Biases in inverse Ising estimates of near-critical behavior

Maximilian B. Kloucek,[1,2] Thomas Machon,[1] Shogo Kajimura,[3] C. Patrick Royall,[4] Naoki Masuda,[5,6] and Francesco Turci[1]

[1]*School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, United Kingdom*
[2]*Bristol Centre for Functional Nanomaterials, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, United Kingdom*
[3]*Faculty of Information and Human Sciences, Kyoto Institute of Technology, Kyoto 606-8585, Japan*
[4]*Gulliver UMR CNRS 7083, ESPCI Paris, Université PSL, 75005 Paris, France*
[5]*Department of Mathematics, State University of New York at Buffalo, Buffalo, New York 14260-2900, USA*
[6]*Computational and Data-Enabled Science and Engineering Program, State University of New York at Buffalo, Buffalo, New York 14260-5030, USA*

(Received 19 January 2023; accepted 27 April 2023; published 7 July 2023)

Inverse Ising inference allows pairwise interactions of complex binary systems to be reconstructed from empirical correlations. Typical estimators used for this inference, such as pseudo-likelihood maximization (PLM), are biased. Using the Sherrington-Kirkpatrick model as a benchmark, we show that these biases are large in critical regimes close to phase boundaries, and they may alter the qualitative interpretation of the inferred model. In particular, we show that the small-sample bias causes models inferred through PLM to appear closer to criticality than one would expect from the data. Data-driven methods to correct this bias are explored and applied to a functional magnetic resonance imaging data set from neuroscience. Our results indicate that additional care should be taken when attributing criticality to real-world data sets.

DOI: 10.1103/PhysRevE.108.014109

I. INTRODUCTION

It is often the case that while interactions between individual constituents of complex systems are unknown, their correlations are measurable. Reconstructing the strength of the interactions from these correlations is an inverse problem. Inverse Ising inference, also known as *pairwise* maximum entropy modeling, is an inference technique used to learn the maximum entropy model (MEM) [1] representing a system of interacting binary variables [2,3]—termed *spins*. Following seminal work on the inference of interactions in retinal neurons [4], the maximum entropy modeling framework has been used in a range of biological settings, from understanding protein interactions [5] to modeling antibody diversity [6] and even to analyze the collective behavior of flocks of birds [7]. In neuroscience, in particular, the inference of pairwise MEMs (i.e., Ising models) from binary data has become common practice and is used both to understand the behavior of neuronal tissue [4,8,9] and to learn functional connectivity networks from coarse-grained functional magnetic resonance imaging (fMRI) studies in humans [10–13]. This procedure provides insight into the structure of the inferred networks, including their sparsity and the heterogeneity of their couplings [3,14,15].

In such Ising models, critical behavior emerges between a disordered high-temperature paramagnetic (P) phase with weak correlations and a low-temperature strongly interacting spin-glass (SG) phase with multiple metastable minima and large correlations [16–22]. The critical state is associated with a range of advantageous properties, including providing optimal sensitivity to inputs [23], enabling coordination between individual elements [24,25], allowing a large range of dynamic responses [26,27] and maximizing computational ability through edge-of-chaos computation [28,29]. In neuroscience, the activation patterns of ensembles of neurons have been shown to display typical signatures of critical behavior, as these so-called "neuronal avalanches" follow power-law distributions with size and dynamics that are consistent with a critical branching process [30–34]. Similar power-law distributions have also been found in human brain imaging studies [25,35], and inverse Ising inference specifically has recently shown that typical brain activity measured by fMRI occurs near the SG-P phase transition [13]. More generally, a range of evidence supports that many complex biological systems exist in a near-critical state [36,37], and it is postulated that the advantageous properties of the critical state may generally cause complex systems to organize towards criticality. With this in mind, maximum entropy techniques provide a valuable tool-set with which to further investigate the criticality hypothesis, allowing the inferred models to be assessed in the well-understood framework of equilibrium statistical physics [1].

In this paper, we focus on the pseudo-likelihood maximization (PLM) [2,38] approach to solving the inverse Ising problem. We chose this logistic regression-based method as it

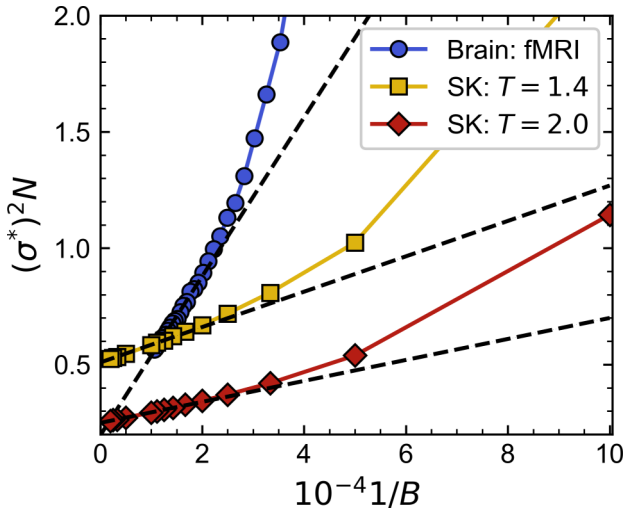

FIG. 1. Convergence of the variance of the inferred parameters $(\sigma^*)^2$ (scaled by system size) vs inverse sample number $1/B$. The dashed lines indicate linear fits to the asymptotic (large B) regime of the data. Asymptotic intercepts of the two SK state-points correspond to $T^* = 1.40$ and 2.00; PLM is able to perfectly reproduce the state-point given infinite data. The gradient b_1 of the asymptotic fits sets the severity of the small sample bias. This gradient depends strongly on the state-point, system size, and topology of the input data, e.g., we find b_1 (SK: $T = 1.4$) = 760, b_1 (SK: $T = 2.0$) = 450, and b_1 (Brain) = 3416.

is widely acknowledged as the state-of-the-art solution to the inverse Ising problem [3,14,15]. In Sec. II we introduce the PLM method and highlight alternative inverse Ising solvers. In Sec. III we show that parameters estimated via PLM are biased and that statistical averages of the parameters, such as the variance, converge to the true value as a linear function of $1/B$; see Fig. 1. We relate this convergence to the standard small-sample bias of maximum likelihood estimators (MLEs). Using the Sherrington-Kirkpatrick (SK) [16,17] spin-glass model as a benchmark, we find that the small B bias causes the inferred model to appear tuned towards criticality; models inferred using PLM show both a lower temperature and enhanced critical fluctuations than the input model from which the data were generated. Similar to previous authors [3,14], we find that the rate at which the bias is dissipated is state-point (i.e., temperature) -dependent and links the failure of PLM at low temperatures (i.e., for highly correlated data) to the separation effect observed in logistic regression [39,40]. In Sec. IV we present two corrective procedures to remove the bias, a self-consistency correction and Firth's penalized logistic regression [40–42], and we compare their performance by measuring how well they capture the temperature and critical fluctuations of the input data set. In Sec. V we explore the repercussions of the small sample bias in interpreting inference results from real data by considering an fMRI data set of brain activity related to meditation. Our results lead us to caution against claims of criticality in PLM models, as we show that small sample size biases tune models inferred from dynamical (i.e. fluctuating) data to a closer-to-critical state.

II. BACKGROUND: INVERSE ISING INFERENCE VIA PSEUDO-LIKELIHOOD MAXIMIZATION

We will consider systems of N interacting binary variables $s_i \in \pm 1, i = 1, \ldots, N$ which we refer to as *spins*. These spins may represent any binary quantities, such as the magnetic moments of spins in a metal (up, down), or the state of a neuron or region of interest (ROI) in the brain (on, off). The spins fluctuate in time, and for each of the N labeled regions, we have time series of length B. The state of the entire spin vector at a time t', $\boldsymbol{s}(t = t')$, is called a *configuration*, and the full data set of $B \times N$ observations will either be referred to as a *trajectory* or as the *data set*. Inverse Ising inference corresponds to learning a set of interaction parameters that is likely to reproduce the observations. Configurations of the inferred model will follow the maximum entropy probability distribution [1], and by measuring the per-spin magnetization $m_i = \langle s_i \rangle$ and cross-correlations,

$$C_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle, \tag{1}$$

where $\langle \cdot \rangle$ indicate time averages, one infers the so called *pairwise* maximum entropy model. This model is defined by the Hamiltonian

$$\mathcal{H} = -\sum_i h_i s_i - \frac{1}{2} \sum_{i \neq j} J_{ij} s_i s_j, \tag{2}$$

where the J_{ij} represent pairwise couplings between the spins and the h_i are external fields, with the summation index $i \neq j$ running over all nonmatching pairs of i and j. The inference problem, therefore, consists in determining the symmetric matrix of couplings $\boldsymbol{J}$ (diagonal entries are zero as there are no self-couplings) and vector of fields $\boldsymbol{h}$ from the correlations $\boldsymbol{C}$ and averages $\boldsymbol{m}$. The probability of observing a given configuration $\boldsymbol{s}$ follows the maximum entropy (Boltzmann) distribution:

$$P(\boldsymbol{s}) = \frac{1}{Z} \exp\left[-\beta \mathcal{H}(\boldsymbol{s})\right], \tag{3}$$

with $\beta = 1/T$ being the inverse temperature and Z the partition function, a normalization constant. The log-likelihood of observing a given trajectory $\{\boldsymbol{s}\}_B$ from the couplings and fields is

$$\mathcal{L}(\{\boldsymbol{s}\}_B | \boldsymbol{h}, \boldsymbol{J}) = \beta \sum_i h_i m_i + \frac{\beta}{2} \sum_{i \neq j} J_{ij} (m_i m_j + C_{ij}) - \log Z. \tag{4}$$

The set of parameters $\{\boldsymbol{h}^*, \boldsymbol{J}^*\}$ which maximizes Eq. (4) is the maximum-likelihood solution to the inverse Ising problem. When the number of spins N is very small (typically a few tens), it is computationally feasible to perform this optimization directly. However, the problem becomes rapidly intractable with increasing N (the number of possible configurations scales as 2^N), and a range of alternative methods have been proposed to perform the maximization. This includes Boltzmann learning [43] which uses Monte Carlo (MC) simulations [44,45] to evaluate the gradients of (4). While it is technically possible to compute these gradients with unbounded accuracy, the random nature and computational

intensity of MC sampling means this process is also limited to small ($N \sim 120$ [9]) system sizes. Various analytical solutions have also been introduced as alternatives, see [3,46,47] for reviews, yet most of these require additional assumptions to be made about the system, and often fail in the low-temperature (strong-coupling) regime, providing more error-prone solutions [3]. A powerful alternative approach, and the method studied here, is pseudo-likelihood maximization (PLM) [2]. In PLM one replaces the log-likelihood (4) by a set of N pseudo-log-likelihoods

$$\mathcal{L}_r(\{\boldsymbol{s}\}_B|h_r,\boldsymbol{J}_r) = \frac{1}{B}\sum_{t=1}^{B} \ln P_{\{h_r,\boldsymbol{J}_r\}}[s_r(t)|\boldsymbol{s}_{\setminus r}(t)], \tag{5}$$

which depend only on the parameter h_r and the rth row of entries $\boldsymbol{J}_r = \{J_{rj}\}_{j\neq r}$ to the coupling matrix. We also introduce the conditional probability distribution

$$P_{\{h_r,\boldsymbol{J}_r\}}(s_r|\boldsymbol{s}_{\setminus r}) = 1/(1+e^{-2\beta s_r[h_r+\sum_{r\neq j}J_{rj}s_j]}), \tag{6}$$

corresponding to the probability of observing the rth spin in state $+1$ or -1 given all other $N-1$ spins. We note that each of the $\mathcal{L}_r$ can be maximized independently for each spin, making the problem highly suitable for parallelization, and that in the limit of $B\to\infty$ the PLM approach to inverse Ising inference is exact. Moreover, the structure of the pseudo-likelihood means that each PLM optimization is formally identical to logistic regression for which efficient computational algorithms exist. For this work, we perform the regressions using the *sklearn.linear_model.LogisticRegression* classifier from the *Scikit-learn* [48] Python package. Note that the coupling matrix inferred this way is not symmetric, and we therefore always perform a postinference symmetrizing step, setting $\boldsymbol{J}^* = \frac{1}{2}[\boldsymbol{J}^*_{\rm PLM} + (\boldsymbol{J}^*_{\rm PLM})^T]$, where T is the transpose of the matrix. $\boldsymbol{J}$ can also be interpreted as the weighted adjacency matrix of a complex network [49,50], which encodes the topology of the model. As such, $l1$-regularized sparsity promoting versions of PLM [38] are commonly used for network reconstruction when a large set of parameters are known (or assumed to be zero), and most extensions to PLM have focused on improving its performance for this purpose [14,15]. When sparsity cannot be assumed to be unregularized, PLM still offers superior performance to other approximate inverse Ising solvers [3], and in this work we focus exclusively on PLM without regularization. We note that the fact that PLM provides independent estimates of J_{ij} and J_{ji} is meaningful for kinetic and dynamic models with nonsymmetric interactions and makes the approach viable for more generic network reconstructions.

III. PLM ACCURACY NEAR CRITICALITY

Previous work found that the accuracy of PLM depends on the temperature (or coupling strength) and topology of the underlying true model [2,3,14,51], and that errors are minimized near the critical point. It should be noted, however, that system sizes of only N from 16 to 64 were analyzed in these studies, and so the critical point is poorly defined. These results are commonly attributed to the theoretical finding that the generalized susceptibility of the Ising model can be related to the Fisher information matrix [52], and that the maximization of these quantities at the critical point corresponds to a high density of distinguishable parametrizations of different models. Notice that statistical distinguishability pertains to the parametrizations of the same model (e.g., the SK Hamiltonian) and does not correspond to our ability to distinguish different physical models.

When inferring a parametrization, finite-sample effects play an important role. The parameter estimates obtained through PLM, like those of any maximum-likelihood estimator (MLE), are well known to depend both on the number of samples B (with a slow $1/B$ convergence) and a prefactor set by the true parameter [41], which is not known *a priori*, so that, for example,

$$J^*_{ij} = J^0_{ij} + \frac{b_{1,ij}(\boldsymbol{h}^0,\boldsymbol{J}^0)}{B} + O(B^{-2}). \tag{7}$$

Here, the superscripts $*$ and 0 denote the inferred and true values, respectively, while $b_{1,ij}(\boldsymbol{h}^0,\boldsymbol{J}^0)$ is the state-dependent prefactor to the leading $1/B$ bias. Collectively, the prefactors $b_{1,ij}(\boldsymbol{h}^0,\boldsymbol{J}^0)$ set the difficulty of learning a given state-point (i.e., model) by increasing or decreasing the amount of data required to dissipate the bias. When expressed this way, we expect that due to the maximized Fisher information at criticality [52], systems near the critical point would have a smaller first-order prefactor making them easier to learn. It is also possible to express averaged quantities of the inferred parameters, such as the standard deviation σ of the couplings $\boldsymbol{J}$ in terms of the bias, so that

$$\sigma^* = \sigma^0 + \frac{b_1(\boldsymbol{h}^0,\boldsymbol{J}^0)}{B} + O(B^{-2}), \tag{8}$$

where b_1 is the combination of several bias terms on the individual J_{ij} couplings. Again, b_1 is an input state-dependent prefactor that sets the bias contribution to the inferred standard deviation. We show that this relation holds for two select SK state-points, as well as for a real-world data set in Fig. 1.

The above bias expressions are general, and hold for any MLE. But PLM specifically involves performing a set of binary logistic regressions, and these are known to be affected by an additional small sample size issue termed separation [39]. Separation occurs when a subset of covariates (e.g., $\boldsymbol{s}_{\rm sep}\subset\boldsymbol{s}_{\setminus r}$) in the logistic regression can perfectly predict the outcome variable (s_r), and leads to (theoretically infinitely) large estimates for the corresponding parameters. Most commonly the separation will be *quasi-complete* and only parameter estimates associated with $\boldsymbol{s}_{\rm sep}$ will be infinite, with the remaining parameter estimates remaining relatively unaffected. In real settings, where the logistic regression is solved numerically, the precise values of the separated parameters will not be infinite and instead depend on the convergence criteria of the numeric optimization scheme [40]. Methods that implicitly remove the first-order bias term through modifying the log-likelihood function [41] have been shown to control separation [40,42].

We perform a study of the first-order bias of PLM, and recontextualize the findings of previous authors in terms of these results. In particular, we connect the effects of

separation in the data sets (also observed in [2]) to criticality. For any SK-like model, the highly correlated nature of data drawn either from the critical point or the low-temperature SG and F phases means that there will always be a B-dependent threshold temperature below which separation will occur, and at which PLM can no longer estimate the parameters correctly. For nonseparated data, our work additionally shows that the bias behaves as

$$b_{1,ij}(\boldsymbol{h}^0, \boldsymbol{J}^0) \approx b_{1,ij}(\sigma^0), \tag{9}$$

that is, the bias is, to a good approximation, purely a function of the variance of the parameters (i.e., the inverse temperature).

A. Inference of SK models

We benchmark the PLM method on a zero-field ($\boldsymbol{h}^0 = 0$), fully connected Sherrington-Kirkpatrick (SK) model [16] with system sizes comparable to typical coarse-grained fMRI brain region analyses [13,53,54], $N = 50, 100, 200, 400, 800$, collecting a total of $B_{\max} = 10\,000$ samples per state-point.

We generate our input couplings $\boldsymbol{J}^0$ by drawing from a Gaussian distribution with mean $\mu^0 = \mu/TN$ and standard deviation $\sigma^0 = \sigma/TN^{1/2}$. Note that we have absorbed the temperature of Eq. (3) into our definition of the model parameters. We do this as for real applications the "temperature" (in a statistical physics sense) of the system is undefined, and only coupling strengths can be extracted. μ and σ are intensive variables, and the state of the system is characterized by the dimensionless average coupling strength μ/σ and temperature T/σ. We fix $\sigma = 1$ and sample from the range $\mu \in [0, 2]$, $T \in [0.5, 2]$, where in the $N \to \infty$ thermodynamic limit, the system explores all of its phases. These phases are a low-temperature disordered spin-glass (SG), low-temperature ordered ferromagnetic (F), and high-temperature disordered paramagnetic (P) phase. From previous findings, we expect the inference to perform best near the phase transitions [2,3]. We produce input time-series for every state-point via standard Monte Carlo simulations [44,45], sampling data every $1000N$ steps. We monitor the autocorrelation time τ to ensure that subsequent samples are decorrelated and can be considered as independent and identically distributed (i.i.d.). From the series, we compute the spin-glass order parameter (also known as the *overlap*)

$$q = \frac{1}{N}\sum_{i=1}^{N}\langle s_i\rangle^2, \tag{10}$$

and the covariance as

$$C^2 = \frac{1}{N}\sum_{i,j=1}^{N} C_{ij}^2, \tag{11}$$

which is related to the spin-glass susceptibility $\chi_{\mathrm{SG}} = C^2/T^2$ [13,55]. When approaching the SG phase from the higher-temperature P phase, the susceptibility (and hence the covariance) increases rapidly, reflecting the development of spontaneous large correlations near the phase transition, i.e., the critical point. We then perform unregularized PLM inference on each trajectory, as the models are not sparse. The quality of the inference is assessed by measuring the error

$$\varepsilon = \sqrt{\frac{\sum_{i\leqslant j}\left(\theta_{ij}^* - \theta_{ij}^0\right)^2}{\sum_{i\leqslant j}\left(\theta_{ij}^0\right)^2}}, \tag{12}$$

a robust aggregate measure of the deviations in parameter estimation previously defined in [3]. Here $\boldsymbol{\theta}$ is a symmetric matrix containing all PLM parameters, with $\theta_{ii} = h_i$ and $\theta_{ij} = J_{ij}$ as all $J_{ii} = 0$ (there are no self-couplings). Note that as the number of couplings $N_J = N(N-1)/2$ is much larger than the number of fields $N_h = N$, the error is dominated by contributions from the couplings. The biases enter the numerator of this expression, and a $1/B$ scaling of the error is expected.

Each generated model will deviate by a small amount from T, and so we define the *measured* temperature of each model realization as $T^0 = 1/(\sigma^0 N^{1/2})$, where σ^0 is the standard deviation computed from the realized couplings. We similarly define the measured inferred temperature as $T^* = 1/(\sigma^* N^{1/2})$, where σ^* is the standard deviation of the inferred couplings. This allows us to define a second global metric on the inference quality, i.e., how well the inferred model reproduces the temperature.

B. Error dependence on state-point

In Fig. 2 we illustrate the overall phase behavior of the SK model and of the inference error for $N = 200$ and $B = 10\,000$, with the phase boundaries of the $N = \infty$ system overlaid. In (a) we recover the known SK phase diagram, with low values for the overlap in the paramagnetic and spin-glass phase compared to the ferromagnetic phase. The phase transitions correspond to regimes of increased susceptibility, which peak at the phase boundaries but are blurred and shifted due to finite-size effects [56] as shown in (b). Deep in the F and P phases, C^2 is low due to the lack of fluctuations in the F phase, and a lack of spin-spin correlations in the P phase.

Figure 2(c) shows the performance of the PLM method as quantified by the error in (12). We observe a minimum in ε in the paramagnetic phase ($\mu = 0.4$, $T = 1.1$), and a rapid increase as the two correlated F and SG phases are approached. This is consistent with previous studies of the fully connected SK model for $N = 64$ [2,3], who found the error to be minimized around $T \sim 1$ for $\mu = 0$. We note importantly that although the error minimum is close to the peak of the critical fluctuations, the two are not coincident—in our simulations we find, e.g., the P-SG C^2 peak at $T \approx 0.6$, while finite-size studies of the SK model have shown the critical fluctuations of the specific heat of the P-SG transition to peak at $T \approx 0.7$ for similar system sizes [57]. The ε color-map is capped at $1.5\varepsilon_{\min}$, but much larger errors, $O(10^3$–$10^4) \times \varepsilon_{\min}$, are observed in the correlated F and SG phases. These large errors are due to the B configurations becoming highly correlated ($\tau \gg 1$), causing the separation effect introduced in Sec. III to lead to (infinitely) large parameter estimates. The numeric values of ε in these regions do not have a real meaning—they simply indicate that no maximum-likelihood solution can be found for some of the PLM logistic regressions.

The region of minimal error is characterized by isocontours that run parallel to the phase transition lines, implying a strong dependence of the error on the *distance* from the

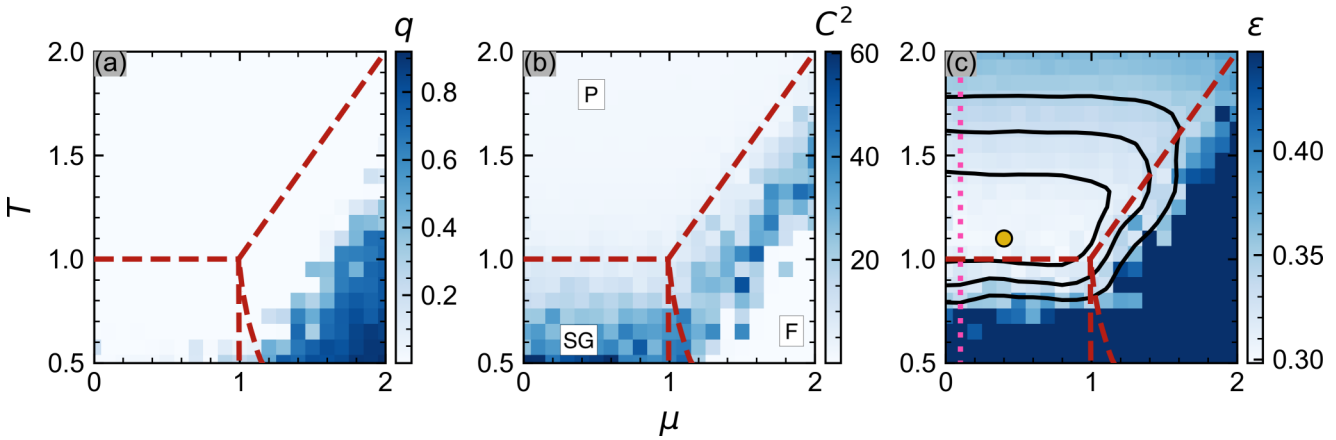

FIG. 2. Overview of the order parameter (a), susceptibility (b), and error (c) for the zero-field SK phase diagram with $N = 200$ and $B = 1 \times 10^4$. The values of the observables shown at each state-point were calculated by averaging over three independent realizations at that state-point. Red dashed lines show phase-transition lines in the $N \to \infty$ limit. (b) Labels P, F, and SG indicate the locations of paramagnetic, ferromagnetic, and spin-glass phases in the thermodynamic limit. (c) Contours of ε are shown in black, with the pink dotted line labeling the line $\mu = 0.1$ across which a more detailed examination of the error is made in Fig. 3. The location of the minimum error is denoted by an orange circle, and it occurs in the P phase above the P-SG boundary. Note that ε is thresholded so that the maximum plotted value is $\varepsilon_{\max} = 1.5\varepsilon_{\min}$.

phase boundary. This is especially clear at low μ, where the ε-contours lie roughly along lines of constant T. To understand the relationship between the inference error and the emergence of criticality, we also study a profile of the phase diagram at fixed $\mu = 0.1$, away from the ferromagnetic phase. In Fig. 3 we compare the inference error and C^2 as a function of T. This confirms that for $N = 200$, ε has a flat minimum, centered around a temperate $T_{\min} \approx 1.1$. The shape of the minimum is asymmetric, with ε diverging slowly as T goes from $T_{\min} \to \infty$ and rapidly as $T_{\min} \to 0$. At the minimum, the fluctuations $C^2(T_{\min})$ are three to four times larger than their high-temperature limit. But as T decreases further, the fluctuations continue to increase while the error rapidly diverges. The divergence occurs without a significant increase in the autocorrelation time, indicating that it is not due to poor sampling. The minimum error is therefore within the regime of enhanced critical fluctuations, and occurs close-to but offset-from the phase transition. At the supposed finite size critical temperature, the inference fails due to the inherently highly correlated nature of the data from this regime (τ is large).

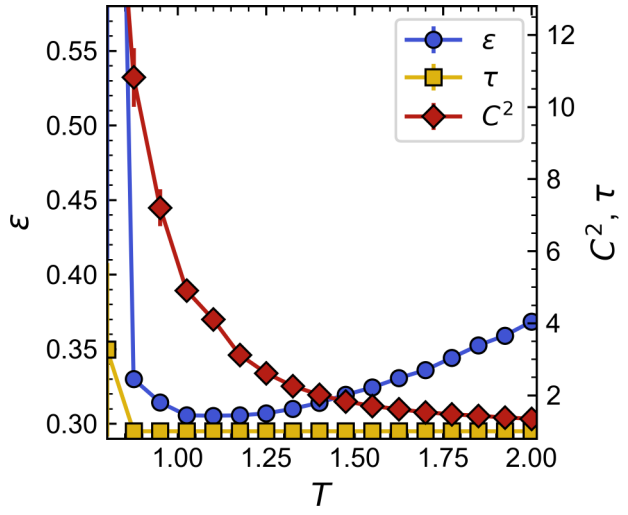

FIG. 3. Plot of the error ε, autocorrelation time τ, and correlation measure C^2 for varying T at fixed $\mu = 0.1$. τ is expressed in units of $1000N$. Each point is calculated by averaging over MC trajectories of length $B = 10^4$ with sampling frequency $1000 \times N$ generated from 21 independent realizations of the SK model at that temperature. Error bars are the standard errors of the observables over these 21 trajectories. Values $T < 0.8$ are not plotted as τ began to substantially deviate from 1 on the approach to the SG regime.

C. Error origin and impact on critical fluctuations

The inference error ε is the combination of the error on each individual parameter, dominated by the couplings J_{ij}. For highly correlated data we know the PLM inference will fail due to separation. We want to better understand the origin of the error when inference is possible, and to do this we directly inspect the probability distribution of $\boldsymbol{J}$ for a selection of sample sizes B. Figure 4(a) shows that the inferred distributions remain symmetric and appear approximately Gaussian, but that they systematically overestimate the variance, gradually converging to that of the input distribution with increasing B. The origin of this spread can be explained in terms of the MLE bias—for logistic regression, the parameter estimates are known to be biased away from 0, i.e., overestimated [41], which in our case spreads the overall distribution of parameters. Since $T^* = 1/(\sigma^* N^{1/2})$, applying PLM to small data sets leads to an inaccurate estimate of the state-point, biased towards a lower temperature. For disordered data sets, this corresponds to biasing the model towards the near-critical regime. The question is, how large is this effect?

In Fig. 4(b) we plot the dependence of the inferred temperature on the input temperature for various B. Taking the worse case, $B = 1000$, as an example, we find that the inference provides incorrect estimates of the state-point in the

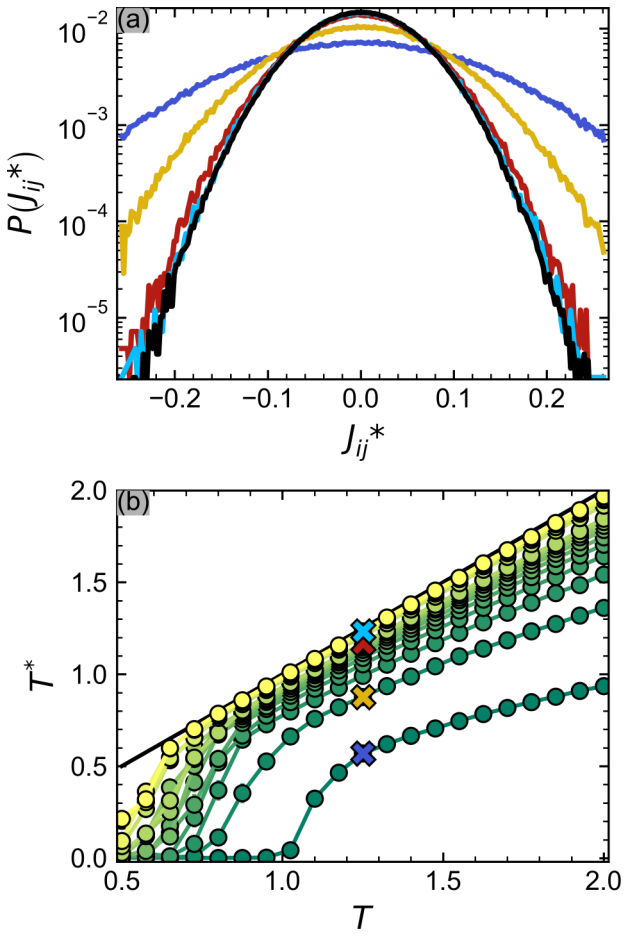

FIG. 4. (a) Probability distributions of the inferred couplings from the PLM inference for different varying B from the same state-point ($\mu = 0.1, T = 1.25$) near the minimum ε in Fig. 3. Dark-blue, orange, red, and light-blue (outer to inner in grayscale) lines correspond to $B = \{1, 2, 10, 50\} \times 10^3$, respectively. The black innermost line shows the ground truth distribution for reference. (b) The inferred temperature as a function of the input temperature for $B = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50\} \times 10^3$. The darkest green (lower) line shows results for the smallest B, and the lightest yellow (upper) line corresponds to the largest B. Colored crosses (lower to upper) indicate temperatures of the respective distributions (outer to inner) in (a). $N = 200$, with points and error bars representing the mean and standard error of the PLM temperature estimate from MC trajectory length B of 21 independent model realizations at each T.

entire range of temperatures considered. Even at the optimal conditions, where the inference error is minimal, $T^* \approx 0.4T^0$ mislabelling the state. Parameter estimates which suffered from separation in Fig. 4(b) are indicated by low $T^* \to 0$, as the anomalously large inferred parameters cause the variance to explode. Collecting more data allows the onset of separation to be delayed, and lower temperature state-points closer to the P-SG transition to be correctly characterized.

We expect that the misattributed temperatures will cause the inferred models to exhibit falsely enhanced critical fluctuations. This is because the spin-glass correlations C^2 and susceptibility χ_{SG} increase rapidly as T decreases and one crosses the P-SG transition line. To investigate this, we resimulate the models inferred via PLM for each T and produce an

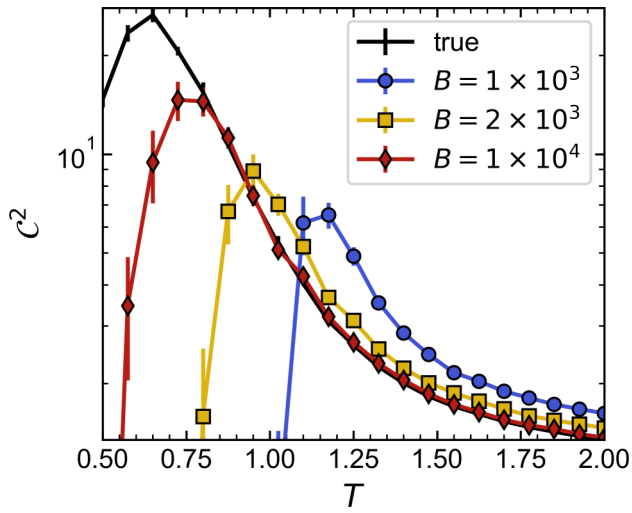

FIG. 5. Susceptibility measure C^2 as a function of input temperature T for three data quality conditions; $B = 1 \times 10^3$ (circles), $B = 2 \times 10^3$ (squares), $B = 1 \times 10^4$ (diamonds). The black line shows the susceptibility as measured from simulations of the true model. Each data point and error bar in the black line represents the mean and standard error calculated over 3 (for each B condition) $\times$ 21 = 63 independent model realizations.

estimate of the correlations C^2 corresponding to the inferred model. This process can be summarized as follows. For every state point:

(i) Produce 21 independent data sets of sample size B. We proceed by initializing the spins randomly and equilibrating initially every run for 10^6 (discarded) MC sweeps, where every sweep corresponds to $1000N$ Monte Carlo steps. We then evolve the system for $B \times 1000 \times N$ steps to collect B, independent configurations.

(ii) Extract 21 PLM models (one per data set).

(iii) Run 6 MC simulations using the PLM estimate for $10^5 \times N$ steps, sampled every $10N$ steps.

(iv) Evaluate C^2 over each simulation and average.

Figure 5 shows the corresponding results for C^2; the black line represents the correlations of the input models, while the colored symbols show those corresponding to the PLM estimates for three different sample sizes. As suspected from the temperature shifts in Fig. 4, we observe that PLM overestimates C^2 for input models generated at high T. With decreasing T, C^2 reaches a peak value, which gets higher and is located closer in T to the true C^2 peak with increasing B. For small sample sizes, the location of the peak corresponds to the onset of separation, as the arbitrarily strong couplings found for $T < T_{\mathrm{peak}}$ fix $C^2 \to 0$. In summary, our numerical experiments on the SK model indicate that PLM on small data sets provides couplings that underpredict T and artificially enhance C^2. PLM will therefore misattribute finite data sets stemming from the fluctuating P phase to a near-critical state point.

D. Temperature dependence on sample size

In the previous section, we qualitatively described that the inferred temperature depends on B. Here we quantitatively

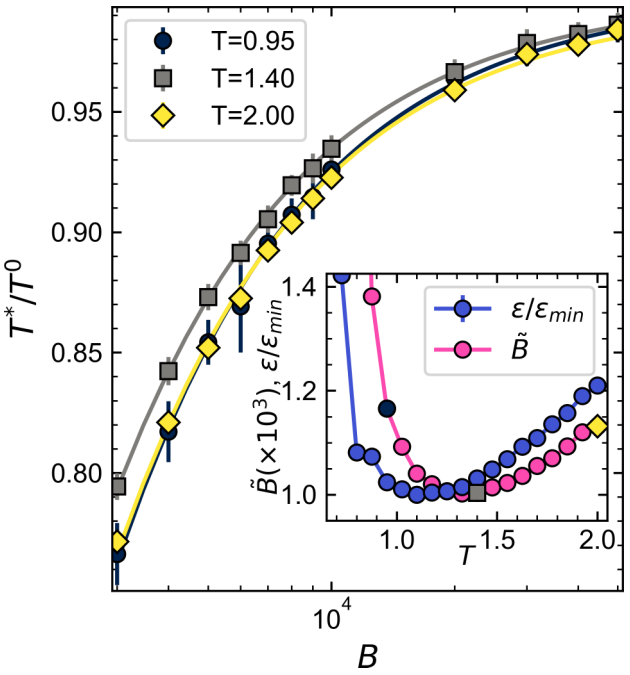

FIG. 6. T^*/T^0 as a function of B for three illustrative input temperatures; $T = 0.95$ (circles), $T = 1.4$ (squares), and $T = 2$ (diamonds). Points and error bars for each B are means and standard errors obtained by repeating the simulation and inference process for 21 independent input model realizations at each T. Inset: The dependence of the empirical scaling parameter $\tilde{B}(T)$ (light pink) and the error $\varepsilon(T)$ (dark blue) when $B = B_{\max} = 5 \times 10^4$ samples are used for PLM. Shaded symbols show $\tilde{B}$ for the corresponding $T^*/T^0(B)$ saturation curves in the main body of the figure. Coefficient of determination $R^2 > 0.980$ of the heuristic arc-tan fit (13) for all $\tilde{B}$ plotted.

demonstrate that this effect is governed by the slow $1/B$ convergence of the MLE bias, and that the temperature of the input model sets the learning difficulty of the problem. Figure 6 shows the dependence of the ratio T^*/T^0 on the sample size for different T, where again $\mu = 0.1$. We find that for increasing B, the curves follow a saturating behavior, which can be fitted with high accuracy to the following heuristic model:

$$T^* = (2T_{B\to\infty}/\pi) \times \arctan(B/\tilde{B}), \tag{13}$$

where $T_{B\to\infty}$ and $\tilde{B}$ are fitting parameters of the heuristic model. $\tilde{B}$ quantifies the rate of the asymptotic first-order bias convergence and is shown in the inset of Fig. 6, alongside with ε. Both share a nonmonotonic behavior indicating a correspondence between the scale $\tilde{B}$ (quantifying the typical sample size to have small deviations in T^*/T^0) and the average error on the couplings ε. The minimum of $\tilde{B}$ is shifted further into the P phase, to $T \approx 1.4$. Note that for $N = 200$, the minimum number of samples is $\tilde{B} \gtrsim 1000$, pointing to the necessity of a minimum of several thousand samples for reliable inference. This highlights a poignant issue for real data sets; the bias dissipation parameter $\tilde{B}$ is not known *a priori* and can vary by orders of magnitude depending on the temperature of the input model. The limit defining a "small" data set therefore depends on the underlying state-point of the data.

We motivate the arc-tan fit by noting that for $x = B/\tilde{B} \geqslant 1$,

$$\arctan(x) = \frac{\pi}{2} - \frac{1}{x} + O(x^{-3}), \tag{14}$$

so that

$$T^* = T_{B\to\infty} - \frac{\tilde{B}'}{B} + O(B^{-3}) \tag{15}$$

follows the same first-order linear dependence on $1/B$ as (7). An equally valid approach would be to plot T as a function of $1/B$ and perform a linear fit to the asymptotic regime as is done in Fig. 1. These results imply that the dependence of the MLE bias prefactors $b_{1,ij}$ on the input model parameters $(\boldsymbol{h}^0, \boldsymbol{J}^0)$ can, to a good extent, be approximated by a statistical average of the parameter distribution, i.e., $b_{1,ij}(\boldsymbol{h}^0, \boldsymbol{J}^0) \approx b_{1,ij}(T^0)$.

We have chosen only to present results for $N = 200$ here as the analysis of other system sizes leads to identical conclusions. Although there is a sharpening of the phase transition with increasing N, the minimum error state-point remains offset from the transition temperature within the P phase. We note that we find $\tilde{B} \propto N$, i.e., the amount of data to solve the problem depends linearly on N. This dependence arises from the fact that although the full inverse Ising problem deals with estimating $N + N(N-1)/2$ parameters, each individual logistic regression equation only infers N parameters.

In summary, the PLM method on the SK model displays biases that scale as the inverse of the sample size B^{-1}. The magnitude of the bias strongly depends on the state-point of the input data. Small sample bias causes the temperature of the inferred model to be underestimated, falsely enhancing the critical fluctuations exhibited by models inferred from near-critical paramagnetic data. Any PLM model inferred from fluctuating (i.e., dynamically varying) data will thus appear as closer-to-critical than it actually is. In the following, we describe data-driven procedures to mitigate these effects.

IV. DATA-DRIVEN BIAS REDUCTION

Due to the $1/B$ convergence of the PLM temperature estimate, we suspect that methods that remove the first-order MLE bias may provide better estimates of the state-point. A range of such methods exists in the literature [58], and they can be largely grouped into explicit methods that correct for the bias correction *after* inferring the parameters, e.g., jackknife resampling [59], and implicit methods that correct the bias *during* inference via a modification (penalization) of the likelihood function [40–42,60,61].

In this section, we propose an explicit correction to the PLM parameter estimates, which aims to best capture the critical properties of the data by enforcing self-consistency with C^2. We benchmark this against an implicit bias correction, Firth's penalized logistic regression [40–42]. We will assess both methods based on their ability to estimate T and C^2 for a range of input temperatures.

A. Correcting the bias via self-consistent C^2

In Sec. III, we demonstrate that (a) the MLE bias of the parameters is captured by a global property of the parameter

distribution (the temperature), and (b) that the PLM models overestimate C^2. We thus propose to correct the bias by requiring that the inferred models display C^2 as close as possible to the one estimated from the input data: in this sense, we aim at inferring models whose fluctuations are self-consistent with those of the input data set. To do so, we perform a second optimization *after* estimating the PLM parameters. Again denoting the PLM parameter estimates by $\boldsymbol{\theta}^*$, we optimize the objective function

$$\mathcal{L}'(T_f) = \left[C^2_{\text{input}} - C^2_{\text{MC}}(T_f)\right]^2, \tag{16}$$

where C^2_{input} is C^2 measured as from the input data set and $C^2_{\text{MC}}(T_f)$ is calculated from MC simulations of a rescaled PLM model with parameters $\boldsymbol{\theta}_{T_f} = \boldsymbol{\theta}^*/T_f$. The rescaling parameter $T_f > 0$ acts as a *fictitious temperature* which can shift the state-point of the inferred model. We denote the optimal value of T_f by $T_f^\dagger$, with the corresponding corrected parameter estimates being $\boldsymbol{\theta}^\dagger$. C^2 has a strong dependence on B, so we match the amount of data in the input and in the MC simulations $B_{\text{data}} = B_{\text{MC}}$. In practice, we calculate C^2_{MC} for six independent MC simulations of length B_{data} for every T_f, and then feed the average over these six runs into Eq. (16).

The results of this corrective procedure are shown in Fig. 7. The correction significantly improves the reconstructed temperature and, by design, perfectly matches the fluctuations of C^2 when separation does not occur. The improvement to the T^* prediction is particularly pronounced for small data sets and at high T. At low T, where separation occurs, the corrective optimization fails to converge and $C^2_{\text{input}} \neq C^2(T_f^\dagger)$. This highlights a pitfall of explicit methods; they inherit any instabilities of the original MLE, such as those that lead to separation in logistic regression [40,58]. We note that the improvement itself can also be used as a score on the reliability of the PLM estimates and as an indication of the necessity of more data, with $T_f \to 1$ as $B \to \infty$.

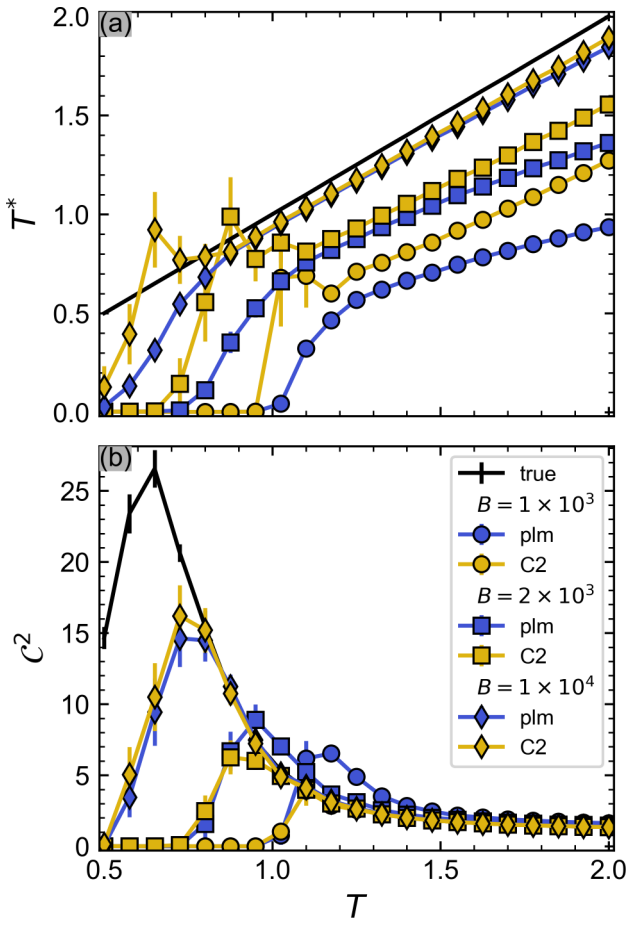

FIG. 7. The inferred temperature T^* [panel (a)] and the susceptibility measure C^2 [panel (b)] at different input temperatures T for three data quality conditions; $B = 1 \times 10^3$ (circles), $B = 2 \times 10^3$ (squares), $B = 1 \times 10^4$ (diamonds). Blue (dark) lines show observables for the PLM inference, orange (light) lines show observables after performing the self-consistency (C2) correction. Black lines are the temperature and susceptibility of the input models at each state point. $N = 200$, $\mu = 0.1$, with all plotted points again corresponding to means and standard errors over 21 model realizations.

B. Firth's penalized logistic regression

One may also remove the first-order bias implicitly through Firth's penalized likelihood maximization [41]. In our notation, this corresponds to maximizing the penalized log-likelihood $\mathcal{L}'_r$:

$$\mathcal{L}'_r(\{\boldsymbol{s}\}_B|h_r, \boldsymbol{J}_r) = \mathcal{L}_r(h_r, \boldsymbol{J}_r|\{\boldsymbol{s}\}_B) + 0.5 \ln |F(h_r, \boldsymbol{J}_r)|, \tag{17}$$

where $|F(h_r, \boldsymbol{J}_r)|$ is the determinant of the Fisher information matrix for each row of parameters r. Full details of this method will not be given here, but see [40,41,58] for appropriate overviews. We implement this computationally by modifying the Python code available at Ref. [62]. The corrective term of Eq. (17) tends to 0 as $B \to \infty$, returning the unpenalized likelihood. For small B the penalty compensates the $O(B^{-1})$ bias, and is known to control separation in logistic regression [40,42].

In Fig. 8 we demonstrate the effect of the penalty on the inferred parameters at a low $T = 0.57$ and a high $T = 1.6$ state-point. At low T separation causes the unpenalized PLM parameters associated with specific s_i to diverge, with $\max(\theta_{ij}^{\text{PLM}}) \approx 400$, leading to a non-Gaussian highly spread PDF. In our language, models with these very strong couplings correspond to zero temperature. Firth's penalty controls this effect, and although we still see large parameter estimates for the same s_i, these are orders of magnitude smaller with $\max(\theta_{ij}^{\text{Firth}}) \approx 4$. Firth's correction reduces the spread of the inferred distribution and largely captures the Gaussian nature of the input coupling distribution, even at low T. It therefore provides better T^* estimates than unpenalized PLM.

C. Comparing methods

In Fig. 9 we compare the performance of the self-consistency (C2) correction and Firth's (firth) correction for the smallest data set, $B = 1000$, where bias effects are most extreme. We again generate data from 21 independent model realizations for each T and then apply each inference method to each data set separately. We naively assess if the separation has occurred by checking if $|\boldsymbol{\theta}^*|_{\max} > \mu^* + 10\sigma^*$, where $|\boldsymbol{\theta}^*|$ is the absolute value of the inferred parameters from each inference scheme. Any input T where *a single* inferred model

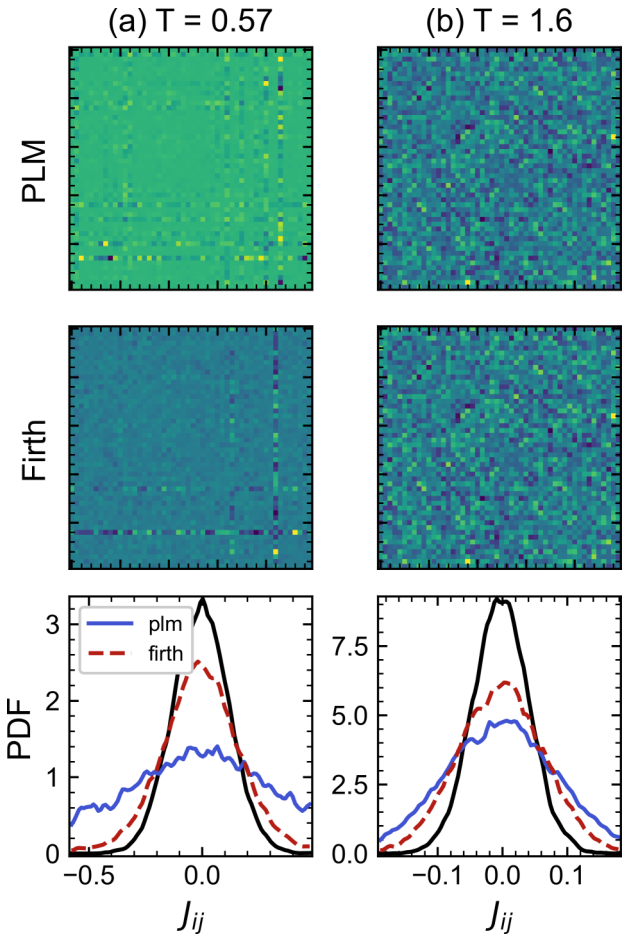

FIG. 8. Parameter matrices $\boldsymbol{\theta}$ for a subset of 50 spins for unpenalized PLM (top) and Firth's penalized PLM (middle) at two different temperatures. (Bottom) probability density functions (PDFs) of the corresponding inferred parameters (light colored lines) along with the true parameter PDF (black line). Firth's correction controls the inference at low T, leading to finite parameters and a PDF that more closely matches the true input model. At high T, Firth's correction shifts the inferred temperature towards the true temperature (by reducing the spread of the parameter PDF). $B = 10^3$, $N = 200$, and $\mu = 0.1$.

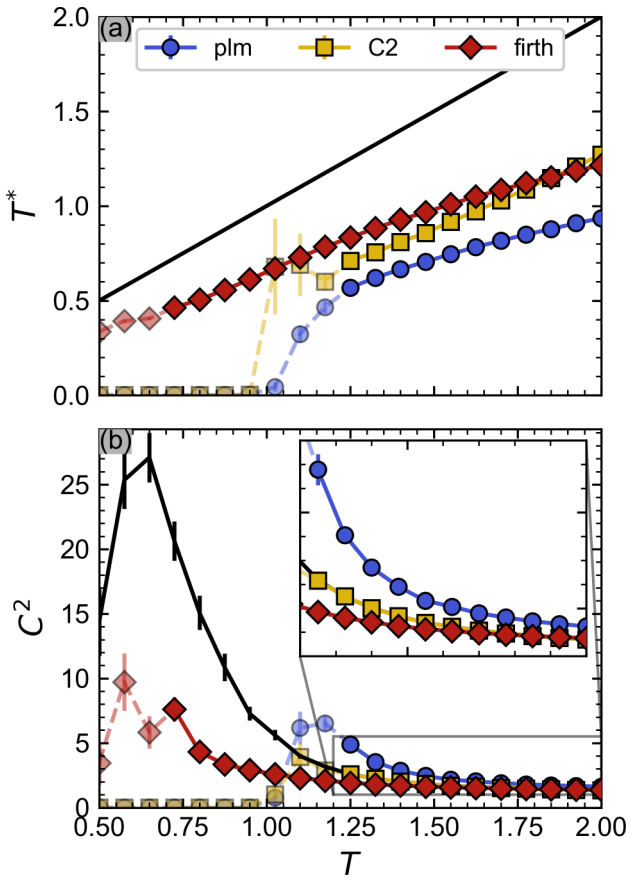

FIG. 9. Comparison of the corrective methods' ability to capture the criticality relevant observables T^* and C^2 for a small sample size $B = 10^3$. Transparent points indicate T where separation occurred. Firth's correction gives reasonable estimates for T^* at much lower T than PLM, highlighting this method's ability to control separation. In contrast to normal PLM, Firth's correction underestimates C^2 for all T. At high T, Firth's correction and the C2 correction perform similarly. Each point represents an average over 21 independent data sets at that T.

satisfied the separation condition (i.e., had "anomalously" large parameters) is indicated by the transparent points in Fig. 9. We note that this is a conservative definition; if even a single inference fails, we characterize the whole state point as separation-prone. We see that (according to our definition) the onset of separation is delayed to much lower T when using Firth's penalty, and that Firth's penalized logistic regression predicts nonzero T^* for all T. At high T both corrections perform similarly well in improving the estimated T^*, although the dependence $T^*(T)$ appears to scale more favorably using the C2 correction as T increases.

Figure 9(b) shows how well each method reproduces the correlations of the input data. We observe an interesting tradeoff; although Firth's correction provides better estimates for T, C^2 of the corresponding models is systematically underpredicted. Firth's correction thus fails to capture the correlations of the data. We note that this contrasts with the unpenalized logistic regression results for PLM, which instead overestimate C^2. The main advantage of Firth's correction over the C2 correction, therefore, appears to be that lower temperature state-points can be estimated.

This naturally raises a question: can we apply the C2 correction to the models inferred using Firth's penalization and improve the estimate of both T and C^2? The answer is negative: unsurprisingly, applying the C2 correction to the penalized parameter estimates increases C^2 at the cost of lowering T^*, and ultimately leads to the same estimates as simply applying the C2 correction to the unpenalized PLM model. We summarize these findings in Table I. There thus appears to be an inherent tradeoff between capturing the temperature and the correlations of a data set. When deciding which method is "best," one must therefore decide which property is most important to encode correctly.

D. Implications for inference around criticality

So far we have shown that small-sample biases influence the determination of the state of Ising models inferred using

TABLE I. Percentage errors of inferred estimates T^* and C^2 for a single model realization at $T = 1.025$, $\mu = 0.1$, with $B = 10^4$ using a range of inference schemes. Firth's correction provides the best estimate of the temperature but the worst estimate of the critical fluctuations. We demonstrate an implicit tradeoff between correctly inferring T and C^2. Applying the C2 correction either to the PLM model or to the Firth corrected model produces the same T^* and C^2 pair. The temperature is underestimated, irrespective of the inference scheme used.

Method	T^* % error	C^2 % error
PLM	−6.9	3.4(8)
PLM → C2	−6.0	0.2(9)
Firth	−4.0	−7.7(5)
Firth → C2	−6.1	−0.1(8)

PLM. The problem is that what constitutes a "small" sample size itself depends on the state-point (and topology) of the true model that generated the data. Studies claiming criticality in Ising models inferred using PLM thus need to control for bias, for example through subsampling their data and performing a similar analysis as in Fig. 6. They should also consider that the PLM model they infer from any data set which is dynamic (i.e., fluctuates) will be biased *towards* the critical point and exhibits enhanced critical fluctuations when simulated. In such cases, the C2 correction may be applied to rescale the couplings and match the empirical correlations. This will also shift the inferred temperature toward the true temperature. We note, however, that the C2 corrected temperature remains systematically smaller than the true temperature, and should only be considered as a lower-bound estimate of the true temperature. It may be appropriate to use Firth's implicitly corrected penalized logistic regression if separation is found to occur. This correction will provide reasonable parameter estimates even for low T state points, but it should be noted the fluctuations displayed by these models are not representative of the data, which is important when considering near-critical phenomena.

V. CASE STUDY: CRITICALITY OF AN FMRI DATA SET

In this section, we demonstrate the effect of the bias in a real setting by analyzing a human brain imaging data set obtained from a *functional magnetic resonance imaging* (fMRI) study. Brain imaging data sets are particularly interesting from the perspective of criticality as a range of advantageous computational properties have been attributed to systems operating in the near-critical state [23–30,34,63–68]. These have given rise to the idea that the brain is tuned towards a critical point, and a range of experimental results support this *critical brain hypothesis* [30–33,37,64,65,69,70]. PLM was recently used to contribute to this body of work, correlating IQ to the spin-glass susceptibility [13], and inverse Ising inference, in general, has previously been used to map different cognitive states to different disordered spin models [12]. Such mapping of the fMRI signals to the Sherrington-Kirkpatrick interaction model allows us to make the brain criticality hypothesis more explicit and quantitative, but it also requires a number of additional assumptions: for example, that the signals can be meaningfully binarized or that the interactions between brain regions are exclusively pair-wise. It is important to note that the SK model is the maximum entropy model reproducing *any* second-order statistics of binary variables [4], so that if our measurements only concern second-order statistics, then the SK model is sufficient. One could also consider alternative models to regress the data, including oscillatory models such as the Hopf model [71–73], the Kuramoto model [74–76], or even the Hawkes point process [52,77]. The presence and form of critical-like fluctuations will depend on the model of choice, meaning that there are additional uncertainties in the estimate of the distance from criticality that we perform using the inverse Ising inference on the SK model.

We consider data from a single-participant resting-state fMRI study, the full experimental details of which can be found in [53]. Imaging sessions were carried out on separate days under two different conditions: those where the participant practiced mindfulness meditation (MM) before undergoing imaging, and those where he did not (noMM). During each session, $B = 236$ samples were collected from $N = 399$ regions of interest (ROIs) within the brain. We will consider each ROI as a spin s_i and perform inverse Ising inference using PLM. Note that the trajectories for each ROI, $s_i(t)$, obtained from the preprocessing in [53] are continuous. We therefore binarized the data by removing the average from the signal and setting $s_i(t) < 0 = -1$ and any $s_i(t) \geqslant 0 = +1$. In total, $B_{\text{noMM}} = 40 \times 236 = 9440$ samples were collected for the noMM condition and $B_{\text{MM}} = 18 \times 236 = 4248$ for the MM condition. We investigate whether PLM inference leads to the identification of a close-to-critical state, and if this state is affected by the biological condition (i.e., practicing mindfulness meditation). Moreover, we study if the inference biases identified here significantly impact our conclusions.

The distributions of the PLM parameters obtained from the *full* data sets are shown in Fig. 10(a). First, as opposed to the SK model, the distributions are skewed, with a long tail at positive values of the couplings. Second, the MM condition corresponds to a larger variance in the couplings than the noMM one. The immediate consequence is that the mapped temperatures of the two full data sets are different, $T^*_{\text{full-MM}} = 0.98$ and $T^*_{\text{full-noMM}} = 1.33$.

However, the two sample sizes are $B_{\text{noMM}} \approx 2B_{\text{MM}}$. Can this lead to a significant statistical effect in the estimation of the corresponding state points? Figure 10(b) provides an affirmative and quantitative answer: as we subsample (ss) the noMM data, we find that the PLM temperature estimate decreases with reducing B, and when $B_{\text{ss-noMM}} = B_{\text{MM}}$, it meets the same estimated temperature of the MM data. Hence, for the data considered here, there is no significant difference between the MM and the noMM data when the statistical bias is taken into account. We further find that below some critical value, $B_c \approx 2000$ separation occurs leading to the failure of the inference.

To refine the estimate of the noMM state-point, we can apply the self-consistency correction (yellow circles) and fit the empirical saturation function, Eq. (13), to the PLM temperature estimates. Fitting data with $B \geqslant 2500$, we find $T_{B\to\infty} = 1.72$ and $\tilde{B} = 3465$. The available data sets are with $B_{\text{MM}} \approx 1.2\tilde{B}$ and $B_{\text{MM}} \approx 2.7\tilde{B}$, indicating that the small sample size bias strongly impacts our conclusions here.

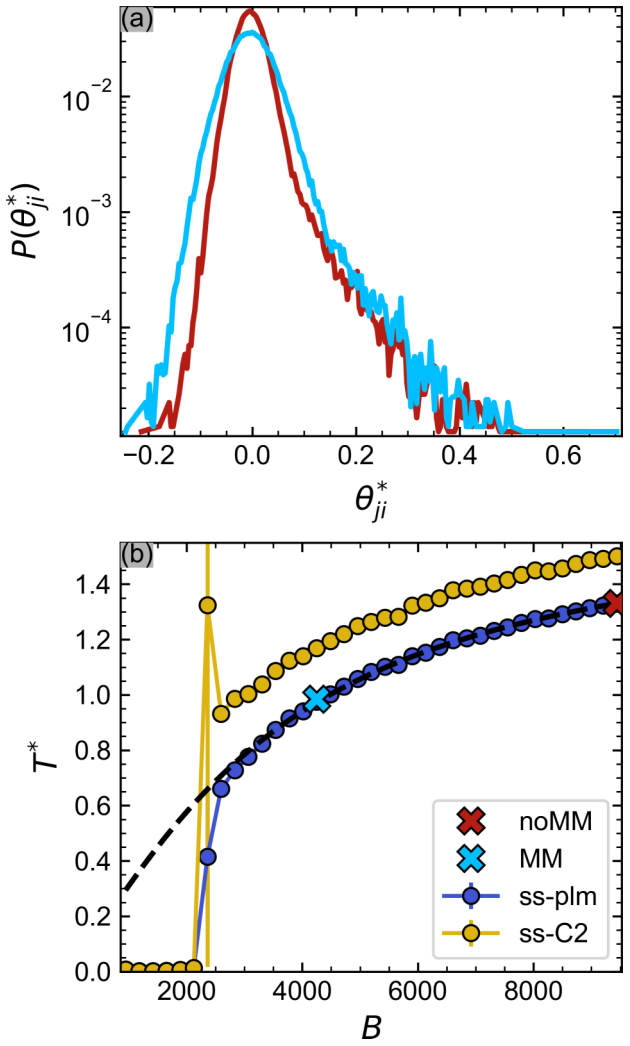

FIG. 10. (a) Inferred parameter distributions for the noMM condition in red (dark gray) and the MM condition in blue (light gray). The bias causes the MM distribution to appear more spread. (b) Subsampling analysis of the noMM data set. Blue (dark gray) points show PLM temperature estimates for each number of subsamples, and orange (light gray) points correspond to the temperatures of the self-consistency corrected model. The fitted empirical model (13) is shown by the dashed black line. Red (dark gray) and blue (light gray) crosses show the temperatures corresponding to the distributions in (a).

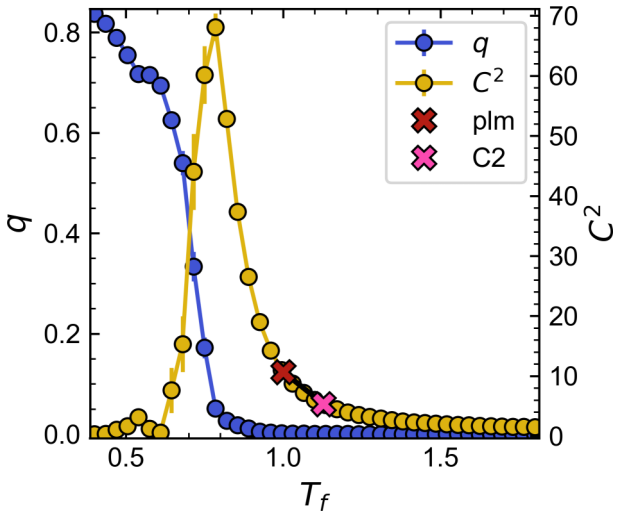

FIG. 11. Spin-glass order parameter q (blue/dark gray data points) and susceptibility C^2 (orange/light gray data points) as functions of the fictitious temperature T_f for the model inferred from the no-MM condition, $\boldsymbol{\theta}^*_{\text{noMM}}$. The ferromagnetic order parameter m was also measured and found to be 0 throughout. This sweep, therefore, characterizes a transition from a low-temperature spin-glass to a high-temperature paramagnetic phase. Points and error bars correspond to means and standard errors of 60 independent MC simulations with $B = 10^4$ samples. Red (dark gray) and pink (light gray) crosses correspond, respectively, to the PLM and self-consistency corrected models of the noMM condition.

To contextualize the meaning of the inferred temperatures in Fig. 10(b) we again introduce the fictitious temperature T_f and perform MC simulations of $\frac{1}{T_f}\boldsymbol{\theta}^*_{\text{noMM}}$ for a range of T_f; see Fig. 11. A peak of C^2 at $T_f = 1$ would mean that the PLM model is situated exactly at the critical point. We instead find the peak at $T_f = T_c = 0.78$ and so $\boldsymbol{\theta}^*_{\text{noMM}}$ is a paramagnetic state-point above the transition, albeit still with substantial critical fluctuations $C^2(T_f = 1) \approx 0.15 C^2_{\text{max}}$. The self-consistency correction shifts the model further from T_c. Hence, the MM and noMM conditions appear to be at best near critical if not paramagnetic.

Summarizing these results, initial analysis of the two conditions would lead to the conclusions that (a) practicing mindfulness meditation changes the state-point of the brain, and (b) the noMM condition represents a near-critical paramagnetic state-point. Carefully accounting for the bias instead reveals that both data sets more likely originate from the same state-point, and the C2 corrected temperature estimate shows that, as a lower bound, the true state-point of the data lies far from the transition in the paramagnetic phase. We find, therefore, no evidence to suggest that the resting state fluctuations in this imaging study correspond to a critical brain state.

VI. CONCLUSIONS

In this work, we have studied the importance of small sample size biases in the pseudo-likelihood maximization (PLM) approach to inverse Ising inference. Although PLM is exact in the limit of a large sample size, we show that this condition is often unlikely to be achieved in real-world data sets. We demonstrate that estimates of important physical quantities (such as the temperature) that define the state of the inferred model depend linearly on $1/B$, similarly to the standard bias of maximum likelihood estimators.

We present a detailed study of the fully connected SK model for $N = 200$. The above biases cause models inferred from paramagnetic data sets to exhibit enhanced critical fluctuations, with this effect worsening with decreasing B. Paramagnetic data are therefore misclassified as near-critical for finite B, i.e., PLM underestimates the distance from criticality. The inference error is minimized close-to but offset from the phase transition in the paramagnetic regime. The

development of strong correlations on the approach to the critical point means that PLM fails due to separation at lower T. We note that, although information-theoretic arguments suggest otherwise [52], for small or intermediate B the regime of failure occurs *before* the finite-size critical temperature T_c is reached.

We describe data-driven approaches to mitigate these effects. The self-consistent correction we propose improves the temperature estimate T^* while matching the critical fluctuations of the data set. It performs well when a PLM solution can be found, i.e., when separation does not occur, and it provides the best improvement to T^* at high T. For low T (or equivalently for small B), Firth's penalized logistic regression may be used to estimate the state point when standard PLM fails. Although this produces T^* closer to T, we caution that the critical fluctuations inferred using Firth's correction are not representative of the data. These models thus fail to capture an essential property of the system. Both the self-consistency correction and Firth's correction provide biased estimates of the temperature, with $T^* \to T^0$ from below as $B \to \infty$. The estimated temperatures T^* should therefore be considered as lower bounds on the true temperature of the data set. In contrast with other regularization techniques, neither correction requires a hyperparameter to be tuned.

We also show that the bias profoundly impacts the estimation of criticality in a real finite B data set from neuroscience. Not accounting for small sample size effects causes state points to be incorrectly classified. For fluctuating, i.e., dynamically varying data, this corresponds to inferring models that are falsely tuned towards the critical point. Applying the self-consistency correction to this data set allows us to counteract this and establish that, as a lower bound, the data are paramagnetic. The above results lead us to conclude that any PLM study claiming criticality in a real data set must carry out a proper analysis of the dependence of the inference on B, e.g., through the subsampling scheme we describe. Otherwise small sample size biases cannot be ruled out as the primary cause of the criticality of the inferred model. This is especially important as we have shown that the bias prefactors, e.g., $\tilde{B}$, setting the learning difficulty of the model are functions of the state (and also topology [14]), and that one thus cannot establish *a priori* what constitutes a "small" sample size.

The finite-sample effects impacting on the parameters inferred via pseudo-likelihood maximization are expected to be generic features, as they are a consequence of the maximum-likelihood procedure that lies at the core of the method. This suggests that similar finite-size effects in the inference of the properties of generic spin systems should be observed around the critical regions.

With a view to the future, experimental evidence for criticality appears to be ripe across biological systems [37]. Maximum entropy approaches, such as the PLM solution for inverse Ising inference, provide an attractive route with which to investigate these, by allowing the criticality of the data to be assessed in the framework of statistical physics. We show here that any such study must make careful consideration of small sample size biases before claiming a system to be critical. This is especially important as the distance to criticality is becoming an increasingly relevant biological variable [70], and, in neuroscience for instance, it is starting to be considered a guiding route for clinical work [78].

The code used for this study is available as a Python package at Ref. [79].

ACKNOWLEDGMENTS

This work was supported by the EPSRC Centre for Doctoral Training in Functional Materials: The Bristol Centre for Functional Nanomaterials (BCFN) with grant code EP/L016648/1. N.M. acknowledges support from the Japan Science and Technology Agency (JST) Moonshot R&D (under Grant No. JPMJMS2021).

[1] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. **106**, 620 (1957).
[2] E. Aurell and M. Ekeberg, Inverse Ising Inference Using All the Data, Phys. Rev. Lett. **108**, 090201 (2012).
[3] H. C. Nguyen, R. Zecchina, and J. Berg, Inverse statistical problems: From the inverse Ising problem to data science, Adv. Phys. **66**, 197 (2017).
[4] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population, Nature (London) **440**, 1007 (2006).
[5] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing, Proc. Natl. Acad. Sci. (USA) **106**, 67 (2009).
[6] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Maximum entropy models for antibody diversity, Proc. Natl. Acad. Sci. (USA) **107**, 5405 (2010).
[7] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Statistical mechanics for natural flocks of birds, Proc. Natl. Acad. Sci. (USA) **109**, 4786 (2012).
[8] A. Tang, D. Jackson, J. Hobbs, W. Chen, J. L. Smith, H. Patel, A. Prieto, D. Petrusca, M. I. Grivich, A. Sher, P. Hottowy, W. Dabrowski, A. M. Litke, and J. M. Beggs, A Maximum Entropy Model Applied to Spatial and Temporal Correlations from Cortical Networks In Vitro, J. Neurosci. **28**, 505 (2008).
[9] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. J. Berry, Searching for collective behavior in a large network of sensory neurons, PLoS Comput Biol **10**, e1003408 (2014).
[10] T. Watanabe, S. Hirose, H. Wada, Y. Imai, T. Machida, I. Shirouzu, S. Konishi, Y. Miyashita, and N. Masuda, A pairwise maximum entropy model accurately describes resting-state human brain networks, Nat. Commun. **4**, 1370 (2013).

[11] T. Watanabe, S. Hirose, H. Wada, Y. Imai, T. Machida, I. Shirouzu, S. Konishi, Y. Miyashita, and N. Masuda, Energy landscapes of resting-state brain networks, Front. Neuroinform. **8**, (2014).
[12] T. Watanabe, N. Masuda, F. Megumi, R. Kanai, and G. Rees, Energy landscape and dynamics of brain activity during human bistable perception, Nat. Commun. **5**, 4765 (2014).
[13] T. Ezaki, E. Fonseca dos Reis, T. Watanabe, M. Sakaki, and N. Masuda, Closer to critical resting-state neural dynamics in individuals with higher fluid intelligence, Commun. Biol. **3**, 52 (2020).
[14] A. Y. Lokhov, M. Vuffray, S. Misra, and M. Chertkov, Optimal structure and parameter learning of Ising models, Sci. Adv. (2018).
[15] A. Decelle and F. Ricci-Tersenghi, Pseudolikelihood Decimation Algorithm Improving the Inference of the Interaction Network in a General Class of Ising Models, Phys. Rev. Lett. **112**, 070603 (2014).
[16] D. Sherrington and S. Kirkpatrick, Solvable Model of a Spin-Glass, Phys. Rev. Lett. **35**, 1792 (1975).
[17] S. Kirkpatrick and D. Sherrington, Infinite-ranged models of spin-glasses, Phys. Rev. B **17**, 4384 (1978).
[18] J. R. L. de Almeida and D. J. Thouless, Stability of the Sherrington-Kirkpatrick solution of a spin glass model, J. Phys. A **11**, 983 (1978).
[19] G. Parisi, Toward a mean field theory for spin glasses, Phys. Lett. A **73**, 203 (1979).
[20] G. Parisi, A sequence of approximated solutions to the S-K model for spin glasses, J. Phys. Math. Gen. **13**, L115 (1980).
[21] G. Parisi, Order Parameter for Spin-Glasses, Phys. Rev. Lett. **50**, 1946 (1983).
[22] T. Castellani and A. Cavagna, Spin-glass theory for pedestrians, J. Stat. Mech. (2005) P05012.
[23] O. Kinouchi and M. Copelli, Optimal dynamical range of excitable networks at criticality, Nat. Phys **2**, 348 (2006).
[24] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale, Scale-free correlations in starling flocks, Proc. Natl. Acad. Sci. (USA) **107**, 11865 (2010).
[25] E. Tagliazucchi, P. Balenzuela, D. Fraiman, and D. R. Chialvo, Criticality in Large-Scale Brain fMRI Dynamics Unveiled by a Novel Point Process Analysis, Front. Physio. **3**, (2012).
[26] P. Rämö, J. Kesseli, and O. Yli-Harja, Perturbation avalanches and criticality in gene regulatory networks, J. Theor. Biol. **242**, 164 (2006).
[27] G. Deco and V. K. Jirsa, Ongoing Cortical Activity at Rest: Criticality, Multistability, and Ghost Attractors, J. Neurosci. **32**, 3366 (2012).
[28] N. Bertschinger and T. Natschläger, Real-time computation at the edge of chaos in recurrent neural networks, Neural Comput. **16**, 1413 (2004).
[29] R. Legenstein and W. Maass, Edge of chaos and prediction of computational performance for neural circuit models, Neural Netw. **20**, 323 (2007).
[30] J. M. Beggs and D. Plenz, Neuronal avalanches in neocortical circuits, J. Neurosci. **23**, 11167 (2003).
[31] J. M. Beggs, The criticality hypothesis: How local cortical networks might optimize information processing, Philos. Trans. R. Soc. A **366**, 329 (2008).
[32] C. Haldeman and J. M. Beggs, Critical Branching Captures Activity in Living Neural Networks and Maximizes the Number of Metastable States, Phys. Rev. Lett. **94**, 058101 (2005).
[33] J. M. Beggs and N. Timme, Being critical of criticality in the brain, Front. Physio. **3** (2012).
[34] J. Wilting and V. Priesemann, 25 years of criticality in neuroscience—established results, open controversies, novel concepts, Curr. Opin. Neurobiol. Computat. Neurosci. **58**, 105 (2019).
[35] V. Priesemann, M. Valderrama, M. Wibral, and M. Le Van Quyen, Neuronal avalanches differ from wakefulness to deep sleep – evidence from intracranial depth recordings in humans, PLoS Comput Biol **9**, e1002985 (2013).
[36] T. Mora and W. Bialek, Are biological systems poised at criticality? J. Stat. Phys. **144**, 268 (2011).
[37] M. A. Muñoz, Colloquium: Criticality and dynamical scaling in living systems, Rev. Mod. Phys. **90**, 031001 (2018).
[38] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, High-dimensional Ising model selection using l1-regularized logistic regression, Ann. Statist. **38**, 1287 (2010).
[39] A. Albert and J. A. Anderson, On the existence of maximum likelihood estimates in logistic regression models, Biometrika **71**, 1 (1984).
[40] C. Zorn, A solution to separation in binary response models, Polit. Anal. **13**, 157 (2005).
[41] D. Firth, Bias reduction of maximum likelihood estimates, Biometrika **80**, 27 (1993).
[42] G. Heinze and M. Schemper, A solution to the problem of separation in logistic regression, Statist. Med. **21**, 2409 (2002).
[43] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, A learning algorithm for Boltzmann machines*, Cogn. Sci. **9**, 147 (1985).
[44] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. **21**, 1087 (1953).
[45] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika **57**, 97 (1970).
[46] Y. Roudi, E. Aurell, and J. Hertz, Statistical physics of pairwise probability models, Front. Comput. Neurosci. **3**, 652 (2009).
[47] Y. Roudi, J. Tyrcha, and J. Hertz, Ising model for neural data: Model quality and approximate methods for extracting functional connectivity, Phys. Rev. E **79**, 051915 (2009).
[48] 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 É. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).
[49] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. **74**, 47 (2002).
[50] M. E. Newman, The structure and function of complex networks, SIAM Rev. **45**, 167 (2003).
[51] A. Montanari and J. Pereira, Which graphical models are difficult to learn? *Advances in Neural Information Processing Systems 22*, edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (NIPS, 2009).
[52] I. Mastromatteo and M. Marsili, On the criticality of inferred models, J. Stat. Mech. (2011) P10012.
[53] S. Kajimura, N. Masuda, J. K. L. Lau, and K. Murayama, Focused attention meditation changes the boundary and configuration of functional networks in the brain, Sci. Rep. **10**, 18426 (2020).

[54] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen, A multi-modal parcellation of human cerebral cortex, Nature (London) **536**, 171 (2016).

[55] K. H. Fischer and J. A. Hertz, *Spin Glasses*, 1st ed. (Cambridge University Press, Cambridge, 1991).

[56] K. Binder, Finite size effects on phase transitions, Ferroelectrics **73**, 43 (1987).

[57] T. Aspelmeier, A. Billoire, E. Marinari, and M. A. Moore, Finite-size corrections in the Sherrington–Kirkpatrick model, J. Phys. A **41**, 324008 (2008).

[58] I. Kosmidis, Bias in parametric estimation: Reduction and useful side-effects, WIREs Comput. Stat. **6**, 185 (2014).

[59] R. G. Miller, The Jackknife–A review, Biometrika **61**, 1 (1974).

[60] I. Kosmidis and D. Firth, Bias reduction in exponential family nonlinear models, Biometrika **96**, 793 (2009).

[61] I. Kosmidis and D. Firth, A generic algorithm for reducing bias in parametric estimation, Electron. J. Statist. **4**, 1097 (2010).

[62] https://github.com/jzluo/firthlogist.

[63] S. Haykin, J. C. Principe, T. J. Sejnowski, J. McWhirter, T. G. Dietterich, M. I. Jordan, and S. Haykin, *New Directions in Statistical Signal Processing: From Systems to Brains* (MIT Press, Cambridge, MA, 2006).

[64] J. M. Beggs, Neuronal avalanches are diverse and precise activity patterns that are stable for many hours in cortical slice cultures, J. Neurosci. **24**, 5216 (2004).

[65] D. Plenz, Neuronal avalanches and coherence potentials, Eur. Phys. J. Spec. Top. **205**, 259 (2012).

[66] S. Yu, H. Yang, O. Shriki, and D. Plenz, Universal organization of resting brain activity at the thermodynamic critical point, Front. Syst. Neurosci. **7**, (2013).

[67] D. Marinazzo, M. Pellicoro, G. Wu, L. Angelini, J. M. Cortés, and S. Stramaglia, Information transfer and criticality in the ising model on the human connectome, PLoS ONE **9**, e93616 (2014).

[68] W. Bialek, Perspectives on theory at the interface of physics and biology, Rep. Prog. Phys. **81**, 012601 (2018).

[69] N. Friedman, S. Ito, B. A. W. Brinkman, M. Shimono, R. E. L. DeVille, K. A. Dahmen, J. M. Beggs, and T. C. Butler, Universal Critical Dynamics in High Resolution Neuronal Avalanche Data, Phys. Rev. Lett. **108**, 208102 (2012).

[70] J. O'Byrne and K. Jerbi, How critical is brain criticality? Trends Neurosci. **45**, 820 (2022).

[71] P. A. Robinson, C. J. Rennie, and D. L. Rowe, Dynamics of large-scale brain activity in normal arousal states and epileptic seizures, Phys. Rev. E **65**, 041924 (2002).

[72] M. Breakspear, J. A. Roberts, J. R. Terry, S. Rodrigues, N. Mahant, and P. A. Robinson, A unifying explanation of primary generalized seizures through nonlinear brain modeling and bifurcation analysis, Cereb. Cortex **16**, 1296 (2006).

[73] F. Freyer, J. A. Roberts, R. Becker, P. A. Robinson, P. Ritter, and M. Breakspear, Biophysical mechanisms of multistability in resting-state cortical rhythms, J. Neurosci. **31**, 6353 (2011).

[74] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence* (Springer, New York, 1984).

[75] J. A. Acebrón, L. L. Bonilla, C. J. P. Vicente, F. Ritort, and R. Spigler, The kuramoto model: A simple paradigm for synchronization phenomena, Rev. Mod. Phys. **77**, 137 (2005).

[76] H. Choi and S. Mihalas, Synchronization dependent on spatial structures of a mesoscopic whole-brain network, PLoS Comput. Biol. **15**, e1006978 (2019).

[77] A. G. Hawkes, Spectra of some self-exciting and mutually exciting point processes, Biometrika **58**, 83 (1971).

[78] V. Zimmern, Why brain criticality is clinically relevant: A scoping review, Front. Neural Circuits **14**, 54 (2020).

[79] https://github.com/maxkloucek/pyplm.