HHS Public Access

Author manuscript

Phys Rev E. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

Phys Rev E. 2019 February ; 99(2-1): 023311. doi:10.1103/PhysRevE.99.023311.

Network inference in stochastic systems from neurons to currencies: Improved performance at small sample size

Danh-Tai Hoang[1,2], **Juyong Song**[3,4,5], **Vipul Periwal**[1,*], **Junghyo Jo**[6,7,†]

[1]Laboratory of Biological Modeling, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

[2]Department of Natural Sciences, Quang Binh University, Dong Hoi, Quang Binh 510000, Vietnam

[3]Asia Pacific Center for Theoretical Physics, Pohang, Gyeongbuk 37673, Korea

[4]Department of Physics, Pohang University of Science and Technology, Pohang, Gyeongbuk 37673, Korea

[5]Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34014 Trieste, Italy

[6]School of Computational Sciences, Korea Institute for Advanced Study, Seoul 02455, Korea

[7]Department of Statistics, Keimyung University, Daegu 42601, Korea

Abstract

The fundamental problem in modeling complex phenomena such as human perception using probabilistic methods is that of deducing a stochastic model of interactions between the constituents of a system from observed configurations. Even in this era of big data, the complexity of the systems being modeled implies that inference methods must be effective in the difficult regimes of small sample sizes and large coupling variability. Thus, model inference by means of minimization of a cost function requires additional assumptions such as sparsity of interactions to avoid overfitting. In this paper, we completely divorce iterative model updates from the value of a cost function quantifying goodness of fit. This separation enables the use of goodness of fit as a natural rationale for terminating model updates, thereby avoiding overfitting. We do this within the mathematical formalism of statistical physics by defining a formal free energy of observations from a partition function with an energy function chosen precisely to enable an iterative model update. Minimizing this free energy, we demonstrate coupling strength inference in nonequilibrium kinetic Ising models, and show that our method outperforms other existing methods in the regimes of interest. Our method has no tunable learning rate, scales to large system sizes, and has a systematic expansion to obtain higher-order interactions. As applications, we infer a functional connectivity network in the salamander retina and a currency exchange rate network from time-series data of neuronal spiking and currency exchange rates, respectively. Accurate small sample size inference is critical for devising a profitable currency hedging strategy.

* vipulp@mail.nih.gov. † jojunghyo@kmu.ac.kr.

I. INTRODUCTION

An explosion in data availability in recent years has ushered in a new era of data-driven research for natural and social sciences. Identifying systems dynamics from observed data, e.g., biochemical reactions [1], gene expression measurements [2], neuronal or brain region activities [3–6], and population dynamics [7], is of fundamental interest in science [8–12]. For complex phenomena, such as human perception, modeling system dynamics in a probabilistic framework became possible with the advent of inexpensive computational resources, and has led to great progress in the last 25 years. Regardless of whether stochasticity is inherent in the system or only apparent due to partial observability [13], many stochastic processes have been analyzed by autoregressive-moving-average models [14] or probabilistic directed acyclic graphical models, often termed Bayesian networks [15].

The structure of such dynamic processes is often unknown and, in the social sciences in particular, there may be no underlying fundamental theory to delineate possible models. Thus, a data-driven approach has merit for the inference of models from time-series data [16]. Machine learning using recurrent neuronal networks is such an approach [17], but it usually requires a large amount of training data and is computationally intensive. Given time series of N variables, network inference rapidly becomes computationally demanding with increasing N. Even restricting to pair-wise interactions requires determining N^2 parameters and demands $L \geqslant N^2$ samples. Including higher-order interactions leads to an exponential increase in the number of model parameters, and a concomitant increase in required sample size. In scientific contexts, however, we often encounter the case that data generated from experiments are not big enough to reconstruct the interaction network for a given system. Theorists contend with the computational difficulties of inferring large systems by positing properties such as sparsity of interactions or specifying distributions of couplings, usually with scant experimental support.

Statistical physics is often used for model inference [18,19], but, in fact, for small sample sizes, the observed configurations of the system may bear no semblance to random sampling or a thermodynamic limit. We develop here an iterative parameter-free model estimator using only the mathematical formalism of statistical physics to define a free energy of data, and show that minimizing this free energy enables a systematic nonparametric model inference.

Over-fitting is a major problem in the analysis of underdetermined systems. Cross validation splits the observation into a training set and a testing set, e.g., in a $(k-1)$-to-1 proportion, namely, k-fold cross validation, for model training and validating, respectively [20]. However, for small sample sizes, it is imperative to avoid further reductions in the data available for training. Approaches such as LASSO [21] and Ridge regression [22,23] add a penalty for nonzero coupling strengths. Regularization terms have been widely applied for inference of sparse networks [24,25]. Here, by decoupling an iterative multiplicative model update step from any cost function minimization, we are free to use the likelihood or any other measure of discrepancy between observation and model expectation as a stopping criterion, so that we can use the entire data set for model inference.

In Sec. I, we explain the theory underlying our approach. We demonstrate that our free energy minimization (FEM) approach infers coupling strengths in nonequilibrium kinetic Ising models, outperforming previous approaches particularly in the large coupling variability and small sample size regimes in Sec. III. Real data are always a stringent test of model inference, so we demonstrate applications of FEM to infer biological and financial networks from neuronal activities and currency fluctuations. Finally, we summarize the computational merits of FEM in Sec. IV. Some mathematical details are explained in the Supplemental Material [26]. We provide complete source code and documentation on GITHUB [27].

II. THEORY

The kinetic Ising model is commonly used as an illustrative example in stochastic model inference. In this model, the N-spin state $\sigma = (\sigma_1,\ldots,\sigma_N)$ at time $t+1$ is stochastically determined from the current state $\sigma(t)$ at time t with the following conditional probability:

$$P(\sigma_i(t+1)|\sigma(t)) = \frac{\exp[\sigma_i(t+1)H_i(\sigma(t))]}{\exp[H_i(\sigma(t))] + \exp[-H_i(\sigma(t))]}, \tag{1}$$

for $i = 1,\ldots,N$. The local field $H_i(\sigma(t))$ represents the influence of the present state $\sigma(t)$ on the future state $\sigma_i(t+1)$. Here, for ease of explanation, we focus on the simplest case $H_i(\sigma(t)) = \sum_j W_{ij}\sigma_j(t)$, with the aim to determine the weight matrix W_{ij}. Of course, $H_i(\sigma(t))$ could include higher-order interactions of $\sigma(t)$ in general, and we show later that the formalism extends to this case with no change. The state $\sigma_i(t+1)$ tends to align with the local field $H_i(\sigma(t))$, so the model expectation defined by

$$\langle\sigma_i(t+1)\rangle_{H_i(\sigma(t))} \equiv \sum_{\rho=\pm 1} \rho\, P(\sigma_i(t+1) = \rho|\sigma(t)) \tag{2}$$

is just $\langle\sigma_i(t+1)\rangle_{H_i(\sigma(t))} = \tanh[H_i(\sigma(t))]$. Our goal is to infer the coupling strength W_{ij} between variables $\sigma_i(t+1)$ and $\sigma_j(t)$ from time series data of $\{\sigma(t)\}_{t=1}^{L}$.

Notice that

$$\left|\frac{\langle\sigma_i(t+1)\rangle_{H_i(\sigma(t))}}{\sigma_i(t+1)}\right| = |\tanh[H_i(\sigma(t))]| \leqslant 1, \tag{3}$$

so, if we define an improved $H_i^{\text{new}}(\sigma(t))$ by

$$H_i^{\text{new}}(\sigma(t)) \leftarrow \frac{\sigma_i(t+1)}{\langle\sigma_i(t+1)\rangle_{H_i(\sigma(t))}} H_i(\sigma(t)), \tag{4}$$

then

$$\left|\langle\sigma_i(t+1)\rangle_{H_i^{\text{new}}(\sigma(t))}\right| \geqslant \left|\langle\sigma_i(t+1)\rangle_{H_i(\sigma(t))}\right|, \tag{5}$$

because $|H_i^{\text{new}}(\sigma(t))| \geqslant |H_i(\sigma(t))|$ and, therefore, $\left|\tanh\left[H_i^{\text{new}}(\sigma(t))\right]\right| \geqslant |\tanh[H_i(\sigma(t))]|$. Then, Eq. (5) means that the model prediction for $\sigma_i(t+1)$ is closer to ±1, and is therefore better. In fact, if the model prediction has the wrong sign, Eq. (4) will even correct the sign of $H_i^{\text{new}}(\sigma(t))$. An important point to note is that the suggested update, Eq. (4), appears to be multiplicative, rather than an incremental additive correction based on error gradients. In actual fact, the update is multiplicative for nonvanishing $H_i(\sigma(t))$ but because $x/\tanh x \to 1$ for $x \to 0$, the update becomes a shift if the local field vanishes, $H_i(\sigma(t)) = 0$, with $H_i^{\text{new}}(\sigma(t)) \leftarrow \sigma_i(t+1)$. In other words, the multiplicative update includes an inhomogeneous update which prevents the iteration from being trapped in an $H_i = 0$ state if it is inconsistent with $\sigma_i(t+1) \neq 0$. This obviously could not happen with a naive multiplicative update.

However, we are considering each t independently of any other if we update using Eq. (4), but the aim is to find the best *functional form* of $H_i(\sigma(t))$ that will determine the system for all t. For the linear example, $H_i(\sigma) = \sum_j W_{ij}\sigma_j$, it is not difficult to find the best W_{ij}^{new} directly from $H_i^{\text{new}}(\sigma(t)) = \sum_j W_{ij}^{\text{new}}\,\sigma_j(t)$ averaged in a principled way over all t:

$$\sum_t H_i^{\text{new}}(\sigma(t))\delta\sigma_k(t) = \sum_t \sum_j W_{ij}^{\text{new}}\,\sigma_j(t)\delta\sigma_k(t), \tag{6}$$

$$\sum_t \delta H_i^{\text{new}}(\sigma(t))\delta\sigma_k(t) = \sum_t \sum_j W_{ij}^{\text{new}}\,\delta\sigma_j(t)\delta\sigma_k(t), \tag{7}$$

$$\left\langle \delta H_i^{\text{new}}\delta\sigma_k \right\rangle = \sum_j W_{ij}^{\text{new}} \left\langle \delta\sigma_j\delta\sigma_k \right\rangle, \tag{8}$$

where we multiplied fluctuations of microstates $\delta\sigma_k(t) \equiv \sigma_k(t) - \langle\sigma_k\rangle$ on both sides. Note that the sample average is defined as $\langle f \rangle \equiv 1/L\sum_{t=1}^{L} f(t)$ and the sample average of fluctuations always vanishes with $\langle \delta t \rangle = 0$. Therefore, one can obtain

$$W_{ij}^{\text{new}} = \sum_k \left\langle \delta H_i^{\text{new}}\,\delta\sigma_k \right\rangle [C^{-1}]_{kj}, \tag{9}$$

by inverting the connected correlation matrix $C_{jk} \equiv \langle \delta\sigma_j\delta\sigma_k \rangle$ on the right-hand side of Eq. (8). The challenge now is to find the appropriate theoretical principles and framework for this kind of update that apply not just to this simple linear form of the local field H_i but also to all functional forms of H_i including higher-order terms.

For this, we turn to Schwinger's famous idea to use generating functions to provide a natural connection between expectation values $m = \langle\sigma\rangle$ of microstates σ and expectation values $\langle H_i^{\text{new}} \rangle_m$ of any observable H_i^{new} conditioned on m [28], which therefore gives the expectation value, $\langle H_i^{\text{new}} \rangle_m$, as a *function* of m. This is the foundation of modern approaches to field theory as described in textbooks, for example, [29]. We start by defining a moment generating function,

$$Z_i(J,\beta) = \sum_t \exp\left[J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t))\right], \tag{10}$$

which is a function of a vector parameter J, a scalar parameter β, and an observable $H_i^{\text{new}}(\sigma(t))$ of data $\sigma(t)$. A convex free energy $F_i \equiv \ln Z_i$ can be used to obtain expectation values of spin activities and observables by differentiation,

$$\frac{\partial F_i}{\partial J_j} = \frac{\sum_t \sigma_j(t) \exp\left[J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t))\right]}{\sum_t \exp\left[J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t))\right]} = \langle \sigma_j \rangle_J \equiv m_j(J), \tag{11}$$

$$\frac{\partial F_i}{\partial \beta} = -\frac{\sum_t H_i^{\text{new}}(\sigma(t)) \exp\left[J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t))\right]}{\sum_t \exp\left[J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t))\right]} = -\langle H_i^{\text{new}} \rangle_J . \tag{12}$$

As usual, a convex dual free energy G_i can be defined to make the expected activity vector m the independent variable, and $J(m)$ the dependent vector, by using the convexity preserving Legendre transform $F_i(J) + G_i(m) = J \cdot m$. By defining a normalized probability, $P(\sigma(t)) \equiv \exp[J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t)) - F_i]$ in Eq. (10), we can show that G_i can be indeed interpreted as a thermodynamic free energy,

$$G_i = \beta \langle H_i^{\text{new}} \rangle_J - S_i \tag{13}$$

with the expectation value of H_i^{new} taking the place of internal energy and the Shannon entropy of data, $S_i = -\sum_t P(\sigma(t)) \ln P(\sigma(t))$. At $\beta = 0$, minimizing the free energy G_i is exactly maximizing the entropy S_i.

The duality between the free energies F_i and G_i through their Legendre transform leads to

$$\frac{\partial G_i}{\partial m_j} = J_j, \tag{14}$$

$$\frac{\partial G_i}{\partial \beta} = -\frac{\partial F_i}{\partial \beta} = \langle H_i^{\text{new}} \rangle_m, \tag{15}$$

where we identify $\langle H_i^{\text{new}} \rangle_{J(m)} \equiv \langle H_i^{\text{new}} \rangle_m$. Therefore, once we know the free energy G_i, it is straightforward to obtain $\langle H_i^{\text{new}} \rangle_m$, the expectation value of observable H_i^{new} conditioned on the expectation value $m = \langle \sigma \rangle$ of microstates σ. For our purposes, however, it will not be necessary to obtain $G_i(m)$ for all possible values of m, as it will suffice to know its derivatives at its minimum for $\beta = 0$. The free energy G_i is minimized when $J(m^*) = \partial_m G(m^*) = 0$ from Eq. (15), which happens at the data expectation:

$$m^* \equiv \langle \sigma \rangle_{J=0} = \frac{1}{L} \sum_{t=1}^{L} \sigma(t) . \tag{16}$$

Therefore, this is the value of m about which we expand in a Taylor series, hence this method is termed free energy minimization. The Taylor expansion of $G_i(m)$ up to second-order terms at $m = m^*$ is

$$G_i(m) = G_i(m^*) + \frac{1}{2} \sum_{j,k} \left[\frac{\partial^2 G_i}{\partial m_j \partial m_k} \right]^* \left(m_j - m_j^* \right)\left(m_k - m_k^* \right), \tag{17}$$

where the derivatives $[\cdot]^*$ are taken at $m = m^*$. Differentiating the expanded $G_i(m)$ with respect to β leads to

$$\frac{\partial G_i(m)}{\partial \beta} = \frac{\partial G_i(m^*)}{\partial \beta} - \sum_{j,k} \frac{\partial m_k^*}{\partial \beta} \left[\frac{\partial^2 G_i}{\partial m_j \partial m_k} \right]^* \left(m_j - m_j^* \right). \tag{18}$$

Here, each derivative in Eq. (18) is calculated as follows:

$$-\frac{\partial m_k}{\partial \beta} = \frac{\partial}{\partial \beta} \left[\frac{\sum_t \sigma_k(t) \exp\left[\left(J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t)) \right) \right]}{\sum_t \exp\left[\left(J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t)) \right) \right]} \right] = \left\langle \delta H_i^{\text{new}} \, \delta \sigma_k \right\rangle, \tag{19}$$

and

$$\frac{\partial^2 G_i}{\partial m_j \partial m_k} = \frac{\partial J_k}{\partial m_j} = [C^{-1}]_{jk}, \tag{20}$$

where

$$C_{jk} = \frac{\partial m_j}{\partial J_k} = \frac{\partial}{\partial J_k} \left[\frac{\sum_t \sigma_j(t) \exp\left[\left(J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t)) \right) \right]}{\sum_t \exp\left[\left(J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t)) \right) \right]} \right] = \left\langle \delta \sigma_j \delta \sigma_k \right\rangle_m . \tag{21}$$

Here, we have used standard abbreviated notation: $\langle f \rangle^* \equiv \langle f \rangle_{J=0}$, and $\langle \delta f \rangle_m \equiv \langle f \rangle_m - \langle f \rangle^*$. Plugging in Eqs. (15), (19), and (20), we obtain

$$\left\langle \delta H_i^{\text{new}} \right\rangle_m = \sum_{j,k} \left\langle \delta H_i^{\text{new}} \delta \sigma_k \right\rangle^* \left[C^{-1} \right]_{kj}^* \left\langle \delta \sigma_j \right\rangle_m, \tag{22}$$

which is valid for *any* choice of observable H_i^{new}.

Now, if we take our observable H_i^{new} to be the right-hand side of Eq. (4), then an improved estimate of W_{ij} for the linear term in H_i is

$$W_{ij}^{\text{new}} \leftarrow \sum_k \left\langle \delta H_i^{\text{new}} \, \delta \sigma_k \right\rangle^* [C^{-1}]_{kj}^*, \tag{23}$$

exactly as suggested by Eq. (9). Moreover, higher-order contributions of σ_j to H_i are obtained simply by expanding to higher orders in the Taylor series in Eq. (17). For instance, when the interactions between variables contain not only linear terms but also quadratic terms, $H_i(\sigma(t)) = \sum_j W_{ij}\sigma_j(t) + \frac{1}{2}\sum_{j,k} Q_{ijk}\sigma_j(t)\sigma_k(t)$, the formalism gives

$$\begin{aligned} Q_{ijk}^{\text{new}} \leftarrow & \sum_{\mu,\nu} \left\langle \delta H_i^{\text{new}} \delta\sigma_\mu \sigma_\nu \right\rangle^* [C^{-1}]^*_{j\mu} [C^{-1}]^*_{k\nu} \\ & - \sum_l \sum_{\lambda,\mu,\nu} \left\langle \delta H_i^{\text{new}}\, \delta\sigma_l \right\rangle^* \left\langle \delta\sigma_\lambda \delta\sigma_\mu \sigma_\nu \right\rangle^* \times [C^{-1}]^*_{j\lambda} [C^{-1}]^*_{k\mu} [C^{-1}]^*_{l\nu}, \end{aligned} \tag{24}$$

and

$$W_{ij}^{\text{new}} \leftarrow \sum_k \left\{ \left\langle \delta H_i^{\text{new}}\, \delta\sigma_k \right\rangle^* [C^{-1}]^*_{kj} - Q_{ijk}^{\text{new}} \langle \sigma_k \rangle^* \right\} \tag{25}$$

(see Supplemental Material, Text I [26]). Therefore, with our choice of observables, H_i^{new}, the Schwinger formalism estimates improved W_{ij}^{new} and higher-order terms like Q_{ijk}^{new} from previous estimates.

Note that we have not made any use of a cost function in our update rule, which was based on the simple observation in Eq. (4). However, overfitting is a major problem for small sample size inference, so we now turn to the crucial issue of a stopping criterion for the update iteration in Eq. (23). We consider the overall discrepancy between the observed $\sigma_i(t+1)$ and the model prediction $\langle \sigma_i(t+1) \rangle_{H_i(\sigma(t))}$:

$$D_i \equiv \sum_t \left[\sigma_i(t+1) - \langle \sigma_i(t+1) \rangle_{H_i(\sigma(t))} \right]^2. \tag{26}$$

Clearly, the parameter update of $W_{ij}, Q_{ijk}, \ldots$ through Eqs. (22), (24), and (25) is completely independent of the computation of D_i. As D_i can be rewritten as

$$D_i \equiv \sum_t \left[1 - \frac{\langle \sigma_i(t+1) \rangle_{H_i(\sigma(t))}}{\sigma_i(t+1)} \right]^2, \tag{27}$$

we see that each term in D_i would be individually reduced by virtue of Eq. (5), consistent with Eq. (4), but clearly the *common* functional form of $H_i(\sigma)$ means that these are not all independent. Therefore, we stop the iteration when D_i starts to increase. This crucial decoupling between the stopping criterion and our multiplicative update is only possible because the update is completely independent of D_i.

To summarize the inference algorithm with FEM:

- **i.** Compute $H_i(\sigma(t)) = \sum_j W_{ij}\sigma_j(t)$ (initialize with a random W_{ij}).
- **ii.** Compute $H_i^{\text{new}}(\sigma(t))$ as the right-hand side of Eq. (4).
- **iii.** Update $W_{ij} = W_{ij}^{\text{new}} \leftarrow \sum_k \left\langle \delta H_i^{\text{new}} \delta\sigma_k \right\rangle^* [C^{-1}]^*_{kj}$.

iv. Repeat (i)–(iii) until D_i starts to increase.

v. Compute (i)–(iv) in parallel for every index $i \in \{1,2,\cdots,N\}$.

The algorithm is similarly applied to the model containing both linear terms W_{ij} and quadratic terms Q_{ijk} with Eqs. (25) and (24).

III. RESULTS

A. Kinetic Ising model

We first tested FEM on the inference of connection weights W_{ij} ($\neq W_{ji}$) in the kinetic Ising model, which is often used as a benchmark for stochastic causality inference. The Sherrington-Kirkpatrick (SK) model assumes W_{ij} are normally distributed with zero mean and variance equal to g^2/N [30]. In the limit of large sample size (large L/N^2), our iterative method decreases the mean-square error, $\mathrm{MSE} = N^{-2}\sum_{i,j=1}^{N}\left(W_{ij} - W_{ij}^{\mathrm{true}}\right)^2$, as the number of iterations increases [Fig. 1(a)]. We obtain good agreement between true and predicted weights [Fig. 1(b)]. In real world problems, W_{ij}^{true} is inaccessible so MSE cannot be defined. However, D_i in Eq. (26) is an alternative measure of the discrepancy between observation $\sigma_i(t+1)$ and model expectation. The discrepancy measures D_i are independent for each spin i. We checked that MSE and $D = N^{-1}\sum_{i=1}^{N} D_i$ change similarly during iterations. More importantly, for small sample sizes (small L/N^2), MSE and D decrease with iterations initially, but start to increase after some number of iterations [Fig. 1(c)]. For the kinetic Ising model, $D_i = 4\sum_t [1 - P(\sigma_i(t+1)|\sigma(t))]^2$ with the transition probability, $P(\sigma_i(t+1)|\sigma(t))$ in Eq. (1). Therefore, decreasing D_i can only result from $P(\sigma_i(t+1)|\sigma(t))$ saturating the causal relation between observations, $\sigma(t)$ and $\sigma_i(t+1)$, through W. Distinct spins indexed by i often require different numbers of iterations. Stopping the iteration for spin i when D_i saturates leads to accurate inference with minimal computation. For limited data (e.g., $L/N^2 = 0.2$), these stopping criteria lead to accurate inference [Fig. 1(d)] without overfitting.

Now we compare the inference performance of our method with other representative methods [31–33]: naïve mean field (nMF), Thouless-Anderson-Palmer mean field (TAP), exact mean field (eMF), and maximum likelihood estimation (MLE). MLE requires maximizing the data likelihood, $\mathscr{P} = \prod_{t=1}^{L-1}\prod_{i=1}^{N} P(\sigma_i(t+1)|\sigma(t))$, and uses gradient ascent to update Wij incrementally through $W_{ij}^{\mathrm{new}} = W_{ij} + \alpha/(L-1)\partial \ln \mathscr{P}/\partial W_{ij}$ [31,33], where the learning rate α is an undetermined parameter controlling the updating speed. In contrast, the maximizing condition ($\partial \ln \mathscr{P}/\partial W_{ij} = 0$) and mean-field approximations provide matrix equations, $W = A^{-1}BC^{-1}$, where matrices $B_{ij} = \langle \delta\sigma_i(t+1)\delta\sigma_j(t)\rangle$ and $C_{ij} = \langle \delta\sigma_i(t)\delta\sigma_j(t)\rangle$ represent time-delayed and equal-time correlations in data, and A are diagonal matrices, which are different for nMF, TAP, and eMF (see Supplemental Material, Text 2 for brief reviews of these mean-field methods [26]).

As shown in our Jupyter notebook [27], our stopping criterion can also help eMF and MLE avoid unnecessary iterations and improve the performance of these methods. Therefore, as a concrete illustration, in the following, we consider the case of eMF and MLE with our stopping criterion. For weak coupling ($g = 1$), TAP, eMF, MLE, and FEM have similar

inference accuracy that increases with sample size [Fig. 1(e)]. nMF showed poor accuracy independent of data size, since the zeroth-order mean-field approximation works only for very weak coupling strengths [31]. As we further increase coupling strength, the other two mean-field methods, TAP and eMF, also start to give less accurate results than MLE and FEM [Fig. 1(f)–1(h)]. The errors at the large coupling strength originate from the approximation of weak coupling expansions in nMF and TAP and the assumption of a Gaussian distribution of $\sum_{j=1}^{N} W_{ij}\sigma_j$ in eMF, developed in the thermodynamic limit ($N \to \infty$). However, our iterative method, FEM, and the standard MLE do not make assumptions on the coupling strength. For large sample size ($L/N^2 > 1$), FEM works as well as MLE, but for small sample size, FEM provides better accuracy than MLE. For example, the inference error (MSE) of FEM is approximately 4 times lower than that of MLE for $L/N^2 = 0.2$ and $g = 4$. As noted above, the separation between model updates and goodness of fit cost, D_i, in FEM is critical for stopping model updates for small sample size.

In addition to inference accuracy, FEM has two advantages in computation. First, the FEM update is multiplicative and not incremental, while MLE updates (using conjugate gradient ascent or some other numerical maximization) have an undetermined parameter, the learning rate α, which needs to be tuned. A very large rate ($\alpha = 3$) leads to loss of convergence, whereas a very small rate ($\alpha = 0.5$) leads to many iterations with infinitesimal updates. We set $\alpha = 1$. Second, FEM requires 20 times fewer updates than MLE [Fig. 2(a)], which reduces computation time 100-fold [Fig. 2(b)]. Note that the matrix inversion of C^*_{ij} is performed only once at the beginning and is not a computational efficiency consideration in either FEM or any of the MLE based methods. There are no other matrix inversions in FEM.

To further demonstrate the effectiveness of FEM, we show two examples of inferred networks when W_{ij} has more general coupling distributions than the SK model, as real systems often deviate strongly from normally distributed coupling strengths. In the first example, the spins have alternating bands of positive and negative couplings modulated by distance as $|W_{ij}| = W_0/\ln(R_{ij})$, where R_{ij} represents the radius of the circle [Fig. 3(a)]. The couplings are non-normally distributed [Fig. 3(b)]. The spin raster scan exhibits nontrivial structure [Fig. 3(c)], reminiscent of binocular rivalry [34]. As the number of observed configurations increases, the predicted coupling strengths [Fig. 3(d)] approach their true values [Fig. 3(a)]. In the second, the photograph of the 2018 Gerber baby, Lucas Warren, was used as the heat map of the coupling matrix [Fig. 3(e)]. These couplings are also non-normally distributed [Fig. 3(f)] with periodic bursting in the simulated spin raster scan [Fig. 3(g)], but the couplings are still predicted well [Fig. 3(h)].

Our formulation, based on the differential geometry of the data free energy, automatically includes higher-order regression equations for the local field $H_i(\sigma)$ in Eq. (24). For example, we checked higher-order inference with FEM by using a generalized kinetic Ising model with linear and quadratic couplings, $H_i(\sigma(t)) = \sum_j W_{ij}\sigma_j(t) + \sum_{j,k} Q_{ijk}\sigma_j(t)\sigma_k(t)/2$, where W_{ij} and Q_{ijk} are normally distributed. The quadratic couplings are symmetric ($Q_{ijk} = Q_{ikj}$) and have no self-interactions ($Q_{ijj} = 0$) since $\sigma_j^2 = 1$. The number of Q_{ijk} parameters is $N^2(N-1)/2$. The recovery of both linear and quadratic couplings is evident (Fig. 4).

B. Neuronal network

We applied our method to infer a neuronal network from temporal neuronal activities in the tiger salamander (*Ambystoma tigrinum*) retina [35]. The multichannel experiment recorded stochastic firing patterns of 160 neurons when the salamander retina was stimulated by a film clip of fish swimming. As in Ref. [36], we considered only the 100 most active neurons. After processing the data [see Supplemental Material, Text 3 [26]; Fig. 5(a)], we inferred the neuronal network governing the local field, $H_i(\sigma(t)) = H_i^{\text{ext}} + \sum_j W_{ij}\sigma_j(t)$. Here we included a constant bias external field H_i^{ext} for neuron i to consider the persistent silence of neurons. We inferred the neuronal network weights W_{ij} [Fig. 5(b)], and the external local fields for each neuron by using $H_i^{\text{ext}} = \langle H_i \rangle - \sum_j W_{ij}\langle \sigma_j \rangle$. The external local fields are mostly negative, which implies that neuronal activities are biased to be silent [Fig. 5(c)].

The true couplings are unknown for this system. As a validation, with the H_i^{ext} and W_{ij} that we had determined, we simulated neuronal activities. We found agreement between the covariances of neuronal activities $C_{ij} = \langle \delta\sigma_i(t+1)\delta\sigma_j(t) \rangle$ of the observed and simulated data [Fig. 5(d)]. For a more stringent validation, we reconstructed the full neuronal activities from specific "pinned" neuron activities, representing inputs. Fixing the time sequences $\sigma_j(t)$ of specific chosen input neurons $j \in I$, we reconstructed the activities $\sigma_i(t+1)$ of the remaining neurons $i \notin I$.

As a control, we selected the input neurons at random and compared them with input neurons selected on the basis of the coupling strength $|W_{ij}|$ as the input set I. As more input neurons are considered, the reconstruction predicts $\sigma_i(t+1)$ more accurately [Figs. 5(e) and S2 [26]]. Pinning the activities of only $|I| = 10$ strongly coupled neurons gave predicted activities of the remaining 90 neurons that were very close to the observed activities [Fig. 5(f)], in contrast to predicted activities obtained by pinning randomly selected sets of 10 input neurons [Fig. 5(g)].

C. Currency network

Finally, we apply our method to another difficult and representative stochastic problem, currency exchange rate fluctuations. We obtained time series of currency exchange rates from January 2000 to December 2017 [37], and examined exchange rates denominated in Euro (EUR) of 11 actively traded currencies [Fig. 6(a)]. First, we concentrate on the daily fluctuations of the exchange rates, since most financial analyses center on price increments rather than absolute prices [38]. We binarize the real-valued rates to concentrate on the sign of their daily fluctuations [Fig. 6(b)]. We defined the binarized rate $\sigma_i(t) = 1$ for a day-to-day increase of exchange rate i at time t [$r_i(t) > r_i(t-1)$], and $\sigma_i(t) = -1$ for the decrease. If there was no change [$r_i(t) = r_i(t-1)$], we set $\sigma_i(t) = \sigma_i(t-1)$. Second, we divide the data for different periods to investigate the time dependence of the couplings between exchange rates. Using the Fourier transform of the binarized time series, we identified a characteristic period, 550 business days (~2 years), of the fluctuations [Fig. 6(c)]. We inferred the currency network weights W_{ij} separately in two-year periods, shown here [Figs. 6(d)–6(f), upper] for the three periods 2012–2013, 2014–2015, and 2016–2017. We found agreement between the covariance $C_{ij} \langle \delta\sigma_i(t+1)\delta\sigma_j(t) \rangle$ of the observed currency data and that of the simulated

currency data using $H_i(\sigma(t)) = H_i^{\text{ext}} + \sum_j W_{ij}\sigma_j(t)$ [Figs. 6(d)–6(f), lower]. In contrast, when we estimated the currency network using the data for the entire period 2000–2017, the network had weaker connections and smaller covariances C_{ij} compared to the time-dependent analysis [Fig. 6(g)].

The raw exchange rate data are continuous. Is our binarized inference of any practical value? To address this, we simulated a currency trade strategy, and checked if the strategy was profitable. Using only data within a time window of a period T, $\{\sigma(t-T+1), \sigma(t-T+2), \ldots, \sigma(t)\}$, we predicted the currency fluctuations $\sigma(t+1)$ on the next day. For the trade simulation, we considered a hedging trader who buys one currency with 1 EUR and sells one currency with 1 EUR. To earn profits, the trader is supposed to sell or buy a currency that has the highest probability of increase or decrease in exchange rate: the currency sell = arg $\max_i P(\sigma_i(t+1) = +1|\sigma(t))$ and the currency buy = arg $\max_i P(\sigma_i(t+1) = -1|\sigma(t))$. Then, a daily profit can be defined as profit$(t) = r_{\text{sell}}(t+1)/r_{\text{sell}}(t) - r_{\text{buy}}(t+1)/r_{\text{buy}}(t)$. We calculated cumulative profits of the trade simulation from 2004 to 2017 with various time window sizes that we considered as past information [Fig. 6(h) for $T = 500$ days]. Hedging strategies profit from market volatility and, indeed, our trade simulation showed large profits when the exchange rates had large fluctuations [Fig. 6(a)]. The window size T had an optimal period of 500–750 business days [Fig. 6(i)]. For a more refined strategy, we considered the quality or accuracy of our inference by probing the discrepancy D_i in Eq. (26). Instead of trading every day, we traded only on the days when the discrepancy at that day, $D(t) \equiv \sum_i [\sigma_i(t) - \langle\sigma_i(t)\rangle_{H_i(\sigma(t-1))}]^2$, was lower than the average $T^{-1}\sum_{t=1}^{T} D(t)$ for a fixed window size T. This strategy doubled the profits per transaction [Figs. 6(h) and 6(i)], showing that the discrepancy D_i is a useful measure of model accuracy.

IV. DISCUSSION

We demonstrated that underdetermined stochastic systems can be inferred in a conceptually simple and computationally efficient manner using the mathematical framework of statistical physics. Since network inference is an important subject, many different approaches have been developed. Equilibrium approaches assume symmetric interactions ($W_{ij} = W_{ji}$) between node i and node j, and estimate the pair-wise interaction strengths that can maximally explain the observed static patterns of network activity in brains [36,39,40], proteins [41,42], and stock markets [43]. In contrast, nonequilibrium approaches do not assume symmetry, and infer asymmetric causal relations between nodes that can better explain dynamic patterns of network activity [33]. Network inference for nonequilibrium models (e.g., using recurrent neuronal networks) is computationally expensive. Although mean-field methods have been introduced to circumvent this practical problem [31,44,45], these approximation methods only work for weak-interaction regimes with large sample size. All small sample size inference must contend with overfitting so the key feature of our approach was to consistently decouple the model update step and a discrepancy measure that is similar to expectation maximization. This decoupling allows us to iterate with a multiplicative model update, and to stop when the discrepancy measure quantifies that the multiplicative update has saturated. We derived this within a standard statistical physics formulation [28,29], so no *ad hoc* averaging or approximation steps were involved. We

demonstrated that our method outperforms others in inferring the asymmetric interactions of the kinetic Ising model, especially in strong-interaction regimes, and particularly when available data are limited. Another aspect of small sample size inference is that longer time-scale modulation of couplings can be uncovered. This is of considerable practical import as we demonstrated with the currency exchange rate network.

FEM has several computational merits. Besides having no incremental learning rate that requires tuning, the method is parallelizable and scalable: We computed results for the kinetic Ising model with up to $N = 5000$ interacting spins, determining 2.5×10^7 parameters (Fig. S3 [26]). We also demonstrated that the method can infer not only linear interactions but also higher-order interactions. Moreover, FEM is generalizable to systems with any number of discrete states, although we focused on binary stochastic systems here.

We have addressed the *inference* of networks but without addressing the *predictive* capabilities of the networks inferred directly. While our profitable trade demonstration shows that the inferred model is generalizing well, we emphasize that just because we have found a good stopping criterion for our iterative update does not imply that the predictions from the inferred model are as accurate as the stopping criterion value would indicate. Finding a stopping criterion that would *include* predictive accuracy is an area that we are investigating. The usual approach is to perform a training-testing split to evaluate predictive performance, but this may not be optimal for small datasets. Finally, uncovering hidden nodes for stochastic network inference [46] is an exciting avenue for future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

Gašper Tkačik generously provided the neuronal activity data. We thank Changbong Hyeon and Arthur Sherman for comments on the manuscript. This work was supported by the Intramural Research Program of the National Institutes of Health, NIDDK (D.-T.H., V.P.), and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B03932264) and the Max Planck Society, Gyeongsangbuk-Do and Pohang City (J.S., J.J.).

References

[1]. Klimovskaia A, Ganscha S, and Claassen M, PLoS Comput. Biol 12, e1005234 (2016). [PubMed: 27923064]

[2]. Bar-Joseph Z, Gitter A, and Simon I, Nat. Rev. Genet 13, 552 (2012). [PubMed: 22805708]

[3]. Dombeck DA, Khabbaz AN, Collman F, Adelman TL, and Tank DW, Neuron 56, 43 (2007). [PubMed: 17920014]

[4]. Schneidman E, Berry II MJ, Segev R, and Bialek W, Nature 440, 1007 (2006). [PubMed: 16625187]

[5]. Nguyen JP, Shipley FB, Linder AN, Plummer GS, Liu M, Setru SU, Shaevitz JW, and Leifer AM, Proc. Natl. Acad. Sci. U.S.A 113, E1074 (2016). [PubMed: 26712014]

[6]. Bernal-Casas D, Lee HJ, Weitz AJ, and Lee JH, Neuron 93, 522 (2017). [PubMed: 28132829]

[7]. Sugihara G, May R, Ye H, Hsieh C.-h., Deyle E, Fogarty M, and Munch S, Science 338, 496 (2012). [PubMed: 22997134]

[8]. Schmidt M and Lipson H, Science 324, 81 (2009). [PubMed: 19342586]

[9]. Brunton SL, Proctor JL, and Kutz JN, Proc. Natl. Acad. Sci. U.S.A 113, 3932 (2016). [PubMed: 27035946]

[10]. Yair O, Talmon R, Coifman RR, and Kevrekidis IG, Proc. Natl. Acad. Sci. U.S.A 114, E7865 (2017). [PubMed: 28831006]

[11]. Nguyen HC, Zecchina R, and Berg J, Adv. Phys 66, 197 (2017).

[12]. Natale JL, Hofmann D, Hernández DG, and Nemenman I, arXiv:170506370.

[13]. Raj A and van Oudenaarden A, Cell 135, 216 (2008). [PubMed: 18957198]

[14]. Hamilton JD, Time Series Analysis, Vol. 2 (Princeton University, Princeton, NJ, 1994).

[15]. Friedman N, Science 303, 799 (2004). [PubMed: 14764868]

[16]. Janes KA and Yaffe MB, Nat. Rev. Mol. Cell Biol 7, 820 (2006). [PubMed: 17057752]

[17]. Connor JT, Martin RD, and Atlas LE, IEEE Trans. Neural Network 5, 240 (1994).

[18]. Sohl-Dickstein J, Battaglino PB, and DeWeese MR, Phys. Rev. Lett 107, 220601 (2011). [PubMed: 22182019]

[19]. Decelle A and Ricci-Tersenghi F, Phys. Rev. Lett 112, 070603 (2014). [PubMed: 24579583]

[20]. Mosteller F and Tukey J, in Revised Handbook of Social Psychology, Vol. 2, edited by Lindzey Gand Aronson E(Addison Wesley, Reading, Massachusetts, 1968), pp. 80–203.

[21]. Tibshirani R, J. Roy. Stat. Soc. B 58, 267 (1996).

[22]. Phillips DL, J. ACM 9, 84 (1962).

[23]. Tikhonov AN, Goncharsky AV, Stepanov VV, and Yagola AG, Numerical Methods for the Solution of Ill-Posed Problems (Springer-Science+Business Media, New York, 1995).

[24]. Bühlmann P and van de Geer S, Statistics for High-Dimensional Data: Methods, Theory and Applications (Springer-Verlag, Berlin, 2011).

[25]. Huang H, Eur. Phys. J. B 86, 484 (2013).

[26]. See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.99.023311 where the derivation for the inference of quadratic interactions, the brief review on mean-field methods, and the description of neuronal data processing are summarized.

[27]. Hoang D-T, Song J, Periwal V, and Jo J, https://nihcompmed.github.io/network-inference/.

[28]. Schwinger J, London Edinburgh Dublin Philos. Mag. J. Sci 44, 1171 (1953).

[29]. Toms DJ, The Schwinger Action Principle and Effective Action (Cambridge University, New York, 2007).

[30]. Sherrington D and Kirkpatrick S, Phys. Rev. Lett 35, 1792 (1975).

[31]. Roudi Y and Hertz J, Phys. Rev. Lett 106, 048702 (2011). [PubMed: 21405370]

[32]. Mézard M and Sakellariou J, J. Stat. Mech (2011) L07001.

[33]. Zeng H-L, Alava M, Aurell E, Hertz J, and Roudi Y, Phys. Rev. Lett 110, 210601 (2013). [PubMed: 23745850]

[34]. Moreno-Bote R, Rinzel J, and Rubin N, J. Neurophysiol 98, 1125 (2007). [PubMed: 17615138]

[35]. Marre O, Tkacik G, Amodei D, Schneidman E, Bialek W, and Berry M, IST Austria (2017), 10.15479/AT:ISTA:61.

[36]. Tkačik G, Marre O, Amodei D, Schneidman E, Bialek W, and Berry II MJ, PLOS Comput. Biol 10(1), e1003408 (2014). [PubMed: 24391485]

[37]. Bank of Italy, https://tassidicambio.bancaditalia.it/timeSeries.

[38]. Pincus S and Kalman RE, Proc. Natl. Acad. Sci. U.S.A 101, 13709 (2004). [PubMed: 15358860]

[39]. Tkačik G, Mora T, Marre O, Amodei D, Palmer SE, Berry MJ, and Bialek W, Proc. Natl. Acad. Sci. U.S.A 112, 11508 (2015). [PubMed: 26330611]

[40]. Watanabe T, Hirose S, Wada H, Imai Y, Machida T, Shirouzu I, Konishi S, Miyashita Y, and Masuda N, Nat. Commun 4, 1370 (2013). [PubMed: 23340410]

[41]. Mora T, Walczak AM, Bialek W, and Callan CG, Proc. Natl. Acad. Sci. U.S.A 107, 5405 (2010). [PubMed: 20212159]

[42]. Weigt M, White RA, Szurmant H, Hoch JA, and Hwa T, Proc. Natl. Acad. Sci. U.S.A 106, 67 (2009). [PubMed: 19116270]

[43]. Bury T, Physica A (Amsterdam) 392, 1375 (2013).

[44]. Huang H and Kabashima Y, J. Stat. Mech (2014) P05020.
[45]. Mahmoudi H and Saad D, J. Stat. Mech (2014) P07001.
[46]. Hoang D-T, Jo J, and Periwal V, arXiv:190104122.

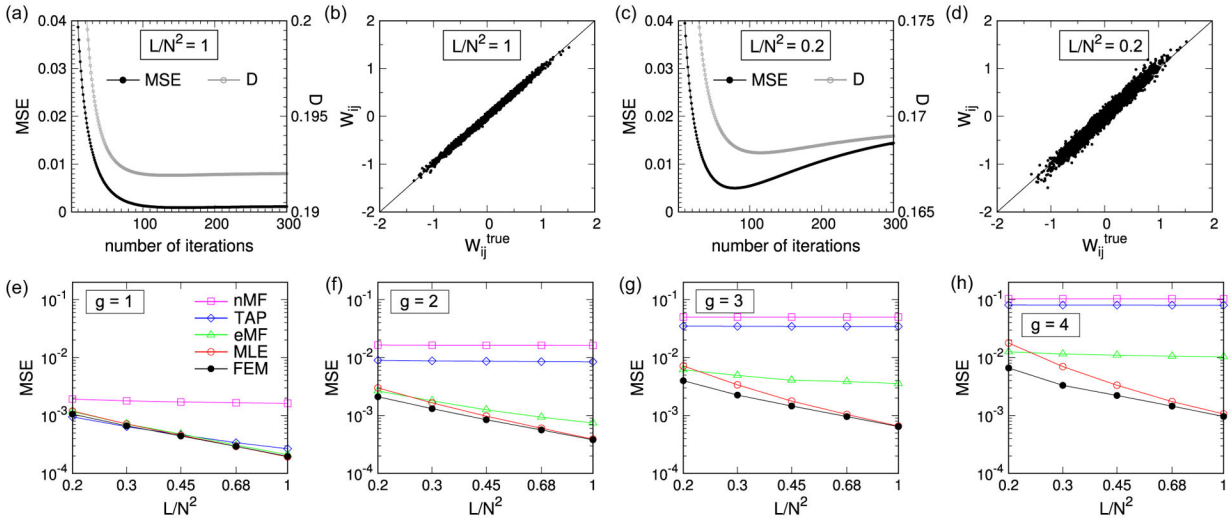

FIG. 1.

Network inference for the kinetic Ising model. Inference mean-square error (MSE, black) and discrepancy (D, gray) are shown as functions of the number of iterations for large observed configurations, $L/N^2 = 1$ (a) and few observed configurations, $L/N^2 = 0.2$ (c). Predicted couplings vs actual couplings for $L/N^2 = 1$ (b) and $L/N^2 = 0.2$ (d). The inference errors are obtained for naïve mean-field (nMF), Thouless-Anderson-Palmer (TAP), exact mean-field (eMF), maximum likelihood estimation (MLE), and free energy minimization (FEM), for various numbers of observed configurations, L/N^2 from 0.2 to 1 in the limit of weak coupling, $g = 1$ (e), and in the limit of stronger coupling, $g = 2$ (f), $g = 3$ (g), and $g = 4$ (h). A system size $N = 100$ is used. A learning rate $\alpha = 1$ is used for MLE.

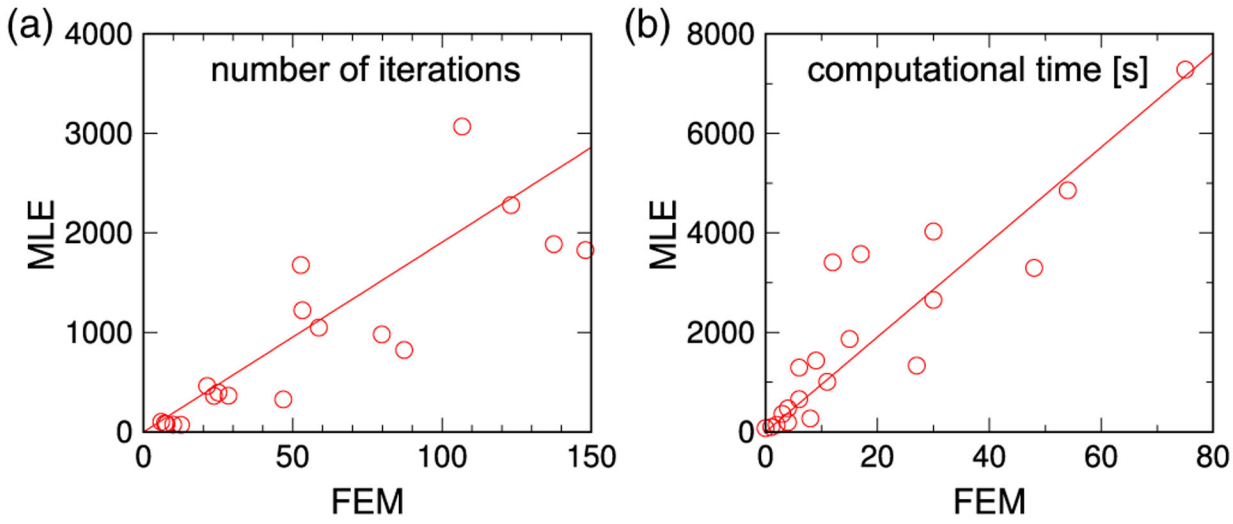

FIG. 2.

Efficiency of inference. Number of iterations per spin (a) and real computational time (b) by using MLE vs FEM for various coupling strengths g from 1 to 4 and number of observed configurations L/N^2 from 0.2 to 1. A system size $N = 100$ is used. A learning rate $\alpha = 1$ is used for MLE.

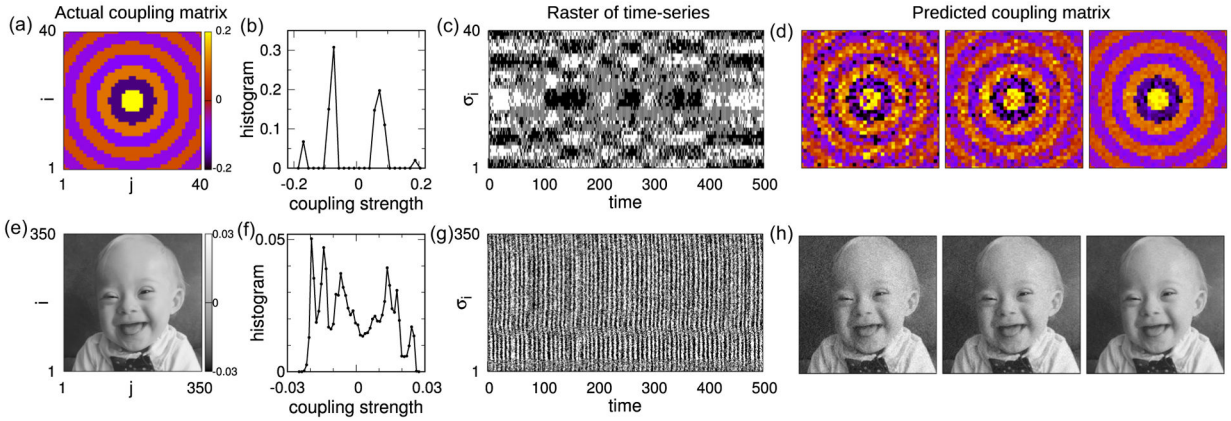

FIG. 3.

Effectiveness of FEM in inferring network with specific structures. Given true coupling weights of $N = 40$ (a) and 350 (e) spin variables with non-Gaussian distributions, typical time series of their activities are generated, (c) and (g). Predicted coupling weights are obtained for different data lengths $L/N^2 = 0.5$, 1, and 4 from left to right, (d) and (h). The image is converted from the photograph of the 2018 Gerber baby, Lucas Warren (with permission from Gerber).

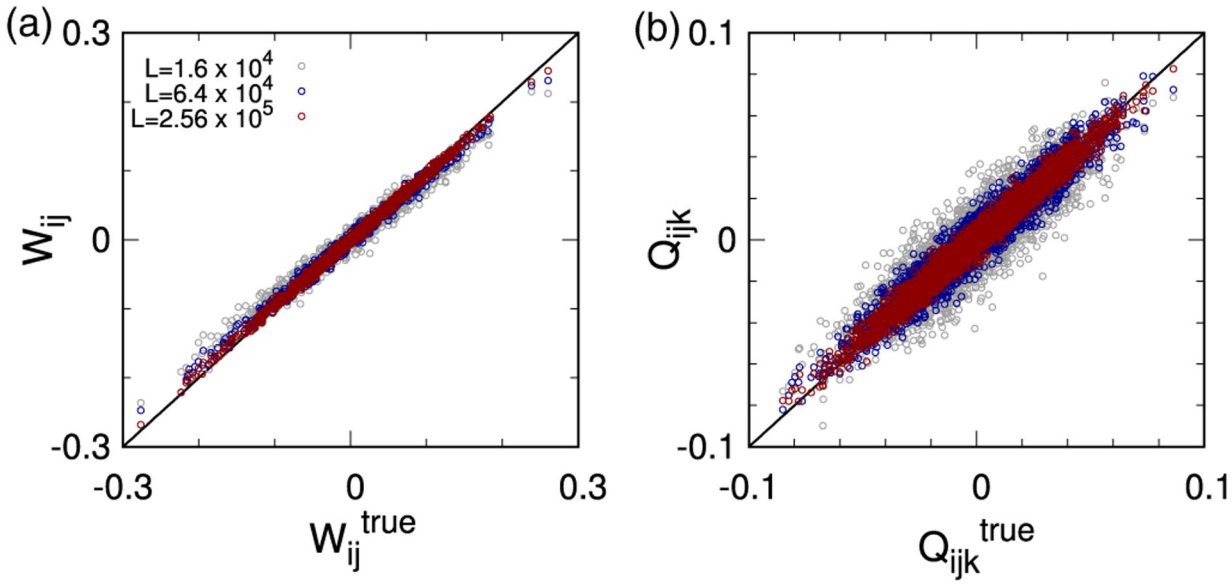

FIG. 4.

Accurate inference of higher-order coupling strengths. Linear (a) and quadratic (b) coupling strengths in the nonlinear kinetic Ising model are predicted from FEM. Here the true coupling strengths are normally distributed with a system size $N = 40$. Three different data lengths, $L = 1.6 \times 10^4$ (gray), 6.4×10^4 (blue), and 2.56×10^5 (red), are examined.

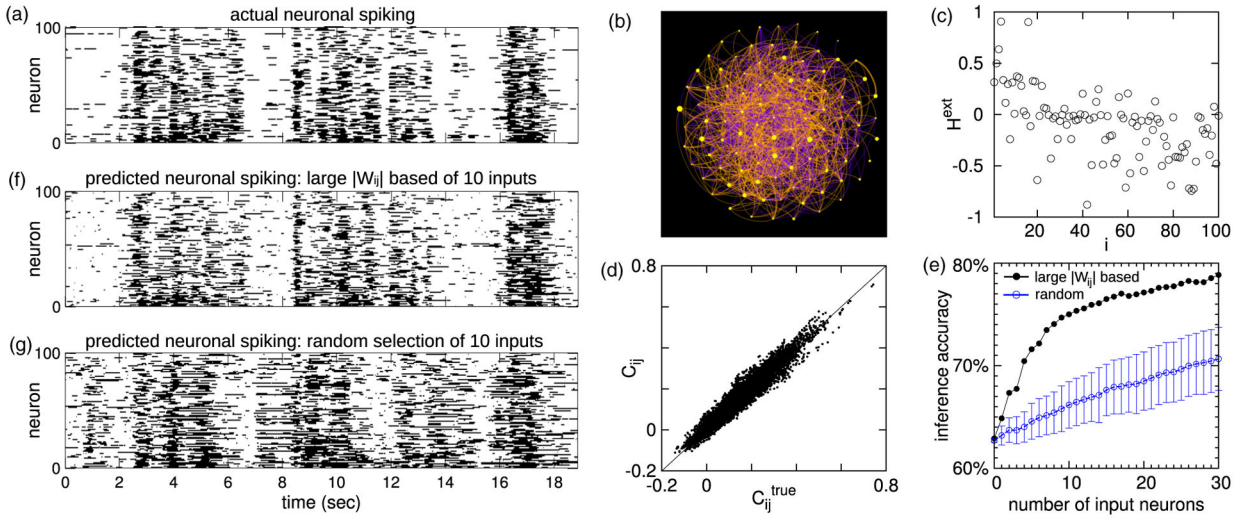

FIG. 5.

Inference of coupling strengths between neurons, external local fields, and neuronal activities. From activities of 100 neurons (a), the neuronal network (b) and external local field H_i^{ext} (c) are predicted. The red and blue edges represent positive and negative couplings, respectively. Edge direction is clock-wise. Inferred correlation covariances C_{ij} are compared with actual correlation covariances C_{ij}^{true} (d). Inference accuracy of remaining neuronal activities vs number of input neurons selected based on large $|W_{ij}|$ (filled black circles), and randomly selected (empty blue circles). Error bars represent the standard deviation from 50 random trials (e). Neuronal activities are reconstructed with 10 input neurons, selected based on large $|W_{ij}|$ (f), and randomly selected (g).

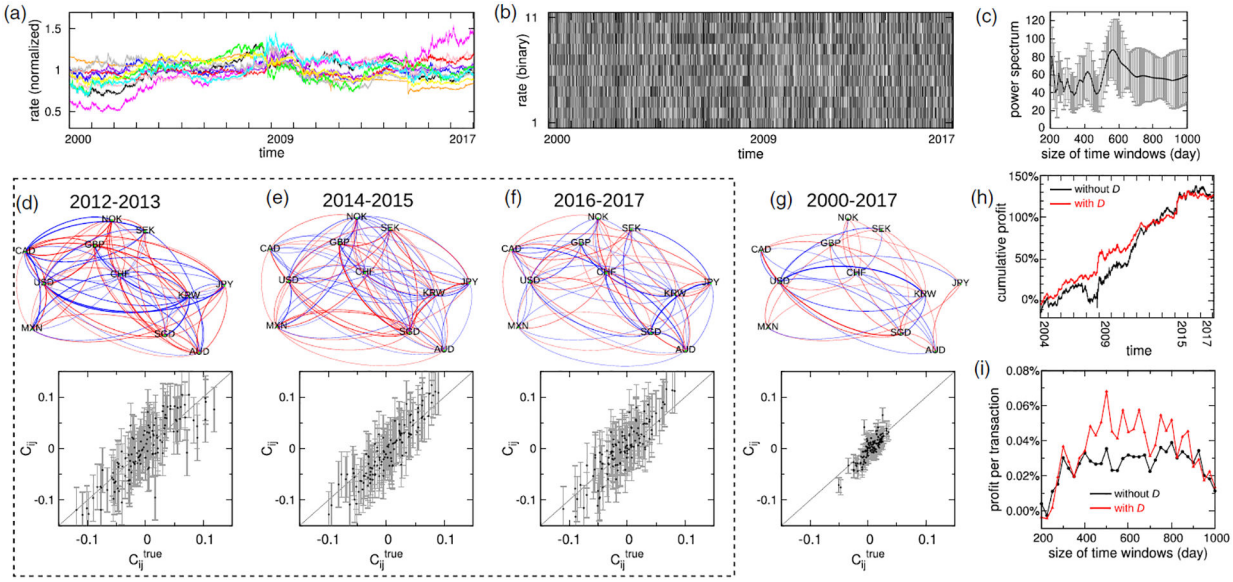

FIG. 6.

Inference of coupling strengths between currency exchange rates. Normalized exchange rates relative to EUR of 11 currencies are plotted with different colors representing distinct currencies (a). A raster representation of binarized exchange rate fluctuations is plotted with black dots representing increase, white dots decrease. Average power spectrum obtained from a Fourier transform of exchange rate fluctuations vs time-window size in which error bar represents standard deviation from different currencies (c). The currency networks are predicted for different periods, e.g., from the years of 2012 to 2013 (d), 2014 to 2015 (e), and 2016 to 2017 (f). The network for the whole data, from 2000 to 2017, is also predicted (g). The red and blue edges represent positive and negative couplings, respectively. Edge direction is clock-wise. Predicted covariances are shown to compare with observed covariances C_{ij}^{true} [(d)–(g), lower]. Cumulative profit vs time period with trading every day (without D, black) and trading only on days specified by lower model discrepancy (with D, red) strategies (h). Profit per transaction using our strategy is plotted as a function of time-window size (i).