Inference of the kinetic Ising model with heterogeneous missing data

Carlo Campajola,[1,*] Fabrizio Lillo,[2] and Daniele Tantari[3]
[1]*Scuola Normale Superiore di Pisa, piazza dei Cavalieri 7, 56126 Pisa, Italy*
[2]*University of Bologna - Department of Mathematics, piazza di Porta San Donato 5, 40126 Bologna, Italy*
[3]*University of Florence - Department of Economics and Management, via delle Pandette 9, 50127 Firenze, Italy*

(Received 22 March 2019; published 28 June 2019)

We consider the problem of inferring a causality structure from multiple binary time series by using the kinetic Ising model in datasets where a fraction of observations is missing. Inspired by recent work on mean field methods for the inference of the model with hidden spins, we develop a pseudo-expectation-maximization algorithm that is able to work even in conditions of severe data sparsity. The methodology relies on the Martin-Siggia-Rose path integral method with second-order saddle-point solution to make it possible to approximate the log-likelihood in polynomial time, giving as output an estimate of the couplings matrix and of the missing observations. We also propose a recursive version of the algorithm, where at every iteration some missing values are substituted by their maximum-likelihood estimate, showing that the method can be used together with sparsification schemes such as lasso regularization or decimation. 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 and when some of the hypotheses that are necessary to the saddle-point approximation are violated, such as the small couplings limit and the assumption of statistical independence between couplings.

DOI: 10.1103/PhysRevE.99.062138

I. INTRODUCTION

Ising-like models and their countless variations have been used throughout recent decades to describe data or model systems with the most diverse nature [1–5] 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 behaviors mimicking the features of these systems [2,6], and use a deductive approach to explain the stylized properties of data that we observe in the real world. On the other hand, one can use these models in the fashion of descriptive and forecasting models [1,4,5,7], 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 regarded as the inverse formulation of the model, while the former is the direct formulation.

A model of this family has recently been revamped for time-series data, i.e., the nonequilibrium or kinetic Ising model [8,9], 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 [10] all the way to traders in a financial market [6,11]. Recent work on the inverse kinetic Ising model has led to the development of exact solutions [12], cavity methods [13], and mean field (MF) [14] 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 [15].

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 [16–18]. The literature on the kinetic Ising model has previously considered mainly the inference problem in the presence of hidden nodes [15], i.e., part of the spins are *never* observed, but it is known that they exist and interact with the visible nodes (i.e., spins). 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 for a significant fraction of the time. 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 that 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

*carlo.campajola@sns.it

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 [19], the interest in the problem has grown and different kinds of deletion [20], imputation [21], and estimation [22–24] 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 [25].

Taking inspiration from the work by Dunn *et al.* [15], 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 time stamps a missing value is recorded. The result is an algorithm closely related to an expectation-maximization (EM) method [22], iteratively alternating a step of log-likelihood gradient ascent [26] and the self-consistent resolution of Thouless-Anderson-Palmer (TAP) equations [14], 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 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.

The paper is organized as follows: in Sec. II, we define the considered kinetic Ising model, we explain the inference method in detail, and describe the approximations needed to make the algorithm converge in feasible time. In Sec. III, we present results on synthetic data and give an overview of the performance that can be expected with different data specifications. Section IV concludes the article.

II. SOLVING THE INVERSE PROBLEM WITH MISSING VALUES

The kinetic Ising model (or nonequilibrium Ising model) [8] is defined on a set of spins $y \in \{-1, +1\}^N$, whose dynamics is described by the transition probability mass function,

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

where $\langle i, j\rangle$ is a sum over neighboring 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 neighbors is substituted by a sum on all the other spins. This recasts the transition probability into the following form:

$$p[y(t+1)|y(t)] = Z^{-1}(t) \exp\left[\sum_{i=1}^{N} y_i(t+1) \tilde{g}_i(t)\right], \quad (2)$$

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, and thus they must have full rank, which 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 $\sum_i'$ as the sum over indices i at time $t+1$, while the regular $\sum_i$ indicates a sum over indices i at time t and $\sum_i^-$ is 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

$$p[\{s', \sigma'\}|\{s, \sigma\}] = Z^{-1} \exp\left[\sum_i{}' s_i' g_i' + \sum_a{}' \sigma_a' g_a'\right]. \quad (3)$$

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

$$g_i = \sum_j J_{ij}^{oo} s_j^- + \sum_b J_{ib}^{oh} \sigma_b^- + h_i,$$
$$g_a = \sum_j J_{aj}^{ho} s_j^- + \sum_b J_{ab}^{hh} \sigma_b^- + h_a, \quad (4)$$

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 the 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), taking the trace over the missing values,

$$p[\{s\}] = \mathrm{Tr}_\sigma \prod_t p[\{s', \sigma'\}|\{s, \sigma\}]. \tag{5}$$

To solve the problem, our approach is closely related to the one developed by Dunn *et al.* [15], where the authors investigate a system where only a subset of spins is observable. The extension to our case is presented below.

The trace of Eq. (5) is computationally intractable for large systems with many hidden variables. However, the Martin-Siggia-Rose (MSR) path integral formulation [27] allows one to decouple spins and perform the trace at the cost of computing a high-dimensional integral. Define the functional

$$\mathcal{L}[\psi] = \ln \mathrm{Tr}_\sigma \prod_t \exp\left[\sum_a \psi_a \sigma_a\right] p[\{s', \sigma'\}|\{s, \sigma\}]. \tag{6}$$

Notice that this is equivalent to the log-likelihood if $\psi_a(t) = 0$ $\forall a, t$, and thus the goal of the calculation will be to efficiently maximize $\mathcal{L}[\psi]$ 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. Let

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

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

$$\mathcal{L}[\psi] = \ln \int \mathcal{D}\mathcal{G} \exp[\Phi], \tag{7}$$

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

$$\Phi = \ln \mathrm{Tr}_\sigma \exp\left[Q + \Delta + \sum_t \sum_a \psi_a \sigma_a \right]. \tag{8}$$

Now the trace can be easily computed since the introduction of the δ function has decoupled the σ's by fixing the value of the local fields g.

As mentioned, the cost is computing the integral of Eq. (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 $\mathcal{G}$.

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 $p[\{\sigma\}|\{s, J, h\}]$.

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. (7) is

$$\delta\mathcal{L} = -\tfrac{1}{2} \ln \det[\nabla^2_{\mathcal{G}} \mathcal{L}],$$

where $\nabla^2_{\mathcal{G}}\mathcal{L}$ is the Hessian matrix in the $\mathcal{G}$ space of $\mathcal{L}$ evaluated at the saddle point. The resulting structure of the matrix, shown in the Appendix for the sake of space, is sparse and almost block diagonal.

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

$$\begin{aligned} \ln\det(\alpha+\beta) &= \ln\det(\alpha) + \ln\det[\mathbb{I} + \alpha^{-1}\beta] \\ &= \ln\det(\alpha) + \mathrm{Tr}\ln[\mathbb{I} + \alpha^{-1}\beta] \\ &\approx \ln\det(\alpha) + \mathrm{Tr}[\alpha^{-1}\beta] \\ &\quad + \tfrac{1}{2}\mathrm{Tr}\{[\alpha^{-1}\beta]^2\} + \cdots . \end{aligned} \tag{9}$$

Given that α is block diagonal, so will 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 the order of $1/N$ and variance of the order of J_1^2/N, which means $\ln\det(\alpha)$ is quadratic in J_1 (see the Appendix). 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,

$$\begin{aligned} \delta\mathcal{L} \approx &-\frac{1}{2}\sum_t {\sum_i}' \left\{ \left[1 - \tanh^2(g'_i)\right] \sum_b \left[J^{oh\prime}_{ib}\right]^2 \left(1 - \mu_b^2\right) \right\} \\ &-\frac{1}{2}\sum_t {\sum_a}' \left\{ \left[\mu'^2_a - \tanh^2(g'_a)\right] \sum_b \left[J^{hh\prime}_{ab}\right]^2 \left(1 - \mu_b^2\right) \right\}. \end{aligned}$$

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

$$l_a(t) = \frac{\partial(\delta\mathcal{L})}{\partial\psi_a(t)}. \tag{10}$$

Implementation of 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 neighborhood, 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

$$\Gamma[\mu] = \mathcal{L} - \sum_t \sum_a \psi_a(t)\mu_a(t) \; s.t. \; -\psi_a(t) = \frac{\partial \Gamma[\mu]}{\partial \mu_a(t)}, \tag{11}$$

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

$$\Gamma_0[\mu] = \sum_t \left\{ \sum_i{}' \left[s_i' g_i^{0\prime} - \ln 2\cosh\left(g_i^{0\prime}\right)\right] + \sum_a{}' \left[\mu_a' g_a^{0\prime} - \ln 2\cosh\left(g_a^{0\prime}\right)\right] + \sum_a S[\mu_a] \right\}, \tag{12}$$

where $S[x] = -\frac{1+x}{2}\ln(\frac{1+x}{2}) - \frac{1-x}{2}\ln(\frac{1-x}{2})$ 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.

Once this approximate log-likelihood is maximized and the final iteration of the expectation part of the algorithm is finished, the result is an (approximated) 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 the gradient ascent step, i.e., 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.

III. 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:

(1) The average observation frequency, taking the Bernoulli probabilities $p_i = p, \forall i = 1, \ldots, N$.

(2) 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.

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

(4) 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 [28] or the decimation procedure [29] to select the links.

(5) 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.

(6) 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 1, 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 that 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 EM (R-EM): by properly iterating the algorithm multiple times, it allows one to boost data artificially, thus achieving good performances even when the fraction of missing values is particularly high.

In Test 2, 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 a 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 1, 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 3 aims to assess 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 modeling perspective it is inversely proportional to the impact of the noise on the dynamics. Given the approximation of Eq. (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 4, 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 5, we study how the independent and identically distributed (i.i.d.) assumption made in Eq. (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 [30]. 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 6 by looking at the dependency of performance metrics on the ratio T/N, which 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 mainly using 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 inferred coupling matrix $\hat{J}$.

A. Test 1: 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 $T = 10\,000$ 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 Fig. 1(a), 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 $p \geqslant 0.8$, 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 $\hat{\sigma}_a = \mathrm{sign}(m_a)$ 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 Figs. 1(b) and 1(c). We defer a more rigorous treatment of this recursive method to future work, while still proposing it here as we find it surprisingly accurate.

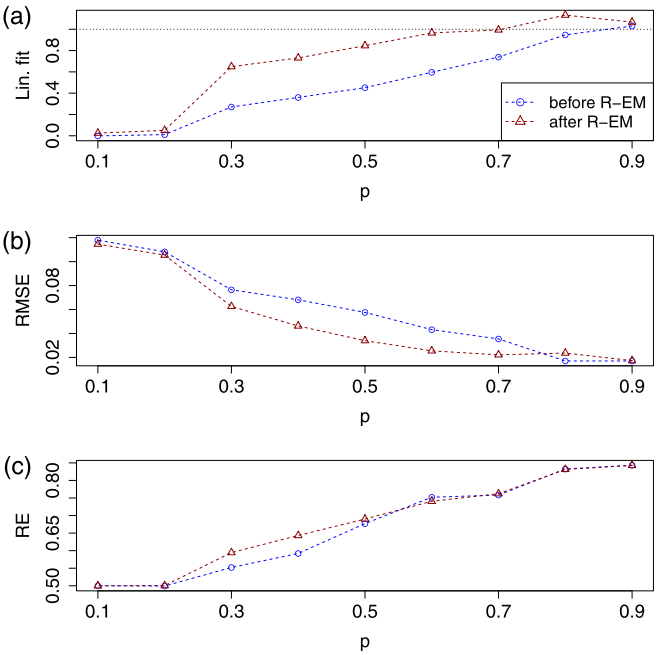

FIG. 1. (a) Angular coefficient of the linear fit $\hat{J}_{ij} = aJ_{ij} + c$ before and after R-EM varying the average observation density p. (b) Root-mean-squared error on the couplings. (c) Reconstruction efficiency.

Figure 1(c) 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 ($p \leqslant 0.2$), 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.

B. Test 2: Heterogeneous p_i

In Test 2, we want to highlight how our model is a generalization of the one studied extensively by Dunn *et al.* [15] 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 1, 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

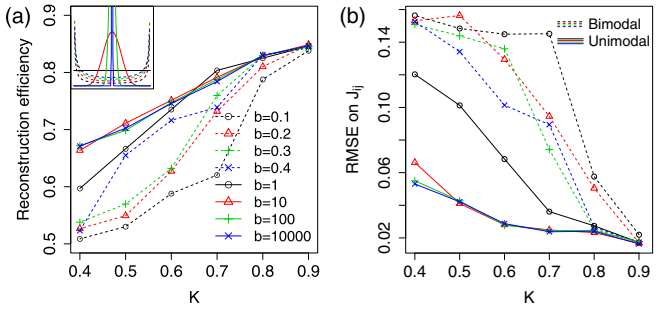

FIG. 2. (a) Reconstruction efficiency as a function of K with different beta parameters. Inset: the probability density function of the adopted beta distributions with $K = 0.5$ (color coding is the same as in the main panel). (b) Root-mean-square error on the couplings as a function of K with different beta parameters.

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-δ 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 Fig. 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.* [15] model approximated by $B[a(K), 0.1]$ is identified with reasonable performance only when $K \geqslant 0.8$. This is extremely mitigated when the observations are more homogeneously distributed, particularly in the case of the coupling coefficients whose estimation seems to require a rather homogeneous distribution of observations among variables in order 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 Fig. 3, we plot the root-mean-square error on couplings conditional on the probability of subsequently observing 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 J's 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 observed pairs of spins and an important loss in accuracy, but with a bias component that is mitigated for the most frequently observed pairs.

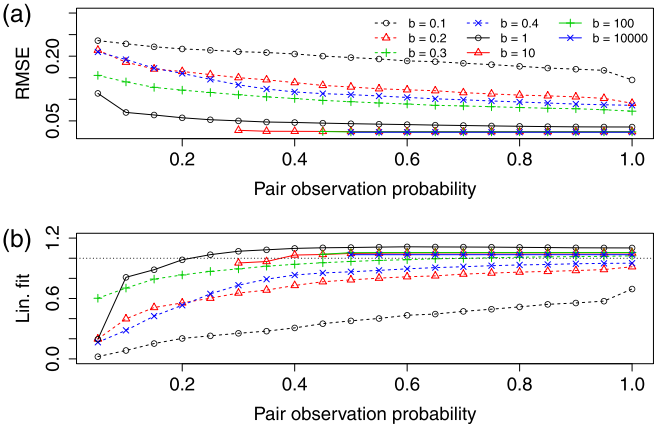

FIG. 3. Quality of inference varying the probability of observing the end nodes at subsequent times. (a) RMSE for different values of the beta b parameter with mean $K = 0.7$. (b) Linear fit coefficient for different values of the b parameter, $K = 0.7$.

C. Test 3: 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 behavior 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 the couplings can 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 $N = 100$, $T = 10\,000$, $p_i = p = 0.8$, and J_1 ranging from 0.05 to 13. We see from Fig. 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 Fig. 4(a), we plot the reconstruction efficiency which has a steady increase and saturates towards 1 after $J_1 \simeq 5$. Figure 4(b) shows the relative RMSE, that is RMSE/J_1, and we

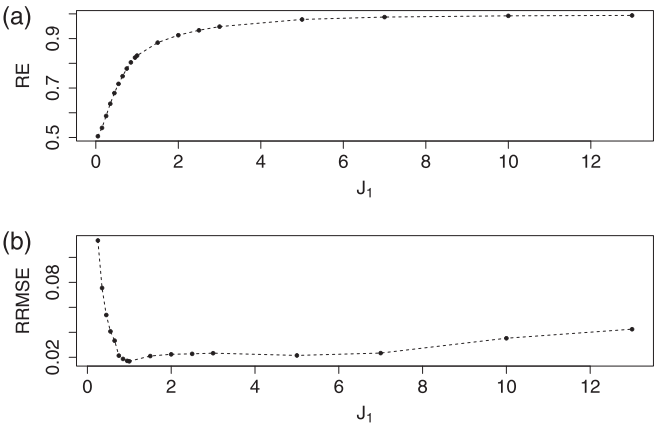

FIG. 4. (a) Reconstruction efficiency as a function of J_1. (b) Rescaled RMSE (by J_1) on the couplings as a function of J_1.

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. (9), the algorithm seems to work efficiently, even in cases where the variance of the couplings J_1^2/N is of the order of 1, albeit a region of optimality for the inference of the couplings seems to lie within $0.5 \leqslant J_1 \leqslant 7$.

D. Test 4: Impact of network structure

We test the algorithm performance on a 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 one 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 $T = 10\,000$, and with a probability of observing the variables of $p \in \{0.8, 0.6, 0.4\}$. One then 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 *et al.* [29]. 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 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 the case $h_i = 0\,\forall\, i$, $\mathcal{L}_1 = -\sum_t M(t) \ln$. The tilted log-likelihood takes the form

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

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 $\mathcal{L}(1) = \mathcal{L}_1$, and 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 Figs. 5 and 6, the results of the test. We observe how the Receiver Operating Characteristic (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 nonsparse network and when the data have 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.

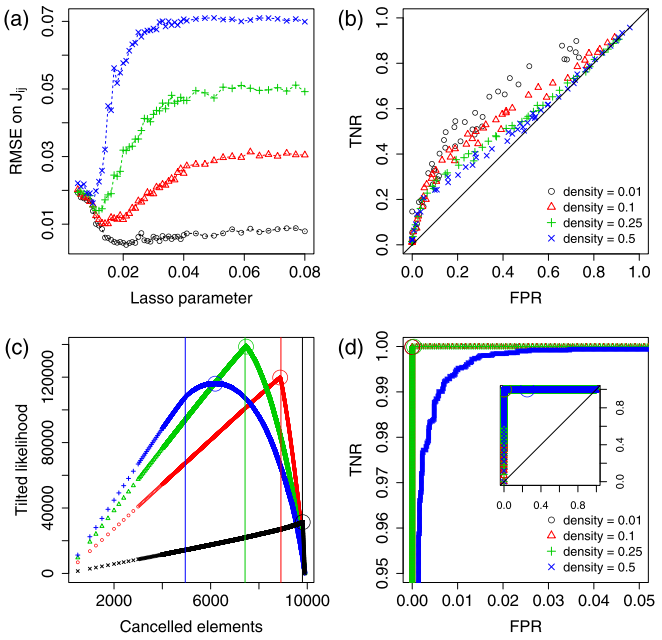

FIG. 5. (a),(b) Results from the lasso with 80% observations: (a) RMSE on couplings as a function of the lasso parameter; (b) ROC curves. (c),(d) Results from the decimation procedure with 80% observations: (c) Tilted likelihood evolution through the decimation process, where vertical lines show the 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.

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 no longer reliable in that the tilted likelihood becomes nonconvex, as shown in Fig. 6, and the maximum is not in the correct 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 nonconvex 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.

E. Test 5: Impact of asymmetricity assumption

Another assumption we made to perform the calculations in Eq. (9) was that the J_{ij} are i.i.d. Gaussian random variables.

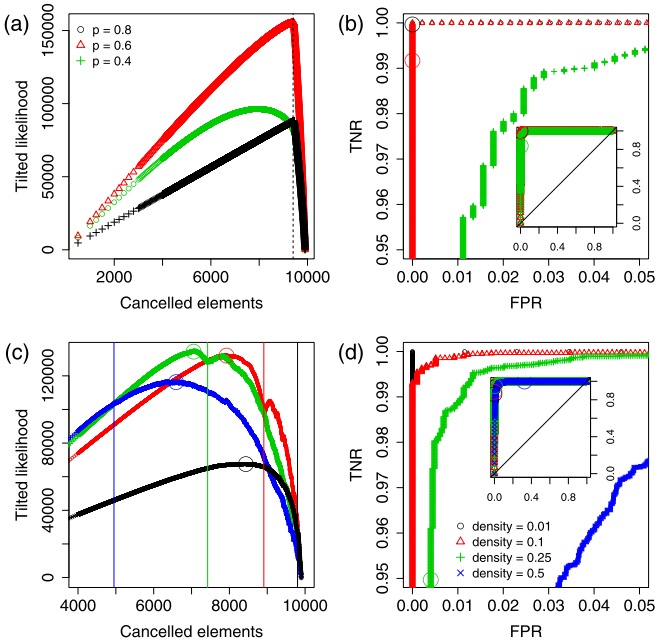

FIG. 6. (a),(b) Results from the decimation procedure with 80%, 60%, and 40% observations available and a network density of 0.05: (a) Tilted likelihood evolution through the decimation, where the vertical line shows the correct number of null elements; (b) ROC curves through the decimation process with different observation densities. (c),(d) Results from the decimation introducing local fields h: (c) Tilted likelihood, where the vertical lines show the correct number of null elements; (d) ROC curves. The introduction of local fields makes the tilted likelihood nonconvex and seriously affects the performance.

In the case of social networks and trade networks, reciprocity, which is the correlation between J_{ij} and J_{ji}, is often found to be much higher than what would be expected in an i.i.d. setting [30]. We ask ourselves how impactful this assumption is on the outcome of the inference and we test the algorithm on data generated from a model with $N = 100$, $T = 10\,000$, $p_i = p = 0.8$, $J_1 = 1$ and such that $\mathrm{Cor}(J_{ij}, J_{ji}) = \rho$, $i \neq j$. We show the results for this series of tests in Fig. 7. What we find is that the ρ parameter barely affects the performance and even makes it easier to infer the hidden variables,

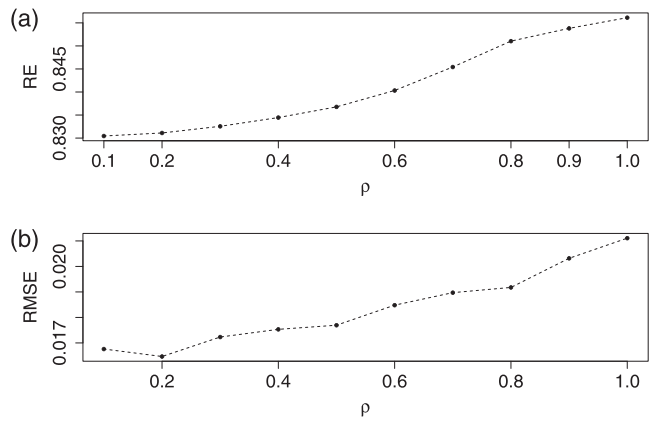

FIG. 7. (a) Reconstruction efficiency varying the correlation between symmetric elements of J. (b) RMSE on the couplings.

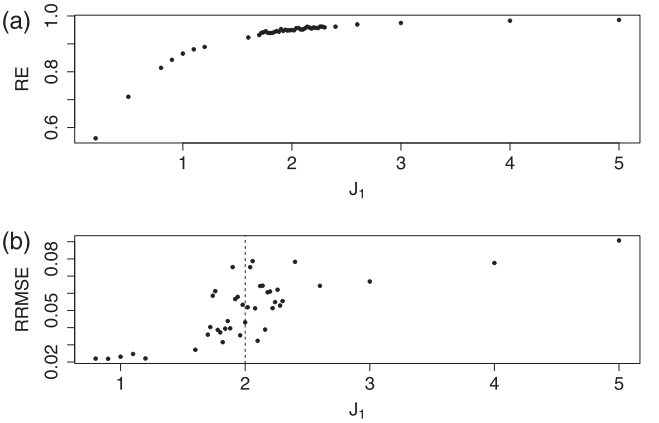

FIG. 8. (a) Reconstruction efficiency as a function of J_1 in the SK model. (b) Rescaled RMSE on couplings as a function of J_1.

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. We then turn our attention to the extreme case of $\rho = 1$, corresponding to the well-known Sherrington-Kirkpatrick (SK) model [31], 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 no longer easy to infer. 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. 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 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.

F. Test 6: 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 modeled, N. We run simulations with $N = 100$, $J_1 = 1$, $p_i = p = 0.8$ and varying T between 100 and 25 000, and report the results in Fig. 9. It can be seen that the RMSE on J_{ij} diminishes, after $T/N = 20$, with what might look like a power-law behavior with exponent close to 0.5, although we

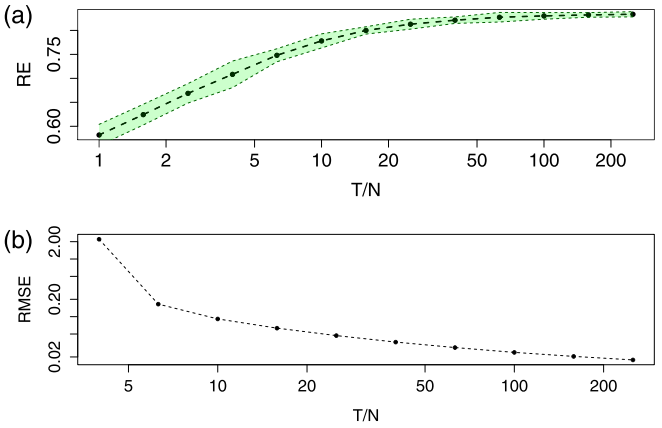

FIG. 9. (a) Reconstruction efficiency as a function of the T/N ratio. (b) RMSE as a function of the T/N ratio. The area in green is one standard deviation from the mean over 30 repetitions.

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.

G. 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, i.e., whether they go "against" the flow (i.e., sell when the price rises and vice versa) or follow the trend. In the model, this is introduced by adding a set of linear parameters β in the local fields that couple the variables to the driver,

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

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 β_k parameters is found for our performance metrics.

IV. CONCLUSIONS

In this article, we develop 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 for future research will be on opinion spreading in financial markets where transactions occur at high frequency, such as the foreign exchange or the cryptocurrency markets. We indeed envision that 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.

ACKNOWLEDGMENTS

The authors are grateful to Prof. Matteo Marsili and to the participants of the 2018 Spring College on the Physics of Complex Systems (Trieste) for insightful comments and discussions. D.T. acknowledges GNFM-Indam for financial support. C.C. and D.T. acknowledge Scuola Normale Superiore for financial support of the project SNS18_A_TANTARI.

APPENDIX A: ZERO-ORDER SADDLE-POINT APPROXIMATION

We start from Eq. (8) in the main text, where we have introduced the Dirac δ function to obtain a functional form of $\mathcal{L}$ for which the trace can be calculated. The result is the functional Φ of Eq. (8), which, once the trace is done, reads

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

This is the function to be extremized to find the saddle point around which the integral is to be computed. Setting $\nabla_{\mathcal{G}}\Phi = 0$ gives

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

which, substituted in Φ, give the zero-order solution to the saddle-point integral. The other ingredient is the vector of magnetizations m which, as stated in the main text, is obtained by exploiting the property of $\mathcal{L}$ being the moment generating functional for σ. Thus, we find

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

APPENDIX B: SECOND-ORDER SADDLE-POINT APPROXIMATION

The second-order approximation requires the calculation of the determinant of the Hessian of the log-likelihood, $\nabla^2_{\mathcal{G}}\mathcal{L}$, taken at the saddle-point coordinates. 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 submatrices $A^{tt'}, \ldots, G^{tt'}$, given by

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

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

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

It is thus clear that the determinant of this matrix, under the approximation in Eq. (9) of the main text, is

$$
\det\left[\nabla^2_{\mathcal{G}}\mathcal{L}\right] \approx \prod_t (\det A^{tt} \det B^{tt} + \mathbb{I})(\det C^{tt} \det D^{tt} + \mathbb{I}),
$$

which leads to the form of the correction reported in the main text.

As mentioned in Eq. (10) in the main text, introducing the Gaussian correction shifts the magnetizations by a quantity

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

Thus we rewrite both Γ_0 and $\delta\mathcal{L}$, substituting $\mu_a(t)|_{\psi_a(t)=0} = m_a(t) - l_a(t)|_{\psi_a(t)=0}$ in the functional and in the saddle-point solutions for g, and obtain

$$\Gamma_0[m] = \sum_t \left\{ \sum_i{}' \left[s'_i g'_i - \ln\cosh(g'_i) \right] + \sum_a{}' \left[m'_a g'_a - \ln\cosh(g'_a) \right] + \sum_a S[m_a] \right.$$
$$\left. - \sum_i{}' \left[s'_i - \tanh(g'_i) \right] \sum_a J^{oh\,\prime}_{ia} l_a - \sum_a{}' \left[m'_a - \tanh(g'_a) \right] \sum_b J^{hh\,\prime}_{ab} l_b - \sum_a{}' l'_a \left[g'_a - \sum_b J^{hh\,\prime}_{ab} l_b \right] + \sum_a l_a \tanh^{-1}(m_a) \right\},$$

where, in this last formula, $g(t)$ have become the fields of Eq. (4) in the main text with m in place of σ. Given this last expression, it can be seen that since $l_a(t)$ is already quadratic in J and always multiplies an object of the order of one, all terms involving $l_a(t)$ are higher order and can be neglected in the current approximation.

Skipping to Eq. (12) in the main text and adding the Gaussian correction to the Γ_0 functional, we obtain the final form of the approximated log-likelihood to be maximized,

$$\Gamma_1[m] = \Gamma_0[m] - \frac{1}{2} \sum_t \sum_i{}' \left\{ [1 - \tanh^2(g'_i)] \sum_b \left[J^{oh\,\prime}_{ib} \right]^2 \left(1 - m_b^2 \right) \right\}$$
$$- \frac{1}{2} \sum_t \sum_a{}' \left\{ \left[m_a^{2\,\prime} - \tanh^2(g'_a) \right] \sum_b \left[J^{hh\,\prime}_{ab} \right]^2 \left(1 - m_b^2 \right) \right\}.$$

The final result is the formulas necessary for the EM-like algorithm, namely, the log-likelihood gradient and the self-consistent relations for the magnetizations. The first takes the form

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

where the fields g and their derivatives are given by

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

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

$$m_a = \tanh \left(g_a + m_a \left\{ \sum_i{}' [1 - \tanh^2(g'_i)] \sum_{kl} G'_{ik} J^2_{kl} F^T_{la} + \sum_b{}' \left[m_b^{2\,\prime} - \tanh^2(g'_b) \right] \sum_{kl} F'_{bk} J^2_{kl} F^T_{la} \right.\right.$$
$$\left. - \sum_c{}^{-} \sum_{kl} F_{ak} J^2_{kl} F^{T\,-}_{lc} \left(1 - m_c^{2\,-} \right) \right\} + \sum_i{}' \left[s'_i - \tanh(g'_i) \right] \sum_{kl} G'_{ik} J_{kl} F^T_{la}$$
$$+ \sum_b{}' \left[m'_b - \tanh(g'_b) \right] \sum_{kl} F'_{bk} J_{kl} F^T_{la} + \sum_i{}' \frac{\tanh(g'_i)}{\cosh^2(g'_i)} \sum_{oqb} G'_{io} J_{oq} F^T_{qb} \left(1 - m_b^{2\,\prime} \right) \sum_{kl} G'_{ik} J_{kl} F^T_{la}$$
$$\left. + \sum_c{}' \frac{\tanh(g'_c)}{\cosh^2(g'_c)} \sum_{oqb} F'_{co} J_{oq} F^T_{qb} \left(1 - m_b^{2\,\prime} \right) \sum_{kl} F'_{ck} J_{kl} F^T_{la} \right). \tag{B1}$$

[1] T. Bury, Market structure explained by pairwise interactions, Physica A: Stat. Mech. Appl. **392**, 1375 (2013).

[2] J.-P. Bouchaud, Crises and collective socio-economic phenomena: Simple models and challenges, J. Stat. Phys. **151**, 567 (2013).

[3] S. Tanaka and H. A. Scheraga, Model of protein folding: Incorporation of a one-dimensional short-range (Ising) model into a three-dimensional model, Proc. Natl. Acad. Sci. **74**, 1320 (1977).

[4] S. Cocco, R. Monasson, L. Posani, and G. Tavoni, Functional networks from inverse modeling of neural population activity, Curr. Opin. Syst. Biol. **3**, 103 (2017).

[5] B. Kadirvelu, Y. Hayashi, and S. J. Nasuto, Inferring structural connectivity using Ising couplings in models of neuronal networks, Sci. Rep. **7**, 8156 (2017).

[6] S. Bornholdt, Expectation bubbles in a spin model of markets: Intermittency from frustration across scales, Int. J. Mod. Phys. C **12**, 667 (2001).

[7] T. Ibuki, S. Higano, S. Suzuki, Jun-ichi Inoue, and A. Chakraborti, Statistical inference of co-movements of stocks during a financial crisis, J. Phys.: Conf. Ser. **473**, 012008 (2013).

[8] B. Derrida, E. Gardner, and A. Zippelius, An exactly solvable asymmetric neural network model, Europhys. Lett. **4**, 167 (1987).

[9] A. Crisanti and H. Sompolinsky, Dynamics of spin systems with randomly asymmetric bonds: Ising spins and glauber dynamics, Phys. Rev. A **37**, 4865 (1988).

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

[11] D. Sornette, Physics and financial economics (1776–2014): Puzzles, Ising and agent-based models, Rep. Prog. Phys. **77**, 062001 (2014).

[12] J. Sakellariou, Inverse inference in the asymmetric ising model, Ph.D. thesis, Université Paris Sud-Paris XI, 2013.

[13] P. Zhang, Inference of kinetic Ising model on sparse graphs, J. Stat. Phys. **148**, 502 (2012).

[14] Y. Roudi and J. Hertz, Dynamical tap equations for nonequilibrium Ising spin glasses, J. Stat. Mech. (2011) P03031.

[15] B. Dunn and Y. Roudi, Learning and inference in a nonequilibrium Ising model with hidden nodes, Phys. Rev. E **87**, 022127 (2013).

[16] Y. Aït-Sahalia, J. Fan, and D. Xiu, High-frequency covariance estimates with noisy and asynchronous financial data, J. Am. Stat. Assoc. **105**, 1504 (2010).

[17] G. Buccheri, G. Bormetti, F. Corsi, and F. Lillo, A score-driven conditional correlation model for noisy and asynchronous data: An application to high-frequency covariance dynamics, https://ssrn.com/abstract=2912438 (unpublished).

[18] F. Corsi, S. Peluso, and F. Audrino, Missing in asynchronicity: A Kalman-EM approach for multivariate realized covariance estimation, J. Appl. Econom. **30**, 377 (2015).

[19] D. B. Rubin, Inference and missing data, Biometrika **63**, 581 (1976).

[20] E. R. Buhi, P. Goodson, and T. B. Neilands, Out of sight, not out of mind: Strategies for handling missing data, Amer. J. Health Behav. **32**, 83 (2008).

[21] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Vol. 81 (John Wiley & Sons, Hoboken, NJ, 2004).

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B (Methodological) **39**, 1 (1977).

[23] B. M. Marlin and R. S. Zemel, Collaborative prediction and ranking with non-random missing data, in *Proceedings of the Third ACM Conference on Recommender Systems* (ACM, New York, NY, 2009), pp. 5–12.

[24] K. Mohan, J. Pearl, and J. Tian, Graphical models for inference with missing data, in *Advances in Neural Information Processing Systems*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (2013), pp. 1277–1285, https://papers.nips.cc/book/advances-in-neural-information-processing-systems-26-2013.

[25] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Vol. 793 (John Wiley & Sons, Hoboken, NJ, 2019).

[26] Yu. Nesterov, Accelerating the cubic regularization of Newton's method on convex problems, Math. Program. **112**, 159 (2008).

[27] P. C. Martin, E. D. Siggia, and H. A. Rose, Statistical dynamics of classical systems, Phys. Rev. A **8**, 423 (1973).

[28] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B (Methodological) **58**, 267 (1996).

[29] A. Decelle and P. Zhang, Inference of the sparse kinetic Ising model using the decimation method, Phys. Rev. E **91**, 052136 (2015).

[30] T. Squartini, F. Picciolo, F. Ruzzenenti, and D. Garlaschelli, Reciprocity of weighted networks, Sci. Rep. **3**, 2729 (2013).

[31] S. Kirkpatrick and D. Sherrington, Infinite-ranged models of spin-glasses, Phys. Rev. B **17**, 4384 (1978).