Causality inference in stochastic systems from neurons to currencies: Profiting from small sample size

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

[1] *Laboratory of Biological Modeling, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA*
[2] *Asia Pacific Center for Theoretical Physics, Pohang, Gyeongbuk 37673, Korea*
[3] *Department of Physics, Pohang University of Science and Technology, Pohang, Gyeongbuk 37673, Korea*
[4] *The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34014 Trieste, Italy*
[5] *School of Computational Sciences, Korea Institute for Advanced Study, Seoul 02455, Korea*
[6] *Department of Statistics, Keimyung University, Daegu 42601, Korea*

(Dated: September 13, 2018)

Success in modeling complex phenomena such as human perception hinges critically on the availability of data and computational power. Significant progress has been made in modeling such phenomena using probabilistic methods, particularly in image analysis and speech recognition. Maximum Likelihood Estimation (MLE) combined with Bayesian model selection is the basis of much of this progress, as MLE converges to the true model with copious data. In the sciences, large enough datasets are rarae aves, so alternatives to MLE must be developed for small sample size. We introduce a data-driven statistical physics approach to model inference based on minimizing a free energy of data and show superior model recovery for small sample sizes. We demonstrate coupling strength inference in non-equilibrium kinetic Ising models, including in the difficult large coupling variability regime, and show scaling to systems of arbitrary size. 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.

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 probabilisitic 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 universal model-free 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 too complex with increasing N. Even considering only pair-wise interactions requires determining N^2 parameters and demands $L \geq N^2$ samples. Including higher-order interactions leads to an exponential increase in the number of model parameters, and a concomitant increase in 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.

Maximum Likelihood Estimation (MLE) is the gold standard for stochastic model parameter inference, as it converges to the true model parameters in the limit of large sample size. On the other hand, MLE is limited by the fact that the likelihood equations are specific to a given estimation problem, that the numerical estimation is usually non-trivial, and most importantly, MLE can be heavily biased for small samples where the optimality properties of MLE may not apply. MLE can also be sensitive to the choice of starting values [18].

According to the Rao-Blackwell theorem [19, 20], the conditional expected value of an estimator given a sufficient statistic is another estimator that is at least as good, and this result applies to MLE estimators as well. The Rao-Blackwell result usually applies for sufficient and complete statistics, and leads to an idempotent improvement, in other words, the improvement requires no iteration. However, for our small sample size purposes, more apropos is the recent result of Galili and Meilij-

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

son [21], which suggests that a Rao-Blackwell–type iterative improvement of a parameter estimator is worth investigating.

Statistical physics is often used for model inference [22, 23], 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 corresponds to linear and higher-order data regressions. Over-fitting is a major problem in the analysis of under-determined systems. By decoupling an iterative Rao-Blackwell estimator update step from an update–consistent stopping criterion, we demonstrate that our Free Energy Minimization (FEM) approach infers coupling strengths in non-equilibrium kinetic Ising models, outperforming previous approaches particularly in the large coupling variability and small sample size regimes. Real data is 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.

II. ITERATIVE STOCHASTIC CAUSALITY INFERENCE FROM FREE ENERGY MINIMIZATION

As a concrete illustration, let us start with a kinetic Ising model in which a vector σ of N spins $\sigma_i(t) = \pm 1$ is stochastically updated based on the following conditional probability

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

with a local field $H_i(\sigma(t)) \equiv \sum_j W_{ij}\sigma_j(t)$. Our goal is to infer the coupling strength W_{ij} that minimizes the discrepancy between observed $\sigma_i(t+1)$ and model expectation $\langle\langle\sigma_i(t+1)\rangle\rangle_{\sigma(t)} \equiv \sum_{\rho=\pm1}\rho P(\sigma_i(t+1) = \rho|\sigma(t))$. For the kinetic Ising model, $\langle\langle\sigma_i(t+1)\rangle\rangle_{\sigma(t)} = \tanh H_i(\sigma(t))$.

To infer W_{ij} from $\{\sigma(t)\}$, we implement a Rao-Blackwell scheme of estimator improvement, $H_i^{\rm new}(m) = \sum_j W_{ij}^{\rm new} m_j$, by obtaining the expectation values $\langle E_i\rangle_m$ of observables $E_i(t)$ conditioned on m. At the moment, we assume that $E_i(t)$ is given for every data, but later we will introduce a well-fitted definition of $E_i(t)$. The elegant mathematical formalism developed by Schwinger provides a natural connection between expectation values $m = \langle\sigma\rangle$ of microstates σ and expectation values $\langle E_i\rangle_m$ of observables E_i conditioned on m [24, 25]. We first define a moment generating function,

$$Z_i(J,\beta) = \sum_t \exp(J\cdot\sigma(t) - \beta E_i(t)), \tag{2}$$

which is a function of a vector parameter J, a scalar parameter β, and a 'data energy' $E_i(t)$. A convex free energy $F_i = \log Z_i$ can be used to obtain expectation values of spin activities and energy by differentiation,

$$\frac{\partial F_i}{\partial J_j} = \frac{\sum_t \sigma_j(t)\exp(J\cdot\sigma(t) - \beta E_i(t))}{\sum_t \exp(J\cdot\sigma(t) - \beta E_i(t))} = \langle\sigma_j\rangle_J \equiv m_j(J), \tag{3}$$

$$\frac{\partial F_i}{\partial \beta} = -\frac{\sum_t E_i(t)\exp(J\cdot\sigma(t) - \beta E_i(t))}{\sum_t \exp(J\cdot\sigma(t) - \beta E_i(t))} = -\langle E_i\rangle_J. \tag{4}$$

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, $P(\sigma(t)) \equiv \exp(J\cdot\sigma(t) - \beta E_i(t) - F_i)$ in Eq. (2), we can show that G_i can be indeed interpreted as a thermodynamic free energy,

$$G_i = \beta\langle E_i\rangle_J - S_i \tag{5}$$

with the expectation value of E_i and the Shannon entropy of data, $S_i = -\sum_t P(\sigma(t))\log P(\sigma(t))$. At $\beta = 0$, minimizing the free energy is exactly maximizing the entropy, making every sample equally valuable. Then, the duality between the free energies F_i and G_i through their Legendre transform leads to

$$\frac{\partial G_i}{\partial m_j} = J_j, \tag{6}$$

$$\frac{\partial G_i}{\partial \beta} = -\frac{\partial F_i}{\partial \beta} = \langle E_i\rangle_m, \tag{7}$$

where we identify $\langle E_i\rangle_{J(m)} \equiv \langle E_i\rangle_m$. Therefore, once we know the free energy G_i, it is straightforward to obtain $\langle E_i\rangle_m$. For our purposes, however, it is not necessary to obtain $G_i(m)$ for all values of m, as it suffices to know the function at minimum, because the free energy is minimized at the data expectation: $m^* = \langle\sigma\rangle_{J=0}$. At its minimum, m^*, we have $J(m^*) = \partial_m G(m^*) = 0$ in Eq. (6) and this is the value of J about which we will expand, hence the term Free Energy Minimization (FEM). Then, we have the Taylor expansion of $G_i(m)$ upto the second-order terms at $m = m^*$:

$$G_i(m) = G_i(m^*) + \frac{1}{2}\sum_{j,k}\left[\frac{\partial^2 G_i}{\partial m_j \partial m_k}\right]^*(m_j - m_j^*)(m_k - m_k^*), \tag{8}$$

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

$$\frac{\partial G_i(m)}{\partial \beta} = \frac{\partial G_i(m^*)}{\partial \beta} - \sum_{j,k}\frac{\partial m_k^*}{\partial \beta}\left[\frac{\partial^2 G_i}{\partial m_j \partial m_k}\right]^*(m_j - m_j^*). \tag{9}$$

Here, each derivative in Eq. (9) is calculated as follows:

$$-\frac{\partial m_k}{\partial \beta} = \frac{\partial}{\partial \beta}\left[\frac{\sum_t \sigma_k(t)\exp(J\cdot\sigma(t) - \beta E_i(t))}{\sum_t \exp(J\cdot\sigma(t) - \beta E_i(t))}\right] = \langle\delta E_i \delta\sigma_k\rangle, \tag{10}$$

and

$$\frac{\partial^2 G_i}{\partial m_j \partial m_k} = \frac{\partial J_k}{\partial m_j} = [C^{-1}]_{jk}, \tag{11}$$

where

$$C_{jk} = \frac{\partial m_j}{\partial J_k} = \frac{\partial}{\partial J_k}\left[\frac{\sum_t \sigma_j(t)\exp(J\cdot\sigma(t)-\beta E_i(t))}{\sum_t \exp(J\cdot\sigma(t)-\beta E_i(t))}\right] = \langle \delta\sigma_j \delta\sigma_k\rangle. \tag{12}$$

Here, we used short notations: $\langle f\rangle \equiv \langle f\rangle_J$, $\langle f\rangle^* \equiv \langle f\rangle_{J=0}$, and $\langle \delta f\rangle \equiv \langle f\rangle - \langle f\rangle^*$. Plugging these derivatives into Eq. (9), we obtain the following equation:

$$\langle \delta E_i\rangle = \sum_{j,k} \langle \delta E_i \delta\sigma_k\rangle^* [C^{-1}]^*_{kj} \langle \delta\sigma_j\rangle. \tag{13}$$

Finally, this implies

$$\langle E_i\rangle_m = \sum_j W^*_{ij}\langle\sigma_j\rangle_m, \tag{14}$$

where

$$W^*_{ij} \equiv \sum_k \langle \delta E_i \delta\sigma_k\rangle^* [C^{-1}]^*_{kj}. \tag{15}$$

This formalism allows to derive higher-order contributions of σ_j to E_i by expanding higher-order Taylor series in Eq. (8) (see SI Text 1).

We now turn to finding an appropriate E_i. Consider

$$E_i(t) \equiv \frac{\sigma_i(t+1)}{\langle\langle\sigma_i(t+1)\rangle\rangle_{\sigma(t)}} H_i(\sigma(t)), \tag{16}$$

and define the Rao-Blackwell conditional expectation update: $H_i(m)^{\text{new}} \leftarrow \langle E_i\rangle_m$. Intuitively, if the observation $\sigma_i(t+1)$ is larger/smaller than the corresponding model expectation $\langle\langle\sigma_i(t+1)\rangle\rangle_{\sigma(t)}$, this update increases/decreases $H_i(\sigma(t))$ proportionally to the discrepancy ratio between the observation and the model expectation, including the sign. The differential geometry of $G_i(m)$ around its minimum m^* then gives $W^{\text{new}}_{ij} = \sum_k \langle \delta E_i \delta\sigma_k\rangle^* [C^{-1}]^*_{kj}$ as a matrix multiplication in Eq (15).

The second crucial aspect for small sample size inference is to find a suitable stopping criterion for the Rao-Blackwell update. We consider the overall discrepancy between $\sigma_i(t+1)$ and $\langle\langle\sigma_i(t+1)\rangle\rangle_{\sigma(t)}$:

$$D_i(W) \equiv \sum_t \left[\sigma_i(t+1) - \langle\langle\sigma_i(t+1)\rangle\rangle_{\sigma(t)}\right]^2. \tag{17}$$

The minimum of $D_i(W)$ is the closest we can approach a fixed point of the update iteration, consistent with Eq. (16) and the Rao-Blackwell expectation. Therefore, we stop the iteration when $D_i(W)$ starts to increase.

To summarize the inference algorithm with FEM:

(i) Compute $H_i(\sigma(t)) \equiv \sum_j W_{ij}\sigma_j(t)$ (initialize with a random W_{ij});

(ii) Compute $E_i(t)$ as defined in Eq. (16);

(iii) Extract $W^{\text{new}}_{ij} = \sum_k \langle \delta E_i \delta\sigma_k\rangle^* [C^{-1}]^*_{kj}$;

(iv) Repeat (i)-(iii) until $D_i(W)$ starts to increase;

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

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 [26]. In the limit of large sample size (large L/N^2), our iterative method decreases the mean square error, $\text{MSE} = N^{-2}\sum_{i,j=1}^N (W_{ij} - W^{\text{true}}_{ij})^2$, as the number of iterations increases (Fig. 1A). We obtain good agreement between true and predicted weights (Fig. 1B). In real world problems, W^{true}_{ij} is inaccessible so MSE cannot be defined. However, $D_i(W)$ in Eq. (17) is an alternative measure of the discrepancy between observation $\sigma_i(t+1)$ and model expectation. The discrepancy measures $D_i(W)$ are independent for each spin i. We checked that MSE and $D = N^{-1}\sum_{i=1}^N D_i(W)$ 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. 1C). For the kinetic Ising model, $D_i(W) = 4\sum_t [1 - P(\sigma_i(t+1)|\sigma(t))]^2$ with the transition probability, $P(\sigma_i(t+1)|\sigma(t))$ in Eq. (1). Therefore, decreasing $D_i(W)$ 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 $D_i(W)$ 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. 1D) without over-fitting.

Now we compare the inference performance of our method with other representative methods [27–29]: 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, $\mathcal{P} = \prod_{t=1}^{L-1}\prod_{i=1}^N P(\sigma_i(t+1)|\sigma(t))$, and uses gradient ascent to update W_{ij} incrementally through $W^{\text{new}}_{ij} = W_{ij} + \alpha/(L-1)\partial\log\mathcal{P}/\partial W_{ij}$ [27, 29], where the learning rate α is an undetermined parameter controlling the updating speed. In contrast, the maximizing condition ($\partial\log\mathcal{P}/\partial W_{ij} = 0$) 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,

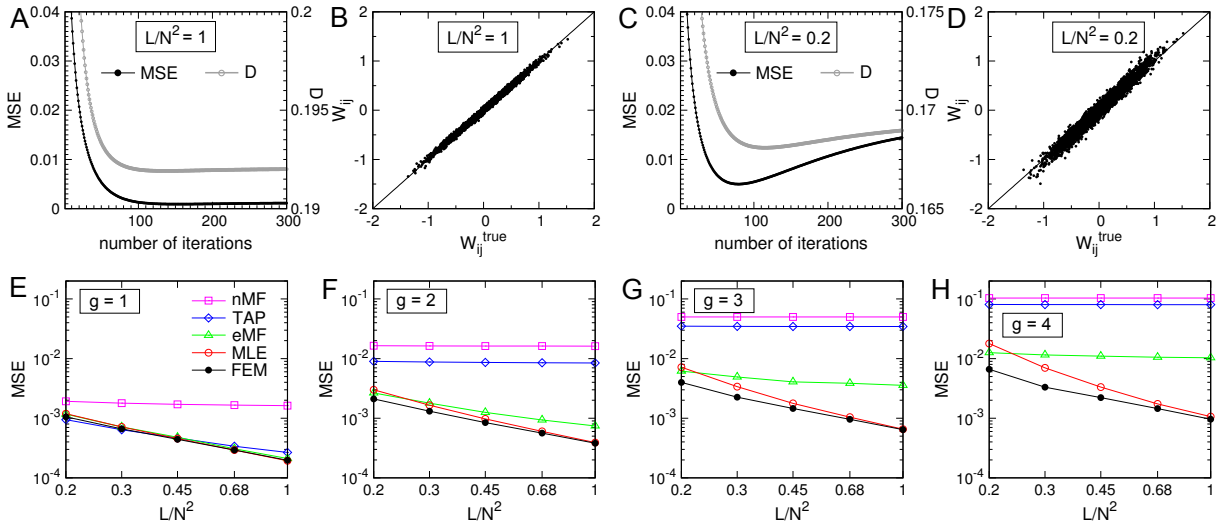

FIG. 1. Network inference for the kinetic Ising model. Inference error (MSE, black) and discrepancy (D, gray) are shown as function of number of iterations for large observed configurations, $L/N^2 = 1$ (A) and few observed configurations, $L/N^2 = 0.2$ (C). Predicted couplings versus actual couplings 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 number 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.

and A are diagonal matrices, which are different for nMF, TAP, and eMF (SI Text 2 has brief reviews of these mean-field methods).

For weak coupling ($g = 1$), TAP, eMF, MLE and FEM have similar inference accuracy that increases with sample size (Fig. 1E). nMF showed poor accuracy independent of data size, since the zeroth-order mean-field approximation works only for very weak coupling strengths [27]. 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. 1F-H). For large sample size ($L/N^2 > 1$), our iterative method, FEM, works as well as standard MLE. For small sample size, however, 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 $g = 4$. 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 determined. 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. 2A), which reduces computation time a 100-fold (Fig. 2B).

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-

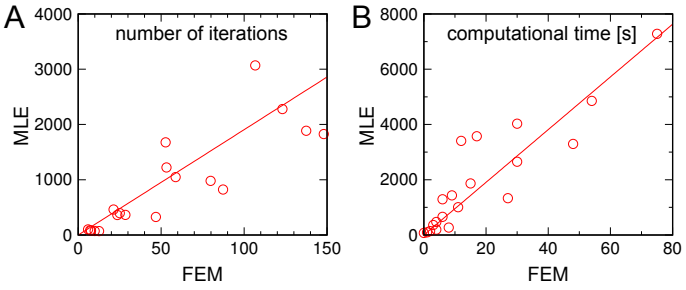

FIG. 2. Efficiency of inference. Number of iterations per spin (A) and real computational time (B) by using MLE versus FEM for various coupling strengths, g from 1 to 4 and number of observed configurations, L/N^2 from 0.2 to 1. A system size $N = 100$ is used. A learning rate $\alpha = 1$ is used for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

distributed coupling strengths. In the first example, the spins have alternating bands of positive and negative couplings modulated by distance as $|W_{ij}| = W_0/\log(R_{ij})$, where R_{ij} represents the radius of the circle (Fig. 3A). The couplings are non-normally distributed (Fig. 3B). The spin raster scan exhibits nontrivial structure (Fig. 3C), reminiscent of binocular rivalry [30]. As the number of observed configurations increases, the predicted coupling strengths (Fig. 3D) approach their true values (Fig. 3A). In the second, the 2018 Gerber baby's photograph was used as the heatmap of the coupling matrix (Fig. 3E). These couplings are also non-normally distributed (Fig. 3F) with periodic bursting in the simulated spin raster scan (Fig. 3G), but the couplings are still predicted well (Fig. 3H).

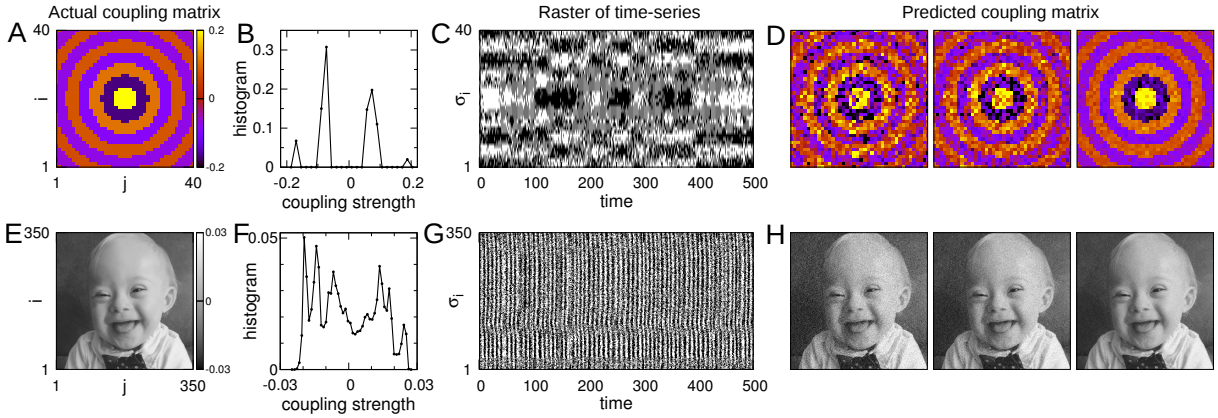

FIG. 3. Effectiveness of FEM in inferring network with specific structures. Given true coupling weights of $N = 40$ (A) and 350 (E) spin variables with non-Gaussian distributions, typical time-series of their activities are generated (C, G). Predicted coupling weights are obtained for different data lengths $L/N^2 = 0.5$, 1, and 4 from left to right (D, H). The image is converted from the 2018 Gerber baby's photograph (with permission from Gerber).

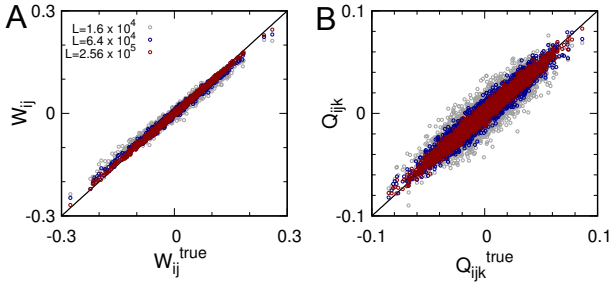

FIG. 4. Accurate inference of higher-order coupling strengths. Linear (A) and quadratic (B) coupling strengths in the nonlinear kinetic Ising model are predicted from FEM. Here the true coupling strengths are normally distributed with a system size $N = 40$. Three different data lengths, $L = 1.6 \times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

Our formulation, based on the differential geometry of the data free energy, automatically includes higher-order regression equations for the local field $H_i(\sigma)$ (SI Text 1). 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 [31]. The multi-channel experiment recorded stochastic firing patterns of 160 neurons when the salamander retina was stimulated by a film clip of fish swimming. As in Ref. [32], we considered only the 100 most active neurons. After processing the data (SI Text 3; Fig. 5A), we inferred the neuronal network governing the local field, $H_i(\sigma(t)) = H_i^{\rm ext} + \sum_j W_{ij}\sigma_j(t)$. Here we included a constant bias external field $H_i^{\rm ext}$ for neuron i to consider the persistent silence of neurons. We inferred the neuronal network weights W_{ij} (Fig. 5B), and the external local fields for each neuron by using $H_i^{\rm ext} = \langle H_i \rangle - \sum_j W_{ij}\langle\sigma_j\rangle$. The external local fields are mostly negative, which implies that neuronal activities are biased to be silent (Fig. 5C).

The true couplings are unknown for this system. As a validation, with the $H_i^{\rm ext}$ and W_{ij} we determined, we simulated neuronal activities. We found agreement between the covariances of neuronal activities $C_{ij} = \langle \delta\sigma_i(t)\delta\sigma_j(t)\rangle$ of the observed and simulated data (Fig. 5D). 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. 5E and S2). 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. 5F), in contrast to predicted activities obtained by pinning randomly selected sets of 10 input neurons (Fig. 5G).

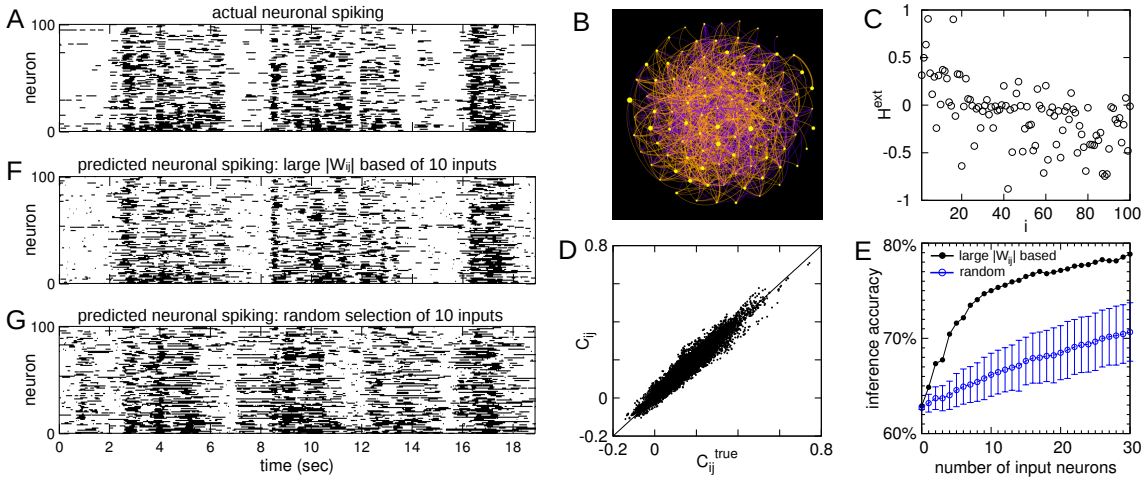

FIG. 5. Inference of coupling strengths between neurons, external local fields and neuronal activities. From activities of 100 neurons (A), neuronal network (B) and external local field $H_i^{\rm ext}$ (C) are predicted. The red and blue edges represent positive and negative couplings, respectively. Edge direction is clock-wise. Inferred correlation covariances C_{ij} are compared with actual correlation covariances $C_{ij}^{\rm true}$ (D). Inference accuracy of remaining neuronal activities versus 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).

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 [33], and examined exchange rates denominated in Euro (EUR) of 11 actively traded currencies (Fig. 6A). First, we concentrate on the daily fluctuations of the exchange rates, since most financial analyses center on price increments rather than absolute prices [34]. We binarize the real-valued rates to concentrate on the sign of their daily fluctuations (Fig. 6B). 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 ($\sim$ 2 years), of the fluctuations (Fig. 6C). We inferred the currency network weights W_{ij} separately in two year periods, shown here (Figs. 6D-F, upper) for the three periods 2012-2013, 2014-2015, and 2016-2017. We found agreement between the covariance $C_{ij} = \langle \delta\sigma_i(t)\delta\sigma_j(t)\rangle$ of the observed currency data and that of the simulated currency data using $H_i(\sigma(t)) = H_i^{\rm ext} + \sum_j W_{ij}\sigma_j(t)$ (Figs. 6D-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 (Figs. 6G)

The raw exchange rate data is 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, $\{\sigma(t-T+1), \sigma(t-T+2), \cdots, \sigma(t)\}$, 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/buy a currency that has the highest probability of increase/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 $\text{profit}(t) = r_{sell}(t+1)/r_{sell}(t) - r_{buy}(t+1)/r_{buy}(t)$. We calculated cumulative profits of the trade simulation from 2004 to 2017 with various time window sizes that we considered as past information (Fig. 6H 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. 6A). The window size T had an optimal period of 500-750 business days (Fig. 6I). For a more refined strategy, we considered the quality or accuracy of our inference by probing the discrepancy $D_i(W)$ in Eq. (17). Instead of trading every day, we traded only on the days when the discrepancy at that day, $D(t) \equiv \sum_i \left[\sigma_i(t) - \langle\langle\sigma_i(t)\rangle\rangle_{\sigma(t-1)}\right]^2$, 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. 6H and 6I), showing that the discrepancy $D_i(W)$ is a useful measure of model accuracy.

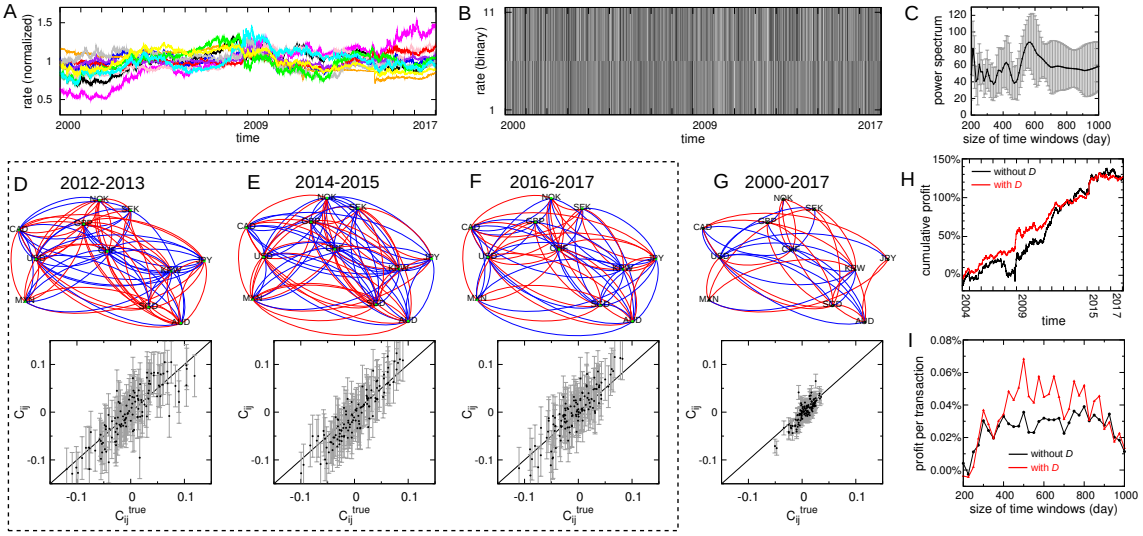

FIG. 6. Inference of coupling strengths between currency exchange rates. Normalized exchange rates relative to EUR of 11 currencies are plotted with different colors representing distinct currencies (A). A raster representation of binarized exchange rate fluctuations is plotted with black dots representing increase, white dots decrease. Average power spectrum obtained from a Fourier transform of exchange rate fluctuations versus 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 $C_{ij}^{\rm true}$ (D-G, lower). Cumulative profit versus time period for various time-window sizes (H). Profit per transaction using our strategy is plotted as a function of time-window size (I).

IV. DISCUSSION

We demonstrated that under-determined 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 [32, 35, 36], proteins [37, 38], and stock markets [39]. In contrast, non-equilibrium approaches do not assume symmetry, and infer asymmetric causal relations between nodes that can better explain dynamic patterns of network activity [29]. Causality inference for non-equilibrium models (e.g., using recurrent neuronal networks) is computationally expensive. Although mean-field methods have been introduced to circumvent this practical problem [27], these approximation methods only work for weak-interaction regimes with large sample size. All small sample size inference must contend with over-fitting 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 allowed 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 [24, 25], so no ad hoc averaging or approximation steps were involved. We demonstrated that our method outperfoms others in inferring the asymmetric interactions of the kinetic Ising model, especially in strong-interaction regimes, and particularly when available data was 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). 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. Uncovering hidden nodes for stochastic network inference [40] is an exciting avenue for future work.

ACKNOWLEDGMENT

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 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.J.).

[1] A. Klimovskaia, S. Ganscha, and M. Claassen, PLoS computational biology **12**, e1005234 (2016).
[2] Z. Bar-Joseph, A. Gitter, and I. Simon, Nature Reviews Genetics **13**, 552 (2012).
[3] D. A. Dombeck, A. N. Khabbaz, F. Collman, T. L. Adelman, and D. W. Tank, Neuron **56**, 43 (2007).
[4] E. Schneidman, M. J. Berry, 2nd, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).
[5] J. P. Nguyen, F. B. Shipley, A. N. Linder, G. S. Plummer, M. Liu, S. U. Setru, J. W. Shaevitz, and A. M. Leifer, Proc Natl Acad Sci U S A **113**, E1074 (2016).
[6] D. Bernal-Casas, H. J. Lee, A. J. Weitz, and J. H. Lee, Neuron **93**, 522 (2017).
[7] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, Science **338**, 496 (2012).
[8] M. Schmidt and H. Lipson, Science **324**, 81 (2009).
[9] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Proc Natl Acad Sci U S A **113**, 3932 (2016).
[10] O. Yair, R. Talmon, R. R. Coifman, and I. G. Kevrekidis, Proceedings of the National Academy of Sciences **114**, E7865 (2017).
[11] H. C. Nguyen, R. Zecchina, and J. Berg, Advances in Physics **66**, 197 (2017), https://doi.org/10.1080/00018732.2017.1341604.
[12] J. L. Natale, D. Hofmann, D. G. Hernández, and I. Nemenman, arXiv preprint arXiv:1705.06370 (2017).
[13] A. Raj and A. van Oudenaarden, Cell **135**, 216 (2008).
[14] J. D. Hamilton, *Time series analysis*, Vol. 2 (Princeton University Press, 1994).
[15] N. Friedman, Science **303**, 799 (2004).
[16] K. A. Janes and M. B. Yaffe, Nature reviews Molecular cell biology **7**, 820 (2006).
[17] J. T. Connor, R. D. Martin, and L. E. Atlas, IEEE transactions on neural networks **5**, 240 (1994).
[18] National Institute of Standards and Technology, *NIST/SEMATECH e-Handbook of Statistical Methods* (Accessed: April 16, 2018), `http://www.itl.nist.gov/div898/handbook/`.
[19] C. Radhakrishna Rao, Bull. Calcutta Math. Soc. **37** (1945).
[20] D. Blackwell, The Annals of Mathematical Statistics **18**, 105 (1947).
[21] T. Galili and I. Meilijson, The American Statistician **70**, 108 (2016).
[22] J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese, Phys Rev Lett **107**, 220601 (2011).
[23] A. Decelle and F. Ricci-Tersenghi, Phys Rev Lett **112**, 070603 (2014).
[24] J. Schwinger, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **44**, 1171 (1953).
[25] D. J. Toms, *The Schwinger action principle and effective action* (Cambridge University Press, 2007).
[26] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[27] Y. Roudi and J. Hertz, Phys Rev Lett **106**, 048702 (2011).
[28] M. Mézard and J. Sakellariou, Journal of Statistical Mechanics: Theory and Experiment **2011**, L07001 (2011).
[29] H.-L. Zeng, M. Alava, E. Aurell, J. Hertz, and Y. Roudi, Phys Rev Lett **110**, 210601 (2013).
[30] R. Moreno-Bote, J. Rinzel, and N. Rubin, Journal of Neurophysiology **98**, 1125 (2007).
[31] O. Marre, G. Tkacik, D. Amodei, E. Schneidman, W. Bialek, and I. Berry, Michael J, IST Austria (2017), 10.15479/AT:ISTA:61.
[32] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. J. Berry, II, PLOS Computational Biology **10**, 1 (2014).
[33] Bank of Italy, *Exchange Rates Portal* (Accessed: December 19, 2017), `https://tassidicambio.bancaditalia.it/timeSeries`.
[34] S. Pincus and R. E. Kalman, Proceedings of the National Academy of Sciences of the United States of America **101**, 13709 (2004).
[35] G. Tkačik, T. Mora, O. Marre, D. Amodei, S. E. Palmer, M. J. Berry, and W. Bialek, Proceedings of the National Academy of Sciences **112**, 11508 (2015), http://www.pnas.org/content/112/37/11508.full.pdf.
[36] T. Watanabe, S. Hirose, H. Wada, Y. Imai, T. Machida, I. Shirouzu, S. Konishi, Y. Miyashita, and N. Masuda, Nat Commun **4**, 1370 (2013).
[37] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Proceedings of the National Academy of Sciences **107**, 5405 (2010), http://www.pnas.org/content/107/12/5405.full.pdf.
[38] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proceedings of the National Academy of Sciences **106**, 67 (2009), http://www.pnas.org/content/106/1/67.full.pdf.
[39] T. Bury, Physica A: Statistical Mechanics and its Applications **392**, 1375 (2013).
[40] D.-T. Hoang, J. Jo, and V. Periwal, (2018).
[41] P. Kara, P. Reinagel, and R. C. Reid, Neuron **27**, 635 (2000).

clearpage

SUPPORTING INFORMATION (SI)

SI Text 1: Schwinger's source formalism

Here, we derive the differential geometry of $\langle E_i \rangle$ in terms of second-order $\langle \sigma \rangle$ dependency by using Schwinger's source formalism [24, 25]. This is a model-free approach, because we do not assume a specific functional form of $\langle E_i \rangle$ at the beginning. Then, we have the

Taylor expansion of $G_i(m)$ at $m = m^*$:

$$G_i(m) = G_i(m^*) + \frac{1}{2}\sum_{j,k}\left[\frac{\partial^2 G_i}{\partial m_j \partial m_k}\right]^*(m_j - m_j^*)(m_k - m_k^*) + \frac{1}{6}\sum_{j,k,l}\left[\frac{\partial^3 G_i}{\partial m_j \partial m_k \partial m_l}\right]^*(m_j - m_j^*)(m_k - m_k^*)(m_l - m_l^*) + \mathcal{O}(\delta^4 m) \tag{S1}$$

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

$$\frac{\partial G_i(m)}{\partial \beta} = \frac{\partial G_i(m^*)}{\partial \beta} - \sum_{j,k}\frac{\partial m_k^*}{\partial \beta}\left[\frac{\partial^2 G_i}{\partial m_j \partial m_k}\right]^*(m_j - m_j^*) + \frac{1}{2}\sum_{j,k}\frac{\partial}{\partial \beta}\left[\frac{\partial^2 G_i}{\partial m_j \partial m_k}\right]^*(m_j - m_j^*)(m_k - m_k^*) - \frac{1}{2}\sum_{j,k,l}\frac{\partial m_l^*}{\partial \beta}\left[\frac{\partial^3 G_i}{\partial m_j \partial m_k \partial m_l}\right]^*(m_j - m_j^*)(m_k - m_k^*) + \mathcal{O}(\delta^3 m). \tag{S2}$$

Now, we calculate each derivative in Eq. (S2):

(i)

$$-\frac{\partial m_k}{\partial \beta} = \frac{\partial}{\partial \beta}\left[\frac{\sum_t \sigma_k(t)\exp(J\cdot\sigma(t) - \beta E_i(t))}{\sum_t \exp(J\cdot\sigma(t) - \beta E_i(t))}\right] = \langle \delta E_i \delta\sigma_k\rangle. \tag{S3}$$

(ii)

$$\frac{\partial^2 G_i}{\partial m_j \partial m_k} = \frac{\partial J_k}{\partial m_j} = [C^{-1}]_{jk}, \tag{S4}$$

where

$$C_{jk} = \frac{\partial m_j}{\partial J_k} = \frac{\partial}{\partial J_k}\left[\frac{\sum_t \sigma_j(t)\exp(J\cdot\sigma(t) - \beta E_i(t))}{\sum_t \exp(J\cdot\sigma(t) - \beta E_i(t))}\right] = \langle \delta\sigma_j \delta\sigma_k\rangle. \tag{S5}$$

(iii)

$$\frac{\partial}{\partial \beta}\left[\frac{\partial^2 G_i}{\partial m_j \partial m_k}\right] = \frac{\partial}{\partial \beta}[C^{-1}]_{jk} = -\sum_{\mu,\nu}[C^{-1}]_{j\mu}\frac{\partial C_{\mu\nu}}{\partial \beta}[C^{-1}]_{\nu k} = \sum_{\mu,\nu}[C^{-1}]_{j\mu}[C^{-1}]_{k\nu}\langle \delta E_i \delta\sigma_\mu\sigma_\nu\rangle. \tag{S6}$$

(iv)

$$\frac{\partial^3 G_i}{\partial m_j \partial m_k \partial m_l} = \frac{\partial}{\partial m_j}[C^{-1}]_{kl} = \sum_\lambda \frac{\partial J_\lambda}{\partial m_j}\frac{\partial}{J_\lambda}[C^{-1}]_{kl} = -\sum_{\lambda,\mu,\nu}[C^{-1}]_{j\lambda}[C^{-1}]_{k\mu}\frac{\partial C_{\mu\nu}}{\partial J_\lambda}[C^{-1}]_{\nu l} = -\sum_{\lambda,\mu,\nu}[C^{-1}]_{j\lambda}[C^{-1}]_{k\mu}[C^{-1}]_{l\nu}\langle \delta\sigma_\lambda\delta\sigma_\mu\sigma_\nu\rangle. \tag{S7}$$

Plugging these derivatives into Eq. (S2), we obtain the following equation up to second order in δm:

$$\begin{aligned}\langle \delta E_i\rangle = &\sum_{j,k}\langle \delta E_i\delta\sigma_k\rangle^*[C^{-1}]^*_{kj}\langle\delta\sigma_j\rangle \\ &+ \frac{1}{2}\sum_{j,k}\sum_{\mu,\nu}\langle \delta E_i\delta\sigma_\mu\sigma_\nu\rangle^*[C^{-1}]^*_{j\mu}[C^{-1}]^*_{k\nu}\langle\delta\sigma_j\rangle\langle\delta\sigma_k\rangle \\ &- \frac{1}{2}\sum_{j,k,l}\sum_{\lambda,\mu,\nu}\langle \delta E_i\delta\sigma_l\rangle^*\langle\delta\sigma_\lambda\delta\sigma_\mu\sigma_\nu\rangle^* \\ &\qquad\times[C^{-1}]^*_{j\lambda}[C^{-1}]^*_{k\mu}[C^{-1}]^*_{l\nu}\langle\delta\sigma_j\rangle\langle\delta\sigma_k\rangle.\end{aligned} \tag{S8}$$

Finally, we obtain the following relation:

$$\langle \delta E_i\rangle = \sum_j W^*_{ij}\langle\delta\sigma_j\rangle + \frac{1}{2}\sum_{j,k}Q^*_{ijk}\langle\delta\sigma_j\rangle\langle\delta\sigma_k\rangle, \tag{S9}$$

where

$$W^*_{ij} \equiv \sum_k \langle \delta E_i\delta\sigma_k\rangle^*[C^{-1}]^*_{kj} \tag{S10}$$

and

$$\begin{aligned}Q^*_{ijk} \equiv &\sum_{\mu,\nu}\langle \delta E_i\delta\sigma_\mu\sigma_\nu\rangle^*[C^{-1}]^*_{j\mu}[C^{-1}]^*_{k\nu} \\ &- \sum_l\sum_{\lambda,\mu,\nu}\langle \delta E_i\delta\sigma_l\rangle^*\langle\delta\sigma_\lambda\delta\sigma_\mu\sigma_\nu\rangle^*[C^{-1}]^*_{j\lambda}[C^{-1}]^*_{k\mu}[C^{-1}]^*_{l\nu}.\end{aligned} \tag{S11}$$

The second term in Eq. (S9) can be approximated as

$$\begin{aligned}\langle\delta\sigma_j\rangle\langle\delta\sigma_k\rangle &= \big(\langle\sigma_j\rangle - \langle\sigma_j\rangle^*\big)\big(\langle\sigma_k\rangle - \langle\sigma_k\rangle^*\big) \\ &\approx \langle\sigma_j\sigma_k\rangle - \langle\sigma_j\sigma_k\rangle^* \\ &\quad - \langle\sigma_j\rangle^*\big(\langle\sigma_k\rangle - \langle\sigma_k\rangle^*\big) - \langle\sigma_k\rangle^*\big(\langle\sigma_j\rangle - \langle\sigma_j\rangle^*\big) \\ &= \langle\delta(\sigma_j\sigma_k)\rangle - \langle\sigma_j\rangle^*\langle\delta\sigma_k\rangle - \langle\sigma_k\rangle^*\langle\delta\sigma_j\rangle,\end{aligned} \tag{S12}$$

where the second line assumes a negligible correlation between σ_j and σ_k: $\langle\sigma_j\sigma_k\rangle \approx \langle\sigma_j\rangle\langle\sigma_k\rangle$. Then, with the Rao-Blackwell conditional expectation update $H_i(m)^{\text{new}} \leftarrow \langle E_i\rangle_{J(m^*)}$, Eq. (S9) implies

$$H_i = \sum_j\left(W^*_{ij} - \sum_k Q^*_{ijk}\langle\sigma_k\rangle^*\right)\sigma_j + \frac{1}{2}\sum_{j,k}Q^*_{ijk}\sigma_j\sigma_k, \tag{S13}$$

where we used $Q_{ijk} = Q_{ikj}$. This formalism allows one to infer the linear and quadratic relations between H_i and σ.

SI Text 2: Review on the mean-field methods for the kinetic Ising model

Maximum likelihood estimation (MLE)

The kinetic Ising model updates spins with the conditional probability,

$$P(\sigma_i(t+1) = \pm 1|\sigma(t)) = \frac{\exp(\pm H_i(\sigma(t)))}{\exp(H_i(\sigma(t))) + \exp(-H_i(\sigma(t)))}, \tag{S14}$$

where $H_i(\sigma(t)) = \sum_j W_{ij}\sigma_j(t)$. Then, the expectation value of $\sigma_i(t+1)$ given $\sigma(t)$ becomes

$$\begin{aligned}\langle\langle \sigma_i(t+1)\rangle\rangle_{\sigma(t)} &= \sum_{\rho=\{1,-1\}} \rho P(\sigma_i(t+1) = \rho|\sigma(t)) \\ &= \tanh(H_i(\sigma(t))).\end{aligned} \tag{S15}$$

Given N-dimensional time-series data $\sigma(t)$ with length L, the data likelihood is defined as

$$\mathcal{P} = \prod_{t=1}^{L-1}\prod_{i=1}^{N} P(\sigma_i(t+1)|\sigma(t)). \tag{S16}$$

Using MLE, one can optimize W_{ij} to increase $\log \mathcal{P}$:

$$W_{ij}^{\text{new}} = W_{ij} + \frac{\alpha}{L-1}\frac{\partial \log \mathcal{P}}{\partial W_{ij}} \tag{S17}$$

with a learning rate α [29]. Here, one can calculate the gradient with Eq. (S14),

$$\frac{\partial \log \mathcal{P}}{\partial W_{ij}} = \sum_{t=1}^{L-1}\Big(\sigma_i(t+1)\sigma_j(t) - \tanh\big(H_i(\sigma(t))\big)\sigma_j(t)\Big). \tag{S18}$$

Naïve mean-field approximation (nMF)

The maximum condition of the log-likelihood ($\partial \log \mathcal{P}/\partial W_{ij}$=0) in Eq. (S18) gives

$$\sum_{t=1}^{L-1}\sigma_i(t+1)\sigma_j(t) = \sum_{t=1}^{L-1}\tanh\big(H_i(\sigma(t))\big)\sigma_j(t) \tag{S19}$$

with $H_i(\sigma(t)) = \sum_k W_{ik}\sigma_k(t)$. For a mean-field approximation, spin activities are represented by the mean field activity plus its residual: $\sigma_i(t) = m_i + \delta\sigma_i(t)$. Then, using the Taylor expansion, one can approximate $\tanh\big(H_i(\sigma(t))\big) \approx \tanh(g_i) + \big(1 - \tanh^2(g_i)\big)\sum_k W_{ik}\delta\sigma_k(t)$ with $g_i = \sum_k W_{ik}m_k$. The zeroth-order expectation of $\langle\langle\sigma_i(t+1)\rangle\rangle_{\sigma(t)} \approx \tanh(g_i)$ gives the self-consistent equation

$$m_i = \tanh\big(\sum_k W_{ik}m_k\big). \tag{S20}$$

Then, using the mean-field approximation, Eq. (S19) becomes

$$\begin{aligned}&\sum_t \big(m_i + \delta\sigma_i(t+1)\big)\big(m_j + \delta\sigma_j(t)\big) \\ &= \sum_t \big(m_i + (1-m_i^2)\sum_k W_{ik}\delta\sigma_k(t)\big)\big(m_j + \delta\sigma_j(t)\big)\end{aligned} \tag{S21}$$

Given the data with length L,

$$\begin{aligned}&\frac{1}{L-1}\sum_{t=1}^{L-1}\delta\sigma_i(t+1)\delta\sigma_j(t) \\ &= (1-m_i^2)\sum_k W_{ik}\frac{1}{L-1}\sum_{t=1}^{L-1}\delta\sigma_k(t)\delta\sigma_j(t).\end{aligned} \tag{S22}$$

One can also derive this equation from $\delta\sigma_i(t+1) = (\partial m_i/\partial m_k)\delta\sigma_k(t)$ with Eq. (S20). The equality gives a matrix equation to infer

$$W_{\text{nMF}} = A_{\text{nMF}}^{-1}BC^{-1}, \tag{S23}$$

where $[A_{\text{nMF}}]_{ij} = (1-m_i^2)\delta_{ij}$ is a diagonal matrix; $B_{ij} = \langle\delta\sigma_i(t+1)\delta\sigma_j(t)\rangle$ is a time-delayed correlation; and the covariance matrix $C_{ij} = \langle\delta\sigma_i(t)\delta\sigma_j(t)\rangle$ is an equal-time correlation [27].

Thousless-Anderson-Palmer mean-field approximation (TAP)

Compared to nMF, TAP considers the second-order correction of the Onsager's reaction term:

$$\begin{aligned}\langle\sigma_i(t+1)\rangle &= \big\langle\tanh\big(\sum_k W_{ik}\sigma_k(t)\big)\big\rangle \\ &\approx \tanh(g_i) + \frac{1}{2}\Big[\frac{\partial^2\tanh(x)}{\partial x^2}\Big]_{x=g_i}\langle\delta g_i^2\rangle \\ &\approx \tanh(g_i) - \big(1-\tanh^2(g_i)\big)\tanh(g_i)\sum_l W_{il}^2(1-m_l^2)\end{aligned} \tag{S24}$$

with $g_i \equiv \sum_k W_{ik}m_k$, $\delta g_i \equiv \sum_k W_{ik}\delta\sigma_k(t)$, and

$$\begin{aligned}\langle\delta g_i^2\rangle &= \sum_{k,l} W_{ik}W_{il}\langle\delta\sigma_k\delta\sigma_l\rangle = \sum_l W_{il}^2\langle\delta\sigma_l^2\rangle \\ &= \sum_l W_{il}^2\langle(\sigma_l - m_l)^2\rangle = \sum_l W_{il}^2(1-m_l^2)\end{aligned} \tag{S25}$$

under the assumption of the negligible correlation between σ_k and σ_l: $\langle\delta\sigma_k\delta\sigma_l\rangle \approx 0$ for $k \neq l$. The correction gives a refined self-consistent equation

$$m_i = \tanh\big(\sum_k W_{ik}m_k - m_i\sum_l W_{il}^2(1-m_l^2)\big). \tag{S26}$$

Then, using $\delta\sigma_i(t+1) = (\partial m_i/\partial m_k)\delta\sigma_k(t)$, one can derive

$$\delta\sigma_i(t+1) = (1-m_i^2)(1-F_i)\sum_k W_{ik}\delta\sigma_k(t) \quad \text{(S27)}$$

with $F_i \equiv (1-m_i^2)\sum_l W_{il}^2(1-m_l^2)$. This leads to

$$\langle\delta\sigma_i(t+1)\delta\sigma_j(t)\rangle = (1-m_i^2)(1-F_i)\sum_k W_{ik}\langle\delta\sigma_k(t)\delta\sigma_j(t)\rangle. \quad \text{(S28)}$$

Therefore, one obtains the TAP estimates

$$W_{\rm TAP} = (1-F_i)^{-1}W_{\rm nMF}. \quad \text{(S29)}$$

Here, one can obtain F_i as a solution of the self-consistent equation [27]:

$$F_i(1-F_i)^2 = (1-m_i^2)\sum_l [W_{\rm nMF}]_{il}^2(1-m_l^2). \quad \text{(S30)}$$

Exact mean-field approximation (eMF)

For random W_{ik} with a large number N of spin components, it is a reasonable assumption that $H_i = \sum_{k=1}^N W_{ik}\sigma_k$ follows a Gaussian distribution with a mean $g_i = \sum_k W_{ik}m_k$ and a variance $\Delta_i = \langle\delta g_i^2\rangle = \sum_l W_{il}(1-m_l^2)$ in Eq. (S25):

$$\langle\delta\sigma_i(t+1)\rangle = \int_{-\infty}^{\infty} dx \frac{e^{-x^2}}{\sqrt{2\pi}}\tanh(g_i + x\sqrt{\Delta_i}). \quad \text{(S31)}$$

Here, the zeroth-order and second-order Taylor expansion of $\tanh(g_i + x\sqrt{\Delta_i})$ with respect to x give the nMF and TAP solutions in Eqs. (S20) and (S26). The multivariable $x \equiv \delta g_i$ and $y \equiv \delta g_j$ may also follow a Gaussian distribution:

$$P(x,y) = \frac{1}{2\pi\sqrt{\Delta_i\Delta_j}}\exp\left[-\frac{x^2}{2\Delta_i} - \frac{y^2}{2\Delta_j} + \Delta_{ij}\frac{xy}{\Delta_i\Delta_j}\right], \quad \text{(S32)}$$

where the covariance Δ_{ij} is defined as

$$\begin{aligned}\Delta_{ij} \equiv \langle\delta g_i\delta g_j\rangle &= \left\langle \sum_k W_{ik}\delta\sigma_k \sum_l W_{jl}\delta\sigma_l\right\rangle \\ &= \sum_{k,l} W_{ik}\langle\delta\sigma_k\delta\sigma_l\rangle W_{lj}^\top = [WCW^\top]_{ij}. \quad \text{(S33)}\end{aligned}$$

Then, the time-delayed correlation matrix B can be approximated as

$$\begin{aligned}B_{ik} &= \langle\delta\sigma_i(t+1)\delta\sigma_k(t)\rangle \\ &= \langle\sigma_i(t+1)\sigma_k(t)\rangle - \langle\sigma_i(t+1)\rangle\langle\sigma_k(t)\rangle \\ &= \langle\sigma_k(t)\tanh(g_i+\delta g_i(t))\rangle - \langle\sigma_k(t)\rangle\langle\tanh(g_i+\delta g_i(t))\rangle. \quad \text{(S34)}\end{aligned}$$

Using B, one can derive $BW^\top = AWCW^\top$ as follows:

$$\begin{aligned}\sum_k W_{jk}B_{ik} &= \big\langle \sum_k W_{jk}\sigma_k(t)\tanh(g_i+\delta g_i(t))\big\rangle \\ &\quad - \big\langle \sum_k W_{jk}\sigma_k(t)\big\rangle\big\langle\tanh(g_i+\delta g_i(t))\big\rangle \\ &= \langle(g_j+\delta g_j(t))\tanh(g_i+\delta g_i(t))\rangle \\ &\quad - \langle g_j+\delta g_j(t)\rangle\langle\tanh(g_i+\delta g_i(t))\rangle \\ &= \langle\delta g_j\tanh(g_i+\delta g_i)\rangle \\ &= \int_{-\infty}^{\infty}\frac{dxdy}{2\pi\sqrt{\Delta_i\Delta_j}} y\tanh(g_i+x) \\ &\quad\quad \times\exp\left[-\frac{x^2}{2\Delta_i} - \frac{y^2}{2\Delta_j} + \Delta_{ij}\frac{xy}{\Delta_i\Delta_j}\right] \\ &\approx \frac{\Delta_{ij}}{\Delta_i\Delta_j}\int_{-\infty}^{\infty}\frac{dx}{\sqrt{2\pi\Delta_i}}\frac{dy}{\sqrt{2\pi\Delta_j}} xy^2\tanh(g_i+x) \\ &\quad\quad \times\exp\left[-\frac{x^2}{2\Delta_i} - \frac{y^2}{2\Delta_j}\right] \\ &= \frac{\Delta_{ij}}{\Delta_i}\int_{-\infty}^{\infty}\frac{dx}{\sqrt{2\pi\Delta_i}} x\tanh(g_i+x)\exp\left[-\frac{x^2}{2\Delta_i}\right] \\ &= \Delta_{ij}\int_{-\infty}^{\infty}\frac{dx}{\sqrt{2\pi\Delta_i}}\exp\left[-\frac{x^2}{2\Delta_i}\right]\big(1-\tanh^2(g_i+x)\big) \\ &= [WCW^\top]_{ij}a_i. \quad \text{(S35)}\end{aligned}$$

This equation gives

$$W_{\rm eMF} = A_{\rm eMF}^{-1}BC^{-1}, \quad \text{(S36)}$$

where $[A_{\rm eMF}]_{ij} = a_i\delta_{ij}$ is a diagonal matrix. In practice, one can obtain $W_{\rm eMF}$ with the following iterations [28]:

(i) Calculate Δ_i (Gguess Δ_i for the first round):

$$\Delta_i = \frac{1}{a_i^2}\sum_j [BC^{-1}]_{ij}^2(1-m_j^2). \quad \text{(S37)}$$

(ii) Find g_i as a solution for the following integral equation:

$$m_i = \int_{-\infty}^{\infty}\frac{dx}{\sqrt{2\pi}}\exp\left[-\frac{x^2}{2}\right]\tanh(g_i + x\sqrt{\Delta_i}). \quad \text{(S38)}$$

(iii) Calculate a_i given g_i and Δ_i:

$$a_i = \int_{-\infty}^{\infty}\frac{dx}{\sqrt{2\pi}}\exp\left[-\frac{x^2}{2}\right]\big(1-\tanh^2(g_i+x\sqrt{\Delta_i})\big). \quad \text{(S39)}$$

SI Text 3: Neuronal data processing

In the original data, neuron i is defined as "active" ($\sigma_i(t) = 1$), if the neuron fires at least once during the

time window $[t, t + \delta t]$, otherwise "silent" ($\sigma_i(t) = -1$) (Fig. S1, upper). To suppress the dependency of the time interval δt for the activity definition, we used a moving average of activities. We examined the past five and future five activities of neuron i, and redefined $\sigma_i(t) = 1$, if neuron i emitted at least one spike in the time window, otherwise $\sigma_i(t) = -1$ (Fig. S1, lower). Since neurons may have a refractory period that prevents consecutive spikes after emitting a spike [41], the moving average can also help infer the genuine interaction between neurons by reducing the effect of the refractory period.

For the estimation of W_{ij} and $H_i^{\rm ext}$, we estimated W_{ij} first with $H_i^{\rm ext} = 0$, and then estimated W_{ij} and $H_i^{\rm ext}$ together because $H_i^{\rm ext}$ turned out to be quite large compared to W_{ij}. These training procedures were repeated for 20 times.

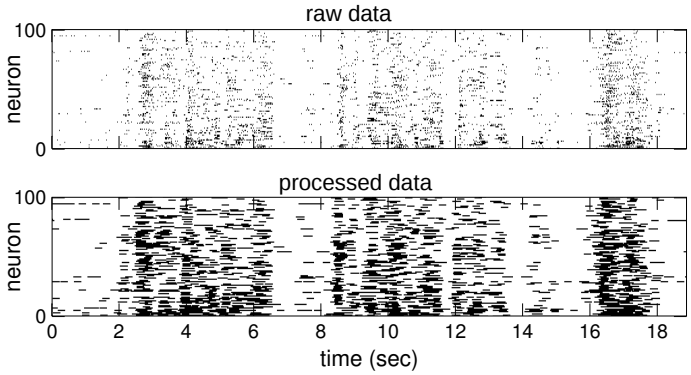

FIG. S1. Data processing. Rasters of 100 neuronal activities from raw (upper) and processed (lower) data are plotted with black dots representing spikes, white dots representing queiscence.

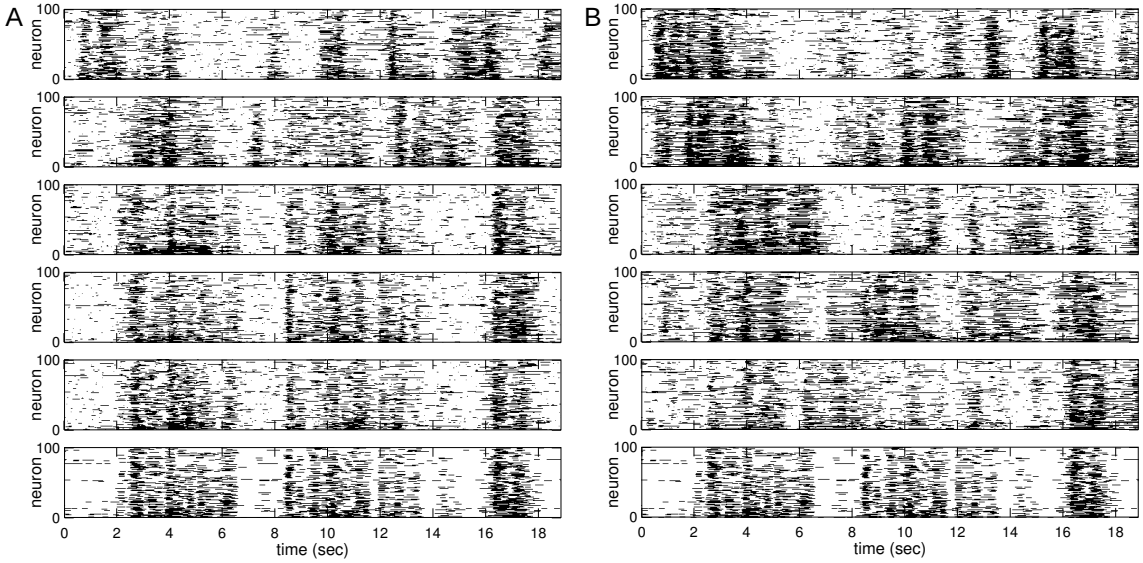

FIG. S2. Reconstruction of neuronal activity. The raster of neuronal activities is recovered using our method, with large-$|W_{ij}|$-based selection (A) and under random selection (B) with various numbers of input neurons, from top to bottom: 1, 2, 8, 10, 12. The actual raster is also shown on the bottom of each column for comparison.

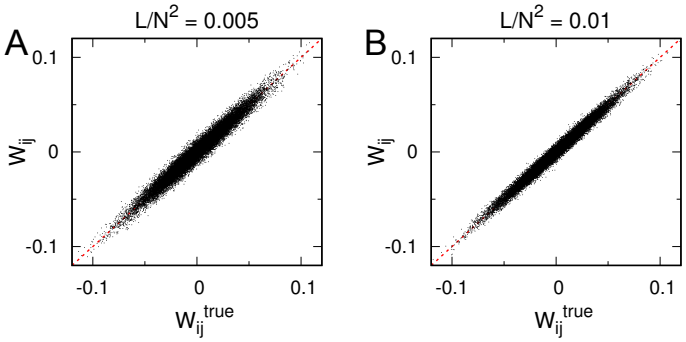

FIG. S3. Inference of coupling strength in large system size, $N = 5000$. Predicted couplings versus actual couplings for $L/N^2 = 0.005$ (A) and 0.01 (B). The actual coupling strengths are normally-distributed with $g = 2$. The computation time for this simulation is approximately 4 days and 8 days, respectively, for $L/N^2 = 0.005$ and 0.01 on a 2.30 GHz processor.