

Inverse Ising techniques to infer underlying mechanisms from data

Hong-Li Zeng(曾红丽), Erik Aurell
Citation: Chin. Phys. B . 2020, 29(8): 080201 . **doi:** 10.1088/1674-1056/ab8da6
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Symmetry properties of fluctuations in an actively driven rotor

He Li(李赫), Xiang Yang(杨翔), Hepeng Zhang(张何朋)
Chin. Phys. B . 2020, 29(6): 060502 . **doi:** 10.1088/1674-1056/ab862b

Ergodicity recovery of random walk in heterogeneous disordered media

Liang Luo(罗亮), Ming Yi(易鸣)
Chin. Phys. B . 2020, 29(5): 050503 . **doi:** 10.1088/1674-1056/ab8212

Accurate reconstruction of the optical parameter distribution in participating medium based on the frequency-domain radiative transfer equation

Yao-Bin Qiao(乔要宾), Hong Qi(齐宏), Fang-Zhou Zhao(赵方舟), Li-Ming Ruan(阮立明)
Chin. Phys. B . 2016, 25(12): 120201 . **doi:** 10.1088/1674-1056/25/12/120201

Co-evolution of the brand effect and competitiveness in evolving networks

Guo Jin-Li
Chin. Phys. B . 2014, 23(7): 070206 . **doi:** 10.1088/1674-1056/23/7/070206

A self-adaptive stochastic resonance system design and study in chaotic interference

Lu Kang, Wang Fu-Zhong, Zhang Guang-Lu, Fu Wei-Hong
Chin. Phys. B . 2013, 22(12): 120202 . **doi:** 10.1088/1674-1056/22/12/120202

TOPICAL REVIEW — Machine learning in statistical physics

Inverse Ising techniques to infer underlying mechanisms from data*

Hong-Li Zeng(曾红丽)[1,2,†] and Erik Aurell[3,4,‡]

[1] School of Science, New Energy Technology Engineering Laboratory of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2] Nordita, Royal Institute of Technology, and Stockholm University, SE-10691 Stockholm, Sweden
[3] KTH – Royal Institute of Technology, AlbaNova University Center, SE-10691 Stockholm, Sweden
[4] Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, 30-348 Kraków, Poland

(Received 9 March 2020; revised manuscript received 9 March 2020; accepted manuscript online 28 April 2020)

As a problem in data science the inverse Ising (or Potts) problem is to infer the parameters of a Gibbs–Boltzmann distributions of an Ising (or Potts) model from samples drawn from that distribution. The algorithmic and computational interest stems from the fact that this inference task cannot be carried out efficiently by the maximum likelihood criterion, since the normalizing constant of the distribution (the partition function) cannot be calculated exactly and efficiently. The practical interest on the other hand flows from several outstanding applications, of which the most well known has been predicting spatial contacts in protein structures from tables of homologous protein sequences. Most applications to date have been to data that has been produced by a dynamical process which, as far as it is known, cannot be expected to satisfy detailed balance. There is therefore no a priori reason to expect the distribution to be of the Gibbs–Boltzmann type, and no a priori reason to expect that inverse Ising (or Potts) techniques should yield useful information. In this review we discuss two types of problems where progress nevertheless can be made. We find that depending on model parameters there are phases where, in fact, the distribution is close to Gibbs–Boltzmann distribution, a non-equilibrium nature of the under-lying dynamics notwithstanding. We also discuss the relation between inferred Ising model parameters and parameters of the underlying dynamics.

Keywords: inverse Ising problem, kinetic Ising model, statistical genetics, fitness reconstruction

PACS: 02.50.Tt, 05.40.–a, 05.45.Tp, 05.90.+m **DOI:** 10.1088/1674-1056/ab8da6

1. Introduction

The Gibbs–Boltzmann distribution of the Ising model on L spin (For later reference we prefer to refer to the number of spins in the model with the letter L, for "loci". The more customary letter N will later be reserved to the number of samples drawn from the distribution, following a convention using in statistics) is

$$P(\boldsymbol{s}) = \frac{\exp\left(-\beta\left(\sum_i \theta_i s_i + \sum_{i<j} J_{ij} s_i s_j\right)\right)}{Z}, \tag{1}$$

where β is the inverse temperature, and Z is the partition function, defined as

$$Z = \sum_{\boldsymbol{s}} \exp\left(-\beta\left(\sum_i \theta_i s_i + \sum_{i<j} J_{ij} s_i s_j\right)\right). \tag{2}$$

The parameters of the model are L external fields $\{\theta_i\}_{i=1}^{L}$ and $\frac{L(L-1)}{2}$ coupling constants or interactions $\{J_{ij}\}_{i<j}$. The Gibbs–Boltzmann distribution of a Potts model is defined in a similar way, except that each variable can take q values ($q = 2$ for the Ising model) and the model parameters are vectors and matrices (by reparametrization invariance the number of independent paramaters is respectively $q-1$ for the vector and $(q-1)^2$ for the matrix, which for $q = 2$ gives only one parameter of each type as in Eq. (1)) ($\{\theta_i^{(\alpha)}\}$ for $1 \le \alpha \le q$ and $\{J_{ij}^{(\alpha,\alpha')}\}$ for $1 \le \alpha, \alpha' \le q$).

From the viewpoint of physics, Eq. (1) gives the equilibrium distribution at inverse temperature β corresponding to the Ising energy function (or Hamiltonian).[1–3] The traditional Ising problem of statistical mechanics is to determine properties of the distribution $P(\boldsymbol{s})$ from the model parameters $\{\theta_i, J_{ij}\}$. The probability distribution $P(\boldsymbol{s})$, or ensemble, will be reflected in samples drawn independently from that distribution. Combining the two steps of estimating the ensemble and sampling from the distribution, the direct Ising problem can be defined as the problem of estimating an empirical probability distribution over samples from model parameters. The inverse Ising problem is then the opposite problem of inferring model parameters from samples drawn from the distribution.[4–6]

To stress the inverse nature of the problem it is useful to

*Project supported partially by the National Natural Science Foundation of China (Grant No. 11705097), the Natural Science Foundation of Jiangsu Province of China (Grant No. BK20170895), the Jiangsu Government Scholarship for Overseas Studies of 2018 and Scientific Research Foundation of Nanjing University of Posts and Telecommunications, China (Grant No. NY217013), and the Foundation for Polish Science through TEAM-NET Project (Grant No. POIR.04.04.00-00-17C1/18-00).

†Corresponding author. E-mail: hlzeng@njupt.edu.cn
‡Corresponding author. E-mail: eaurell@kth.se

© 2020 Chinese Physical Society and IOP Publishing Ltd

introduce some notation from statistics. The class of distributions (1), with values of the external fields and interactions in some set, is called an exponential family (exponential because the parameters all appear in the exponent, and family because a set of parameters are considered). The inverse Ising problem is accordingly called parameter inference in an exponential family.[7] The most basic way to infer parameters from independent samples from one and the same probability distribution is maximum likelihood (ML). For computational reasons, ML is often formulated in logarithmic coordinates as maximum log-likelihood. Given N independent samples from Eq. (1), the maximum log-likelihood amounts to the convex optimization problem,

$$\{\theta_i^*, J_{ij}^*\}^{\mathrm{ML}} = \operatorname{argmax}\Bigg[-\sum_i \theta_i \langle s_i\rangle - \sum_{i<j} J_{ij}\langle s_i s_j\rangle - \frac{1}{\beta}\log Z\Bigg], \tag{3}$$

where $\langle s_i\rangle$ and $\langle s_i s_j\rangle$ are the empirical averages computed from the samples. The star on the parameters on the left-hand side mark that these are inferred, and the superscript ML indicates the inference method. Searching the only reason for Eq. (3) is a difficult task because the forward problem of computing Z from the parameters is difficult. The effect of the parameter β cannot be separated from an overall scale of $\{\theta_i^*\}$ and $\{J_{ij}^*\}$, and therefore only appears in Eq. (3) as a proportionality of the log-partition function $\log Z\left(\beta, \{\theta_i^*\}, \{J_{ij}^*\}\right)$. From now on we will, when not specified otherwise, set β equal to 1.

A fundamental fact of statistical inference, which holds for all exponential families, is that maximum likelihood does not need all the data. Indeed, in Eq. (3) data only appear as empirical averages. That is, if we have a table of N independent samples this means NL data items, but Eq. (3) only depends on $\frac{L(L+1)}{2}$ numbers computed from the data. Those numbers (here means and correlations) are called sufficient statistics for inference in an exponential family.[8,9] A second fundamental fact is that maximum likelihood inference gives the same result as maximizing Shannon entropy conditioned by the sufficient statistics. From the physical point of view this follows directly from Eq. (1) being an equilibrium distribution, which minimizes free energy. Maximizing Shannon entropy conditioned by some chosen set of empirical averages is called the maximum-entropy[10–12] or max-entropy approach to statistical inference. By the above such a set of empirical averages is in one-to-one relation with a set of parameters in an exponential family for which they are sufficient statistics. This relation between exponential parameters and empirical averages is called conjugacy, or, in Information Geometry,[13,14] a duality. The max-entropy approach with a given set of empirical averages is equivalent to maximum likelihood inference in an exponential family with the conjugate parameters.

In physics, Eq. (1) appears as a (canonical) equilibrium distribution of a system interacting with a heat bath. Let two configurations of the system be $\boldsymbol{s}$ and $\boldsymbol{s}'$, and let the probability of the system to make the change from $\boldsymbol{s}$ to $\boldsymbol{s}'$ per unit time be $W_{\boldsymbol{s},\boldsymbol{s}'}$. Then equilibrium will be reached if the transition rates satisfy the detailed balance conditions[15]

$$P(\boldsymbol{s})W_{\boldsymbol{s},\boldsymbol{s}'} = P(\boldsymbol{s}')W_{\boldsymbol{s}',\boldsymbol{s}}. \tag{4}$$

In equilibrium, transitions from $\boldsymbol{s}$ to $\boldsymbol{s}'$ and $\boldsymbol{s}$ to $\boldsymbol{s}'$ are equally likely. As a consequence there cannot be chains of states such that cyclic transitions in one direction ($\boldsymbol{s}^1 \to \boldsymbol{s}^2 \to \cdots \to \boldsymbol{s}^k \to \boldsymbol{s}^1$) is more likely than in the opposite direction ($\boldsymbol{s}^1 \to \boldsymbol{s}^k \to \cdots \to \boldsymbol{s}^2 \to \boldsymbol{s}^1$). Chemistry and Biology have many examples of such cycles appear, from chemical oscillations of the Belouzov–Zhabotinsky type to the cell cycle and circadian rythms.[16,17] This immediately says that not all dynamics on discrete state spaces can satisfy detailed balance, and so cannot be expected to have stationary distributions like Eq. (1).

If we focus on single-spin flips and $P(\boldsymbol{s})$ in Eq. (1) we can write the detailed balance conditions as a relation between spin flip rates $r_i(+, \boldsymbol{s}_{\setminus i})$ and $r_i(-, \boldsymbol{s}_{\setminus i})$,

$$r_i(-, \boldsymbol{s}_{\setminus i}) = r_i(+, \boldsymbol{s}_{\setminus i}) e^{-2\beta\theta_i - 2\beta\sum_j J_{ij}s_j}, \tag{5}$$

where $r_i(-)$ is the rate of spin i to flip from down to up, and $r_i(+)$ is the rate from up to down. Both of them depend on the configurations of all the other spins, written as $\boldsymbol{s}_{\setminus i}$. Alternatively we can rewrite Eq. (5) as

$$r_i(\boldsymbol{s}) = \gamma_i(\boldsymbol{s}_{\setminus i}) e^{-\beta\Delta_i E(\boldsymbol{s})}, \tag{6}$$

where $r_i(\boldsymbol{s})$ is the rate of flipping spin i in configuration $\boldsymbol{s}$, $\Delta_i E(\boldsymbol{s})$ is the energy change when doing so, and $\gamma_i(\boldsymbol{s}_{\setminus i})$ is an overall rate which does not depend on the value of spin i. Different Monte Carlo procedures (or Markov chain Monte Carlo (MCMC) algorithms) differ by this overall rate $\gamma_i(\boldsymbol{s}_{\setminus i})$.

To give an example of a spin-flip dynamics which does not satisfy detailed balance we point to the class of focused algorithms for constraint satisfaction problems, invented by Papadimitriou now three decades ago.[18–26] In such algorithms one imagines that the energy function is a sum of local terms all of which are one or zero. A solution is a configuration where all the energy terms are zero (zero-energy ground state). A focused algorithm is one where the rate of flipping spin i is zero unless at least one of the constraints depending on i is unsatisfied, but otherwise the dynamics remains partly random. It is clear that for such dynamics one can flip into a satisfied state, but once there the dynamics stops; one cannot flip out (the first condition of focusing can be satisfied in the equilibrium algorithm (6) by taking β to infinity (zero temperature); then the algorithm is a deterministic greedy search, and is no longer random). It is well known that focused algorithms such

as "walksat" outperform equilibrium algorithms in many important applications.[19,24]

Let us now go back to the problem of inferring the parameters of the Ising model in Eq. (1) where the data has been generated by some process which may or may not satisfy detailed balance. The inference procedure is at this point treated as a black-box. What does this mean? Does it even make sense? When does it make sense?

In equilibrium statistical mechanics the answer is clear and simple: the process will make sense if the data is generated by a process in detailed balance with an energy function in the same exponential family, and in a phase where sampling is possible. The first condition simply means that if the data is generated from a process with, say, third-order interactions between the spins, those interactions will not be recovered from inferring only first-order and second-order interactions. The second conditions means that parameters have to be such that the dynamics explores enough configurations that there is enough information to infer from. A trivial example when this is not the case is zero temperature where the configuration goes to a local minimum of the energy, and then does not change. A more subtle example is a spin glass phase where for large but not infinite β only part of the Gibbs distribution (1) will be sampled by an MCMC algorithm unless the simulation time is exponentially large in system size.[29] Inference from naturally generated samples, which are "stuck in one valley", has long been known to be impossible by the class of inverse Ising methods surveyed here.[27] For specific problems and with more tailored methods such a task is sometimes nevertheless possible.[28] Inference from samples that are drawn uniformly from such a distribution has on the other hand been shown to be possible, and even easy.[29] Such uniform samples however have to be generated by methods that either needs a large computational effort (long simulation time), or one needs to restart the simulation many times with new random initial values, which corresponds to real data from many separate time series.

Once we step out of the realm of equilibrium dynamics we are much more in the dark. For the specific example of symmetric simple exclusion process (SSEP) it is known that the stationary distribution, i.e. the equivalent of Eq. (1), contains all interactions of all orders,[30,31] meaning all single-spin and pair-wise terms as in Eq. (1), all three-spin interactions, and so on. This is so even though the SSEP dynamics is entirely specified by nearest-neighbor pairwise exclusion, and the non-equilibrium aspects are only the boundary conditions, particle exchanges with reservoirs. When the dynamics can be described as depending on energy changes with some non-equilibrium element such as focusing at every step ("bulk driven non-equilibrium process"), the possibilities for the stationary distributions are wider still. The outcome of an inverse Ising procedure applied to such data may therefore be completely unrelated to the parameters of the mechanisms that gave rise to the data. The computational complexity and number of data required to infer the parameters of any kind of non-equilibrium steady state from snapshots has been shown to be daunting.[32–34] Nevertheless, this is the setting of most successful and interesting applications of inverse Ising techniques to date.[35,36] Why is this?

In this review we will present two cases where the above problem can be analyzed and/or studied in simulations. The first case is kinetic Ising models with possibly different values of pairwise parameters J_{ij} and J_{ji}. When $J_{ij} = J_{ji}$ (symmetric kinetic Ising models) this is nothing by a Monte Carlo procedure to compute the distribution $P(\boldsymbol{s})$ in Eq. (1). Models where $J_{ij} \neq J_{ji}$ (asymmetric kinetic Ising models) have however also been widely studied, e.g. as model systems in neuroscience.[27,37,38] The kinetic Ising models hence interpolate between equilibrium and non-equilibrium systems. They also illustrate that more efficient inference procedures than inverse Ising are available if one can use a time series and not only independent samples from a stationary distribution.

The second case are slightly more involved spin dynamics that model evolution under mutations, Darwinian selection (fitness), finite-N effects (genetic drift) and recombination (sex). We will here see that inverse Ising works in certain ranges of parameters describing the relative strengths of mutations, fitness and sex, but not in others. We will also see that the relation is not trivial; non-trivial theory is needed to translate the results from inverse Ising to inferred fitness that can be compared to model parameters.

This review is organized as follows. In Section 2, we summarize for completeness some inverse Ising techniques. This topic is already covered by excellent reviews to which we refer for more details and a wider palette of techniques. In Section 3 we introduce the kinetic Ising problem in its symmetric and asymmetric form, and present characteristic results, and in Section 4, we present two applications of those techniques taken from earlier work by one of us (HLZ). Section 5 presents on the other hand a class of problems in population genetics, and Section 6 contains an outlook and discussion.

2. Techniques for inverse Ising

The inverse Ising problem has been studied under several different names, such as statistical inference in exponential families (as above), Boltzmann machines, maximum-entropy modeling, direct coupling analysis (DCA), logistic regression techniques, and so on. For small enough system (small enough L) maximum likelihood in Eq. (3) is computationally feasible, for instance by the iterative method also known as Boltzmann machine.[39] The idea is that very widely used method is to adjust the parameters in the exponential family to make empiri-

cal averages and ensemble averages of the conjugate sufficient statistics agree.

For large L, maximum likelihood (ML) is not computationally efficient, meaning that it requires an effort exponentially increasing in L. In other words, for a given fixed L, if ML is computationally feasible depends on time and the development of computer hardware. Nevertheless, for many applications that have been of interest, either ML has not been feasible, or other inference schemes have given comparable results with less effort. In any case, it has been an interesting theoretical challenge to design and analyze schemes that make a different trade-off between accuracy and computational speed than ML.

The state of the art of inverse Ising was recently extensively reviewed in Ref. [6], and we will here only provide a background for the later sections. A first type of inference methods attempts to circumvent the computational challenge of ML by estimating the partition function Z efficiently. Such methods are collectively known as mean-field inference, because they rely on mean-field techniques. The by far most common version of mean-field inference relies on a variational ansatz in terms of magnetizations, which yields the physical mean-field equations of the Ising model,

$$m_i = \tanh\left(h_i + \sum_j J_{ij} m_j\right). \tag{7}$$

In this equation only m_i is taken from the data, and there are only L equations. By using also linear-response

$$c_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle = \frac{\partial m_i}{\partial h_j} \tag{8}$$

one finds the naive mean-field inference formula[40]

$$J_{ij}^{*,\mathrm{nMF}} = -\left(c^{-1}\right)_{ij}. \tag{9}$$

The above expression is computationally quite convenient as it reduces a complicated inference to matrix inversion. One may note that Eq. (9) is the same formula as inferring the interaction matrix of a Gaussian model (precision matrix in information theory) from data. It is an elementary property of multidimensional centered Gaussian distributions that they can be written as $P(x) = \frac{1}{N}\exp\left(-\frac{1}{2}xC^{-1}x\right)$, where C is the covariance matrix. The precision matrix (the model parameters) can therefore be inferred as the inverse matrix of C (the data). The difference is that for an Ising model, Eq. (9) is only approximate, and is not always with good accuracy; for the SK model (to be discussed below) it holds for instance at high-temperature (weak interactions), but not at low temperature. If needed one can combine Eqs. (7) and (9) to estimate also the external fields, i.e.,

$$h_i^{*,\mathrm{nMF}} = \tanh^{-1} m_i - \sum_j J_{ij}^{*,\mathrm{nMF}} m_j. \tag{10}$$

More advanced mean-field methods than naive mean-field are obtained by starting from more advanced approximations than Eq. (7). The best-known of these is TAP (Thouless–Anderson–Palmer)[41] which starts from

$$m_i = \tanh\left(h_i + \sum_j J_{ij} m_j - m_i \sum_j J_{ij}^2 (1 - m_j^2)\right). \tag{11}$$

Using linear response then gives J_{ij}^* as the solution of a quadratic equation

$$J_{ij}^{*,\mathrm{TAP}} + 2 m_i m_j \left(J_{ij}^{*,\mathrm{TAP}}\right)^2 = -\left(c^{-1}\right)_{ij}. \tag{12}$$

A general feature of inference methods of this type is that in the variational ansatz the data is only taken into account through the single-variables marginals, i.e. through the magnetizations. It is only linear-response Eq. (8), which is an exact property of the full Ising model, but not of the variational ansatz, that two-variable marginal are brought back into play.

Another type of mean-field inference equation attempts to find the Ising model which best fits the data. The variational parameters are then magnetizations (m_i) and correlations (c_{ij}), conjugate to model parameters h_i and J_{ij}. This approach was first developed as an iterative procedure called "susceptibility propagation"[42,43] and only later shown to also yield equations like Eqs. (9) and (12), where ratios of hyperbolic functions appear on the left-hand side, but the right-hand side is still just the inverse matrix of correlations.[44] An alternative derivation of this elegant approach can be found in Ref. [6], which also contains a survey of many more methods that have been introduced and tested in the literature.

A different type of inference gives up on the ambition to approximate the partition function, and hence the full probability distribution $P(\boldsymbol{s})$. Instead, one tries to infer the parameters from some other property which can be efficiently computed. The most widely used such method is maximum pseudo-likelihood[45] or pseudo-likelihood maximization (PLM). This starts from the conditional probability of the Ising model

$$P(s_i | \boldsymbol{s}_{\setminus i}) = \frac{\exp\left(-\beta\left(\theta_i s_i + \sum_j J_{ij} s_i s_j\right)\right)}{\sum_{s'=\pm 1}\exp\left(-\beta\left(\theta_i s' + \sum_j J_{ij} s' s_j\right)\right)}. \tag{13}$$

In contrast to Eq. (1) there is now no longer any difficult to compute normalization factor. The denominator in Eq. (13) is the normalization of a distribution over only one Ising spin, and hence has only two terms. When treated in the same way as ML Eq. (3), Eq. (13) leads to L inference problems, one for each spin i

$$\left(\theta_i^*, J_{ij}^*\right)^{\mathrm{PML},i} = \mathrm{argmax}\left[-\theta_i \langle s_i \rangle - \sum_j J_{ij} \langle s_i s_j \rangle - \frac{1}{\beta}\langle \log \zeta_i \rangle\right], \tag{14}$$

where ζ_i is the sum in the denominator in Eq. (13). The left-hand side emphasizes that this is inference "as seen from spin i" (by maximizing conditional probability of spin i). To get the final answer one needs to combine $J_{ij}^{*,\mathrm{PML},i}$ and $J_{ij}^{*,\mathrm{PML},j}$, typically by taking their average.

In the limit of infinite data, PLM will almost surely find the same parameters as ML, a property referred to as statistical consistency (The formal definition of statistical consistency is that as the number of samples goes to infinity, the argmin of the estimator converges in probability to the right answer. this holds for ML and PLM and some other inference methods to be discussed below, but does not hold for mean-field inference methods. In the limit of infinite data the sample averages used in mean-field will always surely be the same as ensemble averages, but the recovered parameters will not be the true ones because physical mean-field is in itself approximate. For a discussion, see e.g.[6] and references cited therein.). In applications PLM has often been found to outperform both naive and advanced mean-field inference.[6] Why that is so cannot be said to be completely known, since the number of samples in real data sets is finite. The error of mean-field inference compared to PLM in the infinite sample limit (lack of statistical consistency) could therefore be compensated by the error in PLM when used on a finite number of samples. Empirically this has mostly not been found to be the case, but that may partially be a consequence of the kinds of data sets that have been considered in the literature.

2.1. Undersampling, regularization, prior information and evaluation criteria

High-dimensional statistics is the branch of modern statistics where the number of samples (N) is assumed to grow together with or slower than the number of parameters (here $\frac{L(L+1)}{2}$). Common sense says that if there are fewer samples than parameters and no other information, then the parameters cannot be fully determined by the data. This rule-of-thumb has to be applied with care, because often there is other information, used explicitly or implicitly; we will refer to a few such cases below.

Nevertheless, the rule-of-thumb points to something important, namely that in the important application of inverse Potts methods to contact prediction in protein structures,[46,47] the number of parameters (for 20 types of amino acids in a protein of 100 residues) is typically about $20^2 \cdot 100^2$, which is four million, while the number of samples is rarely more than a hundred thousand. All inference methods outlined above are therefore in this application used in regimes where they are under-sampled, and so need to be regularized. For naive mean-field inference a regularization by pseudo-counts (adding fictitious uniformly distributed samples) was used in Refs. [46,47], while an L_1-regularization was used in Ref. [48], and an L_2-regularization in Ref. [49]. For PLM similarly L_2-regularization was used in Refs. [50–52].

An important aspect of all inference is what is the family from which one tries to infer parameters. This can be given in a Bayesian interpretation as an *a priori* distribution of parameters; the more one knows in that direction, the better the inference can be. Many regularizers can be seen as logarithms of Bayesian prior distributions such that the analogy also works the other way: regularized inference is equivalent to inference with a prior (exponential of the regularizer), and can therefore work better because it uses more information. For instance, if all parameters are supposed to be either zero or bounded away from zero by some lower threshold value, and if the ones that are non-zero are sparse, then the authors of Ref. [53] showed that L_1-regularized PLM can find the graph structure using relatively few samples, given certain assumption that were later shown to be restrictive.[54] Nevertheless, using and analyzing thresholding also in the retained predictions, the authors of Ref. [55] were able to show that L_1-regularized PLM can indeed find the graph structure using order of $\log L$ samples (the same authors also showed that L_1-regularized PLM with thresholding, as used in the plmDCA software of Ref. [52] can recover parameters in L_2 norm using order of $\log L$ samples).

A second and equally important aspect is the evaluation criteria. The criterion in Ref. [53] is graphical: the objective is to infer properties of the model (the non-zero interactions) which can be represented as a graph. It is obvious that inference under this criterion will be difficult without a gap in the distribution of interaction parameters away from zero. Information theory imposes limits on the smallest couplings that can be retrieved from the finite amount of data;[55,56] given finite data it is simply not possible to distinguish a parameter which is strictly zero from one which is only very small. Another type of criterion is metrical, most often the squared differences of the actual and inferred parameter values.[6,27] Yet another is probabilistic by determining some difference between the two probability distributions as in Eq. (1), one with the actual parameters and one with the inferred parameters. Two examples of probabilistic criteria are Kullback–Leibler divergence and variational distance. An advantage of probabilistic criteria is that they focus on typical differences of samples, and not on parameter differences which may (sometimes) not matter so much as to the samples observed. However, as this requires sampling from the distributions, it is also a disadvantage.

Important results have been obtained as to how many samples are required for successful inference when both the a priori distributions and and the criteria are varied, first in Ref. [53] under strong assumptions, and more recently in Refs. [55,57]. These two latter papers also introduced a different objective function, interaction screening objective (ISO),

that has dependence on the same local quantities as pseudo-likelihood, and which provably outperforms PLM in terms of expected error for the given number of samples, providing near sample-optimal guarantee. ISO has also more recently been generalized to learn Ising models in the presence of samples corrupted by independent noise,[58] and to the case of Potts models and beyond pairwise interactions.[59]

In practice and in many successful applications to real data, criteria have been of the type "correctly recovering k largest interactions", colloquially known as "top-k". Performance under such criteria is straight-forward to analyze empirically when there is a known answer; one simply compiles two lists of k largest parameters and what interactions they refer to, and then compares the two lists. For instance, one can check what fraction of k largest inferred interactions can also be found among the k largest actual interactions, which is known as k-true positive rate, or TPR(k). In the application of inverse Potts methods to contact prediction in protein structures,[46,47] k has commonly been taken to be around 100. The inequality that the number of retained parameters is less than the number of samples has hence been respected, with a large margin. The theoretical analysis of performance under this type of criterion is however more involved, as the distribution of the largest values of a random background is an extreme deviations problem. One approach is to leverage an L_∞ norm guarantee,[55,59] for another using large deviation theory, see Refs. [60,61].

2.2. Time series and alltime inverse Ising techniques

In Section 3 we will consider inference from data generated by a kinetic Ising model, and in Section 4 we will consider applications of this technique to data in Neuroscience and from Finance. The main message of these sections will be that if you have time series data, it will be usually better to do inference on the time-labeled data. As we will show, even when the dynamics is of the type (6), respects detailed balance, and has stationary distribution (1), it can be faster and easier to infer J_{ij} from the dynamical law than by inverse Ising techniques.

Nevertheless, even if the data was generated in a dynamic process, we do not always have time series data. In Sections 5.1 and 5.4 we will consider models of evolution, intended as stylized descriptions of the kind of genetic/protein data on which inverse Ising (Potts) techniques have been applied successfully.[36,46,47,62] The underlying dynamics is then of the type of $N_{\rm tot}$ individuals (genomes/genetic signatures/proteins) of size (genomic length) L evolving for a time T, while the data is on N individuals (genomes/genetic signatures/proteins) sampled at one time (or at uneven times so that the time information is hard to use, or the time at which they were sampled is unknown, the cases may differ depending on the data set).

Averages at any given time will have errors which go down as $(N_{\rm tot})^{-\frac{1}{2}}$, typically a very small number for real data sets, but not necessarily very small in a simulation. For the evaluation of how simulations match theory it is therefore of interest to also consider as input data to inverse Ising variants of naive mean-field Eq. (9) and PLM Eq. (14), where the averages are computed both over samples and over time. We refer to these variants as alltime versions of the respective algorithms.

3. A model: kinetic asynchronous Ising dynamics

A standard approach to sample the equilibrium Ising model is Glauber dynamics.[63,64] On the level of probability distributions it is formulated as master equations

$$\frac{\mathrm{d}}{\mathrm{d}t}p(s_1,\ldots,s_L;t) = \sum_i \omega_i(-s_i)p(s_1,\ldots,-s_i,\ldots,s_L;t) - \sum_i \omega_i(s_i)p(\boldsymbol{s};t), \tag{15}$$

where $\omega_i(s_i)$ is the flipping rate, i.e., the probability for the state of ith spin to changes from s_i to $-s_i$ per unit time while the other spins are momentarily unchanged. Equation (15) shows that the configuration $s_1,\ldots,s_L$ is destroyed by a flip of any spin s_i (a loss term), but it can also be created by the flip from any configuration with the form $s_1,\ldots-s_i,\ldots,s_L$ (a gain term). The flipping rate of spin i is

$$\omega_i(\boldsymbol{s}) = \frac{\gamma}{1+\exp\left[2s_i\left(\theta_i+\sum_j J_{ij}s_j\right)\right]} = \frac{\gamma}{2}\left[1-s_i\tanh\left(\theta_i+\sum_j J_{ij}s_j\right)\right]. \tag{16}$$

The parameter γ is an overall rate which in Glauber dynamics is assumed to be the same for all spins. The left-hand side depends on the whole configuration $\boldsymbol{s}$ because the values of all spins enter on the right-hand side. The inverse temperature β is here set to be 1; as noted above it can be absorbed in the parameters.

For small enough systems (small L), Eq. (15) can be simulated by solving 2^L linear ordinary differential equations. For larger L, Eq. (15) can only be simulated by Monte Carlo procedure. This means that one considers N separate spin configurations $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N$, each of which is evolved in time. The empirical probability distribution

$$P^e(\boldsymbol{s},t) = \frac{1}{N}\sum_{s=1}^{N} 1_{\boldsymbol{s},\boldsymbol{s}_s(t)} \tag{17}$$

is then an approximation of $P(\boldsymbol{s},t)$ in Eq. (15). We note (trivially) that for large systems, $P^e(\boldsymbol{s},t)$ will typically be either zero or $\frac{1}{N}$; the chance that among N separate spin configurations $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N$ two are exactly equal will be very small. $P^e(\boldsymbol{s},t)$ hence approximates $P(\boldsymbol{s},t)$ as to certain summary

statistics such as single-spin averages (magnetizations), but typically cannot approximate $P(\boldsymbol{s},t)$ very well as to the values for individual configurations.

For simplicity of presentation we will here focus on the time-homogeneous case in which all parameters are time independent. Distributions will then eventually relax to a stationary state, and we will assume that this process has taken place. Inference can then by carried out by treating samples at different times independently, i.e., by the type of alltime algorithms discussed in Section 2.2. For the rest of this section, N (the number of different time series) will hence be 1. Indeed, as in the Monte Carlo procedure the different samples do not interact, one can limit oneself to just one time series, as long as one is interested in properties of the statistically stationary state reached at large times.

The dynamics of a configuration $\boldsymbol{s}(t)$ is governed by the same rates as in Eq. (15). In the Monte Carlo simulation scheme it is convenient to consider spin i as responding to an effective field from the external field θ_i and the interactions from all the other spins. This effective field is time-dependent because the configurations of the other spins change in time, viz.

$$H_i(t) = \sum_j J_{ij} s_j(t) + \theta_i, \tag{18}$$

and the instantaneous rates are then

$$\omega_i(\boldsymbol{s},t) = \frac{\gamma}{2}\left[1 - s_i(t)\tanh\left(H_i(t)\right)\right]. \tag{19}$$

One approach to simulation is to introduce a small time step increment δt and to flip each spin at each time with probability $\omega_i(\boldsymbol{s},t)$. For this scheme to simulate Eq. (15) one must take δt so small that the chance of any other spin to flip in the same short time interval is negligible. This scheme can be said to rely on $L\cdot t/\delta t$ random variables, one for the decision whether or not to flip each spin in each time interval. Since on average less than one spin will flip in each time interval the probabilities of these variables have to be very biased towards not flipping.

A computationally more efficient scheme is to first consider the rate of the event of flipping any spin. That is,

$$\omega_{\mathrm{TOT}}(\boldsymbol{s},t) = \sum_i \omega_i(\boldsymbol{s},t). \tag{20}$$

As long as no spin flips this overall rate does not change. The waiting time until any spin has flipped is therefore an exponentially distributed random variable with rate ω_{TOT}, and the chance that it was spin i that flipped is $\omega_i/\omega_{\mathrm{TOT}}$. The dynamics can then be simulated in discrete steps starting from a configuration $\boldsymbol{s}_0$ at t_0 such that flips take place at times $t_1, t_2, \ldots$ Initially the rates are $\{\omega_i(\boldsymbol{s}_0)\}$, and $t_1 - t_0$ is an exponentially distributed random variable with rate $\omega_{\mathrm{TOT}}(\boldsymbol{s}_0) = \sum_i \omega_i(\boldsymbol{s}_0)$. The first spin to flip will be the j-th spin with probability $\omega_j(\boldsymbol{s}_0)/\omega_{\mathrm{TOT}}(\boldsymbol{s}_0)$, and after the flip all rates are updated to $\{\omega_i(\boldsymbol{s}_1)\}$, and the process is repeated. This algorithm is called the Gillespie algorithm,[65] and relies on $L\cdot t/\overline{\Delta t}$ random variables where $\overline{\Delta t}$ is some characteristic time interval between the flips. At the price of a slightly more complicated structure it is thus faster than the first algorithm by a ratio $\overline{\Delta t}/\delta t$. Furthermore this method is exact; $\overline{\Delta t}$ is a property of the dynamics and not of the simulation scheme.

A third approach is to update at each step a spin i picked uniformly at random with probability $\gamma\delta t$. After such an update, which may or may not change the spin value, the new value will be

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

From this we can evaluate the rate of flipping of spin i per unit time to be

$$\begin{cases} \gamma/\{1+\exp[2\beta H_i(t)]\}, & \text{when } \ s_i(t) = 1, \\ \gamma/\{1+\exp[-2\beta H_i(t)]\}, & \text{when } \ s_i(t) = -1. \end{cases}$$

which gives the same rate as in Eq. (16). Since two random numbers are called for each spin at each time interval, this scheme can be said to rely on $2L\cdot t/\delta t$ random variables.

3.1. Symmetric and asymmetric Sherrington–Kirkpatrick (SK) models

As illustrative examples we will now look at symmetric and asymmetric SK models,[66] which are defined as follows. First we introduce J_{ij} with no restriction on i and j. Such a matrix can be split into its symmetric and asymmetric parts. We write

$$J_{ij} = J^s_{ij} + kJ^{as}_{ij}, \quad k \geq 0, \tag{21}$$

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

$$\begin{aligned} J^s_{ij} &= J^s_{ji}, \\ J^{as}_{ij} &= -J^{as}_{ji}. \end{aligned} \tag{22}$$

The parameter k in Eq. (21) measures the asymmetric degree of the interactions J_{ij}. With $k = 0$, J_{ij}'s are a fully symmetric model the stationary distribution of which is Eq. (1). Any $k \neq 0$ means that the J_{ij} and J_{ji} are not the same, and we have a non-equilibrium dynamics. The SK kinetic model, extended to non-equilibrium,[67] means to take both the symmetric and the asymmetric couplings to be identically and independently Gaussian distributed random variables with means zero and variances

$$\langle {J^s_{ij}}^2 \rangle = \langle {J^{as}_{ij}}^2 \rangle = \frac{g^2}{N}\frac{1}{1+k^2}. \tag{23}$$

This parametrization is chosen such that the total coupling matrix J follows a Gaussian distribution

$$p(J_{ij}) \propto \exp\left(-\frac{(J_{ij}-\mu)^2}{2\sigma^2}\right) \tag{24}$$

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

The interactions J_{ij} define spin update rates (16) or (19). To see that asymmetric interactions do not lead to Gibbs distributions (1), it is useful to temporarily change the parametrization so that there are three only non-zero interactions $J_{ij} = J_{jk} = J_{ki} = J$, all large. All other J_{ij} are zero, and all θ_i are also zero. Assume that initially the three spins s_i, s_j and s_k are all up, i.e., $+++$. They will then have the same (small) flip rate $\gamma/\left(1+e^{J}\right)$, and one of them will flip first, let that be spin i, so that the next state is $-++$. After this has happened the (much larger) rate for either i to flip back or for k to flip will be $\gamma/\left(1+e^{-J}\right)$. A flip of spin i will hence almost surely either go back to the starting state $+++$ after two flips, or lead to the configuration $-+-$. This second state will in turn almost surely lead to $++-$ or $---$. The first of these is a shift of the state after the first flip to the left, and by circular permutation symmetry it must be more likely that the shifts continue in that direction rather than to the right. The second is on the other hand obviously the mirror image of the starting state, and all rates are again low. Flipping out of $---$ would lead to $+--$, which would give $+-+$ and then $+++$ or $--+$, which is also a shift to the left. A dynamics which has some similarities to the above where motion surely goes only in one direction is the basis of Dijkstra's famous self-stabilizing system under distributed control,[68] for a physics perspective, see Ref. [69].

3.2. Inference for asynchronous Ising models

Many techniques for inverse Ising as discussed above in Section 2 have been applied to data from asynchronous Ising (or similar) dynamics, mainly for neuroscience applications.[4,70–72] Since our purpose here is to compare to inference using a time series we will for the equilibrium case just consider the simplest method, which is naive mean-field (nMF) (9). On the methodological side, much work has been performed on applying inverse Ising techniques to synchronous versions of Ising dynamics.[73–76] This work will not be covered here. Dynamic mean-field inference as used below was originally developed for synchronous updates in Ref. [77], also see Ref. [78]. Inference in more realistic (and more complex) models from neuroscience has also been carried out, but is beyond the scope of this review, see Refs. [72,79,80].

3.3. Mean-field inference

We now derive versions of nMF and TAP inference for asynchronously updated kinetic models following.[81]

For kinetic Ising model with Glauber dynamics, the state of spin i is time dependent $s_i(t)$, thus the time-dependent means and correlations are naturally defined as

$$\begin{aligned} &m_i(t) = \langle s_i(t)\rangle\,, \\ &c_{ij}(t_0+\tau,t_0) = \left\langle s_i(\tau+t_0)s_j(t_0)\right\rangle - m_i(\tau+t_0)m_j(t_0). \end{aligned} \tag{25}$$

Then, with the master equation (15) and the flipping rate (16), we have the equations of motion for means and correlations as

$$\frac{\mathrm{d}m_i(t)}{\mathrm{d}t} = -m_i(t) + \langle\tanh\left[H_i(t)\right]\rangle, \tag{26}$$

$$\frac{\mathrm{d}\langle s_i(t)s_j(t_0)\rangle}{\mathrm{d}t} = -\langle s_i(t)s_j(t_0)\rangle + \langle\tanh\left[H_i(t)s_j(t_0)\right]\rangle. \tag{27}$$

In the forward problem of statistical physics we would here have the closure problem: the left-hand side is the time derivative of an average while the right-hand side contains terms of an average of a higher order. In the inverse problem we start by observing that the term on the left-hand side and the first terms on the right-hand side of Eqs. (26) and (27) can be taken from data. The second term on the right-hand side contains averages of the tanh function and involves all kinds of higher-order correlations. The equations thus have to be closed with respect to these terms, but in a slightly different way in the forward problem.

We introduce the notation

$$b_i = \theta_i + \sum_j J_{ij}m_j \tag{28}$$

for the non-fluctuating part of the argument of the tanh and rewrite $H_i(t) = \theta_i + \sum_j J_{ij}s_j(t)$ as

$$H_i \equiv b_i + \sum_j J_{ij}\delta s_j(t), \tag{29}$$

where the sum depends on the fluctuating term $\delta s_i(t) = s_i(t) - m_i$. In lowest order we neglect fluctuations in altogether and close the equation for magnetizations as

$$\frac{\mathrm{d}m_i(t)}{\mathrm{d}t} = -m_i(t) + \tanh b_i(t) \quad \text{(lowest order closure)}. \tag{30}$$

If this equation reaches a stationary state it must satisfy $m_i = \tanh b_i$, which we recognize as the equation of physical mean-field, i.e., Eq. (7). To the same lowest order, Eq. (27) is $\frac{\mathrm{d}\langle s_i(t)s_j(t_0)\rangle}{\mathrm{d}t} = -\langle s_i(t)s_j(t_0)\rangle + m_i(t)m_j(t)$, which relaxes to the uncorrelated state.

The first non-trivial equation is obtained by expanding Eq. (27) to first order which gives

$$\begin{aligned} &\langle s_i(t)s_j(t_0)\rangle + \frac{\mathrm{d}\langle s_i(t)s_j(t_0)\rangle}{\mathrm{d}t} \\ &= m_im_j + (1-m_i^2)\left(\sum_j J_{ik}\langle\delta s_k(t)\delta s_j(t_0)\rangle\right), \end{aligned} \tag{31}$$

where we have used Eq. (30) and stationarity to identify the derivative of the tanh function as $(1-m_i^2)$. Introducing

$$C_{ij}(t,t_0) = \langle \delta s_i(t) \delta s_j(t_0) \rangle = \langle s_i(t) s_j(t_0) \rangle - m_i m_j, \quad (32)$$

$$D_{ij}(t,t_0) = C_{ij}(t,t_0) + \frac{\mathrm{d}C_{ij}(t,t_0)}{\mathrm{d}t}, \quad (33)$$

we have

$$D_{ij}(t,t_0) = (1-m_i^2) \sum_k J_{ik} C_{kj}(t,t_0). \quad (34)$$

While this equation holds (to this order) for any two times t and t_0 it is especially convenient in the limit $t \to t_0$. Similarly to the procedure in naive mean-field inference (9) we can then invert Eq. (34) to arrive at an asynchronous mean field inference formula

$$J^{*,\mathrm{asyn-nMF}} = A^{-1} D C^{-1}, \quad (35)$$

where A is the diagonal matrix given by $A_{ij} = \delta_{ij}(1-m_i^2)$. Equation (35) is a linear matrix equation with respect to J_{ij}. We can solve it for J_{ij} directly for asynchronous Ising models.

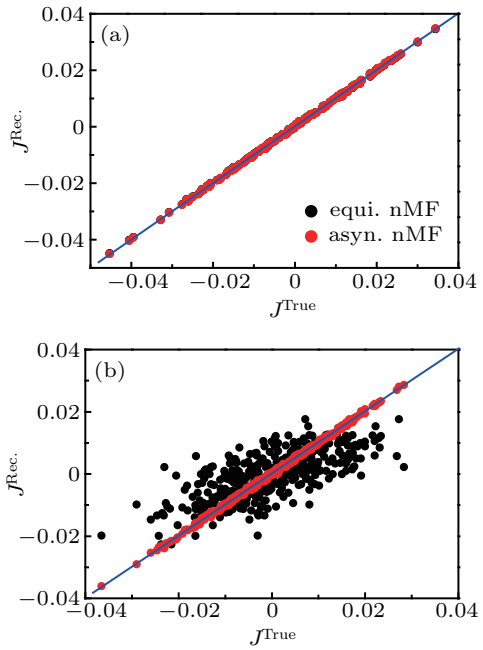


Fig. 1. The scatter plots for the true tested couplings versus the reconstructed ones. (a) Reconstruction for the symmetric SK model with $k = 0$; (b) inference for the asymmetric SK model with $k = 1$. Red dots, inferred couplings with asynchronous nMF approximation; black dots, inferred ones with equilibrium nMF approximation. The recovered asynchronous J_{ij}'s in (a) are symmetrized while no symmetrization for them in (b). The other parameters for both panels are $g = 0.3$, $N = 20$, $\theta = 0$, $L = 20 \times 10^7$.

Figure 1 shows the scatter plots for the tested couplings versus the recovered ones. The tested model for Fig. 1(a) is the symmetric SK model with $k = 0$ in Eq. (21) while fully asymmetric SK with $k = 1$ for Fig. 1(b). The couplings are reconstructed by the equilibrium nMF (9) (black dots) and the asynchronous nMF (35) method (red dots), respectively. As shown in Fig. 1(a), both the methods have the same ability to recover the tested symmetric SK model. Here, the data length $L = 20 \times 10^7$. Nevertheless, the couplings inferred by the asynchronous nMF needs to be symmetrized to keep the same results with that from equilibrium nMF, especially for short data length (not shown here). Figure 1(b) shows that, for the fully asymmetric SK model with $k = 1$, the asynchronous nMF works much better than the equilibrium nMF. This clearly shows that equilibrium inference methods are typically not suitable for non-equilibrium processes, while asynchronous inference works for both equilibrium and non-equilibrium process.

By a similar procedure we can also derive a higher-order approximation, which we refer to as dynamic TAP. The starting point is to redefine the $b_i(t)$ term in the tanh to include a term analogous to the static TAP Eq. (11). We then first have

$$H_i(t) = b_i - m_i \sum_{k \neq i} J_{ik}^2 (1-m_k^2) + \sum_k J_{ik} \delta s_k(t), \quad (36)$$

from which the lowest-order equation for the stationary state is of the TAP form. The second step is to expand the tanh function in Eq. (27) around $b_i - m_i \sum_{k \neq i} J_{ik}^2 (1-m_k^2)$ to the third order and to keep terms up to third order in J. In this way we get an inference formula, which is formally the same as in the nMF approximation,

$$J^{*,\mathrm{asyn-TAP}} = A^{-1} D C^{-1}, \quad (37)$$

where only the matrix A is different

$$A_{ij} = \delta_{ij}(1-m_i^2) \left[1 - (1-m_i^2) \sum_j J_{ij}^2 (1-m_j^2) \right]. \quad (38)$$

Equation (37) is a function of the couplings J, and therefore it is a nonlinear equation for matrix J.

Equation (37) could be solved for J though two approaches. One iterative way is starting from reasonable initial values J_{ij}^0, and inserting them in the RHS of formula (37). The resulting J_{ij}^1 is the solution after one iteration. They can be again replaced in the RHS to get the second iteration results and so on.

$$J^{t+1} = A(J^t)^{-1} D C^{-1}. \quad (39)$$

An alternative way is solving it by casting the inference formula to a set of cubic equations. For Eq. (38), denoting

$$F_i = (1-m_i^2) \sum_j J_{ij}^2 (1-m_j^2) \quad (40)$$

and plugging it into Eq. (37), and then we get the following equation for J_{ij}:

$$J_{ij}^{\mathrm{asyn-TAP}} = \frac{V_{ij}}{(1-m_i^2)(1-F_i)}, \quad (41)$$

where $V_{ij} = [DC^{-1}]_{ij}$. Substituting Eq. (41) into Eq. (40), we obtain the cubic equation for F_i as

$$F_i(1-F_i)^2 - \frac{\sum_j V_{ij}^2(1-m_j^2)}{1-m_i^2} = 0. \tag{42}$$

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

$$J_{ij}^{\rm asyn-TAP} = \frac{J_{ij}^{\rm asyn-nMF}}{1-F_i}. \tag{43}$$

3.4. Maximum-likelihood inference

To emphasize how different is inference from a time series compared to from samples, we will now show that maximum likelihood inference of such dynamics from such data is possible. We will also show that this approach admits approximation schemes different from mean-field. The presentation will follow.[82]

The log-likelihood of observing a full time series of a set of interacting spins is analogous to the probability of a history of a Poisson point process.[15] The probability space of events in some time period $[0:t]$ consists of the number of jumps (n), the times of these jumps $(t_1, t_2, \ldots, t_n)$ and which spin jumps at each time $(i_1, i_2, \ldots, i_n)$. The measure over this space is proportional to the uniform measure over n times a weight

$$\begin{aligned} &\mu_{i_1}^{(1)} {\rm d}t_1 \cdots \mu_{i_n}^{(n)} {\rm d}t_n \\ \times\ &\exp\left(-\mu^{(1)}t_1 - \mu^{(2)}(t_2-t_1) - \cdots - \mu^{(n+1)}(t-t_n)\right), \end{aligned}$$

where $\mu_{i_n}^{(n)}$ is the jump rate in open time interval $(t_{i-1}:t_i)$ of the event that actually took place at time t_i, and $\mu^{(n)} = \sum_i \mu_i^{(n)}$. We recall from the discussion of the Gillespie algorithm that in the open time interval $(t_{i-1}:t_i)$ all the rates stay the same, and that the length of the interval is an exponentially distributed random variable with parameter which is the sum of all the rates. In another time interval some or all of the rates can be different.

A rigorous construction of the above path probability can be found in Appendix A of Ref. [83]. Here we will follow a more heuristic approach and introduce a small finite time δt such that we can use the first simulation approach discussed in Section 3. The objective function to maximize is then

$$\mathcal{L} = \sum_{i,t} \log\left[(1-\gamma\delta t)\delta_{s_i(t+\delta t),s_i(t)} + \gamma\delta t \frac{{\rm e}^{s_i(t+\delta t)H_i(t)}}{2\cosh H_i(t)}\right]. \tag{44}$$

The sums in Eq. (44) go over all spins i and all times separated by the small increment δt. The terms in Eq. (44) can be understood as the lowest order approximation (linear in δt) of $\log\prod_{it} P_i(s_i(t+\delta t)|\boldsymbol{s}(t))$, where P_i is the conditional probability of spin i at time $t+\delta t$, conditioned on the configuration of all spins at time t. Maximum likelihood inference of dynamics from a time series is therefore analogous to pseudo-maximum likelihood in Eq. (14) from independent samples. At the price of potentially very many and very biased samples (at most times no spin will jump) this points to that inference from a time series is a fundamentally easier task.

Separating times with and without spin flips (45), the resulting learning rules will be

$$\begin{aligned} \delta J_{ij} &\propto \frac{\partial \mathcal{L}}{\partial J_{ij}} \\ &= \sum_{flips} [s_i(t+\delta t) - \tanh(H_i(t))] s_j(t) \\ &\quad + \frac{\gamma\delta t}{2} \sum_{no\ flips} q_i(t) s_i(t+\delta t) s_j(t), \end{aligned} \tag{45}$$

with $q_i(t) \equiv [1-\tanh^2(H_i(t))]$, and it includes the rule for the θ_i with the convention $J_{i0} = \theta_i$, $s_0(t) = 1$. Following Ref. [82] where we also considered the case that the times where nothing happens are known, we will refer Eq. (45) as the "spin-history-only" (SHO) algorithm.

Similarly to mean-field inference, Eq. (45) can also be averaged, which gives the learning rule

$$\delta J_{ij} \propto \gamma^{-1}\dot{C}_{ij}(0) + C_{ij}(0) - \langle \tanh(H_i(t)) s_j(t)\rangle, \tag{46}$$

which we refer to as AVE.[82] AVE requires knowing equal-time correlations, their derivatives at $t=0$, and $\langle \tanh(H_i(t)) s_j(t)\rangle$. This latter quantity depends on the model parameters (through $H_i(t)$), so, in practice, estimating it at each learning step requires knowing the entire spin history, the same data as needs SHO leaning.

All of four methods now introduced to infer parameters from a time series (nMF, TAP, SHO and AVE) will produce a fully connected network structure. Similarly to inverse Ising from samples we may want to include L_1 penalties to get the graphical structure.[84] Such effects are considered in Ref. [85], showing that inferring the sparsity structure from time series data is both a feasible and reliable procedure.

3.5. Performance of kinetic Ising inference methods

In this section, performance tests of the four above introduced algorithms for recovering parameters in asynchronous Ising models are presented. We compared the performance of two ML algorithms SHO, and AVE to each other and to two mean-field algorithms nMF and TAP.

The tested model is as discussed above the fully asymmetric SK model (J_{ij} is independent of J_{ji}), J_{ij}'s are identically and independently distributed Gaussian variables with zero means and variance g^2/N. As a performance measure, we use the mean square error (ε) which measures the L_2 distance between the inferred parameters and the underlying parameters used to generate the data

$$\varepsilon = \frac{\sum_{i\neq j}(J_{ij}^* - J_{ij}^{\rm True})^2}{N(N-1)}, \tag{47}$$

where $J_{ij}^{\rm True}$ are the true values of interactions and J_{ij}^{*} are the inferred ones. We study the reconstruction error for different data length L, system size N, external field θ and coupling strength g.

Figure 2 shows the performance of these algorithms. Each panel also shows two ML-based learning methods SHO and AVE appear to perform equally well for large enough L since they effectively use the same data (the spin history). Note however the opposite trend in Fig. 2(a) shows the reconstruction getting better with longer data length L for both ML and mean-field based methods. Figure 2(b) shows that the MSE for the ML algorithms is insensitive to N, while two mean-field algorithms improve as N becomes larger; in these calculations, the average numbers of updates and flips per spin were kept constant, taking $L = 5 \times 10^5 N$). Figure 2(c) shows that the performance of two ML algorithms is also not sensitive at all to θ, while nMF and TAP work noticeably less well with a non-zero θ. The effects of (inverse) g are depicted in Fig. 2(d). For fixed L, all the algorithms do worse at strong couplings (large g). The nMF and TAP do so in a much more clear fashion at smaller g, growing approximately exponentially with g for g greater than ≈ 0.2. In the weak-coupling limit, all algorithms perform roughly similarly, as already seen in Fig. 2(a).

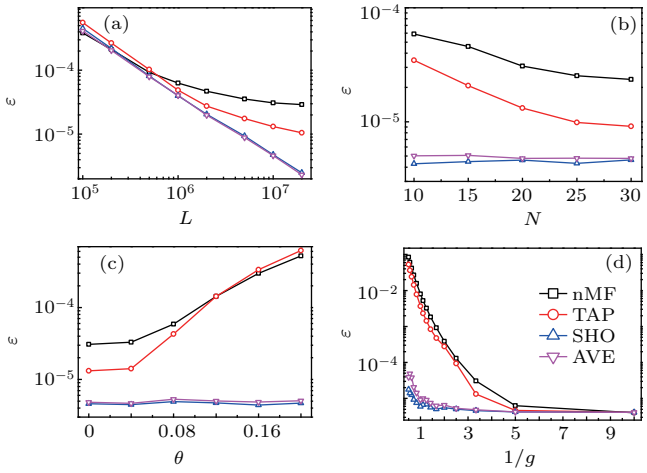


Fig. 2. Mean square error (ε) versus (a) data length L, (b) system size N, (c) external field θ and (d) temperature $1/g$. Black squares show nMF, red circles, TAP, blue up triangle SHO and pink down triangle AVE respectively. The parameters are $g = 0.3$, $N = 20$, $\theta = 0$, $L = 10^7$ except when varied in a panel.

To summarize, the ML methods recover the model better, but in general more slowly. The mean-field based learning rules (nMF and TAP) are much faster in inferring the couplings but have worse accuracy compared with that of the ML-based iterative learning rules (AVE, SHO).

4. Example applications of asynchronous Ising model

Inverse Ising problems have been applied to a wide rage of data analysis, ranging from equilibrium reconstruction methods to kinetic ones. In this section, based on Refs. [81,86], we will present as illustrations applications to one data set of neuronal spike trains, and one data set on transaction data of stocks on financial market. Both areas have been investigated extensively in the last ten years. We refer to Refs. [87–93] for more recent neuronal data and discussions of inference in this context and to Refs. [94–105] for a sample of contributions considering financial data.

For the neuronal data, we show two ML-based learning rules. When considering the data as a time series we use the AVE method of Eq. (35). However, when considering the same data as independent samples from the Gibbs distribution (1) we use Boltzmann machine (BM) (introduced below). We find that for this data the couplings between the neurons obtained are comparable. This means that although there is no a priori for this to be so, the dynamic process of this neuron system apparently satisfies detailed balance or has a stationary distribution of the form of Eq. (1) for other reasons. One clear difference is the self-couplings from one neuron to itself, which are absent in Eq. (1) and are typically present in the dynamic model. A further difference is that to infer parameters from Eq. (1) using samples, those samples have to be generated by Monte Carlo procedure. Although both the methods are based on ML, the dynamic version is thus considerably faster than BM.

For the financial stock trade data we show two mean-field-based algorithms. When considering the data as a time series we use the asynchronous nMF method of Eq. (35). However, when considering the same data as independent samples from Eq. (1), we use naive mean-field inference (here equilibrium nMF) of Eq. (9). We note that we here apply inverse Ising inference to binary data obtained by transforming a time series of financial transactions (see below). Again we find that the results from the two procedures are comparable, except that asynchronous nMF allows the inference of self-couplings, as well as directed links (asymmetric couplings).

4.1. Case 1: Reconstruction of a neuron network from spiking trains

Neurons are the computational units of the brain. While each neuron is actually a cell with complicated internal structure, there is a long history of considering simplified models where the state of a neuron at a given time is a Boolean variable. Zero (or down, or -1) then means resting, and one (or up, or $+1$) means firing, or having a spike of activity. In most neural data most neurons are resting most of the time.

Data description and representation of data The neuronal spike trains are from salamander retina under stimulation by a repeated 26.5-second movie clip. This data set records the spiking times for neurons and has a data length of 3180 s (120 repetitions of the movie clip). Here, only the first $N = 20$ neurons with highest firing rates in the data set are considered.

The data has been binned with time windows of 20 ms (the typical time scale of the auto-correlation function of a neuron) in the previous study.[106] However, since we are using the kinetic model, we could study this data set using a much shorter time bin which leads low enough firing rates and (almost) never more than one spike per bin. Then, the temporal correlations with time delays between neuron pairs as well as the self-correlations become important.

For the asynchronous Ising model, the time bins are $\delta t = 1/(\gamma N)$. For neuronal data, γ can be interpreted as the inverse of the time length of the auto-correlation function, which is typically 10 ms or more.[106] To generate the binary spin history from this spike train data set, the spike trains should be separated into time bins with length $\gamma\delta t = 1/20$. This means that the size of time bins should be chosen as $\delta t = 1/(20\gamma) = 0.5$ ms. The spin trains can be transformed in to binaries as follows: a $+1$ is assigned to every time bin in which there is a spike and a -1 when there is no spikes. To avoid the case that the translation always end up with isolated instances of $+1$ and superfluous -1 s, the memory process for each neuron is introduced to the data set. It is a time period with an exponential distribution with mean of $1/\gamma$ in the data translation. Denote the total firing number of neuron i as F_i, and t_i^f as the firing time of fth spike for neuron i, where $i = 1,\ldots,N$ and $f = 1,\ldots,F_i - 1$, then the mapping of the spike history is follows:

$$s_i(t) = \begin{cases} 1, & \text{if } t \in \left[t_i^f, \min\left(t_i^{f+1}, t_i^n + X\right)\right) \text{ with } X \sim \exp(\gamma^{-1}), \\ -1, & \text{otherwise}, \end{cases} \tag{48}$$

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

Inference methods For this fairly small system we use two types of ML to learn the parameters of Eq. (1). In the equilibrium case of Eq. (3) this can be carried out with the iterative method called Boltzmann machine (BM) which is defined as follows:

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

In above η "learning rate" is a relaxation parameter. For larger systems BM does not scale since computing the ensemble averages $\langle s_i\rangle_{\text{Model}}$ and $\langle s_i s_j\rangle_{\text{Model}}$ is costly, but for the data under consideration here it is a feasible method. When retaining the time series nature of the data we on the other hand use the AVE learning rule of Eq. (46).

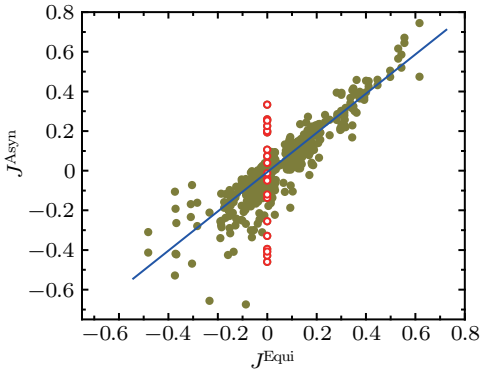


Fig. 3. Inferred asynchronous versus equilibrium couplings for retinal data. Red open dots show the self-couplings which by convention are equal to zero for the equilibrium model.

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

As presented in Fig. 3, the inferred couplings by BM and asynchronous kinetic Ising model are very close to each other. We also tested what happens to the couplings of the asynchronous model if during learning we symmetrized the couplings matrix at each iteration by adding its transpose to itself and dividing by two and also putting the self-couplings to zero. We find that the resulting asynchronous couplings get even closer to the equilibrium ones, which is consistent to the conclusion for kinetic Ising data.

However, the asynchronous model allows the inference of self-couplings (diagonal elements of the coupling matrix) which are not present in the equilibrium model. As shown in Fig. 3, the diagonals from the equilibrium model equals to zeros by convention and denoted by the open red dots. Furthermore, to be different from the symmetric couplings by the equilibrium model, the asynchronous model provides more details as the inferred couplings are directed and asymmetric.

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

4.2. Case 2: Reconstruction of a finance network

In this study, we present equilibrium nMF (9)) and asynchronous nMF (35) algorithm to infer a financial network from trade data with 100 stocks. The recorded time series are transformed into binaries by local averaging and thresholding. This introduces additional parameters that have to be studied extensively to understand the behavior of the system. The inferred couplings from asynchronous nMF method is quite similar to the equilibrium ones. Both produced network communities have similar industrial features. However, the asynchronous method is more detailed as they are directed compared with that from the equilibrium ones.

Data description and representation The data was transactions recordings on the New York Stock Exchange (NYSE) over a few years. Each trade is characterized by a time, a traded volume, and a price. We only focus on the trades for 100 trading days between 02.01.2003 and 30.05.2003. However, trading volume and trading time only are utilized in the study. To avoid the opening and closing periods of the stock exchange, 10^4 central seconds of each day are employed as in Ref. [107]. Two parameters are introduced to the data transform as the sliding window is adopted. One is the size of the sliding time window (denoted as Δt), the other one is the shifting constant which is fixed as 1 s.

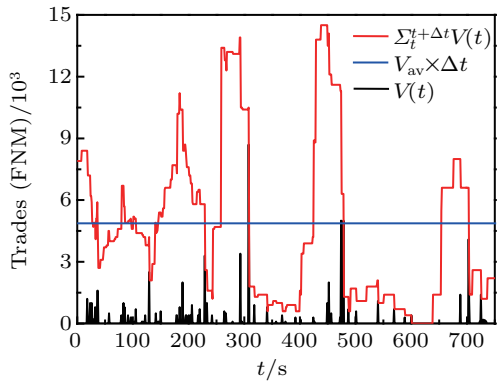


Fig. 4. Traded volume data for the stock of Fannie Mae (FNM), a mortgage company. Black line for time series of traded volumes $V_i(t)$, red for summed volumes during time interval Δt, blue for the threshold $V_i^{\rm th} = \chi V_i^{\rm av} \times \Delta t$. Parameters: $\Delta t = 50$ s and $\chi = 1$.

For stock i, the sum of the volumes $V_i(t,\Delta t)$ traded in window $[t, t+\Delta t)$ is compared with a given volume threshold $V_i^{\rm th} = \chi V_i^{\rm av}\Delta t$, where $V_i^{\rm av}$ is the average (over the whole time series) volume of the considered stock traded per second, and χ a parameter controlling our volume threshold:

$$s_i(t) = \begin{cases} 1, & \text{if } V_i(t,\Delta t) \geq V_{\rm th}^i, \\ -1, & \text{if } V_i(t,\Delta t) < V_{\rm th}^i. \end{cases} \tag{49}$$

We explored the parameters Δt and χ systematically for the inference with the goal of finding values of the parameters which yield inferred couplings containing interesting information.

Figure 4 shows the traded volume information for a mortgage company Fannie Mae (FNM). With the mapping approach described in Eq. (49), we have $+1$'s above the blue threshold line in Fig. 4 while -1's below that line. Then the asynchronous data is ready for the network reconstruction.

Inference methods With the transformed binaries, the magnetization m_i and correlations $C_{ij}(\tau)$ are defined as Eq. (25). With them, two different inference methods with nMF approximation are utilized for the reconstruction.

Equilibrium nMF ($i \neq j$), which only focuses on equal time correlations[108]

$$J_{ij} = -C(0)^{-1}_{ij}.$$

Asynchronous nMF (Ref. [81]) uses the derivative of the time-lagged correlations $\dot{C}_{ij}(\tau)$, as shown in Eq. (35) and rewritten as

$$J_{ij} = \frac{1}{1-m_i^2}\left(\frac{{\rm d}C(\tau)}{{\rm d}\tau}|_{\tau=0}C(0)^{-1}\right)_{ij}.$$

Reconstruction results Massive explorations over different values of the window size Δt and χ's are complimented to achieve meaningful interactions between stocks. A natural rough approach is to consider that couplings will contain interesting information if they are big in absolute value: they indicate a strong interaction between stocks. For asynchronous inference, the derivative of the time-lagged correlations $\dot{C}_{ij}(\tau)$ is computed through a linear fitting of this function $C_{ij}(\tau)$ using four points: $C(0)$, $C(\Delta t/5)$, $C(2\Delta t/5)$ and $C(3\Delta t/5)$.

Figure 5 shows that both the inference methods give similar distributions of couplings. For comparison, the distributions are re-scaled so as to have the same standard deviation. It can be remarked that the inferred couplings have a strictly positive mean and a long positive tail. This prevalence of positive couplings can intuitively be linked with the market mode phenomenon:[109–112] a large eigenvalue appears, corresponding to a collective activity of all stocks, as illustrated in Fig. 6.

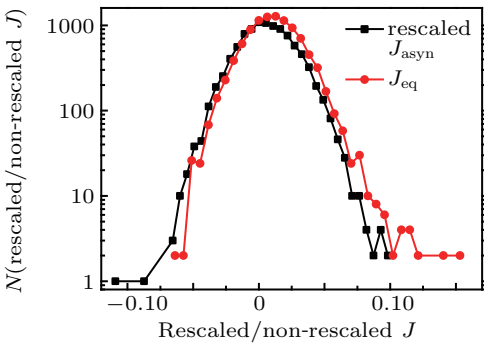


Fig. 5. Histograms of inferred couplings by equilibrium nMF and re-scaled asynchronous nMF. Black squares for re-scaled $N(J_{\rm asyn})$ to have the same standard deviation as $N(J_{\rm eq})$. Here $\chi = 0.5$ and $\Delta t = 200$ s for both the methods.

The similarity of interaction matrices J and J' inferred

from different methods can be measured by a similarity quantity $Q_{J,J'}$, which is defined as

$$Q_{J,J'} = \frac{\sum_{i,j} J_{ij} J'_{ij}}{\sum_{i,j} \max(J_{ij}, J'_{ij})^2}. \tag{50}$$

This measurement compares elements of two matrices one by one and gives a global similarity measure. It takes real values between 1 (when $J_{ij} = J'_{ij}$ for all i and j) and -1 ($J_{ij} = -J'_{ij}$ for all i and j), and values close to zero indicate uncorrelated couplings. The values of Q is smaller than 0.02 in absolute value when all elements of the vectors J_{ij} and J'_{ij} are drawn independently at random from a same Gaussian distribution, of mean 0, and for different values of the standard deviation of this distribution. Here, the value of Q for the inferred J_{ij}'s by equilibrium nMF and asynchronous nMF is about 0.5 with $\chi = 0.5$ and $\Delta t = 50$ s, which indicates these two methods are not independent of each other.

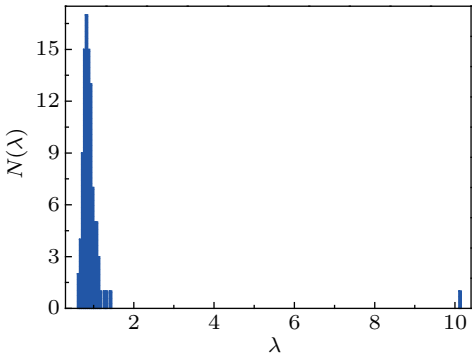


Fig. 6. Histograms of the eigenvalues of the equal time connected correlation matrix. Parameters: $\chi = 0.5$ and $\Delta t = 100$ s.

Next, we will present two inferred financial networks that recovered by equilibrium and asynchronous nMF method, respectively. As the inferred finance networks are densely connected, we focus only on the largest couplings, which can be explained by closely related activities of the considered stocks. Figure 7(a) shows that with equilibrium inference, more than half the stocks in the data can be displayed on a network where almost all links have simple economical interpretations.

The network of Fig. 7(a) presents different communities, each color represents one industrial sector. They are mostly determined by a common industrial activity. Some of the links are very easy to explain with the proximity of activities (and often quite robust). For instance, the pairs FNM-FRE (Fannie Mae-Freddie Mac, active in home loan and mortgage), UNP-BNI (Union Pacific Corporation-Burlington Northern Santa Fe Corporation, railroads), BLS-SBC (BellSouth-SBC Communications, two telecommunications companies now merged in AT&T), DOW-DD (Dow-DuPont, chemical companies), MRK-PFE (Merck & Co.-Pfizer, pharmaceutical companies), KO-PEP (The Coca-Cola Company-PepsiCo, beverages). These two last companies display strong links with the medical sector at different scales of volume and time, as KO here with MDT (Medtronic) and JNJ (Johnson & Johnson). This medical sector is itself linked to the pharmaceutical sector with PFE, MRK, LLY (Lilly), BMY (Bristol-Myers Squibb) and SGP (Schering-Plough). Telecommunications (BLS, SBC) are linked to electric power with DUK (Duke Energy).

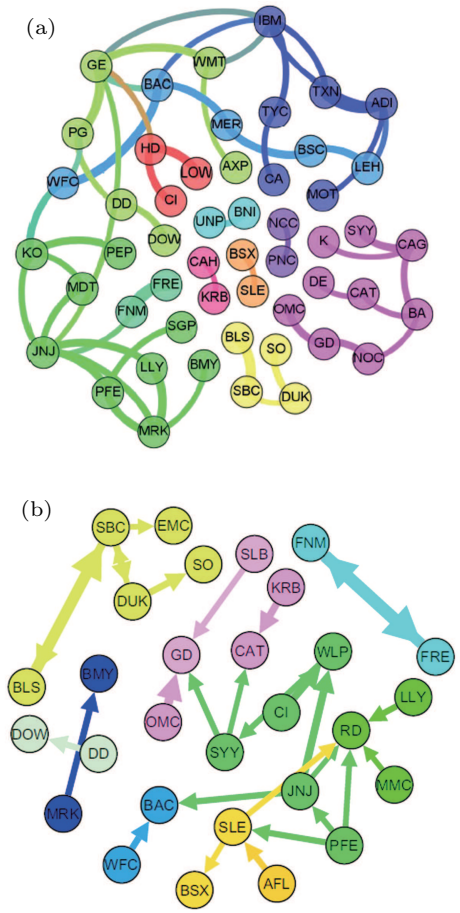


Fig. 7. Inferred financial networks, showing only the largest interaction strengths (proportional to the width of links and arrows). Colors are indicative, and chosen by a modularity-based community detection algorithm.[16] Parameters: $\chi = 0.5$ and $\Delta t = 100$ s. (a) Equilibrium inference (the figure originates from Ref. [86]). (b) Asynchronous inference with $\tau = 20$ s.

GE (General Electrics) is for a large range of parameters a very central node, which is consistent with its diversified activities. Figure 7(a) from Ref. [86] presents the relation between PG (Procter & Gamble) and WMT (Walmart), both retailers of consumer goods come at this level of interaction strength through GE.

The banking sector as shown by a chain with light blue color is linked to the sector of electronic technology (with dark blue color). Moreover, the defense and aerospace sector as shown in magenta is linked to engines and machinery with (CAT) (Caterpillar Inc.) and DE (John Deere), and more strangely, to packaged food with CAG (ConAgra Foods), SYY (Sysco) and K (Kellogg Company).

Figure 7(b) presents the results from asynchronous nMF

under the same conditions. It shows that the results of equilibrium and asynchronous inference are consistent, and that asynchronous inference provides additional information, as it infers an directed network. For instance, the financial sector is directed and also influenced by the medical sector. The detailed descriptions for each stock can be found in Ref. [113].

From the network samples, we have the following two basic conclusions. First, they show market mode (most of the interaction strengths found are usually positive, which indicates that the financial market has a clear collective behavior)[110,111] even only trade and volume information is considered. Stocks tend to be traded or not traded at the same time.

In addition, the strongest inferred interactions can be easily understood by similarities in the industrial activities of the considered stocks. This means that financial activity tends to concentrate on a certain activity sector at a certain time. For price dynamics this phenomenon is well-known,[109,112,114] but it is perhaps more surprising that it appeared also based on only information of traded volumes.

5. Fitness inference of population genetics

We now turn to inverse Ising (Potts) techniques applied to sequence-type biological data. This has variously been called direct coupling analysis (DCA)[36,43,46,47] and max-entropy modeling;[35,115] as noted above, other names are also in use. We will use the terms inverse Ising and DCA interchangeably.

A common feature of all these applications is that the input is a static table of $N \cdot L$ symbols. Each row is a sequence of L symbols from data, and there are N such rows (N samples). A breakthrough application has been to identify residues (amino acid molecules) that are spatially close in proteins (chains of amino acids). The table then represents a family of proteins with supposedly similar structure and supposedly same origin, and each row is the amino acid sequence of a member of that family.[36,43,46,47] The basic idea is that two columns in the table (two positions in the protein structure) have non-trivial statistical dependency if their joint variation influence biological fitness. Such co-dependency in biological fitness is called epistasis. The most immediate cause of epistatsis among loci inside one gene coding for one protein is through structure.[116] Often this is pictorially motivated by a mutation changing charge, hydrophilic/hydrophobic or size of one member of a residue pair, which then changes the relative fitness of variants (alleles) of the other member of the pair. In certain other cases dependencies discovered by DCA can be attributed to other causes than structure[117,118] but those cases appear to be relatively rare.

Many details are needed to turn the above to powerful tool in protein structure prediction. One aspect is that proteins in a family typically have different lengths, and that therefore the $N \cdot L$ table is not directly taken from the data, but only after multiple sequence alignment, which has to be carried out with the help of bionformatics software, or the ready alignment taken from a data base such as PFAM.[119,120] Another is that predicting contacts are only one ingredient in a much larger computational pipeline which uses intermolecular force fields, predictions on secondary structure and solvability and know-how developed in the protein science community over many years. Still, impressive results have been achieved.[121–123] It should be noted that if the goal is to predict protein structure a purely data-driven approach is possible, where a model of the deep neural network type is trained on large training sets comprised of sequence-structure pairs. As has been widely reported, such an approach from Google Deep Mind currently outperforms model-based learning methods such as DCA for this task.[124,125] The price is computational cost beyond what most academic researchers can afford, and lack interpretability of the inferred model, which could be close to Eq. (1), but could also be very different.

Beyond protein structures, DCA has been used to predict nucleotide-nucleotide contacts of RNAs,[126] multiple-scale protein-protein interactions,[118] amino acid–nucleotide interaction in RNA-protein complexes,[127] interactions between HIV and the host immune system,[128–130] and other synergistic effects not necessarily related to spatial contacts.[131–133] Of particular relevance for the following are applications of DCA to whole-genome sequence data from bacterial populations, in Ref. [134] on Streptoccoccus pneumoniae and in Ref. [135] on Neisseria gonorrhoeae. Standard versions of DCA are rather compute-intensive for genome-scale inference tasks, but methodological speed-ups[136,137] and alternative approaches[138] have been quickly developed. Antibiotic resistance is an important medical problem throughout the world, and so is the relative paucity of new drugs. Combinatorial drug combinations are therefore promising avenues to look for new treatment strategies. The obstacle is the combinatorial explosion of combinations: if there are L potential individual targets there will be L^2 potential target pairs, and so on. The hope is that DCA could be one way (one out of many) to predict which combinations may have an effect on the grounds that they are already reflected as epistasis in natural sequence data. In that respect it was promising that Skwark *et al.* in Ref. [134] were able to retrieve interactions between members of the penicillin-binding protein (PBP) family of proteins; resistance to antibiotics in the β-lactam family of compounds is in S. pneumoniae associated to alterations in their target enzymes, which are the PBPs.[139]

Evolution is a dynamic process. We should imagine that the biological sequence data used in DCA are as in Section 2.2 (or more involved). The underlying dynamics is of N_{tot} sequences (a number which could change with time, but

which we will assume constant) of length L (which could also change, but which we will also assume constant), and which evolve evolving for a time T. At the end of the process we sample N sequences. In protein data, T is typically of the order of hundreds of millions of years, and the model is obviously simplified. In the bacterial whole-genome data of Refs. [134,135], T may be as short as years or decades, and the model may be closer to reality. In any case, the goal is to infer fitness from the sampled sequences, and to understand when it can (or cannot) reasonably be done by DCA.

We will structure the discussion as follows. In Section 5.1, we will discuss dynamics of a population in a fitness landscape on which there is a large literature both in population genetics and in statistical physics. We will there define what we mean by fitness, and introduce recombination. In Section 5.2, we will present the important concept of quasi-linkage equilibrium (QLE), originally due to Kimura, and in Section 5.3 we will state the relation between inferred interactions and underlying fitness that hold in QLE. Numerical examples and tests are presented in Section 5.4.

5.1. Dynamics of a sexually reproducing population in a fitness landscape

That there exist formal similarities between the dynamics of genomes in a population and entities (spins) in statistical physics has been known for a long time. Fokker–Planck equations to describe the change of probability distributions over allele frequencies were introduced by Fisher almost a century ago,[140,141] and later, in a very clear a concise manner, by Kolmogorov.[142] The link has been reviewed several times from the side of statistical physics, for instance in Refs. [143,144]. Central to the discussion in the following will be recombination (or sex), by which two parents give rise to an offspring, the genome of which is a mixture of the genome of the parents. From the point of view of statistical physics, recombination is a kind of collision phenomena. It therefore cannot be described by linear equations (Fokker–Planck-like equations) but can conceivably be described by nonlinear equations (Boltzmann-like equations). The mechanisms to be discussed are of this type, where Boltzmann's Stosszahlansatz is used to factorize the collision operator.

All mammals reproduce sexually, as do almost all birds, reptiles and fishes, and most insects. Many plants and fungi can reproduce both sexually and asexually. Recombination in bacteria is much less of a all-or-none affair. Typically only some genetic material is passed from a donor to a recipient, directly or indirectly. The main forms of bacterial recombination are conjugation (direct transfer of DNA from a donor to a recipient), transformation (ability to take up DNA from the surroundings), and transduction (transfer of genetic material by the intermediary of viruses). The relative rate of recombination in bacteria varies greatly between species, and also within one species, depending on conditions. As one example we quote a tabulation of the ratio of recombination to mutation rate in S. pneumoniae, which has been measured to vary from less than one to over forty.[145] There are long-standing theoretical arguments against the possibility of complex life without sex, as a consequence of Eigen's "error catastrophe".[146,147] It is likely that most forms of life use some form of recombination, albeit perhaps not all the time, though the relative rate of recombination to other processes may be small. In the following we will eventually assume that recombination is faster than other processes, which may be as much the exception as the norm in bacteria and other microorganisms. Such a "dense-gas" (using the analogy with collisions) is however where there is an available theory which can be used at the present time.

The driving forces of evolution are hence assumed to be genetic drift, mutations, recombination, and fitness variations. The first refers to the element of chance; in a finite population it is not conformed which genotypes will reproduce and leave descendants in later generations. The last three describe the expected success or failure of different genotypes.

Genetic drift can be explained by considering N different genomes $s_1,\ldots,s_N$. Under neutral evolution all genomes have equal chance to survive into the next generation, but that does not mean all will do so. In a Wright–Fisher model one considers a new generation with N new genomes $s'_1,\ldots,s'_N$, where each one is a drawn randomly with uniform probability from the previous generation. The chance (or risk) that an individual does not survive from one generation to the next is then $\left(1-\frac{1}{N}\right)^N$, which is about $\mathrm{e}^{-1}\approx 37\%$. Monte Carlo simulations of finite populations necessarily include such effects where some successful individuals crowd out other less fortunate ones.

Mutations are random genome changes described by mean rates. A model of N individuals evolving under mutations and genetic drift which happen synchronously is also called a Wright–Fisher model, and when they happen asynchronously a Moran model.[144] If the genome (or the variability of the genome) consists of only one biallelic locus (one Ising spin, $L=1$) then the state of a population will be given by the number k of individuals where the allele is "up" ($N-k$ individuals then have the "down" allele). The dynamics of the Moran model can be seen as the dynamics of N spins where each spin can flip on its own or can copy the state of another spin (or do nothing). It can also be seen as a transitions in a finite lattice where $k=0$ means all spins are down, and $k=N$ means all spins are up. The probability distribution over this variable k changes by a Master equation where the variable can take values $0,1,\ldots,N$. If mutation rate is zero the two end states in the lattice will be absorbing: eventually all individuals will be up, or all will be down. If mutation rate is non-zero but

small, the stationary probability distribution will be centered on small and large k and transitions between the two macroscopic states will happen only rarely. For a very pedagogical discussion of these classical facts we can refer to Ref. [144].

The evolution of the distribution over L biallelic loci under mutations has many similarities to Eq. (15), and always satisfies detailed balance (this is not generally true for more than one allele per locus and a general mutation matrix). As the rate $r_i(\boldsymbol{s})$ in Eq. (6) the rate of mutations can and generally does depend on genomic position ("mutation hotspots") and on the alleles at other loci ("genomic context"). For theoretical discussion and simulations it is however more convenient to assume an overall uniform flipping rate as in Eq. (16).

A fitness landscape means a propensity for a given genotype to propagate its genomic material to the next generation. This propensity is a function of genotype, and called a fitness function. It is important to note that this concept does not cover all that fitness can mean in biology. Excluded effects are for instance cyclical dominance where A beats B, B beats C and C beats A.[148–152] We will assume that the fitness of genotype $\boldsymbol{s}$ which carries allele s_i on locus i depends on single-locus variations and pair-wise co-variations, that is,

$$F(\boldsymbol{s}) = F_0 + \sum_i f_i s_i + \sum_{ij} f_{ij} s_i, s_j. \tag{51}$$

The first term is an overall constant. The second term, linear in the genome, is called additive component of fitness. The last term, quadratic in the genome, is called the epistatic component of fitness. The dynamics due to Darwinian selection on the level of populations is thus

$$\frac{\partial P(\boldsymbol{s})}{\partial t}|_{\text{sel}} = P(\boldsymbol{s})\left(F(\boldsymbol{s}) - \langle F(\boldsymbol{s})\rangle\right), \tag{52}$$

where $\langle F(\boldsymbol{s})\rangle = \sum_{\boldsymbol{s}} F(\boldsymbol{s}P(\boldsymbol{s})$ is the average fitness over the population. Evolutionary dynamics due to fitness in a fitness landscape is thus quadratic in the distribution function ($^8P(\boldsymbol{s})\langle F(\boldsymbol{s})\rangle$ is quadratic in $P(\boldsymbol{s})$). The conditions under which the combined dynamics under mutations and fitness satisfy detailed balance are a kind of integrability conditions. On the level of dynamics on allele frequencies such a condition is known as the existence of a Svirezhev–Shahshahani potential,[153–156] also see Ref. [157].

Recombination (or sex) is the mixing of genetic material between different individuals. In diploid organisms (such as human) every individual has two copies of each separate component of its genetic material (chromosome), where one comes from the father and one comes from the mother, each of whom also has two copies, one from each grandparent. When passing from the parents to the child the material from the grandparents is mixed in the process called cross-over, so that one chromosome of the child inherited from one parent typically consists of segments alternately taken from the two chromosomes of that parent.

In haploid organisms the situation is both simpler since each organism only has one copy of its genetic material, and also more complicated since the mixing of information can happen in many different ways. It is convenient to postulate a dynamics like a physical collision process

$$\frac{\partial P(\boldsymbol{s})}{\partial t}|_{\text{rec}} = r\sum_{\xi,\boldsymbol{s}'} C(\xi)\big[Q(\boldsymbol{s}_1,\boldsymbol{s}_2)P_2(\boldsymbol{s}_1,\boldsymbol{s}_2) - Q(\boldsymbol{s},\boldsymbol{s}')P_2(\boldsymbol{s},\boldsymbol{s}')\big], \tag{53}$$

where r is an overall rate of sex compared to other processes, $Q(\boldsymbol{s}_1,\boldsymbol{s}_2))$ is the chance of individuals $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ mating (reaction probability) and $C(\xi)$ is the chance that they produce an offspring given by a pattern ξ [158,159] (probability of outcome of reaction). On the left hand side we have the single-genome distribution function P, and on the right-hand side the two-genome distribution function P_2; the equations are closed by a Stosszahlansatz

$$P_2(\boldsymbol{s}_1,\boldsymbol{s}_2) = P(\boldsymbol{s}_1)P(\boldsymbol{s}_2). \tag{54}$$

The complexities of recombination can then be accommodated by the two functions Q and C. A method to infer recombination hotspots in bacterial genomes was discussed in Ref. [160], and the issue was also discussed in Ref. [161], in relation to the same "Maela" data set used in Ref. [134]. Detailed descriptions of the relation between on the one hand $(\boldsymbol{s},\boldsymbol{s}')$ and on the other $(\boldsymbol{s}_1,\boldsymbol{s}_2)$ as parameterized by ξ can be found in Refs. [158,159].

5.2. The quasi-linkage equilibrium phase

The concept of quasi-linkage equilibrium (QLE) and its relation to sex were discovered by population geneticist Kimura,[162–164] and later developed by Neher and Shraiman in two influential papers.[158,165] We will refer to this theory of QLE as the Kimura–Neher–Shraiman (KNS) theory. To define QLE and to state the main result of KNS we must first introduce the simpler concept of linkage equilibrium (LE), which goes back to the work of Hardy and Weinberg more than a century ago.[166,167]

Consider two loci A and B where there can be, respectively, n_{A} and n_{B} alleles. The configuration of one genome with respect to A and B is then $(x_{\text{A}},x_{\text{B}})$, where x_{A} takes values in $\{1,\ldots,n_{\text{A}}\}$ and x_{B} takes values in $\{1,\ldots,n_{\text{B}}\}$. The configuration of a population of N individuals is the set $[(x_{\text{A}}^{(s)},x_{\text{B}}^{(s)})]$, where s ranges from 1 to N. This set defines the empirical probability distribution with respect to A and B as

$$P_{\text{AB}}(x_{\text{A}},x_{\text{B}}) = \frac{1}{N}\sum_{s=1}^{N} 1_{x_{\text{A}}^{(s)},x_{\text{A}}} 1_{x_{\text{B}}^{(s)},x_{\text{B}}}, \tag{55}$$

where $1_{a,b}$ is the Kronecker delta. Similarly we can define distributions over one locus as $P_{\rm A}(x_{\rm A}) = \frac{1}{N}\sum_{s=1}^{N} 1_{x_{\rm A}^{(s)},x_{\rm A}}$, and $P_{\rm B}(x_{\rm B})$. The distribution of genomes in a population over loci A and B is said to be in linkage equilibrium (LE) if the alleles $a_{\rm A}$ and $x_{\rm B}$ are independent under the empirical distribution, i.e., if $P_{\rm AB}(x_{\rm A},x_{\rm B}) = P_{\rm A}(x_{\rm A})P_{\rm B}(x_{\rm B})$. All other distributions are in linkage disequilibrium (LD).

Specifying for completeness to the case of interest here where all loci are biallelic and epistatic contributions to fitness is quadratic (pairwise dependencies), and quasi-linkage equilibrium is a subset of distributions in LD where the joint distribution over loci is the Gibbs distribution in Eq. (1). The fundamental insight of Kimura was that such distributions appear naturally in sexually reproducing populations where recombination is fast.[162–164] In this setting, epistatic contribution to fitness is a small effect since there is a lot of mixing of genomes between individuals from one generation to the next. The dependencies (parameters J_{ij}) are also small, such that the distributions over alleles are almost independent. In other words, the distributions in QLE which appear in KNS theory are close to being in linkage equilibrium. Nevertheless, the parameters h_i and J_{ij} in Eq. (1) are hence here consequences of a dynamical evolution law.

The derivation of Eq. (1) from the dynamics described above in Section 5 has been given in the literature[158,159] and will be therefore not be repeated here. We will instead just state the most important result of the KNS theory. This is

$$J_{ij} = \frac{f_{ij}}{rc_{ij}}, \tag{56}$$

where c_{ij} characterizes the amount of recombination between loci i and j. Referring to the dynamics (53) this quantity is defined as

$$c_{ij} = \sum_{\xi} C(\xi)\left(\xi_i(1-\xi_j) + (1-\xi_i)\xi_j\right). \tag{57}$$

In words, c_{ij} is simply the probability that the alleles at loci i and j were inherited from different parents. In most models of recombination this will depend on the genomic distance between i and j such that c_{ij} will be close to zero when i and j are close, and then grow to $\frac{1}{2}$ when they are far apart.

5.3. Inferred interactions and underlying fitness

Turning around the concepts, Eq. (56) can be interpreted by a inference formula of epistatic fitness from genomic data:

$$f_{ij}^{*} = J_{ij}^{*} \cdot rc_{ij}, \tag{58}$$

where $*$ indicates the inferred value.

The parameter J_{ij}^{*} can be determined from data by DCA while the parameters r and c_{ij} have to be determined by other means. However, since the QLE phase is characterized by J_{ij} being small, or, alternatively, f_{ij} being smaller than rc_{ij}, we can make the simplifying assumption that $c_{ij} \approx \frac{1}{2}$ for all pairs of loci we consider. Formula (58) then says that underlying fitness parameters f_{ij} are proportional to inferred Ising parameters J_{ij}, where the proportionality is $r/2$.

Nevertheless, Eq. (58) will also work when the variation of c_{ij} is taken into account, as long as the product rc_{ij} remains smaller than f_{ij}. It is not currently clear if there also exists an extension of Eq. (58) which also holds when rc_{ij} is of the order of or larger than f_{ij}, including the case when i and j are close ("hitch-hiking mutations").

5.4. Fitness inference for synthetic Ising genomic data

We here describe results obtained from simulating a finite population using the FFPopSim software.[168] Partial results in the same direction were reported in Ref. [159]; more complete results, though not using the single-time versions of algorithms as we will here, in Ref. [169].

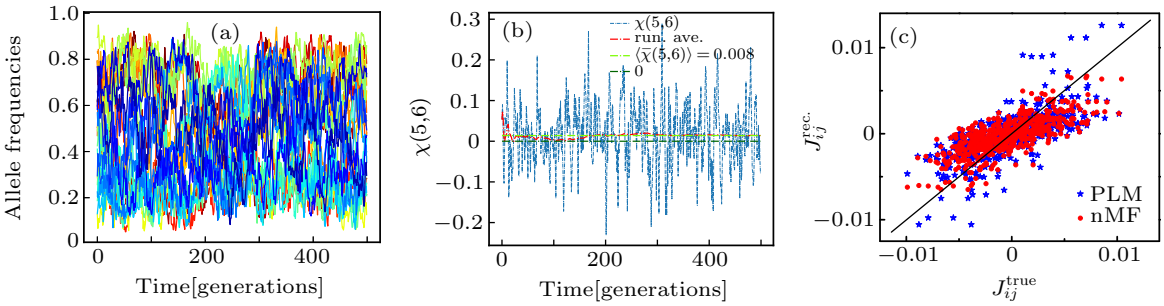


Fig. 8. (a) Temporal behavior of all allele frequencies defined as $f_i[1]$. Data recorded every 5 generations. (b) An example of pairwise correlation changing with time. With finite population size, there exists strong fluctuations in the system. (c) Scatter plot for the reconstructed against the tested fitness with DCA-nMF (red dots) and DCA-PLM (blue dots) algorithm for J_{ij}'s. Parameters: the number of loci $L = 25$, the number of individuals $N = 200$, mutation rate $\mu = 0.01$, recombination rate $r = 0.1$, crossover rate $\rho = 0.5$, standard deviation of epistatic fitness $\sigma = 0.002$.

In a finite population, statistical genetics as described above only holds on the average. When following one population in time, fluctuations of order $N^{-\frac{1}{2}}$ appear for observables such as single-locus frequencies and pair-wise loci-loci correlations. Figures 8(a) and 8(b) show the simulations using the FFPopSim software for allele frequencies and a specified pair-wise loci-loci correlations that these fluctuations can in practice (in simulations) be quite large. Figure 8(c) presents

the reconstructed fitness by DCA-PLM (blue dots) and DCA-nMF (red dots) against the tested fitness. Both the methods exhibit clear trend along the diagonal direction though with fluctuations.

The inference of fitness is governed by a set of parameters during the population evolutionary process. The illustrated examples in simulation below contain some fixed parameters, which are the number of loci $L = 25$, the number of individuals $N = 500$ (to avoid the singularity of correlation matrices of single generation), the length of generations $T = 500 \times 5$, the crossover rate $\rho = 0.5$. The varied parameters are the mutation rate μ, the recombination rate r and the strength of fitness σ. In the following we discuss what one can observe by systematically varying these three parameters.

Furthermore, it is of interest to see how the KNS inference theory performs by averaging the results from singletime data. This means that we infer parameters from snapshots, and then average those inferred parameters over the time of the snapshot. The authors of Ref. [169] in contrast studied the inference using alltime versions of the data, where inference was performed only once. In Fig. 9, we show the phase diagrams of epistatic fitness inference. The color indicates the relative root-mean-square error of the fitness reconstruction, where lighter color means larger error. However, the mean square error ε as shown in Eq. (47) is used for consistence in the following scatter plots. Figure 9(a) shows the mutation rate μ versus the recombination rate r. Figure 9(b) shows the fitness strength σ versus r. Both of them have wide broad ranges of parameters where the KNS theory works well for the fitness recovery. The inference phase diagrams based on sigletime here are quite close to those presented in Ref. [47] using alltime.

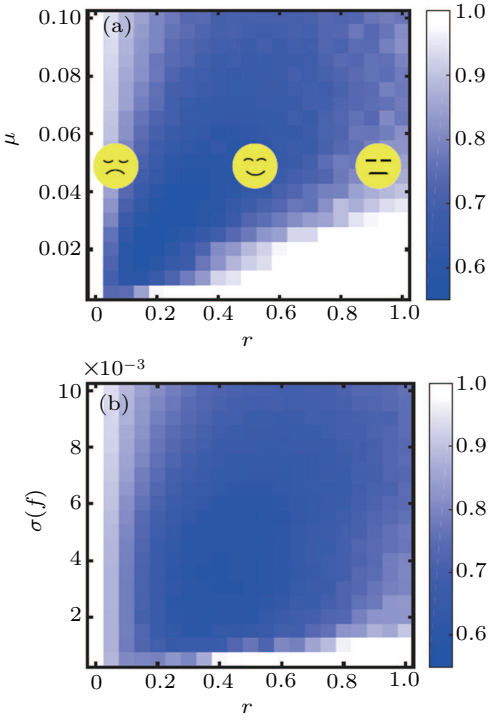


Fig. 9. Phase diagram for epistatic fitness recovery with DCA-nMF J_{ij}'s from the average of singletime data. (a) Mutation rate μ versus recombination rate r. For large recombination while low mutation, KNS inference does not work. However, for small r, the KNS inference theory does not satisfied. (b) Epistatic fitness strength σ with r. For large recombination and very small fitness, KNS inference does not work.

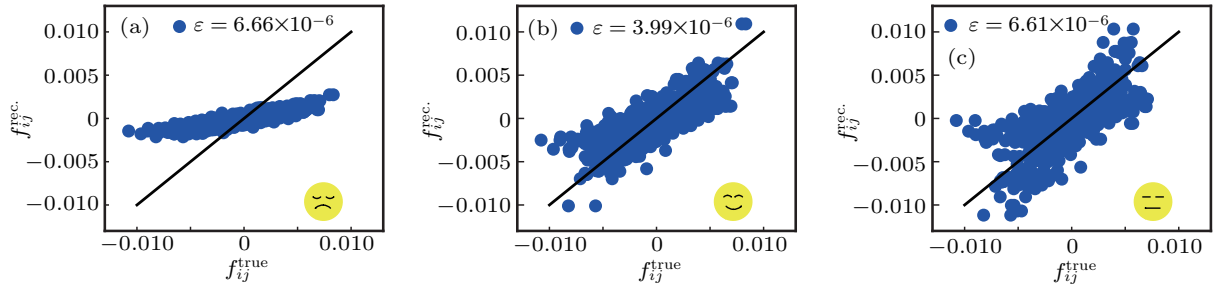


Fig. 10. Scatter-plots of inferred epistatic fitness against the true fitness based on the averaged results from singletime: (a) with sad face: $r = 0.1$, where the KNS theory cannot be satisfied here, (b) with smiling face: $r = 0.5$, where inference works, (c) with not-that-sad face: $r = 0.9$, where the KNS theory works but with very heavy fluctuations. Here the DCA-nMF algorithm for J_{ij}'s is utilized.

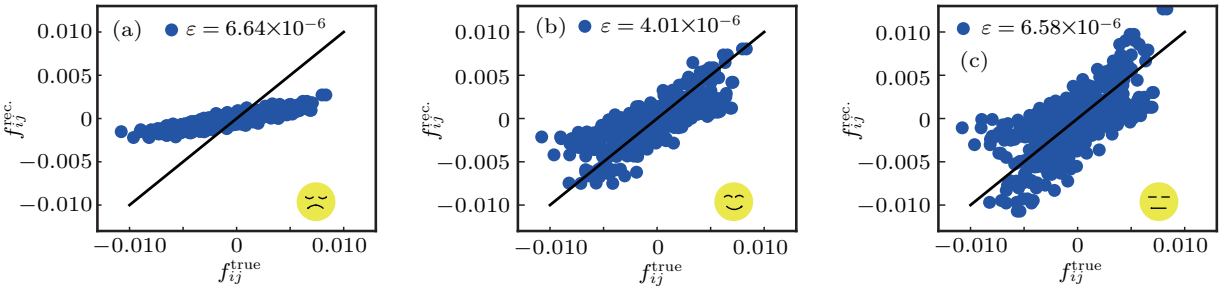


Fig. 11. Corresponding scatter plots by alltime averages with Fig. 10. The DCA-nMF algorithm for J_{ij}'s is also used here. The parameters for each sub-panel are the same as those in Fig. 10.

When mutation rates are very low, the frequencies of most loci is frozen to 0 or 1 for most of the time. This is a classical fact for evolution of one single locus, as discussed above, but also holds more generally. For an evolving population simulated with the FFPopSim software it was demonstrated in Ref. [169]. In this regime fitness recovery is hence impossible as there is not enough variation. On the other hand, the KNS inference theory does not hold for high enough μ, as one of the assumptions is that recombination is a faster process than mutations. Thus, three points on the μ–r phase diagram are picked with the same μ and differing r's, marked as sad face, smiling face and not-that-sad face, respectively. The corresponding scatter plots are presented in Fig. 10. As is expected, KNS inference works but with very heavy fluctuations for very high r, whereas does not hold for low r.

To see if there are differences between the inference with average over singletime and alltime, the corresponding scatter plots of Fig. 10 (singletime) are presented in Fig. 11 (alltime). With the parameters illustrated here, the difference between the two approaches is small.

6. Discussion and perspectives

Inverse Ising/Potts or DCA has emerged as a powerful paradigm of biological data analysis which has helped to revolutionize protein structure prediction. For the first time it has been shown to be possible to predict protein structure from sequence, though crucially from many similar sequences, not from a single one. The central idea which has made this possible is to exploit statistical dependencies encoded by a postulated Gibbs distribution (1) of the Ising/Potts form over sequence space. While DCA recently has been overtaken by more complex AI learning methods of the deep learning type, it remains the case that it was the success of DCA that showed this to be possible. Many other applications have appeared, some of them in areas where AI learning methods are not likely to succeed due to lack of training examples.

In this review we have striven to put these developments in the context of statistical physics. On some level a distribution over sequences must be arrived at by a evolutionary process, which, though it may be complicated, shares aspects of non-equilibrium spin dynamics. Indeed, these analogies have been noted for a long time, and have been explored from both the viewpoint of (theoretical) population genetics, and statistical physics. We have here added the dimension of learning, how knowledge of the type of dynamics and inference techniques can be used together to deduce biological parameters from data. We have also considered more direct applications of kinetic Ising models to model the evolution of neurons and of economic data, and how to infer connections from such data.

The main conclusions are as follows. First, we have stressed that dynamics that does not fulfill detailed balance can have practically arbitrarily complicated stationary states, even if interactions is only pair-wise. It can therefore not be the case that inverse Ising/Potts can generally give useful information: in the wrong parameter phase it is instead much more likely to yield garbage. The simplest example is inference in asynchronous kinetic Ising models discussed in Section 3: those models contain parameters (the anti-symmetric combination $J_{ij} - J_{ji}$) that are not simply present in the Ising distribution (1). DCA, by whichever algorithm, therefore will never be able to find them. Even more, the stationary distribution in such models is quite different from Eq. (1), and DCA is also not able to find the symmetric combination $J_{ij} + J_{ji}$ either (unless the anti-symmetric combination is relatively small). On the other hand, straight-forward methods relying on inference from time series are able to recover symmetric and anti-symmetric combinations equally easy. The moral of this part of our review is simple: if you have time series data, you should use it to infer dynamic models; it is both a more general and an easier procedure.

Second, we have considered evolutionary dynamics in finite populations under selection, mutations and recombination. Following the pioneering work of Kimura and more recently Neher and Shraiman, we discussed how the high-recombination regime leads to a distribution of the type (1), where the parameters can be inferred by DCA. We have noted that in the same high-recombination regime the effective interaction parameters are small, which corresponds to the high-temperature regime in inverse Ising. Hence inference in the high-recombination regime is limited by finite sample noise. Given finite data inference therefore works best in an intermediate regime, not too high recombination (because then statistical co-variance will be too weak), and not too low recombination (because then the Kimura–Neher–Shraiman theory does not apply). Crucially, we have observed that though the parameters inferred by DCA on such data are related to fitness, they are not the fitness parameters governing the evolutionary dynamics itself. The relation is albeit a simple proportionality, at least for pairs of loci far enough apart on the genome, but it is not an identity. The moral of this part of our review is thus: if you have a theory connecting the underlying mechanism which you want to clarify to the data which you can use, then you should be well advised to analyze the data using that theory.

Many open questions remain in the field of DCA, out of which we will but discuss some that are closely connected to the main thrust of our argument. The Kimura–Neher–Shraiman (KNS) theory is a huge step forward to an understanding of what is actually inferred for such a procedure, but is obviously only a first step. Most directly, both Figs. 10(a)

and 11(a) strongly suggest a functional relationship. These plots were obtained in parameter regions where the KNS theory cannot be expected to be valid, and indeed it is not: the mean square error of inferred and underlying fitness parameters is large. Since the plots suggest a functional relationship, there should however be another theory, which at this point is unknown. In other words, KNS is not the end, but should be the starting point for developing theories connecting fitness (and other evolutionary parameters) to distributions over sequences in much wider settings, and ways to learn such parameters from sequence data. In particular, since the KNS theory is only valid when fitness parameters f_{ij} are smaller than compounded recombination parameters rc_{ij}, KNS is likely not valid for the very strongest epistatic effects which are potentially the most interesting and biologically relevant.

Much work further deserves to be performed to incorporate further biological realism in KNS and/or its successor theories and software. Among the many important effects (most discussed above) which have not been taken into account in this review we list:

Multi-allele loci

Realistic mutation matrices that vary over a genome and depending on the transitions

Mutations that do not act on single loci, i.e., insertions and deletions (indels)

Other models of fitness and other distributions of, e.g., pair-wise parameters f_{ij}

More realistic models of recombination incorporating also recombination hotspots

More types of recombination, as appropriate for bacterial evolution

Effects of population growth and bottle-necks.

Many kinds of simulation software has been developed in the computational biology community, for instance, the fwdpp[170] software suitably used recently in Ref. [171]. To objective would not be to redo or replace such software packages, but to reuse them in the context of theory-driven inference.

One further direction important to pursue is the effect of spatial and environmental separation, believed to be a main mechanism behind speciation and the emergence of biological variation in general. Its effects in models of the Wright–Fisher–Moran type were emphasized in Ref. [144]. Spatial separation would in general tend to counter-act recombination, in that individuals which could recombine if they would meet actually are not likely to meet. For instance, a bacterium with one of highest known recombination rates is the human pathogen Helicobacter pylori (the cause of stomach ulcers), but two such bacteria actually can only recombine when they find themselves in the stomach of the same host. Strains of H. pylori can thus be distinguished on a global scale, and only merge when their human host populations overlap.[172]

Acknowledgments

The material in Sections 2, 3, and 4 is partly based on the PhD work of one of us (HLZ, Aalto University, Finland, 2014). We thank colleagues and collaborators during that time, particularly Mikko Alava, John Hertz, Hamed Mahmoudi, Matteo Marsili and Yasser Roudi. The material in Section 5 was motivated by many discussions in the statistical physics community for which we thank Johannes Berg, Simona Cocco, Bert Kappen, David Lacoste, Andrey Lokhov, Rémi Monasson, Roberto Mulet, Andrea Pagnani, Luca Peliti, Federico Ricci-Tersenghi, Chris Sander, Alex Schug, Erik Van Nimwegen, Martin Weigt, Riccardo Zecchina, and many others. For the actual development of the material in Section 5, one of us (EA) is much indebted to a discussion with Boris Shraiman, and collaborations and/or discussions with Magnus Ekeberg, Yueheng Lan, Cecilia Lövkvist, Martin Weigt, Marcin Skwark, Andrea Pagnani, Christoph Feinauer, Hai-Jun Zhou, Chen-Yi Gao, Angelo Vulpiani, Fabio Cecconi, Timo Koski, Arne Elofsson, Daniel Falush, Kaisa Thorell, Yoshiyuke Kabashima, Jukka Corander, Santeri Puranen, Yingying Xu, Alexander Mozeika, Rémi Lemoy and Onur Dikmen. The material in Section 5.4 is the outcome of ongoing collaboration with Simona Cocco, Eugenio Mauro and Rémi Monasson whom we thank for many fruitful discussions and suggestions on various stages of the work. For this part we also thank Richard Neher for the FFPopSim software package, and for valuable comments. We finally thank Johannes Berg and Andrey Lokhov for constructive comments on the manuscript.

References

[1] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and beyond: An Introduction to the Replica Method and Its Applications* (Singapore: World Scientific)
[2] Fischer K and Hertz J A 1991 *Spin Glasses* (Cambridge: Cambridge University Press)
[3] Mézard M and Montanari A 2009 *Information, Physics, and Computation* (Oxford: Oxford University Press)
[4] Schneidman E, Berry M J, Segev R and Bialek W 2006 *Nature* **440** 1007
[5] Roudi Y, Aurell E and Hertz J A 2009 *Front. Comput. Neurosci.* **3** 22
[6] Nguyen H C, Zecchina R and Berg J 2017 *Adv. Phys.* **66** 197
[7] Wainwright M J and Jordan M I 2007 *Found. Trends Mach. Learn.* **1** 1
[8] Darmois G 1935 *C. R. Acad. Sci. Paris* **200** 1265
[9] Koopman B 1936 *Trans. Am. Math. Soc.* **39** 399
[10] Jaynes E T 1957 *Phys. Rev.* **106** 620
[11] Aurell E 2016 *PLOS Comput. Biol.* **12** 1
[12] van Nimwegen E 2016 *PLOS Comput. Biol.* **12** e1004726
[13] Amari S I, Barndorff-Nielsen O E, Kass R E, Lauritzen S L and Rao C R 1987 *Differential Geometry in Statistical Inference* in *Lecture Notes–Monograph Series* Vol. 10 Chap. 2
[14] Amari S I and Nagaoka H 2000 *Methods Inf. Geom. Translations Mathematical Monographs V. 191* (New York: American Mathematical Society)
[15] Kampen N V 2007 *Stochastic Processes in Physics and Chemistry* 3rd edn (Amerstamdam: Elsevier)
[16] Kuramoto Y 1984 *Chemical Oscillations, Waves, and Turbulence* (Springer Series in Synergistics) (Berlin: Springer)

[17] Goldbeter A 2010 *Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour* (Cambridge: Cambridge University Press)

[18] Papadimitriou C H 1991 *Proceedings 32nd Annual Symposium of Foundations of Computer Science* October 1–4, 1991 San Juan, Puerto Rico, USA, pp. 163–169

[19] Selman B, Kautz H and Cohen B 1996 *Local Search Strategies for Satisfiability Testing* in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* Vol. 26

[20] Barthel W, Hartmann A K and Weigt M 2003 *Phys. Rev. E* **67** 066104

[21] Aurell E, Gordon U and Kirkpatrick S 2004 *Comparing Beliefs, Surveys and Random Walks* in *Neural Information Processing Systems 2004* (Vancouver, Canada)

[22] Seitz S, Alava M and Orponen P 2005 *J. Stat. Mech.: Theory Exp.* **2005** P06006

[23] Alava M, Ardelius J, Aurell E, Kaski P, Krishnamurthy S, Orponen P and Seitz S 2008 *Proc. Natl. Acad. Sci. USA* **105** 15253

[24] Kautz H and Selman B 2007 *Disc. Appl. Math.* **155** 1514

[25] Lemoy R, Alava M and Aurell E 2015 *Phys. Rev. E* **91** 013305

[26] Aurell E, Domínguez E, Machado D and Mulet R 2019 *Phys. Rev. Lett.* **123** 230602

[27] Kree R and Zippelius A 1987 *Phys. Rev. A* **36** 4421

[28] Braunstein A, Ramezanpour A, Zecchina R and Zhang P 2011 *Phys. Rev. E* **83** 056114

[29] Nguyen H C and Berg J 2012 *Phys. Rev. Lett.* **109** 050602

[30] Bertini L, De Sole A, Gabrielli D, Jona-Lasinio G and Landim C 2002 *J. Stat. Phys.* **107** 635

[31] Derrida B 2007 *J. Stat. Mech.: Theory Exp.* **2007** P07023

[32] Dettmer S L, Nguyen H C and Berg J 2016 *Phys. Rev. E* **94** 052116

[33] Berg J 2017 *J. Stat. Mech.: Theory Exp.* **2017** 083402

[34] Dettmer S L and Berg J 2018 *J. Stat. Mech.: Theory Exp.* **2018** 023403

[35] Tkačik G, Marre O, Amodei D, Schneidman E, Bialek W and Berry Michael J n 2014 *PLOS Comput. Biol.* **10** e1003408

[36] Cocco S, Feinauer C, Figliuzzi M, Monasson R and Weigt M 2018 *Rep. Prog. Phys.* **81** 032601

[37] Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L675

[38] Roudi Y, Tyrcha J and Hertz J 2009 *Phys. Rev. E* **79** 051915

[39] Ackley D H, Hinton G E and Sejnowski T J 1985 *Cogn. Sci.* **9** 147

[40] Kappen H and Rodríguez F B 1998 *Neural Comput.* **10** 1137

[41] Thouless D J, Anderson P W and Palmer R G 1977 *Philos. Mag.* **35** 593

[42] Mézard M and Mora T 2009 *J. Physiol.-Paris* **103** 107

[43] Weigt M, White R A, Szurmant H, Hoch J A and Hwa T 2009 *Proc. Natl. Acad. Sci. USA* **106** 67

[44] Ricci-Tersenghi F 2012 *J. Stat. Mech.: Theory Exp.* **2012** P08015

[45] Besag J 1975 *J. R. Stat. Soc. D* **24** 179

[46] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D S, Sander C, Zecchina R, Onuchic J N, Hwa T and Weigt M 2011 *Proc. Natl. Acad. Sci. USA* **108** E1293

[47] Marks D S, Colwell L J, Sheridan R, Hopf T A, Pagnani A, Zecchina R and Sander C 2011 *PLoS ONE* **6** e28766

[48] Jones D T, Buchan D W A, Cozzetto D and Pontil M 2012 *Bioinformatics* **28** 184

[49] Andreatta M, Laplagne S, Li S C and Smale S 2014 arXiv 1311.1301v3

[50] Kamisetty H, Ovchinnikov S and Baker D 2013 *Proc. Natl. Acad. Sci.* **110** 15674

[51] Ekeberg M, Lövkvist C, Lan Y, Weigt M and Aurell E 2013 *Phys. Rev. E* **87** 012707

[52] Ekeberg M, Hartonen T and Aurell E 2014 *J. Comput. Phys.* **276** 341

[53] Ravikumar P, Wainwright M J and Lafferty J D 2010 *Ann. Stat.* **38** 1287

[54] Bento J and Montanari A 2009 *Proceedings of the 22nd International Conference on Neural Information Processing Systems* NIPS-09 (Red Hook, NY, USA: Curran Associates Inc.) pp. 1303–1311

[55] Lokhov A Y, Vuffray M, Misra S and Chertkov M 2018 *Sci. Adv.* **4** e1700791

[56] Santhanam N P and Wainwright M J 2012 *IEEE Trans. Inf. Theory* **58** 4117

[57] Vuffray M, Misra S, Lokhov A and Chertkov M 2016 *Advances in Neural Information Processing Systems 29* edited by Lee D D, Sugiyama M, Luxburg U V, Guyon I and Garnett R (Curran Associates, Inc.) pp. 2595–2603

[58] Goel S, Kane D M and Klivans A R 2019 *Proceedings of the Thirty-Second Conference on Learning Theory* in *Proceedings of Machine Learning Research* edited by Beygelzimer A and Hsu D (Phoenix, USA: PMLR) Vol. 99 pp. 1449–1469

[59] Vuffray M, Misra S and Lokhov A Y 2019 *Efficient learning of discrete graphical models* arXiv:1902.00600

[60] Xu Y, Aurell E, Corander J and Kabashima Y 2017 arXiv: 1704.01459[physics.data-an]

[61] Xu Y, Puranen S, Corander J and Kabashima Y 2018 *Phys. Rev. E* **97** 062112

[62] Stein R R, Marks D S and Sander C 2015 *PLOS Comput. Biol.* **11** e1004182

[63] Glauber R J 1963 *J. Math. Phys.* **4** 294

[64] Suzuki M and Kubo R 1968 *J. Phys. Soc. Jpn.* **24** 51

[65] Gillespie D T 1977 *J. Phys. Chem.* **81** 2340

[66] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792

[67] Crisanti A and Sompolinsky H 1987 *Phys. Rev. A* **36** 4922

[68] Dijkstra E W 1974 *Commun. ACM* **17** 643

[69] Aurell E 2013 *J. Phys. Conf.* **473** 012017

[70] Decelle A and Zhang P 2015 *Phys. Rev. E* **91** 052136

[71] Shlens J, Field G, Gauthier J, Grivich M, Petrusca D, Sher A, Litke A and Chichilnisky E 2006 *J. Neurosci.* **26** 8254

[72] Cocco S, Leibler S and Monasson R 2009 *Proc. Natl. Acad. Sci. USA* **106** 14058

[73] Roudi Y and Hertz J 2011 *Phys. Rev. Lett.* **106** 048702

[74] Roudi Y and Hertz J 2011 *J. Stat. Mech.: Theory Exp.* **2011** P03031

[75] Mézard M and Sakellariou J 2011 *J. Stat. Mech.: Theory Exp.* **2011** L07001

[76] Zhang P 2012 *J. Stat. Phys.* **148** 502

[77] Kappen H J and Spanjers J J 2000 *Phys. Rev. E* **61** 5658

[78] Aurell E and Mahmoudi H 2012 *Phys. Rev. E* **85** 031119

[79] Pillow J W, Shlens J, Paninski L, Sher A, Litke A M, Chichilnisky E and Simoncelli E P 2008 *Nature* **454** 995

[80] Mastromatteo I and Marsili M 2011 *J. Stat. Mech.: Theory Exp.* **2011** P10012

[81] Zeng H L, Aurell E, Alava M and Mahmoudi H 2011 *Phys. Rev. E* **83** 041135

[82] Zeng H L, Alava M, Aurell E, Hertz J and Roudi Y 2013 *Phys. Rev. Lett.* **110** 210601

[83] Kipnis C and Landim C 1999 *Scaling Limits of Interacting Particle Systems* (Berlin: Springer-Verlag) Vol. 320

[84] Wainwright M J, Ravikumar P and Lafferty J D 2007 *Adv. Neural. Inf. Process. Syst.* **19** 1465

[85] Zeng H L, Hertz J and Roudi Y 2014 *Phys. Scr.* **89** 105002

[86] Zeng H L, Lemoy R and Alava M 2014 *J. Stat. Mech.: Theory Exp.* **2014** P07008

[87] Roudi Y, Dunn B and Hertz J 2015 *Curr. Opin. Neurobiology* **32** 38

[88] Cocco S, Monasson R, Posani L and Tavoni G 2017 *Curr. Opin. Struct. Biol.* **3** 103

[89] Huang H 2017 *J. Stat. Mech.: Theory Exp.* **2017** 033501

[90] Poli D, Pastore V P, Martinoia S and Massobrio P 2016 *J. Neural. Eng.* **13** 026023

[91] Latimer K W, Rieke F and Pillow J W 2019 *ELife* **8** e47012

[92] Sadeghi K and Berry M J 2020 *BioRxiv*

[93] Hoang D T, Song J, Periwal V and Jo J 2019 *Phys. Rev. E* **99** 023311

[94] Bacry E, Mastromatteo I and Muzy J F 2015 *Market Microstruct. Liquidity* **01** 1550005

[95] Ma J, Wang L and Wang T 2015 *The 27th Chinese Control and Decision Conference (2015 CCDC)* (IEEE) May 23–25, 2015, Qingdao, China, pp. 238–243

[96] Borysov S S, Roudi Y and Balatsky A V 2015 *Eur. Phys. J. B* **88** 321

[97] Zarinelli E, Treccani M, Farmer J D and Lillo F 2015 *Market Microstruct. Liquidity* **01** 1550004

[98] Li S, He J and Song K 2016 *Entropy* **18** 331

[99] Fan Y, Yu G, He Z, Yu H, Bai R, Yang L and Wu D 2017 *Entropy* **19** 51

[100] Zhao L, Bao W and Li W 2018 *J. Phys. Conf.* **1113** 012009

[101] Becker A P 2018 *Maximum entropy and network approaches to systemic risk and foreign exchange*, PhD thesis (Boston: Boston University)

[102] Alossaimy A N M and Stemler T 2019 *Using Complex Networks to Uncover Interaction in Stock Markets*, PhD thesis (The University of Western Australia)

[103] Bucci F, Benzaquen M, Lillo F and Bouchaud J P 2019 arXiv:1901.05332
[104] Hoffmann T, Peel L, Lambiotte R and Jones N S 2020 *Sci. Adv.* **6** eaav1478
[105] Ikeda Y and Takeda H 2020 arXiv:2001.04097
[106] Segev R, Puchalla J and Berry M J 2006 *J. Neurophysiol.* **95** 2277
[107] Mastromatteo I and Marsili M 2011 *J. Stat. Mech.: Theory Exp.* **2011** P10012
[108] Tanaka T 2000 *Neural Comput.* **12** 1951
[109] Bury T 2013 *Eur. Phys. J. B* **86** 1
[110] Bouchaud J P and Potters M 2003 *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management* (Cambridge: Cambridge University Press)
[111] Mantegna R N and Stanley H E 2003 *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge: Cambridge University Press)
[112] Biely C and Thurner S 2008 *Quant. Finance* **8** 705
[113] Zeng H L 2014 *Connectivity Inference with Asynchronously Updated Kinetic Ising Models*, PhD thesis (Finland: Aalto University)
[114] Kullmann L, Kertész J and Kaski K 2002 *Phys. Rev. E* **66** 026125
[115] Mann J K, Barton J P,Ferguson A L,Omarjee S, Walker B D, Chakraborty A and Ndung'u T 2014 *PLoS Comput. Biol.* **10** e1003776
[116] Phillips P C 2008 *Nat. Rev. Genet.* **9** 855
[117] Gueudré T, Baldassi C, Zamparo M, Weigt M and Pagnani A 2016 *Proc. Natl. Acad. Sci. USA* **113** 12186
[118] Uguzzoni G, John Lovis S, Oteri F, Schug A, Szurmant H and Weigt M 2017 *Proc. Natl. Acad. Sci. USA* **114** E2662–E2671
[119] Pfam 32.0 2018 http://pfam.xfam.org/
[120] El-Gebali S, Mistry J, Bateman A, Eddy S R, Luciani A, Potter S C, Qureshi M, Richardson L J, Salazar G A, Smart A, Sonnhammer E L, Hirsh L, Paladin L, Piovesan D, Tosatto S C and Finn R D 2019 *Nucleic Acids Res.* **47** D427
[121] Ovchinnikov S, Park H, Varghese N, Huang P S, Pavlopoulos G A, Kim D E, Kamisetty H, Kyrpides N C and Baker D 2017 *Science* **355** 294
[122] Michel M, Menéndez Hurtado D, Uziela K and Elofsson A 2017 *Bioinformatics* **33** i23
[123] Ovchinnikov S, Park H, Kim D E, DiMaio F and Baker D 2018 *Proteins* **86** 113
[124] Senior A W, Jumper J, Hassabis D and Kohli P 2020 *AlphaFold: Using AI for Scientific Discovery*
[125] Senior A W, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson A W, Bridgland A *et al.* 2020 *Nature* **577** 706
[126] De Leonardis E, Lutz B, Ratz S, Simona C, Monasson R, Weigt M and Schug A 2015 *Nucleic Acids Res.* **43** 10444
[127] Weinreb C, Riesselman A J, Ingraham J B, Gross T, Sander C and Marks D S 2016 *Cell* **165** 963
[128] Ferguson A L, Mann J K, Omarjee S, Ndung'u T, Walker B D and Chakraborty A K 2013 *Immunity* **38** 606
[129] Shekhar K, Ruberman C F, Ferguson A L, Barton J P, Kardar M and Chakraborty A K 2013 *Phys. Rev. E* **88** 062705
[130] Louie R H Y, Kaczorowski K J, Barton J P, Chakraborty A K and McKay M R 2018 *Proc. Natl. Acad. Sci. USA* **115** E564
[131] Figliuzzi M, Jacquier H, Schug A, Tenaillon O and Weigt M 2016 *Mol. Biol. Evol.* **33** 268
[132] Hopf T A, Ingraham J B, Poelwijk F J, Scharfe C P I, Springer M, Sander C and Marks D S 2017 *Nat. Biotechnol.* **35** 128
[133] Couce A, Caudwell L V, Feinauer C, Hindré T, Feugeas J P, Weigt M, Lenski R E, Schneider D and Tenaillon O 2017 *Proc. Natl. Acad. Sci. USA* **114** E9026–E9035
[134] Skwark M J, Croucher N J, Puranen S, Chewapreecha C, Pesonen M, Xu Y Y, Turner P, Harris S R, Beres S B, Musser J M, Parkhill J, Bentley S D, Aurell E and Corander J 2017 *PLoS Genet.* **13** e1006508
[135] Schubert B, Maddamsetti R, Nyman J, Farhat M R and Marks D S 2018 *BioRxiv* 325993
[136] Puranen S, Pesonen M, Pensar J, Xu Y Y, Lees J A, Bentley S D, Croucher N J and Corander J 2018 *Microb. Genom.* **4**
[137] Gao C Y, Zhou H J and Aurell E 2018 *Phys. Rev. E* **98** 032407
[138] Pensar J, Puranen S, Arnold B, MacAlasdair N, Kuronen J, Tonkin-Hill G, Pesonen M, Xu Y, Sipola A, Sánchez-Busó L, Lees J A, Chewapreecha C, Bentley S D, Harris S R, Parkhill J, Croucher N J and Corander J 2019 *Nucleic Acids Res.* **47** e112
[139] Hakenbeck R, Brückner R, Denapaite D and Maurer P 2012 *Future Microbiol.* **7** 395
[140] Fisher R A 1922 *P. Roy. Soc. Edinb.* **42** 321
[141] Fisher R 1930 *The Genetical Theory of Natural Selection* (Oxford: Clarendon)
[142] Kolmogorov A N 1935 *Dokl. Akad. Nauk. SSSR* **3** 129
[143] Peliti L 1997 arXiv:cond-mat/9712027
[144] Blythe R A and McKane A J 2007 *J. Stat. Mech.: Theory Exp.* **2007** P07018
[145] Chaguza C, Andam C P, Harris S R *et al.* 2016 *MBio* **7** e01053
[146] Eigen M 1971 *Naturwissenschaften* **58** 465
[147] Eigen M 2002 *Proc. Natl. Acad. Sci. USA* **99** 13374
[148] Maynard Smith J 1982 *Evolution and the Theory of Games* (Cambridge: Cambridge University Press)
[149] Nowak M A and Sigmund K 2004 *Science* **303** 793
[150] Claussen J C and Traulsen A 2008 *Phys. Rev. Lett.* **100** 058104
[151] Wang Z, Xu B and Zhou H J 2015 *Sci. Rep.* **4** 5830
[152] Liao M J, Din M O, Tsimring L and Hasty J 2019 *Science* **365** 1045
[153] Shahshahani S 1979 *A New Mathematical Framework for the Study of Linkage and Selection* (New York: American Mathematical Society)
[154] Bürger R 2000 *The Mathematical Theory of Selection, Recombination, and Mutations* (New York: Wiley) Vol. 228
[155] Svirezhev Y and Passekov V 2012 *Fundamentals of Mathematical Evolutionary Genetics* (Berlin: Springer) Vol. 22
[156] Huillet T E 2017 *J. Stat. Phys.* **168** 15
[157] Aurell E, Ekeberg M and Koski T 2019 arXiv:1906.00716
[158] Neher R A and Shraiman B I 2011 *Rev. Mod. Phys.* **83** 1283
[159] Gao C Y, Cecconi F, Vulpiani A, Zhou H J and Aurell E 2019 *Phys. Biol.* **16** 026002
[160] Yahara K, Didelot X, Ansari M A, Sheppard S K and Falush D 2014 *Mol. Biol. Evol.* **31** 1593
[161] Chewapreecha C, Harris S R, Croucher N J *et al.* 2014 *Nat. Genet.* **46** 305
[162] Kimura M 1956 *Evolution* **10** 278
[163] Kimura M 1964 *J. Appl. Probab.* **1** 177
[164] Kimura M 1965 *Genetics* **52** 875
[165] Neher R A and Shraiman B I 2009 *Proc. Natl. Acad. Sci. USA* **106** 6866
[166] Hardy G H, *et al.* 1908 *Science* **28** 49
[167] Weinberg W 1908 *Jahresh. Ver. Vaterl. Naturkd. Württemb* **64** 368
[168] Neher R and Zanini F 2012 *FFPopSim*
[169] Zeng H L and Aurell E 2020 *Phys. Rev. E* **101** 052409
[170] Thornton K R 2014 *Genetics* **198** 157
[171] Arnold B, Sohail M, Wadsworth C, Corander J, Hanage W P, Sunyaev S and Grad Y H 2020 *Mol. Biol. Evol.* **37** 417
[172] Thorell K, Yahara K, Berthenet E *et al.* 2017 *PLoS Genet.* **13** e1006730