Inference of stochastic time series with missing data

Sangwon Lee [1,*] Vipul Periwal,[2,†] and Junghyo Jo [3,4,‡]

[1]*Department of Physics and Astronomy, Seoul National University, Seoul 08826, Korea*

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

[3]*Department of Physics Education and Center for Theoretical Physics and Artificial Intelligence Institute, Seoul National University, Seoul 08826, Korea*

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

(Received 28 January 2021; revised 2 June 2021; accepted 22 July 2021; published 16 August 2021)

Inferring dynamics from time series is an important objective in data analysis. In particular, it is challenging to infer stochastic dynamics given incomplete data. We propose an expectation maximization (EM) algorithm that iterates between alternating two steps: E-step restores missing data points, while M-step infers an underlying network model from the restored data. Using synthetic data of a kinetic Ising model, we confirm that the algorithm works for restoring missing data points as well as inferring the underlying model. At the initial iteration of the EM algorithm, the model inference shows better model-data consistency with observed data points than with missing data points. As we keep iterating, however, missing data points show better model-data consistency. We find that demanding equal consistency of observed and missing data points provides an effective stopping criterion for the iteration to prevent going beyond the most accurate model inference. Using the EM algorithm and the stopping criterion together, we infer missing data points from a time-series data of real neuronal activities. Our method reproduces collective properties of neuronal activities such as correlations and firing statistics even when 70% of data points are masked as missing points.

DOI: 10.1103/PhysRevE.104.024119

I. INTRODUCTION

System identification is one of the most important tasks in data science [1]. To be specific, suppose we have observations $\{\vec{\sigma}(t)\}$ of a sample $\vec{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_N)$, where t can denote a time index for time-series data, or a sample index for independent data. Considering stochastic systems, statistical mechanics has been adopted to construct the least structured probabilistic model of $P(\vec{\sigma})$ from observations. In particular, Ising-like models of $P(\vec{\sigma}) \propto \exp\left(\sum_{i,j} W_{ij}\sigma_i\sigma_j\right)$ incorporate the pairwise interactions between variables. Its applicability ranges from neuroscience and biology [2–7] to economics and sociology [8–10]. Such pairwise interactions are generally sufficient to explain complex higher-order patterns in many cases [2,11]. To consider time-dependent data arising from kinetic interactions, another type of Ising model has been proposed [12]. Unlike the equilibrium model of $P(\vec{\sigma})$, the kinetic Ising model has a probabilistic relation between $\sigma_i(t+1)$ and $\sigma_j(t)$ with the conditional probability of $P[\sigma_i(t+1)|\vec{\sigma}(t)] \propto \exp\left[\sum_j W_{ij}\sigma_i(t+1)\sigma_j(t)\right]$. The kinetic model has been used to reconstruct neural networks from temporal neuronal activities [13,14].

Both equilibrium and kinetic (or nonequilibrium) models have network parameters W_{ij}. In the equilibrium model, it is symmetric ($W_{ij} = W_{ji}$) to represent undirected correlation between σ_i and σ_j. However, in the kinetic model, W_{ij} is not necessarily symmetric ($W_{ij} \neq W_{ji}$) as it represents directed causality from $\sigma_j(t)$ to $\sigma_i(t+1)$. Due to the wide applications of these models, it is an important inverse problem to infer W_{ij} from observations $\{\vec{\sigma}(t)\}$. As concretely established in the equilibrium model [15], the kinetic model also has many inference methods including various mean-field approximations [13,16], maximum likelihood estimation [13,14], and the recent expectation-reflection (ER) method [17].

In real-world problems, it is common that only a part of a network is observable. For example, it is impossible to observe every neuron in the brain. Therefore, one should consider not only observed visible units but also unobserved hidden units. Much effort has been devoted to infer both hidden variables and the network parameters [18–22]. These methods basically rely on the expectation-maximization (EM) algorithm [23]. The EM method is composed of two iterative steps: E-step predicts hidden variables by using mean-field approximations [18,19] or replica-based approaches [20,21], or a likelihood-based method [22], and M-step optimizes the network parameters from the reconstrcted data.

In addition to the issue of hidden units, another practical issue is that even visible units are not always observable throughout an experiment. At each time point, some units become accessible to observers and the others become inaccessible. For example, a large-scale neural network can be partially scannable in neuroscience experiments [24]. This scenario is also common in finance and social science. For a

*Present address: Max Planck School Matter to Life, University of Göttingen, Göttingen 37077, Germany.

†vipulp@mail.nih.gov

‡jojunghyo@snu.ac.kr

trading network, trade records are available only when traders are active [25]. Although some pioneering work exists, research on the kinetic Ising model with missing or partially masked data is still in its infancy. Campajola *et al.* [26] addressed this issue inspired by the mean-field approach of Dunn *et al.* [18] that was originally developed for the problem of hidden units. The mean-field approach imposes an *a priori* assumption that interactions are weak and dense [19]. Thus, it is imperative to develop an approach free from such constraints. Of course, any such method has to be tested against real-world problems beyond the proof-of-concept with synthetic data.

Here we develop a general method to infer the parameters of a kinetic Ising model from data with sporadic missing values of any measured variable, extending our recent study on hidden units [22]. The algorithm uses the EM method: the E-step restores missing time points stochastically using a likelihood ratio, while the M-step infers network parameters from observed and restored data. Approximating the E-step with a stochastic realization is called the stochastic approximation EM (SAEM) [27]. By repeating the alternation between E- and M-steps, the algorithm infers network parameters. However, too much iteration can overfit the model, preventing accurate inference of model parameters. It is therefore crucial to find a criterion for stopping the EM iterations at the right time. We find an effective stopping criterion for such an optimal inference. It is based on the intuition that once missing data is properly restored, there should not be any distinction between observed and restored data. Therefore, we stop the iteration when the stochastic model shows equal uncertainty for the prediction of observed and missing data points. Using the EM method with such an optimal number of iterations, we demonstrate that our method successfully infers interactions from synthetic data of the kinetic Ising model. We then apply the method to infer the kinetics of a neuronal network from real neuronal activity data [28,29], and confirm that the inference reproduces collective behaviors such as time correlation and firing statistics of neuronal activities.

This paper is organized as follows. In Sec. II, we describe our EM-based method and the stopping criterion for optimal iterations. In Sec. III, we validate this method with simulated data using the kinetic Ising model. We then apply the method to infer dynamics from a recording of neuronal activity, and compare the inference performance with equilibrium and simple imputation models. Finally, we summarize and discuss our findings in Sec. IV. To keep the paper self-contained, we briefly introduce the mean-field method developed by Campajola *et al.* [26] as a benchmark of our method in Appendix A. We also consider various conditions to validate our method in Appendix B. Furthermore, we explain the equilibrium model in Appendix C, and describe additional experimental data on neuronal activities in Appendix D.

II. METHOD

We consider stochastic dynamics of the kinetic Ising model with N binary variables. The ith spin $\sigma_i(t+1)$ at time $t+1$ is stochastically determined by the conditional probability:

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

where the local field $H_i(t)=\sum_j W_{ij}\sigma_j(t)+b_i$ integrates the influence of connected spins as well as external bias b_i. We denote the model parameters as $\theta=(W_{ij},b_i)$. Here we choose a synchronous update for simplicity; refer to Refs. [14,30] for an asynchronous update. Given binary time-series data $\{\vec{\sigma}(t)\}_{t=0}^{L}$ of length $L+1$, various methods exist to infer θ in Eq. (1) [13,15–17]. In particular, the ER method provides faster inference than the standard maximum likelihood estimation, since the iterative algorithm is based on a multiplicative and parallelizable parameter update [22]. In this study, however, we used standard logistic regression included in the Python package, *scikit-learn* [31], because logistic regression shows similar performance with the ER method with high accuracy and fast computation. Note that Eq. (1) implies the logistic function:

$$P[\sigma_i(t+1)=1|\vec{\sigma}(t),\theta]=\frac{1}{1+\exp\left[-2\sum_j W_{ij}\sigma_j(t)-2b_i\right]}. \tag{2}$$

In the absence of regularization, the logistic regression is equivalent to the maximization of a total likelihood of data,

$$\mathcal{L}_{\rm tot}(\theta)=\prod_{i=1}^{N}\prod_{t=0}^{L-1}P[\sigma_i(t+1)|\vec{\sigma}(t),\theta]. \tag{3}$$

Now suppose that some of the data points are missing. Let $\mathcal{M}$ denote the set of missing data points. We explicitly distinguish between missing and observed data points by denoting $\sigma_i^m(t)$ for $(i,t)\in\mathcal{M}$ and $\sigma_i^o(t)$ for $(i,t)\notin\mathcal{M}$. However, when the distinction is not necessary, we omit the superscripts m and o. To recover these missing values, we first assign random binary values for $\sigma_i^m(t)$ at every $(i,t)\in\mathcal{M}$. Then, from the observed and randomly assigned values of $\{\vec{\sigma}(t)\}_{t=0}^{L}$, we can infer an initial value of θ using logistic regression in Eq. (2), or maximizing $\mathcal{L}_{\rm tot}(\theta)$ in Eq. (3). The next step is to update the missing data points. Previous EM-based algorithms used a mean-field approach to approximate missing data points as mean values [18,26]. For the E-step, however, we stochastically assign missing data points following SAEM. The likelihood of missing values $\sigma_i^m(t)=\pm 1$ can be derived from the total likelihood $\mathcal{L}_{\rm tot}$ as follows:

$$\mathcal{L}_{i,t}^{\pm}\equiv P\left[\sigma_i^m(t)=\pm 1|\vec{\sigma}(t-1),\theta\right]\times\prod_{j=1}^{N}P[\sigma_j(t+1)|F_i^{\pm}[\vec{\sigma}(t)],\theta], \tag{4}$$

where $F_i^{\pm}[\vec{\sigma}(t)]=[\sigma_1(t),\cdots,\sigma_i^m(t)=\pm 1,\cdots,\sigma_N(t)]$ for $(i,t)\in\mathcal{M}$. $\mathcal{L}_{i,t}^{\pm}$ is a product of the likelihoods determined by the one-step backward state of $\vec{\sigma}(t-1)$ and the one-step forward state $\vec{\sigma}(t+1)$. The likelihoods for $t=0$ and L involve only forward and backward states, respectively: $\mathcal{L}_{i,0}^{\pm}\equiv\prod_{j=1}^{N}P\{\sigma_j(1)|F_i^{\pm}[\vec{\sigma}(0)],\theta\}$ and $\mathcal{L}_{i,L}^{\pm}\equiv P[\sigma_i^m(L)=\pm 1|\vec{\sigma}(L-1),\theta]$. Using these likelihood values, we stochastically re-assign ± 1 to $\sigma_i^m(t)$ with a probability of

$\mathcal{L}^{\pm}_{i,t}/(\mathcal{L}^{+}_{i,t}+\mathcal{L}^{-}_{i,t})$ for every missing data point of $(i,t)\in\mathcal{M}$ with random order. Alternatively, we have updated the missing data points with a Metropolis-like manner [32], and confirmed no noticeable changes (data not shown). Also note that updating one missing point is affected by other missing points that may or may not have been updated in that step.

After updating all missing data points, we optimize θ from the observed and restored data $\sigma_i(t)$ by maximizing $\mathcal{L}_{\rm tot}(\theta)$ in Eq. (3) (M-step). Repeating these E- and M-steps, θ is expected to converge to the true parameter values. However, excess iteration ends up with worse inference for θ, especially when a larger fraction of data is missing, a well-known issue in the latent variable field [33]. How to stop the EM iteration at a right time?

To address this issue, we consider model-data consistency. It is not entirely straightforward to check consistency in stochastic models. In the kinetic Ising model, the future state $\sigma_i(t+1)$ probabilistically depends on the current state $\vec{\sigma}(t)$. Therefore, we consider the discrepancy between $\sigma_i(t+1)$ and its expectation value,

$$\begin{aligned}\mathbb{E}[\sigma_i(t+1)] &= \sigma_i(t+1)P[\sigma_i(t+1)|\vec{\sigma}(t),\theta]\\ &\quad-\sigma_i(t+1)\{1-P[\sigma_i(t+1)|\vec{\sigma}(t),\theta]\}\\ &= \sigma_i(t+1)\{2P[\sigma_i(t+1)|\vec{\sigma}(t),\theta]-1\},\end{aligned}\tag{5}$$

as a quality-of-fit measure. The expectation value is simply obtained as $\mathbb{E}[\sigma_i(t+1)]=\tanh H_i(t)$ for the kinetic Ising model following the conditional probability of Eq. (1). Here the mean discrepancy for every data point can be quantified as

$$\begin{aligned}D &= \frac{1}{NL}\sum_{i=1}^{N}\sum_{t=0}^{L-1}\{\sigma_i(t+1)-\mathbb{E}[\sigma_i(t+1)]\}^2\\ &= \frac{1}{NL}\sum_{i=1}^{N}\sum_{t=0}^{L-1}\sigma_i^2(t+1)\{2-2P[\sigma_i(t+1)|\vec{\sigma}(t),\theta]\}^2\\ &= \frac{4}{NL}\sum_{i=1}^{N}\sum_{t=0}^{L-1}\{1-P[\sigma_i(t+1)|\vec{\sigma}(t),\theta]\}^2.\end{aligned}\tag{6}$$

For the derivation of the second line in Eq. (6), we used Eq. (5). The discrepancy D effectively measures the loss of the likelihood $\mathcal{L}_{\rm tot}$ of data $\{\vec{\sigma}(t)\}_{t=0}^{L}$, as large D corresponds to small $\mathcal{L}_{\rm tot}$ in general. Now we separately examine the model-data discrepancy for observed and missing data points by defining two measures:

$$\begin{aligned}D_{\rm obs} &= \frac{1}{NL-|\mathcal{M}|}\sum_{(i,t+1)\notin\mathcal{M}}\{\sigma_i(t+1)-\mathbb{E}[\sigma_i(t+1)]\}^2,\\ D_{\rm mis} &= \frac{1}{|\mathcal{M}|}\sum_{(i,t+1)\in\mathcal{M}}\{\sigma_i(t+1)-\mathbb{E}[\sigma_i(t+1)]\}^2,\end{aligned}\tag{7}$$

where $|\mathcal{M}|$ is the set size of missing data points. Note that we assume that initial data points do not include missing values as $(i,t=0)\notin\mathcal{M}$. During the iterations of the EM steps, both $D_{\rm obs}$ and $D_{\rm mis}$ keep decreasing because the M-step optimizes θ to minimize the model-data discrepancy. However, the relative size of $D_{\rm obs}$ and $D_{\rm mis}$ changes with iterations as follows. At the beginning, $D_{\rm mis}$ is larger than $D_{\rm obs}$ because the missing data points of $\sigma_i^m(t)$ are just randomly assigned. After some iterations, however, the missing data points can become overly fine-tuned whereas the observed data points are always fixed. This makes $D_{\rm mis}$ smaller than $D_{\rm obs}$. Therefore, it is intuitive to halt the iterations when the $D_{\rm obs}$ and $D_{\rm mis}$ curves cross each other to avoid overestimation of model parameters. The equal model-data consistency of $D_{\rm obs}=D_{\rm mis}$ implies that the stochastic model has the same uncertainty for predicting observed and missing data points. In other words, the restoration of missing data points becomes good enough for the model not to distinguish the observed and restored data points. We conduct various numerical experiments to demonstrate the power of this stopping criterion.

Finally we summarize the overall procedure:

(i) Initialize missing data $\sigma_i^m(t)$ for $(i,t)\in\mathcal{M}$ with random binary values;

(ii) M-step: Infer model parameters θ from the whole data using the logistic regression in Eq. (2);

(iii) E-step: Re-assign the missing data $\sigma_i^m(t)$ based on the likelihood ratio $\mathcal{L}^{\pm}_{i,t}/(\mathcal{L}^{+}_{i,t}+\mathcal{L}^{-}_{i,t})$ in Eq. (4);

(iv) Repeat (ii) and (iii) until $D_{\rm mis}-D_{\rm obs}<\epsilon$ holds with a small threshold ϵ.

In this way, we can infer the model parameters, and restore missing data points.

III. RESULTS

We test the performance of our method with simulated data using the kinetic Ising model. After checking the concordance of inferred and true couplings with the simulated data, we apply the method to examine experimental recordings of neuronal activities.

A. Inference of kinetic Ising model

We simulated a stochastic time series using the Sherrington-Kirkpatrick (SK) model [34]. The SK model follows the conditional probability in Eq. (1), where the local field is defined as $H_i=\sum_j W_{ij}\sigma_j+b_i$. We turn off the bias $b_i=0$ for simplicity, and consider asymmetric interactions $W_{ij}\neq W_{ji}$. In the SK model, each value of W_{ij} is independently drawn from a Gaussian distribution, $\mathcal{N}(0,g^2/N)$, with zero mean and the variance of g^2/N. We set the overall scale of the interaction as $g=1$. With the selected $W_{ij}^{\rm true}$, we obtain $L=10\,000$ time points of $\vec{\sigma}(t)=[\sigma_1(t),\sigma_2(t),\cdots,\sigma_N(t)]$ with a system size of $N=100$. Then the total number of data values of $(i,t)\in\mathcal{A}$ has a set size $|\mathcal{A}|=(L+1)N$. We then mask some data points of $(i,t)\in\mathcal{M}$ as missing data points $\sigma_i^m(t)$, and hide the true $W_{ij}^{\rm true}$. The fraction of missing data is defined as $p=|\mathcal{M}|/|\mathcal{A}|$. Our task is two-fold: to restore probable values of missing data points $\sigma_i^m(t)$ and to infer the true $W_{ij}^{\rm true}$.

As the EM algorithm iterates, we evaluated the quality of the inference by measuring the root-mean-square error of W_{ij}, RMSE $=N^{-1}\sqrt{\sum_{i,j}(W_{ij}-W_{ij}^{\rm true})^2}$, as in Ref. [26]. Too many iterations usually lead to increasing RMSE [Fig. 1(a), left panels]. In particular, the overshooting becomes more evident when more data points are masked as missing data. At early iterations, W_{ij} is underestimated because the missing data restored with random values reduce the correlation between variables [Fig. 1(b), left panel]. However, after too

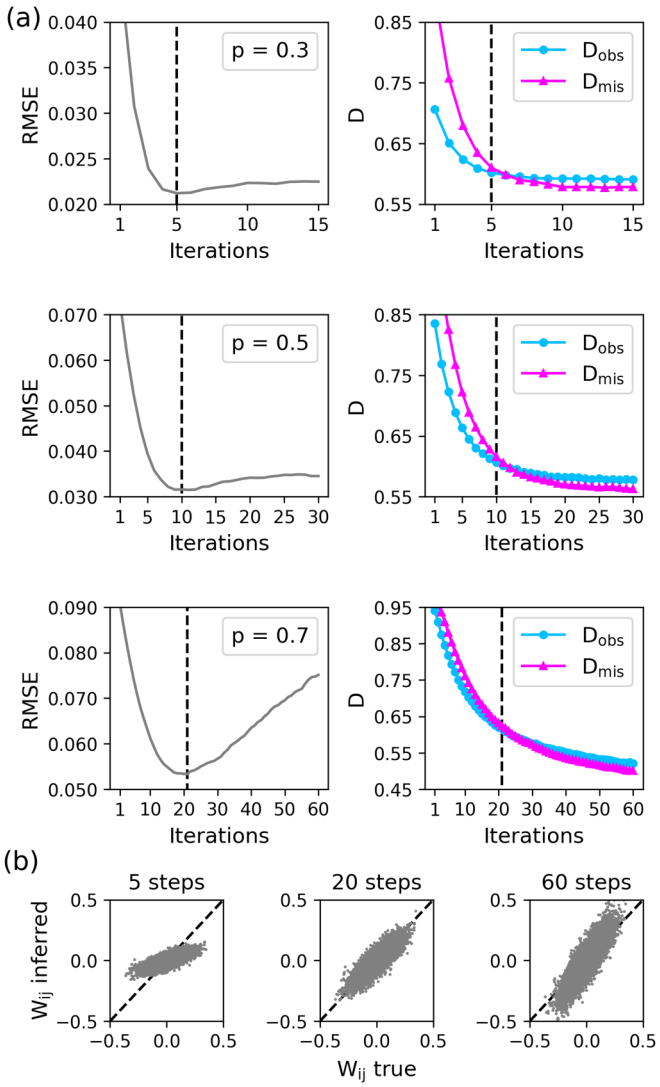

FIG. 1. EM-based inference with missing data. (a) The evolution of root-mean-square error (RMSE) of W_{ij} (left panels) and two quality measures of inference, D_{obs} and D_{mis} (right panels), over the iteration of E- and M-steps. Dashed lines denote the time when the algorithm stops according to our stopping criterion. The fraction of missing observations is $p = 0.3$, 0.5, and 0.7. (b) Comparison between the true and inferred W_{ij} at different times: 5 (left), 20 (center), and 60 (right) steps after the iteration starts, with $p = 0.7$. The correct stopping point is near 20 steps according to our stopping criterion. A system size $N = 100$ and a data length $L = 10\,000$ are used.

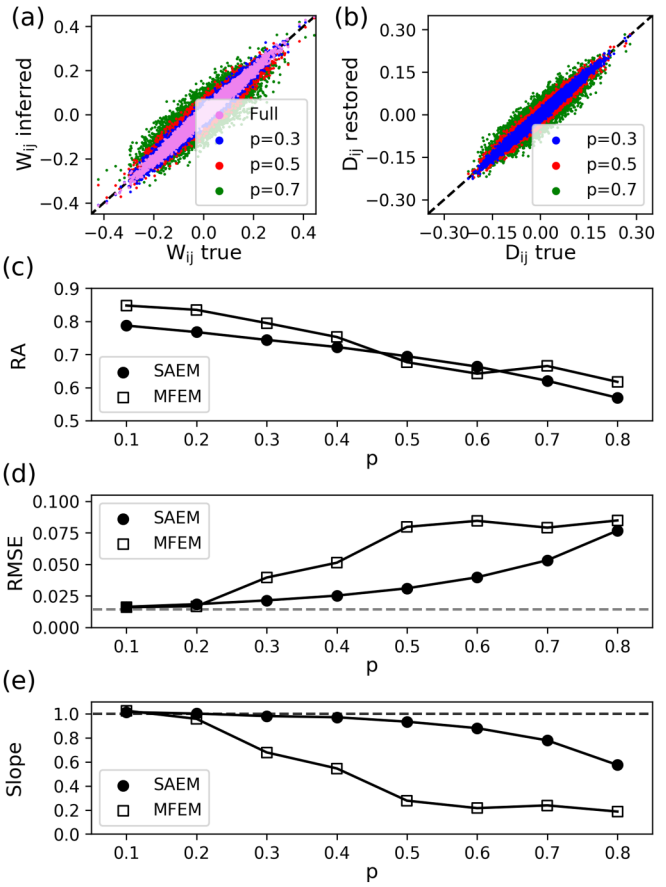

FIG. 2. Quality of inference. (a) Comparison between the true and inferred W_{ij} with a fraction of missing observations $p = 0.3$ (blue), 0.5 (red), and 0.7 (green). A result from the full data (purple) is also shown. (b) One-step lagged correlation D_{ij} obtained from true data and restored data. (c–e) Comparison of our stochastic approximation EM (SAEM, filled circles) and the mean-field EM (MFEM, empty squares) by Campajola *et al.* [26], in terms of restoration accuracy (RA) of missing data (c), RMSE of W_{ij} (d), and the slope of the linear regression between W_{ij} and W_{ij}^{true} (e). The dashed lines in (d) and (e) correspond to the RMSE and the slope inferred from the full data. Here we show the average results of three independent experiments. The standard deviation is comparable to the marker size and thus not shown. A system size $N = 100$ and a data length $L = 10\,000$ are used.

many iterations, W_{ij} is overestimated because the discrete values of the restored missing data are even better fitted to the model, with exaggerated H_i or larger values of W_{ij} (right panel). An optimal iteration is where we find a reasonably unbiased estimation of W_{ij} (center panel).

Together with the RMSE, we examined the model-data discrepancy measures of D_{obs} and D_{mis} for the observed and missing data points [Fig. 1(a), right panels]. As expected, D_{mis} is larger than D_{obs} at the beginning of iterations when the restoration of missing data points is more or less random. However, after some iterations, D_{mis} becomes smaller than D_{obs} since the restored data is excessively fine-tuned. Interestingly, the equality of D_{mis} and D_{obs} occurs near the optimal iteration that minimizes the RMSE. In practice, the crossing iteration can be monitored by an inequality condition, $D_{\text{mis}} - D_{\text{obs}} < \epsilon$ with a small positive ϵ. Here we take $\epsilon = 0.01$.

After providing evidence for this stopping criterion, we checked the performance of our algorithm with varying fraction of missing data, p. The result for $p = 0.3$ is visibly indistinguishable from the one for full data corresponding to $p = 0$. As expected, the inferred W_{ij} deviates farther from its true value as p increases [Fig. 2(a)]. Then we examined the collective behavior of data, particularly the one-step lagged correlation: $D_{ij} = \langle \sigma_i(t)\sigma_j(t+1)\rangle_t - \langle \sigma_i(t)\rangle_t \langle \sigma_j(t)\rangle_t$. Here $\langle f(t)\rangle_t \equiv 1/L \sum_{t=0}^{L-1} f(t)$ represents a time-averaged value of $f(t)$. Note that $\langle \sigma_j(t+1)\rangle_t \approx \langle \sigma_j(t)\rangle_t$ for large L.

The one-step lagged correlations are successfully recovered with our inference, even though a large fraction ($p = 0.7$) of data is missing [Fig. 2(b)].

We measured the accuracy of restoration and inference. First, restoration accuracy (RA) measures what fraction of the restored missing data $\sigma_i^m(t)$ is matched with the unmasked original data $\sigma_i(t)$. Second, RMSE measures how well the inferred W_{ij} is matched with true W_{ij}^{true}. In addition to the RMSE, we also measure the slope of the linear regression between W_{ij} and W_{ij}^{true}. As shown in Fig. 1(b), slope smaller than 1 represents underestimation of W_{ij}, whereas slope larger than 1 represents overestimation of W_{ij}. Using these metrics, we compared our SAEM method to the mean-field EM (MFEM) method developed by Campajola *et al.* [26] (see Appendix A for a brief summary of the method).

RA is the ratio of correctly restored missing data points, and it generally decreases with p [Fig. 2(c)]. Our method restores nearly 80% of the masked data points when 10% of the full data is masked as missing data. However, if 80% of the data is missing, then the restoration is more or less the same as a random restoration. While MFEM achieves a slightly better RA than our SAEM, ours shows clearly better model inference with smaller RMSE especially in the intermediate range of $0.3 \leqslant p \leqslant 0.7$ [Fig. 2(d)]. Furthermore, it ensures an unbiased estimation of W_{ij} at least when the p is less than 0.5 [Fig. 2(e)]. In contrast, MFEM suffers from severe underestimation of W_{ij} when more than 20% of the data is missing.

Now we test the performance of SAEM in harsher inference conditions with too weak or strong couplings or an insufficient data size. Throughout the experiments, we used a fixed system size $N = 80$. Unless mentioned otherwise, the data length is $L = 6400$ and the scale of couplings is $g = 1.0$.

First, we evaluated the effect of coupling scale g. Strong spin interactions lower stochasticity in the kinetic Ising model. When g is too small ($\approx$0.1), every missing data point has a flip probability close to 50%, so accurate restoration of missing data is fundamentally impossible. This is why D_{mis} becomes smaller than D_{obs} right after random initialization [Fig. 3(a), upper-right panel]. At this point, we can also achieve the lowest error of W_{ij}, as the RMSE keeps increasing over the iterations (upper-left panel). However, large g ($\approx$4.0) makes D_{mis} and D_{obs} converge more slowly (bottom panels). Therefore, the exact value of ϵ, the threshold value of $D_{\text{mis}} - D_{\text{obs}}$, affects the stopping point of the EM iterations more significantly, although RMSE does not change much with iterations for the converging case. A larger g results in a higher RA [Fig. 3(b)]. This is expected because strong g makes the inference of missing data less stochastic. For a large $g > 4$ and small $p < 0.5$, almost all missing data are correctly restored. For a small $g < 0.25$, however, RA becomes around 50%. Next we examined the quality of model inference depending on g. To ensure a fair comparison between results with different scales of coupling, we used a rescaled RMSE, $\text{RRMSE} = \text{RMSE}/g$, as used in Ref. [26] [Fig. 3(c)]. For $p < 0.5$, RRMSE keeps decreasing with g. Similar to the worse restoration, this is because small g induces larger fluctuations with a flip probability close to 50%, making the inference more difficult. When $p > 0.6$, the RRMSE curves deviate from those of $p < 0.5$, especially when g is large. Taken together, these results indicate that our SAEM is applicable

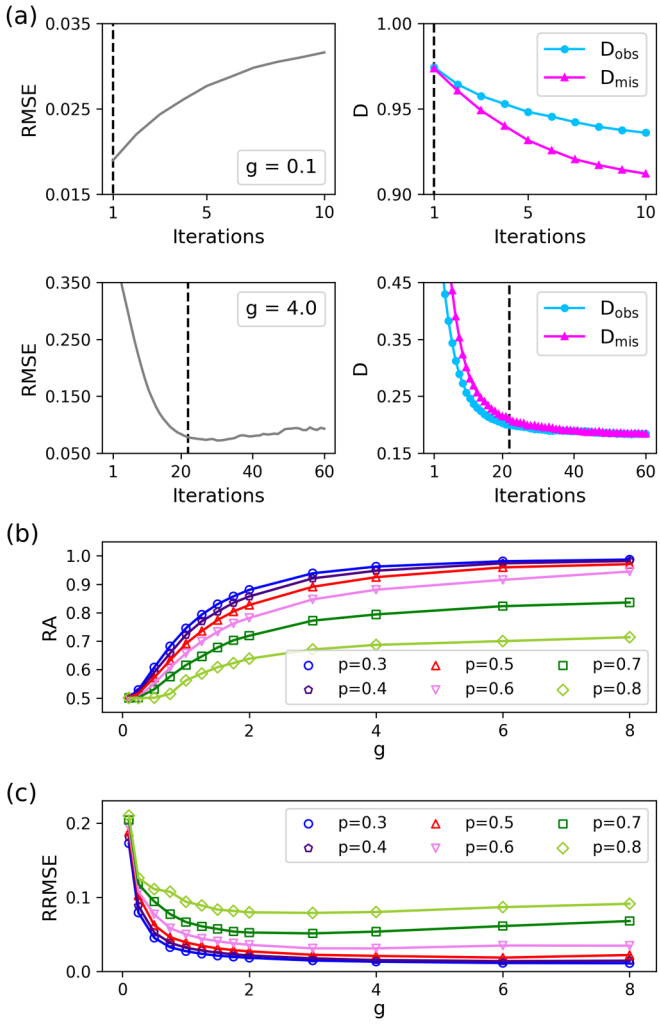

FIG. 3. Inference for strong coupling. (a) The evolution of RMSE and D over the iterations when the scale of couplings, g, is small ($g = 0.1$, top panels) or large ($g = 4.0$, bottom panels) with the fraction of missing observations $p = 0.5$. (b-c) RA of missing data (b) and rescaled RMSE of W_{ij} (c) as a function of g from 0.1 to 8. A system size $N = 80$ and a data length $L = 6400$ are used.

to a wide range of $0.5 < g < 8$, at least when less than 50% of data is missing.

Next, we checked the dependency of our SAEM on the data length L. When L is small ($L/N^2 = 0.25$), RMSE increases quickly after the optimal iteration, which emphasizes an important role of our stopping criterion [Fig. 4(a), top panels]. When L is large ($L/N^2 = 8$), however, both RMSE and data-model discrepancy D converge (bottom panels). As more data is provided, RA increases and finally saturates [Fig. 4(b)], and RMSE of W_{ij} decreases [Fig. 4(c)]. It is of interest that RMSE shows a power-law-like behavior in the wide range of L/N^2, which was also observed in Ref. [26]. The power-law exponent does not depend on p. Therefore, this scaling behavior can suggest an appropriate sample size for the applicability of our SAEM.

To further validate SAEM and our stopping criterion, we surveyed its performance under various conditions. We investigated the effects of sparsity or symmetry of coupling W_{ij} and the presence of external bias b_i. Then we confirmed that this

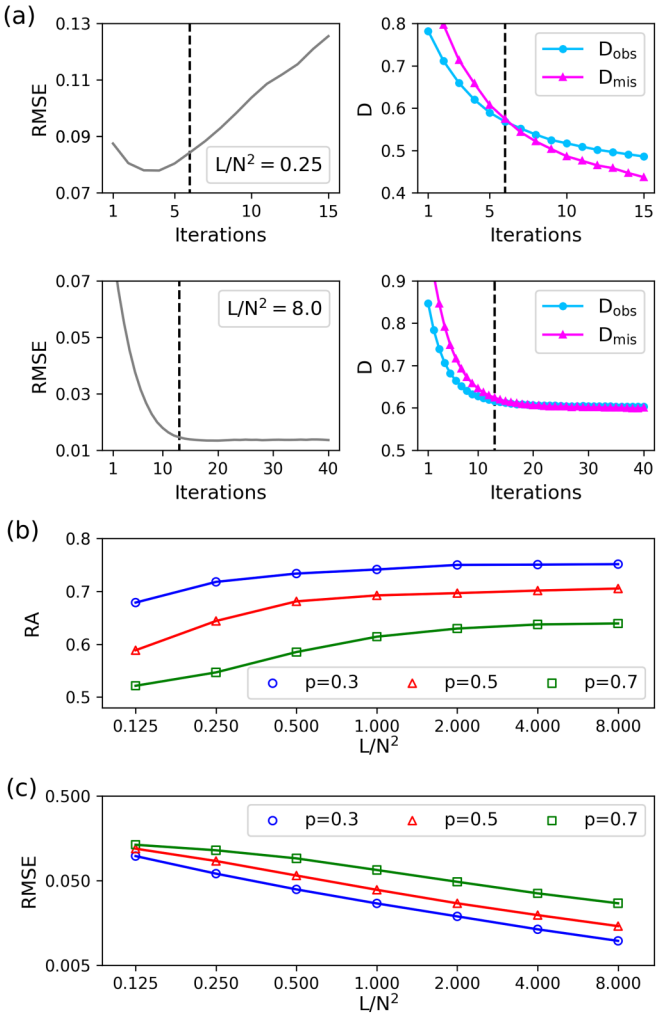

FIG. 4. Inference and data size. (a) The evolution of RMSE and D over the iterations when the data length, L, is short ($L/N^2 = 0.25$, top panels) or long ($L/N^2 = 8.0$, bottom panels) with the fraction of missing observations $p = 0.5$. (b, c) RA of missing data (b) and RMSE of W_{ij} (c) as a function of L/N^2 from 0.125 to 8. Here we use a system size of $N = 80$, and the scale of couplings $g = 1$.

method shows accurate restoration and inference under these conditions (refer Appendix B for details).

B. Inference of neuronal dynamics

Having confirmed that our algorithm robustly works for synthetic data, we now apply the algorithm to explore real data. We examined temporal data of neuronal activities recorded in salamander retinal ganglion cells [35]. The data consist of neuronal spike trains of 160 neurons, whose bin size is 20 ms. The state of each bin is considered *active* [$\sigma_i(t) = +1$] when at least one spike is present in the time bin, and *inactive* [$\sigma_i(t) = -1$] otherwise. We assumed that the effect of measurement noise is negligible in the binarized data. The total length of the time series is $L = 28\,3041$. In this study, we randomly selected 60 neurons among the 80 most active neurons to exclude silent neurons. Figure 5(a) summarizes our problem setting. We have neuronal activity recording data (top panel). After masking 70% of them randomly as missing data (center panel), we restore the missing data using our inference method (bottom panel).

First, we monitored the evolution of D_{obs} and D_{mis} during the iterations of the SAEM method [Fig. 5(b)]. The gap between two model-data discrepancies decreases with the iterations. However, D_{mis} never falls below D_{obs}. We have also observed that the crossover between D_{mis} and D_{obs} diminishes as data size gets larger [Fig. 4(a), bottom panels]. The experimental data has a pretty large data size ($L/N^2 \approx 80$). In this case, the specific value of ϵ becomes crucial to determine the stopping point of the EM iteration. Here we used $\epsilon = 0.01$ as in the previous application for the synthetic data because it is good enough to accurately restore missing data and reproduce statistical features in data.

To evaluate the inference performance, we compare four inference methods: SAEM, EQEM, MEAN, and FREQ. SAEM and EQEM are based on physical models. SAEM implements the kinetic or nonequilibrium Ising model. We have elaborated on this algorithm in previous sections. EQEM utilizes the standard equilibrium Ising model, which has been widely used to model collective behavior of neuronal networks [2,3,28]. This algorithm uses an EM-based approach similar to SAEM, but it considers the neural activities at different times separately (see Appendix C for the details of EQEM). For both algorithms, we include external bias b_i as well as neighboring interactions W_{ij}. Contrary to SAEM and EQEM, MEAN and FREQ use simple imputation schemes. MEAN stochastically assigns $\sigma_i^m(t) = \pm 1$ under a constraint that their mean activity becomes equal to the mean activity of observed data points $\sigma_i^o(t)$. FREQ simply assigns $\sigma_i^m(t) = -1$ for every missing data point considering that silent neurons are dominant in the observed data.

First, we compared RA of the four methods [Fig. 5(c)]. Unexpectedly, FREQ, the simplest restoration method, achieves the highest RA (94.6%). Other methods show lower accuracy: 91.8% (SAEM), 90.4% (EQEM), and 89.9% (MEAN). Despite this fact, the fundamental problem of FREQ is that it cannot consider the inherent stochasticity of missing data points. Although this naive method performs the best restoration in terms of RA, it overlooks a small portion of data points with $\sigma_i^m(t) = +1$, and thus perturbs all the collective properties present in the data.

Therefore, we focused on other descriptive statistics of data related to the collective behavior of neurons, to compare the inference performance. First, we measured the number of simultaneous spikes, $K(t) = \sum_{i=1}^{N} (\sigma_i(t) + 1)/2$, at different times, and obtained their relative frequencies $P(K)$ [Fig. 5(d)]. We found that SAEM matches the original $P(K)$ much better than any other methods. MEAN and FREQ underestimate a probability of large K, while EQEM predicts a much heavier tail for $P(K)$ than the original one. This tendency for EQEM was also observed by Tkačik *et al.* [28]. They tried to solve the issue by using an equilibrium model with pairwise interactions and an additional potential parameterized by K. Here we emphasize that SAEM, having nonequilibrium dynamics into the system, naturally captures the pattern of simultaneous firing without further assumptions. The marked underestimation of $P(K)$ by FREQ is expected because it considers all missing data points as inactive [$\sigma_i^m(t) = -1$].

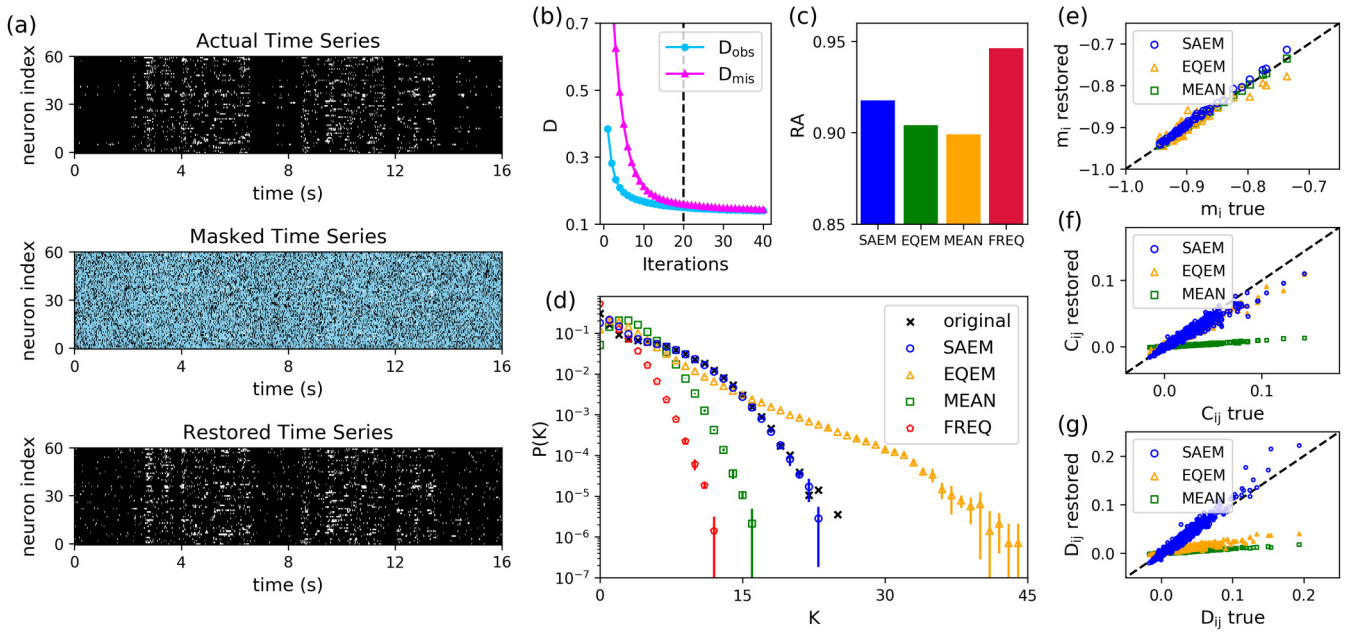

FIG. 5. Inference with real neural activity data. (a) Restoration of neuronal activity data published by Marre *et al.* [35]. Among neural activities of 60 neurons (top), we randomly mask a part of them with a fraction of $p = 0.7$ (center). Then we recover the masked data using our stochastic EM-based inference method (bottom). (b) The evolution of model-data discrepancy D for observed and missing data over the EM iterations. (c) Restoration accuracy (RA) of missing data points is compared between four methods: our nonequilibrium EM model (SAEM), equilibrium EM model (EQEM), imputation by mean activities of observed neurons (MEAN), and imputation by the most dominant activity (FREQ). Five independent experiments are averaged. (d) Relative frequencies of K simultaneous spikes obtained from the original unmasked data and restored data from four inference methods. (e–g) Comparison between true and restored mean activity m_i (e), equal-time activity correlation C_{ij} (f), and one-step lagged activity correlation D_{ij} (g), for three inference methods except for FREQ.

Next, we examined mean activities $m_i = \langle \sigma_i(t) \rangle_t$, equal-time correlations $C_{ij} = \langle \sigma_i(t)\sigma_j(t) \rangle_t - \langle \sigma_i(t) \rangle_t \langle \sigma_j(t) \rangle_t$, and one-step lagged correlations $D_{ij} = \langle \sigma_i(t)\sigma_j(t+1) \rangle_t - \langle \sigma_i(t) \rangle_t \langle \sigma_j(t) \rangle_t$ from the restored data [Figs. 5(e)–5(g)]. SAEM correctly infers all the statistics, m_i, C_{ij}, and D_{ij}. The correct estimation of C_{ij} is surprising because SAEM matches m_i and D_{ij} by tuning b_i and W_{ij} and does not directly affect C_{ij}. Similarly, the correct estimation of m_i and C_{ij} by EQEM is not surprising because EQEM models m_i and C_{ij} by tuning b_i and W_{ij}. However, EQEM severely underestimates D_{ij}, perhaps because neuronal firing is far from an equilibrium process in that it is a manifestation of information transmission. Finally, MEAN by definition fits only m_i and obviously fails to account for pairwise correlations between spins, both C_{ij} and D_{ij}. We omit the results for FREQ, since it is obvious that FREQ cannot reproduce all of the statistics.

As a further corroboration of the validity of SAEM, we examined another data set of neuronal activities [36], and confirmed that SAEM reproduces the collective behavior of neurons as shown here (see Appendix D). These results clearly demonstrate that our method (SAEM) is an effective approach for understanding real experimental data.

IV. DISCUSSION

We developed a stochastic approximation EM algorithm that infers the kinetic Ising model from stochastic time series with missing data. The algorithm alternates between an E-step stochastically restoring missing data points and an M-step optimizing model parameters. Using this SAEM with an appropriate stopping criterion for the EM iterations, we could successfully infer model parameters without under- or over-estimation even when up to 70% of the data is missing. We demonstrated the performance of the inference with synthetic data from extensive simulations of the kinetic Ising model, and with real neuronal data. In particular, our algorithm, based on a nonequilibrium model, outperforms equilibrium models in reproducing collective behavior in neuronal activities.

We found an effective scheme for determining the optimal number of iterations that provides the best model inference. Under easy inference conditions such as a moderate coupling regime, sufficient data, and a small fraction of missing data, excessive EM iterations do no harm since the inferred parameter values ultimately converges. However, under difficult inference conditions, the excessive iterations lead to worse inference of coupling strengths in the kinetic Ising model. Here, for the best inference, it is crucial to stop SAEM at the right iteration. Our key idea is that the stochastic model must show equal uncertainty to predict observed and missing data points because seamless restoration of missing parts implies no distinction between observed and restored parts. This condition can be practically monitored by the equality of the model-data discrepancy of observed and missing data ($D_{\rm obs} = D_{\rm mis}$). We confirmed that the stopping criterion works well in various conditions. However, it still lacks the-

oretical justification or derivation from fundamental physical principles.

This study applied the SAEM algorithm for the kinetic Ising model. However, other nonequilibrium models can be considered as the underlying stochastic model. For instance, Marre *et al.* [37] proposed an Ising model incorporating both spatial and temporal correlations among neurons, and showed that this model predicts spatio-temporal patterns of neuronal activities significantly better than the standard Ising model (also see Refs. [38,39]). Moreover, Tyrcha *et al.* [40] showed that applying nonstationary external bias can explain a structure of neuronal spike trains even without direct interactions between neurons.

Our methodology to deal with incomplete kinetic Ising data can be applied to other fundamental scenarios. For example, we can apply our algorithm to data with irregular observation times. Unequally spaced time-series data is common in astronomy [41,42], paleoclimatology [43], biology [44], and many other fields in which regular observations are infeasible. In the context of the kinetic Ising model, sequences at unobserved times can be correctly inferred using exactly the same approach as this paper. Other applications would be the reconstruction of interaction networks when experimental artifacts make sporadic missing observations. We expect that our new algorithm can be generalized for many situations.

ACKNOWLEDGMENTS

We thank Carlo Campajola for kindly sharing the code for the mean-field EM method. This work was supported by the Intramural Research Program of the National Institutes of Health, NIDDK (V.P.), and the New Faculty Startup Fund from Seoul National University, and the National Research Foundation of Korea (NRF) Grant funded by the Korea government (MSIT) (Grant No. 2019R1F1A1052916) (J.J.).

APPENDIX A: MEAN-FIELD EM

We briefly summarize the mean-field EM (MFEM) developed by Campajola *et al.* [26]. The first step of MFEM is to initialize the coupling parameters, W_{ij}, with random values. An analytic approximation of expected log-likelihood, with the current coupling parameters and observed data points, can be derived using the Martin-Siggia-Rose path integral formalism [45] and the saddle-point approximation. This gives a self-consistent equation for calculating the mean magnetization, $m_i(t)$, for each missing data point. After solving this equation to compute $m_i(t)$ (E-step), the gradient of the expected log-likelihood with regard to W_{ij} is obtained. Applying the gradient ascent gives the update rule for W_{ij} (M-step). Repeating these two steps until convergence, one can get both the coupling parameters W_{ij} and the missing data points $\sigma_i^m(t) = \text{sign}[m_i(t)]$. The authors also observed an underestimation of W_{ij} [see Fig. 2(e)]. To mitigate this issue, they proposed a recursive procedure that iterates the whole EM algorithm several times. Every time the algorithm ends, they fix the most polarized $m_i(t)$ for each time t according to its sign. This can reduce the extent of underestimation, and noticeably improve the quality of inference [26]. However, this procedure requires multiple implementations of the original algorithm, requiring longer computation time.

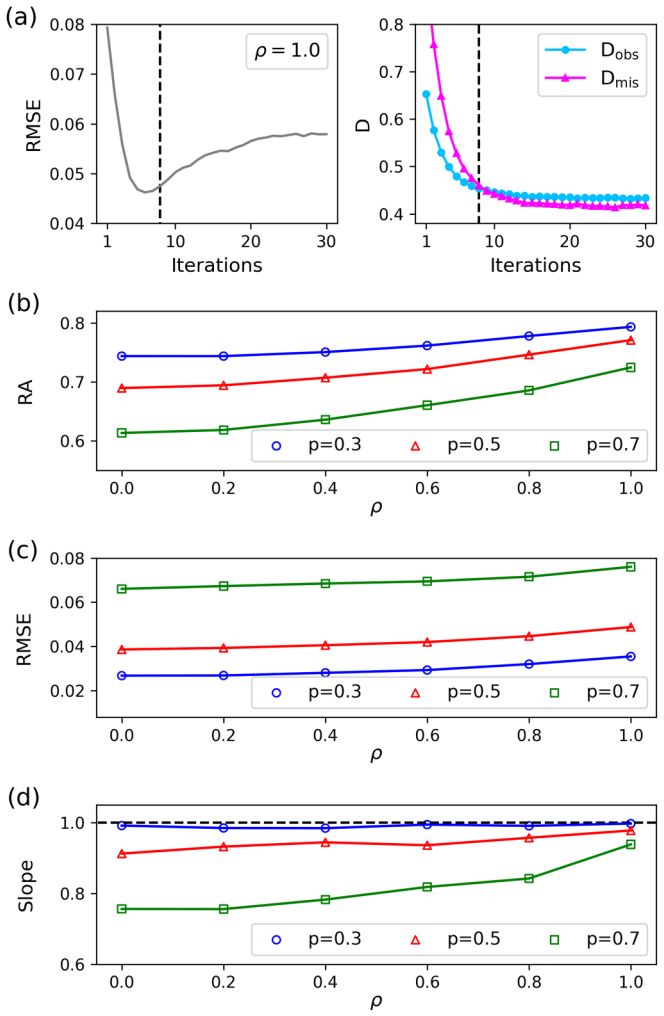

FIG. 6. Inference of symmetric coupling. (a) The evolution of RMSE and D over the iterations when a coupling matrix W is fully symmetric (symmetry parameter $\rho = 1$) and a fraction of missing data is $p = 0.5$. (b–d) RA of missing data (b), RMSE of W_{ij} (c), and the linear regression slope (d) as a function of ρ.

APPENDIX B: ADDITIONAL VALIDATIONS

We further tested the robustness of our SAEM under various conditions such as sparsity or symmetry of coupling strengths, and the presence of external biases.

First, we investigated the effect of symmetry of coupling strengths. The original SK model [34] assumes that all couplings are independent and identically distributed. We mitigated this assumption by inducing symmetry to the coupling matrix: $W = \sqrt{1-\rho}W_{\text{rand}} + \sqrt{\rho}W_{\text{sym}}$. The elements of W_{rand} and W_{sym} are drawn from a Gaussian distribution, $\mathcal{N}(0, g^2/N)$, but W_{sym} is symmetric. $\rho \in [0, 1]$ is a control parameter tuning the degree of symmetry. For example, $\rho = 1$ makes the coupling matrix fully symmetric. We confirmed that the stopping criterion ($D_{\text{obs}} - D_{\text{mis}} < \epsilon$) worked to find an optimal iteration when a coupling matrix is fully symmetric [Fig. 6(a)]. It is of interest that RA slightly increases

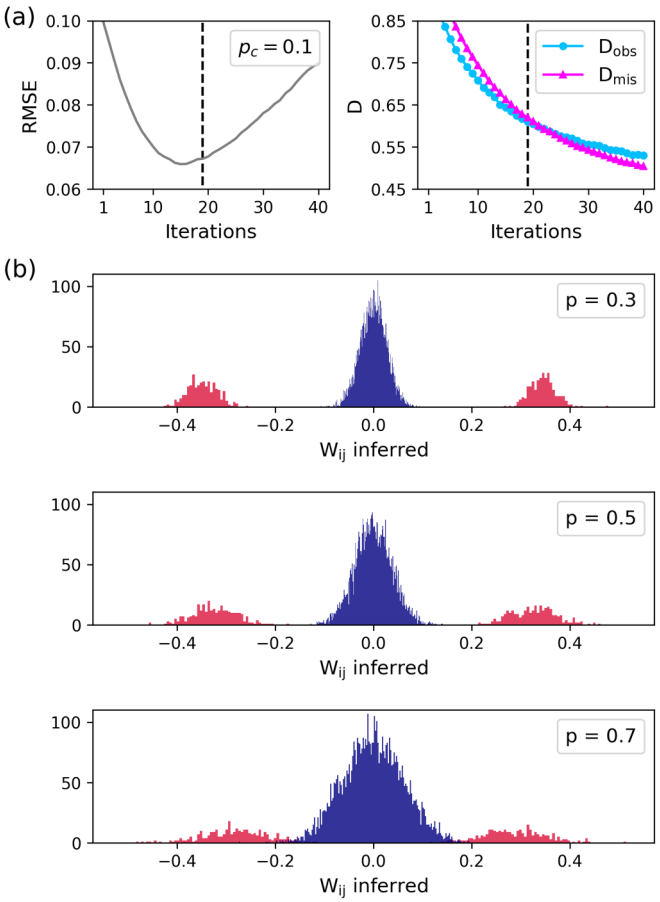

FIG. 7. Inference of sparse coupling. (a) The evolution of RMSE and D over the iterations when interactions are sparse (sparsity $p_c = 0.1$). A fraction of missing data is $p = 0.7$. (b) Inference of connected (red) and disconnected (blue) couplings with $p = 0.3$ (top), 0.5 (center), and 0.7 (bottom).

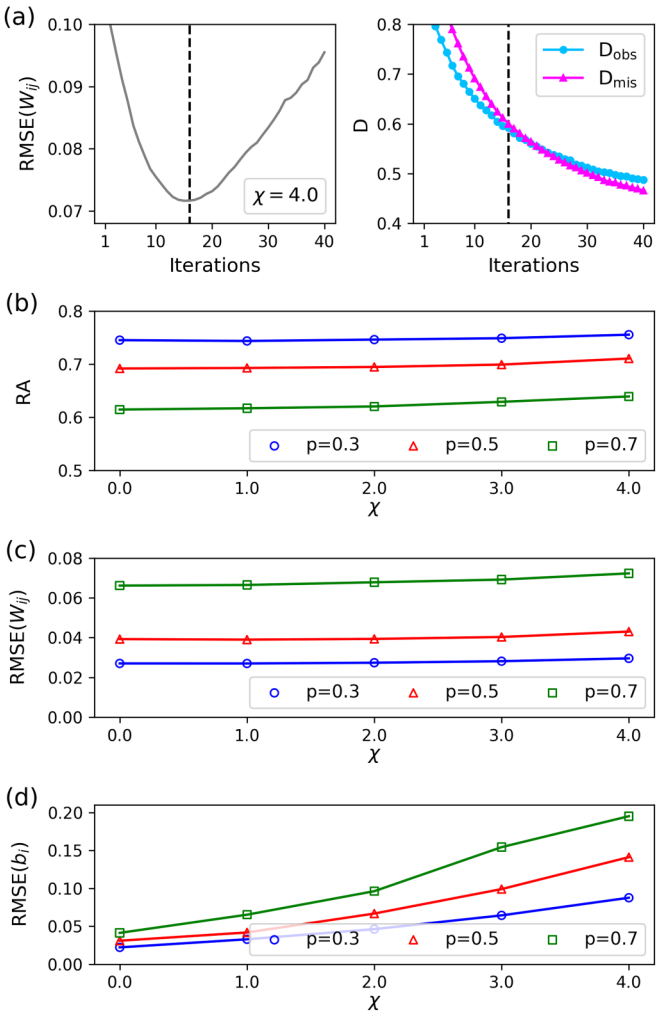

FIG. 8. Inference of strong bias. (a) The evolution of RMSE (of W_{ij}) and D over the iterations when there is a constant bias ($b_i = \chi g/\sqrt{N}$, where $\chi = 4.0$) for spins. A fraction of missing data is $p = 0.7$. (b–d) RA of missing data (b), RMSE of W_{ij} (c), and b_i (d) as a function of χ.

as W becomes more symmetric [Fig. 6(b)], while RMSE gets slightly worse [Fig. 6(c)]. We also measured the linear regression slope of W_{ij}^{true} on W_{ij} and found that it increases with ρ [Fig. 6(d)]. Given a small fraction of missing data ($p < 0.5$), the slope is always close to 1.0, implying no under- or overestimation.

Second, we examined the effect of sparsity of coupling strengths. We consider a discrete SK model where the couplings are $W_{ij} \in \{-g/\sqrt{p_c N}, 0, g/\sqrt{p_c N}\}$ with a probability of $p_c/2$, $1 - p_c$, and $p_c/2$, respectively, following [46]. Here p_c is a control parameter tuning the sparsity of a network, i.e., a fraction of connected pairs. For a sparse network with $p_c = 0.1$, we applied our algorithm to infer missing time series. Our stopping criterion again found a near-optimal point [Fig. 7(a)], even when a large amount of data is masked ($p = 0.7$). The inferred couplings are clearly separated [Fig. 7(b)]. We can easily distinguish connected pairs from disconnected ones regardless of a fraction of missing data p. Note that we did not use other techniques such as ℓ_1-regularization [47,48] or decimation of couplings [49,50].

Last, we activated the bias b_i. When b_i was drawn from a Gaussian distribution with zero mean, our algorithm found a good stopping point as well as reasonable estimates of W_{ij} and b_i. We did not show the results because they are not significantly different from the case when $b_i = 0$. How about when all b_i have the same sign, either plus or minus? For instance, neurons are usually inactive, so b_i for a neuronal network would be mostly negative. To deal with this problem, we set a fixed $b_i = \chi g/\sqrt{N}$, where χ denotes the scale of a bias. Our stopping criterion worked well with a moderately high bias [$\chi = 4.0$, Fig. 8(a)]. RA and RMSE of W_{ij} do not significantly change with χ [Figs. 8(b) and 8(c)]. Meanwhile, RMSE of b_i increases with χ as expected [Fig. 8(d)].

APPENDIX C: EQUILIBRIUM ISING MODEL

In this section, we describe the equilibrium EM (EQEM) method based on the equilibrium Ising model, which we have briefly mentioned in Sec. III B. In the standard Ising model, a probability of spin configuration $\vec{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_N)$ with $\sigma_i = \pm 1$ is parameterized by an energy $E(\vec{\sigma})$, $P(\vec{\sigma}) = Z^{-1} \exp[-E(\vec{\sigma})]$. Here, $Z = \sum_{\vec{\sigma}} \exp[-E(\vec{\sigma})]$ is a normalization factor and called a partition function. Energy is

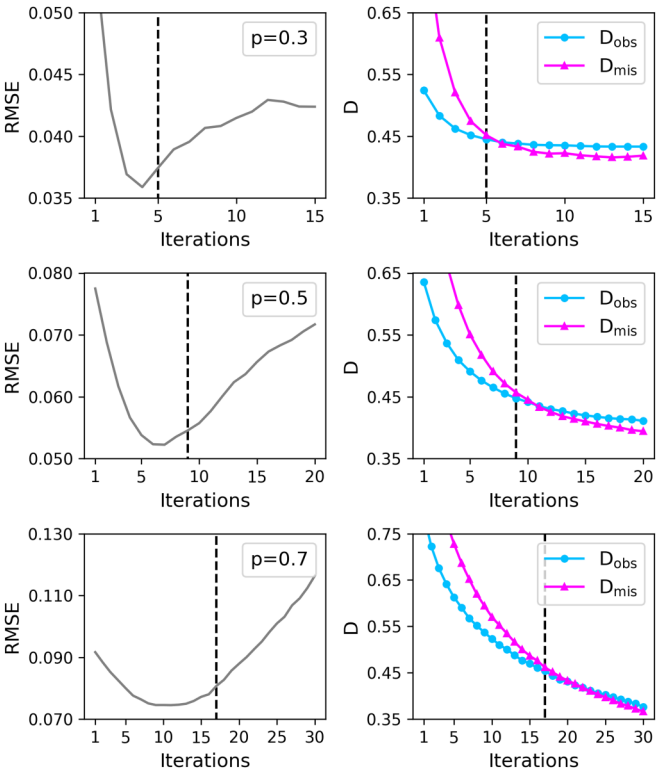

FIG. 9. The same as Fig. 1(a) but with the equilibrium Ising model.

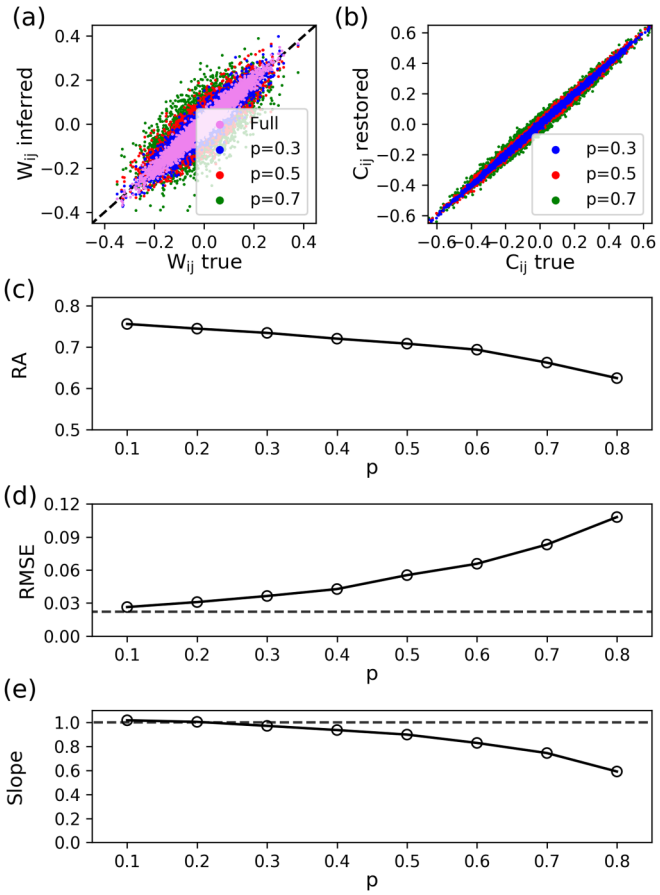

FIG. 10. The same as Fig. 2 but with the equilibrium Ising model. The one-step lagged correlation D_{ij} in Fig. 2(b) is replaced by the equal-time correlation C_{ij}.

described by the first and second moments of spins, i.e., $E(\vec{\sigma}) = -\sum_{i<j} W_{ij}\sigma_i\sigma_j - \sum_i b_i\sigma_i$, where W_{ij} is an interaction parameter and b_i an external bias. The task is to reconstruct missing data points and find the true $\theta = (W_{ij}, b_i)$ when some data points are missing.

We again split the algorithm into E- and M-steps. First, we initialize missing data points with random binary values. Note that in the equilibrium model, the time index t merely acts as a label to distinguish different sequences of $\vec{\sigma}$. From the randomly assigned data, we infer an error-prone θ. Then we restore the missing data points with ± 1 using the ratio of two probabilities, $P\{F_i^+[\vec{\sigma}(t)]|\theta\}/P\{F_i^-[\vec{\sigma}(t)]|\theta\}$ (E-step). $F_i^{\pm}[\vec{\sigma}(t)]$ is defined similar to the main text (see Sec. II). Calculation of the ratio of probabilities does not require the computation of partition function and therefore can be easily done. We update all of the missing data points one by one.

Given the restored data points, we can infer the optimal θ (M-step) by maximizing the data likelihood, called the Boltzmann machine [51]. However, this requires an extensive calculation of partition function Z, which entails the sum of 2^N configurations. Physicists have developed a number of methods to circumvent the exact computation of Z, including mean-field [52,53], Bethe approximation [54], Sessak-Monasson [55], and adaptive cluster expansions [56] (see also a pedagogical review [15]). In this study, we choose a maximum pseudolikelihood approach [46] that maximizes the *local* pseudolikelihood, $\mathcal{L}_i(\theta) = \prod_t P[\sigma_i(t)|\vec{\sigma}_i^c(t), \theta]$ for each i, where $\vec{\sigma}_i^c$ denotes all spins besides the ith spin. Since $P[\sigma_i(t)|\vec{\sigma}_i^c(t), \theta]$ is determined by a local field $H_i(t) = \sum_{j\neq i} W_{ij}\sigma_j(t) + b_i$, the same logistic regression function [31] can be utilized, allowing fast and accurate inference. After optimizing each $\mathcal{L}_i(\theta)$ separately, we put $W_{ij} \leftarrow (W_{ij} + W_{ji})/2$ to ensure symmetric couplings.

Running this EQEM, we observed the overshooting similar to what we have observed from the algorithm with kinetic Ising model (SAEM, see Fig. 1). To avoid overshooting, we define the same model-data discrepancy measures for observed and missing data points, $D_{\rm obs}$ and $D_{\rm mis}$, in which $\sigma_i(t+1)$ is replaced to $\sigma_i(t)$ [see Eq. (7)]. We track the evolution of $D_{\rm obs}$ and $D_{\rm mis}$ and halt the EM iteration when the stopping condition $D_{\rm mis} - D_{\rm obs} < \epsilon$ is met (we use $\epsilon = 0.01$ as in the main text). Overall, EQEM and SAEM are nearly the same, but the main differences are (i) we apply a new stochastic rule to restore missing data points and (ii) we use a pseudolikelihood maximization scheme to infer W_{ij}.

Figure 9 illustrates our stopping criterion for EQEM. Samples were generated using the Metropolis-Hastings algorithm [32] with a system size $N = 100$ and a sample size $L = 10\,000$. Especially when a fraction of missing data is large ($p = 0.7$), the iteration stops several steps after we achieve the minimum error of W_{ij}, which means that the model inference of EQEM is suboptimal. In contrast, SAEM showed coincidence of the minimum of RMSE and the intersection of $D_{\rm obs}$ and $D_{\rm mis}$ in general cases. The mismatch in EQEM may result from the pseudolikelihood approximation. EQEM

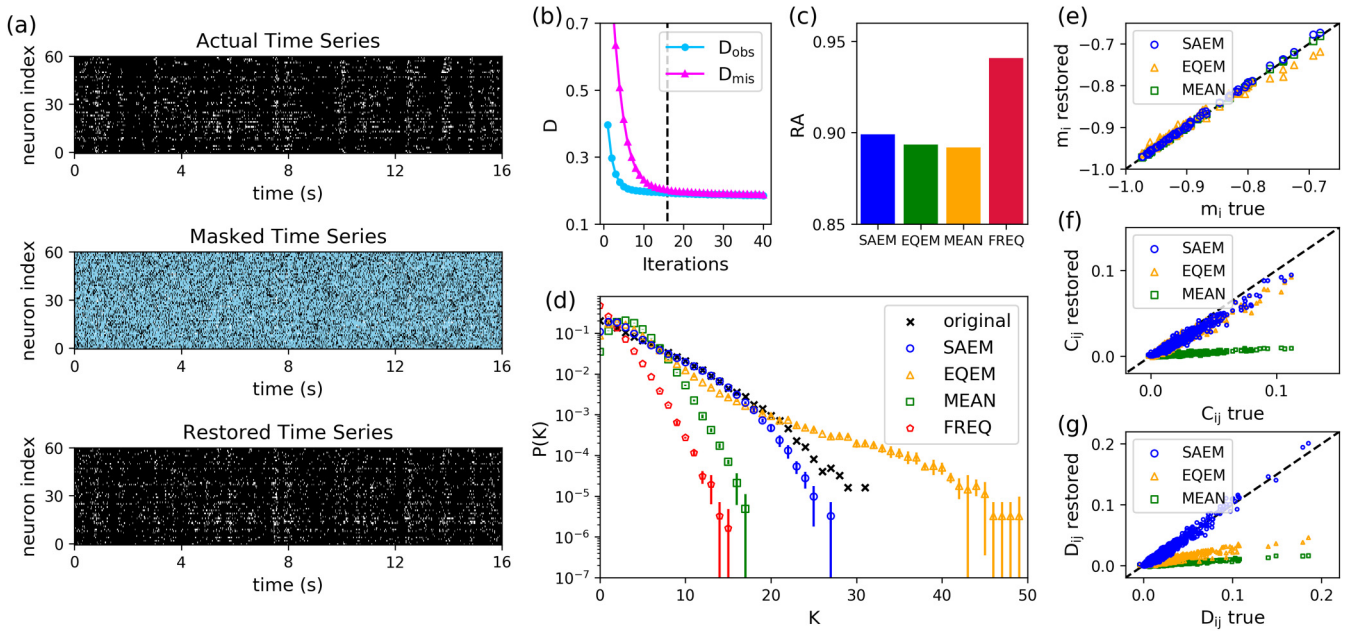

FIG. 11. The same as Fig. 5 but with the data published by Kohn and Smith [36].

used the pseudolikelihood for the inference (M-step), whereas it used the exact likelihood for the restoration (E-step).

Even with this seemingly improper stopping criterion, EQEM is capable of inferring the true W_{ij} [Fig. 10(a)]. Furthermore, an excellent agreement between the pair correlations C_{ij} of true and restored data was found [Fig. 10(b)], despite a relatively large deviation of the inferred W_{ij}. Next, we addressed the quality of data and model inference with restoration accuracy of missing data, root-mean-square error of W_{ij}, and linear regression slope of true W_{ij} on inferred W_{ij}. RA is above 0.6 even when 80% of the data is missing [Fig. 10(c)]. RMSE decreases with p as expected [Fig. 10(d)]. When a small fraction of data is missing ($p < 0.4$), we can accurately infer W_{ij} with an error less than 0.05. In this regime, the slope of linear regression is close to 1.0, which means no under- or overestimation [Fig. 10(e)]. These results are similar to SAEM (Fig. 2).

APPENDIX D: OTHER EXPERIMENTAL DATA

In addition to Sec. III B, we tested our algorithms with other neuronal spike train data published by Kohn and Smith [36]. They recorded spiking activities of 70-100 neurons in the visual cortex of adult monkeys under various visual stimuli. Here, we chose a "spontaneous" dataset where a uniform gray screen had been used. We binned the neural activity with 20 ms and selected 60 neurons randomly from the 80 most active neurons. The data length was $L = 123\,404$. For the detailed experimental procedure, see Refs. [57,58]. We summarize the results in Fig. 11 with the same format as Fig. 5. The results are similar.

[1] S. Brunton and N. Kutz, *Data-driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, Cambridge, UK, 2019).

[2] E. Schneidman, I. I. Michael J. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).

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

[4] A. Lapedes, B. Giraud, and C. Jarzynski, arXiv:1207.2484.

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

[6] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Jr., Proc. Natl. Acad. Sci. USA **107**, 5405 (2010).

[7] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, Proc. Natl. Acad. Sci. USA **103**, 19033 (2006).

[8] T. Bury, Physica A **392**, 1375 (2013).

[9] Y. Shemesh, Y. Sztainberg, O. Forkosh, T. Shlapobersky, A. Chen, and E. Schneidman, eLife **2**, e00759 (2013).

[10] E. D. Lee, C. P. Broedersz, and W. Bialek, J. Stat. Phys. **160**, 275 (2015).

[11] M. Figliuzzi, P. Barrat-Charlaix, and M. Weigt, Mol. Biol. Evol. **35**, 1018 (2018).

[12] B. Derrida, E. Gardner, and A. Zippelius, Europhys. Lett. **4**, 167 (1987).

[13] Y. Roudi and J. Hertz, Phys. Rev. Lett. **106**, 048702 (2011).

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

[15] H. C. Nguyen, R. Zecchina, and J. Berg, Adv. Phys. **66**, 197 (2017).

[16] M. Mézard and J. Sakellariou, J. Stat. Mech.: Theory Exp. (2011) L07001.

[17] D.-T. Hoang, J. Song, V. Periwal, and J. Jo, Phys. Rev. E **99**, 023311 (2019).
[18] B. Dunn and Y. Roudi, Phys. Rev. E **87**, 022127 (2013).
[19] J. Tyrcha and J. Hertz, Math. Biosci. Eng. **11**, 149 (2014).
[20] L. Bachschmid-Romano and M. Opper, J. Stat. Mech.: Theory Exp. (2014) P06013.
[21] C. Battistin, J. Hertz, J. Tyrcha, and Y. Roudi, J. Stat. Mech.: Theory Exp. (2015) P05021.
[22] D.-T. Hoang, J. Jo, and V. Periwal, Phys. Rev. E **99**, 042114 (2019).
[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, J. R. Stat. Soc.: Ser. B (Methodol.) **39**, 1 (1977).
[24] D. Soudry, S. Keshri, P. Stinson, M.-h. Oh, G. Iyengar, and L. Paninski, PLoS Comput. Biol. **11**, e1004464 (2015).
[25] C. Campajola, F. Lillo, and D. Tantari, Quant. Finance **20**, 1765 (2020).
[26] C. Campajola, F. Lillo, and D. Tantari, Phys. Rev. E **99**, 062138 (2019).
[27] B. Dylyon, M. Lavielle, and E. Moulines, Ann. Statist. **27**, 94 (1999).
[28] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. J. B. I. I., PLoS Comput. Biol. **10**, e1003408 (2014).
[29] X. Chen, F. Randi, A. M. Leifer, and W. Bialek, Phys. Rev. E **99**, 052418 (2019).
[30] H.-L. Zeng, E. Aurell, M. Alava, and H. Mahmoudi, Phys. Rev. E **83**, 041135 (2011).
[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Research **12**, 2825 (2011).
[32] W. K. Hastings, Biometrika **57**, 97 (1970).
[33] R. Little and D. Rubin, *Statistical Analysis with Missing Data* (Wiley, New York, NY, 2002).
[34] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[35] O. Marre, G. Tkacik, D. Amodei, E. Schneidman, W. Bialek, and M. Berry, *Multi-electrode Array Recording from Salamander Retinal Ganglion Cells* (IST, Austria, 2017).
[36] A. Kohn and M. A. Smith, Utah array extracellular recordings of spontaneous and visually evoked activity from anesthetized macaque primary visual cortex (V1), CRCNS.org (2016), http://dx.doi.org/10.6080/K0NC5Z4X.
[37] O. Marre, S. El Boustani, Y. Frégnac, and A. Destexhe, Phys. Rev. Lett. **102**, 138101 (2009).
[38] H. Nasser, O. Marre, and B. Cessac, J. Stat. Mech.: Theory Exp. (2013) P03006.
[39] H. Nasser and B. Cessac, Entropy **16**, 2244 (2014).
[40] J. Tyrcha, Y. Roudi, M. Marsili, and J. Hertz, J. Stat. Mech.: Theory Exp. (2013) P03005.
[41] N. R. Lomb, Astrophys. Space Sci. **39**, 447 (1976).
[42] J. D. Scargle, Astrophys. J. **263**, 835 (1982).
[43] M. Schulz and M. Mudelsee, Comput. Geosci. **28**, 421 (2002).
[44] T. Ruf, Biol. Rhythm Res. **30**, 178 (1999).
[45] P. C. Martin, E. D. Siggia, and H. A. Rose, Phys. Rev. A **8**, 423 (1973).
[46] E. Aurell and M. Ekeberg, Phys. Rev. Lett. **108**, 090201 (2012).
[47] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, Ann. Stat. **38**, 1287 (2010).
[48] A. Y. Lokhov, M. Vuffray, S. Misra, and M. Chertkov, Sci. Adv. **4**, e1700791 (2018).
[49] A. Decelle and F. Ricci-Tersenghi, Phys. Rev. Lett. **112**, 070603 (2014).
[50] A. Decelle and P. Zhang, Phys. Rev. E **91**, 052136 (2015).
[51] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cognit. Sci. **9**, 147 (1985).
[52] H. Kappen, F. Rodríguez, and F. B. Rodr'iguez, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 1998), pp. 280–286.
[53] T. Tanaka, Phys. Rev. E **58**, 2302 (1998).
[54] F. Ricci-Tersenghi, J. Stat. Mech.: Theory Exp. (2012) P08015.
[55] V. Sessak and R. Monasson, J. Phys. A: Math. Theor. **42**, 055001 (2009).
[56] J. P. Barton, E. De Leonardis, A. Coucke, and S. Cocco, Bioinformatics **32**, 3089 (2016).
[57] M. A. Smith and A. Kohn, J. Neurosci. **28**, 12591 (2008).
[58] R. C. Kelly, M. A. Smith, R. E. Kass, and T. S. Lee, J. Comput. Neurosci. **29**, 567 (2010).