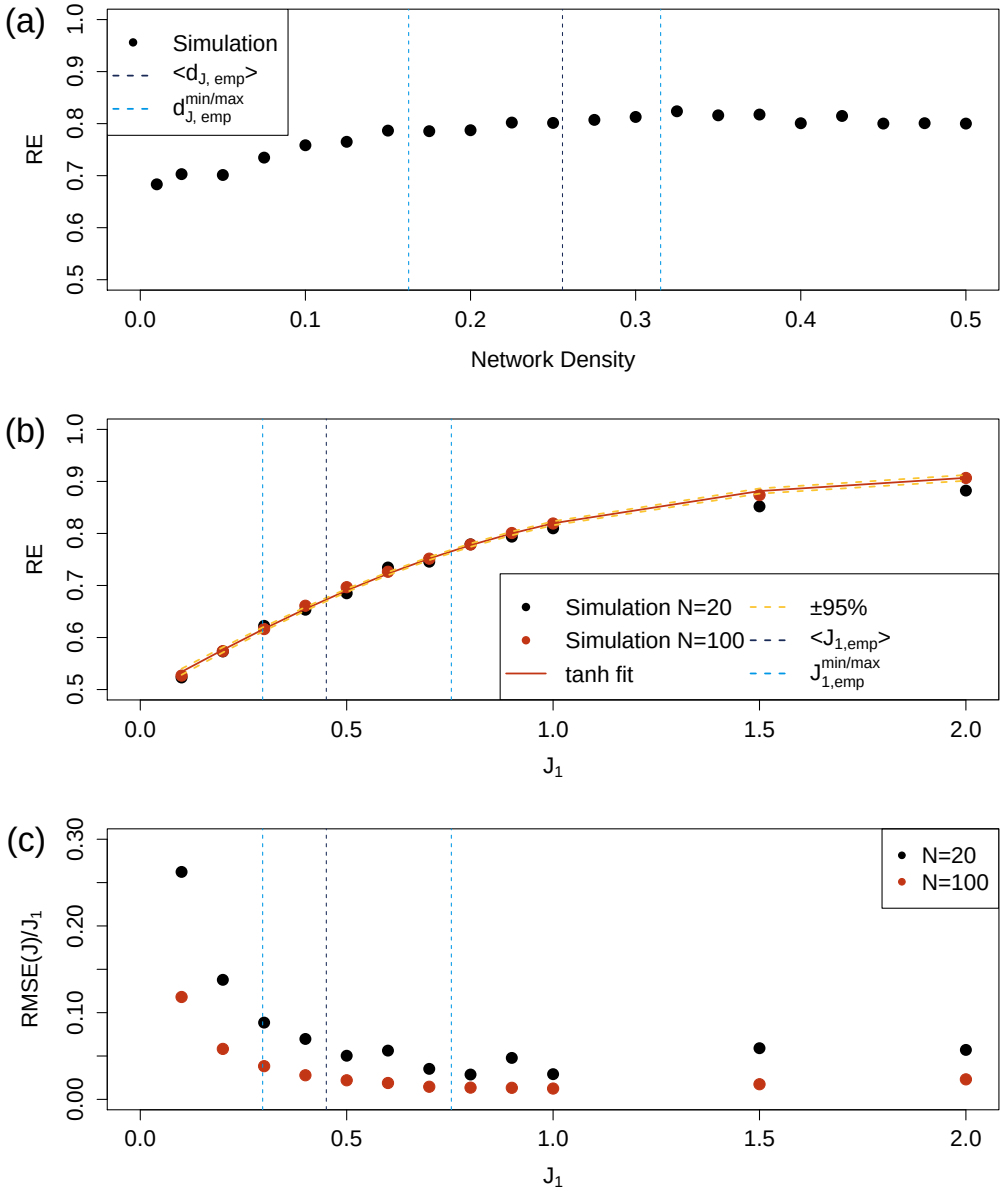

Figure 4.10: Results from the simulation study. (a) Reconstruction Efficiency (RE) varying the network density d_J. We see that, besides a slightly decreasing efficiency at very low densities, the expected performance is more or less constant. (b) RE varying the magnitude of couplings J_1. Here we see a clear relation between the two, highlighted by the hyperbolic tangent fit; (c) Rescaled Root Mean Squared Error (RMSE) on J elements. In all panels the blue lines show the region where the models inferred from trading data are situated.

which we find our inferred parameters there is a RMSE of roughly 5% in simulations, giving an idea of the error one could expect on the estimates.

Of course this is an ideal case, where the data generating process and the model coincide, meaning that these results have to be interpreted as upper bounds in performance. We indeed see that our out-of-sample performance results of Figure 4.5 are below the RE we get from simulations, but we argue that they are not that far from those given the size of the inferred parameters, meaning that even if the model is more than likely misspecified and an oversimplification of reality it still captures significant features from the data.

4.4 Conclusions

In this chapter we proposed the Kinetic Ising Model as a method to infer causal relationships between trader activities in a financial market at high frequency and to achieve a better estimate of the aggregate supply and demand at any point in time. We applied the model to a proprietary dataset offered to us by a major dealer, selecting the most active traders on their electronic foreign exchange trading platform to study the lead-lag relationships that occur among them and how their behaviour affects the state of liquidity on another market, the EBS inter-dealer electronic exchange. We showed that several market players can be identified as influencers, that is they are typically leading the order flow on the 5 minutes time-scale, and that their trading activity and opinion explains liquidity imbalances on the EBS market. Studying the persistence in time of the network structure on the local scale, we notice that some nodes have directed neighbourhoods that replicate through months, an effect that further validates the inferred

lead-lag relations and that matches quite well with the results from the influence analysis. We also test the forecasting performance that can be achieved with this model, finding that both the model and the LAR benchmark are not particularly well-suited for the purpose and the inclusion of the lead-lag relationships does not change the forecast significantly. We do not investigate the nature of these lead-lag relationships, but we propose they should be interpreted as the effect of different traders following similar strategies with different reaction times, leading to one or more traders consistently forerunning the others, rather than a more "direct" type of influence caused by actual social interactions. Finally we defined a herding measure based on the inferred opinions, which we show has much stronger Granger Causality relations with the state of liquidity on the inter-dealer market than just observed trade signs, exposing a mechanism that highlights the role of the dealer in providing immediacy to her clients and absorbing the cost of liquidity.

Chapter 5

The Score-Driven Kinetic Ising Model

The contents of this chapter are the result of a joint work with Domenico Di Gangi, Prof. Fabrizio Lillo and Prof. Daniele Tantari, to appear soon in an online pre-print and submitted for publication.

5.1 Introduction

In this last research chapter we will explore the possibility of giving a time evolution to some of the parameters of the Kinetic Ising Model. One complication that is ubiquitous to real complex systems, and particularly to the ones that cannot be reproduced in laboratory experiments, is non-stationarity. It is often the case in fact that systems change over time, possibly even in response to their own dynamics: traders in financial markets continuously adapt their strategic decision-making to each other's actions (Challet et al. (2016)) and to new information (Lillo et al. (2015)); preys change their behavior to avoid predators (Schmitz (2017)); neurons rein-

force (or inhibit) connections in response to stimuli (Tavoni et al. (2017)). Making accurate descriptions assuming that all parameters are constant is then frequently very hard if not impossible, resulting either in very strong limitations to sample selection and experimental design or in the necessity to develop models that are able to capture this non-stationarity with reasonable effort and accuracy.

There are examples of successful attempts to overcome this issue: for instance the introduction of temporal networks (Holme and Saramäki (2012)) as the space in which interactions are embedded has provided suitable methods to account for relations that are confined in time. More generally these network models refer to the broader literature on Hidden Markov Models (Ghahramani (2001); Sewell and Chen (2016); Nystrup et al. (2017)), where the basic assumption is that the observations come from a model whose parameters are dependent on an underlying, hidden Markovian dynamics that makes the system state evolve in time. While these approaches shine when the network structure is known, as is the case for instance in transportation networks (Gallotti and Barthelemy (2015)) or interbank networks (Mazzarisi et al. (2020a)), when the network structure is unknown its inference can be cumbersome and dictates important model selection decisions on how to characterize the hidden Markov dynamics.

Here we tackle the problem of how to efficiently model non-stationarity in the KIM: we focus on its applications to time series analysis and extend it to allow the presence of time-varying parameters with score-driven dynamics (Creal et al. (2013); Harvey (2013)), which is a relatively recent and extremely effective method to describe non-stationary time series.

In its standard form we presented in Chapter 3 the Kinetic Ising Model for time series involves three main sets of parameters: a $N \times N$ interaction

or coupling matrix J, a N-dimensional vector h and a $N \times K$ matrix b characterizing the interaction with external covariates $x(t) \in \mathbb{R}^K$. The model is Markovian with synchronous dynamics, characterized by the transition probability

$$p(s(t+1)|s(t), x(t); \beta, J, h, b) = \\ = Z^{-1}(t) \exp\left[\beta \sum_i s_i(t+1)\left[\sum_j J_{ij} s_j(t) + h_i + \sum_k b_{ik} x_k(t)\right]\right] \quad (5.1)$$

where $Z(t)$ is a normalizing constant and β is a parameter that determines the amount of noise in the dynamics, known as the *inverse temperature*; the smaller is β, the more the dynamics of the $s(t)$ evolves randomly, to the point that, in the limit $\beta \to 0$, $s(t)$ becomes a vector of independent Bernoulli random variables with parameter 0.5, while if $\beta \to +\infty$ the dynamics becomes fully deterministic. Typically the quantity inside the inner brackets of Eq. 5.1 is called the *effective field* perceived by spin i at time t, and in the following we will refer to it as $g_i(t) = \sum_j J_{ij} s_j(t) + h_i + \sum_k b_{ik} x_k(t)$. For ease of notation we can also define the set of static parameters of the KIM, $\Theta = (J, h, b)$.

There are two main reasons that motivate our interest in developing an effective non-stationary version of this model: the first is that, as we will argue in the following paragraphs, the introduction of a time-varying noise parameter $\beta(t)$ allows to better understand the role of noise in the dynamics, quantifying the level of noise at any point in time and thus leading to more informed forecasts; the second is that by introducing a convenient factorization for the model parameters it is possible to discriminate whether an observation is more or less explained by endogenous interactions with other variables or by exogenous effects, offering better insight on the dy-

namics that generated the data even when these effects are not constant over time. As mentioned, the non-stationarity of parameters is a common problem to complex systems such as financial markets, where for instance it is widely accepted that the volatility of returns is time-dependent, but also to brain networks where the processing of time-varying stimuli (Nghiem et al. (2017); Ferrari et al. (2018); Nghiem et al. (2020)) or the spontaneous emergence of thought (Mooneyham et al. (2017)) have been investigated in recent years with more quantitative methods.

To expand on the first point made above, a more practical representation of the effect of having different noise levels is obtained by deriving the theoretical Area Under the ROC Curve (AUC) for the KIM and observing how it varies as a function of β. The AUC is a standard metric to evaluate the performance of binary classifiers (Hanley and McNeil (1982); Bradley (1997)), which the Kinetic Ising Model de facto is, and relies on the generation of the Receiver Operating Characteristic (ROC) curve based on the predictions $\hat{s}_i(t+1)$ provided by the model.

A ROC curve is a set of points $(FPR(\alpha), TPR(\alpha))$, with $\alpha \in [0,1]$ being a free parameter determining the minimum value of $p(s_i(t+1) = +1|s(t), x(t); \beta, \Theta)$ which is considered to predict $\hat{s}_i(t+1) = 1$. If the prediction $\hat{s}_i(t+1)$ matches the realization $s_i(t+1)$ then the classification is identified as a True Positive (or Negative, if $p < \alpha$), otherwise it is identified as a False Positive (Negative). The True Positive Rate (TPR) is the ratio of True Positives to the total number of realized Positives, that is True Positives plus False Negatives. Similarly the False Positive Rate (FPR) is the ratio of False Positives to the total number of realized Negatives. Summarizing

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN}$$

We can explicitly derive the analytical form of the theoretical AUC, that is the area that lies below the set of points $(FPR(\alpha), TPR(\alpha))$, assuming the data generating process is well specified and performing some assumptions on the distribution of the model parameters. As a reminder, a classifier having $AUC = 0.5$ is called an *uninformed classifier*, meaning it makes predictions statistically indistinguishable from random guessing, while values of AUC greater than 0.5 are a sign of good forecasting capability. Following the definition of TPR and FPR one can compute their expected values

$$TPR_\phi(\alpha, \beta) = \frac{1}{Z_\phi^+(\beta)} \int_{g_i : p^+ > \alpha} dg_i \phi(g) p^+(\beta, g_i) \tag{5.2a}$$

$$FPR_\phi(\alpha, \beta) = \frac{1}{Z_\phi^-(\beta)} \int_{g_i : p^+ > \alpha} dg_i \phi(g) p^-(\beta, g_i) \tag{5.2b}$$

where $Z_\phi^\pm(\beta) = p(s_i = \pm 1)$ is a normalization function, $\phi(g)$ is the unconditional distribution of the effective fields g_i (which we discuss in more detail in Appendix B) and we have abbreviated the probability of sampling a positive or negative value as

$$p^\pm(\beta, g_i) = \frac{e^{\pm \beta g_i}}{2 \cosh(\beta g_i)}$$

The definition of the theoretical AUC then reads as

$$AUC_\phi(\beta) = \int_1^0 TPR_\phi(\alpha, \beta) \frac{\partial FPR_\phi(\alpha, \beta)}{\partial \alpha} d\alpha$$

that is the area below the set of points $(FPR(\alpha), TPR(\alpha))$. The lower limit to the integration in Eqs. 5.2 is $g_{min} : p^+(g_{min}) = \alpha$, which is found to be

$$g_{min}(\alpha, \beta) = \frac{1}{2\beta} \log \frac{\alpha}{1-\alpha}$$

Then applying the partial derivative to the definition of FPR it follows that

$$\frac{\partial FPR}{\partial \alpha} = -\frac{1}{Z_\phi^-(\beta)} \frac{\partial g_{min}}{\partial \alpha} \phi(g_{min})(1-\alpha)$$

where we have substituted $p^-(\beta, g_{min}) = 1 - \alpha$. Plugging all the above results in the definition of AUC_ϕ we then find

$$AUC_\phi(\beta) = \frac{1}{Z_\phi^+(\beta) Z_\phi^-(\beta)} \int_0^1 d\alpha \left[\int_{g_{min}(\alpha,\beta)}^{+\infty} dg \phi(g) \frac{e^{\beta g}}{2\cosh \beta g} \right] \times \\ \times \left[\frac{1}{2\alpha\beta} \phi(g_{min}(\alpha, \beta)) \right] \quad (5.3)$$

In Figure 5.1 we show the result assuming $\phi(g)$ is a Gaussian distribution with mean g_0 and standard deviation g_1. This is the case for instance if the J_{ij} entries are Gaussian distributed with zero mean as we show in Appendix B, since g would become a sum of Gaussian variables with random signs given by the values of $s(t)$. We see that the AUC is monotonically increasing with β, but also that the distribution of the static parameters affects the slope with which the curve converges towards 1. Indeed the smaller the mean and variance of the effective fields g_i, the slower the growth of $AUC(\beta)$.

This result would prove extremely useful if it wasn't for the fact that, in the standard form with static parameters of the KIM, β is not identifiable

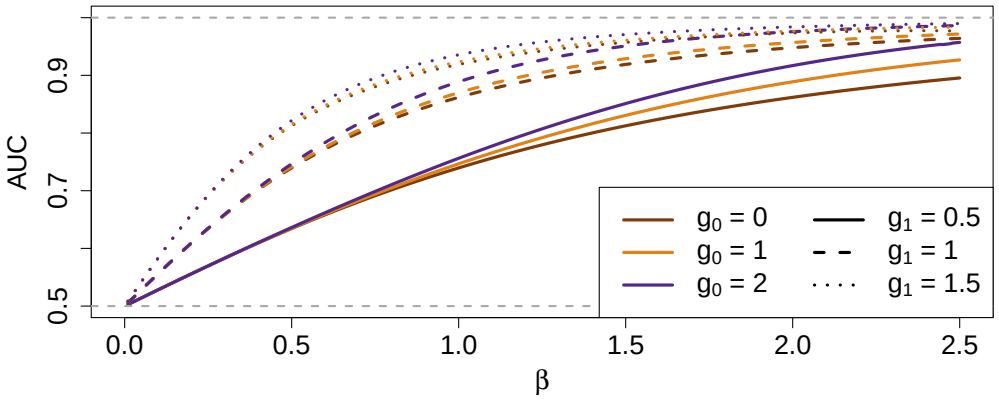

Figure 5.1: Theoretical AUC as a function of β assuming g_i is Gaussian distributed with mean g_0 and standard deviation g_1. We see that increasing β has the effect of reducing the uncertainty on the random variable $s_i(t+1)$, keeping g_i unchanged. Grey dashed lines at AUC $= 0.5$ and AUC $= 1$ are guides to the eye.

(Sakellariou (2013)): indeed it is a common multiplying factor to all the other parameters, meaning that for any two values β_1 and β_2 there are also two sets of parameters Θ_1 and Θ_2 such that $p(s(t+1)|s(t);\beta_1,\Theta_1) = p(s(t+1)|s(t);\beta_2,\Theta_2)$ for all s(t). For this reason in inference problems it is typically assumed that $\beta = 1$ incorporating its effect in the size of the other parameters.

As we will see in more detail in Section 5.2 there is a way in which the β can be identified, and it relies on relaxing the assumption that β is constant throughout the whole sample. If the β of Eq. 5.1 is allowed to be time-varying the identification problem is limited to its average value (which still needs to be assumed equal to 1), while its local value can be inferred from the data using suitable methods. It is clear that the presence of a time-varying parameter implies the necessity to complicate the model to describe the dynamical laws of the parameter, but thanks to the score-driven methodology we propose it is actually both very easy and very efficient to

do so.

This result has implications particularly for forecasting applications: a forecast should be considered more or less reliable by looking at the value of $\beta(t)$ at the previous instant in time and considering how well above 0.5 the corresponding expected AUC is. In Section 5.2 we introduce a dynamic β specification of the KIM which is designed to capture this effect, which we then apply in Section 5.4 to a financial setting.

Having stated some of the motivations that move us towards the development of a non-stationary KIM, let us set the stage to introduce score-driven models by briefly reviewing the theory of time-varying parameters models in discrete time. There is a rich literature on the topic, which has been summarized in the review by Tucci (1995) and more recently by Koopman et al. (2016). In general, a time-varying parameters model can be written as

$$y(t) \sim p(y(t)|f(t), \mathcal{Y}(t-1), \Phi) \tag{5.4a}$$

$$f_t = \psi(f(t-1), f(t-2), ..., \mathcal{Y}(t-1), \epsilon(t), \Phi) \tag{5.4b}$$

where $y(t)$ is a vector of observations sampled from the probability distribution function p , $\mathcal{Y}(t-1)$ is the set of all observations up to time $t-1$ and $f(t)$ are the parameters which are assumed to be time varying. The dynamics of those parameters can either depend on past observations, on past values of the same parameters, on some external noise $\epsilon(t)$ and on a set of static parameters Φ.

If the function ψ only contains past values of the time-varying parameters, a noise term and the static parameters, then the model is called a *parameter-driven* model, whereas if the function ψ can be written as a deter-

ministic function of past observations only, it is called an *observation-driven* model (Cox et al. (1981)).

Examples for parameter-driven models can be found in the financial econometrics literature looking at the Stochastic Volatility models (Tauchen and Pitts (1983); Shephard (2005)), as well as other examples as Bauwens and Veredas (2004) or Hafner and Manner (2012).

The other family is the one of observation-driven models, whose probably most celebrated example is the Generalized AutoRegressive Conditional Heteroscedasticity (GARCH) model of Bollerslev (1986), where a time series of log-returns is modelled using a time-varying volatility parameter depending deterministically on squared observations up to that time and past values of volatilities.

The main advantage of adopting an observation-driven model rather than a parameter-driven one lies in its estimation: having time-varying parameters that only depend on observations through a set of static parameters results in a strong reduction of complexity in writing the likelihood of the model, whereas the calculations for most non-trivial parameter-driven models are typically extremely convoluted and computationally intensive.

In this work we focus on one specific class of observation-driven models, the one of score-driven or Generalized Autoregressive Score (GAS) models, and their implementation in the case of the Kinetic Ising Model. Originally introduced by Creal et al. (2013) and Harvey (2013), they postulate that time-varying parameters depend on observations through the score of the conditional likelihood, that is its gradient.

To better introduce the score-driven methodology, let us consider a sequence of observations $\{y(t)\}_{t=1}^{T}$, where each $y(t) \in \mathbb{R}^N$, and let us define a model with conditional probability density $p(y(t)|f(t))$ depending

on a vector of time-varying parameters $f(t) \in \mathbb{R}^M$. Defining the score as $\nabla_t = \frac{\partial \log p(y(t)|f(t))}{\partial f(t)}$, a score-driven model assumes that the time evolution of $f(t)$ is ruled by the recursive relation

$$f(t+1) = w + Bf(t) + A\mathcal{I}^{-1/2}(t)\nabla_t \tag{5.5}$$

where w, B and A are a set of static parameters. In this generic form, w is a M-dimensional vector, while A and B are $M \times M$ matrices. $\mathcal{I}^{-1/2}(t)$ is also a $M \times M$ matrix, that we choose to be the inverse of the square root of the Fisher information matrix associated with $p(y(t)|f(t))$. This is not the only possible choice for this rescaling matrix (Creal et al. (2013)) but we will keep it this way throughout this article as it is the most intuitive way of rescaling the score.

As is clear from Eq. 5.5, the score drives the time evolution of $f(t)$. This means that given a form of $p(y(t)|f(t))$ the sampling of the observations from this distribution results in a deterministic update of the time-varying parameters. The update can remind the reader of a Newton-like method for optimization, in that the parameters are moved towards the maximum of the likelihood at each realization of the observations while keeping track of the time evolution through the B static parameter.

Another reason to implement a score-driven model is provided by results (Blasques et al. (2015, 2017)) from information theory about the optimality of this approach compared to any other observation-driven method.

Finally, the score-driven modelling approach provides access to a simple statistical test, developed by Calvori et al. (2017), which tests whether it is reasonable to assume that a given parameter is time-varying. This is of crucial importance when estimating a model parameters on data, as knowing whether the parameter can be considered static or should be assumed to

be time-varying helps in the definition of models that extract more relevant informations from the data and are less prone to overfitting or underfitting problems.

The chapter is structured as follows: in Section 5.2 we formalize two implementations of score-driven Kinetic Ising Models, the Dynamical Noise KIM (DyNoKIM) and the Dynamic Endogeneity KIM (DyEKIM); then in Section 5.3 we provide a number of tests on simulated data to assess the consistency of the estimation and to showcase the utility of score-driven modelling; in Section 5.4 we offer three example applications to financial data of the two models; Section 5.5 concludes the chapter.

5.2 The Score-Driven KIM

The Dynamical Noise KIM

In this section we define the Dynamical Noise Kinetic Ising Model (DyNoKIM), where as anticipated the noise parameter β of Eq. 5.1 is considered to be time-varying, which we assume to be modelled by a score-driven dynamics, To keep the formulas concise, we impose that $h_i = b_{ik} = 0$ for all i, k as it is straightforward to extend the results for any value of h and b. This leads to writing the transition probability as

$$p(s(t+1)|s(t); J, \beta(t)) = \ = Z^{-1}(t) \prod_i \exp\left[\beta(t) \sum_j s_i(t+1) J_{ij} s_j(t)\right] \tag{5.6}$$

with $Z(t) = \prod_i 2\cosh\left[\beta(t) \sum_j J_{ij} s_j(t)\right]$.

The interpretation for this model is simple yet extremely useful: the higher the value of β, the smaller the uncertainty over the realization of

$s(t+1)$ or, in other words, the more accurate a prediction of the value of $s(t+1)$, as we have shown in Fig. 5.1.

We still have not explicitly introduced the dynamic rule of motion for the time-varying parameter $\beta(t)$, which, as was stated above, we choose to be score-driven. We define the parameter to be positive to represent the inverse of a noise, and thus we define the update equation for its logarithm, letting $f(t) = \log \beta(t)$

$$\log \beta(t+1) = w + B \log \beta(t) + A\mathcal{I}^{-1/2}(t)\nabla_t \tag{5.7}$$

where w, B and A are parameters to be inferred by Maximum Likelihood Estimation (MLE) and $\mathcal{I}$ is the Fisher Information matrix.

The last term in Eq. 5.7 includes the score, which is the derivative of the log-likelihood $\mathcal{L}$ at a given time t with respect to the time-varying parameter $\log \beta(t)$, reading

$$\nabla_t = \beta(t) \sum_i \left(s_i(t+1) - \tanh \left[\beta(t) \sum_j J_{ij} s_j(t) \right] \right) \sum_j J_{ij} s_j(t) \tag{5.8}$$

The score is rescaled by the inverse of the square root of the Fisher Information, which is used to regularize its impact at different times by considering the convexity of the log-likelihood. The Fisher Information corresponds to the expectation of the Hessian of the log-likelihood, changed in sign and evaluated at time t

$$\mathcal{I}(t) = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}(t)}{\partial (\log \beta)^2} \right]_{\beta(t)} = -\beta(t)^2 \frac{\partial^2 \mathcal{L}(t)}{\partial \beta^2} \Big|_{\beta(t)}$$

where

$$\frac{\partial^2 \mathcal{L}(t)}{\partial \beta^2}\Big|_{\beta(t)} = -\sum_i \left(1 - \tanh^2\left[\beta(t)\sum_j J_{ij}s_j(t)\right]\right)\left(\sum_j J_{ij}s_j(t)\right)^2$$

and the expectation can be dropped as the above equation does not depend on the observation $s(t+1)$.

In the statistical physics literature there have been several attempts to study similar models: some examples are Penney et al. (1993) where a model very similar to the one of Eq. 5.1 is considered, or the literature on superstatistics of Beck and Cohen (2003) and Beck et al. (2005) which provides a general theory for physical systems with non-static parameters and in particular studies models where a time-varying noise parameter takes the role of $\beta(t)$ in Eq. 5.6. There is however one important difference, which is related to the assumption of local equilibrium and time scale separation that is common to all the cited works. The authors assume that the sampling of the observations and of the time-varying parameters take place on two separated time scales, meaning that the time-varying parameters are locally constant when the observations are sampled. This is not true for score-driven models, which are in fact designed to not require this assumption, intuitively formalized by the values of the parameters B and A. If $B \gg A$ then the evolution of f is indeed slower than the one of observations, while if $B \ll A$ they evolve on the same time scale.

The estimation of the model can be done in two steps, first estimating a static version where $\beta(t) = 1 \; \forall \, t$ in order to infer the static parameters J, and then proceeding to the estimation of the score-driven part. In short, the model is estimated first as a static Kinetic Ising Model, fitting only the J and h parameters following the procedure of Sakellariou (2013), and then a standard gradient descent algorithm for optimization (Kingma and Ba

(2014)) is used to fit the w, B and A parameters related to the score-driven dynamics. While in principle a joint estimation procedure of all parameters would be possible, this two-step procedure is to be preferred as it does not require to apply the filter of Eq. 5.7 at every iteration of the gradient descent method when estimating J and h, a feature that significantly reduces the computational cost of the inference.

In Section 5.3 we provide simulation results to validate this estimation procedure, while later in Section 5.4 we show an empirical application of the DyNoKIM to forecasting stock price changes at high frequency.

The Dynamic Endogeneity KIM

The second specification of the score-driven Kinetic Ising Model we explore in this article is the Dynamic Endogeneity Kinetic Ising Model (DyEKIM). In the DyEKIM we let the number of time-varying parameters be a bit larger, assuming that the J, h and b parameters each have their own specific time-varying factorization. In principle these choices are up to the modeller, depending on the specific application and data: here we present one factorization we believe is a reasonable choice for the financial applications we propose in Section 5.4, albeit other implementations could be possible too. Going back to Eq. 5.1, we now impose the following structure to each of the time-varying parameters:

$$
\begin{aligned}
J_{ij}(t) &= \beta_{diag}(t) J_{ii}\delta_{ij} + \beta_{off}(t) J_{ij}(1-\delta_{ij}) \\
h_i(t) &= \beta_h(t)(h_i + h_0(t)) \\
b_{ik}(t) &= \beta_k(t) b_{ik} x_k(t)
\end{aligned}
$$

where δ_{ij} here represents the Kronecker symbol which is 1 if $i = j$ and

0 otherwise. The conditional probability density for this model, calling $\boldsymbol{\beta}(t) = (\beta_{diag}, \beta_{off}, \beta_h, \{\beta_k\})$, reads

$$
\begin{aligned}
&p(s(t+1)|s(t), x(t); J, h, b, h_0(t), \boldsymbol{\beta}(t)) = \\
&= Z^{-1}(t) \exp\Big[\sum_i s_i(t+1)\Big[\beta_{diag}(t) J_{ii} s(t) + \beta_{off}(t) \sum_{j\neq i}(t) J_{ij} s_j(t) + \\
&+ \beta_h(t)(h_i + h_0(t)) + \sum_k \beta_k(t) b_{ik} x_k(t)\Big]\Big] \qquad (5.9)
\end{aligned}
$$

This change in the form of the model radically changes the interpretation one gives to the values of $\boldsymbol{\beta}(t)$. While it still has the role of modulating the relevance of parameters, and thus the entropy is still smaller when increasing any component of $\boldsymbol{\beta}$, the main effect is establishing how important are autocorrelations and lagged cross-correlations among spins compared to idiosyncratic or external effects at any point in time. This model can then be used to describe data where the dynamics of the variables is dependent on others at intermittent times, disentangling network effects from idiosyncratic dynamics or exogenous effects in a time-varying fashion.

We will discuss in more detail the specific interpretation for each of the time-varying parameters in the empirical applications of the second and third part of Section 5.4. The intuition behind this choice however is that we want to be able to discriminate between different components of the dynamics observed in a set of variables: one associated to external inputs (β_k), one to the idiosyncratic properties of variable i (β_h), as well as general trends (h_0), one for autocorrelations (β_{diag}) and finally one for lagged cross-correlations among variables (β_{off}). In this formulation then each of these time-varying parameters provides insight on the relative importance of one term over the others in the generation of the data, highlighting periods of higher or lower endogeneity of the dynamics (when correlations have higher

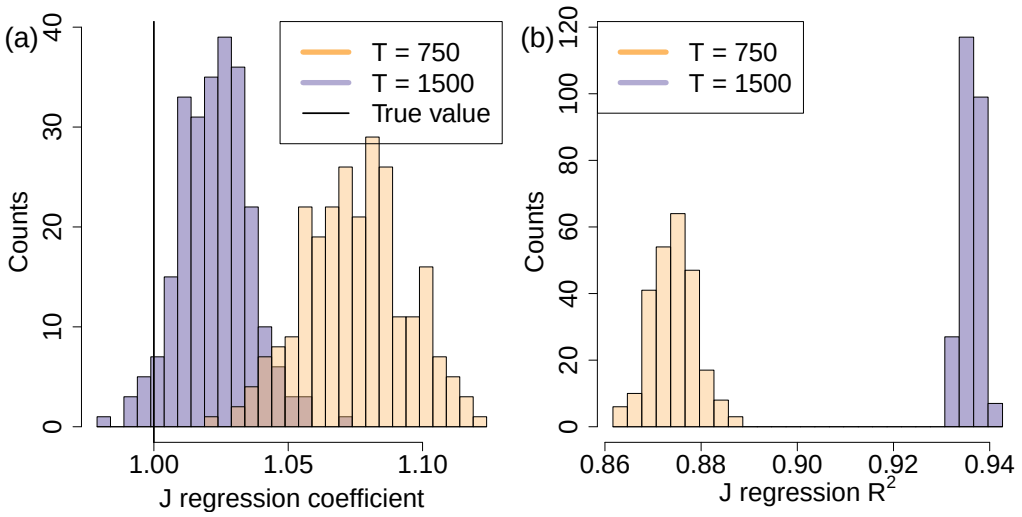

Figure 5.2: Consistency of the J matrix estimation. (a) Histogram of linear regression coefficients b between inferred and true values of J_{ij} over 250 samples for $N = 50$, $T = 750$ and $T = 1500$; (b) Histogram of coefficients of determination (R^2) for the same set of models. The convergence of both values towards 1 when increasing T is a sign of consistency of the estimation.

importance) rather than periods where the dynamics is more idiosyncratic or exogenously driven.

Regarding the estimation, the procedure is largely the same as the one for the previous model. There are however a couple of subtleties that need to be pointed out, regarding the structure of the B and A parameters and of the Fisher Information $\mathcal{I}$, which are now matrices. In order to make the estimation less computationally demanding in our example applications we choose to assume A, B and $\mathcal{I}$ diagonal, disregarding the dependencies between time-varying parameters: this will likely make our estimates less precise, but it also reduces the number of static parameters to be inferred, letting us bypass model selection decisions which are outside the scope of this article.

As a last remark, notice that the DyNoKIM and the DyEKIM are equivalent when $h_0(t) = 0 \; \forall \, t$ and $\beta_{diag} = \beta_{off} = \beta_h = \beta_k = \beta$. In the next sec-

tion we mainly present simulation results for the DyNoKIM alone to keep the manuscript concise, as we found no significant differences between the two models when it comes to the reliability of the estimation process, and later apply them to real-world scenarios where their interpretation is much more meaningful.

5.3 Estimation on simulated data

DyNoKIM - consistency, filtering and forecasting

We start our analysis from a consistency test on simulated data, aimed at understanding whether the two-step estimation procedure we outlined above is able to recover the values of the parameters of the model when the model itself generated the data.

Here we report results for simulations run with parameters $N = 50$, $T = 750$ or $T = 1500$, $J_{ij} \sim \mathcal{N}(0, 1/\sqrt{N})$, $h_i = 0 \ \forall \, i$, $B = 0.95$ and $A = 0.01$. We see from Fig. 5.2 that the estimation of the elements of J is indeed consistent: we estimate a linear regression model between the estimated and the true values of J_{ij}, namely $J_{ij}^{est} = bJ_{ij}^{true} + a$, and plot the histogram of the values of b and of the coefficient of determination R^2 of the resulting model from 250 simulations and estimations. In the ideal case where for any i, j $J_{ij}^{est} = J_{ij}^{true}$ one would have $b = R^2 = 1$, which is what we aim for in the limit $T \to \infty$. We see from our results that there is indeed a convergence of both values towards 1 when increasing sample size, reducing both the bias and the variance of the regression parameters.

Turning to the score-driven dynamics parameters A and B, the situation does not change significantly. In Fig. 5.3 we show the histograms of estimated values of B and A over 250 simulations of $N = 50$ variables for both

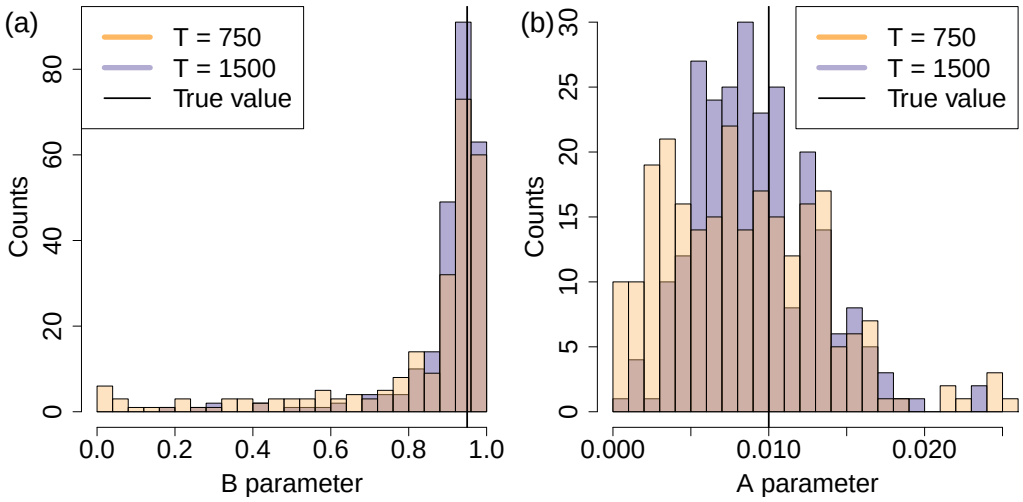

Figure 5.3: Consistency of the score-driven dynamics parameters. (a) Histogram of estimated values of B over 250 samples for $N = 50$, $T = 750$ and $T = 1500$; (b) Histogram of estimated values of A over 250 samples for the same set of models. The convergence towards the true value by increasing T is a sign of consistency of the estimation.

$T = 750$ and $T = 1500$. It again appears clearly that when increasing the sample size the bias and variance of the estimators converge towards 0, with the estimated parameter converging towards its simulated value. Thanks to these results we are able to confidently apply the two-step estimation method without needing to estimate all the parameters at once.

Having shown that the model can be estimated consistently and efficiently, we want to test its performance when the β dynamics is not produced with the score-driven data generating process. Indeed there is little reason to believe that this sort of dynamics is significant for real-world applications, where the dynamics of β might follow exogenous and unknown rules. The power of score-driven models lies also in this feature, in that they are able to estimate time-varying parameters such as $\beta(t)$ without actually needing any assumption on their true dynamical laws. In this sense they behave as filters for the underlying, unknown dynamics of the parameter.

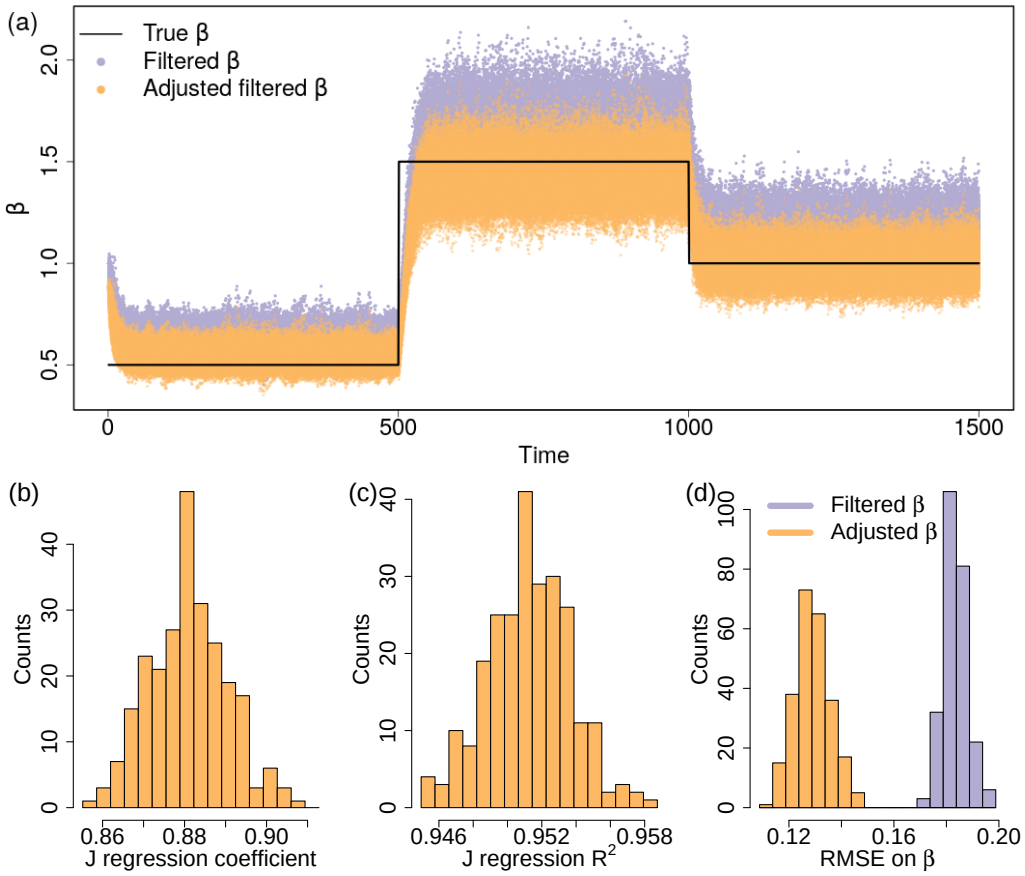

Figure 5.4: Estimation of $\beta(t)$ under model misspecification. (a) Simulated dynamics of $\beta(t)$ in the shape of a double step function and estimated (adjusted by b) β over 250 trials; (b-c) Consistency analysis for J parameters (linear regression coefficient and R^2); (d) Root Mean Squared Error (RMSE) on the estimated $\beta(t)$ compared to the simulated one over 250 trials.

We show in Fig. 5.4 the results for a set of simulations where $\beta(t)$ follows a deterministic piecewise constant dynamics, taking values 0.5, 1.5 and 1 each for 500 timesteps, while the other parameters are $N = 30$, $J_{ij} \sim \mathcal{N}(0, 1/\sqrt{N})$, $h_i = 0 \; \forall \, i$ and the time evolution of $s(t)$ is given by the DyNoKIM with the enforced $\beta(t)$. The estimator is then fed with just the resulting simulated $s(t)$ time series, having the task to reconstruct J and

find a pair of parameters A and B able to produce a dynamics of $\beta(t)$ that resembles the one that produced the data. The consistency analysis for the elements of J shows that there is a bias to underestimate the value of the parameters (in absolute value), however the explained variance expressed by the R^2 coefficient is large, meaning the quality of the fit is high enough. The bias is indeed incorporated in the estimated β, shown in Fig. 5.4a, where one can see that the filtered β is typically above the true value. However, once it gets rescaled by the b parameter of the linear regression $J_{ij}^{inf} = bJ_{ij}^{true} + a$ it fits nicely on the data generating process and the Root Mean Squared Error (RMSE), reported in Fig. 5.4d, reduces accordingly.

Clearly there is no way to determine the rescaling coefficient b from data as there is no J^{true} to compare the inferred parameters with, but it is comforting to see that when the magnitude of J_{ij} is underestimated it is compensated by an equal overestimation of $\langle\beta\rangle$, meaning the overall effect is unchanged (as β always multiplies J).

In Fig. 5.5 we report the analogue of Fig. 5.4a for two other types of $\beta(t)$ dynamics, one a deterministic sine function and the other an AutoRegressive model of order 1 (AR(1)). The frequency of the sine function is chosen to have exactly 5 periods in the simulation length, with an amplitude of 0.5 and mean 1, while the AR(1) model reads

$$\beta^{AR}(t+1) = a_0 + a_1\beta^{AR}(t) + \epsilon(t)$$

where $\epsilon(t) \sim \mathcal{N}(0, \Sigma^2)$ with parameters $a_0 = 0.005$, $a_1 = 0.995$, $\Sigma = 0.01$ so to have $\langle\beta^{AR}\rangle = 1$.

As mentioned in the previous section, DyNoKIM is able to identify time periods when the data is more predictable using a Kinetic Ising Model approach. In Fig. 5.6 we show how the forecasting performance depends on

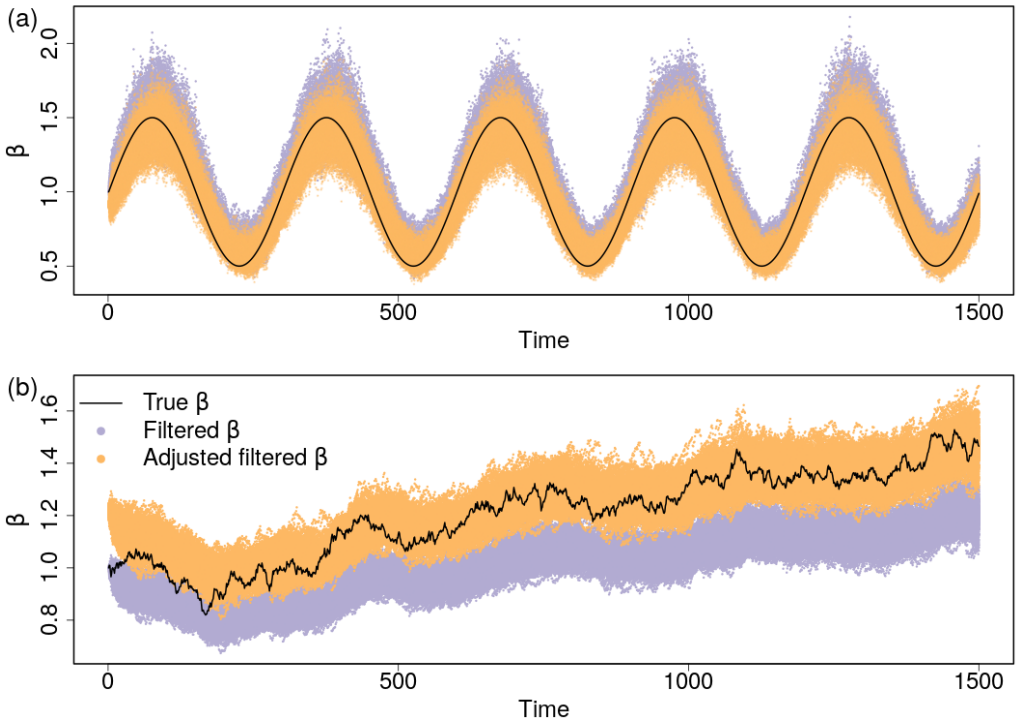

Figure 5.5: Simulation and estimation of a misspecified score-driven model. (a) Deterministic $\beta(t)$ following a sin function; (b) Stochastic $\beta(t)$ following an AutoRegressive model of order 1.

the estimated values of β on a simulated dataset, where an underlying piecewise constant β is used to generate configurations with the DyNoKIM, using the Area Under Curve (AUC) performance metric. We see that the Area Under the Curve of the ROC is significantly dependent on the estimated values of β, meaning that forecasts made at high β values are significantly more reliable than those made at low β values, as predicted by the theory.

DyEKIM - separating multiple effects

In this section we briefly show some simulations results from tests on the DyEKIM, where we want to show that different effects are correctly separated and identified when estimating the model on a misspecified data

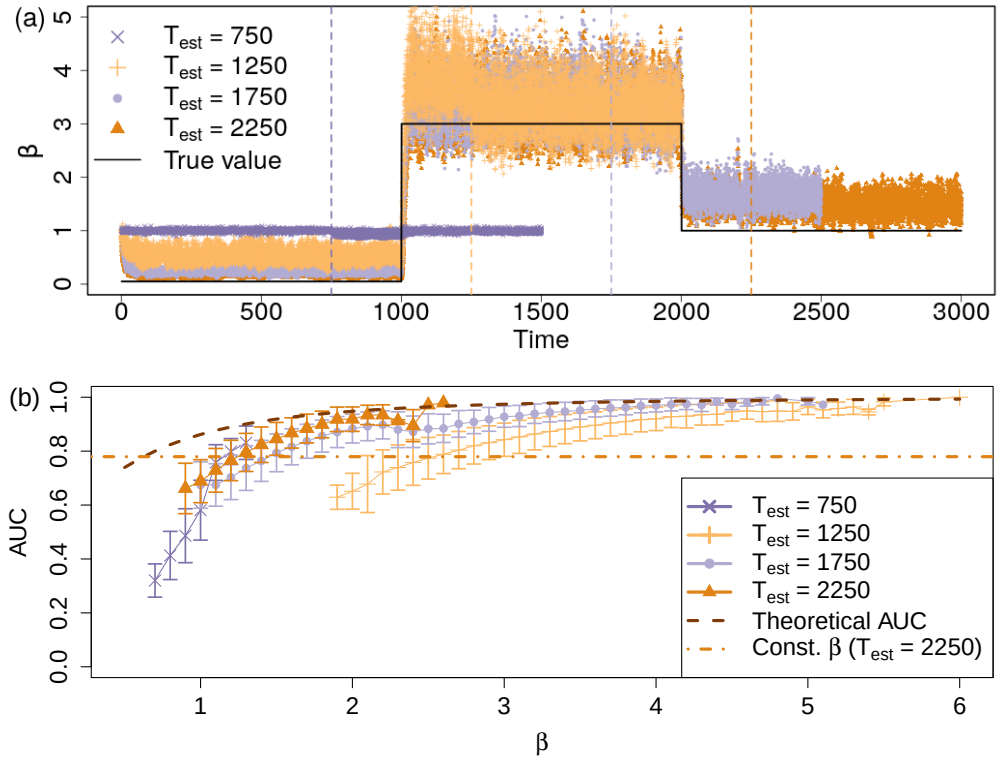

Figure 5.6: Forecasting performance on increasingly long estimation windows on a misspecified data generating process for $\beta(t)$. (a) Estimation of a piecewise constant $\beta(t)$ using as estimation windows 750, 1250, 1750 and 2250 steps; (b) Area Under Curve (AUC) as a function of the estimated β, aggregated in bins and reporting means and 1 standard deviation error bars. The lines indicate the theoretical value prescribed by the theory and the average AUC one would have using no time-varying parameter.

generating process. In fact while the consistency analysis largely resembles the one we reported for the DyNoKIM in Figures 5.2 and 5.3 and for this reason we omit it, the effect of filtering multiple time-varying parameters is something that cannot be predicted by the simulations on the DyNoKIM alone.

In Figure 5.7 we show the results when estimating the DyEKIM on a dataset generated by a Kinetic Ising Model with time-varying $\beta_{diag}(t)$, $\beta_{off}(t)$ and $\beta_h(t)$ as in Eq. 5.9 but where the dynamics of the parameters

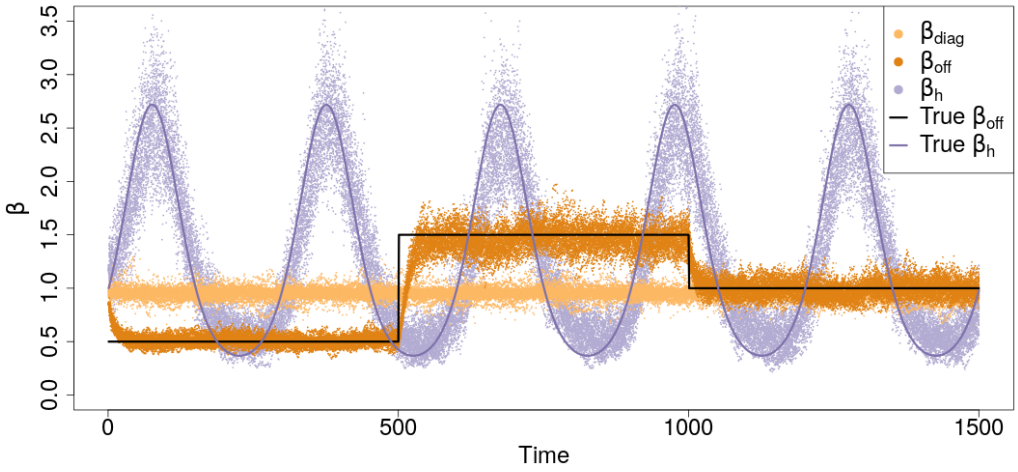

Figure 5.7: Estimation of $\beta_{diag}(t)$, $\beta_{off}(t)$ and $\beta_h(t)$ under model misspecification. The model was simulated with a constant $\beta_{diag}(t) = 1$, a piece-wise constant $\beta_{off}(t)$ and an exponentiated sinusoidal $\beta_h(t) = \exp[\sin(\omega t)]$, with $\omega = 5\frac{2\pi}{T}$. The points are the result of 30 different simulations and estimations, the lines show the values of β_{off} and β_h used to generate the data.

is predetermined instead of following the score-driven update rule. We arbitrarily choose to take a constant $\beta_{diag}(t) = 1$, a piece-wise constant $\beta_{off}(t)$ and an exponentiated sinusoidal $\beta_h(t) = \exp[\sin(\omega t)]$, with $\omega = 5\frac{2\pi}{T}$, $T = 1500$ and $N = 30$. The results show that the filter works correctly and that the different time-varying parameters are consistently estimated, regardless of the kind of dynamics given to each of them.

Having provided evidence that both the DyNoKIM and the DyEKIM can be consistently estimated and have a specific interpretation, in the following section we propose three simple real world applications for our modelling approach, which we apply to high-frequency trading data from the US stock market and from the Foreign Exchange (FX) market.

5.4 Empirical applications

Forecasting stock activity with the DyNoKIM

The first dataset we use is a selection of 11 trading days (November 6 to November 20, 2019) in the 100 largest capitalization stocks in the NASDAQ and NYSE[1], for which we track the events of mid-price change in the Limit Order Book (LOB) at a frequency of 5 seconds. The mid-price is the average of the best bid and best ask occupied price levels in the LOB of a stock, defined for stock i at time t as

$$M_i(t) = \frac{P_i^b(t) + P_i^a(t)}{2}$$

where $P_i^b(t)$ and $P_i^a(t)$ are the best bid and ask prices available in the LOB of stock i at time t. We discretize time in slices of 5 seconds and define for each stock a binary time series $s_i(t)$, taking value $+1$ if the mid-price has changed in the previous 5 seconds and -1 otherwise. The choice of time scale is largely arbitrary: we choose 5 seconds to obtain a set of variables that have unconditional mean as close to 0 as possible to have a balanced dataset. The mid-price movements have been used in the past in the modelling of intensity bursts in market activity (Rambaldi et al. (2015, 2018)), where the authors used Hawkes point processes to investigate how Foreign Exchange markets behave around macroeconomic news, as well as to study how endogenous the price formation mechanism is in financial markets measuring what has been called the "market reflexivity" (Filimonov and Sornette (2012); Hardiman et al. (2013); Filimonov and Sornette (2015); Hardiman and Bouchaud (2014); Wheatley et al. (2019)). It has to be noted

[1]Data provided by LOBSTER academic data - powered by NASDAQ OMX.

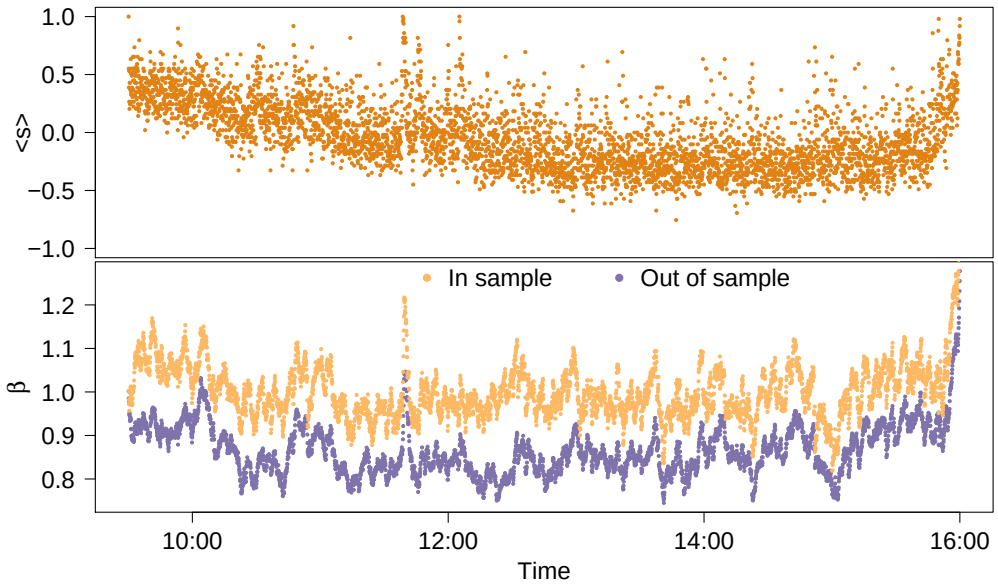

Figure 5.8: Results applying the DyNoKIM to series of US stocks mid-price change events at 5 seconds time scale on November 19, 2019. (top) Cross-sectional mean value of $s(t)$. A value of $\langle s \rangle$ closer to $+1$ indicates a large fraction of stocks has changed price in those 5 seconds and vice-versa if $\langle s \rangle \to -1$; (bottom) Estimated $\beta(t)$ in and out of sample. The estimated amplitude of the dynamics is not huge but still significant and exhibits a sharp rise towards the end of the trading day, which we find in all the analysed days.

that our approach differs from the literature in the time discretization we perform, while the cited approaches all consider continuous-time models.

There are multiple reasons for which the mid-price can change: it can be the arrival of a new limit order at a price which is between the best bid/ask, the cancellation of the last order at the best bid/ask or the execution of a market order which consumes all the limit orders at the best bid/ask. We ignore what causes the movement of the mid-price and focus our attention on the lagged interdependencies among different stocks, by applying the DyNoKIM to the multivariate time series $s(t)$.

We test whether there is reason to assume a time-varying β by perform-

Significance of LM tests				
	SP100 - DyNoKIM	FC - DyEKIM	FOMC	FX
$p < 0.001$	100%	100%	100%	88%
$p < 0.01$	-	-	-	-
$p < 0.05$	-	-	-	4%
$p > 0.05$	-	-	-	8%

Table 5.1: Percentage of p-values of Lagrange Multiplier tests below significance thresholds, divided by dataset. The first column refers to the application of the first part of Section 5.4, the second and third to the second part and the last to the last part. The only non-rejected nulls regard two β_b parameters in the FX dataset, meaning traders don't show significant changes in strategy when it comes to their reactions to prices in those months.

ing the Lagrange Multiplier (LM) test proposed by Calvori et al. (2017) as a generalization of the method by White (1987). In short, the LM test consists in testing the null hypothesis that $\log \beta = f$ is constant in time, that is $f = w$ and $A = B = 0$, against the alternative hypothesis of a time-varying parameter. Calvori et al. (2017) show that the test statistic of the LM test can be written as the Explained Sum of Squares (ESS) of the auxiliary linear regression

$$\mathbf{1} = c_w \nabla_t^0 + c_A \mathcal{S}_{(t-1)}^0 \nabla_t^0 \tag{5.10}$$

where ∇_t^0 is the time t element of the score under the null hypothesis that $f(t) = w \; \forall \, t$, $\mathcal{S}_t^0$ is the time t element of the rescaled score (i.e. $\mathcal{I}^{-1/2}(t)\nabla_t$) under the null, the constants c_w and c_A are estimated by standard linear regression methods and the resulting LM test statistic is distributed as a χ^2 random variable with one degree of freedom. If the null is rejected, the hypothesis that β is time varying is a valid alternative and we can proceed to estimate the score-driven dynamics parameters.

All our empirical results are validated by this preliminary test, for which

we have strong rejections of the null on all samples as reported in the first column of Table 5.1.

In Fig. 5.8 we show the activity we see on a trading day, Tuesday November 19th, 2019. The top panel shows the cross-sectional average of s, that represents the fraction of stocks for which the mid-price has changed in a given time window, while the bottom panel shows the values of $\beta(t)$ we obtain estimating the model parameters on the same day (in sample, yellow points) or on the previous day and using observations to obtain the filtered β value (out of sample, purple points).

We see that there is a pattern in the fraction of moving stock prices over the day, with higher values at the opening and closing times (we do not include opening and closing auctions in our data) as it is typically observed for stock markets, and that a similar pattern is observed for β throughout the day. We also notice that the out of sample β is always lower than the in sample one - as should be expected, since it utilizes parameters which are not MLE - but it follows the same overall behaviour.

Our theoretical and simulation results from Figures 5.1 and 5.6 suggest to use our estimates of β to quantify the reliability of forecasts using this model: we thus estimate the model parameters once per day and use them to filter $\beta(t)$ on the next day, while checking the accuracy with which the model predicts mid-price movements out of sample using the AUC metric.

The forecasts $\hat{s}_i(t+1)$ are produced according to

$$\hat{s}_i(t+1) = \text{sign}\left[p\left(s_i(t+1)=1|s(t), J, h, \beta(t-1)\right) - \alpha\right] \tag{5.11}$$

and the ROC curves are obtained varying the value of α between 0 and 1. Notice that we take $\beta(t-1)$ instead of $\beta(t)$ as in the original Eq. 5.6: the reason is that in order to estimate $\beta(t)$ we need the observation of

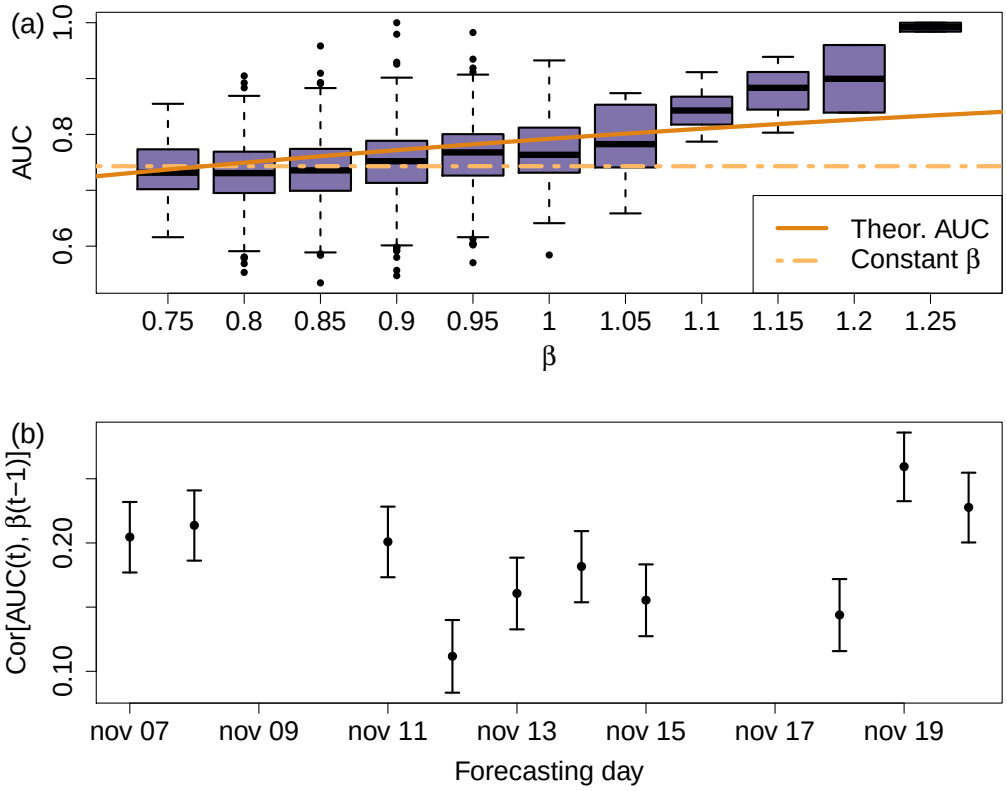

Figure 5.9: AUC statistics compared to $\beta(t-1)$ forecasting US stocks mid-price change events at 5 seconds time scale. (a) AUC values for November 19, 2019 aggregated for different values of $\beta(t-1)$ compared to the theoretical AUC in the hypothesis of Gaussian effective fields g_i and to the average performance with a constant β; (b) Pearson's correlation coefficient between $\beta(t-1)$ and $AUC(t)$ estimated daily with 95% confidence intervals.

$s(t+1)$, which is what we are trying to predict instead, thus we take the last available estimate of β as a proxy for the current value. In this way our prediction is fully causal.

We show results of this analysis in Fig. 5.9. When looking at a single day we see an upward trend in the AUC score as a function of the value of $\beta(t-1)$, meaning that the higher the estimated β the more reliable the forecast can be considered. Comparing our results to the theoretical value that the AUC should take if the distribution of effective fields g_i were

Gaussian we see that the empirical results are in good agreement with the theoretical prediction, however since the actual fields we measure are non-Gaussian the match is not perfect. A further aggregated measure is shown in Fig. 5.9b by looking at the correlation coefficient between AUC and $\beta(t-1)$, estimated daily. Again we see how the correlation is significantly positive, with 95% confidence bands well above 0.

This simple example proves that our theoretical results for the DyNoKIM are indeed verified in realistic applications and that using this method - which we believe could be applied even to more sophisticated models - can result in a significant gain in the use of forecasting models, giving a simple criterion to discriminate when to trust (or not) the forecasts.

Endogenous vs exogenous price activity

In another application to a stock prices dataset, we analyze two events that caused turmoil in the stock markets on the intraday level as an example application of our second kind of score-driven Kinetic Ising Model, the Dynamic Endogeneity KIM (DyEKIM). The two events we choose to analyze are the Flash Crash of May 6, 2010 and the Federal Open Market Committee announcement of July 31, 2019. The Flash Crash marked a historic event for electronic markets, when a seemingly unjustifiable sudden drop in the price of E-mini S&P 500 futures contracts caused all major stock indices to plummet in a matter of a few minutes, including the biggest to date one-day point decline for the Dow Jones Industrial Average and an overall loss of over 5% value across markets. The markets then stabilized and recovered most of the losses when circuit breakers came into place in the original venue (the Chicago Mercantile Exchange) (Securities et al. (2010)). Multiple explanations of what happened have been offered by a large num-

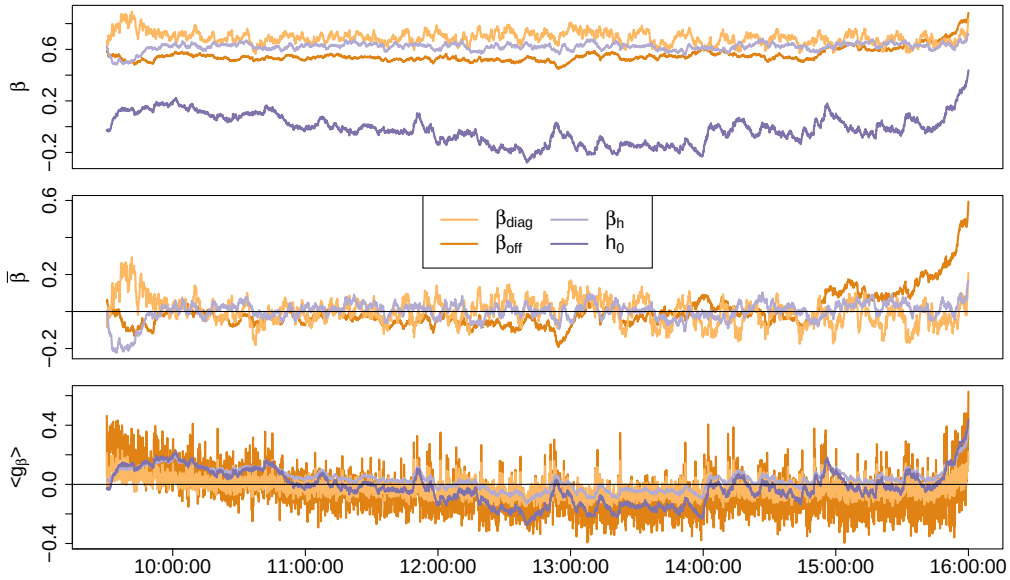

Figure 5.10: Values of $\boldsymbol{\beta}(t)$, $\bar{\boldsymbol{\beta}}(t)$, $\langle g_\beta \rangle(t)$ and $h_0(t)$ on a regular trading day, November 12, 2019. We see that the endogenous components of g have larger values at the beginning and the end of the day, while the exogenous g_{β_h} only grows towards closing. The most varying β parameter is the one related to cross-correlations, β_{off}, which has a very significant increase towards market closure.

ber of academics, regulators and practitioners: CFTC-SEC officials initially attributed responsibility to a "fat-finger trade" by a mutual fund unloading its inventory through an unsophisticated sell algorithm, triggering a liquidity crisis in the futures and stock markets.

Following the official report, alternative explanations challenging this view were presented, as in Easley et al. (2011), where they argue that the state of liquidity had deteriorated prior to the start of the crash and that liquidity providers, in the form of High-Frequency Traders (HFT) and market makers, turned their backs on the market as soon as the distress rose, becoming liquidity consumers. Madhavan (2012) does not take position on the cause but argues that market fragmentation, that is the fact that

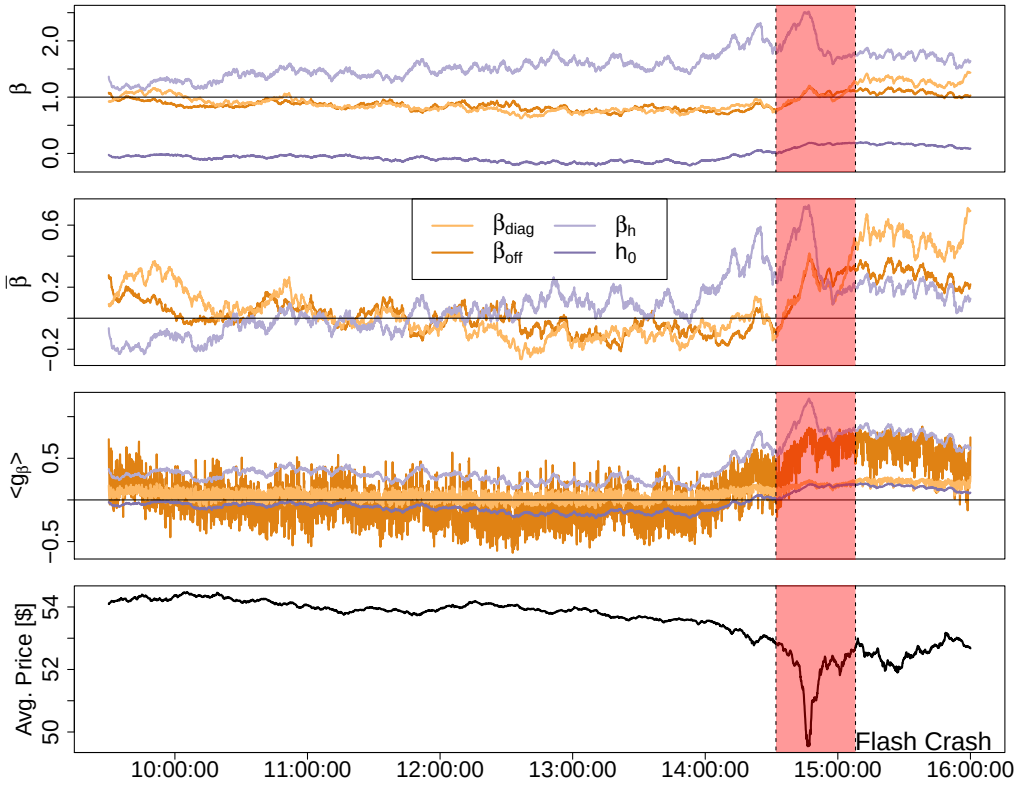

Figure 5.11: Values of $\boldsymbol{\beta}(t)$, $\bar{\boldsymbol{\beta}}(t)$, $\langle g_\beta \rangle(t)$ and $h_0(t)$ on May 6, 2010, along with the average midprice across the S&P100 stocks. The red shade highlights the time window (14:32:00 to 15:08:00 EST) where the Flash Crash takes place. We see that the average activity parameter h_0 starts increasing in the 45 minutes preceding the crash, while during the crash a bigger role is played by the correlation parameters β_{diag} and β_{off}.

the same financial instrument can be traded on multiple markets, causes liquidity provision to be more susceptible to transitory order imbalances, a view that is confirmed by Menkveld and Yueshen (2019). Finally, Kirilenko et al. (2017) analyze trading records by market participants and find that in terms of *executed* orders the behavior of HFTs had not changed during the Flash Crash, while traditional intermediaries acted according to their limited risk-bearing capacity and did not absorb the shock in full. This difference although does not mean that HFTs did not contribute to the

amplification of the liquidity crisis, as the authors argue that they operate significantly different strategies from traditional market makers, including *quote sniping* (or *latency arbitrage*) which is harmful to liquidity provision (Aquilina et al. (2020)).

The other event we analyze is the announcement following the Federal Open Market Committee (FOMC) meeting of July 31, 2019. In this recent meeting the Federal Reserve operated its first interest rate cut in over a decade, the last one dating back to the 2008 financial crisis, encountering mixed reactions in both the news and the markets. In particular an answer to a question in the Q&A press conference by the Fed Chairman Powell has been highlighted by news agencies, when being asked whether further cuts in the future meetings were an option, he answered "we're thinking of it essentially as a midcycle adjustment to policy" (Powell (2019)). This answer triggered turmoil in the equity markets, with all major indices dropping around 2% in a few minutes.

Our analysis focuses again for both events on midprice movements for the then S&P100-indexed stocks at the 5 seconds time scale[2]. Differently from the previous example, here we apply the DyEKIM methodology to study variations in the relative importance of different sets of parameters as events unfold, as defined in Eq. 5.9. In this setting we include no covariates $x_k(t)$, so there is no b parameter matrix and consequently no $\beta_k(t)$. As usual we begin by running the Lagrange Multiplier test on each of the hypothesized time-varying parameters, obtaining that all the nulls are rejected on both datasets as summarized in the second and third columns of Table 5.1. To exclude dependencies between the tests we take as null models both the completely static model (i.e. where all the time-varying

[2]Data provided by LOBSTER academic data - powered by NASDAQ OMX.

parameters are constant) and the model where all the parameters are time-varying except the one being tested, obtaining similar results regardless of the choice.

In order to better understand the results of this modelling approach, we first need to define two quantities based on the filtered values of β that capture different aspects of the effect the time-varying parameter has. First of all, we want to understand how each β varies compared to its own average value: different βs might differ in their unconditional mean, and studying their variation with respect to that base level can highlight effects that can be overshadowed by the fact that a β has a larger average value. To this end we introduce the quantity

$$\bar{\beta}(t) = \frac{\beta(t) - \mathbb{E}[\beta(t)]}{\mathbb{E}[\beta(t)]} \tag{5.12}$$

where $\mathbb{E}[\beta(t)]$ is the sample mean of $\beta(t)$. Another quantity we study is the value taken by the components of the effective fields $g_i(t)$, which can be subdivided in their components related to each of our time-varying parameters. In particular, we define

$$\begin{aligned} g_i(t) &= g_{i,\beta_{diag}}(t) + g_{i,\beta_{off}}(t) + g_{i,\beta_h}(t) \\ g_{i,\beta_{diag}}(t) &= \beta_{diag}(t) J_{ii} s_i(t) \\ g_{i,\beta_{off}}(t) &= \beta_{off}(t) \sum_j J_{ij} s_j(t) \\ g_{i,\beta_h}(t) &= \beta_h(t)(h_i + h_0(t)) \end{aligned}$$

which we then mediate across all indices i, obtaining the quantities $\langle g_{\beta_{diag}} \rangle(t)$ and so on.

The way to interpret these quantities follows from the interpretation the various time-varying parameters have: as mentioned in the definition of the

model, the β_{diag} parameter captures the level of endogeneity in the dynamics related to auto-correlation in the time series; β_{off} is related to endogeneity in the form of lagged cross-correlations; β_h instead models the level to which the observations are close to realizations of independent Bernoulli random variables, unconditional of previously observed values - that is, they are not dependent from any other modelled variable, thus linking to exogenous effects - and h_0 shifts up or down the mean of these independent Bernoulli, thus capturing purely exogenous effects on the dynamics. What the $\langle g_\beta \rangle(t)$ quantities show then is intuitively related to what the explained sum of squares means for linear regression models, in the sense that the more a $\langle g_\beta \rangle(t)$ is far from 0 relative to others the more the data reflect a dynamics that is modelled by that subset of parameters. We choose to show these quantities as a simple way of assessing the relevance of the components, a problem that is not easily solved in these kinds of models. One potential candidate to better quantify these effects is provided by dominance analysis (Budescu (1993); Azen and Budescu (2003)), which to the best our knowledge has only been applied in the framework of multiple logistic regressions but never to autoregressive models and whose generalization goes beyond the scope of this article.

Since the baseline model is applied to stock midprice changes at high frequency, typically called the *activity* of a stock which is taken as a proxy of high-frequency volatility (Filimonov and Sornette (2012); Hardiman et al. (2013)), the interpretation of these time-varying parameters relates to volatility clustering in the case of β_{diag}, to volatility spillovers for β_{off}, to higher or lower market-wise volatility for h_0 and the relevance of exogenous effects is given by β_h.

In Figure 5.10 we show results for these quantities on a regular trading

day, November 12, 2019, which show the typical intraday patterns one can observe from the values of $\boldsymbol{\beta}$(t). We see that the J-related parameters, β_{diag} and β_{off}, as well as the corresponding g components, show a U-shaped pattern throughout the trading day, having higher values at the opening and closing, while the h-related parameter β_h only shows an increase towards the end of the day. The h_0 parameter, which captures the average exogenous price activity across all stocks, shows itself a U-shaped pattern which is more pronounced at closing, consistent with the intraday pattern typical of traded volume.

Figure 5.11 shows the same quantities during the Flash Crash of May 6, 2010. Here the situation appears to be radically different from the one of Figure 5.10: the parameters show a huge variation around the crash, with an abnormal increase in quantities related to β_h in the 45 minutes preceding the crash followed by a similar increase of the endogeneity parameters β_{diag} and β_{off} during the event, which then stay relevant until market close. The intraday pattern is overshadowed by the effect of the crash, but the picture at the beginning of the day is similar to normal trading days. These measurements are consistent with the reconstruction of how events unfolded, with an abnormal exogenous increase in activity starting the crash, which is then amplified by endogenous mechanisms of volatility spillovers. Of note, the endogeneity parameters persist at relatively high values in the aftermath of the crash, indicating that the turmoil induced by the Flash Crash reverberated for the remainder of the trading hours, even after the prices had recovered at pre-crash levels.

Moving on to the recent FOMC announcement of July 31, 2019, in Figure 5.12 we show the values of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, $\langle g_\beta \rangle$ and the average stock price of the S&P100-listed stocks we consider in the analysis (which are a different set

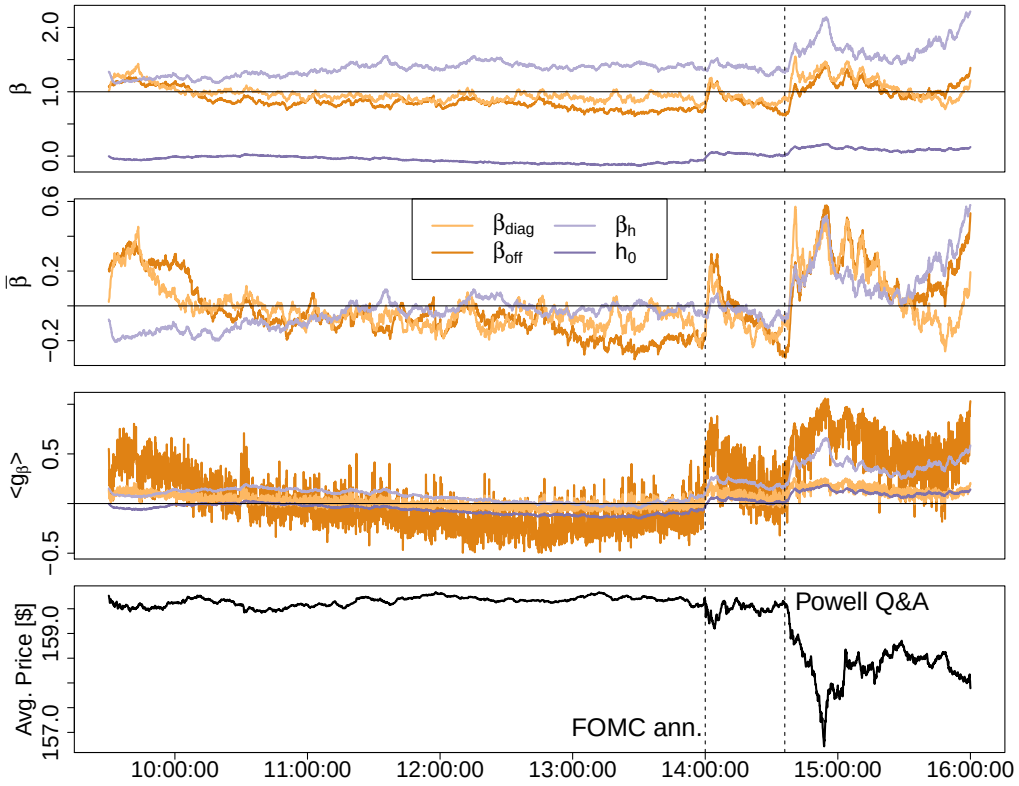

Figure 5.12: Values of $\boldsymbol{\beta}(t)$, $\bar{\boldsymbol{\beta}}(t)$, $\langle g_\beta \rangle(t)$ and $h_0(t)$ on July 31, 2019, along with the average midprice across the S&P100 stocks. We highlight the time of at which the announcement becomes public (14:00:00 EST) and the time at which the press Q&A with Chairman Powell begins (14:36:00 EST).

from the one in the Flash Crash example). The announcement went public at 14:00:00 EST and is followed by a press conference at 14:30:00 EST, with a Q&A starting at around 14:36:00 EST. Again we see that the usual intraday pattern shown in Figure 5.10 is interrupted by the news, which however, differently from the Flash Crash, is a scheduled announcement. This difference leads to the complete absence of any sort of "unusual" effect in the earlier hours of the day, as typically analysts provide forecasts regarding these announcements in the previous days and this information is already incorporated in the prices. What then happens is that, if the news

does not meet market expectations, a correction in prices will occur as soon as the information is made public, leading to higher market volatility in the minutes and hours following the announcement (Chuliá et al. (2010); Hautsch et al. (2011)). In this specific case, forecasts were mixed between a 25 and a 50 basis points interest rates cut scenario[3].

The published announcement at 14:00 EST mostly matched these forecasts, with the FOMC lowering the interest target rate by 25 basis points, and we indeed see that the price levels are not particularly affected by the news. However an increase in volatility, and in particular the endogenous components, can still be observed in the few minutes following the announcement, quickly returning to average levels though. What is actually interesting is to see the reaction to the press conference held 30 minutes after the release, and in particular to the answers the Chairman of the Fed Jerome H. Powell gives to journalists in the Q&A. We see in fact that as soon as the Q&A starts, around 14:36 EST, prices begin to plummet in response to the Chairman's answers, possibly reacting to the statement that this interest rates cut was only intended as a "midcycle adjustment to policy" rather than as the first of a series. Expectations of further rates cuts in the later months of the year could be a reason for this adjustment in the prices when these forecasts are not met, as usually lower interest rates push the stock prices up. We see however that this unexpected event causes a behavior in the time-varying parameters estimates much more similar to what we have seen in the Flash Crash, albeit the endogenous components are even more significant here.

Overall, these two examples show that our model captures different reactions to events in stock volatilities depending whether at least part

[3]This information can be found on any finance-focused media outlet such as finance.yahoo.com, bloomberg.com or zacks.com

of the new information is already incorporated in the price, as is the case for the FOMC decision release, or whether the event is unpredictable in nature and triggered by external causes, as in the Flash Crash or the press conference of July 31, 2019.

Identifying traders strategy changes around macroeconomic news

Another example application we propose is an extension to the study presented in Chapter 4, where we utilized the standard Kinetic Ising Model to infer a network of lead-lag relationships among traders and to estimate the opinions held by traders on the underlying asset price, in this case the spot exchange rate between Euro and US Dollar. Here the time series represent buy (+1) or sell (-1) trades performed by individual traders in the period May - September 2013, on a time scale of 5 minutes on the electronic Foreign Exchange platform of a major dealer in the market, which provided the data. Briefly summarizing the process to produce the time series, we start from the trading records, containing information about the time of trade with millisecond precision, anonymized identity of the trader, volume of EUR purchased in exchange for USD (negative in case the trade goes in the other direction) and price paid, and we aggregate the traded volume for each trader in 5 minute time windows, and take the sign of the total volume as the binary variable to feed the Kinetic Ising Model. In the model we also include the log-returns on the exchange rate as an external covariate, thus letting $x(t) = r(t)$ and introducing the covariate coupling parameters b_i. We also split the data monthly in order to account for the non-stationary sample of traders of the platforms, which enter and exit the dataset thus rendering the data too incomplete on longer time scales.

We then proceed to the estimation of the J and h parameters alongside the imputation of the unobserved trades as discussed in the previous chapters, and we add the score-driven dynamics to the model as in Eq. 5.9. After performing the usual LM test reported in the fourth column of Table 5.1, we infer the score-driven dynamics parameters. We then obtain a set of time series for $\boldsymbol{\beta}$, now including also a time-varying $\beta_b(t)$ parameter for the log-returns couplings, and h_0. As mentioned in Chapter 4, this model should be interpreted as a way to put in relation the strategic decisions made by traders, highlighting which market participants can carry information about short-term trends in demand and supply as well as identifying the relations between traders adopting different strategies. In this extended score-driven version, the time-varying parameters allow a more refined interpretation of the model results by making explicit when the considered traders are more "coupled" to others or to price variations in their strategic behavior.

We compare our results to a dataset of macroeconomic announcement times from the website *www.dailyfx.com*, which provides a calendar of scheduled announcements (e.g. interest rate decisions by central banks, quarterly unemployment rate reports, ...) with labels characterizing which currencies are mostly affected by the announcement and the level of importance (low, medium or high) of the news. We restrict our analysis to the news that are labeled as highly important and involving either EUR or USD, the pair traded by the traders in our dataset. We obtain a total of 474 non-overlapping announcement events, of which 283 are referred to the US Dollar and the remaining 201 to the Euro.

As we have different models for different months which we need to compare, we first need to standardize the time series of our time-varying pa-

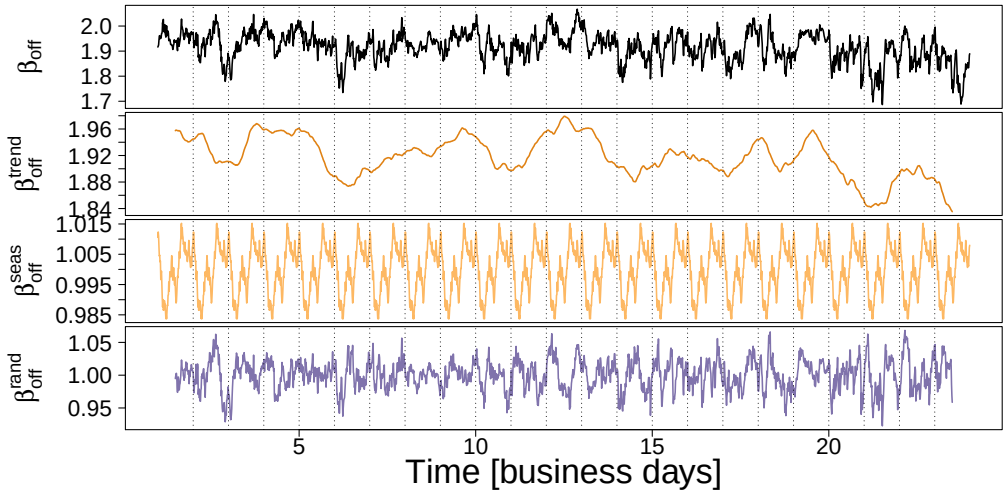

Figure 5.13: Multiplicative seasonal-trend decomposition of the inferred β_{off} series from a month (May 2013) of trader activity data in the Foreign Exchange market, with daily seasonality. Dotted lines mark the overnights.

rameters. To do so we proceed to operate a multiplicative trend-seasonal decomposition on the $\boldsymbol{\beta}$s and an additive decomposition on h_0. Then we have

$$\boldsymbol{\beta}_k(t) = \boldsymbol{\beta}_k^{seas}(t)\boldsymbol{\beta}_k^{trend}(t)\boldsymbol{\beta}_k^{rand}(t)$$
$$h_0(t) = h_0^{seas}(t) + h_0^{trend}(t) + h_0^{rand}(t)$$

where $\boldsymbol{\beta}(t) = (\beta_{diag}(t), \beta_{off}(t), \beta_h(t), \beta_b(t))$. The seasonal component is assumed to have daily periodicity, thus capturing any intraday pattern the parameters might show, while the trend component is the moving average of the parameter with a two-side square filter with bandwidth of one day, to match the seasonality. The remaining components $\boldsymbol{\beta}_k^{rand}$ and h_0^{rand} are what we are actually interested in, as they are the residual part of our parameters that is not explained by either the intraday pattern or the local average value. To check that we were not neglecting other possible choices,

we measure seasonality in the data by computing the Fourier transform of the time series to extract the principal spectral component (not shown here for the sake of space) and found it to be typically around the daily frequency: we decided to enforce daily seasonality in order to make the decomposition homogeneous across months. In Figure 5.13 we show one example of the trend-seasonal decomposition for β_{off} in the month of May 2013.

Having done this decomposition, we focus on the behavior of the residual parts $\boldsymbol{\beta}_k^{rand}$ and h_0^{rand} in the vicinity of news announcements. Defining the news timestamp t^*, we select the values of $\boldsymbol{\beta}_k^{rand}$ and h_0^{rand} in the interval $[t^* - 60m, t^* + 60m]$, that is one hour before and after the event. In order to be able to compare the residuals coming from different months, since we find that they are distributed similarly to a Gaussian by inspecting the quantile-quantile plots, we normalize them by subtracting the mean and dividing by the standard deviation, obtaining

$$\hat{\boldsymbol{\beta}}_k^{rand}(t) = \frac{\boldsymbol{\beta}_k^{rand}(t) - \mathbb{E}[\boldsymbol{\beta}_k^{rand}]}{\text{Stdev}[\boldsymbol{\beta}_k^{rand}]} \tag{5.13}$$

We then take, for each lag l in the time window around the events, the average value of the normalized $\hat{\boldsymbol{\beta}}_k^{rand}$ and $\hat{h}_0$ across all events, that is

$$\langle \hat{\boldsymbol{\beta}}_k^{rand}(l) \rangle = \frac{1}{N_e} \sum_{e=1}^{N_e} \hat{\boldsymbol{\beta}}_k^{rand}(t_e^* + l)$$

where N_e is the number of macroeconomic news events, $l \in [-60m, 60m]$ and t_e^* is the timestamp within which event e takes place, that is the announcement happens in the time window $(t_e^* - 5m, t_e^*]$ identified by the bin labeled 0 in the figures.

In Figure 5.14 we show these average values for all the different $\boldsymbol{\beta}$ components, along with 95% confidence intervals obtained considering the prob-

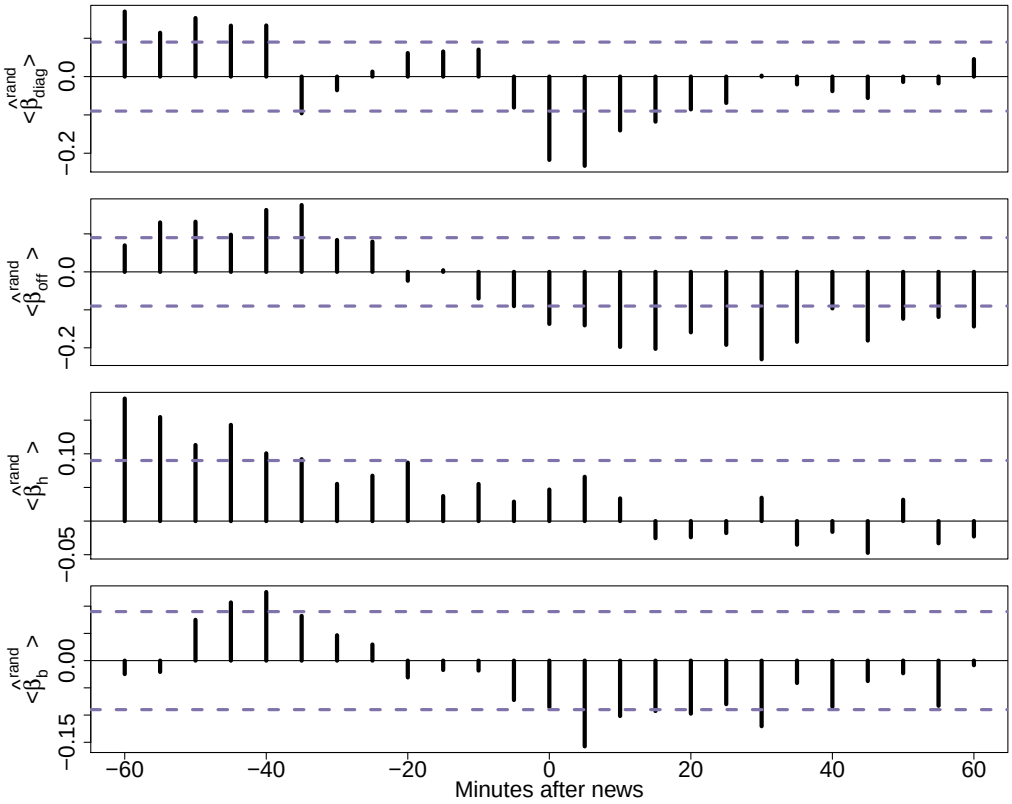

Figure 5.14: Patterns of the residual normalized components of $\boldsymbol{\beta}$ around macroeconomic news announcements, with 95% confidence bands.

ability that a sample of N_e Gaussian random variables sampled with zero population mean and unit variance has an empirical mean different from zero. We clearly see that there are significant patterns in the proximity of the news announcement, where both the β_{diag} and the β_{off} parameters show a reduction in the importance of both autocorrelation and cross-correlation effects in the trading behavior by traders, while the exogenous component β_h is mostly unchanged. The β_b parameter is also marginally smaller in the immediate vicinity of the announcement, possibly meaning that in that time frame the traders are less focused on following the price dynamics and more on reacting to the news.

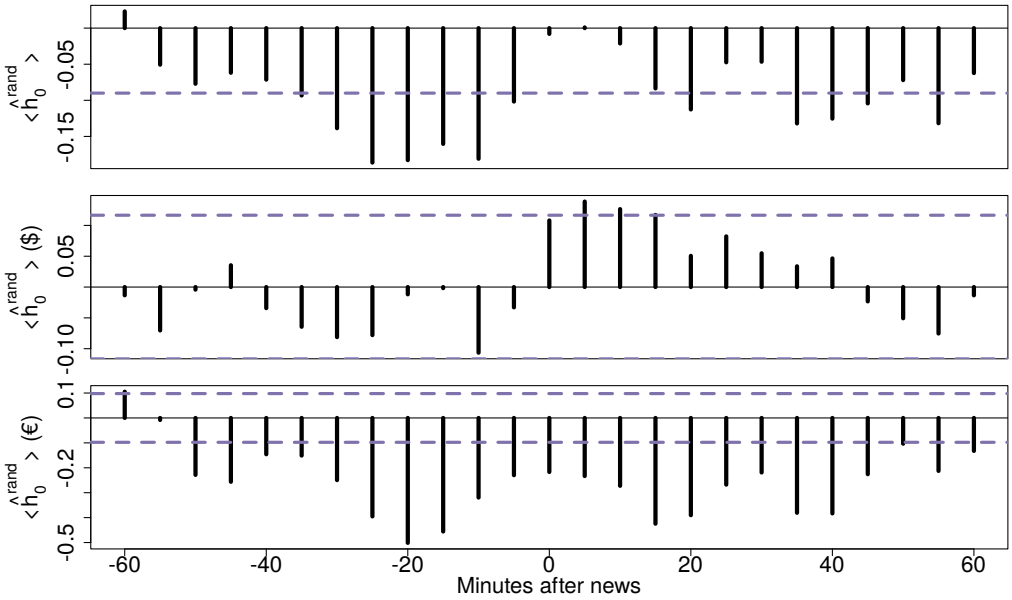

Figure 5.15: Patterns of the normalized residual component of h_0 around macroeconomic news. (top) all events; (mid) only USD-related events; (bottom) only EUR-related events.

As further evidence that this modelling approach captures meaningful effects, in Figure 5.15 we show the pattern of h_0 around the news events. Again, when $h_0 > 0$ it means that most of the trading activity of traders is directed towards purchasing EUR in exchange for USD, and viceversa when $h_0 < 0$. While when looking at the top panel of Figure 5.15 it might seem that around news traders tend to buy more USD, while it is actually more subtle than that. Indeed as we mentioned our dataset contains news affecting either USD or EUR, so we can condition our averaging procedure to this information.

In the middle and bottom panels of Figure 5.15 we show the average pattern of $\hat{h}_0^{rand}$ around USD-affecting news and EUR-affecting news, respectively. What we find is that traders are actually more likely to drop the affected currency from their inventory when they know a news is coming,

and that the effect is stronger for EUR than USD. We believe there are at least two plausible explanations for this: one is that the market we analyze is a European market in London opening times, thus the majority of the traders are likely to be European; another is that the year is 2013 and the Euro debt crisis is affecting EUR-related news, causing more risk-aversion in traders. While this effect is not particularly surprising and could be showed by simply looking at the net order flow, which is what h_0 is designed to capture, it proves that our modelling approach can be used to cleanly filter these effects when describing these systems.

We also argue that this particular behavior is consistent with the way we select traders: as we only consider in our data traders that are active in more than 30% of the timestamps, we are very likely excluding the traders that follow a news-trading strategy, as they are unlikely to trade that often outside of these time intervals. In this sense it is not surprising to see other traders be risk-averse with respect to the news, thus dropping the affected currency before the news comes to avoid adverse selection and higher volatility.

5.5 Conclusions

We have applied the score-driven methodology to extend the Kinetic Ising Model to a time-varying parameters formulation, introducing two new models for non-stationary time series, the Dynamical Noise Kinetic Ising Model (DyNoKIM) and the Dynamic Endogeneity Kinetic Ising Model (DyEKIM). We showed that the DyNoKIM, characterized by a time-varying noise level parameter $\beta(t)$, has a clear utility in forecasting applications, as the Area Under the ROC Curve can be showed to be a growing function of $\beta(t)$,

while the DyEKIM can be used to discriminate between endogenous and exogenous effects in the evolution of a time series. We then provided three example applications of the two models: in the first the DyNoKIM is successfully used to quantify the forecasting accuracy of stock activities in the US stock market; in the second we applied the DyEKIM to describe the high-frequency volatilities of US stocks in proximity of extreme events such as the Flash Crash of May 6, 2010 or around scheduled announcements as the FOMC report of July 31, 2019; in the last empirical application we built upon a previous work (Campajola et al. (2020)) on traders lead-lag networks to describe how trading strategies affect one another around macroeconomic announcements, showing that the DyEKIM effectively captures some interesting features of the data. Our empirical applications have been focused on financial systems, but we envision our approach can be useful also in other fields of application such as neuroscience and machine learning, where the static version of the Kinetic Ising Model has been in use for a long time.

Chapter 6

Conclusive remarks

In this thesis we have provided innovative contributions to the literature on the Kinetic Ising Model, as well as explored some of its potential financial applications.

We began by developing an inference algorithm for the estimation of the KIM parameters on datasets with missing observations, successfully adapting a known approximation to the presence of randomly distributed missing values in the sample.

The proposed methodology, based on an Expectation-Maximization-like strategy, is shown to be resilient to noise and relatively high fractions of missing data, and applicable to a wide range of problems thanks to the possibility to pair it with model selection techniques such as LASSO regularization or Decimation.

We then proceeded to apply our methodology in Chapter 4, studying a dataset of trading records in a financial market at high frequency. The data contains information about trading by clients of a major dealer in the foreign exchange market, specifically trading on the EUR/USD spot rate. For each trade we have information about the sign of the trade (whether the client

has bought or sold EUR for USD), the pseudonymous identity of the client and the exact time at which the trade took place. We discretized time to the 5 minutes time-scale and used traded volumes to determine the opinion the trader holds about the exchange rate in each time window, proceeding then to estimate the KIM parameters on the resulting dataset. Since traders are not constantly active in the market, the data has a significant fraction of missing values, which were taken into account and estimated thanks to the method developed in Chapter 3; this has been done assuming that even when a trader is inactive she still holds an opinion about the rate, which can be informative of the trading she or similar traders might be operating on other platforms.

The resulting models were mapped into networks of influence between traders, which we have studied through influencer detection techniques, showing there are some agents that are typically leading the order flow, and we also showed that the imputation of missing values provides a clearer picture about the state of supply and demand in the market as a whole, as justified by our Granger Causality analysis with the state of liquidity on another market venue, the centralized interdealer EBS exchange.

We found that the lead-lag networks are persistent in time, by fitting the model monthly over two years of data and measuring neighbourhood similarity, however the methodology is not well-suited for forecasting purposes due to the necessity, during the imputation of missing values, of including information from the future.

As a general remark, we interpret these lead-lag networks as the effect of traders following similar strategies with different reaction times, despite our choice of terminology referring to opinion spreading which we intended to provide intuition about the system without diving too deep in the financial

lexicon, which might be obscure to a broader audience.

The final contribution of this thesis is an extension of the KIM to a time-varying parameters formulation using the score-driven methodology. We proposed two such extentions, the Dynamical Noise KIM (DyNoKIM) and the Dynamic Endogeneity KIM (DyEKIM), each suited for a specific application. We provided evidence that the DyNoKIM has clear utility in forecasting applications, as we have shown that the Area Under the ROC Curve is an increasing function of the inverse noise parameter $\beta(t)$, an effect we measure empirically on a dataset of midprice variations of US stocks.

On the other hand the DyEKIM can be used to study the level of endogeneity of the dynamics observed in data, which we describe in two empirical applications, one to the high-frequency volatility of US stocks in proximity of extreme events or news and the other to the same traders dataset of Chapter 4, this time focusing on the time frames surrounding macroeconomic news announcements. We find that in both cases the DyEKIM captures interesting features in the data, providing a useful tool for the analysis of financial time series and not only, as the formulation of the score-driven KIM is general enough that opportune adaptations can be devised for all sorts of application fields, such as computational neuroscience and the theory of machine learning.

In conclusion, this thesis has provided a new bridge between several strands of literature, mainly the ones on statistical mechanics and financial econometrics, but also taking inspiration from machine learning, neuroscience and social science. We hope that our work can be beneficial to these communities, which have already done much by growing our understanding of the world around us and the world we shape ourselves to live in.

Bibliography

Aït-Sahalia, Y., Fan, J., and Xiu, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Al-Osh, M. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (inar (1)) process. *Journal of Time Series Analysis*, 8(3):261–275.

Alfarano, S., Lux, T., and Wagner, F. (2005). Estimation of agent-based models: the case of an asymmetric herding model. *Computational Economics*, 26(1):19–49.

Alfi, V., Cristelli, M., Pietronero, L., and Zaccaria, A. (2009a). Minimal agent based model for financial markets i. *The European Physical Journal B*, 67(3):385–397.

Alfi, V., Cristelli, M., Pietronero, L., and Zaccaria, A. (2009b). Minimal agent based model for financial markets ii. *The European Physical Journal B*, 67(3):399–417.

Aquilina, M., Budish, E., and O'Neill, P. (2020). Quantifying the high-frequency trading "arms race": A simple new methodology and estimates. *United Kingdom Financial Conduct Authority Occasional Paper*.

Azen, R. and Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological methods*, 8(2):129.

Banerjee, A. V. (1992). A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817.

Barde, S. (2016). Direct comparison of agent-based models of herding in financial markets. *Journal of Economic Dynamics and Control*, 73:329–353.

Barigozzi, M. and Brownlees, C. (2019). Nets: Network estimation for time series. *Journal of Applied Econometrics*, 34(3):347–364.

Bauwens, L. and Veredas, D. (2004). The stochastic conditional duration model: a latent variable model for the analysis of financial durations. *Journal of econometrics*, 119(2):381–412.

Beck, C. and Cohen, E. G. (2003). Superstatistics. *Physica A: Statistical mechanics and its applications*, 322:267–275.

Beck, C., Cohen, E. G., and Swinney, H. L. (2005). From time series to superstatistics. *Physical Review E*, 72(5):056133.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Biely, C., Hanel, R., and Thurner, S. (2009). Socio-economical dynamics as a solvable spin system on co-evolving networks. *The European Physical Journal B*, 67(3):285–289.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026.

Bikhchandani, S. and Sharma, S. (2000). Herd behavior in financial markets. *IMF Staff papers*, 47(3):279–310.

Blasques, F., Koopman, S. J., and Lucas, A. (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika*, 102(2):325–343.

Blasques, F., Lucas, A., and van Vlodrop, A. (2017). Finite sample optimality of score-driven volatility models. *Tinbergen Institute Discussion Paper*, 17-111/III.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327.

Bornholdt, S. (2001). Expectation bubbles in a spin model of markets: Intermittency from frustration across scales. *International Journal of Modern Physics C*, 12(05):667–674.

Bossomaier, T., Barnett, L., Harré, M., and Lizier, J. T. (2016). An introduction to transfer entropy. *Cham: Springer International Publishing*, pages 65–95.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

Bouchaud, J.-P. (2013). Crises and collective socio-economic phenomena: Simple models and challenges. *Journal of Statistical Physics*, 151(3):567–606.

Bouchaud, J.-P., Farmer, J. D., and Lillo, F. (2009). How markets slowly digest changes in supply and demand. In *Handbook of financial markets: dynamics and evolution*, pages 57–160. Elsevier.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Brock, W. A. and Durlauf, S. N. (1999). A formal model of theory choice in science. *Economic theory*, 14(1):113–130.

Buccheri, G., Bormetti, G., Corsi, F., and Lillo, F. (2020). A score-driven conditional correlation model for noisy and asynchronous data: An application to high-frequency covariance dynamics. *Journal of Business & Economic Statistics*, pages 1–17.

Budescu, D. V. (1993). Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, 114(3):542.

Buhi, E. R., Goodson, P., and Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American journal of health behavior*, 32(1):83–92.

Bury, T. (2013). Market structure explained by pairwise interactions. *Physica A: Statistical Mechanics and its Applications*, 392(6):1375–1385.

Cai, F., Han, S., Li, D., and Li, Y. (2019). Institutional herding and its price impact: Evidence from the corporate bond market. *Journal of Financial economics*, 131(1):139–167.

Calvori, F., Creal, D., Koopman, S. J., and Lucas, A. (2017). Testing for parameter instability across different modeling frameworks. *Journal of Financial Econometrics*, 15(2):223–246.

Campajola, C., Lillo, F., and Tantari, D. (2019). Inference of the kinetic ising model with heterogeneous missing data. *Physical Review E*, 99(6):062138.

Campajola, C., Lillo, F., and Tantari, D. (2020). Unveiling the relation between herding and liquidity with trader lead-lag networks. *Quantitative Finance*, page in press.

Capone, C., Filosa, C., Gigante, G., Ricci-Tersenghi, F., and Del Giudice, P. (2015). Inferring synaptic structure in presence of neural interaction time scales. *PloS one*, 10(3):e0118412.

Castellano, C., Muñoz, M. A., and Pastor-Satorras, R. (2009). Nonlinear q-voter model. *Physical Review E*, 80(4):041129.

Cenesizoglu, T. and Grass, G. (2018). Bid-and ask-side liquidity in the nyse limit order book. *Journal of Financial Markets*, 38:14–38.

Challet, D., Chicheportiche, R., Lallouache, M., and Kassibrakis, S. (2016). Trader lead-lag networks and order flow prediction. *Available at SSRN 2839312*.

Challet, D., Chicheportiche, R., Lallouache, M., and Kassibrakis, S. (2018). Statistically validated lead-lag networks and inventory prediction in the foreign exchange market. *Advances in Complex Systems*, 21(08):1850019.

Challet, D., Marsili, M., Zhang, Y.-C., et al. (2013). Minority games: interacting agents in financial markets. *OUP Catalogue*.

Chuliá, H., Martens, M., and van Dijk, D. (2010). Asymmetric effects of federal funds target rate changes on s&p100 stock returns, volatilities and correlations. *Journal of Banking & Finance*, 34(4):834–839.

Cocco, S., Monasson, R., Posani, L., and Tavoni, G. (2017). Functional networks from inverse modeling of neural population activity. *Current Opinion in Systems Biology*, 3:103–110.

Cont, R. (2007). Volatility clustering in financial markets: empirical facts and agent-based models. In *Long memory in economics*, pages 289–309. Springer.

Coolen, A. (2001a). Statistical mechanics of recurrent neural networks i: Statics. *Handbook of biological physics*, 4:531–596.

Coolen, A. (2001b). Statistical mechanics of recurrent neural networks iidynamics. *Handbook of biological physics*, 4:619–684.

Cordi, M., Challet, D., and Kassibrakis, S. (2019). The market nanostructure origin of asset price time reversal asymmetry. *arXiv preprint arXiv:1901.00834*.

Corsi, F., Peluso, S., and Audrino, F. (2012). Missing in Asynchronicity: A Kalman-EM Approach for Multivariate Realized Covariance Estimation.

Economics Working Paper Series 1202, University of St. Gallen, School of Economics and Political Science.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., Juselius, K., and Lauritzen, S. L. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 93–115.

Cox, J. C. (1996). The constant elasticity of variance option pricing model. *Journal of Portfolio Management*, page 15.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Crisanti, A. and Sompolinsky, H. (1988). Dynamics of spin systems with randomly asymmetric bonds: Ising spins and glauber dynamics. *Physical Review A*, 37(12):4865.

Curme, C., Tumminello, M., Mantegna, R. N., Stanley, H. E., and Kenett, D. Y. (2015). Emergence of statistically validated financial intraday lead-lag relationships. *Quantitative Finance*, 15(8):1375–1386.

Davis, M. H. (2016). Verification of internal risk measure estimates. *Statistics & Risk Modeling*, 33(3-4):67–93.

De Dominicis, C. (1978). Dynamics as a substitute for replicas in systems with quenched random impurities. *Physical Review B*, 18(9):4913.

De Vincenzo, I., Giannoccaro, I., Carbone, G., and Grigolini, P. (2017). Criticality triggers the emergence of collective intelligence in groups. *Physical Review E*, 96(2):022309.

Decelle, A., Ricci-Tersenghi, F., and Zhang, P. (2016). Data quality for the inverse lsing problem. *Journal of Physics A: Mathematical and Theoretical*, 49(38):384001.

Decelle, A. and Zhang, P. (2015). Inference of the sparse kinetic ising model using the decimation method. *Physical Review E*, 91(5):052136.

Dehnert, M., Helm, W., and Hütt, M.-T. (2003). A discrete autoregressive process as a model for short-range correlations in dna sequences. *Physica A: Statistical Mechanics and its Applications*, 327(3-4):535–553.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Derrida, B., Gardner, E., and Zippelius, A. (1987). An exactly solvable asymmetric neural network model. *EPL (Europhysics Letters)*, 4(2):167.

Dunn, B. and Roudi, Y. (2013). Learning and inference in a nonequilibrium ising model with hidden nodes. *Physical Review E*, 87(2):022127.

Easley, D., De Prado, M. M. L., and O'Hara, M. (2011). The microstructure of the "flash crash": flow toxicity, liquidity crashes, and the probability of informed trading. *The Journal of Portfolio Management*, 37(2):118–128.

Edwards, S. F. and Anderson, P. W. (1975). Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965.

Fagiolo, G., Giachini, D., and Roventini, A. (2019a). Innovation, finance, and economic growth: an agent-based approach. *Journal of Economic Interaction and Coordination*, pages 1–34.

Fagiolo, G., Guerini, M., Lamperti, F., Moneta, A., and Roventini, A. (2019b). Validation of agent-based models in economics and finance. In *Computer Simulation Validation*, pages 763–787. Springer.

Farmer, J. D. (2002). Market force, ecology and evolution. *Industrial and Corporate Change*, 11(5):895–953.

Farmer, J. D. and Foley, D. (2009). The economy needs agent-based modelling. *Nature*, 460(7256):685–686.

Ferrari, U., Deny, S., Chalk, M., Tkačik, G., Marre, O., and Mora, T. (2018). Separating intrinsic interactions from extrinsic correlations in a network of sensory neurons. *Physical Review E*, 98(4):042410.

Ferrenberg, A. M. and Landau, D. (1991). Critical behavior of the three-dimensional ising model: A high-resolution monte carlo study. *Physical Review B*, 44(10):5081.

Filimonov, V. and Sornette, D. (2012). Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108.

Filimonov, V. and Sornette, D. (2015). Apparent criticality and calibration issues in the hawkes self-excited point process model: application to high-frequency financial data. *Quantitative Finance*, 15(8):1293–1314.

Fisher, M. E. (1964). Magnetism in one-dimensional systems—the heisenberg model for infinite spin. *American Journal of Physics*, 32(5):343–346.

Galariotis, E. C., Krokida, S.-I., and Spyrou, S. I. (2016). Bond market investor herding: Evidence from the european financial crisis. *International Review of Financial Analysis*, 48:367–375.

Gallotti, R. and Barthelemy, M. (2015). The multilayer temporal network of public transport in great britain. *Scientific data*, 2(1):1–8.

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. In *Hidden Markov models: applications in computer vision*, pages 9–41. World Scientific.

Glosten, L. R. and Milgrom, P. R. (1985). Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14:71–100.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Grinblatt, M. and Keloharju, M. (2000). The investment behavior and performance of various investor types: a study of finland's unique data set. *Journal of Financial Economics*, 55:43–67.

Grinblatt, M., Titman, S., and Wermers, R. (1995). Momentum investment strategies, porfolio performance, and herding: A study of mutual fund behavior. *American Economic Review*, 85:1088–1105.

Guéant, O. (2016). *The Financial Mathematics of Market Liquidity: From optimal execution to market making*, volume 33. CRC Press.

Gutiérrez-Roig, M., Borge-Holthoefer, J., Arenas, A., and Perelló, J. (2019). Mapping individual behavior in financial markets: synchronization and anticipation. *EPJ Data Science*, 8(1):10.

Hafner, C. M. and Manner, H. (2012). Dynamic stochastic copula models: Estimation, inference and applications. *Journal of Applied Econometrics*, 27(2):269–295.

Hamilton, W. D. (1971). Geometry for the selfish herd. *Journal of theoretical Biology*, 31(2):295–311.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Hardiman, S. J., Bercot, N., and Bouchaud, J.-P. (2013). Critical reflexivity in financial markets: a hawkes process analysis. *The European Physical Journal B*, 86(10):442.

Hardiman, S. J. and Bouchaud, J.-P. (2014). Branching-ratio approximation for the self-exciting hawkes process. *Physical Review E*, 90(6):062807.

Harras, G., Tessone, C. J., and Sornette, D. (2012). Noise-induced volatility of collective dynamics. *Physical Review E*, 85(1):011150.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. Econometric Society Monographs. Cambridge University Press.

Hautsch, N., Hess, D., and Veredas, D. (2011). The impact of macroeconomic news on quote adjustments, noise, and informational volatility. *Journal of Banking & Finance*, 35(10):2733–2746.

Hertz, J. A., Roudi, Y., Thorning, A., Tyrcha, J., Aurell, E., and Zeng, H.-L. (2010). Inferring network connectivity using kinetic ising models. *BMC neuroscience*, 11(1):P51.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies*, 6(2):327–343.

Hinton, G. E. (2012). A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer.

Ho, T. and Stoll, H. R. (1980). On dealer markets under competition. *The Journal of Finance*, 35(2):259–267.

Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics reports*, 519(3):97–125.

Hong, Y., Liu, Y., and Wang, S. (2009). Granger causality in risk and detection of extreme risk spillover between financial markets. *Journal of Econometrics*, 150(2):271–287.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

Ibuki, T., Higano, S., Suzuki, S., Inoue, J.-i., and Chakraborti, A. (2013). Statistical inference of co-movements of stocks during a financial crisis. In *Journal of Physics: Conference Series*, volume 473, page 012008. IOP Publishing.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258.

Jacobs, P. A. and Lewis, P. A. (1978). Discrete time series generated by mixtures. iii. autoregressive processes (dar (p)). Technical report, NAVAL POSTGRADUATE SCHOOL MONTEREY CALIF.

Jacobs, P. A. and Lewis, P. A. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis*, 4(1):19–36.

Janssen, H.-K. (1976). On a lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties. *Zeitschrift für Physik B Condensed Matter*, 23(4):377–380.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Jegadeesh, N. and Titman, S. (1995). Overreaction, delayed reaction, and contrarian profits. *The Review of Financial Studies*, 8(4):973–993.

Kadirvelu, B., Hayashi, Y., and Nasuto, S. J. (2017). Inferring structural connectivity using ising couplings in models of neuronal networks. *Scientific reports*, 7(1):8156.

Kaizoji, T. (2000). Speculative bubbles and crashes in stock markets: an interacting-agent model of speculative activity. *Physica A: Statistical Mechanics and its Applications*, 287(3-4):493–506.

Kaizoji, T., Bornholdt, S., and Fujiwara, Y. (2002). Dynamics of price and trading volume in a spin model of stock markets with heterogeneous agents. *Physica A: Statistical Mechanics and its Applications*, 316(1-4):441–452.

Kim, J., Kim, B., and Sohraby, K. (2008). Mean queue size in a queue with discrete autoregressive arrivals of order p. *Annals of Operations Research*, 162(1):69–83.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998.

Kirkpatrick, S. and Sherrington, D. (1978). Infinite-ranged models of spin-glasses. *Physical Review B*, 17(11):4384.

Kiss, C. and Bichler, M. (2008). Identification of influencers—measuring influence in customer networks. *Decision Support Systems*, 46(1):233–253.

Koopman, S. J., Lucas, A., and Scharth, M. (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *Review of Economics and Statistics*, 98(1):97–110.

Kosterlitz, J. (1974). The critical properties of the two-dimensional xy model. *Journal of Physics C: Solid State Physics*, 7(6):1046.

Kristoufek, L. and Vosvrda, M. (2018). Herding, minority game, market clearing and efficient markets in a simple spin model framework. *Communications in Nonlinear Science and Numerical Simulation*, 54:148–155.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53:1315–1335.

Lakonishok, J., Shleifer, A., and Vishny, R. W. (1992). The impact of institutional trading on stock prices. *Journal of Financial Economics*, 32:23–43.

Leal, S. J., Napoletano, M., Roventini, A., and Fagiolo, G. (2016). Rock around the clock: An agent-based model of low-and high-frequency trading. *Journal of Evolutionary Economics*, 26(1):49–76.

Lillo, F., Miccichè, S., Tumminello, M., Piilo, J., and Mantegna, R. N. (2015). How news affects the trading behaviour of different categories of investors in a financial market. *Quantitative Finance*, 15(2):213–229.

Lillo, F., Moro, E., Vaglica, G., and Mantegna, R. N. (2008). Specialization and herding behavior of trading firms in a financial market. *New Journal of Physics*, 10:043019.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. Wiley.

Lux, T. et al. (2001). Turbulence in financial markets: the surprising explanatory power of simple cascade models. *Quantitative finance*, 1(6):632–640.

Lux, T. and Sornette, D. (2002). On rational bubbles and fat tails. *Journal of Money, Credit and Banking*, pages 589–610.

Madhavan, A. (2012). Exchange-traded funds, market structure, and the flash crash. *Financial Analysts Journal*, 68(4):20–35.

Marlin, B. M. and Zemel, R. S. (2009). Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*, pages 5–12. ACM.

Martin, P. C., Siggia, E., and Rose, H. (1973). Statistical dynamics of classical systems. *Physical Review A*, 8(1):423.

Mazzarisi, P., Barucca, P., Lillo, F., and Tantari, D. (2020a). A dynamic network model with persistent links and node-specific latent variables, with an application to the interbank market. *European Journal of Operational Research*, 281(1):50–65.

Mazzarisi, P., Zaoli, S., Campajola, C., and Lillo, F. (2020b). Tail granger causalities and where to find them: extreme risk spillovers vs. spurious linkages. *arXiv preprint arXiv:2005.01160*.

Menkveld, A. J. and Yueshen, B. Z. (2019). The flash crash: A cautionary tale about highly fragmented markets. *Management Science*, 65(10):4470–4488.

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. In *Advances in neural information processing systems*, pages 1277–1285.

Montes, L. (2003). Smith and newton: some methodological issues concerning general economic equilibrium theory. *Cambridge Journal of Economics*, 27(5):723–747.

Mooneyham, B. W., Mrazek, M. D., Mrazek, A. J., Mrazek, K. L., Phillips, D. T., and Schooler, J. W. (2017). States of mind: characterizing the neural bases of focus and mind-wandering through dynamic functional connectivity. *Journal of cognitive neuroscience*, 29(3):495–506.

Musciotto, F., Marotta, L., Piilo, J., and Mantegna, R. N. (2018). Long-term ecology of investors in a financial market. *Palgrave Communications*, 4(1):92.

Nesterov, Y. (2008). Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181.

Nghiem, T.-A., Marre, O., Destexhe, A., and Ferrari, U. (2017). Pairwise ising model analysis of human cortical neuron recordings. In *International Conference on Geometric Science of Information*, pages 257–264. Springer.

Nghiem, T.-A. E., Tort-Colet, N., Górski, T., Ferrari, U., Moghimyfiroozabad, S., Goldman, J. S., Teleńczuk, B., Capone, C., Bal, T., Di Volo, M., et al. (2020). Cholinergic switch between two types of slow waves in cerebral cortex. *Cerebral Cortex*, 30(6):3451–3466.

Nicosia, V., Tang, J., Mascolo, C., Musolesi, M., Russo, G., and Latora, V. (2013). Graph metrics for temporal networks. In *Temporal networks*, pages 15–40. Springer.

Nietzsche, F. (1997). *Nietzsche: untimely meditations*. Cambridge University Press.

Nietzsche, F. W. (1974). *The gay science: With a prelude in German rhymes and an appendix of songs*, volume 985. Vintage.

Nietzsche, F. W. and Common, T. (1950). *Thus Spake Zarathustra*. Modern Library New York.

Nofsinger, J. R. and Sias, R. W. (1999). Herding and feedback trading by institutional and individual investors. *The Journal of finance*, 54(6):2263–2295.

Novelli, L., Wollstadt, P., Mediano, P., Wibral, M., and Lizier, J. T. (2019). Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Network Neuroscience*, 3(3):827–847.

Nystrup, P., Madsen, H., and Lindström, E. (2017). Long memory of financial time series and hidden markov models with time-varying parameters. *Journal of Forecasting*, 36(8):989–1002.

Onsager, L. (1944). Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Physical Review*, 65(3-4):117.

Opper, M. and Saad, D. (2001). *Advanced mean field methods: Theory and practice*. MIT press.

Palmer, M., Eickhoff, M., and Muntermann, J. (2018). Detecting herding behavior using topic mining: The case of financial analysts. *Research Papers*, 97.

Peierls, R. (1936). On ising's model of ferromagnetism. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(3):477–481.

Penney, R., Coolen, A., and Sherrington, D. (1993). Coupled dynamics of fast spins and slow interactions in neural networks and spin systems. *Journal of Physics A: Mathematical and General*, 26(15):3681.

Pham-Gia, T. and Turkkan, N. (1992). Determination of the beta distribution from its lorenz curve. *Mathematical and computer modelling*, 16(2):73–84.

Powell, J. (2019). Transcript of chair powell's press conference. *Federal Open Market Committee, July*, 31.

Rambaldi, M., Filimonov, V., and Lillo, F. (2018). Detection of intensity bursts using hawkes processes: An application to high-frequency financial data. *Physical Review E*, 97(3):032318.

Rambaldi, M., Pennesi, P., and Lillo, F. (2015). Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach. *Physical Review E*, 91(1):012819.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Roehner, B. and Sornette, D. (2000). "thermometers" of speculative frenzy. *The European Physical Journal B-Condensed Matter and Complex Systems*, 16(4):729–739.

Roudi, Y. and Hertz, J. (2011a). Dynamical tap equations for non-equilibrium ising spin glasses. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(03):P03031.

Roudi, Y. and Hertz, J. (2011b). Mean field theory for nonequilibrium network reconstruction. *Physical review letters*, 106(4):048702.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Sakellariou, J. (2013). *Inverse inference in the asymmetric ising model*. PhD thesis, Université Paris Sud-Paris XI.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186.

Schmitz, O. (2017). Predator and prey functional traits: understanding the adaptive machinery driving predator–prey interactions. *F1000Research*, 6.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Securities, U., Commission, E., Commission, C. F. T., et al. (2010). Findings regarding the market events of may 6, 2010. *Washington DC*.

Sewell, D. K. and Chen, Y. (2016). Latent space models for dynamic networks with weighted edges. *Social Networks*, 44:105–116.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shephard, N. (2005). *Stochastic volatility: selected readings*. Oxford University Press on Demand.

Shiller, R. J. (2015). *Irrational exuberance: Revised and expanded third edition*. Princeton university press.

Sood, V. and Redner, S. (2005). Voter model on heterogeneous graphs. *Physical review letters*, 94(17):178701.

Sornette, D. (2014). Physics and financial economics (1776–2014): puzzles, ising and agent-based models. *Reports on progress in physics*, 77(6):062001.

Squartini, T., Picciolo, F., Ruzzenenti, F., and Garlaschelli, D. (2013). Reciprocity of weighted networks. *Scientific reports*, 3:2729.

Stanley, H. and Kaplan, T. (1966). Possibility of a phase transition for the two-dimensional heisenberg model. *Physical Review Letters*, 17(17):913.

Tanaka, S. and Scheraga, H. A. (1977). Model of protein folding: incorporation of a one-dimensional short-range (ising) model into a three-dimensional model. *Proceedings of the National Academy of Sciences*, 74(4):1320–1323.

Taranto, D. E., Bormetti, G., and Lillo, F. (2014). The adaptive nature of liquidity taking in limit order books. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(6):P06002.

Tauchen, G. E. and Pitts, M. (1983). The price variability-volume relationship on speculative markets. *Econometrica: Journal of the Econometric Society*, pages 485–505.

Tavoni, G., Ferrari, U., Battaglia, F. P., Cocco, S., and Monasson, R. (2017). Functional coupling networks inferred from prefrontal cortex activity show experience-related effective plasticity. *Network Neuroscience*, 1(3):275–301.

Thouless, D. J., Anderson, P. W., and Palmer, R. G. (1977). Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Toth, B., Palit, I., Lillo, F., and Farmer, J. D. (2015). Why is equity order flow so persistent? *Journal of Economic Dynamics and Control*, 51:218–239.

Tucci, M. P. (1995). Time-varying parameters: a critical introduction. *Structural Change and Economic Dynamics*, 6(2):237–260.

Tumminello, M., Lillo, F., Piilo, J., and Mantegna, R. N. (2012). Identification of clusters of investors from their real trading activity in a financial market. *New Journal of Physics*, 14(1):013041.

Tumminello, M., Micciche, S., Lillo, F., Piilo, J., and Mantegna, R. N. (2011). Statistically validated networks in bipartite complex systems. *PloS one*, 6(3):e17994.

Wheatley, S., Wehrli, A., and Sornette, D. (2019). The endo–exo problem in high frequency financial price fluctuations and rejecting criticality. *Quantitative Finance*, 19(7):1165–1178.

White, H. (1987). Specification testing in dynamic models. In *Advances in econometrics-Fifth world congress*, volume 1, pages 1–58.

Williams, O. E., Lillo, F., and Latora, V. (2019). Effects of memory on spreading processes in non-markovian temporal networks. *New Journal of Physics*, 21(4):043028.

Zhang, P. (2012). Inference of kinetic ising model on sparse graphs. *Journal of Statistical Physics*, 148(3):502–512.

Appendix A

Equivalence between KIM and V-DAR(1) models

The contents of this appendix are the result of a joint work with Dr. Piero Mazzarisi, Prof. Fabrizio Lillo and Prof. Daniele Tantari, to appear soon in an online pre-print and submitted for publication.

In this appendix we show the equivalence, under appropriate reparametrization, between the Kinetic Ising Model and the Vector Discrete AutoRegressive model of order 1, the V-DAR(1).

The V-DAR(1) model $\{\{\boldsymbol{X}_t\}, p_{VDAR}, \boldsymbol{\pi}\}$ is defined for a set of T observations, each constituted by a vector of binary random variables $\boldsymbol{X}_t \in \{0,1\}^N$ which are independent conditionally on past observations. This model has been proposed originally in its univariate version in Jacobs and Lewis (1978) and followed by several extensions such as the Discrete AutoRegressive Moving Average (DARMA) model of Jacobs and Lewis (1983), the INteger valued AutoRegressive (INAR) model of Al-Osh and Alzaid (1987) and recently proposed in its multivariate formulation by Mazzarisi et al. (2020b),

the V-DAR model indeed. Models from this family have seen applications in genetics (Dehnert et al. (2003)), queueing theory (Kim et al. (2008)), temporal networks (Williams et al. (2019)) and recently in financial systems, as methods to forecast order flows (Taranto et al. (2014)) or to identify preferential lending between banks (Mazzarisi et al. (2020a)).

It is, like the KIM, a Markovian model of order 1, with transition probability

$$p_{VDAR}(\boldsymbol{X}_t|\boldsymbol{X}_{t-1};\boldsymbol{\pi}) = \prod_{i=1}^{N}\left[\nu_i\left(\sum_{j=1}^{N}\lambda_{ij}\delta_{X_t^i,X_{t-1}^j}\right) + (1-\nu_i)(\chi_i)^{X_t^i}(1-\chi_i)^{1-X_t^i}\right] \tag{A.1}$$

where $\delta_{X_t^i,X_{t-1}^j}$ is the Kronecker delta symbol and $\boldsymbol{\pi} = \{(\nu_i, \{\lambda_{ij}\}_j, \chi_i)\}$. The parameters in $\boldsymbol{\pi}$ are to be intended as probabilities of Bernoulli random variables: ν_i reflects the probability that the value X_t^i is copied from the past; if X_t^i is copied, λ_{ij} is the probability that it takes the value of X_{t-1}^j; otherwise, if X_t^i is not copied, X_t^i is sampled as a Bernoulli random variable with parameter χ_i.

It then follows that $\boldsymbol{\pi}$ is defined in a space $\Pi = ([0,1] \times \mathcal{S}_N \times [0,1])^N$ where $\mathcal{S}_N = \{\lambda_i \in [0,1]^N : \sum_j^N \lambda_{ij} = 1\}$. The space Π has dimension $N(N+1)$, exactly as the space of parameters of the Kinetic Ising Model, and the VDAR(1) also maps the evolution of binary variables with lagged dependency of order 1. It is thus immediate to ask the question whether a mapping between the two exists, as well as finding under which conditions the two models can be considered equivalent.

The transition probability of the Kinetic Ising Model given by Eq. 2.3 can be stated in terms of the same binary variables $\boldsymbol{X}_t \in \{0,1\}^N \ \forall i,t$, through the relation $X_t^i = \frac{1+s_i(t)}{2}$, reading

$$p_{KIM}(\boldsymbol{X}_t|\boldsymbol{X}_{t-1};\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{\exp\left[2X_t^i\left(h_i + \sum_{j=1}^{N} J_{ij}(2X_{t-1}^j - 1)\right)\right]}{1 + \exp\left[2\left(h_i + \sum_{j=1}^{N} J_{ij}(2X_{t-1}^j - 1)\right)\right]} \quad \text{(A.2)}$$

where $\boldsymbol{\theta} = (J, h)$, thus the KIM can be summarized as model $\{\{\boldsymbol{X}_t\}, p_{KIM}, \boldsymbol{\theta}\}$. Calling $\Theta = \mathbb{R}^{N\times N} \times \mathbb{R}^N$ the space of all possible KIM parameters $\boldsymbol{\theta}$, if one is able to show that there exists a unique and injective map $f : \Pi \to \Theta$ such that

$$p_{KIM}(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}; f(\boldsymbol{\pi})) = p_{VDAR}(\boldsymbol{X}_t|\boldsymbol{X}_{t-1};\boldsymbol{\pi}) \quad \text{(A.3)}$$

for any $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t-1}$ then the two models are equivalent in the range of f, which does not necessarily coincide with the whole codomain Θ.

Before stating the theorem, let us show that this map exists in the trivial cases of $N = 1$ and $N = 2$. In the case $N = 1$, where both J and h are scalars and there is only one $\lambda = 1$ by design, this mapping is easily found to be

$$h = \frac{1}{4}\log\left(\frac{\frac{\chi}{1-\chi} + \nu}{\frac{1}{\chi} - (1-\nu)}\right) \quad \text{(A.4)}$$

$$J = \frac{1}{4}\log\left(1 + \frac{\nu}{(1-\nu)^2\chi(1-\chi)}\right) \quad \text{(A.5)}$$

One can notice that here J is strictly positive as long as $\nu, \chi > 0$ and is $J = 0$ if and only if $\nu = 0$: this points to the idea that the V-DAR(1) model is indeed a restricted version of the KIM, with the elements of the coupling matrix restricted to positive values. Intuitively this is due to the fact that, while $J_{ij} < 0$ implies that spin i tends to take the opposite value of j, there is not a probability of "negated copying" in the V-DAR model.

In the case $N = 2$ there are two λ_i parameters. By considering three independent configurations of $\boldsymbol{X_{t-1}}$ and one possible realization of $\boldsymbol{X_t}$, one can extract from Eq. A.3 the system

$$\begin{pmatrix} 1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} J_{i1} \\ J_{i2} \\ h_i \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \log\left(\frac{1}{(1-\nu_i)\chi_i} - 1\right) \\ \log\left(\frac{1}{\nu_i(1-\lambda_i)+(1-\nu_i)\chi_i} - 1\right) \\ \log\left(\frac{1}{\nu_i+(1-\nu_i)\chi_i} - 1\right) \end{pmatrix}, \quad \forall i = 1, 2 \tag{A.6}$$

What we learn from this case is that the mapping is in fact the solution to a linear system of equations involving the inverse logistic functions on the right hand side and the vector of parameters of the KIM on the left hand side. Again it is easy to find that also in this case any value of $\boldsymbol{\pi}$ maps to a value of $\boldsymbol{\theta}$ where $J_{ij} \geq 0 \ \forall i, j$.

As it will be useful in the following, let us define a subspace of the space of KIM parameters, $\Theta^+ = (\mathbb{R}_+^{N \times N} \times \mathbb{R}^N) \subset \Theta$, where $\mathbb{R}_+$ is the set of positive real numbers including 0, hence Θ^+ only includes matrices J with positive entries.

Given these premises, we can now move to the main result of this appendix, by stating

Theorem 1. *Given a set of observations $\{\boldsymbol{X_t}\}$ for $t = 1, \ldots, T$, a Kinetic Ising Model $\{\{\boldsymbol{X}_t\}, p_{KIM}, \boldsymbol{\theta}\}$ and a V-DAR(1) model $\{\{\boldsymbol{X}_t\}, p_{VDAR}, \boldsymbol{\pi}\}$, there exists a unique and invertible map $f : \Pi \to \Theta^+ \subset \Theta$, i.e. the two models are equivalent if $\theta = f(\pi) \in \Theta^+$.*

In order to prove the theorem above, let us first construct the system of equations generating the mapping for the generic case $N > 2$. Following the same procedure used to construct Eq. A.6, defining $\mathrm{Log}(x) = \frac{e^{2x}}{1+e^{2x}}$ we find

$$M_n \cdot \begin{pmatrix} J_{i1} \\ J_{i2} \\ J_{i3} \\ \vdots \\ J_{iN} \\ h_i \end{pmatrix} \equiv \begin{pmatrix} 1 & 1 & ... & 1 & 1 & -1 \\ 1 & 1 & ... & 1 & -1 & -1 \\ 1 & 1 & ... & -1 & -1 & -1 \\ ... & ... & ... & ... & ... & ... \\ 1 & -1 & ... & -1 & -1 & -1 \\ -1 & -1 & ... & -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} J_{i1} \\ J_{i2} \\ J_{i3} \\ \vdots \\ J_{iN} \\ h_i \end{pmatrix} = -\tfrac{1}{2} \begin{pmatrix} \text{Log}^{-1}\left[(1-\nu_i)\chi_i\right] \\ \vdots \\ \vdots \\ \vdots \\ \text{Log}^{-1}\left[(\nu_i+(1-\nu_i)\chi_i\right] \end{pmatrix} \tag{A.7}$$

$\forall i = 1, ..., N.$

The above system is constructed by considering $n = N + 1$ independent states for the vector $\boldsymbol{X}_{t-1}$ and one possible realization of the variable X_t^i. Then, matching the probabilities of Eq. A.2 and A.1 for each of the possible independent combinations summarized in the matrix M_n, one finds Eq. A.7 for variable i. It is thus sufficient to apply the same constraints to all is to achieve N systems of n equations in n unknowns, each characterized by the matrix $M_n \in \{-1, 1\}^{(N+1)\times(N+1)}$.

For the sake of clarity, the n independent conditions giving Eq. A.7 read

$$\begin{cases} \text{Log}(h_i + \sum_{j\geq 1} J_{ij}) = \nu_i + (1-\nu_i)\chi_i & \text{if } X_{t-1}^1 = 1, X_{t-1}^2 = 1, ..., X_{t-1}^N = 1 \\ \text{Log}(h_i - J_{i1} + \sum_{j\geq 2} J_{ij}) = \nu_i(\sum_{j=2}^N \lambda_{ij}) + (1-\nu_i)\chi_i & \text{if } X_{t-1}^1 = 0, X_{t-1}^2 = 1, ..., X_{t-1}^N = 1; \\ ... & ... \\ \text{Log}(h_i - \sum_{j\leq n} J_{ij}) = (1-\nu_i)\chi_i & \text{if } X_{t-1}^i = 0, X_{t-1}^i = 0, ..., X_{t-1}^N = 0 \end{cases} \tag{A.8}$$

Given this result, if we call $L_i(\boldsymbol{\pi})$ the vector on the right hand side of Eq. A.7, as soon as M_n is invertible we obtain the mapping $f : \Pi \to \Theta$ as

$$f_i(\boldsymbol{\pi}) = -\frac{1}{2} M_n^{-1} L_i(\boldsymbol{\pi}) \;\; \forall i \tag{A.9}$$

where with a slight abuse of notation we call f_i the mapping onto the subspace of Θ indexed by i, that is $\boldsymbol{\theta}_i = (\{J_{ij}\}_j, h_i)$. The map is injective thanks to the linearity of the systems A.7. To prove that the inverse of M_n exists, we need to prove that its determinant is non-zero. We then start by proving the following

Proposition 1. *Given the determinant of the matrix M_{n-1}, then the determinant of the matrix M_n is*

$$\det(M_n) = (-1)^n 2 \det(M_{n-1}) \tag{A.10}$$

Proof. By means of the *minor expansion formula* (by using the minors associated with the elements of the first row), the determinant of M_n can be computed as

$$\begin{aligned}
\det(M_n) =& (+1)1 \begin{vmatrix} 1 & ... & 1 & -1 & -1 \\ 1 & ... & -1 & -1 & -1 \\ ... & ... & ... & ... & ... \\ -1 & ... & -1 & -1 & -1 \\ -1 & ... & -1 & -1 & -1 \end{vmatrix} + (-1)1 \begin{vmatrix} 1 & ... & 1 & -1 & -1 \\ 1 & ... & -1 & -1 & -1 \\ ... & ... & ... & ... & ... \\ 1 & ... & -1 & -1 & -1 \\ -1 & ... & -1 & -1 & -1 \end{vmatrix} + \\
&+ ... + (-1)^n (1) \begin{vmatrix} 1 & 1 & ... & 1 & -1 \\ 1 & 1 & ... & -1 & -1 \\ ... & ... & ... & ... & ... \\ 1 & -1 & ... & -1 & -1 \\ -1 & -1 & ... & -1 & -1 \end{vmatrix} + (-1)^{n+1}(-1) \begin{vmatrix} 1 & 1 & ... & 1 & -1 \\ 1 & 1 & ... & -1 & -1 \\ ... & ... & ... & ... & ... \\ 1 & -1 & ... & -1 & -1 \\ -1 & -1 & ... & -1 & -1 \end{vmatrix}
\end{aligned} \tag{A.11}$$

Here one notices that the first $n-2$ minors of the sum in Eq. A.11 are zero, because the last two columns of each $n-1 \times n-1$ matrix are indeed

equal (two $n-1$-dimensional vectors of -1). Thus, Eq. A.11 is simplified as

$$\det(M_n) = (-1)^n 2 \begin{vmatrix} 1 & 1 & ... & 1 & -1 \\ 1 & 1 & ... & -1 & -1 \\ ... & ... & ... & ... & ... \\ 1 & -1 & ... & -1 & -1 \\ -1 & -1 & ... & -1 & -1 \end{vmatrix} = (-1)^n 2 \det(M_{n-1}) \quad \text{(A.12)}$$

where we notice that the last two minors of (A.11) are equal to each other and correspond to the determinant of M_{n-1}. Eq. (A.12) then completes the proof of the proposition. □

Thanks to this result, we are now able to tackle the existence problem for the mapping between the two models, expressed by

Proposition 2. *There exists a solution of the problem of Eq. A.7 for any* $N \in \mathbb{N} \setminus 0$ *and this solution is unique.*

Proof. For $N = 1$, the solution can be explicitly computed as showed in Eqs. A.4 and A.5. For $N = 2$ the problem in Eq. A.7 is equivalent to Eq. A.6 and $\det(M_3) = 4$, thus there exists the inverse of the matrix M_3 and the solution is uniquely determined by solving the linear system of Eq. A.6. Because of Proposition 1, the determinant of M_n is different from zero, in particular

$$\det(M_n) = (-1)^{\sum_{l=4}^{n} l}(2^{n-3}) \det(M_3) = (-1)^{\sum_{l=4}^{n} l}(2^{n-3})4,$$

$\forall n > 3$ (or, equivalently, $\forall N > 2$), thus resulting in the existence of the inverse matrix of M_n. Hence, the solution of the problem in Eq. A.7 can be uniquely determined. This completes the proof of the proposition. □

Proof of Theorem 1. Given Propositions 1 and 2 we have proved that an equivalence mapping $f : \Pi \to \Theta$ exists between the two models and is injective. We still have left to prove that f is bijective in Θ^+, or that $f(\Pi) = \Theta^+$.

Let us start by proving that $f(\Pi) \subseteq \Theta^+$, or in other words that for any $\boldsymbol{\pi} \in \Pi$ the corresponding $f(\boldsymbol{\pi})$ has $J_{ij} \geq 0 \ \forall i, j$. In order to do so let us go back to Eq. A.8 and notice that, combining the equations by taking the difference between the first and the second, between the second and the third and so on, we obtain the N relations

$$\begin{cases} \nu_i(1 - \sum_{j\geq 2} \lambda_{ij}) = \text{Log}(h_i + \sum_{j\geq 2} J_{ij} + J_{i1}) - \text{Log}(h_i + \sum_{j\geq 2} J_{ij} - J_{i1}) \\ \quad \dots \\ \nu_i \lambda_{ik} = \text{Log}(h_i - \sum_{j<k} J_{ij} + \sum_{j\geq k+1} J_{ij} + J_{ik}) - \text{Log}(h_i - \sum_{j<k} J_{ij} + \sum_{j\geq k+1} J_{ij} - J_{ik}) \\ \quad \dots \\ \nu_i \lambda_{iN} = \text{Log}(h_i - \sum_{j<N} J_{ij} + J_{iN}) - \text{Log}(h_i - \sum_{j<N} J_{ij} - J_{iN}) \end{cases} \tag{A.13}$$

Being by definition $\nu\lambda_{ij} \geq 0$ for any i, j then it is always true that

$$\text{Log}(C + J_{ij}) - \text{Log}(C - J_{ij}) \geq 0$$

and, since $\text{Log}(x)$ is a monotonically increasing function of x, this can be true if and only if $J_{ij} \geq 0 \ \forall i, j$. Thus this condition is necessarily true if $\boldsymbol{\pi}$ is in the domain of f, meaning $f(\Pi) \subseteq \Theta^+$.

By following the same steps in the opposite direction it is straightforward to prove the reversed relation, that is $J_{ij} \geq 0$ is a sufficient condition to have $f^{-1}(\boldsymbol{\theta}) \in \Pi$ or equivalently $f^{-1}(\Theta^+) \subseteq \Pi$. Indeed for any $J_{ij} \geq 0$, the product $\nu_i\lambda_{ij}$ is $0 \leq \nu_i\lambda_{ij} \leq 1 \ \forall i, j$ given the system A.13 and that

$\mathrm{Log}(x) \in [0,1]$ for any x. Then, by summing all the equations in system A.13 one obtains

$$\nu_i = \mathrm{Log}(h_i + \sum_j J_{ij}) - \mathrm{Log}(h_i - \sum_j J_{ij})$$

which is also positive and smaller than 1 if $J_{ij} \geq 0 \ \forall j$. It then follows that all the λ_{ij} are $0 \leq \lambda_{ij} \leq 1 \ \forall i,j$. Finally, combining the first and last lines of Eq. A.8 one finds that $0 \leq \chi \leq 1$, thus $f^{-1}(\Theta^+) \subseteq \Pi$. Then, being both true that $f(\Pi) \subseteq \Theta^+$ and $f^{-1}(\Theta^+) \subseteq \Pi$ it follows that $f(\Pi) = \Theta^+$, which proves the theorem. □

In conclusion, the V-DAR(1) model for binary random variables $\{X_t^i\}$ is equivalent to the Kinetic Ising Model for spins $\{s_i(t)\}$ thanks to the existence of a unique mapping for both the random variables and the parameters as long as the J parameters of the Kinetic Ising Model are positive, as a consequence of the fact that the ν and λ parameters only account for positive lagged correlations among random variables.

Appendix B

Theoretical results on the distribution of effective fields

In this appendix we expand on what is the unconditional distribution $\phi(g)$ in Eq. 5.2 and how its parameters depend on the static parameters of the model. From an operational perspective this is the distribution that the effective fields show cross-sectionally across the whole sample, that is $g_i(t) \sim \phi(g)\ \forall i, t$, but it can also be calculated by giving a prior distribution to the static parameters of the model, $\Theta = (J, h, b)$. Finding this distribution can be useful to provide an easier and more accurate evaluation of the expected AUC of a forecast at a given β value, as it provides a bridge from the model parameters to the $AUC(\beta)$ we derived in Eq. 5.3 and shown in Fig. 5.1 in the main text.

Let us assume, as is standard in the literature (Crisanti and Sompolinsky (1988); Roudi and Hertz (2011b); Sakellariou (2013)), that the parameters Θ are structured in such a way that

$$J_{ij} \overset{iid}{\sim} \mathcal{N}(J_0/N, J_1^2/N - J_0^2/N^2), \quad J_{ii} = 0 \; \forall \, i$$
$$h_i \overset{iid}{\sim} \mathcal{N}(h_0, h_1^2)$$

while $b_{ik} = 0$ for simplicity. If that is the case then the distribution of $g_i(t)$ is itself a Gaussian, as $g_i(t)$ is now a sum of independent Gaussian random variables J_{ij} and h_i with random coefficients $s_j(t)$. Let us also define two average operators: the average $\langle \cdot \rangle$ over the distribution p of Eq. 2.3, also called the *thermal* average, and the average $\overline{\cdot}$ over the distribution of parameters, also known as the *disorder* average. Following Sakellariou (2013) we can then find the unconditional mean of s_i which reads

$$m_i = \langle s_i(t) \rangle = \langle \tanh \left[\beta g_i(t) \right] \rangle \tag{B.1}$$

where we have substituted the conditional mean value of $s_i(t)$ inside the brackets. This depends from the distribution of $g_i(t)$: assuming stationarity and calling $g_i^0 = \langle g_i(t) \rangle$ and $\Delta_i^2 = \langle g_i^2(t) \rangle - \langle g_i(t) \rangle^2$ we find that they are

$$g_i = \langle \sum_j J_{ij} s_j(t) + h_i \rangle = \sum_j J_{ij} m_j + h_i \tag{B.2a}$$

$$\begin{aligned} \Delta_i^2 &= \left\langle \left(\sum_j J_{ij} s_j(t) + h_i \right)^2 \right\rangle - \left\langle \sum_j J_{ij} s_j(t) + h_i \right\rangle^2 = \\ &= \sum_{j,k} J_{ij} J_{ik} \left[\langle s_j(t) s_k(t) \rangle - m_j m_k \right] \end{aligned} \tag{B.2b}$$

In Eq. B.2b spins $s_j(t)$ and $s_k(t)$ are mutually conditionally independent under p: this means that the only surviving terms are for $j = k$, thus finding

$$\Delta_i^2 = \sum_j J_{ij}^2 (1 - m_j^2) \tag{B.3}$$

Having determined the value of the mean and variance of the effective field of spin i we can now proceed to average over the disorder and find the unconditional distribution of effective fields at any time and for any spin, $\phi(g)$. First we can realize that the average of Eq. B.1 can now be substituted by a Gaussian integral

$$m_i = \int Dx \tanh\left[\beta\left(g_i + x\Delta_i\right)\right] \tag{B.4}$$

where Dx is a Gaussian measure of variable $x \sim \mathcal{N}(0,1)$. Then we can see that the unconditional mean of the fields distribution $\phi(g)$ is

$$\overline{\langle g_i(t)\rangle} = g_0 = \overline{\sum_j J_{ij} m_j + h_i} \tag{B.5}$$

Given the above results, if $J_{ii} = 0\ \forall i$ then the dependency between J_{ij} and m_j vanishes as $N \to \infty$, which means that the two can be averaged over the disorder separately. This results in the following expression for the unconditional mean of $g_i(t)$

$$g_0 = J_0\overline{m_j} + h_0 = J_0 m + h_0 \tag{B.6}$$

where

$$m = \overline{m_i} = \overline{\int Dx \tanh\left[\beta(g_i + x\Delta_i)\right]}$$

both the integral and the average here are of difficult solution and results have been provided by Crisanti and Sompolinsky (1988): they show that in the limit $N \to \infty$ and with $h_i = 0\ \forall i$ the system can be in one of two phases, a paramagnetic phase where $m = 0$ if β is smaller than a critical threshold $\beta_c(J_0)$ and $J_0 < 1$, and a ferromagnetic phase where $m \neq 0$ otherwise. In the following we report results for simulations in the paramagnetic phase,

as the inference is not possible in the ferromagnetic phase. To give better intuition let us consider the integral above in the limit $\beta \to 0$: then we can expand the hyperbolic tangent around 0 to find (since x has zero mean)

$$m \approx \overline{\beta g_i} = \overline{\beta \left(\sum_j J_{ij} m_j + h_0 \right)} = \beta (J_0 m + h_0) \tag{B.7}$$

which in turn leads to an approximated solution for g_0 in the limit $\beta \to 0$

$$g_0 \approx h_0 \left(\frac{\beta J_0}{1 - \beta J_0} + 1 \right)$$

Moving on to the variance of g the calculation is straightforward. Adding the mean over the disorder to Eq. B.2b we find

$$\begin{aligned} g_1^2 &= \overline{\left\langle \left[\sum_j J_{ij} s_j(t) + h_i \right]^2 \right\rangle} - \overline{\left\langle \sum_j J_{ij} s_j(t) + h_i \right\rangle}^2 = \\ &= \overline{\sum_j J_{ij}^2 + h_i^2 + 2h_i \sum_j J_{ij} m_j} - \overline{\sum_j J_{ij} m_j + h_i}^2 = \\ &= J_1^2 + h_1^2 - J_0^2 m^2 \end{aligned} \tag{B.8}$$

Equations B.6 and B.8 can then be used to calculate, given the parameters of the distribution generating Θ, the values of g_0 and g_1 that are to be plugged in the distribution $\phi(g)$ of Eq. 5.3.

We simulated a Kinetic Ising Model with $N = 100$ spins for $T = 2000$ time steps at different constant values of β and then measured the AUC of predictions assuming the parameters are known. In Fig. B.1 we report a comparison between these simulated values and the theoretical ones provided by Eq. 5.3 varying β and the hyperparameters J_0, J_1, h_0 and h_1 in the Gaussian setting we just discussed and adopting the expansion for $\beta \to 0$. We see that the approximation for small β of Eq. B.7 does not affect the

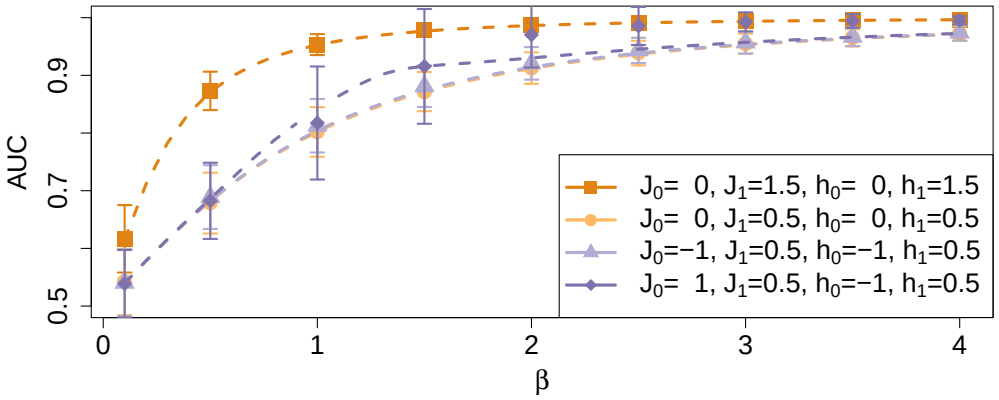

Figure B.1: Comparison between the AUC estimated on data simulated from a Kinetic Ising Model and the theoretically derived AUC with Gaussian distribution of the J and h parameters, varying β and the hyperparameters J_0, J_1, h_0 and h_1. Plot points report average simulated values for a given β with error bars at ± 1 standard deviation, dashed lines report theoretical values predicted by Eq. 5.3.

accuracy of the theoretical prediction for larger values of β and that the mean is correctly captured by Eq. 5.3. The only exception to this is found for $\beta > 1$ and $J_0 = 1$, which according to the literature is close to the line of the ferromagnetic transition: in this case the small β approximation fails to predict the simulated values. Larger values of N and T (not shown here) produce narrower error bars.

The general effect we see from Fig. B.1 is that higher variance of the J and h parameters leads to higher AUC values leaving all else unchanged (orange squares and yellow circles), while moving the means has little effect as long as the system is in its paramagnetic phase.

These results are easy to obtain thanks to the assumption that the model parameters J and h have Gaussian distributed entries, but in principle the distribution $\phi(g)$ can be derived also for other distributions, albeit probably necessiting numerical solutions rather than the analytical ones we presented here.