Mean-field theory of Bayesian clustering

Alexander Mozeika[1] and Anthony C. C. Coolen[1,2]

[1]*Institute for Mathematical and Molecular Biomedicine, King's College London, Hodgkin Building, London SE1 1UL, United Kingdom*
[2]*Department of Mathematics, King's College London, The Strand, London WC2R 2LS, United Kingdom*

(Received 7 September 2017; revised manuscript received 20 January 2018; published 19 October 2018)

We show that model-based Bayesian clustering, the probabilistically most systematic approach to the partitioning of data, can be mapped into a statistical physics problem for a gas of particles and as a result becomes amenable to a detailed quantitative analysis. A central role in the resulting statistical physics framework is played by an entropy function. We demonstrate that there is a relevant parameter regime where mean-field analysis of this function is exact and that, under natural assumptions, the lowest entropy state of the hypothetical gas corresponds to the optimal clustering of data. The by-product of our analysis is a simple but effective clustering algorithm, which infers both the most plausible number of clusters in the data and the corresponding partitions. Describing Bayesian clustering in statistical mechanical terms is found to be natural and surprisingly effective.

DOI: 10.1103/PhysRevE.98.042133

I. INTRODUCTION

The need for clustering analysis in scientific data exploration has grown significantly in recent years, due to the emergence of large high-dimensional data sets in areas such as high-energy physics, astrophysics, biology, and postgenome medicine. The aim of clustering analysis is to allocate similar data items, such as stars [1], galaxies [2], bacterial communities [3], or amino acid sequences [4], to the same category (or "cluster") in an unsupervised way. Inferring the true number of clusters reliably is crucial for the discovery of new data categories. Most current clustering methods, such as Refs. [5–7], make no assumptions about the data distribution and are based on heuristic measures of similarity. Some allow for estimation of the number of clusters but use empirical approaches to do so and ad hoc evaluation criteria tested on benchmark data sets.

Model-based clustering assumes that each data point comes from one of a postulated number of populations with known distributions. The archetypal example is the Gaussian Mixture Model (GMM) [5], which assumes Gaussian distributions. In such models Maximum likelihood (ML) inference is typically used to find data partitions [8], but this is prone to overfitting [5]. The number of clusters K is found on adding a "penalty" term to the log-likelihood function, such as Akaike's Information Criterion (AIC) or Bayesian information criterion (BIC) [8], sometimes with conflicting results [2]. Bayesian inference of GMM-generated data cures overfitting and provides a systematic way to find K [5]. However, computing the posteriors is analytically intractable, and one tends to resort to either variational mean-field approximation [5] or computationally intensive Markov chain Monte Carlo (MCMC) [9].

A more general model-based Bayesian clustering protocol (SPD) was introduced in Ref. [10]. Unlike GMM, it uses priors on the partitions to compute a maximum *a posteriori* probability (MAP) estimate of the data partitioning. Both SPD and GMM Bayesian methods are usually evaluated by clustering synthetic and benchmark real-world data. This is not satisfactory; one would prefer our knowledge and our confidence in clustering outcomes to be based on more than empirical tests.

As a first step in this direction, in this paper we use statistical physics to study model-based Bayesian clustering. This strategy was used in the past to study optimization problems, see, e.g., Ref. [11], and clustering [12,13], but not Bayesian clustering. Starting from the SPD model, we show that data partition inference can be formulated in terms of a quantity that can be seen as the entropy of a gas of a particles (data points), distributed over K reservoirs (clusters). In the regime of a large number of particles we derive a mean-field theory to describe this gas and show that its lowest entropy state corresponds to the optimal MAP clustering of data.

II. MODEL OF DATA AND BAYESIAN CLUSTERING

Let us assume that we observe the sample $\mathbf{X} = \{\mathbf{x}_1,\dots,\mathbf{x}_N\}$, with $\mathbf{x}_i \in \mathbb{R}^d$ for all i, from the distribution

$$p(\mathbf{X}|\boldsymbol{\Theta},\Pi)=\prod_{\mu=1}^{|\Pi|}\prod_{i_\mu\in S_\mu}p\big(\mathbf{x}_{i_\mu}\big|\boldsymbol{\theta}_\mu\big). \tag{1}$$

This distribution is generated by the set (or "partition") $\Pi = \{S_1,\dots,S_{|\Pi|}\}$, with disjunct index sets (or "clusters") $S_\mu \neq \emptyset$, such that $S_\mu \cap S_\nu = \emptyset$ for $\mu \neq \nu$ and $\cup_{\mu=1}^{|\Pi|} S_\mu = [N]$ with $[N] = \{1,\dots,N\}$. Any partition of data into K clusters can be specified by binary "cluster allocation" variables $c_{i\mu} = \mathbb{1}[i \in S_\mu]$, where $i \in [N]$ and $\mu \in [K]$, forming an $N \times K$ partitioning matrix $\mathbf{c}$. This matrix satisfies by construction the following constraints: $\sum_{\mu=1}^{K} c_{i\mu} = 1$ for all $i \in [N]$ and $\sum_{i=1}^{N} c_{i\mu} \geqslant 1$ for all $\mu \in [K]$. Conversely, any $N \times K$ matrix $\mathbf{c} \in \{0,1\}^{NK}$ with binary entries that satisfies these constraints induces a partition $\Pi(\mathbf{c}) = \{S_1(\mathbf{c}),\dots,S_K(\mathbf{c})\}$ of cardinality K. If we also know the prior distributions of model parameters, $p(\boldsymbol{\theta}_\mu)$, $p(\mathbf{c}|K)$, and $p(K)$, then we can use Bayes's theorem (see Appendix A for details) to derive the posterior distribution

$$p(\mathbf{c},K|\mathbf{X}) = \frac{e^{-N\hat{F}_N(\mathbf{c},\mathbf{X})}p(\mathbf{c}|K)p(K)}{\sum_{\tilde{K}=1}^{N}p(\tilde{K})\sum_{\tilde{\mathbf{c}}}e^{-N\hat{F}_N(\tilde{\mathbf{c}},\mathbf{X})}p(\tilde{\mathbf{c}}|\tilde{K})}, \qquad (2)$$

in which

$$\hat{F}_N(\mathbf{c},\,\mathbf{X}) = -\frac{1}{N}\log\langle e^{\sum_{\mu=1}^{K}\sum_{i=1}^{N}c_{i\mu}\log p(\mathbf{x}_i|\boldsymbol{\theta}_\mu)}\rangle_{\boldsymbol{\Theta}}, \qquad (3)$$

with $\langle f(\boldsymbol{\Theta})\rangle_{\boldsymbol{\Theta}} = \int[\prod_{\mu=1}^{K}p(\boldsymbol{\theta}_\mu)\,d\boldsymbol{\theta}_\mu]f(\boldsymbol{\Theta})$. Expression (2) can be used to infer the most probable partition Π for each data sample. First, for each $K\in[N]$ one computes

$$\hat{\mathbf{c}}\,|K = \mathrm{argmax}_{\mathbf{c}}\{e^{-N\hat{F}_N(\mathbf{c},\mathbf{X})}p(\mathbf{c}|K)\}. \qquad (4)$$

Then one uses (4) to determine the estimate $\hat{\Pi}$ of Π:

$$\hat{\Pi} = \mathrm{argmax}_{\hat{\mathbf{c}}\,|K}\{e^{-N\hat{F}_N(\hat{\mathbf{c}},\mathbf{X})}p(\hat{\mathbf{c}}|K)p(K)\}. \qquad (5)$$

Clearly, a key role in our formulas is played by the function (3), which can be seen as an entropy of a gas of N "particles" (the data points) distributed over K "reservoirs" (clusters). The particles can move from one reservoir to another; $c_{i\mu}$ tells us if particle i is in reservoir μ, and the coordinates $\mathbf{x}_i$ act as a "quenched" disorder [14]. We are then interested in the minimum entropy state $\mathrm{argmin}_{\mathbf{c}}\hat{F}_N(\mathbf{c},\,\mathbf{X})$.

III. MEAN-FIELD ANALYSIS OF BAYESIAN CLUSTERING

Let us first consider the case where the cluster parameters are known. In this case the parameter prior $p(\boldsymbol{\theta}_\mu)$ is a δ function, and (3) hence becomes

$$\hat{F}_N(\mathbf{c},\mathbf{X}) = -\sum_{\mu=1}^{K}\frac{M_\mu(\mathbf{c})}{N}\int d\mathbf{x}\;\hat{Q}_\mu(\mathbf{x}|\mathbf{c},\mathbf{X})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu), \qquad (6)$$

which is now written in terms of the number of particles in cluster μ, $M_\mu(\mathbf{c}) = \sum_{i=1}^{N}c_{i\mu}$, and the density of particles in cluster μ, defined as

$$\hat{Q}_\mu(\mathbf{x}|\mathbf{c},\,\mathbf{X}) = \frac{1}{M_\mu(\mathbf{c})}\sum_{i=1}^{N}c_{i\mu}\delta(\mathbf{x}-\mathbf{x}_i). \qquad (7)$$

Suppose there are L distributions $q_\nu(\mathbf{x})$, such that for each ν we find N_ν particles with $\mathbf{x}_i$ sampled from $q_\nu(\mathbf{x})$, with $\sum_{\nu=1}^{L}N_\nu = N$ and $\lim_{N\to\infty}N_\nu/N = \gamma(\nu)$. For large N the density (7) will then typically converge to

$$Q_\mu(\mathbf{x}) = \sum_{\nu=1}^{L}\alpha(\nu|\mu)\,q_\nu(\mathbf{x}). \qquad (8)$$

Here $\alpha(\nu|\mu) = \alpha(\nu,\mu)/\alpha(\mu)$ is a conditional probability, defined by $\alpha(\mu) = \lim_{N\to\infty}M_\mu(\mathbf{c})/N$ and $\alpha(\nu,\mu) = \lim_{N\to\infty}M_{\nu,\mu}(\mathbf{c})/N$, where $M_\mu(\mathbf{c})$ is the number of particles in cluster μ and $M_{\nu,\mu}(\mathbf{c}) = \sum_{i_\nu\in S_\mu(\mathbf{c})}\mathbb{1}[\mathbf{x}_{i_\nu}\sim q_\nu(\mathbf{x})]$ is the number of those particles drawn from the distribution $q_\nu(\mathbf{x})$ that are allocated by $\mathbf{c}$ to cluster μ. Clearly, $\sum_{\mu\leqslant K}\alpha(\nu,\mu) = \gamma(\nu)$, $\sum_{\nu\leqslant L}\alpha(\nu,\mu) = \alpha(\mu) > 0$ and $\sum_{\nu\leqslant L}\sum_{\mu\leqslant K}\alpha(\nu,\mu) = 1$. If (8) holds for $N\to\infty$, then $\hat{F}_N(\mathbf{c},\,\mathbf{X})$ will for $N\to\infty$ converge to

$$F(\boldsymbol{\alpha}) = \sum_{\mu=1}^{K}\sum_{\nu=1}^{L}\alpha(\nu,\mu)D(q_\nu\|p_\mu) + \sum_{\nu=1}^{L}\gamma(\nu)H(q_\nu). \qquad (9)$$

Here $D(q_\nu\|p_\mu)$ is the Kullback-Leibler distance between $q_\nu(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\theta}_\mu)$, and $H(q_\nu)$ is a differential entropy [15]. The transparent and intuitive result (9) can be seen as a mean-field (MF) theory of $\hat{F}_N(\mathbf{c},\,\mathbf{X})$ (see Appendix C for details). The $L\times K$ matrix $\boldsymbol{\alpha}$, with entries $\alpha(\nu,\mu)$, acts as order parameter. More generally one would have $P(F) = \int d\boldsymbol{\alpha}\;P(\boldsymbol{\alpha})\,\delta[F-F(\boldsymbol{\alpha})]$, where

$$P(\boldsymbol{\alpha}) = \lim_{N\to\infty}\sum_{\mathbf{c},\tilde{\mathbf{c}}}p(\mathbf{c}|K)\,q(\tilde{\mathbf{c}}|L)\prod_{\mu=1}^{K}\prod_{\nu=1}^{L}\delta\left[\alpha(\nu,\mu) - \frac{1}{N}\sum_{i=1}^{N}\tilde{c}_{i\nu}c_{i\mu}\right]. \qquad (10)$$

Here $p(\mathbf{c}|K)$ and $q(\tilde{\mathbf{c}}|L)$ are the assumed and the "true" distributions of partitions, respectively. We can limit ourselves to working with expression (9), as opposed to the more involved (10), if $P(\boldsymbol{\alpha})$ is a δ function.

We are interested in the state $\boldsymbol{\alpha}$ for which the function $F(\boldsymbol{\alpha})$ is minimal. First, from $D(q_\nu\|p_\mu)\geqslant 0$ it follows that $F(\boldsymbol{\alpha})\geqslant\sum_{\nu=1}^{L}\gamma(\nu)H(q_\nu)$. The lower bound is saturated when $D(q_\nu\|p_\mu) = 0$, i.e., when $q_\nu(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}_\mu)$ for all (μ,ν), and the mapping between the sets $[L]$ and $[K]$ labeling these distributions is bijective. This can only happen when $L = K$ and $\alpha(\nu,\mu) = \gamma(\nu)\mathbb{1}[D(q_\nu\|p_\mu) = 0]$, i.e., when the "true" partitioning of the data is recovered.

Second, from $D(q_\nu\|p_\mu)\geqslant\min_{\tilde{\mu}}D(q_\nu\|p_{\tilde{\mu}})$ we deduce

$$F(\boldsymbol{\alpha})\geqslant\sum_{\nu=1}^{L}\gamma(\nu)\min_{\tilde{\mu}}D(q_\nu\|p_{\tilde{\mu}}) + \sum_{\nu=1}^{L}\gamma(\nu)H(q_\nu). \qquad (11)$$

This lower bound is saturated when $\alpha(\nu,\mu) = \gamma(\nu)\mathbb{1}[\mu = \mathrm{argmin}_{\tilde{\mu}}D(q_\nu\|p_{\tilde{\mu}})]$ for all (μ,ν). For $K\leqslant L$, this state can be seen as the result of the following "macroscopic" clustering protocol: For all $\nu\in\{1,\dots,L\}$, find the distribution $p(\mathbf{x}|\boldsymbol{\theta}_\mu)$ with the smallest distance $D(q_\nu\|p_\mu)$ to $q_\nu(\mathbf{x})$, and assign all members of ν to cluster μ. If $K<L$, then this recipe will occasionally result in the data from more than one distribution being assigned to the same clusters, see Fig. 1(a), but for $K = L$, each cluster would hold only one distribution. Hence, the protocol is able to recover the true partitioning even when the distributions $q_\nu(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\theta}_\mu)$ are nonidentical.

The inequality $D(q_\nu\|p_\mu)\geqslant\min_{\tilde{\nu}}D(q_{\tilde{\nu}}\|p_\mu)$ gives the lower bound $F(\boldsymbol{\alpha})\geqslant\sum_{\mu=1}^{K}\alpha(\mu)\min_{\tilde{\nu}}D(q_{\tilde{\nu}}\|p_\mu) + \sum_{\nu=1}^{L}\gamma(\nu)H(q_\nu)$, which is saturated when $\alpha(\nu,\mu) = \alpha(\mu)\mathbb{1}[\nu = \mathrm{argmin}_{\tilde{\nu}}D(q_{\tilde{\nu}}\|p_\mu)]$ for all (μ,ν). This state would result from to the following protocol: For all $\nu\in\{1,\dots,L\}$, find the distribution $p(\mathbf{x}|\boldsymbol{\theta}_\mu)$ with the smallest distance $D(q_\nu\|p_\mu)$ to $q_\nu(\mathbf{x})$, and assign all members of μ to cluster ν. For $K>L$, this algorithm could allocate more than one distribution to the same cluster, see Fig. 1(b). Furthermore, since $\sum_{\nu=1}^{L}\alpha(\mu)\mathbb{1}[\nu = \mathrm{argmin}_{\tilde{\nu}}D(q_{\tilde{\nu}}\|p_\mu)] = \alpha(\mu)$, the properties of $\alpha(\nu,\mu)$ imply validity of the set of L linear equations $\sum_{\mu=1}^{K}\alpha(\mu)\mathbb{1}[\nu = \mathrm{argmin}_{\tilde{\nu}}D(q_{\tilde{\nu}}\|p_\mu)] = \gamma(\nu)$, which is underdetermined and hence has either infinitely many solutions or no solutions at all.

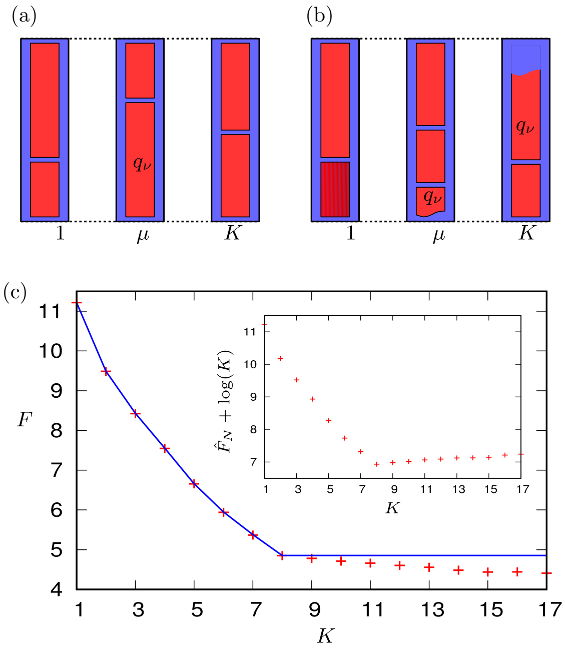

FIG. 1. Bayesian clustering: Data (red rectangles) from L different distributions $q_\nu(\mathbf{x})$ are allocated to K clusters (blue rectangles). (a) For $K \leqslant L$, data from $q_\nu(\mathbf{x})$ occupy *at most* one cluster μ. (b) For $K > L$, data from $q_\nu(\mathbf{x})$ occupy *at least* one cluster. (c) Minimum $F \equiv \min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$ of the mean-field entropy (blue line), shown as a function of K and compared with the ground-state entropy $\hat{F}_N \equiv \min_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$ (red crosses), computed for the data of Fig. 2. The horizontal line corresponds to the lower bound $\sum_{\nu \leqslant L} \gamma(\nu) H(q_\nu) = 4.853905$. Inset: The sum of $\hat{F}_N$ and $\log(K)$, where $\log(K) \equiv \log_e(K)$ shown as a function of K. The minimum of this sum is obtained when $K = L$.

We now consider the case where the cluster parameters are unknown, and $p(\boldsymbol{\theta}_\mu) > 0$ for all $\{\boldsymbol{\theta}_\mu\}$. For $N \to \infty$, the entropy (3) is now strictly dominated via steepest descent by the following set of saddle point equations (see Appendix B for details):

$$\frac{\partial}{\partial \theta_\mu(\ell)} \frac{1}{N} \sum_{i=1}^{N} c_{i\mu} \log p(\mathbf{x}_i|\boldsymbol{\theta}_\mu) = 0. \tag{12}$$

Solving (12) for Gaussian distributions $p(\mathbf{x}_i|\boldsymbol{\theta}_\mu) \equiv \mathcal{N}(\mathbf{x}|\mathbf{m}_\mu, \boldsymbol{\Lambda}_\mu^{-1})$, with mean $\mathbf{m}_\mu$ and inverse covariance matrix $\boldsymbol{\Lambda}_\mu$, gives us (see Appendix B):

$$\hat{F}_N(\mathbf{c}, \mathbf{X}) = \sum_{\mu=1}^{K} \frac{M_\mu(\mathbf{c})}{2N} \log\left[(2\pi e)^d \left|\boldsymbol{\Lambda}_\mu^{-1}(\mathbf{c}, \mathbf{X})\right|\right], \tag{13}$$

where $\boldsymbol{\Lambda}_\mu^{-1}(\mathbf{c}, \mathbf{X})$ is the empirical covariance matrix of the data in cluster μ.

Since (13) represents an average of K entropies of Gaussian distributions, which for $N \to \infty$ will converge to the following mean-field entropy (see Appendix C):

$$F(\boldsymbol{\alpha}) = \sum_{\mu=1}^{K} \alpha(\mu) \frac{1}{2} \log\left[(2\pi e)^d \left|\boldsymbol{\Lambda}_\mu^{-1}(\boldsymbol{\alpha})\right|\right], \tag{14}$$

in which $\boldsymbol{\Lambda}_\mu^{-1}(\boldsymbol{\alpha})$ denotes the covariance matrix

$$\boldsymbol{\Lambda}_\mu^{-1}(\boldsymbol{\alpha}) = \sum_{\nu=1}^{L} \alpha(\nu|\mu) \langle [\mathbf{x} - \mathbf{m}_\mu(\boldsymbol{\alpha})][\mathbf{x} - \mathbf{m}_\mu(\boldsymbol{\alpha})]^T \rangle_\nu, \tag{15}$$

with $\mathbf{m}_\mu(\boldsymbol{\alpha}) = \sum_{\nu=1}^{L} \alpha(\nu|\mu) \langle \mathbf{x} \rangle_\nu$, and the short-hand $\langle \{\cdots\} \rangle_\nu = \int d\mathbf{x}\, q_\nu(\mathbf{x}) \{\cdots\}$. Note that (14) also equals

$$F(\boldsymbol{\alpha}) = \sum_{\mu,\nu} \alpha(\nu, \mu) D[q_\nu \| \mathcal{N}_\mu(\boldsymbol{\alpha})] + \sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu), \tag{16}$$

where $\mathcal{N}_\mu(\boldsymbol{\alpha}) \equiv \mathcal{N}(\mathbf{x}|\mathbf{m}_\mu(\boldsymbol{\alpha}), \boldsymbol{\Lambda}_\mu^{-1}(\boldsymbol{\alpha}))$. Moreover, as shown in Appendix D,

$$F(\boldsymbol{\alpha}) \geqslant \sum_{\mu=1}^{K} \alpha(\mu) H(Q_\mu) \geqslant \sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu). \tag{17}$$

The second inequality in (17) has two consequences. First, if $K \leqslant L$, then for any state $\boldsymbol{\alpha}$ that corresponds to either of the scenarios depicted in Figs. 1(a) and 1(b), we will have $F(\boldsymbol{\alpha}) \geqslant \min_K \min_{\tilde{\boldsymbol{\alpha}}} F(\tilde{\boldsymbol{\alpha}}) = \sum_{\nu \leqslant L} \gamma(\nu) H(q_\nu)$. The lower bound is satisfied when $L = K$ and $q_\nu(\mathbf{x})$ is Gaussian. The "true" parameters $\boldsymbol{\alpha}$ thus represent a locally stable state. Second, when $K > L$, the entropy $F(\boldsymbol{\alpha})$ can only increase with L. This follows from (16) and $D(q_\nu \| \mathcal{N}_\mu(\boldsymbol{\alpha})) \geqslant 0$. If $q_\nu(\mathbf{x})$ is not Gaussian, then $F(\boldsymbol{\alpha}) \geqslant \sum_{\nu=1}^{L} \gamma(\nu) \frac{1}{2} \log[(2\pi e)^d |\mathbf{c}_\nu|]$, where $\mathbf{c}_\nu$ is the covariance matrix of $q_\nu(\mathbf{x})$ (see Appendix D). Equality corresponds to the state shown in the Fig. 1(a) with $L = K$, i.e., here the "true" data partitioning is recovered.

The first inequality in (17) has an appealing geometric interpretation. The entropy $H(Q_\mu)$ of each cluster μ can for large N be estimated by $[d/M_\mu(\mathbf{c})] \sum_{i=1}^{N} c_{i\mu} \log \rho_{i\mu}(\mathbf{c}) + \log [M_\mu(\mathbf{c}) - 1] + \text{const}$, where $\rho_{i\mu}(\mathbf{c}) = \min_{i \in S_\mu(\mathbf{c}) \setminus i} \|\mathbf{x}_i - \mathbf{x}_j\|$ (i.e., the Euclidean distance between particle i and its nearest neighbor) [16]. The average entropy $\sum_{\mu=1}^{K} \alpha(\mu) H(Q_\mu)$ is hence estimated by $(d/N) \sum_{\mu=1}^{K} \sum_{i=1}^{N} c_{i\mu} \log \rho_{i\mu}(\mathbf{c}) + \sum_{\mu=1}^{K} [M_\mu(\mathbf{c})/N] \log [M_\mu(\mathbf{c})/N] + \text{const}$. This is minimized by any state $\mathbf{c}$ which simultaneously maximizes the entropy $-\sum_{\mu=1}^{K} [M_\mu(\mathbf{c})/N] \log [M_\mu(\mathbf{c})/N]$, i.e., "disperses" particles maximally over clusters, and minimizes the nearest neighbor distances $\{\rho_{i\mu}(\mathbf{c})\}$, i.e., favors high particle "densities" in each cluster.

The lower bound $\sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu)$ in (17) is saturated on choosing any bijective map $\alpha : \nu \to \mu$, since this immediately gives us $F(\boldsymbol{\alpha}) = \sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu)$. Such maps are special instances of the more general family

$$\alpha(\nu|\mu) = \frac{\mathbb{1}[\nu \in S_\mu] \gamma(\nu)}{\sum_{\tilde{\nu} \in S_\mu} \gamma(\tilde{\nu})}, \tag{18}$$

where $\Pi = \{S_1, \ldots, S_K\}$ is any partitioning of $[L]$ into K subsets. Finding $\min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$ over all possible matrices of the form (18) by enumeration of all partitions of $[L]$ into K subsets is feasible only for small L, since the number of such partitions is given by the Stirling number of the second kind $\mathcal{S}(L, K)$ which grows as K^L for large L [17].

One can also compute $\min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$ via the following "greedy" algorithm. Start with any partition Π and compute $F(\boldsymbol{\alpha})$. For all $x \in [L]$: Consider all possible moves which do

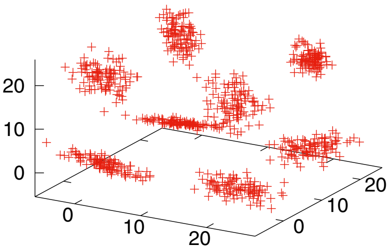

FIG. 2. Data used in our numerical experiments. We generated $L = 8$ clusters with 1000 data points each, of which 100 are shown here. The data in each cluster (i, j, k) are generated from a distinct Gaussian distribution, with mean $(\Delta i, \Delta j, \Delta k)$, where $i, j, k \in \{0, 1\}$ $(\Delta = 20)$, and with covariance matrix sampled from the Wishart distribution with 4 degrees of freedom and precision matrix **1**.

not create empty clusters, and execute the one which gives the largest decrease in $F(\boldsymbol{\alpha})$ and then update Π. Continue the last two steps until convergence of $F(\boldsymbol{\alpha})$ is observed. This macroscopic algorithm can also be implemented "microscopically." At each step: For all $i \in [N]$, consider all possible moves of the particle i from its current cluster $S_\mu(\mathbf{c})$ to a new cluster $S_\nu(\mathbf{c})$ and select the one which reduces $\hat{F}_N(\mathbf{c}, \mathbf{X})$ most. To evolve from a nonordered state as in Fig. 1(b) to an "ordered" state as in Fig. 1(a), this microscopic algorithm has to move on average at least $N(K-1)/K$ particles (see Appendix E). Each move was selected from among $N(K-1)$ possible moves, so the numerical complexity is at least of order $N^2(K-1)^2/K$.

IV. RESULTS OF NUMERICAL EXPERIMENTS

Our mean-field theory for was derived under the assumption that $\hat{F}_N(\mathbf{c}, \mathbf{X})$ is self-averaging for $N \to \infty$. To investigate the correctness of its predictions for finite sample sizes N, we studied low entropy states of (13) as obtained by the gradient descent algorithm on the data of the Fig. 2. For each $K \in [17]$ we ran the algorithm from 100 different random initial states $\mathbf{c}\,(0)$, and computed $\hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$ and the mean-field entropy $F(\boldsymbol{\alpha})$ (14) for each.

For $K \leqslant L$, most final states $\mathbf{c}\,(\infty)$ allocate data from the same distribution correctly to the same cluster, see Fig. 1(a). The values of $\hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$ are those predicted by $F(\boldsymbol{\alpha})$, and indeed correspond to local minima and saddle points of $F(\boldsymbol{\alpha})$ (see Appendix F). Also, according to Fig. 1(c), the value $\hat{F}_N = \min_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$ as estimated from $\mathbf{c}\,(\infty)$ is predicted accurately by $F = \min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$. Residual differences between $\hat{F}_N$ and F reflect finite-size effects. These can be computed exactly when $K = L$, and when $\mathbf{c}\,(\infty)$ represents the true partitioning of the data: The average and variance of $\hat{F}_N$ are in that case given by $\sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu) + Kd(d+1)/4N$ and $d/2N$, respectively (see Appendix G). Finally, we note that the number of particles "moved" by the algorithm in going from $\mathbf{c}\,(0)$ to $\mathbf{c}(\infty)$ is consistent with the lower bound $N(K-1)/K$, so the algorithmic complexity is quadratic in N; see Fig. 3.

If $K > L$, then the states $\mathbf{c}\,(\infty)$ will allocate data from the same distribution to multiple clusters; see Fig. 1(b). Such states are already present for small $K \leqslant L$, and proliferate as K is increased (see Appendix F). The lower bound $\sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu)$ is now violated, and the gap between this bound and the value of $\hat{F}_N$ as obtained by gradient descent increases with K; see Fig. 1(c). While some of the $\hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$ values are consistent with $F(\boldsymbol{\alpha})$ (see Appendix F), the mean-field theory fails to predict $\min_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$ in this regime, due to the noncommutation of the $N \to \infty$ limit and the min operator.

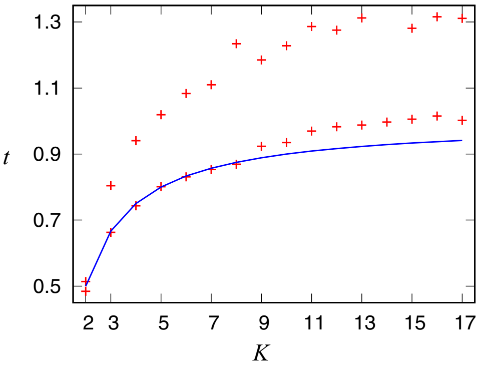

FIG. 3. Total (normalized) number of "moves" t used by the gradient descent algorithm to travel from a random unbiased partition to a final partition, i.e., the effective algorithmic runtime, shown as a function of the assumed number of clusters K. The minimum and maximum time (red crosses) obtained in 100 runs on the data of Fig. 2 are compared with the average lower bound $(K-1)/K$ (blue line).

Our estimate of $\hat{F}_N = \min_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$ can also be used to infer the true number of clusters L. Assuming uniform prior distributions of partitions $p(\mathbf{c}|K) = [K!\mathcal{S}(L, K)]^{-1}$ and cluster sizes $p(K) = N^{-1}\mathbb{1}[K \in [N]]$ in the Bayesian formulas (2)–(5), the total entropy $\hat{F}_N + \frac{1}{N}\log[K!\mathcal{S}(L, K)] \approx \hat{F}_N + \log(K)$ has its minimum at the correct value $K = L$; see inset in Fig. 1(c).

An interesting and important question, from a practical and a theoretical point view, is how Bayesian clustering is affected by the "separation" between different clusters. The simplest nontrivial case is to consider the clustering of d-dimensional data sampled from two isotropic Gaussian distributions $\mathcal{N}(\mathbf{m}_1, \mathbf{1})$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{1})$. Here one can use the Euclidean distance $\|\mathbf{m}_1 - \mathbf{m}_2\| = \Delta$, measured relative to the natural scale $\sqrt{d}$, as a measure of the degree of separation [18] between the "clusters" centered at $\mathbf{m}_1$ and $\mathbf{m}_2$. For large d, most of the vectors $\mathbf{x}$ sampled from $\mathcal{N}(\mathbf{m}, \mathbf{1})$ will be found in the "sphere" of radius $\sqrt{d}$ centered at $\mathbf{m}$, reflecting "concentration" phenomena observed for large d. In particular if we assume that $\mathbf{x}$ is sampled from $\mathcal{N}(\mathbf{m}, \boldsymbol{\Lambda})$, then $\langle \|\mathbf{x} - \mathbf{m}\|^2 \rangle = \mathrm{Tr}\,\boldsymbol{\Lambda}$, and for $\lambda, \epsilon > 0$:

$$\begin{aligned}
&\mathrm{Prob}(\|\mathbf{x}-\mathbf{m}\|^2 \geqslant \mathrm{Tr}\,\boldsymbol{\Lambda} + d\epsilon)\\
&\quad = \mathrm{Prob}[e^{\frac{\lambda}{2}\|\mathbf{x}-\mathbf{m}\|^2} \geqslant e^{\frac{\lambda}{2}(\mathrm{Tr}\,\boldsymbol{\Lambda}+d\epsilon)}]\\
&\quad \leqslant \langle e^{\frac{\lambda}{2}\|\mathbf{x}-\mathbf{m}\|^2}\rangle e^{-\frac{\lambda}{2}(\mathrm{Tr}\,\boldsymbol{\Lambda}+d\epsilon)}\\
&\quad = e^{-\frac{1}{2}[\log|\mathbf{1}-\lambda\boldsymbol{\Lambda}|+\lambda(\mathrm{Tr}\,\boldsymbol{\Lambda}+d\epsilon)]}.
\end{aligned} \tag{19}$$

The upper bound in the above expression was obtained using Markov's inequality and properties of Gaussian integrals. For the choice $\boldsymbol{\Lambda} = \mathbf{1}$, the above inequality, after

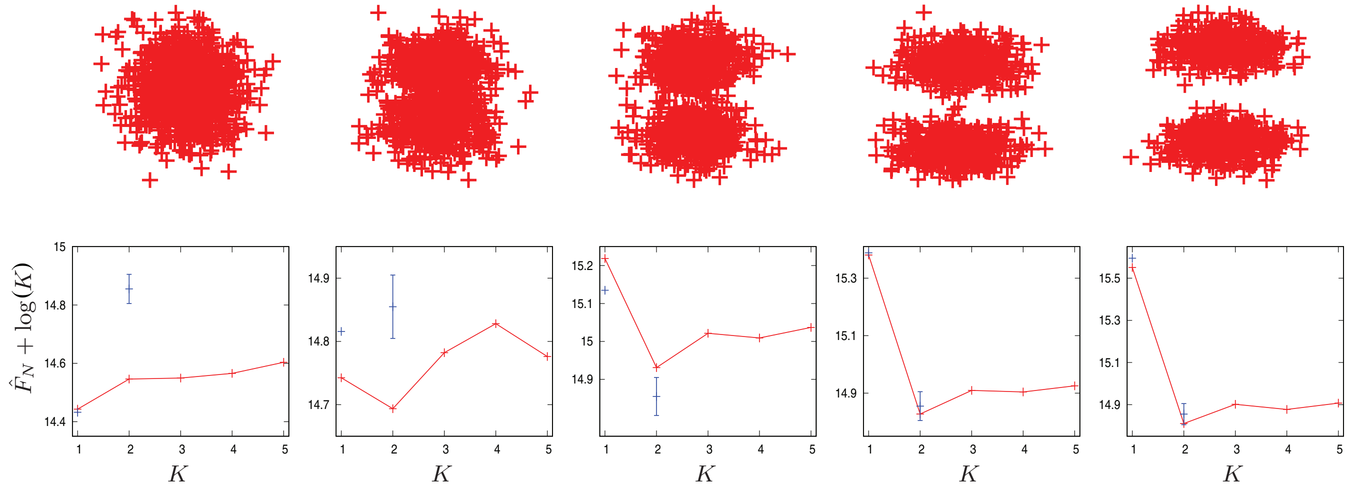

FIG. 4. Bayesian clustering of data $\mathbf{x} \in \mathbb{R}^d$ generated from the Gaussian distributions $\mathcal{N}(\mathbf{m}_1, \mathbf{1})$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{1})$, with separation $\Delta = \|\mathbf{m}_1 - \mathbf{m}_2\|$. The data sample, split equally between the constituent distributions, is of size $N = 2000$ and has $d = 10$. The data were generated for cluster separations $\Delta/\sqrt{d} \in \{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}\}$. Top: Data projected into two dimensions. The separation Δ of the clusters is increasing from the left to the right. Bottom: The sum $\hat{F}_N + \log(K)$ (red crosses connected by lines), where $\hat{F}_N \equiv \min_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$ and $\log(K) \equiv \log_e(K)$, shown as a function of the assumed number of clusters K, and compared with the mean-field prediction $\min_\alpha F(\boldsymbol{\alpha})$ (blue crosses). For $K = 2$, the mean-field prediction $\min_\alpha F(\boldsymbol{\alpha}) = \frac{d}{2}\log(2\pi e)$ is plotted with the finite-size corrections (error bars indicate one standard deviation).

optimizing the upper bound with respect to λ, gives us $\text{Prob}[\|\mathbf{x} - \mathbf{m}\|^2 \geqslant d(1+\epsilon)] \leqslant e^{-\frac{d}{2}(\log\frac{1}{1+\epsilon} - \epsilon)}$.

Let us now consider the MF entropy $\min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$ for the distributions $\mathcal{N}(\mathbf{m}_1, \mathbf{1})$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{1})$, with separation $\|\mathbf{m}_1 - \mathbf{m}_2\| = \Delta$. For the assumed number of clusters $K = 1$ this entropy is given by

$$F_1 = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\left|\mathbf{1} + \sum_{\nu=1}^{2}\gamma(\nu)(\mathbf{m}_\nu - \mathbf{m})(\mathbf{m}_\nu - \mathbf{m})^T\right|, \tag{20}$$

where $\mathbf{m} = \sum_{\nu=1}^{2}\gamma(\nu)\mathbf{m}_\nu$ and $\gamma(\nu)$ is the fraction of data sampled from $\mathcal{N}(\mathbf{m}_\nu, \mathbf{1})$. For $K = 2$ we obtain

$$F_2 = \frac{d}{2}\log{(2\pi e)}, \tag{21}$$

which corresponds to the situation where the true clustering of data is recovered. Furthermore, on choosing $\mathbf{m}_1 = \mathbf{0}$ and $\gamma(\nu) = \frac{1}{2}$ we obtain $F_1 = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log[1 + (\frac{\Delta}{2})^2]$. Thus in this case $F_1 \geqslant F_2$, as required. However, if $\log(2) \geqslant \frac{1}{2}\log[1 + (\frac{\Delta}{2})^2]$, then $F_2 + \log(K) \geqslant F_1$ (note that we minimize $\min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}) + \log(K)$ to infer true number of clusters), so that here we are unable to recover the correct number $K = 2$ of clusters due to the cluster separation Δ being too small. This happens when $\Delta \leqslant 2\sqrt{3} \approx 3.46$. We expect that a similar analysis can be also performed for more general scenarios.

Numerical experiments are in qualitative agreement with the predicted separation boundary $\Delta = 2\sqrt{3}$, as can be seen in Fig. 4. In this figure we also compare the mean-field theory results [(20) and (21)] with the results of numerical simulations. For $K = 1$ the discrepancy between theory and simulations is a finite-size effect. In contrast, for $K = 2$ it is a combination of finite-size effects and the inability of the mean-field theory to account for correlations between the data in clusters for small separations Δ. Such correlations are also responsible for a breakdown of the mean-field theory when $K > L$ (see Fig. 1). For larger separations Δ the theory is in good agreement with the simulations, see Fig. 4, and discrepancies again reflect only finite-size effects.

The magnitude of the finite-size effects can be estimated when $K = L$ for any $d/N < 1$, by the following argument. For the empirical covariance matrix $\hat{\boldsymbol{\Lambda}}$ of a sample of M d-dimensional data vectors generated from the Gaussian distribution $\mathcal{N}(\mathbf{m}, \boldsymbol{\Lambda})$ the random quantity $\log|\hat{\boldsymbol{\Lambda}}|$ will for large M be described by the distribution $\mathcal{N}[\log|\boldsymbol{\Lambda}| + \tau(M,d), \sigma^2(M,d)]$, where $\tau(M,d) = \sum_{\ell=1}^{d}\psi(\frac{M-\ell+1}{2}) - d\log(\frac{M}{2})$ and $\sigma^2(M,d) = \sum_{\ell=1}^{d}\frac{2}{M-\ell+1}$ [19]. Assuming that $K = L$ and that the clustering is perfect allows us to compute, by following steps similar to those followed in the Appendix G, the average and variance of the entropy (13). They are found to be given by $\min_\alpha F(\boldsymbol{\alpha}) + \sum_{\nu=1}^{L}\gamma(\nu)\,\tau(\gamma(\nu)N, d)$ and $\frac{1}{4}\sum_{\nu=1}^{L}\gamma^2(\nu)\,\sigma^2(\gamma(\nu)N, d)$, respectively.

When evaluated for real data sets, the entropy function (13) may also have value as an exploratory tool. To show this, we consider the Wisconsin Diagnostic Breast Cancer (WDBC) data set [20], which describes characteristics of cell nuclei in the images of cells extracted from tumors [21] and contains $N = 569$ data points of dimension $d = 30$. This data set has two (linearly separable) classes, which we assume to be the "true" clusters, one is "benign," represented by 357 data points, and the other is "malignant," represented by 212 data points [21]. A first simple unsupervised method which one might apply to this data set is hierarchical clustering, which uses pairwise distances between the data points to build a hierarchy of clusters, see, e.g., Ref. [22]. The agglomerative version of this algorithm, with Euclidean distances, separates this data into clusters of sizes 549 and 20 at the $K = 2$ clusters

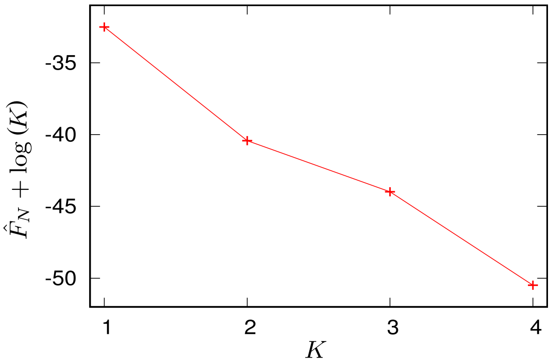

FIG. 5. The sum $\hat{F}_N + \log(K)$, where $\hat{F}_N \equiv \min_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$, as computed for the Wisconsin Diagnostic Breast Cancer data [20] (red crosses connected by lines), shown as a function of the assumed number of clusters K. These results suggest that the true number of clusters in this data set is at least $K = 4$.

level of hierarchy, into clusters of sizes 549, 19, and 1 at the $K = 3$ clusters level of hierarchy, into clusters of sizes 438, 111, 19, and 1 at the $K = 4$ clusters level of hierarchy, etc. Hence, on assuming (correctly) that $K = 2$, one cannot recover the true clusters of the WDBC data with this algorithm. Alternatively, the K-means clustering algorithm, see, e.g., Ref. [5], which minimizes the squared Euclidean distance between the points in a cluster, "finds" in the WDBC data set (again on assuming $K = 2$) clusters of sizes 438 and 131. On comparing these with the true clusters, we observe that K-means "misclassifies" 83 data points in total. It is interesting that the clusters found by K-means were also present in the four clusters generated via hierarchical clustering.

Using instead the gradient descent minimization of (13) as a clustering protocol suggests that there are more than $K = 4$ clusters[1] in the WDBC data set (see Fig. 5). For $K = 2$ the algorithm outputs clusters of sizes 328 and 241, which is, compared with the hierarchical and K-means results, much closer to the true sizes 357 and 212 of the WDBC data set. Now 57 data points were misclassified, which can be explained by the nonsphericity of clusters in this data set. In particular, for any data covariance matrix $\hat{\Sigma}$ the ratio $\mathcal{S}(\hat{\Sigma}) = \mathrm{Tr}^2(\hat{\Sigma})/d\mathrm{Tr}(\hat{\Sigma}^2)$ can be used as a measure of "sphericity" of data, see, e.g., Ref. [23]. We note that $1/d \leqslant \mathcal{S}(\hat{\Sigma}) \leqslant 1$, and that the lower bound $1/d$ is saturated only when a few eigenvalues dominate all others for large d, i.e., when only a few "directions" in $\mathbb{R}^d$ contribute to the variability in the data. The upper bound is saturated when all eigenvalues are equal, i.e., all directions in $\mathbb{R}^d$ contribute equally to the variability. The sphericity values of the "benign" and "malignant" clusters in the WDBC data set are given by 0.034 and 0.036, respectively, so the data in these clusters is highly nonspherical. This indeed suggests that the entropy function (13), derived on assuming arbitrary multivariate Gaussian distributions of a data in the clusters, is better equipped to deal with this scenario than hierarchical or K-means clustering.

V. SUMMARY

In conclusion, in this paper we have demonstrated that mapping Bayesian clustering of data to a statistical mechanical problem is not only possible, but in fact also quite intuitive and fruitful. It enables us to identify objectively the most plausible number of clusters in a data set, and to obtain transparent interpretations and explanations of why and how conventional clustering methods (which are quite often based on ad hoc definitions) may or may not fail to detect clusters correctly, dependent on the quantitative features of the data.

One possible extension of this work, currently in progress, is a more general analytical treatment of this Bayesian clustering problem, in which the distribution $P(F) = \int d\boldsymbol{\alpha}\, P(\boldsymbol{\alpha})\, \delta[F - F(\boldsymbol{\alpha})]$ is no longer assumed to converge to a δ distribution for large N. This will allow us to tackle also the nontrivial regime where $N, d \to \infty$ with N/d finite, and to correct the present mean-field theory in the $K > L$ regime.

ACKNOWLEDGMENT

This work was supported by the Medical Research Council of the United Kingdom (Grant No. MR/L01257X/1).

APPENDIX A: MODEL OF DATA AND BAYESIAN CLUSTERING

Let us assume that we observe the sample $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ for all i, drawn from the distribution

$$p(\mathbf{X}|\boldsymbol{\Theta}, \Pi) = \prod_{\mu=1}^{|\Pi|} \prod_{i_\mu \in S_\mu} p\big(\mathbf{x}_{i_\mu}\big|\boldsymbol{\theta}_\mu\big), \tag{A1}$$

generated by the partition $\Pi = \{S_1, S_2, \ldots, S_{|\Pi|}\}$, where the index sets $S_\mu \neq \emptyset$ obey $S_\mu \cap S_\nu = \emptyset$ for $\mu \neq \nu$, and $\cup_{\mu=1}^{|\Pi|} S_\mu = [N]$, with the short-hand $[N] = \{1, \ldots, N\}$. Furthermore, we assume that each parameter $\boldsymbol{\theta}_\mu$ is sampled randomly and independently from the distribution $p(\boldsymbol{\theta}_\mu)$, and that we are also given the prior distribution of Π, $P(\Pi)$. This allows us to write down the joint distribution

$$p(\mathbf{X}, \boldsymbol{\Theta}, \Pi) = p(\mathbf{X}|\boldsymbol{\Theta}, \Pi) p(\Pi) \prod_{\mu=1}^{|\Pi|} p(\boldsymbol{\theta}_\mu), \tag{A2}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{|\Pi|}\}$. On integrating out the parameters $\boldsymbol{\theta}_\mu$ in the above we obtain the distribution

$$p(\mathbf{X}, \Pi) = \langle p(\mathbf{X}|\boldsymbol{\Theta}, \Pi)\rangle_{\boldsymbol{\Theta}|\Pi}\, p(\Pi), \tag{A3}$$

where $\langle f(\boldsymbol{\Theta})\rangle_{\boldsymbol{\Theta}|\Pi} = \int f(\boldsymbol{\Theta}) \{\prod_{\mu=1}^{|\Pi|} p(\boldsymbol{\theta}_\mu)\, d\boldsymbol{\theta}_\mu\}$. From this follows the conditional distribution

$$p(\Pi|\mathbf{X}) = \frac{p(\mathbf{X}|\Pi) p(\Pi)}{\sum_{\tilde{\Pi}} p(\mathbf{X}|\tilde{\Pi}) p(\tilde{\Pi})} \tag{A4}$$

with

$$p(\mathbf{X}|\Pi) = \langle p(\mathbf{X}|\boldsymbol{\Theta}, \Pi)\rangle_{\boldsymbol{\Theta}|\Pi}. \tag{A5}$$

[1] For $K > 4$, this approach favors small clusters, i.e., we are in nonasymptotic regime, which suggests that a full Bayesian framework is more appropriate for this data.

Let us next consider the "partition function"

$$\begin{aligned}\sum_{\Pi} p(\mathbf{X}|\Pi)p(\Pi) &= \sum_{K=1}^{N}\sum_{\Pi} p(\mathbf{X}|\Pi)p(\Pi)\mathbb{1}[|\Pi|=K]\\ &= \sum_{K=1}^{N}\sum_{\Pi} p(\mathbf{X}|\Pi)p(\Pi|K)p(K),\end{aligned}\tag{A6}$$

where we have defined the two distributions:

$$\begin{aligned}p(\Pi|K) &= \frac{p(\Pi)\mathbb{1}[|\Pi|=K]}{\sum_{\tilde{\Pi}} p(\tilde{\Pi})\mathbb{1}[|\tilde{\Pi}|=K]}\\ p(K) &= \sum_{\Pi} p(\Pi)\mathbb{1}[|\Pi|=K].\end{aligned}\tag{A7}$$

Furthermore, if we define Π_K to be a partition Π with $|\Pi| = K$, i.e., $\Pi_K = \{S_1, \dots, S_K\}$, then

$$\sum_{\Pi} p(\mathbf{X}|\Pi)p(\Pi) = \sum_{K=1}^{N} p(K)\sum_{\Pi_K} p(\mathbf{X}|\Pi_K)p(\Pi_K|K)\tag{A8}$$

and the distribution of Π_K is given by

$$p(\Pi_K|\mathbf{X}) = \frac{p(\mathbf{X}|\Pi_K)p(\Pi_K|K)p(K)}{\sum_{\tilde{K}=1}^{N} p(\tilde{K})\sum_{\tilde{\Pi}_{\tilde{K}}} p(\mathbf{X}|\tilde{\Pi}_{\tilde{K}})p(\tilde{\Pi}_{\tilde{K}}|\tilde{K})}.\tag{A9}$$

The mode of this distribution is located at

$$\hat{\Pi}_K = \text{argmax}_{\Pi_K}\{p(\mathbf{X}|\Pi_K)p(\Pi_K|K)\},\tag{A10}$$

from which, in turn, it follows that the mode of the distribution (A4) is located at

$$\hat{\Pi} = \text{argmax}_{\hat{\Pi}_K}\{p(\mathbf{X}|\hat{\Pi}_K)p(\hat{\Pi}_K|K)p(K)\}.\tag{A11}$$

To see this one considers

$$\begin{aligned}\hat{\Pi} &= \text{argmax}_{\Pi}\{p(\mathbf{X}|\Pi)p(\Pi)\}\\ &= \text{argmax}_{\Pi}\{\{p(\mathbf{X}|\Pi_1)p(\Pi_1)\},\dots\{p(\mathbf{X}|\Pi_K)p(\Pi_K)\},\dots\\ &\quad\{p(\mathbf{X}|\Pi_N)p(\Pi_N)\}\},\end{aligned}$$

where $\{p(\mathbf{X}|\Pi_K)p(\Pi_K)\}$ is a set generated by $\{\Pi_K\}$. Clearly, $\max_{\Pi_K}\{p(\mathbf{X}|\Pi_K)p(\Pi_K)\} = p(\mathbf{X}|\hat{\Pi}_K)p(\hat{\Pi}_K)$, in which $\hat{\Pi}_K = \text{argmax}_{\Pi_K}\{p(\mathbf{X}|\Pi_K)p(\Pi_K)\}$, from which follows that

$$\begin{aligned}\hat{\Pi} &= \text{argmax}_{\hat{\Pi}_K}\{p(\mathbf{X}|\hat{\Pi}_K)p(\hat{\Pi}_K)\}\\ &= \text{argmax}_{\hat{\Pi}_K}\{p(\mathbf{X}|\hat{\Pi}_K)p(\hat{\Pi}_K|K)p(K)\}.\end{aligned}\tag{A12}$$

Any partition Π_K of the data into K clusters can be specified by the binary "allocation" variables $c_{i\mu} = \mathbb{1}[i \in S_\mu]$, where $i \in [N]$ and $\mu \in [K]$, forming the matrix $\mathbf{c}$ with $[\mathbf{c}]_{i\mu} = c_{i\mu}$. Hence $\Pi_K \equiv \Pi_K(\mathbf{c}) = \{S_1(\mathbf{c}), \dots, S_K(\mathbf{c})\}$. Conversely, an $N \times K$ matrix $\mathbf{c}$ with binary entries is a partition only if it satisfies the constraints $\sum_{\mu=1}^{K} c_{i\mu} = 1$ for all $i \in [N]$ and $\sum_{i=1}^{N} c_{i\mu} \geqslant 1$ for all $\mu \in [K]$. The simplest distribution implementing these constraints is the uniform distribution

$$p(\mathbf{c}|K) = \frac{\{\prod_{i=1}^{N}\mathbb{1}[\sum_{\nu=1}^{K} c_{i\nu} = 1]\}\{\prod_{\mu=1}^{K}\mathbb{1}[\sum_{j=1}^{N} c_{j\mu} \geqslant 1]\}}{\sum_{\tilde{\mathbf{c}}}\{\prod_{i=1}^{N}\mathbb{1}[\sum_{\nu=1}^{K} \tilde{c}_{i\nu} = 1]\}\{\prod_{\mu=1}^{K}\mathbb{1}[\sum_{j=1}^{N} \tilde{c}_{j\mu} \geqslant 1]\}}.\tag{A13}$$

The denominator in this expression gives the total number of partitions of the set $[N]$ into K subsets $\mathcal{S}(N, K)$, i.e., it equals the Stirling number of the second kind times the number $K!$ of subset permutations. Thus the probability of each individual partition $\mathbf{c}$ is given by $1/K!\mathcal{S}(N, K)$. We note that for $N \to \infty$ and $K \in O(N^0)$ we have $N^{-1}\log(K!\mathcal{S}(N, K)) \to \log(K)$ [17].

Using this new notation allows us to write the distribution $p(\mathbf{X}|\Pi_K)$ as

$$\begin{aligned}p(\mathbf{X}|\Pi_K) &\equiv p(\mathbf{X}|\mathbf{c}, K)\\ &= \langle e^{\sum_{\mu=1}^{K}\sum_{i=1}^{N} c_{i\mu}\log p(\mathbf{x}_i|\boldsymbol{\theta}_\mu)}\rangle_{\boldsymbol{\Theta}}\\ &= e^{-N\hat{F}_N(\mathbf{c}, \mathbf{X})},\end{aligned}\tag{A14}$$

where $\langle f(\boldsymbol{\Theta})\rangle_{\boldsymbol{\Theta}} = \int f(\boldsymbol{\Theta})\{\prod_{\mu=1}^{K} p(\boldsymbol{\theta}_\mu)d\boldsymbol{\theta}_\mu\}$, and we defined the log-likelihood

$$\hat{F}_N(\mathbf{c}, \mathbf{X}) = -\frac{1}{N}\log\langle e^{\sum_{\mu=1}^{K}\sum_{i=1}^{N} c_{i\mu}\log p(\mathbf{x}_i|\boldsymbol{\theta}_\mu)}\rangle_{\boldsymbol{\Theta}}.\tag{A15}$$

Furthermore, combining $p(\mathbf{c}, K) = p(\mathbf{c}|K)p(K)$ with (A14) gives us the joint distribution

$$p(\mathbf{X}, \mathbf{c}, K) = e^{-N\hat{F}_N(\mathbf{c}, \mathbf{X})}p(\mathbf{c}, K)\tag{A16}$$

from which we can derive the conditional distribution

$$p(\mathbf{c}, K|\mathbf{X}) = \frac{e^{-N\hat{F}_N(\mathbf{c}, \mathbf{X})}p(\mathbf{c}|K)p(K)}{\sum_{\tilde{K}=1}^{N} p(\tilde{K})\sum_{\tilde{\mathbf{c}}} e^{-N\hat{F}_N(\tilde{\mathbf{c}}, \mathbf{X})}p(\tilde{\mathbf{c}}|\tilde{K})}.\tag{A17}$$

For $K \in [N]$ the mode of this distribution is located at

$$\begin{aligned}\hat{\mathbf{c}}\,|K &= \text{argmax}_{\mathbf{c}}\, p(\mathbf{c}, K|\mathbf{X})\\ &= \text{argmax}_{\mathbf{c}}\{e^{-N\hat{F}_N(\mathbf{c}, \mathbf{X})}p(\mathbf{c}|K)\}\end{aligned}\tag{A18}$$

and hence the mode of (A4) is given by

$$\hat{\Pi} = \text{argmax}_{\hat{\mathbf{c}}|K}\{e^{-N\hat{F}_N(\mathbf{c}, \mathbf{X})}p(\hat{\mathbf{c}}|K)p(K)\},\tag{A19}$$

which is our MAP estimator of the partition of data Π.

APPENDIX B: LAPLACE APPROXIMATION

Let us consider the log-likelihood density (3). We note that $\hat{F}_N(\mathbf{c}, \mathbf{X}) = \sum_{\mu=1}^{K}\hat{F}_\mu^N(\mathbf{c}, \mathbf{X})$, where

$$\begin{aligned}\hat{F}_\mu^N(\mathbf{c}, \mathbf{X}) &= -\frac{1}{N}\log\int e^{-N\Phi_\mu(\boldsymbol{\theta}_\mu|\mathbf{c}, \mathbf{X})}p(\boldsymbol{\theta}_\mu)d\boldsymbol{\theta}_\mu\\ \Phi_\mu(\boldsymbol{\theta}_\mu|\mathbf{c}, \mathbf{X}) &= -\frac{1}{N}\sum_{i=1}^{N} c_{i\mu}\log p(\mathbf{x}_i|\boldsymbol{\theta}_\mu),\end{aligned}\tag{B1}$$

$\hat{F}^N_\mu(\mathbf{c},\,\mathbf{X})$ is a log-likelihood density of cluster μ. For large N it can be evaluated by the Laplace method [24]:

$$\begin{aligned}\hat{F}^N_\mu(\mathbf{c},\,\mathbf{X}) &= -\frac{1}{N}\log\Bigg[\frac{\int e^{-N\Phi_\mu(\boldsymbol{\theta}_\mu|\mathbf{c},\,\mathbf{X})}p(\boldsymbol{\theta}_\mu)d\boldsymbol{\theta}_\mu}{\int e^{-N\Phi_\mu(\tilde{\boldsymbol{\theta}}_\mu|\mathbf{c},\,\mathbf{X})}d\tilde{\boldsymbol{\theta}}_\mu}\\ &\quad\times\int e^{-N\Phi_\mu(\tilde{\boldsymbol{\theta}}_\mu|\mathbf{c},\,\mathbf{X})}d\tilde{\boldsymbol{\theta}}_\mu\Bigg]\\ &= \Phi_\mu(\boldsymbol{\theta}^*_\mu|\mathbf{c}),\end{aligned} \tag{B2}$$

where

$$\boldsymbol{\theta}^*_\mu = \mathrm{argmin}_{\boldsymbol{\theta}}\Phi_\mu(\boldsymbol{\theta}|\mathbf{c},\,\mathbf{X}). \tag{B3}$$

The stationarity condition $\frac{\partial}{\partial\theta_\mu(\ell)}\Phi_\mu(\boldsymbol{\theta}|\mathbf{c},\,\mathbf{X})=0$ for all ℓ, from which to solve $\boldsymbol{\theta}^*_\mu$, gives us the equations

$$\frac{\partial}{\partial\theta_\mu(\ell)}\frac{1}{N}\sum_{i=1}^N c_{i\mu}\log p(\mathbf{x}_i|\boldsymbol{\theta}_\mu)=0. \tag{B4}$$

Let us now evaluate (B4) for the multivariate Gaussian distributions

$$\mathcal{N}\big(\mathbf{x}\big|\mathbf{m}_\mu,\,\boldsymbol{\Lambda}_\mu^{-1}\big) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_\mu)^T\boldsymbol{\Lambda}_\mu(\mathbf{x}-\mathbf{m}_\mu)}}{\big|2\pi\boldsymbol{\Lambda}_\mu^{-1}\big|^{\frac{1}{2}}}, \tag{B5}$$

with the means $\mathbf{m}_\mu$ and the inverse covariance matrices $\boldsymbol{\Lambda}_\mu$. On assuming that $p(\mathbf{x}_i|\boldsymbol{\theta}_\mu)\equiv\mathcal{N}(\mathbf{x}|\mathbf{m}_\mu,\,\boldsymbol{\Lambda}_\mu^{-1})$, the desired log-likelihood density becomes

$$\begin{aligned}-\frac{1}{N}\sum_{i=1}^N c_{i\mu}\log\mathcal{N}\big(\mathbf{x}_i|\mathbf{m}_\mu,\,\boldsymbol{\Lambda}_\mu^{-1}\big)\\ = \frac{1}{2N}\sum_{i=1}^N c_{i\mu}(\mathbf{x}_i-\mathbf{m}_\mu)^T\boldsymbol{\Lambda}_\mu(\mathbf{x}_i-\mathbf{m}_\mu)\\ -\frac{M_\mu(\mathbf{c})}{2N}\log[(2\pi)^{-d}|\boldsymbol{\Lambda}_\mu|],\end{aligned} \tag{B6}$$

Here $M_\mu(\mathbf{c})=\sum_{i=1}^N c_{i\mu}=|S_\mu(\mathbf{c})|$ denotes the number of data points in cluster μ. Solving the equations $\frac{\partial}{\partial m_{\mu\ell}}\sum_{i=1}^N c_{i\mu}\log\mathcal{N}(\mathbf{x}_i|\mathbf{m}_\mu,\,\boldsymbol{\Lambda}_\mu^{-1})=0$ and

$$\frac{\partial}{\partial[\boldsymbol{\Lambda}_\mu]_{s\ell}}\sum_{i=1}^N c_{i\mu}\log\mathcal{N}\big(\mathbf{x}_i|\mathbf{m}_\mu,\,\boldsymbol{\Lambda}_\mu^{-1}\big)=0$$

gives us

$$\mathbf{m}_\mu = \frac{1}{M_\mu(\mathbf{c})}\sum_{i=1}^N c_{i\mu}\mathbf{x}_i \tag{B7}$$

$$\boldsymbol{\Lambda}_\mu^{-1} = \frac{1}{M_\mu(\mathbf{c})}\sum_{i=1}^N c_{i\mu}(\mathbf{x}_i-\mathbf{m}_\mu)(\mathbf{x}_i-\mathbf{m}_\mu)^T, \tag{B8}$$

i.e., the empirical mean and covariance of the data in cluster μ. Using the above results in equation (B6) we then obtain the log-likelihood density (13).

APPENDIX C: DISTRIBUTION OF LOG-LIKELIHOOD—A "FIELD THEORY" APPROACH

Let us assume that the data $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ are sampled from the distribution

$$p(\mathbf{X}|L)=\sum_{\tilde{\mathbf{c}}}q(\tilde{\mathbf{c}}|L)\left\{\prod_{\nu=1}^L\prod_{i_\nu\in S_\nu(\tilde{\mathbf{c}})}q_\nu\big(\mathbf{x}_{i_\nu}\big)\right\}, \tag{C1}$$

where $q(\tilde{\mathbf{c}}|L)$ is the "true" distribution of the partitions $\tilde{\mathbf{c}}$ of size L. We are interested in computing the distribution of log-likelihoods

$$P_N(F)=\sum_{\mathbf{c}}p(\mathbf{c}|K)\int d\mathbf{X}\;p(\mathbf{X}|L)\delta[F-\hat{F}_N(\mathbf{c},\mathbf{X})] \tag{C2}$$

$$\hat{F}_N(\mathbf{c},\,\mathbf{X})=-\frac{1}{N}\sum_{\mu=1}^K\sum_{i=1}^N c_{i\mu}\log p(\mathbf{x}_i|\boldsymbol{\theta}_\mu). \tag{C3}$$

Here $p(\mathbf{c}|K)$ is our "assumed" distribution of the partition $\mathbf{c}$ of size K. Let us now evaluate $P_N(F)$ further:

$$P_N(F)=\sum_{\mathbf{c},\tilde{\mathbf{c}}}p(\mathbf{c}|K)q(\tilde{\mathbf{c}}|L)\int\left\{\prod_{\nu=1}^L\prod_{i_\nu\in S_\nu(\tilde{\mathbf{c}})}q_\nu\big(\mathbf{x}_{i_\nu}\big)\right\}\delta[F-\hat{F}_N(\mathbf{c},\mathbf{X})]d\mathbf{X}. \tag{C4}$$

We note that the sum over $\tilde{\mathbf{c}}$ inside the function $\hat{F}_N(\mathbf{c},\mathbf{X})$ can be written in the following form:

$$\begin{aligned}-\hat{F}_N(\mathbf{c},\mathbf{X}) &= \sum_{\mu=1}^K\frac{|S_\mu(\mathbf{c})|}{N}\int\frac{1}{|S_\mu(\mathbf{c})|}\sum_{i_\mu\in S_\mu(\mathbf{c})}\delta(\mathbf{x}-\mathbf{x}_{i_\mu})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\\ &= \sum_{\mu=1}^K\frac{|S_\mu(\mathbf{c})|}{N}\int\frac{1}{|S_\mu(\mathbf{c})|}\sum_{\nu=1}^L\sum_{i_{\mu\nu}\in S_\mu(\mathbf{c})\cap S_\nu(\tilde{\mathbf{c}})}\delta\big(\mathbf{x}-\mathbf{x}_{i_{\mu\nu}}\big)\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\\ &= \sum_{\mu=1}^K\frac{|S_\mu(\mathbf{c})|}{N}\int Q_\mu(\mathbf{x}|\mathbf{c},\tilde{\mathbf{c}},\mathbf{X})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x},\end{aligned} \tag{C5}$$

where we have defined the density

$$Q_\mu(\mathbf{x}|\mathbf{c},\tilde{\mathbf{c}},\mathbf{X})=\frac{1}{|S_\mu(\mathbf{c})|}\sum_{\nu=1}^{L}\sum_{i_{\mu\nu}\in S_\mu(\mathbf{c})\cap S_\nu(\tilde{\mathbf{c}})}\delta\big(\mathbf{x}-\mathbf{x}_{i_{\mu\nu}}\big). \tag{C6}$$

Using the above form in (C4) we obtain

$$\begin{aligned}P_N(F)&=\sum_{\mathbf{c},\tilde{\mathbf{c}}}p(\mathbf{c}|K)\,q(\tilde{\mathbf{c}}|L)\int d\mathbf{X}\left\{\prod_{\nu=1}^{L}\prod_{i_\nu\in S_\nu(\tilde{\mathbf{c}})}q_\nu\big(\mathbf{x}_{i_\nu}\big)\right\}\delta\left[F+\sum_{\mu=1}^{K}\frac{|S_\mu(\mathbf{c})|}{N}\int Q_\mu(\mathbf{x}|\mathbf{c},\tilde{\mathbf{c}},\mathbf{X})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\right]\\&=\sum_{\mathbf{c},\tilde{\mathbf{c}}}p(\mathbf{c}|K)\,q(\tilde{\mathbf{c}}|L)\left\{\prod_{\mu=1}^{K}\prod_{\mathbf{x}}\int dQ_\mu(\mathbf{x})\right\}P_N[\{Q_\mu(\mathbf{x})\}|\mathbf{c},\tilde{\mathbf{c}}\,]\,\delta\left[F+\sum_{\mu=1}^{K}\frac{|S_\mu(\mathbf{c})|}{N}\int Q_\mu(\mathbf{x})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\right],\end{aligned} \tag{C7}$$

where we have defined the (functional) distribution

$$P_N[\{Q_\mu(\mathbf{x})\}|\mathbf{c},\tilde{\mathbf{c}}\,]=\int\left\{\prod_{\nu=1}^{L}\prod_{i_\nu\in S_\nu(\tilde{\mathbf{c}})}q_\nu\big(\mathbf{x}_{i_\nu}\big)\right\}d\mathbf{X}\prod_{\mu=1}^{K}\prod_{\mathbf{x}}\delta[Q_\mu(\mathbf{x})-Q_\mu(\mathbf{x}|\mathbf{c},\tilde{\mathbf{c}},\mathbf{X})]. \tag{C8}$$

Let us next consider

$$\begin{aligned}P_N[\{Q_\mu(\mathbf{x})\}|\mathbf{c},\tilde{\mathbf{c}}\,]&=\int d\mathbf{X}\left\{\prod_{\nu=1}^{L}\prod_{i_\nu\in S_\nu(\tilde{\mathbf{c}})}q_\nu\big(\mathbf{x}_{i_\nu}\big)\right\}\left\{\prod_{\mu=1}^{K}\prod_{\mathbf{x}}\int\frac{d\hat{Q}_\mu(\mathbf{x})}{2\pi/N}\right\}e^{iN\sum_{\mu=1}^{K}\int\hat{Q}_\mu(\mathbf{x})[Q_\mu(\mathbf{x})-Q_\mu(\mathbf{x}|\mathbf{c},\tilde{\mathbf{c}},\mathbf{X})]d\mathbf{x}}\\&=\left\{\prod_{\mu=1}^{K}\prod_{\mathbf{x}}\int\frac{d\hat{Q}_\mu(\mathbf{x})}{2\pi/N}\right\}e^{iN\sum_{\mu=1}^{K}\int\hat{Q}_\mu(\mathbf{x})Q_\mu(\mathbf{x})d\mathbf{x}}\\&\quad\times\int d\mathbf{X}\left\{\prod_{\nu=1}^{L}\prod_{i_\nu\in S_\nu(\tilde{\mathbf{c}})}q_\nu\big(\mathbf{x}_{i_\nu}\big)\right\}e^{\sum_{\nu=1}^{L}\sum_{\mu=1}^{K}\frac{N}{|S_\mu(\mathbf{c})|}\sum_{i_{\mu\nu}\in S_\mu(\mathbf{c})\cap S_\nu(\tilde{\mathbf{c}})}-i\hat{Q}_\mu(\mathbf{x}_{i_{\mu\nu}})}\\&=\left\{\prod_{\mu=1}^{K}\prod_{\mathbf{x}}\int\frac{d\hat{Q}_\mu(\mathbf{x})}{2\pi/N}\right\}e^{iN\sum_{\mu=1}^{K}\int\hat{Q}_\mu(\mathbf{x})Q_\mu(\mathbf{x})d\mathbf{x}}\prod_{\nu=1}^{L}\prod_{\mu=1}^{K}\prod_{i_{\mu\nu}\in S_\mu(\mathbf{c})\cap S_\nu(\tilde{\mathbf{c}})}\int q_\nu(\mathbf{x}_{i_{\mu\nu}})e^{-i\frac{N}{|S_\mu(\mathbf{c})|}\hat{Q}_\mu(\mathbf{x}_{i_{\mu\nu}})}d\mathbf{x}_{i_{\mu\nu}}\\&=\left\{\prod_{\mu=1}^{K}\prod_{\mathbf{x}}\int\frac{d\hat{Q}_\mu(\mathbf{x})}{2\pi/N}\right\}e^{iN\sum_{\mu=1}^{K}\int\hat{Q}_\mu(\mathbf{x})Q_\mu(\mathbf{x})d\mathbf{x}+N\sum_{\mu=1}^{K}\sum_{\nu=1}^{L}\frac{|S_\mu(\mathbf{c})\cap S_\nu(\tilde{\mathbf{c}})|}{N}\log\int d\mathbf{x}\,q_\nu(\mathbf{x})e^{-i\frac{N\hat{Q}_\mu(\mathbf{x})}{|S_\mu(\mathbf{c})|}}}.\end{aligned} \tag{C9}$$

Thus for $P_N[Q|\boldsymbol{\alpha}(\mathbf{c},\tilde{\mathbf{c}})]\equiv P_N[\{Q_\mu(\mathbf{x})\}|\mathbf{c},\tilde{\mathbf{c}}\,]$ we have

$$P_N[Q|\boldsymbol{\alpha}(\mathbf{c},\tilde{\mathbf{c}})]=\int\mathcal{D}\hat{Q}\,e^{N\Psi[Q,\hat{Q}|\boldsymbol{\alpha}(\mathbf{c},\tilde{\mathbf{c}})\,]}, \tag{C10}$$

where

$$\Psi[Q,\hat{Q}|\boldsymbol{\alpha}(\mathbf{c},\tilde{\mathbf{c}})\,]=i\sum_{\mu=1}^{K}\int\hat{Q}_\mu(\mathbf{x})Q_\mu(\mathbf{x})d\mathbf{x}+\sum_{\mu=1}^{K}\sum_{\nu=1}^{L}\alpha(\nu,\mu|\mathbf{c},\tilde{\mathbf{c}})\log\int q_\nu(\mathbf{x})\,e^{\frac{-i}{\alpha(\mu|\mathbf{c})}\hat{Q}_\mu(\mathbf{x})}d\mathbf{x}, \tag{C11}$$

with the usual short-hand for the path integral measure, $\int\mathcal{D}\hat{Q}\equiv\{\prod_{\mu=1}^{K}\prod_{\mathbf{x}}\int[d\hat{Q}_\mu(\mathbf{x})/(2\pi/N)]\}$. In the above formula we have also introduced the matrix $\boldsymbol{\alpha}(\mathbf{c},\tilde{\mathbf{c}})$, with entries $[\boldsymbol{\alpha}(\mathbf{c},\tilde{\mathbf{c}})]_{\nu\mu}=\alpha(\nu,\mu|\mathbf{c},\tilde{\mathbf{c}})$, where in turn $\alpha(\nu,\mu|\mathbf{c},\tilde{\mathbf{c}})=N^{-1}|S_\mu(\mathbf{c})\cap S_\nu(\tilde{\mathbf{c}})|$. We note that $\cup_{\mu=1}^{K}[S_\mu(\mathbf{c})\cap S_\nu(\tilde{\mathbf{c}})]=S_\nu(\tilde{\mathbf{c}})$ and that $\cup_{\nu=1}^{L}[S_\mu(\mathbf{c})\cap S_\nu(\tilde{\mathbf{c}})]=S_\mu(\mathbf{c})$. From these properties it follows that the entries $\alpha(\nu,\mu|\mathbf{c},\tilde{\mathbf{c}})\geqslant 0$ can be interpreted as representing a joint distribution, i.e., $\sum_{\mu=1}^{K}\sum_{\nu=1}^{L}\alpha(\nu,\mu|\mathbf{c},\tilde{\mathbf{c}})=1$, with the marginals $\sum_{\nu=1}^{L}\alpha(\nu,\mu|\mathbf{c},\tilde{\mathbf{c}})=\alpha(\mu|\mathbf{c})=|S_\mu(\mathbf{c})|/N$ and $\sum_{\mu=1}^{K}\alpha(\nu,\mu|\mathbf{c},\tilde{\mathbf{c}})=\alpha(\nu|\tilde{\mathbf{c}})=|S_\nu(\tilde{\mathbf{c}})|/N$. Using all these ingredients in Eq. (C7) then leads us to

$$\begin{aligned}P_N(F)&=\sum_{\mathbf{c},\tilde{\mathbf{c}}}p(\mathbf{c}|K)\,q(\tilde{\mathbf{c}}|L)\int\mathcal{D}Q\;P_N[Q|\boldsymbol{\alpha}(\mathbf{c},\tilde{\mathbf{c}})]\delta\left[F+\sum_{\mu=1}^{K}\frac{|S_\mu(\mathbf{c})|}{N}\int Q_\mu(\mathbf{x})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\right]\\&=\int d\boldsymbol{\alpha}\;P_N(\boldsymbol{\alpha})\int\mathcal{D}Q\;P_N[Q|\boldsymbol{\alpha}]\delta\left[F+\sum_{\mu=1}^{K}\alpha(\mu)\int Q_\mu(\mathbf{x})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\right],\end{aligned} \tag{C12}$$

where we have defined the integral measure $\int \mathcal{D}Q \equiv \{\prod_{\mu=1}^{K}\prod_{\mathbf{x}}\int dQ_\mu(\mathbf{x})\}$ as well as the short-hand $\int d\boldsymbol{\alpha} \equiv \prod_{\mu=1}^{K}\prod_{\nu=1}^{L}\int d\alpha(\nu,\mu)$. The distribution of $\boldsymbol{\alpha}$ is given by

$$P_N(\boldsymbol{\alpha}) = \sum_{\mathbf{c},\tilde{\mathbf{c}}} p(\mathbf{c}|K)\,q(\tilde{\mathbf{c}}|L)\prod_{\mu=1}^{K}\prod_{\nu=1}^{L}\delta[\alpha(\nu,\mu) - \alpha(\nu,\mu|\mathbf{c},\tilde{\mathbf{c}})]. \tag{C13}$$

Now for any smooth function g we can consider the following average:

$$\begin{aligned}\int P_N(F)\,g(F)\,dF &= \int d\boldsymbol{\alpha}\;P_N(\boldsymbol{\alpha})\int \mathcal{D}Q\;P_N[Q|\boldsymbol{\alpha}]\;g\left[-\sum_{\mu=1}^{K}\alpha(\mu)\int Q_\mu(\mathbf{x})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\right]\\ &= \int d\boldsymbol{\alpha}\;P_N(\boldsymbol{\alpha})\frac{\int \mathcal{D}Q\;P_N[Q|\boldsymbol{\alpha}]}{\int \mathcal{D}\tilde{Q}\;P_N[\tilde{Q}|\boldsymbol{\alpha}]}\;g\left[-\sum_{\mu=1}^{K}\alpha(\mu)\int Q_\mu(\mathbf{x})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\right]\\ &= \int d\boldsymbol{\alpha}\;P_N(\boldsymbol{\alpha})\frac{\int \mathcal{D}Q\;\int \mathcal{D}\hat{Q}\;e^{N\Psi[Q,\hat{Q}|\boldsymbol{\alpha}]}}{\int \mathcal{D}\tilde{Q}\;\int \mathcal{D}\hat{Q}\;e^{N\Psi[\tilde{Q},\hat{Q}|\boldsymbol{\alpha}]}}\;g\left[-\sum_{\mu=1}^{K}\alpha(\mu)\int Q_\mu(\mathbf{x})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\right].\end{aligned} \tag{C14}$$

Let us assume that $P_N(\boldsymbol{\alpha}) \to P(\boldsymbol{\alpha})$ as $N \to \infty$. Furthermore, we expect that in this limit the functional integral in the above equation is dominated by the extremum of the functional Ψ and hence for the distribution $P(F) = \lim_{N\to\infty} P_N(F)$ we obtain

$$\int P(F)\,g(F)\,dF = \int P(\boldsymbol{\alpha})\;g\left[-\sum_{\mu=1}^{K}\alpha(\mu)\int Q_\mu(\mathbf{x}|\boldsymbol{\alpha})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\right]d\boldsymbol{\alpha}, \tag{C15}$$

where $Q_\mu(\mathbf{x}|\boldsymbol{\alpha})$ is a solution of the saddle-point equations $\delta\Psi[Q,\hat{Q}|\boldsymbol{\alpha}]/\delta Q_\mu(\mathbf{x}) = 0$ and $\delta\Psi[Q,\hat{Q}|\boldsymbol{\alpha}]/\delta\hat{Q}_\mu(\mathbf{x}) = 0$. Solving the latter gives us the following two equations:

$$i\hat{Q}_\mu(\mathbf{x}) = 0, \tag{C16}$$

$$Q_\mu(\mathbf{x}) = \sum_{\nu=1}^{L}\frac{\alpha(\nu,\mu)}{\alpha(\mu)}\,\frac{q_\nu(\mathbf{x})e^{\frac{-i}{\alpha(\mu)}\hat{Q}_\mu(\mathbf{x})}}{\int q_\nu(\mathbf{x}')\,e^{\frac{-i}{\alpha(\mu)}\hat{Q}_\mu(\mathbf{x}')}d\mathbf{x}'} \tag{C17}$$

from which follows the equation

$$Q_\mu(\mathbf{x}|\boldsymbol{\alpha}) = \sum_{\nu=1}^{L}\alpha(\nu|\mu)q_\nu(\mathbf{x}), \tag{C18}$$

where $\alpha(\nu|\mu) = \alpha(\nu,\mu)/\alpha(\mu)$ is a conditional distribution. From the above we conclude that

$$P(F) = \int d\boldsymbol{\alpha}\;P(\boldsymbol{\alpha})\delta\left[F + \sum_{\mu=1}^{K}\sum_{\nu=1}^{L}\alpha(\nu,\mu)\int q_\nu(\mathbf{x})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}\right]. \tag{C19}$$

If we assume that $P(\boldsymbol{\alpha})$ is a δ function, this gives us the MF log-likelihood

$$F(\boldsymbol{\alpha}) = -\sum_{\mu=1}^{K}\sum_{\nu=1}^{L}\alpha(\nu,\mu)\int q_\nu(\mathbf{x})\log p(\mathbf{x}|\boldsymbol{\theta}_\mu)d\mathbf{x}, \tag{C20}$$

which is seen to be equivalent to (9). Let us next consider the distribution (C2) of the log-likelihood density (13):

$$P_N(F) = \sum_{\mathbf{c}} p(\mathbf{c}|K)\int d\mathbf{X}\;p(\mathbf{X}|L)\delta\left\{F - \sum_{\mu=1}^{K}\frac{|S_\mu(\mathbf{c})|}{2N}\log\left[(2\pi e)^d\left|\boldsymbol{\Lambda}_\mu^{-1}(\mathbf{c},\mathbf{X})\right|\right]\right\}, \tag{C21}$$

where $\boldsymbol{\Lambda}_\mu^{-1}(\mathbf{c},\mathbf{X})$ is the covariance matrix of the data in cluster μ, which can be written in the form

$$\boldsymbol{\Lambda}_\mu^{-1}(\mathbf{c},\mathbf{X}) = \frac{1}{|S_\mu(\mathbf{c})|}\sum_{i_\mu\in S_\mu(\mathbf{c})}\left[\mathbf{x}_{i_\mu} - \mathbf{m}_\mu(\mathbf{c})\right]\left[\mathbf{x}_{i_\mu} - \mathbf{m}_\mu(\mathbf{c})\right]^T, \tag{C22}$$

where $\mathbf{m}_\mu(\mathbf{c}) = \frac{1}{|S_\mu(\mathbf{c})|}\sum_{i_\mu \in S_\mu(\mathbf{c})} \mathbf{x}_{i_\mu}$. Further manipulation of $P_N(F)$ gives

$$P_N(F) = \sum_{\mathbf{c},\tilde{\mathbf{c}}} p(\mathbf{c}|K)\, q(\tilde{\mathbf{c}}|L) \int d\mathbf{X} \left\{ \prod_{\nu=1}^{L} \prod_{i_\nu \in S_\nu(\tilde{\mathbf{c}})} q_\nu\big(\mathbf{x}_{i_\nu}\big) \right\} \delta \left\{ F - \sum_{\mu=1}^{K} \frac{|S_\mu(\mathbf{c})|}{2N} \log\left[(2\pi e)^d \left|\mathbf{\Lambda}_\mu^{-1}(\mathbf{c},\mathbf{X})\right|\right] \right\} \tag{C23}$$

and the covariance matrix can be written in the form

$$\begin{aligned}\mathbf{\Lambda}_\mu^{-1}(\mathbf{c},\mathbf{X}) &= \frac{1}{|S_\mu(\mathbf{c})|} \sum_{\nu=1}^{L} \sum_{i_{\nu\mu} \in S_\mu(\mathbf{c}) \cap S_\nu(\tilde{\mathbf{c}})} \big[\mathbf{x}_{i_{\nu\mu}} - \mathbf{m}_\mu(\mathbf{c})\big]\big[\mathbf{x}_{i_{\nu\mu}} - \mathbf{m}_\mu(\mathbf{c})\big]^T \\ &= \int d\mathbf{x}\, Q_\mu(\mathbf{x}|\mathbf{c},\tilde{\mathbf{c}},\mathbf{X}) \left[\mathbf{x} - \int Q_\mu(\mathbf{y}|\mathbf{c},\tilde{\mathbf{c}},\mathbf{X})\, \mathbf{y}\, d\mathbf{y}\right]\left[\mathbf{x} - \int Q_\mu(\mathbf{z}|\mathbf{c},\tilde{\mathbf{c}},\mathbf{X})\, \mathbf{z}\, d\mathbf{z}\right]^T. \end{aligned} \tag{C24}$$

From the above it is clear that $\hat{F}_N$ is a functional of the density $Q_\mu(\mathbf{x}|\mathbf{c},\tilde{\mathbf{c}},\mathbf{X})$, defined in (C6), and the matrix $\boldsymbol{\alpha}(\mathbf{c},\tilde{\mathbf{c}})$. Following the same steps as in deriving equations (C7)–(C12) gives us

$$P_N(F) = \int d\boldsymbol{\alpha}\, P_N(\boldsymbol{\alpha}) \int \mathcal{D}Q\, P_N[Q|\boldsymbol{\alpha}]\delta \left\{ F - \sum_{\mu=1}^{K} \alpha(\mu) \frac{1}{2} \log\left[(2\pi e)^d \left|\mathbf{\Lambda}_\mu^{-1}[Q]\right|\right] \right\}, \tag{C25}$$

where

$$\mathbf{\Lambda}_\mu^{-1}[Q] = \int d\mathbf{x}\, Q_\mu(\mathbf{x})[\mathbf{x} - \int Q_\mu(\mathbf{y})\, \mathbf{y}\, d\mathbf{y}]\left[\mathbf{x} - \int Q_\mu(\mathbf{z})\, \mathbf{z}\, d\mathbf{z}\right]^T. \tag{C26}$$

Furthermore, for $N \to \infty$, using a similar argument as outlined in Eqs. (C14)–(C19), we obtain

$$P(F) = \int d\boldsymbol{\alpha}\, P(\boldsymbol{\alpha})\delta \left\{ F - \sum_{\mu=1}^{K} \alpha(\mu) \frac{1}{2} \log\left[(2\pi e)^d \left|\mathbf{\Lambda}_\mu^{-1}(\boldsymbol{\alpha})\right|\right] \right\}, \tag{C27}$$

where the covariance matrix $\mathbf{\Lambda}_\mu^{-1}(\boldsymbol{\alpha})$ is defined by

$$\mathbf{\Lambda}_\mu^{-1}(\boldsymbol{\alpha}) = \sum_{\nu=1}^{L} \alpha(\nu|\mu) \langle [\mathbf{x} - \mathbf{m}_\mu(\boldsymbol{\alpha})][\mathbf{x} - \mathbf{m}_\mu(\boldsymbol{\alpha})]^T \rangle_\nu, \tag{C28}$$

where $\mathbf{m}_\mu(\boldsymbol{\alpha}) = \sum_{\nu=1}^{L} \alpha(\nu|\mu)\langle \mathbf{x} \rangle_\nu$ is the mean, and we used the short-hand $\langle \{\cdots\} \rangle_\nu = \int q_\nu(\mathbf{x})\{\cdots\} d\mathbf{x}$. Assuming that $P(\boldsymbol{\alpha})$ is a δ function subsequently gives us the MF log-likelihood expression (14).

APPENDIX D: PROOFS OF INFORMATION-THEORETIC INEQUALITIES

In this section we compute lower bounds for the MF entropy (14). First, we show that $F(\boldsymbol{\alpha})$ satisfies the inequalities

$$F(\boldsymbol{\alpha}) \geqslant \sum_{\mu=1}^{K} \alpha(\mu) H(Q_\mu) \geqslant \sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu). \tag{D1}$$

Let us consider the Kullback-Leibler distance [15] $D(Q_\mu \| \mathcal{N}_\mu)$ between the mixture $Q_\mu(\mathbf{x}) = \sum_{\nu=1}^{L} \alpha(\nu|\mu) q_\nu(\mathbf{x})$ and the Gaussian distribution $\mathcal{N}(\mathbf{x}|\mathbf{m}_\mu, \mathbf{\Lambda}_\mu^{-1})$:

$$\begin{aligned} D(Q_\mu \| \mathcal{N}_\mu) &= \int d\mathbf{x} \sum_{\nu=1}^{L} \alpha(\nu|\mu) q_\nu(\mathbf{x}) \log\left[\frac{\sum_{\nu=1}^{L} \alpha(\nu|\mu) q_\nu(\mathbf{x})}{\mathcal{N}\big(\mathbf{x}|\mathbf{m}_\mu, \mathbf{\Lambda}_\mu^{-1}\big)}\right] \\ &= -H(Q_\mu) - \sum_{\nu=1}^{L} \alpha(\nu|\mu) \int d\mathbf{x}\, q_\nu(\mathbf{x}) \log \mathcal{N}\big(\mathbf{x}|\mathbf{m}_\mu, \mathbf{\Lambda}_\mu^{-1}\big) \\ &= -H(Q_\mu) - \sum_{\nu=1}^{L} \alpha(\nu|\mu) \int d\mathbf{x}\, q_\nu(\mathbf{x}) \log\left[\frac{e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_\mu)^T \mathbf{\Lambda}_\mu (\mathbf{x}-\mathbf{m}_\mu)}}{\left|2\pi \mathbf{\Lambda}_\mu^{-1}\right|^{\frac{1}{2}}}\right] \end{aligned}$$

$$= -H(Q_\mu) + \frac{1}{2}\log\left[(2\pi)^d\left|\boldsymbol{\Lambda}_\mu^{-1}\right|\right] + \frac{1}{2}\sum_{\nu=1}^{L}\alpha(\nu|\mu)\int d\mathbf{x}\, q_\nu(\mathbf{x})(\mathbf{x}-\mathbf{m}_\mu)^T\boldsymbol{\Lambda}_\mu(\mathbf{x}-\mathbf{m}_\mu)$$

$$= -H(Q_\mu) + \frac{1}{2}\log\left[(2\pi)^d\left|\boldsymbol{\Lambda}_\mu^{-1}\right|\right] + \frac{1}{2}\mathrm{Tr}\left\{\boldsymbol{\Lambda}_\mu\sum_{\nu=1}^{L}\alpha(\nu|\mu)\int q_\nu(\mathbf{x})(\mathbf{x}-\mathbf{m}_\mu)(\mathbf{x}-\mathbf{m}_\mu)^T d\mathbf{x}\right\}.$$

Let us define the mean and covariance of the distribution $Q_\mu(\mathbf{x}) = \sum_{\nu=1}^{L}\alpha(\nu|\mu)q_\nu(\mathbf{x})$ as $\mathbf{m}_\mu = \int Q_\mu(\mathbf{x})\,\mathbf{x}\,d\mathbf{x}$ and $\boldsymbol{\Lambda}_\mu^{-1} = \int Q_\mu(\mathbf{x})(\mathbf{x}-\mathbf{m}_\mu)(\mathbf{x}-\mathbf{m}_\mu)^T d\mathbf{x}$. Then $D(Q_\mu\|\mathcal{N}_\mu) = -H(Q_\mu) + \frac{1}{2}\log[(2\pi e)^d|\boldsymbol{\Lambda}_\mu^{-1}|]$ and from the simple property $D(Q_\mu\|\mathcal{N}_\mu) \geqslant 0$ we immediately deduce that

$$F(\boldsymbol{\alpha}) \geqslant \sum_{\mu=1}^{K}\alpha(\mu)H(Q_\mu). \tag{D2}$$

Furthermore, for the average entropy we find the following inequality

$$\begin{aligned}\sum_{\mu=1}^{K}\alpha(\mu)H(Q_\mu) &= \sum_{\mu=1}^{K}\alpha(\mu)\sum_{\nu_1=1}^{L}\alpha(\nu_1|\mu)\int q_{\nu_1}(\mathbf{x})\log\left[1/\sum_{\nu_2=1}^{L}\alpha(\nu_2|\mu)\,q_{\nu_2}(\mathbf{x})\right]d\mathbf{x}\\ &= \sum_{\mu=1}^{K}\sum_{\nu_1=1}^{L}\alpha(\nu_1,\mu)\int q_{\nu_1}(\mathbf{x})\left\{\log q_{\nu_1}(\mathbf{x}) - \log q_{\nu_1}(\mathbf{x}) + \log\left[1/\sum_{\nu_2=1}^{L}\alpha(\nu_2|\mu)\,q_{\nu_2}(\mathbf{x})\right]\right\}d\mathbf{x}\\ &= \sum_{\mu=1}^{K}\sum_{\nu=1}^{L}\alpha(\nu,\mu)D(q_\nu\|Q_\mu) + \sum_{\nu=1}^{L}\gamma(\nu)H(q_\nu) \geqslant \sum_{\nu=1}^{L}\gamma(\nu)H(q_\nu).\end{aligned} \tag{D3}$$

Second, for the average entropy

$$F_0 = \sum_{\nu=1}^{L}\gamma(\nu)\frac{1}{2}\log[(2\pi e)^d|\mathbf{c}_\nu|], \tag{D4}$$

where $\mathbf{c}_\nu = \langle\mathbf{x}\,\mathbf{x}^T\rangle_\nu - \langle\mathbf{x}\rangle_\nu\,\langle\mathbf{x}\rangle_\nu^T$ is the covariance matrix of $q_\nu(\mathbf{x})$, we can show that the following holds:

$$F(\boldsymbol{\alpha}) \geqslant F_0 \tag{D5}$$

for all $\boldsymbol{\alpha}$. The above equality follows from properties of the covariance matrix

$$\boldsymbol{\Lambda}_\mu^{-1}(\boldsymbol{\alpha}) = \sum_{\nu=1}^{L}\alpha(\nu|\mu)\,\mathbf{c}_\nu + \sum_{\nu=1}^{L}\alpha(\nu|\mu)[\langle\mathbf{x}\rangle_\nu - \mathbf{m}_\mu(\boldsymbol{\alpha})][\langle\mathbf{x}\rangle_\nu - \mathbf{m}_\mu(\boldsymbol{\alpha})]^T. \tag{D6}$$

To prove (D5) we first derive the inequality

$$\log\left|\sum_{\nu=1}^{L}\alpha(\nu)\mathbf{D}_\nu\right| \geqslant \sum_{\nu=1}^{L}\alpha(\nu)\log|\mathbf{D}_\nu| \tag{D7}$$

for symmetric positive definite matrices $\mathbf{D}_\nu$ and $\sum_{\nu=1}^{L}\alpha(\nu) = 1$, where $\alpha(\nu) \geqslant 0$. This inequality can be derived by repeated application of Minkowski's inequality for determinants, viz. $|\mathbf{D}+\mathbf{B}|^{\frac{1}{d}} \geqslant |\mathbf{D}|^{\frac{1}{d}} + |\mathbf{B}|^{\frac{1}{d}}$ for symmetric positive definite matrices $\mathbf{D}$ and $\mathbf{B}$:

$$\left|\sum_{\nu=1}^{L}\alpha(\nu)\mathbf{D}_\nu\right|^{\frac{1}{d}} = \left|\alpha(1)\mathbf{D}_1 + \sum_{\nu=2}^{L}\alpha(\nu)\mathbf{D}_\nu\right|^{\frac{1}{d}} \geqslant \alpha(1)|\mathbf{D}_1|^{\frac{1}{d}} + \left|\sum_{\nu=2}^{L}\alpha(\nu)\mathbf{D}_\nu\right|^{\frac{1}{d}} \geqslant \sum_{\nu=1}^{L}\alpha(\nu)|\mathbf{D}_\nu|^{\frac{1}{d}} \tag{D8}$$

from which follows the result

$$\log\left|\sum_{\nu=1}^{L}\alpha(\nu)\mathbf{D}_\nu\right| \geqslant d\log\left[\sum_{\nu=1}^{L}\alpha(\nu)|\mathbf{D}_\nu|^{\frac{1}{d}}\right] \geqslant \sum_{\nu=1}^{L}\alpha(\nu)\log|\mathbf{D}_\nu|. \tag{D9}$$

The last step in this argument relied on Jensen's inequality [15]. Let us now apply (D7) to the difference of entropies

$$\begin{aligned}2(F(\boldsymbol{\alpha}) - F_0) &= -\sum_{\nu=1}^{L}\gamma(\nu)\log|\mathbf{c}_\nu| + \sum_{\mu=1}^{K}\alpha(\mu)\log\left|\boldsymbol{\Lambda}_\mu^{-1}(\boldsymbol{\alpha})\right|\\ &= -\sum_{\nu=1}^{L}\gamma(\nu)\log|\mathbf{c}_\nu| + \sum_{\mu=1}^{K}\alpha(\mu)\log\left|\sum_{\nu=1}^{L}\alpha(\nu|\mu)\{\mathbf{c}_\nu + [\langle\mathbf{x}\rangle_\nu - \mathbf{m}_\mu(\boldsymbol{\alpha})][\langle\mathbf{x}\rangle_\nu - \mathbf{m}_\mu(\boldsymbol{\alpha})]^T\}\right|\end{aligned}$$

$$\geqslant -\sum_{\nu=1}^{L}\gamma(\nu)\log|\mathbf{c}_\nu| + \sum_{\mu=1}^{K}\alpha(\mu)\sum_{\nu=1}^{L}\alpha(\nu|\mu)\log|\mathbf{c}_\nu + [\langle\mathbf{x}\rangle_\nu - \mathbf{m}_\mu(\boldsymbol{\alpha})][\langle\mathbf{x}\rangle_\nu - \mathbf{m}_\mu(\boldsymbol{\alpha})]^T|$$

$$\geqslant -\sum_{\nu=1}^{L}\gamma(\nu)\log|\mathbf{c}_\nu| + d\sum_{\mu=1}^{K}\alpha(\mu)\sum_{\nu=1}^{L}\alpha(\nu|\mu)\log\{|\mathbf{c}_\nu|^{\frac{1}{d}} + |[\langle\mathbf{x}\rangle_\nu - \mathbf{m}_\mu(\boldsymbol{\alpha}])[\langle\mathbf{x}\rangle_\nu - \mathbf{m}_\mu(\boldsymbol{\alpha})]^T|^{\frac{1}{d}}\}. \quad \text{(D10)}$$

The last line in the above, obtained by Minkowski's inequality, is equal to zero, and hence $F(\boldsymbol{\alpha}) \geqslant F_0$ for all $\boldsymbol{\alpha}$.

APPENDIX E: ALGORITHMIC COST OF ORDERING RANDOM UNBIASED PARTITIONS

Let us assume that we have N "particles" of L different "colors" which are distributed into K different reservoirs. The probability that a particle has color $\nu \in [L]$ is $\gamma(\nu)$ and that it is in the reservoir μ is $1/K$. Assuming that color and reservoir allocation are independent events, the probability of "configuration" $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_N)$, where $\mathbf{a}_i = (a_i(1), a_i(2))$ with the color $a_i(1) \in [L]$ and reservoir number $a_i(2) \in [K]$ of the particle i, is given by

$$P(\mathbf{A}) = \prod_{i=1}^{N} P(\mathbf{a}_i), \quad \text{(E1)}$$

$$P(\mathbf{a}_i) \equiv P(a_i(1) = \nu, a_i(2) = \mu) = \frac{\gamma(\nu)}{K}. \quad \text{(E2)}$$

The total number of particles in reservoir μ is given by $N_\mu(\mathbf{A}) = \sum_{i=1}^{N}\delta_{\mu;a_i(2)}$. Let us now consider the joint distribution of particle numbers in reservoirs

$$\begin{aligned} P(N_1, \dots, N_K) &= \sum_{\mathbf{A}} P(\mathbf{A}) \prod_{\mu=1}^{K} \delta_{N_\mu;N_\mu(\mathbf{A})} \\ &= K^{-N} \sum_{a_1(2),\dots,a_N(2)} \prod_{\mu=1}^{K} \delta_{N_\mu;\sum_{i=1}^{N}\delta_{\mu;a_i(2)}} \\ &= K^{-N} \frac{N!}{\prod_{\mu=1}^{K} N_\mu!}, \end{aligned} \quad \text{(E3)}$$

where $\sum_{\mu=1}^{K} N_\mu = N$. The probability of observing the event that at least one reservoir is empty is given by

$$\begin{aligned} &1 - P(N_1 > 0, \dots, N_K > 0) \\ &= 1 - \sum_{N_1>0,\dots,N_K>0} K^{-N} \frac{N!}{\prod_{\mu=1}^{K} N_\mu!} \\ &= K^{-N}\left(\sum_{N_1\geqslant 0,\dots,N_K\geqslant 0} \frac{N!}{\prod_{\mu=1}^{K} N_\mu!} \right. \\ &\quad \left. - \sum_{N_1>0,\dots,N_K>0} \frac{N!}{\prod_{\mu=1}^{K} N_\mu!}\right) \\ &= \sum_{\ell=1}^{K-1} \binom{K}{\ell}\left(1 - \frac{\ell}{K}\right)^N. \end{aligned} \quad \text{(E4)}$$

Thus the probability of this event decays exponentially with increasing N and, as $N \to \infty$, the sequence $a_1(2), \dots, a_N(2)$, sampled from the distribution (E2) is, with high probability, a *partition* of the set $[N]$ into K subsets (or clusters). Furthermore, the entropy density $N^{-1}\log(K^N) = \log(K)$ of such sequences approaches the entropy density $N^{-1}\log[K!\,\mathcal{S}(N, K)]$ of the random partitions sampled uniformly from (A13).

Let us assume that $K \leqslant L$. The total number of particles of color ν, and the number of particles of color ν in reservoir μ are given, respectively, by $N_\nu(\mathbf{A}) = \sum_{i=1}^{N}\delta_{\nu;a_i(1)}$ and $N_{\nu\mu}(\mathbf{A}) = \sum_{i=1}^{N}\delta_{\nu;a_i(1)}\delta_{\mu;a_i(2)}$. The number of particles of color ν which are *not* in reservoir μ is the difference $N_\nu(\mathbf{A}) - N_{\nu\mu}(\mathbf{A})$. Suppose that each reservoir has a preference for particles of a particular color (or colors), i.e., there is an onto mapping $\nu \to \mu(\nu)$ between colors and reservoirs, then the total number of particles which are not in "their" reservoirs, i.e., the number of particles which are to be "moved" in order for all particles to be in reservoirs to which they belong, is given by the difference $\sum_{\nu=1}^{L}[N_\nu(\mathbf{A}) - N_{\nu\mu(\nu)}(\mathbf{A})] = N - \sum_{\nu=1}^{L} N_{\nu\mu(\nu)}(\mathbf{A})$.

We are interested in the average and variance of $N - \sum_{\nu=1}^{L} N_{\nu\mu(\nu)}(\mathbf{A})$. The average is given by

$$\begin{aligned} &\left\langle N - \sum_{\nu=1}^{L} N_{\nu\mu(\nu)}(\mathbf{A}) \right\rangle_{\mathbf{A}} \\ &= N - \sum_{\nu=1}^{L}\sum_{i=1}^{N}\langle\delta_{\nu;a_i(1)}\delta_{\mu(\nu);a_i(2)}\rangle_{\mathbf{A}} = N - \sum_{\nu=1}^{L}\sum_{i=1}^{N}\frac{\gamma(\nu)}{K} \\ &= N\frac{K-1}{K} \end{aligned} \quad \text{(E5)}$$

and the variance is given by

$$\begin{aligned} &\mathrm{Var}\left\{N - \sum_{\nu=1}^{L} N_{\nu\mu(\nu)}(\mathbf{A})\right\} \\ &= \mathrm{Var}\left\{\sum_{\nu=1}^{L} N_{\nu\mu(\nu)}(\mathbf{A})\right\} = \left\langle\left[\sum_{\nu=1}^{L} N_{\nu\mu(\nu)}(\mathbf{A}) - \frac{N}{K}\right]^2\right\rangle_{\mathbf{A}} \\ &= \frac{N}{K}\left(1 - \frac{1}{K}\right). \end{aligned} \quad \text{(E6)}$$

The average in the penultimate line of the above was computed as follows:

$$\begin{aligned} &\left\langle\left[\sum_{\nu=1}^{L} N_{\nu\mu(\nu)}(\mathbf{A})\right]^2\right\rangle_{\mathbf{A}} \\ &= \sum_{\nu}\sum_{i_1,i_2}\langle\delta_{\nu;a_{i_1}(1)}\delta_{\mu(\nu);a_{i_1}(2)}\delta_{\nu;a_{i_2}(1)}\delta_{\mu(\nu);a_{i_2}(2)}\rangle_{\mathbf{A}} \\ &\quad + \sum_{\nu_1\neq\nu_2}\sum_{i_1,i_2=1}\langle\delta_{\nu_1;a_{i_1}(1)}\delta_{\mu(\nu_1);a_{i_1}(2)}\delta_{\nu_2;a_{i_2}(1)}\delta_{\mu(\nu_2);a_{i_2}(2)}\rangle_{\mathbf{A}} \end{aligned}$$

$$
\begin{aligned}
&= \frac{N}{K} + \frac{N(N-1)}{K^2}\sum_{\nu=1}^{L}\gamma^2(\nu)\\
&\quad + \frac{N(N-1)}{K^2}\sum_{\nu_1\neq\nu_2}\gamma(\nu_1)\gamma(\nu_2)\\
&= \frac{N}{K} + \frac{N(N-1)}{K^2}. \qquad \text{(E7)}
\end{aligned}
$$

From the above derivations it follows that for a random unbiased partition to be ordered, i.e., for particles of the same color to occupy at most one reservoir, a fraction of particles has to be moved that is on average $\langle 1 - \frac{1}{N}\sum_{\nu=1}^{L} N_{\nu\mu(\nu)}(\mathbf{A})\rangle_{\mathbf{A}} = (K-1)/K$, with variance $\mathrm{Var}\{1 - \frac{1}{N}\sum_{\nu=1}^{L} N_{\nu\mu(\nu)}(\mathbf{A})\} = (1-K^{-1})/NK$.

APPENDIX F: DETAILS OF NUMERICAL EXPERIMENTS

In this section we study the performance of the simplest algorithm that minimizes the log-likelihood function (13) via gradient descent, for the data described in Fig. 2. The algorithm is implemented as follows:

(i) Start with any initial partition $\Pi(\mathbf{c}\,(0)) = \{S_1(\mathbf{c}\,(0)), \ldots, S_K(\mathbf{c}\,(0))\}$ and compute the log-likelihood $\hat{F}_N(\mathbf{c}\,(0), \mathbf{X})$.

(ii) For all $i \in [N]$, consider all possible moves of i from its current cluster $S_\mu(\mathbf{c})$ to a new cluster $S_\nu(\mathbf{c})$ and compute the new value $\hat{F}_N(\mathbf{c}, \mathbf{X})$ for each.

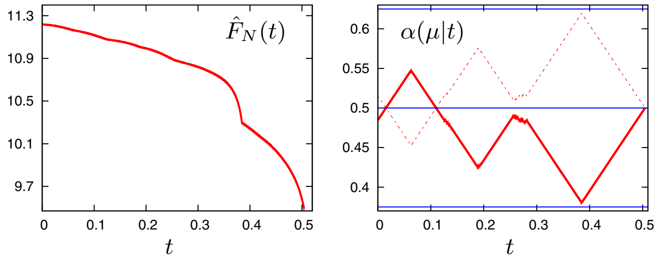

FIG. 7. Evolution of the log-likelihood, $\hat{F}_N(t) \equiv \hat{F}_N(\mathbf{c}(t), \mathbf{X})$, and the fraction of data in cluster μ, $\alpha(\mu|t) \equiv \alpha(\mu|\mathbf{c}(t))$, where $\mu = \{1, 2\}$, shown as functions of time (normalized number of "moves") in the gradient descent algorithm evolving from a random unbiased initial partition. The assumed number of clusters is $K = 2$. Blue horizontal lines correspond to the levels 3/8, 4/8, and 5/8.

(iii) Select and execute a move which gives the largest decrease in $\hat{F}_N(\mathbf{c}, \mathbf{X})$, and update $\Pi(\mathbf{c})$.

(iv) Continue the last two steps while the value of $\hat{F}_N(\mathbf{c}, \mathbf{X})$ continues to change.

(v) Output the partition $\Pi[\mathbf{c}\,(\infty)]$ and the value of $\hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$.

Using as initial states random partitions of data $\mathbf{c}\,(0)$, where each $i \in [N]$ has a probability $1/K$ of being allocated to one

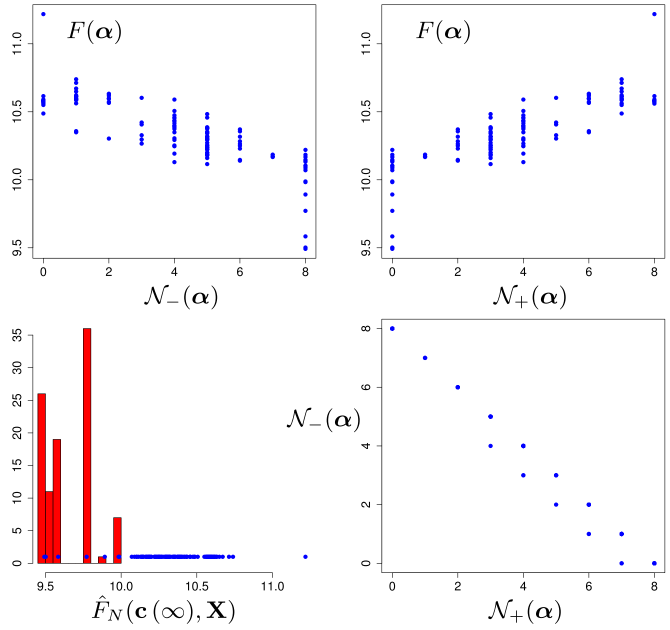

FIG. 6. Top left: $F(\boldsymbol{\alpha})$ as a function of the number of F-increasing directions $\mathcal{N}_-(\boldsymbol{\alpha})$. Top right: $F(\boldsymbol{\alpha})$ as a function of the number of F-decreasing directions $\mathcal{N}_+(\boldsymbol{\alpha})$. Bottom left: Histogram of log-likelihood values $\hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$, obtained by running gradient descent from a 100 different random unbiased partitions, with the assumed number $K = 2$ of clusters. Blue filled circles correspond to the MF log-likelihood, $F(\boldsymbol{\alpha})$, computed for all possible values of $\alpha(\nu, \mu) = \mathbb{1}[\nu \in S_\mu]\gamma(\nu)$. Bottom right: $\mathcal{N}_-(\boldsymbol{\alpha})$ as a function of $\mathcal{N}_+(\boldsymbol{\alpha})$.

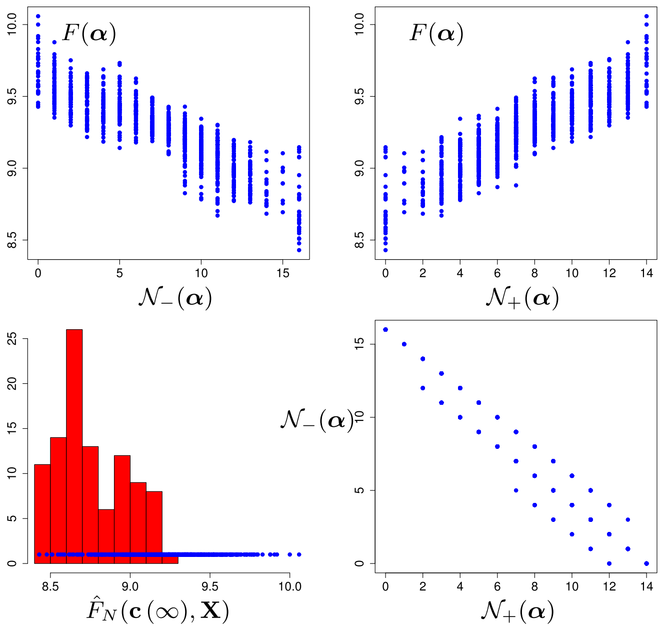

FIG. 8. Top left: $F(\boldsymbol{\alpha})$ as a function of the number of F-increasing directions $\mathcal{N}_-(\boldsymbol{\alpha})$. Top right: $F(\boldsymbol{\alpha})$ as a function of the number of F-decreasing directions $\mathcal{N}_+(\boldsymbol{\alpha})$. Bottom left: Histogram of log-likelihood values $\hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$, obtained by running gradient descent from a 100 different random unbiased partitions, with the assumed number $K = 3$ of clusters. Blue filled circles correspond to the MF log-likelihood, $F(\boldsymbol{\alpha})$, computed for all possible values of $\alpha(\nu, \mu) = \mathbb{1}[\nu \in S_\mu]\gamma(\nu)$. Bottom right: $\mathcal{N}_-(\boldsymbol{\alpha})$ as a function of $\mathcal{N}_+(\boldsymbol{\alpha})$.

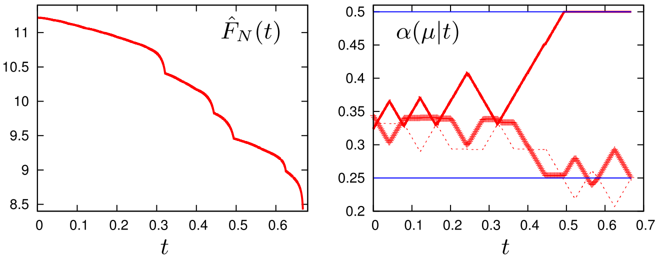

FIG. 9. Evolution of the log-likelihood, $\hat{F}_N(t) \equiv \hat{F}_N(\mathbf{c}(t), \mathbf{X})$, and the fraction of data in cluster μ, $\alpha(\mu|t) \equiv \alpha(\mu|\mathbf{c}(t))$, where $\mu = \{1, 2, 3\}$, shown as functions of time (normalized number of "moves") in the gradient descent algorithm evolving from a random unbiased initial partition. The assumed number of clusters is $K = 3$. Blue horizontal lines correspond to the levels $2/8$ and $4/8$.

of the K clusters,[2] we run the above algorithm for each value of $K \in [17]$ for 100 different initalizations $\mathbf{c}\,(0)$ and select the final partition, $\mathbf{c}\,(\infty)$, with the smallest value of $\hat{F}_N \equiv \hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$. The latter is our estimate of $\min_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$. We also compute, with the same parameters used to generate our data, the mean-field log-likelihood $F(\boldsymbol{\alpha})$ via Eq. (14).

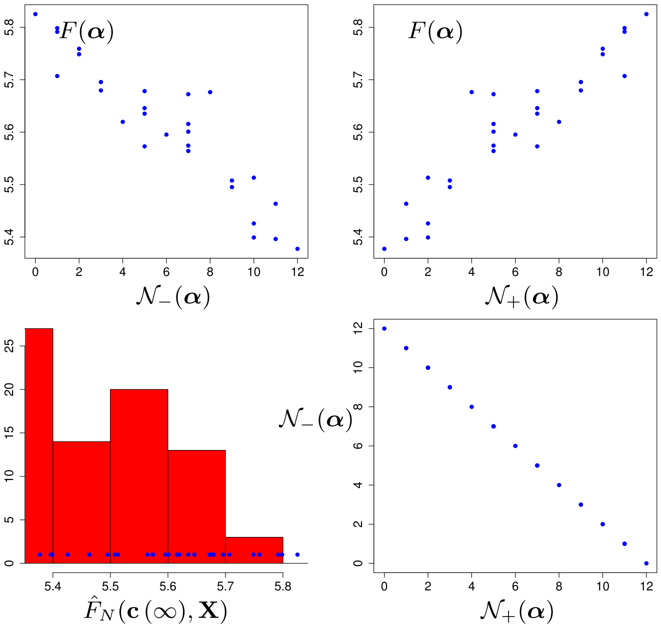

FIG. 10. Top left: $F(\boldsymbol{\alpha})$ as a function of the number of F-increasing directions $\mathcal{N}_-(\boldsymbol{\alpha})$. Top right: $F(\boldsymbol{\alpha})$ as a function of the number of F-decreasing directions $\mathcal{N}_+(\boldsymbol{\alpha})$. Bottom left: Histogram of log-likelihood values $\hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$, obtained by running gradient descent from a 100 different random unbiased partitions, with the assumed number $K = 7$ of clusters. Blue filled circles correspond to the MF log-likelihood, $F(\boldsymbol{\alpha})$, computed for all possible values of $\alpha(\nu, \mu) = \mathbb{1}[\nu \in S_\mu]\gamma(\nu)$. Bottom right: $\mathcal{N}_-(\boldsymbol{\alpha})$ as a function of $\mathcal{N}_+(\boldsymbol{\alpha})$.

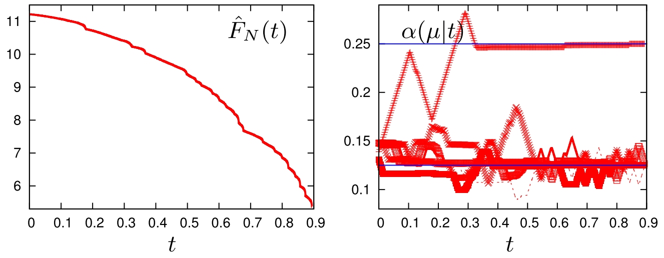

FIG. 11. Evolution of the log-likelihood, $\hat{F}_N(t) \equiv \hat{F}_N(\mathbf{c}(t), \mathbf{X})$, and the fraction of data in cluster μ, $\alpha(\mu|t) \equiv \alpha(\mu|\mathbf{c}(t))$, where $\mu = \{1, 2, \ldots, 7\}$, shown as functions of time (normalized number of "moves") in the gradient descent algorithm evolving from a random unbiased initial partition. The assumed number of clusters is $K = 7$. Blue horizontal lines correspond to the levels $1/8$ and $2/8$.

When $K \leqslant L$, the log-likelihood $\hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$ is dominated by partitions $\mathbf{c}\,(\infty)$ corresponding to local minima and saddlepoints of $F(\boldsymbol{\alpha})$. The matrix $\boldsymbol{\alpha}$ is defined by the entries $[\boldsymbol{\alpha}]_{\nu\mu} = \mathbb{1}[\nu \in S_\mu]\gamma(\nu)$, generated by partitions $\Pi = \{S_1, \ldots, S_K\}$ of the set $[L]$ into K subsets. The total number of partitions is given by $\mathcal{S}(L, K)$. To enumerate all partitions we use the algorithm in Ref. [25]. We classify turning points of $F(\boldsymbol{\alpha})$ as follows. For a given Π and its associated matrix $\boldsymbol{\alpha}$ we count the number $\mathcal{N}_+(\boldsymbol{\alpha})$ of elementary "moves" into the new partition $\tilde{\Pi}$ and $\tilde{\boldsymbol{\alpha}}$ (in a single elementary "move," a member of the set S_μ, with $|S_\mu| > 1$, is moved into the set S_ν) for which $F(\boldsymbol{\alpha}) > F(\tilde{\boldsymbol{\alpha}})$ and the number $\mathcal{N}_-(\boldsymbol{\alpha})$ of moves for which $F(\boldsymbol{\alpha}) < F(\tilde{\boldsymbol{\alpha}})$. If $\mathcal{N}_+(\boldsymbol{\alpha}) = 0$, then the state $\boldsymbol{\alpha}$ is a (possibly local) minimum, and if $\mathcal{N}_-(\boldsymbol{\alpha}) = 0$, then the state $\boldsymbol{\alpha}$ is a (possibly local) maximum. All other cases are saddle points. In Figs. 6, 8, 10, 12, and 14 we compare $\hat{F}_N(\mathbf{c}\,(\infty), \mathbf{X})$ with $F(\boldsymbol{\alpha})$.

Those turning points of $F(\boldsymbol{\alpha})$ that are of the form $[\boldsymbol{\alpha}]_{\nu\mu} = \mathbb{1}[\nu \in S_\mu]\gamma(\nu)$ also act as dynamic "attractors." This can be seen by comparing Fig. 6 to Fig. 7, and Fig. 8 to Fig. 9, etc. Here $\hat{F}_N(t) \equiv \hat{F}_N(\mathbf{c}\,(t), \mathbf{X})$, as computed during the simulated process, is seen to evolve from plateau to plateau by a succession of rapid relaxations, and the value of $\hat{F}_N(t)$ at the beginning of each plateau can be (approximately) mapped to the value of $F(\boldsymbol{\alpha})$ via the fractions $\alpha(\mu) = \sum_{\nu=1}^{L} \mathbb{1}[\nu \in$

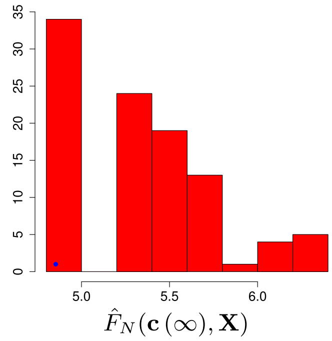

FIG. 12. Histogram of the log-likelihood values obtained by running the gradient descent algorithm from a 100 different random unbiased partitions, with the assumed number $K = 8$ of clusters. The blue filled circle corresponds to the MF lower bound $\sum_{\nu=1}^{L} \gamma(\nu)H(q_\nu) = 4.853905$.

[2] In Appendix E we proved that for $N \to \infty$ the matrix $\mathbf{c}$ constructed in this way is, with high probability, a *partition* of the set $[N]$ into the K subsets.

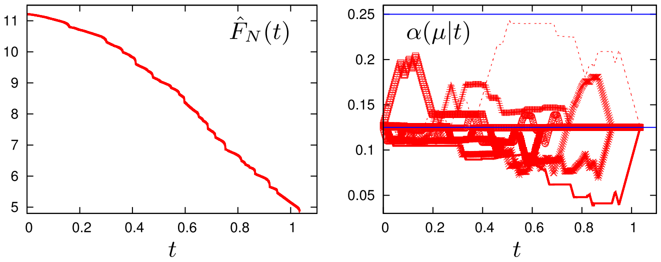

FIG. 13. Evolution of the log-likelihood, $\hat{F}_N(t) \equiv \hat{F}_N(\mathbf{c}(t), \mathbf{X})$, and the fraction of data in cluster μ, $\alpha(\mu|t) \equiv \alpha(\mu|\mathbf{c}(t))$, where $\mu = \{1, 2, \ldots, 8\}$, shown as functions of time (normalized number of "moves") in the gradient descent algorithm evolving from a random unbiased initial partition. The assumed number of clusters is $K = 8$. Blue horizontal lines correspond to the levels $1/8$ and $2/8$.

$S_\mu]\gamma(\nu)$ of data in clusters μ (Fig. 10). However, as K is increased, more and more attractors are not of the form $\mathbb{1}[\nu \in S_\mu]\gamma(\nu)$ (see Figs. 9, 11, 12, 13, and 15).

The predictions of the mean-field log-likelihood $F(\boldsymbol{\alpha})$ for $\min_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$ are incorrect when $K > L$. The log-likelihood $F(\boldsymbol{\alpha})$ is bounded from below by the average entropy $\sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu)$, but in this regime the gap between this lower bound and $\min_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$ is widening as we increase the number of assumed clusters K. This effect can be clearly seen in Fig. 14. We also see in this figure that $\sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu)$ separates the low entropy states obtained by gradient descent into two sets (Fig. 15). The first set, which includes $\text{argmin}_{\mathbf{c}} \hat{F}_N(\mathbf{c}, \mathbf{X})$, is given by[3] $\{\mathbf{c} : \hat{F}_N(\mathbf{c}, \mathbf{X}) \leqslant \sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu)\}$, and the second set is given by $\{\mathbf{c} : \hat{F}_N(\mathbf{c}, \mathbf{X}) > \sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu)\}$. Since for $K > L$ we have $F(\boldsymbol{\alpha}) > \sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu)$, we expect that $\min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$ gives correct predictions for at least some of the low entropy states in the second set.

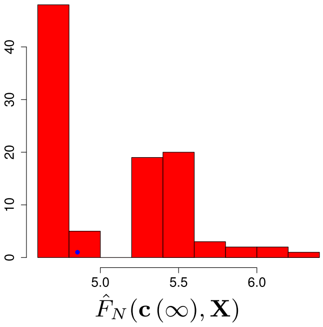

FIG. 14. Histogram of the log-likelihood values obtained by running the gradient descent algorithm from a 100 different random unbiased partitions, with the assumed number $K = 9$ of clusters. The blue filled circle corresponds to the MF lower bound $\sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu) = 4.853905$.

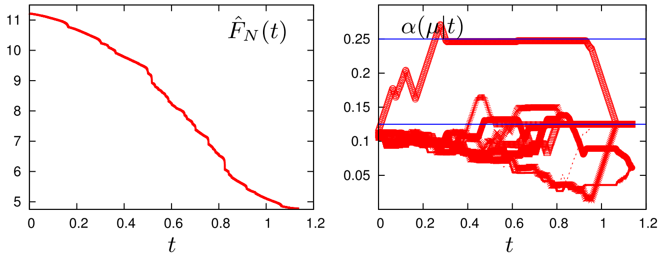

FIG. 15. Evolution of the log-likelihood, $\hat{F}_N(t) \equiv \hat{F}_N(\mathbf{c}(t), \mathbf{X})$, and the fraction of data in cluster μ, $\alpha(\mu|t) \equiv \alpha(\mu|\mathbf{c}(t))$, where $\mu = \{1, 2, \ldots, 9\}$, shown as functions of time (normalized number of "moves") in the gradient descent algorithm evolving from a random unbiased initial partition. The assumed number of clusters is $K = 9$. Blue horizontal lines correspond to the levels $1/8$ and $2/8$.

APPENDIX G: ESTIMATION OF DIFFERENTIAL ENTROPY

In this section we compute the finite sample-size corrections to the MF entropy (14). In order to do this we first note that for a sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where each $\mathbf{x}_i \in \mathbb{R}^d$ is drawn from the multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\mathbf{m}, \boldsymbol{\Lambda})$, the empirical covariance matrix $\hat{\boldsymbol{\Lambda}} = N^{-1} \sum_{i=1}^{N} (\mathbf{x}_i - \hat{\mathbf{m}})(\mathbf{x}_i - \hat{\mathbf{m}})^T$, where $\hat{\mathbf{m}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$ is the empirical mean, obeys the following asymptotic law: $[\log|\hat{\boldsymbol{\Lambda}}| - \log|\boldsymbol{\Lambda}| - d(d+1)/2N]/\sqrt{2d/N} \to \mathcal{N}(0, 1)$ as $N \to \infty$ (see Ref. [19] and references therein). This is equivalent to stating $\log|\hat{\boldsymbol{\Lambda}}| \to \log|\boldsymbol{\Lambda}| + d(d+1)/2N + z\sqrt{2d/N}$, where $z \sim \mathcal{N}(0, 1)$.

Let us assume that the above is true for the empirical covariance matrices that feature in the log-likelihood (13) and evaluate $\hat{F}_N(\mathbf{c})$ for large N:

$$\begin{aligned}
\hat{F}_N(\mathbf{c}) &= \sum_{\mu=1}^{K} \frac{M_\mu(\mathbf{c})}{N} \frac{1}{2} \log\left[(2\pi e)^d \left|\boldsymbol{\Lambda}_\mu^{-1}(\mathbf{c})\right|\right] \\
&= \sum_{\mu=1}^{K} \frac{M_\mu(\mathbf{c})}{2N} \left\{\log\left[(2\pi e)^d \left|\boldsymbol{\Lambda}_\mu^{-1}(\boldsymbol{\alpha})\right|\right]\right. \\
&\quad \left. + \frac{d(d+1)}{2M_\mu(\mathbf{c})} + z_\mu \sqrt{\frac{2d}{M_\mu(\mathbf{c})}}\right\} \\
&= F(\boldsymbol{\alpha}) + \sum_{\mu=1}^{K} \frac{M_\mu(\mathbf{c})}{2N} \left\{\frac{d(d+1)}{2M_\mu(\mathbf{c})} + z_\mu \sqrt{\frac{2d}{M_\mu(\mathbf{c})}}\right\} \\
&= F(\boldsymbol{\alpha}) + \frac{Kd(d+1)}{4N} + \sum_{\mu=1}^{K} z_\mu \sqrt{\frac{d\,\alpha(\mu)}{2N}}. \qquad \text{(G1)}
\end{aligned}$$

The average and variance of the above random variable are given by $F(\boldsymbol{\alpha}) + Kd(d+1)/4N$ and $d/2N$, respectively. We expect the above result to be exact when $F(\boldsymbol{\alpha}) = \sum_{\nu=1}^{L} \gamma(\nu) H(q_\nu)$, which can only happen when $K = L$, and all $q_\nu(\mathbf{x})$ are Gaussian distributions.

[3] The equality in this definition can only be true when $K = L$ (see Fig. 12).

[1] M. A. Kuhn, E. D. Feigelson, K. V. Getman, A. J. Baddeley, P. S. Broos, A. Sills, M. R. Bate, M. S. Povich, K. L. Luhman, H. A. Busk, T. Naylor, and R. R. King, Astrophys. J. **787**, 107 (2014).

[2] R. de Souza, M. Dantas, M. Costa-Duarte, E. Feigelson, M. Killedar, P.-Y. Lablanche, R. Vilalta, A. Krone-Martins, R. Beck, and F. Gieseke, Mon. Notices Royal Astron. Soc. **472**, 2808 (2017).

[3] W. P. Hanage, C. Fraser, J. Tang, T. R. Connor, and J. Corander, Science **324**, 1454 (2009).

[4] V. G. Martin, Y.-C. B. Wu, C. L. Townsend, G. H. C. Lu, J. S. O'Hare, A. Mozeika, A. C. C. Coolen, D. Kipling, F. Fraternali, and D. K. Dunn-Walters, Front. Immunol. **7**, 546 (2016).

[5] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, Berlin, 2006).

[6] B. J. Frey and D. Dueck, Science **315**, 972 (2007).

[7] A. Rodriguez and A. Laio, Science **344**, 1492 (2014).

[8] C. Fraley and A. E. Raftery, J. Am. Stat. Assoc. **97**, 611 (2002).

[9] A. Nobile and A. T. Fearnside, Stat. Comput. **17**, 147 (2007).

[10] J. Corander, M. Gyllenberg, and T. Koski, Adv. Data Anal. Classi. **3**, 3 (2009).

[11] M. Mézard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).

[12] K. Rose, E. Gurewitz, and G. C. Fox, Phys. Rev. Lett. **65**, 945 (1990).

[13] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).

[14] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, Vol. 9 (World Scientific Publishing, Singapore, 1987).

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 2012).

[16] L. Kozachenko and N. N. Leonenko, Probl. Inf. Transm. **23**, 9 (1987).

[17] B. Rennie and A. Dobson, J. Comb. Theory **7**, 116 (1969).

[18] S. Dasgupta, in *Proceedings of the 40th Annual Symposium on Foundations of Computer Science* (IEEE, Los Alamitos, CA, 1999), pp. 634–644.

[19] T. T. Cai, T. Liang, and H. H. Zhou, J. Multivar. Anal. **137**, 161 (2015).

[20] D. Dheeru and E. Karra Taniskidou, *UCI Machine Learning Repository* (University of California, Irvine, School of Information and Computer Sciences, 2017), http://archive.ics.uci.edu/ml.

[21] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, in *Biomedical Image Processing and Biomedical Visualization*, Vol. 1905 (International Society for Optics and Photonics, San Jose, CA, 1993), pp. 861–871.

[22] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, Vol. 1 (Springer, Berlin, 2001).

[23] S. Jung, J. S. Marron *et al.*, Ann. Stat. **37**, 4104 (2009).

[24] N. G. De Bruijn, *Asymptotic Methods in Analysis* (Dover, New York, 1981).

[25] B. Djokić, M. Miyakawa, S. Sekiguchi, I. Semba, and I. Stojmenović, Comput. J. **32**, 281 (1989).