Inverse finite-size scaling for high-dimensional significance analysis

Yingying Xu,[1,2,*] Santeri Puranen,[1,2,4] Jukka Corander,[3,4] and Yoshiyuki Kabashima[5]
[1]*Department of Computer Science, School of Science, Aalto University, 00076 Espoo, Finland*
[2]*Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland*
[3]*Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland*
[4]*Department of Biostatistics, University of Oslo, 0317 Oslo, Norway*
[5]*Department of Mathematical and Computing Science, School of Computing, Tokyo Institute of Technology, Tokyo 152-8552, Japan*

(Received 14 October 2017; revised manuscript received 17 April 2018; published 6 June 2018)

We propose an efficient procedure for significance determination in high-dimensional dependence learning based on surrogate data testing, termed inverse finite-size scaling (IFSS). The IFSS method is based on our discovery of a universal scaling property of random matrices which enables inference about signal behavior from much smaller scale surrogate data than the dimensionality of the original data. As a motivating example, we demonstrate the procedure for ultra-high-dimensional Potts models with order of 10^{10} parameters. IFSS reduces the computational effort of the data-testing procedure by several orders of magnitude, making it very efficient for practical purposes. This approach thus holds considerable potential for generalization to other types of complex models.

DOI: 10.1103/PhysRevE.97.062112

I. INTRODUCTION

Learning about dependencies among stochastic entities from very high-dimensional data is a central task in data science [1]. Despite extensive research efforts in this field over the last decade, selection of statistically significant signals characterized by rare events remains a considerable challenge.

Statistical testing against null hypotheses, which are defined by assuming no dependency among the entities, is the standard approach for selecting the statistically significant signals. Signals for which the probability that a summary statistic would be the same as or of greater magnitude than the actually observed value under a null hypothesis (p value) is lower than a threshold value are regarded as significant. Unfortunately, analytical expressions for such a significance threshold are intractable for most complex models. Therefore, the role of computational methods that assess the threshold from observed data is becoming more and more important in the current era of big data.

Surrogate data testing [2,3] is a representative example of such methods. In this method, the original data are repeatedly randomly shuffled or regenerated under a null hypothesis, and the model is refitted to each generated data instance. The signals from the original data model that do not deviate substantially from typical surrogate data signals (here termed the background distribution) can be considered noise and discarded from further analysis. After its introduction two decades ago, surrogate data testing has become a widely established inference procedure for complex models. It is conceptually similar to the approximate Bayesian computation (ABC) and likelihood-free inference techniques that have become popular in genetics, econometrics, and astronomy [4–10]. Despite its conceptual simplicity, such a procedure may remain impractical if the considered model is expensive to learn, as is often the case with complex models [11,12].

As a motivating example, we consider high-dimensional inverse Ising and Potts modeling in which model parameters, i.e., fields and couplings of Ising or Potts models, are determined to fit with given data. This model class has been intensively studied in the field of direct-coupling analysis (DCA), primarily in the context of protein sequence analysis [13–18], and more recently also for whole-genome analysis of bacterial populations [19,20]. An established practice in DCA applications to protein data is to retain a smaller number of top predictions by comparison of estimated couplings against biological ground truth deduced from crystal structure experiments. However, from the inference perspective, it would be attractive and necessary to have a formal statistical rule for determining which learned model parameters can be classified as noise and which as signal when the ground truth is not available.

Here we propose an efficient procedure to solve the high-dimensional significance analysis problem for DCA, termed the inverse finite-size scaling (IFSS) of a surrogate data test. Our method is based on the discovery that the distribution of parameter estimates under a null hypothesis after normalization by the standard deviation can be characterized by a single function determined only by a limited number of system parameters: the data dimensionality ratio (aspect ratio) $\alpha \equiv n/L$, where n and L represent the number of samples and features in data, respectively, and the data bias distribution $P(\boldsymbol{f})$, which is a collection of summary statistics of features calculated from data (Sec. IV). When the finite-size scaling (FSS) property holds, both the number of features L and samples n present in a data set can be reduced, while keeping their ratio α fixed, such that the resulting distribution of scaled parameter estimates remains the same. Such a property is particularly desirable because it can be applied to surrogate data generation and subsequent parameter learning. The dimensionality reduction

*yingying.xu@aalto.fi

as a consequence of the FSS property implies that significant signals can be extracted from the learned model much easier, as the computational and memory requirements may decrease up to several orders of magnitude, which is demonstrated here using high-dimensional real data.

II. PROBLEM SETTING

Our goal is to identify meaningful pairwise dependencies from L sites by analyzing n samples for those sites. Suppose we have a $n \times L$ dimension data matrix $X = \{x_{\mu i}\}$ ($\mu = 1,2,\ldots,n$ and $i = 1,2,\ldots,L$). Elements in the data matrix can take discrete values from Q states or categories. Assume that the entity relation is described by a Potts model (or an Ising model when $x_{\mu i} = \pm 1$)

$$P(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x};\boldsymbol{J},\boldsymbol{h})} \exp\left[\sum_{i}^{L}\sum_{a}^{Q}\delta(x_i,a)h_i(a) + \sum_{1\leqslant i<j\leqslant L}\sum_{a,b=1}^{Q}\delta(x_i,a)\delta(x_j,b)J_{ij}(a,b)\right], \quad (1)$$

where $Z(\boldsymbol{x};\boldsymbol{J},\boldsymbol{h})$ is the partition function. In the model, $J_{ij}(a,b)$ represents the direct dependencies between states a,b at sites i,j. Fitting the model to data X, regularized inference provides estimates of the model parameters capturing direct dependencies.

Unfortunately, in practical situations, the number of samples n is much smaller than the total number of estimated parameters $J_{ij}(a,b)$ and $h_i(a)$, which grows as $O(Q^2L^2)$. This implies that most estimates are representing only statistical noise and we need to screen a small portion of statistically meaningful couplings among them. The aim of this paper is to develop a computationally efficient procedure for such screening purposes utilizing the FSS property that holds for surrogate data.

III. ILLUSTRATION WITH SYNTHETIC DATA

We demonstrate first that the surrogate data testing approach is valid for inverse Ising or Potts modeling problems by a moderate dimensional synthetic example with $L = 1000, n = 500$, where the ground truth is known. We developed a sparse restricted Boltzmann machine (sRBM) simulator, a variant of the restricted Boltzmann machine technique (see the Appendix), which allows efficient generation of example data. The underlying coupling matrix is extremely sparse with only a small number of nonzero elements $J_{ij} \equiv \sqrt{\sum_{a,b} J_{ij}(a,b)^2}$. We used the pseudolikelihood maximization (plmDCA) [17] algorithm for learning $\boldsymbol{J}$ in the Potts model, which is described in Sec. IV A.

Figure 1(a) shows a result for the sRBM-generated data together with the randomly generated surrogate data. Since we are primarily interested in the large deviations of the coupling distribution, a rank plot is used to visualize the signal behavior. By ranking the parameter values in descending order and plotting in log scale, the tail behavior of the distribution is clearly visible. In Fig. 1(a) the red curve represents the average ranking behavior of 10 independent surrogate data tests. Here

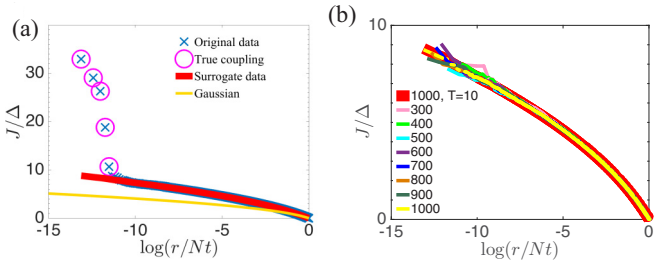

FIG. 1. plmDCA inference for synthetic binary data simulated by sRBM. On the horizontal axis, r is the rank in descending order and Nt refers to the total number of couplings. The function of log represents the natural logarithm. The vertical axis represents the coupling value divided by the standard deviation of all couplings. The simulated system size is $L = 1000$, $\alpha = 0.5$, and the surrogate data tests have been repeated $T = 10$ times. Synthetic data generation is done using bias distributions uniformly distributed over (0,1). Nonzero coupling strength is $J_{ij} = 2$ for all coupled pairs of variables $\{(i,j)\}$. (a) Full size surrogate data tests provide a significance threshold on rank plot. (b) Finite-size scaling property of the surrogate data for synthetic data.

the surrogate data sets are generated by randomly shuffling columns of the synthetic sRBM data; therefore the size are the same as the given data. The true nonzero couplings all deviate from the background distribution and remain above the maximum value of the surrogate data curve. Figure 1(b) shows a curve collapsing phenomenon for the surrogate data generated from the synthetic sRBM data. The procedure for creating surrogate data of different sizes is explained in Sec. V and Fig. 2. When the system size is varied while keeping the bias distribution (explained in Sec. IV) and aspect ratio n/L fixed, all the estimated coupling curves follow the same background distribution. The central tendencies especially are almost identical, while the tails exhibit more random variation.

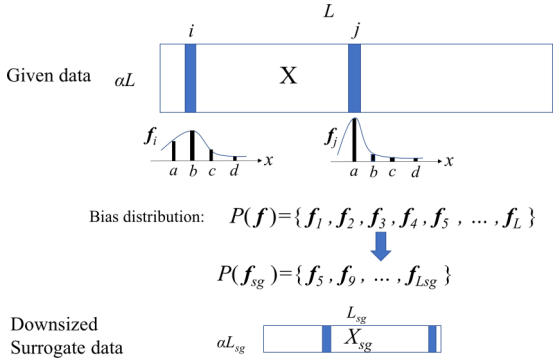

FIG. 2. Concept of bias distribution and generating downsized surrogate data from given data. Bias distribution is a collection of empirical sample (column) distributions ($\boldsymbol{f}$ vectors) of given big data. In the downsized surrogate data, bias distribution is kept by taking a small set of random samples of vector $\boldsymbol{f}$ from $P(\boldsymbol{f})$, which gives a collection of column bias vectors denoted as $P(\boldsymbol{f}_{\rm sg})$. Therefore, $P(\boldsymbol{f}_{\rm sg})$ represents the same distribution with $P(\boldsymbol{f})$. Keeping aspect ratio α the same with given data, one generates small surrogate data from $P(\boldsymbol{f}_{\rm sg})$, where the column position and the correspondence with $\boldsymbol{f}_i$ are random.

IV. FINITE-SIZE SCALING PROPERTY OF SURROGATE DATA

Suppose that samples of site i of the data set, which are represented as entries of column i in X, result in the state or category $q \in \{1,2,\ldots,Q\}$ count k_{qi} and that the ith column of X is denoted as $\boldsymbol{x}_i$. This feature is statistically characterized by the frequency vector $\boldsymbol{f}_i$ composed of Q elements $f_{qi} = k_{qi}/n$, where $q = 1,2,\ldots,Q$, and the collection $\{\boldsymbol{f}_1,\boldsymbol{f}_2,\ldots,\boldsymbol{f}_L\}$ are the summary statistics of the features, which we denote as $P(\boldsymbol{f})$ (Fig. 2) and term the *bias distribution*. In order to create a surrogate data under the null hypothesis one randomly shuffles or regenerates a surrogate sample for $\boldsymbol{x}_i$ of the same dimensionality n, while keeping $\boldsymbol{f}_i$ fixed, repeating this for all columns of X. We denote the obtained surrogate data matrix of the same dimensionality $n \times L$ as $X_{\rm s}$.

We discovered that given an algorithm $\mathcal{A}$, the standardized (scaled by standard deviation) distribution of learned interaction parameters from surrogate data depends only on a limited number of summaries of the data, namely, the aspect ratio and the bias distribution. This relation can be described as a function

$$P(\tilde{J}_{\rm sg}) = g(\tilde{J}_{\rm sg}|\alpha, P(\boldsymbol{f})), \tag{2}$$

where $\tilde{J}_{\rm sg} = J_{\rm sg}/\Delta$ and Δ is the standard deviation of the estimated coupling values for surrogate data. We term relation (2) as the finite-size scaling (FSS) property, where in low-dimensional surrogate data mimics the properties of $P(\tilde{J}_{\rm sg})$ of high-dimensional surrogate data $X_{\rm s}$ given that specific scaling criteria are met.

Employing the FSS property, we propose a procedure to create downsized surrogate data as

$$X \to P(\boldsymbol{f}) \to P(\boldsymbol{f}_{\rm sg}) \to X_{\rm sg}, \tag{3}$$

where $P(\boldsymbol{f}) = \{\boldsymbol{f}_1,\boldsymbol{f}_2,\ldots,\boldsymbol{f}_L\}$ and $P(\boldsymbol{f}_{\rm sg}) = \{\boldsymbol{f}_{s_1},\boldsymbol{f}_{s_2},\ldots,\boldsymbol{f}_{s_{L_{\rm sg}}}\}$ as shown in Fig. 2. $L_{\rm sg}$ is a feature number of the shrunken surrogate data which satisfies $L_{\rm sg} \ll L$. Elements in set $\{s_1,s_2,\ldots,s_{L_{\rm sg}}\}$ are randomly chosen numbers from feature indices $\{1,2,\ldots,L\}$ of given data X. We compute an empirical bias distribution $P(\boldsymbol{f})$ from the data matrix X, where $\boldsymbol{f}$s are $Q \times 1$ dimensional vectors and Q is the number of states in categorical data. By keeping the aspect ratio $\alpha = n/L$ of data matrix X, one can define a convenient surrogate test data size $L_{\rm sg}$ which is much smaller than the original data size and then randomly sample $L_{\rm sg}$ relative frequency vectors $\{\boldsymbol{f}_{s_1},\boldsymbol{f}_{s_2},\ldots,\boldsymbol{f}_{s_{L_{\rm sg}}}\}$ from $P(\boldsymbol{f})$, which provides the bias distribution $P(\boldsymbol{f}_{\rm sg})$ for the reduced-scale surrogate data. One can generate downsized surrogate data $X_{\rm sg}$ according to $P(\boldsymbol{f}_{\rm sg})$ with size $L_{\rm sg}$ and $n_{\rm sg} = \alpha L_{\rm sg}$, where the column position and the correspondence with index i of $\boldsymbol{f}_i$ from the given data are random.

A. Analyzing FSS behavior empirically

In order to show that the FSS property holds in general and does not depend on the type of learning algorithm used in DCA, we tested two distinct types of algorithms: regularized least squares (RLS) [21] and pseudolikelihood maximization (plmDCA) [17]. RLS is an ℓ_2-regularized inference method for DCA based on variational "naive mean-field" inference [1] where $\boldsymbol{J} = -\boldsymbol{C}^{-1}$ and $\boldsymbol{C}$ is the covariance matrix. For categorical data, the elements of $\boldsymbol{C}$ are defined as

$$C_{ij}(a,b) = F_{ij}(a,b) - F_i(a)F_j(b), \tag{4}$$

where

$$F_i(a) = \frac{1}{n_{\rm eff}}\left[\sum_{\mu=1}^{n}\omega(\mu)\delta(x_{\mu i},a)\right], \tag{5}$$

$$F_{ij}(a,b) = \frac{1}{n_{\rm eff}}\left[\sum_{\mu=1}^{n}\omega(\mu)\delta(x_{\mu i},a)\delta(x_{\mu j},b)\right] \tag{6}$$

are the frequencies calculated from data. In (5) and (6), $\omega(\mu)$ denotes the weight for sample μ (row μ in X) and $n_{\rm eff}$ is the effective sample number and $n_{\rm eff} = \sum_{\mu}^{n}\omega(\mu)$. There are several ways to calculate the weights for real data; however, for surrogate data, since the dependence between samples are destroyed, one could see the samples in surrogate data as independent from each other. Therefore, all samples have equal weights,

$$\forall\mu, \quad \omega(\mu) = 1 \quad \text{and} \quad n_{\rm eff} = n, \tag{7}$$

for surrogate data. RLS provides the estimated coupling matrix by the simple matrix equation

$$\boldsymbol{J}^{\rm RLS} = -\boldsymbol{C}(\eta\boldsymbol{1} + \boldsymbol{C}^2)^{-1}, \tag{8}$$

where $\boldsymbol{1}$ is the identity matrix and η is a positive regularization parameter.

On the other hand, another method, plmDCA, learns $\boldsymbol{J}$ in the Potts model so as to maximize the pseudo- (conditional) likelihood $P(x_i|\boldsymbol{x}_{\setminus i};\boldsymbol{J},\boldsymbol{h})$ for each element x_i given all the other elements $\boldsymbol{x}_{\setminus i}$ in conjunction with the regularization by the ℓ_2 norms of the couplings $\boldsymbol{J}$ and the external fields $\boldsymbol{h} = [h_i(a)]$ [17]. More precisely, the pseudolikelihood on site i for sample μ is defined as

$$P(x_i = x_{\mu i}|\boldsymbol{x}_{\setminus i} = \boldsymbol{x}_{\mu,\setminus i};\boldsymbol{J},\boldsymbol{h}) = \frac{\exp\left[h_i(x_{\mu i}) + \sum_{j\neq i}J_{ij}(x_{\mu j},x_{\mu i})\right]}{\sum_{a}^{Q}\exp\left[h_i(a) + \sum_{j\neq i}J_{ij}(a,x_{\mu i})\right]}, \tag{9}$$

and the regularized negative pseudo-log-likelihood function on site i is given by

$$l(\boldsymbol{h}_i,\boldsymbol{J}_i) = -\frac{1}{n_{\rm eff}}\sum_{\mu}^{n}\omega(\mu)\left(h_i(x_{\mu i}) + \sum_{j\neq i}J_{ij}(x_{\mu j},x_{\mu i}) - \log\left\{\sum_{a}^{Q}\exp\left[h_i(a) + \sum_{j\neq i}J_{ij}(a,x_{\mu i})\right]\right\}\right) + \lambda_h||\boldsymbol{h}_i||_2^2 + \lambda_J\sum_{j\neq i}||\boldsymbol{J}_{ji}||_2^2, \tag{10}$$

where $\boldsymbol{J}_i$ denotes $\{\boldsymbol{J}_{ji}\}_{j\neq i}$ and $||\boldsymbol{h}_i||_2^2 = \sum_a^Q h_i(a)^2$, $||\boldsymbol{J}_{ij}||_2^2 = \sum_{a,b}^Q J_{ij}(a,b)^2$. The plmDCA algorithm minimizes the total contribution $\sum_{i=1}^{L} l(\boldsymbol{h}_i,\boldsymbol{J}_i)$, which naively yields

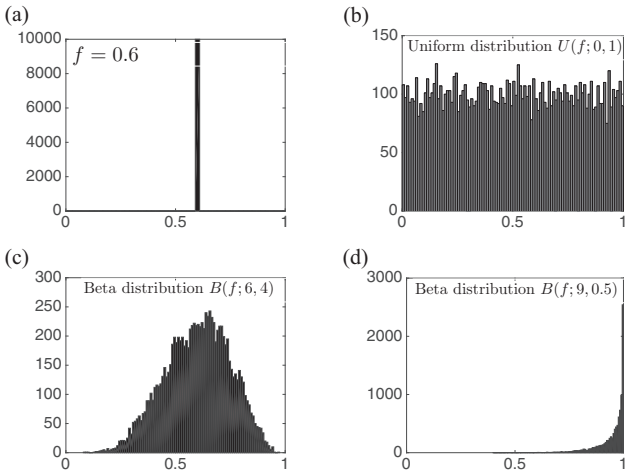

FIG. 3. Histogram of bias distribution of $q = 1$ dimension (data take first type of value). $f_{1i} = \frac{1}{n}\sum_{\mu=1}^{n}\delta(x_{\mu i} = 1)$ is the frequency of $x_{\mu i}$ taking value 1 in column i. These are histograms of $\{f_{1i}\}_{i=1,2,\ldots,L}$ when $n = 100$, $L = 10\,000$: (a) $f_{1i} = 0.6$ for all columns; (b) $\{f_{1i}\}_{i=1,2,\ldots,L}$ are generated from uniform distribution in interval (0,1); (c) $\{f_{1i}\}_{i=1,2,\ldots,L}$ are generated by beta distribution $B(\cdot;a,b)$ where parameters are set as $a = 6, b = 4$; (d) $\{f_{1i}\}_{i=1,2,\ldots,L}$ are generated by beta distribution $B(\cdot;9,0.5)$.

asymmetric couplings $J_{ij}(a,b) \neq J_{ji}(a,b)$. For resolving this drawback, $\boldsymbol{J}$ is symmetrized after the maximization.

We generated synthetic data from four representative types of bias distributions $P(\boldsymbol{f})$ and demonstrated the FSS property of the corresponding surrogate data. Figure 3 shows histograms of the four types of bias distributions for binary data. Figure 4 presents the log scale normalized histogram of standardized couplings learned by RLS for data generated using the bias distributions in Fig. 3. After scaled by the standard deviation of each coupling sets, all empirical distributions of the couplings collapse to a single curve. Similar results were also obtained by the plmDCA algorithm as shown in Fig. 5. For categorical data that learned using the plmDCA algorithm, the FSS property also holds for both $\{J_{ij}(a,b)\}$ (first row in Fig. 5) and coupling scores $\{J_{ij}\}$ (second row in Fig. 5).

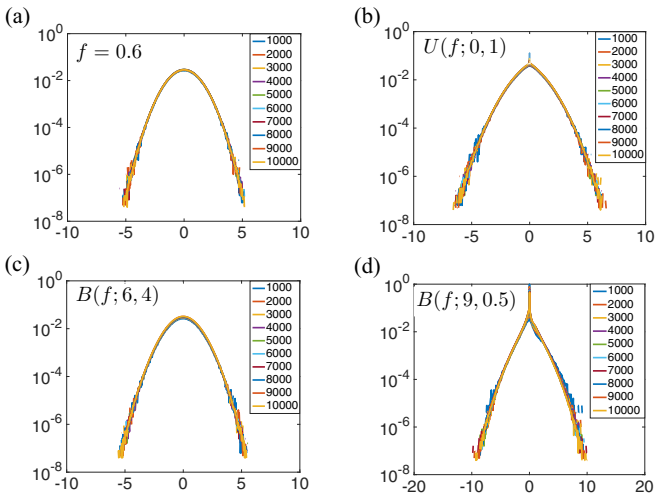

FIG. 4. Probability density plot in log scale of RLS learned couplings for binary data generated by four types of bias distributions shown in Fig. 3. Data $\{x_{\mu i}\}$ take the value of $\{+1,-1\}$. The feature number used in this series is $L = 1000,2000,\ldots,10\,000$, and the data aspect ratio is $\alpha = 0.1$. The regularization parameter used in the RLS algorithm is $\eta = 0.1$. All distributions are scaled by the standard deviation of each instance.

B. Analytically solvable case

Although the FSS property of surrogate data has been confirmed only empirically so far, it can be intuitively understood using random matrix theory. For the Gaussian special case, the model problem of RLS inference can be completely understood as follows [22]. Assume that the elements in the data matrix $X \in \mathbb{R}^{n\times L}$ are real and Gaussian distributed $\mathcal{N}(0,\sigma^2)$. The covariance matrix $\boldsymbol{C} = \frac{1}{n}X^\top X$ is then a Wishart matrix. The spectrum of $\boldsymbol{C}$ converges by the Marcenko-Pastur law almost surely to a limit when n and L tend simultaneously to infinity, and since $\boldsymbol{J}^{\mathrm{RLS}}$ is related to $\boldsymbol{C}$ by (8) the spectrum of $\boldsymbol{J}^{\mathrm{RLS}}$ is almost surely a nonlinear transformation of the Marcenko-Pastur distribution. Furthermore, the distribution of the individual elements of $\boldsymbol{J}^{\mathrm{RLS}}$ can be obtained from a singular value decomposition of matrix X,

$$X = USV^\top, \tag{11}$$

and regarding the left and right eigenbases as samples from the uniform distributions of orthogonal matrices. The covariance matrix can then be expressed as

$$\begin{aligned}\boldsymbol{C} &= \frac{1}{n}VS^2V^\top \\ &= V\Lambda V^\top \\ &= \left(\sum_{k=1}^{L} u_{ik}\lambda_k u_{jk}\right),\end{aligned} \tag{12}$$

where λ_k is the kth eigenvalue of covariance matrix $\boldsymbol{C}$ and u_{ik} is the kth element of the ith eigenbase of matrix $\boldsymbol{C}$. From (8) the inferred interaction matrix obtained by RLS is

$$\left(J_{ij}^{\mathrm{RLS}}\right) = \left(\sum_{k=1}^{L}\frac{\lambda_k}{\eta+\lambda_k^2}u_{ik}u_{jk}\right). \tag{13}$$

In this situation u_{ik} are samples from the uniform distributions of orthogonal matrices, and when the dimension of the matrix $\boldsymbol{C}$ goes to infinity u_{ik} can be handled as random numbers that satisfy

$$\overline{u_{ik}} = 0, \overline{u_{ik}u_{jl}} = \frac{1}{L}\delta_{ij}\delta_{kl}. \tag{14}$$

Condition (14) implies that each diagonal component converges to an $O(1)$ constant as

$$J_{ii}^{\mathrm{RLS}} = \left\langle\frac{\lambda}{\eta+\lambda^2}\right\rangle, \tag{15}$$

where the brackets $\langle\cdot\rangle$ denote the expectation with respect to eigenvalue distribution $\rho(\lambda)$ of the covariance matrix $\boldsymbol{C}$. The

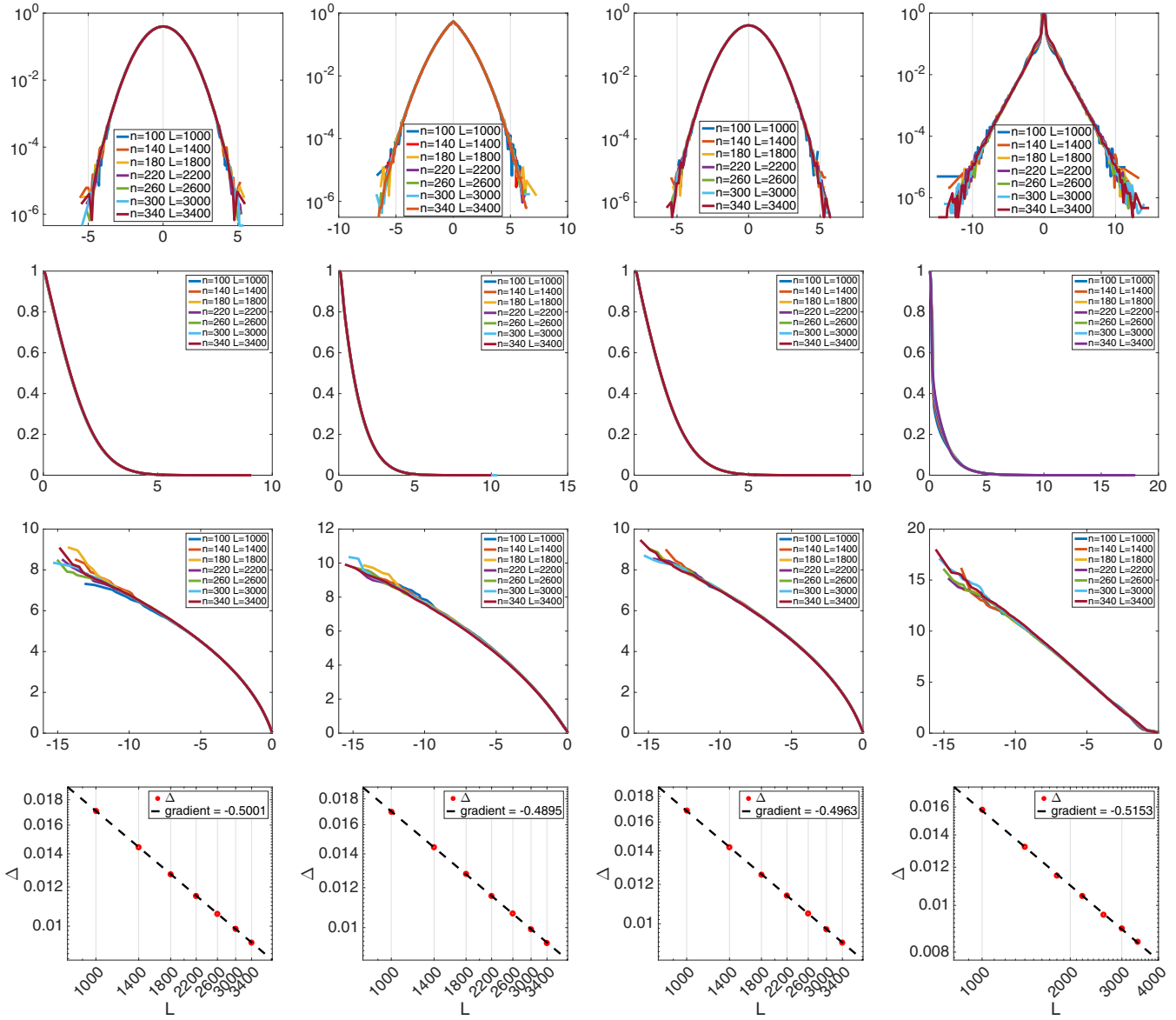

FIG. 5. Finite-size scaling property demonstration by plmDCA algorithm for synthetic data with four types of bias distributions as in Fig. 3. Column of figures from left to right corresponds to results from bias distribution of $f = 0.6$, $U(f;0,1)$, $B(f;6,4)$, and $B(f;9,0.5)$, respectively. Data for two categories were generated, and aspect ratio is set as $\alpha = 0.1$. Regularization parameters in plmDCA are set as $\lambda_J = 0.1, \lambda_h = 0.1$. The first row shows probability density plots of $\{J_{ij}(a,b)\}$ ($i \leqslant j, a, b$ correspond to two states in data, which are all normalized by the standard deviation of each instance). The second row stands for the cumulative coupling value distribution where coupling values are evaluated as $J_{ij} = \sqrt{\sum_{a,b}^{Q} J_{ij}(a,b)^2}$. They are also scaled by the standard deviation of each instance. The third row displays rank plots of the standardized coupling value distribution in the second row, which emphasizes the tail behavior of the distributions. The horizontal axis denotes $\log(r/N_t)$ where r is the ranking in descend order and N_t is the total number of couplings. The last row shows the relation between standard deviations of coupling values, which are used for standardization in the first three rows, and the system size L. The plots are in log-log scale, a clear linear trend appears for all bias types, and the gradients of the lines are all close to 0.5, which confirms the generality of the theoretical result in Eq. (18).

off-diagonal elements similarly follow a zero mean Gaussian distribution

$$J_{ij}^{\mathrm{RLS}} \sim \mathcal{N}(0, \Delta^2) \tag{16}$$

with variance

$$\Delta^2 \simeq \frac{1}{L-1}\left[\left\langle\left(\frac{\lambda}{\eta+\lambda^2}\right)^2\right\rangle - \left\langle\frac{\lambda}{\eta+\lambda^2}\right\rangle^2\right]. \tag{17}$$

For standard Gaussian data, $J_{ii}^{\mathrm{RLS}} = 0.0111$. The simulation shown in Fig. 6 confirmed these results for Gaussian data. The simplicity of RLS makes it tractable to define some of its properties analytically. In this special case, for any α, $P(\tilde{J}_{\mathrm{sg}})$ collapses to a *single function* of $\mathcal{N}(0,1)$. In contrast to RLS, the relationship between the estimated couplings J_{ij} obtained by the pseudolikelihood maximization algorithm is complicated and lacks an analytical expression. However, the

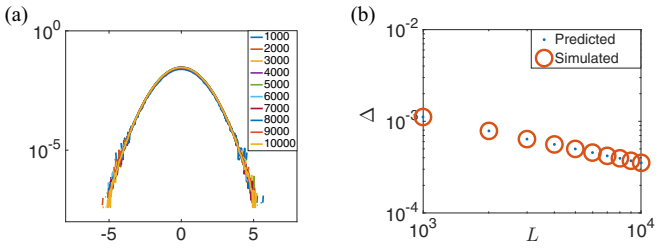

FIG. 6. (a) Probability density plot in log scale of inferred off-diagonal interactions $\boldsymbol{J}^{\mathrm{RLS}}$ obtained by the RLS algorithm for $\mathcal{N}(0,1)$ i.i.d. Gaussian data, $\eta = 0.1$. The number of loci in this series was $L = 1000, 2000, \ldots, 10\,000$, $\alpha = 0.1$. All distributions are scaled by the standard deviation of each instance. (b) $\Delta - L$ relation in log-log scale. Δ is the standard deviation of RLS inferred off-diagonal interactions. One-dimensional polynomial fitting of the simulated data points gives the gradient -0.4982, which is consistent with the predicted relation given by (17). Diagonal terms are 0.11 for all simulations matching the value predicted by (15).

Gaussian conclusion also holds for unbiased categorical data where elements in f are fixed constants for each category, as confirmed by RLS and plmDCA algorithms using simulated data as shown in Figs. 4 and 5.

Furthermore, the scaling property between Δ and L

$$\Delta \propto L^{-1/2} \tag{18}$$

is not only limited to Gaussian data, but holds for surrogate data for any given data in general. Figure 7 shows RLS algorithm simulated results for binary data generated from the four types of bias distribution in Fig. 3 (same data as in Fig. 4), which confirms the scaling relation (18). The same behavior also confirmed by the plmDCA algorithm in Fig. 5 (the last row). The functional relationship between Δ and L can be estimated by fitting the line $y = ax + b$ to the their logarithms, which

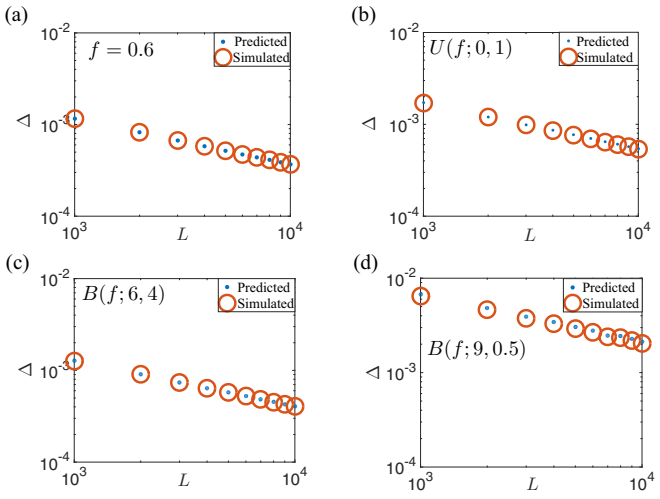

FIG. 7. Log-log $\Delta - L$ relation by RLS algorithm for binary data generated from the four types of bias distribution in Fig. 3 (same data as in Fig. 4), which confirms the scaling relation (18). Fitted gradient on the log-log scale is (a) -0.4983, (b) -0.4991, (c) -0.4966, (d) -0.4990.

gives

$$\Delta = e^b L^a. \tag{19}$$

The fitted gradient a values for these lines are all very close to 0.5, which is consistent with the scaling property predicted by theory (18). The standard deviation of large-scale surrogate data can consequently be predicted by obtaining parameters a and b from a few simulations using small surrogate data and then solving (19).

V. INVERSE FINITE-SIZE SCALING PROCEDURE

In conventional surrogate data testing one would randomly shuffle columns and/or rows of the original data matrix X, or use its empirical marginal distributions to generate random data under a null hypothesis. Instead of learning a model from such high-dimensional surrogate data, we consider a procedure to extract the statistically relevant characteristics of the surrogate data distribution using reduced dimensionality as described in (3) and Fig. 2. Computing any inferences or statistical tests on the downsized surrogate data X_{sg} is much more efficient since $L_{\mathrm{sg}} \ll L$. Invoking the finite-size scaling property of surrogate data requires only that bias distribution $P(\boldsymbol{f})$ and aspect ratio α are preserved.

The scaling property allows us to infer the background distribution of model parameter estimates for large L using a Monte Carlo approach for a smaller L. A drawback of this approach is that the distribution of the extremely large predictions occurring with a probability smaller than $2/[L(L-1)]$ cannot be accurately evaluated. However, we have observed that the tail of the normalized coupling distribution typically decreases as a convex curve [Fig. 1(b), Fig. 5 (the third row), Fig. 10(a), and Fig. 11(a)], which indicates that the straight line extrapolated from the tail in the rank plot acts as an upper bound for the extremely rare predictions. This is directly useful for screening relevant couplings from the background.

We translate the function in (2), which describes the probability distribution of standardized coupling values for surrogate data to a function describing curve behavior on the rank plot of the same distribution,

$$\tilde{J}_{\mathrm{sg}} = \bar{G}_{\mathrm{sg}}(f(r;L)), \tag{20}$$

where $f(r;L) = \log(r/N_t)$ represents the value on the x axis of the rank plot and $r = 1, 2, \ldots, N_t$ is the ranking in descending order. Since the total number of couplings is $N_t = L(L-1)/2$, $f(r;L)$ is a function dependent on system size L. A single test of surrogate data results in a curve $G_{\mathrm{sg}}(\cdot)$ on the rank plot and is subject to random fluctuations. Therefore, taking an average of the curves of many independent surrogate data sets gives a more stable curve $\bar{G}_{\mathrm{sg}}(\cdot)$ on the rank plot, which represents the distribution (2). After averaging a sufficient number of independent finite-size surrogate data tests with system size $L_{\mathrm{sg}}(\ll L)$, we obtain a curve $\bar{G}_{\mathrm{sg}}(f(r;L\mathrm{sg}))$ on the rank plot. Note that $\bar{G}_{\mathrm{sg}}(f(r;L\mathrm{sg}))$ is with system size L_{sg}, which is a shorter curve compared to $\bar{G}_{\mathrm{sg}}(f(r;L))$; however, it overlaps with $\bar{G}_{\mathrm{sg}}(f(r;L))$ because of the FSS property. When $\bar{G}_{\mathrm{sg}}(f(r;L\mathrm{sg}))$ shows a convex tail, a line $l(f(r;L))$ can be fitted on the tail, which produces an actual upper bound estimate for $\bar{G}_{\mathrm{sg}}(f(r;L))$. To decide a significance threshold, we focus on the largest value (corresponding to $r = 1$) of $\tilde{J}_{\mathrm{sg}}$

Algorithm: INVERSE FINITE-SIZE SCALING($\mathbf{X}(n \times L), \tilde{\mathbf{J}}; L_{\mathrm{sg}}, T$)

1) **Extract statistical feature from data** :
   $\alpha, \mathbf{P}(\mathbf{f}) \leftarrow \mathbf{X}$
2) **Finite-size scaled surrogate data test** :
   Sample $\mathbf{P}(\mathbf{f}_{\mathrm{sg}})$ (set number L_{sg}) : $\mathbf{P}(\mathbf{f}_{\mathrm{sg}}) \leftarrow \mathbf{P}(\mathbf{f})$
   Generate $\mathbf{X}_{\mathrm{sg}}(n_{\mathrm{sg}\times L_{\mathrm{sg}}})$: $\mathbf{X}_{\mathrm{sg}} \leftarrow \alpha, \mathbf{P}(\mathbf{f}_{\mathrm{sg}})$
   Learn $\mathbf{J}_{\mathrm{sg}}$ by algorithm $\mathcal{A}$: $\mathbf{J}_{\mathrm{sg}} \leftarrow \mathbf{X}_{\mathrm{sg}}$
   Standardization : $\tilde{\mathbf{J}}_{\mathrm{sg}} \leftarrow \mathbf{J}_{\mathrm{sg}}/\boldsymbol{\Delta}_{\mathrm{sg}}$
3) **Repeat step 2)** T **times**
4) **Prepare rank plot** :
   Averaging : $[\tilde{\mathbf{J}}_{\mathrm{sg}}] \leftarrow \{\tilde{\mathbf{J}}_{\mathrm{sg}}\}$
   Plot together
   $\bar{G}_{\mathrm{sg}}(f(r; L_{\mathrm{sg}}))$: $[\tilde{\mathbf{J}}_{\mathrm{sg}}]$ VS $\log(r/Nt_{\mathrm{sg}})$
   $G(f(r; L))$: $\tilde{\mathbf{J}}$ VS $\log(r/Nt)$
5) **Draw significance threshold** :
   Tail fitting : $l(f(r; L)) \leftarrow \bar{G}_{\mathrm{sg}}(f(\mathrm{tail}; L_{\mathrm{sg}}))$
   Threshold for $\tilde{\mathbf{J}}$ is $l(f(1; L))$.

FIG. 8. Inverse finite-size scaling (IFSS) procedure for deciding a significance threshold for dependence learning. X is the original data matrix, and the model parameters $\tilde{\mathbf{J}}$ are learned by algorithm $\mathcal{A}$. System size of surrogate data L_{sg} is a user-definable hyperparameter. For high-dimensional $X(n \times L)$, $L_{\mathrm{sg}} \ll L$ is expected to lead to a stable behavior. The number of independent surrogate tests T can be decided using Monte Carlo principles; however, in our experiments values $T = 10$–100 have shown reasonable behavior when $G_{\mathrm{sg}}(f(r; L_{\mathrm{sg}}))$ is stable.

from the full size surrogate data, which gives

$$\bar{G}_{\mathrm{sg}}(f(1; L)) < l(f(1; L)). \tag{21}$$

Therefore, we choose $l(f(1; L))$ as a relaxed threshold. This process is termed as inverse finite-size scaling (IFSS) procedure for identifying significant estimates, as summarized in Fig. 8.

VI. APPLICATION TO HIGH-DIMENSIONAL GENOMIC DATA

We confirmed that IFSS also holds for two high-dimensional real genome data sets which contain large numbers of *Pneumococcus* genomes sampled from the Maela refugee camp in Thailand and the Massachusetts pediatric population in the USA [20]. Data dimensionalities are $n = 3042, L = 94028$ for Maela data and $n = 670, L = 78733$ for Massachusetts data, where $\alpha = 0.0324$ and $\alpha = 0.0085$, respectively. Each feature (allele) is categorized according to being major, mid-, minor, or gap in the data column (locus), and the full data set is then statistically characterized by the resulting empirical bias distribution. The major allele dimension of the bias distribution is shown in Fig. 9. One can see the data contain many column samples which are dominated by one type of an allele. These data have been recently analyzed in Ref. [20] using the SuperDCA algorithm, which implements a computationally efficient pseudolikelihood maximization algorithm for fitting ultra-high-dimensional Potts models with an order of $L = 10^5$ features. Standardized SuperDCA learned interaction parameters from different sizes of surrogate data generated from Maela data all collapse to the same distribution [Fig. 10(a)]. The red curve, which represents the average of 100 independent tests, is as expected more stable than each test curve individually. When the surrogate data size changes, the region representing the lower values, which has a high probability density, stays stable while the upper tails representing the rare events fluctuate around the red average data curve. Figure 11(a) shows that analogous results are obtained for the Massachusetts data.

Figure 10(c) demonstrates how the IFSS procedure efficiently determines a statistical significance threshold for high-dimensional genome data sampled in the Maela refugee camp. The main aim of DCA in this application context is to detect epistatic interactions or co-selection between mutations present in bacteria sampled densely from their host population. The small-scale surrogate data curve fits well with the low-ranked region of the coupling curve for the original data. Stronger signals diverge from the background distribution similarly to the pattern seen in the synthetic data tests. Using the IFSS procedure, we are able to filter out 4 420 387 378 coupling parameters from the total of 4 420 585 378 estimates, leaving only a small set of statistically significant interacting pairs of positions to be evaluated for potential biological significance. Figure 11(c) shows that the IFSS procedure also works well for the Massachusetts data.

Note that the surrogate data size we chose here is exceedingly small ($L_{\mathrm{sg}} = 500, n_{\mathrm{sg}} = 16$) compared to the original data ($L = 94\,028, n = 304\,216$ for Maela data), and

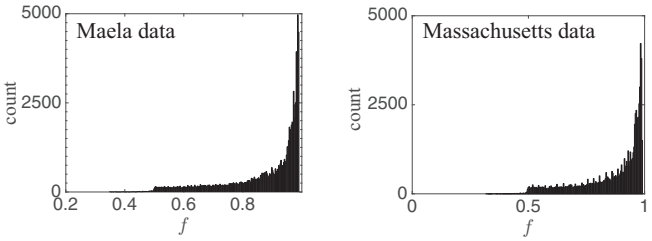

FIG. 9. Empirical bias distribution of Maela and Massachusetts data in the major allele dimension. In the x axis, f means the major allele frequency of one chosen column (locus) in data.

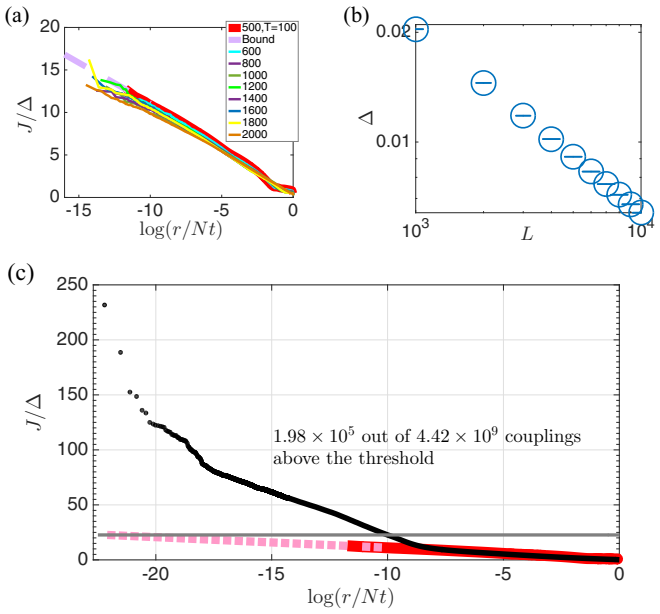

FIG. 10. (a) FSS property test on Maela data. (b) $\Delta - L$ relation of surrogate data. The log-log plot shows a clear linear dependence. Each mark represents a mean of 50 tests, where the horizontal short lines are error bars. Maela surrogate data are generated with system size varying across $L_{\mathrm{sg}} = 1000, 2000, \ldots, 10\,000$. The fitted gradient of the line is -0.5053. Regularization parameters λ_J, λ_h are set as 0.01. (c) IFSS procedure on Maela data. Solid red line: average of 100 simulations of small surrogate data with size $n_{\mathrm{sg}} = 16, L_{\mathrm{sg}} = 500$. Dashed pink line: fitted bound on the tail of the solid red line. By extending the dashed pink line to the same horizontal position with the top ranked coupling for real data, the value of the vertical axis at that point defines the significance threshold shown as the gray line.

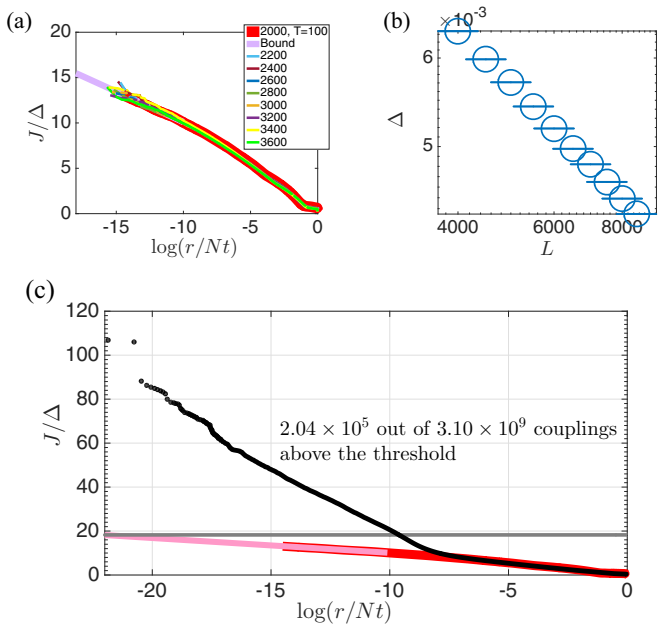

FIG. 11. (a) FSS property of Massachusetts surrogate data. (b) $\Delta - L$ relation of Massachusetts surrogate data. Massachusetts surrogate data results are shown for sizes ranging over $L_{\mathrm{sg}} = 4000, 4500, \ldots, 8500$. Each mark represents a mean of 50 tests, where the horizontal short lines are error bars. The fitted gradient of the line is -0.4993. (c) IFSS procedure on Massachusetts data. Solid red line: average of 100 simulations with small-scale surrogate data with size $n_{\mathrm{sg}} = 17, L_{\mathrm{sg}} = 2000$. For the meaning of markers and other lines see caption of Fig. 10.

consequently the SuperDCA runtime for the 100 simulation runs was negligible (a couple of minutes) in comparison to the time required to perform parameter inference on the full data (several days on a 20-core server). Thus, the IFSS procedure is very useful for practical applications as it enables rapid significance testing even when the original data are extremely high-dimensional and parameter inference is computationally demanding.

Furthermore, since the relaxed threshold $l(f(1; L))$ can be obtained independently from learning the high-dimensional data, one can also obtain a significance threshold estimate for the coupling values of real data from only their standard deviation. Figure 10(b) shows a log-log plot of the coupling standard deviation and data length of the Maela surrogate data, exhibiting a clear linear trend. By fitting a line $y = ax + b$ to the $(\log(\Delta), \log(L))$ data, one can estimate a functional relationship between Δ and L as in (19). This relation enables a prediction of the standard deviation of estimates for a large size surrogate data with the same dimensionality of the original data separately from the process of learning the original high-dimensional data. For Maela data, by fitting a line on the points displayed in Fig. 10(b), we obtain $a = -0.5053, b = -0.3960$, which gives a prediction $\Delta \approx 0.0021$ for surrogate data of equal size as the Maela data ($L = 94\,028$). For Massachusetts data, the fitted line on Fig. 11(b) has $a = -0.4993, b = -0.9175$, yielding $\Delta \approx 0.0014$ for surrogate data of equal size as the Massachusetts data ($L = 78\,733$). The true standard deviations of the original Maela and Massachusetts data are $\Delta_{\mathrm{maela}} \approx \Delta_{\mathrm{mass}} \approx 0.0004$, respectively. It is worth noting that the standard deviations of couplings learned for the original data and the surrogate data with the same dimensionality will not fully match in general, because correlations exist in the original data. Although the surrogate predictions overestimate the true values, they are nevertheless in the same ballpark and are therefore useful approximations, especially given that they can be relatively cheaply computed using a system size of less than 1% of the original problem. Such offline estimates allow one to discard couplings sufficiently smaller than the estimates even online during the learning, which significantly reduces the memory and storage complexity of the inference algorithm. Similarly, by updating the summary statistics of the estimated couplings in an online manner one could still retain the information about the true standard deviation for later use, such as for constructing a rank plot.

VII. DISCUSSION

Our results indicate that for high-dimensional dependence learning using Ising or Potts models using a regularized inference algorithm, the distribution of parameter estimates under a null hypothesis can after normalization by the standard deviation be characterized by a single function determined only by a limited number of system parameters: the data dimensionality ratio $\alpha = n/L$, and the data feature bias $\{\boldsymbol{f}_i\}$

distribution. The proposed inverse finite-size scaling procedure has a high potential for practical applications by enabling accurate prediction of large system behavior from simulations of small system sizes. Our proposed procedure is general and expected to be applicable to a wide range of dependence learning problems. The ability to rescale the given data properly to small sizes has the potential to spark further research of ways to dealing with high-dimensional data. In future work, it would be a valuable target for further research to identify general conditions under which the convergence behavior relating to the finite-size scaling is expected to hold.

ACKNOWLEDGMENTS

We thank Erik Aurell for insightful discussions, and acknowledge the Triton cluster at Aalto University for providing computational resources. We also acknowledge the Tokyo Institute of Technology for hosting Y.X. as a visiting researcher for three months and providing meeting space for the authors (Y.X., S.P., and J.C.). This research is supported by the COIN Centre of Excellence, Academy of Finland Grant No.251170 (Y.X., S.P., and J.C.), ERC Grant No. 742158 (J.C.), and JSPS KAKENHI Grants No. 25120013 and No. 17H00764 (Y.K.).

APPENDIX: SPARSE RESTRICTED BOLTZMANN MACHINE

A nonzero discrete-valued pairwise Markov random field (MRF) over the variables $\boldsymbol{x}$ can be represented by the marginal distribution of a restricted Boltzmann machine (RBM). We use a sparse restricted Boltzmann machine (sRBM) as a synthetic model, by which the data generation can be performed as follows: (1) define the number of nonzero edges n_E, (2) introduce an n_E dimensional vector $\boldsymbol{z}$, each element of which follows the standard Gaussian prior, and (3) prepare an $n_E \times L$ dimensional matrix $\boldsymbol{W} = (W_{lv})$, where W_{lv} represents the nonzero edge strength for edge lv in a bipartite graph in which nodes of two types correspond to elements of $\boldsymbol{z}$ and $\boldsymbol{x}$. Setting $\boldsymbol{W}$ so that W_{lv} are nonzero just for two indices of v for each l, one can introduce a small number of couplings for the original MRF in a controlled manner as $\boldsymbol{J} = \boldsymbol{W}^\top \boldsymbol{W}$.

More specifically, we set the joint distribution of binary sample $\boldsymbol{x} \in \{-1,+1\}^L$ and random vector $\boldsymbol{z} \in \mathbb{R}^{n_E}$ as

$$P(\boldsymbol{x},\boldsymbol{z}) = \frac{1}{Z}\exp\left(-\frac{|\boldsymbol{z}|^2}{2} + \boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{x} + \boldsymbol{h}\cdot\boldsymbol{x}\right),$$

where Z is a normalizing constant. Marginalizing $P(\boldsymbol{x},\boldsymbol{z})$ with respect to $\boldsymbol{z}$ yields

$$P(\boldsymbol{x}) = \int P(\boldsymbol{x},\boldsymbol{z})d\boldsymbol{z} \propto \exp\left(\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{J}\boldsymbol{x} + \boldsymbol{h}\cdot\boldsymbol{x}\right).$$

This indicates that RBM is reduced to a Boltzmann machine of the original MRF after the marginalization.

Sampling from RBM is very efficient by using the Gibbs sampler. Iterating

$$\boldsymbol{x} \sim P(\boldsymbol{x}|\boldsymbol{z}) = \prod_{i=1}^{L} \frac{1 + x_i \tanh\left(h_i + \sum_{l=1}^{n_E} W_{li} z_l\right)}{2},$$

$$\begin{aligned} \boldsymbol{z} \sim P(\boldsymbol{z}|\boldsymbol{x}) &= \frac{1}{Z}\exp\left(-\frac{|\boldsymbol{z}|^2}{2} + \boldsymbol{z}^\top W\boldsymbol{x}\right) \\ &= \prod_{l=1}^{n_E} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(z_l - \sum_{i=1}^{L} W_{li}x_i\right)^2\right], \end{aligned}$$

both of which can be easily parallelized, for a sufficient number of times yields a set of samples from RBM. Sampled values of $\boldsymbol{x}$ correspond to an observed data set from the Boltzmann machine. The possibility of using this kind of a block-Gibbs sampling algorithm is one of the main advantages of an RBM over a fully connected Boltzmann machine.

[1] M. J. Wainwright and M. I. Jordan, Found. Trends Machine Learn. **1**, 1 (2008).
[2] J. Theiler *et al.*, Physica D (Amsterdam) **58**, 77 (1992).
[3] P. Grassberger, Nature (London) **323**, 609 (1986).
[4] J. Lintusaari *et al.*, Syst. Biol. **66**, e66 (2017).
[5] C. C. Drovandi *et al.*, Stat. Sci. **30**, 72 (2015).
[6] E. Cameron and A. N. Pettitt, Mon. Not. R. Astron. Soc. **425**, 44 (2012).
[7] A. Weyant, C. Schafer, and W. M. Wood-Vasey, Astrophys. J. **764**, 116 (2013).
[8] E. Ishida *et al.*, Astron. Comput. **13**, 1 (2015).
[9] G. Louppe and K. Cranmer, arXiv:1707.07113.
[10] K. Cranmer, J. Pavez, and G. Louppe, arXiv:1506.02169.
[11] H. C. Nguyen, R. Zecchina, and J. Berg, Adv. Phys. **66**, 197 (2017).
[12] A. Montanari, Electron. J. Statist. **9**, 2370 (2015).
[13] M. Weigt *et al.*, Proc. Natl. Acad. Sci. USA **106**, 67 (2009).
[14] L. Burger and E. van Nimwegen, PLoS Comput. Biol. **6**, e1000633 (2010).
[15] F. Morcos *et al.*, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).
[16] M. Ekeberg *et al.*, Phys. Rev. E **87**, 012707 (2013).
[17] M. Ekeberg, T. Hartonen, and E. Aurell, J. Comput. Phys. **276**, 341 (2014).
[18] D. S. Marks, T. A. Hopf, and C. Sander, Nat. Biotechnol. **30**, 1072 (2012).
[19] M. Skwark *et al.*, PLos. Genet. **13**, e1006508 (2017).
[20] S. Puranen, M. Pesonen, J. Pensar, Y. Y. Xu, J. A. Lees, S. D. Bentley, N. J. Croucher, and J. Corander, Microb. Genom. (2018), doi: 10.1099/mgen.0.000184.
[21] M. Andreatta *et al.*, arXiv:1311.1301.
[22] Y. Xu *et al.*, arXiv:1704.01459.