IOP Publishing Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **53** (2020) 034001 (32pp) https://doi.org/10.1088/1751-8121/ab59af

Replica analysis of Bayesian data clustering

Alexander Mozeika[1] and **Anthony C C Coolen**[2,3]

[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
[3] London Institute for Mathematical Sciences, 35a South St, Mayfair, London, W1K 2XF, United Kingdom

E-mail: alexander.mozeika@kcl.ac.uk and ton.coolen@kcl.ac.uk

Received 18 December 2018, revised 15 November 2019
Accepted for publication 20 November 2019
Published 23 December 2019

Abstract

We use statistical mechanics to study model-based Bayesian data clustering. In this approach, each partition of the data into clusters is regarded as a microscopic system state, the negative data log-likelihood gives the energy of each state, and the data set realisation acts as disorder. Optimal clustering corresponds to the ground state of the system, and is hence obtained from the free energy via a low 'temperature' limit. We assume that for large sample sizes the free energy density is self-averaging, and we use the replica method to compute the asymptotic free energy density. The main order parameter in the resulting (replica symmetric) theory, the distribution of the data over the clusters, satisfies a self-consistent equation which can be solved by a population dynamics algorithm. From this order parameter one computes the average free energy, and all relevant macroscopic characteristics of the problem. The theory describes numerical experiments perfectly, and gives a significant improvement over the mean-field theory that was used to study this model in past.

Keywords: clustering, Bayesian inference, replica

(Some figures may appear in colour only in the online journal)

1. Introduction

Analytical tools of statistical mechanics are nowadays applied widely to statistical inference problems (see e.g. [1] and references therein). The central object of study in parameter inference is an expression for the likelihood of the data, which encodes information about the model that generated the data and the sampling process. The traditional maximum likelihood (ML) method infers model parameters from the data, but is often intractable (see e.g. [2]) or can lead to overfitting [3]. The Bayesian framework represents a more rigorous approach to parameter inference. It requires assumptions about the 'prior probability' of model parameters, and expresses the 'posterior probability' of the parameters, given the data, in terms of the data likelihood. In the so-called maximum *a posteriori* probability (MAP) method, one computes the most probable parameters, according to the posterior probability. MAP cures overfitting in ML partially by providing a 'regulariser' [1]. Both ML and MAP methods can be seen as optimisation problems, in which the data likelihood and posterior parameter probability, respectively, play the role of the objective function. With a trivial sign change this objective function can be mapped into an 'energy' function to be minimised, so that ML and MAP parameter inference can both equivalently be seen as computing a ground state in statistical mechanics [4, 5].

Clustering is a popular type of inference where one seeks to allocate statistically similar data points to the same category (or cluster), in an unsupervised way. It is used in astrophysics [6], biology [7], and many other areas. The assumed data likelihood in ML and Bayesian model-based clustering methods is usually a Gaussian Mixture Model (GMM) [6, 8]. The GMM likelihood, however, is analytically intractable, and one hence tends to resort to variational approximations [8] or computationally intensive Monte Carlo methods [9]. Furthermore, the number of model parameters, in particular the number of partitions of the data, is *extensive*, even if we fix the dimension of the data to be finite, which leads to additional difficulties [10].

For this reason, not many analytical results are available for model-based clustering (MBC), leaving mostly (many) numerical studies. Here, even when the number of parameters d is kept finite, the matrix of 'allocation' variables $\mathbf{C}$ [8] which we ultimately want to infer is growing with the sample size N. The situation is complicated further if, in addition to $\mathbf{C}$, we are also inferring the number of true clusters K. In the GMM approach, the number of clusters is usually found by adding a 'penalty' term to the log-likelihood function, such as for the Bayesian information criterion (BIC) or the integrated complete-data likelihood (ICL) [11]. These penalty based approaches sometimes lead to conflicting results [6].

A direct solution to the above problems is to follow the approach of statistical mechanics and compute the partition function [4, 5]. This approach is usually not pursued by statisticians, and in this case has not yet been pursued fully by physicists either (in spite of their familiarity with such calculations). Popular Machine Learning textbooks written by physicists, such as [8] or the more recent [12], cover only the (algorithmic) variational mean-field approach for the case when K is unknown, and the (non-Bayesian) expectation-maximisation algorithm for the case when K is known. Most statistical mechanics approaches to data clustering [13–15] use some heuristic measure of data dissimilarity as an energy function, rather than an actual statistical model of the data, or limit themselves to the simple case of assuming only two clusters [16–18] in the high dimensional regime where $d \to \infty$ and $N \to \infty$, with d/N finite.

The work of [17] and [18] is mainly concerned with the inference of parameters of two isotropic Gaussians from a balanced sample, i.e. a very restricted model of the data which does not take into account correlations, different cluster sizes, data with more than two clusters, etc. The former is concerned with the inference of the centres of the assumed Gaussians, and the latter with finding a single 'direction' in the data. Hence both studies do not formally address the MBC problem. Furthermore, in [16] the Bayesian approach is used to infer 'prototype vectors', such as centres of Gaussians, etc, of the same dimension as the data, so also this work is not addressing the MBC problem systematically either. Finally, we note that none of the above papers refer to previous work on MBC in the low-dimensional regime of finite d and $N \to \infty$. To our knowledge, only one study considers the high-dimensional regime of a specific Bayesian GMM clustering problem, namely [19]. A systematic statistical mechanical treatment of the Bayesian clustering problem is still lacking.

In this paper we consider a more general model-based Bayesian clustering protocol, which allows for simultaneous inference of the number of clusters in the data and their components, based on stochastic partitions of the data (SPD) [20]. SPD assumes priors on the partitions to compute the MAP estimate of data partitions. The mean-field (MF) theory of Bayesian SPD inference was developed recently in [21]. That study used the negative log-likelihood as the energy function, and computed its average over the data and the partitions. It led to a simple and intuitive analytical framework, which makes non-trivial predictions about low energy states and the corresponding (MAP) data partitions. However, these predictions are only correct in the regime of 'weak' correlations [21]. In this paper we pursue a full statistical mechanical treatment of the Bayesian clustering problem covering *all* correlation regimes. To this end we analyse the free energy, and we use the replica method [22] to compute its average over the data. This, unlike MF, allows us to compute the average energy of the *optimal* partitions. Furthermore, the present analysis produces a simple algorithmic framework, with the population dynamics [22] clustering algorithm at its heart, for the *simultaneous* inference of the number of clusters in the data and their components. This can be seen as a first *non-variational* result for this type of problems [8].

2. Model of the data and Bayesian cluster inference

Let us assume that we observe a data sample $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, where $\mathbf{x}_i\in\mathbb{R}^d$ for all i. Each vector $\mathbf{x}_i$ are assumed to have been generated independently from one of K distributions, which are members of a parametrized family $P(\mathbf{x}|\boldsymbol{\theta})$. M_1 data-points are sampled from $P(\mathbf{x}|\boldsymbol{\theta}_1)$, with parameter $\boldsymbol{\theta}_1$, M_2 data-points are sampled from $P(\mathbf{x}|\boldsymbol{\theta}_2)$, etc. We clearly have the constraint $\sum_{\mu=1}^K M_\mu=N$, and we assume that $M_\mu\geqslant 1$ for all μ. We will say that $\mathbf{x}_i$ (or its index i) belongs to 'cluster' μ if $\mathbf{x}_i$ was sampled from $P(\mathbf{x}|\boldsymbol{\theta}_\mu)$. The above sampling scenario can be described by the following distribution:

$$P(\mathbf{X}|\mathbf{C},K,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)=\prod_{\mu=1}^{K}\prod_{i=1}^{N}P^{c_{i\mu}}(\mathbf{x}_i|\boldsymbol{\theta}_\mu) \tag{1}$$

which is parametrised by the the partition matrix, or 'allocation' matrix [8], $\mathbf{C}$. Each element of this matrix $[\mathbf{C}]_{i\mu}=c_{i\mu}$ computes an indicator function $\mathbb{1}\left[\mathbf{x}_i\sim P(\mathbf{x}|\boldsymbol{\theta}_\mu)\right]$, i.e. is nonzero if and only if $\mathbf{x}_i$ is sampled from $P(\mathbf{x}|\boldsymbol{\theta}_\mu)$. Furthermore, we have $\sum_{\mu\leqslant K}c_{i\mu}=1$ for all $i\in[N]$[4], i.e. $\mathbf{x}_i$ belongs to only one cluster, and $M_\mu(\mathbf{C})=\sum_{i\leqslant N}c_{i\mu}\geqslant 1$ for all $\mu\in[K]$, i.e. empty clusters are not allowed[5].

Suppose we now want to infer the partition matrix $\mathbf{C}$ and the number of clusters K. The Bayesian approach to this problem (see e.g. [8]) would be to assume prior distributions for parameters and partitions, $P(\boldsymbol{\theta}_\mu)$ and $P(\mathbf{C},K)=P(\mathbf{C}|K)P(K)$[6], and to consider subsequently the posterior distribution

[4] Throughout this paper the notation $[N]$ will be used to represent the set $\{1,\ldots,N\}$.

[5] We note that the distribution (1) could be also defined by using set notation, see e.g. [21].

[6] The simplest route, following the 'principle of insufficient reason', is to choose uniform $P(\mathbf{C}|K)$ and $P(K)$. The former is then given by $P(\mathbf{C}|K)=1/K!\,\mathcal{S}(N,K)$, where $\mathcal{S}(N,K)$ is the Stirling number of the second kind ($\mathcal{S}(N,K)\simeq K^N/K!$ as $N\to\infty$ [23]), and the latter is given by $P(K)=1/N$.

$$\begin{aligned} P(\mathbf{C},K|\mathbf{X}) &= \frac{P(\mathbf{X}|\mathbf{C},K)P(\mathbf{C}|K)P(K)}{\sum_{\tilde{K}=1}^{N} P(\tilde{K}) \sum_{\tilde{\mathbf{C}}} P(\mathbf{X}|\tilde{\mathbf{C}},\tilde{K})P(\tilde{\mathbf{C}}|\tilde{K})} \\ &= \frac{\mathrm{e}^{-N\hat{F}_N(\mathbf{C},\mathbf{X})}P(\mathbf{C}|K)P(K)}{\sum_{\tilde{K}=1}^{N} P(\tilde{K}) \sum_{\tilde{\mathbf{C}}} \mathrm{e}^{-N\hat{F}_N(\tilde{\mathbf{C}},\mathbf{X})}P(\tilde{\mathbf{C}}|\tilde{K})}, \end{aligned} \tag{2}$$

where we have defined the log-likelihood density

$$\hat{F}_N(\mathbf{C},\mathbf{X}) = -\frac{1}{N}\sum_{\mu=1}^{K} \log \left\langle \mathrm{e}^{\sum_{i=1}^{N} c_{i\mu} \log P(\mathbf{x}_i|\boldsymbol{\theta}_\mu)} \right\rangle_{\boldsymbol{\theta}_\mu} \tag{3}$$

and the short-hand $\langle f(\boldsymbol{\theta}_\mu)\rangle_{\boldsymbol{\theta}_\mu} = \int \mathrm{d}\boldsymbol{\theta}_\mu \; P(\boldsymbol{\theta}_\mu)f(\boldsymbol{\theta}_\mu)$. Expression (2) can be used to infer the most probable partition $\mathbf{C}$ [21]. For each $K \leqslant N$ we can compute

$$\begin{aligned} \hat{\mathbf{C}}\,|K &= \mathrm{argmax}_{\mathbf{C}}\, P(\mathbf{C}|\mathbf{X},K) \\ &= \mathrm{argmax}_{\mathbf{C}} \left[\mathrm{e}^{-N\hat{F}_N(\mathbf{C},\mathbf{X})}P(\mathbf{C}|K)\right] \end{aligned} \tag{4}$$

and the MAP estimator

$$\begin{aligned} (\hat{\mathbf{C}},\hat{K}) &= \mathrm{argmax}_{\mathbf{C},K}\, P(\mathbf{C},K|\mathbf{X}) \\ &= \mathrm{argmax}_{\mathbf{C},K} \left[\mathrm{e}^{-N\hat{F}_N(\mathbf{C},\mathbf{X})}P(\mathbf{C}|K)P(K)\right]. \end{aligned} \tag{5}$$

Furthermore, we can use (2) to compute the distribution of cluster sizes

$$P(K|\mathbf{X}) = \frac{\mathrm{e}^{-Nf_N(K,\mathbf{X})}P(K)}{\sum_{\tilde{K}=1}^{N} P(\tilde{K})\,\mathrm{e}^{-Nf_N(\tilde{K},\mathbf{X})}}, \tag{6}$$

where

$$f_N(K,\mathbf{X}) = -\frac{1}{N}\log\left[\sum_{\mathbf{C}} \mathrm{e}^{-N\hat{F}_N(\mathbf{C},\mathbf{X})}P(\mathbf{C}|K)\right]. \tag{7}$$

3. Statistical mechanics and replica approach

3.1. Size independent identities

When the prior $P(\mathbf{C},K) = P(\mathbf{C}|K)P(K)$ is chosen to be uniform[7], MAP inference of clusters and cluster numbers according to (4) and (5) requires finding the minimum $\min_{\mathbf{C}} \hat{F}_N(\mathbf{C},\mathbf{X})$ of the negative log-likelihood (3), which is a function of the data $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)$. Here we assume that $\mathbf{X}$ is sampled from the distribution

$$q(\mathbf{X}|L) = \sum_{\mathbf{C}} q(\mathbf{C}|L)\left\{\prod_{\nu=1}^{L}\prod_{i=1}^{N} q_\nu^{c_{i\nu}}(\mathbf{x}_i)\right\}, \tag{8}$$

where $q(\mathbf{C}|L)$ and $q_\nu(\mathbf{x})$ are, respectively, the 'true' distribution of partitions, of size L, and the true distribution of data in these partitions. We note that the above expression will generally differ from the form (2), which allows to study various scenarios describing 'mismatch' between the assumed model and the actual data.

[7] For non-uniform $P(\mathbf{C}|K)$ we have to minimise $\hat{F}_N(\mathbf{C},\mathbf{X}) - N^{-1}\log P(\mathbf{C}|K)$ instead of $\hat{F}_N(\mathbf{C},\mathbf{X})$.

The minimum of $\hat{F}_N(\mathbf{C},\mathbf{X})$ can be computed within the statistical mechanics framework (see e.g. [5]), via the zero 'temperature' limit of the 'free energy' (density), using $\min_{\mathbf{C}} \hat{F}_N(\mathbf{C},\mathbf{X}) = \lim_{\beta\to\infty} f_N(\beta,\mathbf{X})$, with

$$f_N(\beta,\mathbf{X}) = -\frac{1}{\beta N}\log\sum_{\mathbf{C}} \mathrm{e}^{-\beta N\hat{F}_N(\mathbf{C},\mathbf{X})}. \tag{9}$$

Although the free energy $f_N(\beta,\mathbf{X})$ is a function of the randomly generated data $\mathbf{X}$, we expect that in the thermodynamic limit $N\to\infty$, i.e. for inference with an infinite amount of data, it will be self-averaging, i.e. $\lim_{N\to\infty}\left\{\langle f_N^2(\beta,\mathbf{X})\rangle_{\mathbf{X}} - \langle f_N(\beta,\mathbf{X})\rangle_{\mathbf{X}}^2\right\} = 0$. This implies that instead of (9) we can work with the average free energy density

$$f_N(\beta) = -\frac{1}{\beta N}\Big\langle \log\sum_{\mathbf{C}} \mathrm{e}^{-\beta N\hat{F}_N(\mathbf{C},\mathbf{X})}\Big\rangle_{\mathbf{X}}, \tag{10}$$

where the average $\langle\cdots\rangle_{\mathbf{X}}$ is generated by the distribution $q(\mathbf{X}|L)$. We note that if the prior $P(\mathbf{C}|K)$ is uniform, i.e. $P(\mathbf{C}|K) = 1/K!\,\mathcal{S}(N,K)$, then $f_N(\beta)$ is equivalent to

$$f_N(\beta) = -\frac{1}{\beta N}\Big\langle \log\sum_{\mathbf{C}} P(\mathbf{C}|K)\mathrm{e}^{-\beta N\hat{F}_N(\mathbf{C},\mathbf{X})}\Big\rangle_{\mathbf{X}} + \phi_N(\beta) \tag{11}$$

with $\phi_N(\beta) = -\frac{1}{\beta N}\log[K!\mathcal{S}(N,K)]$. The replica identity $\langle\log z\rangle = \lim_{n\to0} n^{-1}\log\langle z^n\rangle$ allows us to write the relevant part of the average free energy density as

$$f_N(\beta) - \phi_N(\beta) = -\lim_{n\to0}\frac{1}{\beta Nn}\log\Big\langle\Big[\sum_{\mathbf{C}} P(\mathbf{C}|K)\mathrm{e}^{-\beta N\hat{F}_N(\mathbf{C},\mathbf{X})}\Big]^n\Big\rangle_{\mathbf{X}}. \tag{12}$$

The standard route for computing averages via the replica method [22] is to evaluate the above for integer n, following by taking $n\to0$ via analytical continuation. So

$$\begin{aligned}\Big\langle\Big[\sum_{\mathbf{C}} P(\mathbf{C}|K)\mathrm{e}^{-\beta N\hat{F}_N(\mathbf{C},\mathbf{X})}\Big]^n\Big\rangle_{\mathbf{X}} &= \sum_{\mathbf{C}^1}\cdots\sum_{\mathbf{C}^n}\Big[\prod_{\alpha=1}^n P(\mathbf{C}^\alpha|K)\Big]\Big\langle \mathrm{e}^{-\beta N\sum_{\alpha=1}^n \hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})}\Big\rangle_{\mathbf{X}}\\ &= \Big\langle\Big\langle \mathrm{e}^{-\beta N\sum_{\alpha=1}^n \hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})}\Big\rangle_{\{\mathbf{C}^\alpha\}}\Big\rangle_{\mathbf{X}},\end{aligned} \tag{13}$$

where the average $\langle\cdots\rangle_{\{\mathbf{C}^\alpha\}}$ refers to the replicated distribution $\prod_{\alpha=1}^n P(\mathbf{C}^\alpha|K)$. We next compute the average over $\mathbf{X}$ (see appendix A for details) which leads us to the following integral

$$\Big\langle\Big\langle \mathrm{e}^{-\beta N\sum_{\alpha=1}^n \hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})}\Big\rangle\Big\rangle_{\{\mathbf{C}^\alpha\},\mathbf{X}} = \int\{\mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}}\,\mathrm{d}A\,\mathrm{d}\hat{A}\}\,\mathrm{e}^{N\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}]}, \tag{14}$$

with

$$\begin{aligned}\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}] = {}&\mathrm{i}\sum_{\alpha=1}^n\sum_{\mu=1}^K\int\mathrm{d}\mathbf{x}\,\hat{Q}_\mu^\alpha(\mathbf{x})Q_\mu^\alpha(\mathbf{x}) + \mathrm{i}\sum_{\nu,\boldsymbol{\mu}}\hat{A}(\nu,\boldsymbol{\mu})A(\nu,\boldsymbol{\mu})\\ &+\beta\sum_{\alpha=1}^n\sum_{\mu=1}^K\frac{1}{N}\log\langle\mathrm{e}^{N\int\mathrm{d}\mathbf{x}\,Q_\mu^\alpha(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)}\rangle_{\boldsymbol{\theta}_\mu}\\ &+\sum_{\nu,\boldsymbol{\mu}}A(\nu,\boldsymbol{\mu})\log\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\,\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^n\hat{Q}_{\mu_\alpha}^\alpha(\mathbf{x})}\\ &+\frac{1}{N}\log\Big\langle\mathrm{e}^{-\mathrm{i}N\sum_{\nu,\boldsymbol{\mu}}\hat{A}(\nu,\boldsymbol{\mu})A(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\})}\Big\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}},\end{aligned} \tag{15}$$

where the average $\langle\cdots\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}}$ refers to the distribution $q(\mathbf{C}|L)\prod_{\alpha=1}^{n}P(\mathbf{C}^\alpha|K)$. Finally, using the above result in our formula for the average free energy (12) gives us

$$f_N(\beta)=-\lim_{n\to0}\frac{1}{\beta Nn}\log\int\{\mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}}\,\mathrm{d}A\,\mathrm{d}\hat{A}\}\,\mathrm{e}^{N\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}]}+\phi_N(\beta). \tag{16}$$

3.2. Inference for large N

For finite N, equation (16) is as complicated as its predecessor (11). The former can, however, be computed via saddle-point integration when $N\to\infty$, provided we are allowed to take this limit first and the replica limit $n\to0$ later. Now we obtain

$$f(\beta)=-\frac{1}{\beta}\lim_{n\to0}\frac{1}{n}\mathrm{extr}_{\{\mathbf{Q},\hat{\mathbf{Q}},A,\hat{A}\}}\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}]+\phi(\beta), \tag{17}$$

where $\phi(\beta)=\lim_{N\to\infty}\phi_N(\beta)$. The further calculation requires knowledge of the average in the last term of the functional (15), which can be written in the form

$$\left\langle\mathrm{e}^{-\mathrm{i}N\sum_{\nu,\boldsymbol{\mu}}\hat{A}(\nu,\boldsymbol{\mu})A(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\})}\right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}}=\sum_{\{N(\nu,\boldsymbol{\mu})\}}P_N\left[\{N(\nu,\boldsymbol{\mu})\}\right]\mathrm{e}^{-\mathrm{i}\sum_{\nu,\boldsymbol{\mu}}\hat{A}(\nu,\boldsymbol{\mu})N(\nu,\boldsymbol{\mu})}, \tag{18}$$

where the set of variables $\{N(\nu,\boldsymbol{\mu})\}$, which are governed by the distribution

$$P_N\left[\{N(\nu,\boldsymbol{\mu})\}\right]=\sum_{\mathbf{C}}\sum_{\{\mathbf{C}^\alpha\}}q(\mathbf{C}|L)\left\{\prod_{\alpha=1}^{n}p(\mathbf{C}^\alpha|K)\right\}\prod_{\nu,\boldsymbol{\mu}}\delta_{N(\nu,\boldsymbol{\mu});NA(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\})}, \tag{19}$$

are subject to the hard constraints $\sum_{\nu,\boldsymbol{\mu}}N(\nu,\boldsymbol{\mu})=N$ (the sample size), $\sum_{\boldsymbol{\mu}}N(\nu,\boldsymbol{\mu})=N(\nu)$ (the sample size of a data generated from $q_\nu(\mathbf{x})$), and $\sum_{\nu,\boldsymbol{\mu}\setminus\mu_\alpha}N(\nu,\boldsymbol{\mu})=N(\mu_\alpha)>0$ (the size of the cluster μ_α in replica α). To compute the average (18) we will assume that for $N\to\infty$ the distribution $P_N\left[\{N(\nu,\boldsymbol{\mu})\}\right]$ approaches the associated (soft constrained) multinomial distribution

$$\tilde{P}_N\left[\{N(\nu,\boldsymbol{\mu})\}\right]=\frac{N!}{\prod_{\nu,\boldsymbol{\mu}}N(\nu,\boldsymbol{\mu})!}\prod_{\nu,\boldsymbol{\mu}}\tilde{A}(\nu,\boldsymbol{\mu})^{N(\nu,\boldsymbol{\mu})}, \tag{20}$$

where $\sum_{\nu,\boldsymbol{\mu}}\tilde{A}(\nu,\boldsymbol{\mu})=1$ and $\tilde{A}(\nu,\boldsymbol{\mu})>0$. In this case we would find simply

$$\left\langle\mathrm{e}^{-\mathrm{i}N\sum_{\nu,\boldsymbol{\mu}}\hat{A}(\nu,\boldsymbol{\mu})A(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\})}\right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}}=\left\{\sum_{\nu,\boldsymbol{\mu}}\tilde{A}(\nu,\boldsymbol{\mu})\,\mathrm{e}^{-\mathrm{i}\hat{A}(\nu,\boldsymbol{\mu})}\right\}^N. \tag{21}$$

The above assumption can by justified by the following large deviations argument.

3.3. Particle gas representation of replicated partitions

The multinomial distribution (20) describes n copies, i.e. replicas, of N 'particles' distributed over K reservoirs. For $\mathbf{A}=(\mathbf{a}_1,\ldots,\mathbf{a}_N)$ this distribution is given by

$$P(\mathbf{A})=\prod_{i=1}^{N}P(\mathbf{a}_i), \tag{22}$$

where $P(\mathbf{a}_i)=\tilde{A}(\nu,\boldsymbol{\mu})=\mathrm{Prob}(a_i(1)=\nu,a_i(2)=\mu_1,\ldots,a_i(n+1)=\mu_n)$ denotes the probability that a particle i has 'colour' $\nu\in[L]$ and is in 'reservoir' $\mu_1\in[K]$ of replica $n=1$, reservoir $\mu_2\in[K]$ of replica $n=2$, etc. The state $\mathbf{A}$ of this 'gas' of particles is a 'partition' if the reservoirs are not empty, i.e. if $N^{\alpha}_{\mu_\alpha}(\mathbf{A})=\sum_{i\leqslant N}\delta_{\mu_\alpha;a_i(\alpha+1)}>0$ for all α and μ_α. If $\mathbf{A}$ is sampled from the distribution $P(\mathbf{A})$, this will happen with high probability as $N\to\infty$ if the marginal $\tilde{A}(\mu_\alpha)=\sum_{\nu,\boldsymbol{\mu}\setminus\mu_\alpha}\tilde{A}(\nu,\boldsymbol{\mu})>0$. To show this we first compute the average $\left\langle N^{\alpha}_{\mu_\alpha}(\mathbf{A})\right\rangle_{\mathbf{A}}=\sum_{\mathbf{A}}P(\mathbf{A})\,N^{\alpha}_{\mu_\alpha}(\mathbf{A})$:

$$\begin{aligned}\left\langle N^{\alpha}_{\mu_\alpha}(\mathbf{A})\right\rangle_{\mathbf{A}}&=\sum_{i=1}^{N}\sum_{\mathbf{a}_i}P(\mathbf{a}_i)\,\delta_{\mu_\alpha;a_i(\alpha+1)}\\&=N\sum_{\mathbf{a}_i}P(\mathbf{a}_i)\delta_{\mu_\alpha;a_i(\alpha+1)}=N\sum_{\nu,\boldsymbol{\mu}\setminus\mu_\alpha}\tilde{A}(\nu,\boldsymbol{\mu})=N\tilde{A}(\mu_\alpha).\end{aligned}\tag{23}$$

Thus the average $\left\langle N^{\alpha}_{\mu_\alpha}(\mathbf{A})\right\rangle_{\mathbf{A}}>0$. Secondly, for $\epsilon>0$ we consider the probability of observing the event $N^{\alpha}_{\mu_\alpha}(\mathbf{A})\notin(N(\tilde{A}(\mu_\alpha)-\epsilon),N(\tilde{A}(\mu_\alpha)+\epsilon))$. Clearly,

$$\begin{aligned}&\mathrm{Prob}\left(N^{\alpha}_{\mu_\alpha}(\mathbf{A})\notin(N(\tilde{A}(\mu_\alpha)-\epsilon),N(\tilde{A}(\mu_\alpha)+\epsilon))\right)\\&=\mathrm{Prob}\left(N^{\alpha}_{\mu_\alpha}(\mathbf{A})/N\leqslant\tilde{A}(\mu_\alpha)-\epsilon\right)+\mathrm{Prob}\left(N^{\alpha}_{\mu_\alpha}(\mathbf{A})/N\geqslant\tilde{A}(\mu_\alpha)+\epsilon\right).\end{aligned}\tag{24}$$

For any $\lambda>0$, the second term can be bounded using Markov's inequality, as follows

$$\begin{aligned}\mathrm{Prob}\left(N^{\alpha}_{\mu_\alpha}(\mathbf{A})/N\geqslant\tilde{A}(\mu_\alpha)+\epsilon\right)&=\mathrm{Prob}\left(\mathrm{e}^{\lambda N^{\alpha}_{\mu_\alpha}(\mathbf{A})}\geqslant\mathrm{e}^{\lambda N(\tilde{A}(\mu_\alpha)+\epsilon)}\right)\\&\leqslant\langle\mathrm{e}^{\lambda N^{\alpha}_{\mu_\alpha}(\mathbf{A})}\rangle_{\mathbf{A}}\,\mathrm{e}^{-\lambda N(\tilde{A}(\mu_\alpha)+\epsilon)},\end{aligned}\tag{25}$$

with the average

$$\begin{aligned}\langle\mathrm{e}^{\lambda N^{\alpha}_{\mu_\alpha}(\mathbf{A})}\rangle_{\mathbf{A}}&=\sum_{\mathbf{A}}P(\mathbf{A})\,\mathrm{e}^{\lambda N^{\alpha}_{\mu_\alpha}(\mathbf{A})}=\prod_{i=1}^{N}\left\{\sum_{\mathbf{a}_i}P(\mathbf{a}_i)\,\mathrm{e}^{\lambda\delta_{\mu_\alpha;a_i(\alpha+1)}}\right\}\\&=\left[1+\tilde{A}(\mu_\alpha)(\mathrm{e}^{\lambda}-1)\right]^{N}.\end{aligned}\tag{26}$$

Hence

$$\mathrm{Prob}\left(N^{\alpha}_{\mu_\alpha}(\mathbf{A})\geqslant N(\tilde{A}(\mu_\alpha)+\epsilon)\right)\leqslant\mathrm{e}^{-N\mathrm{I}(\lambda,\epsilon)},\tag{27}$$

where $\mathrm{I}(\lambda,\epsilon)=-\log(1+\tilde{A}(\mu_\alpha)(\mathrm{e}^{\lambda}-1))+\lambda(\tilde{A}(\mu_\alpha)+\epsilon)$ is a *rate function*. The latter has its maximum at $\lambda^{*}=\log[(\tilde{A}(\mu_\alpha)^2+\tilde{A}(\mu_\alpha)\epsilon-\tilde{A}(\mu_\alpha)-\epsilon)/(\tilde{A}(\mu_\alpha)(\tilde{A}(\mu_\alpha)-1+\epsilon)]$, and $\mathrm{I}(\lambda^{*},\epsilon)=D(\tilde{A}(\mu_\alpha)+\epsilon\,||\tilde{A}(\mu_\alpha))$, where $D(p\,||q)=p\log(\frac{p}{q})+(1-p)\log(\frac{1-p}{1-q})\geqslant0$ is the Kullback–Leibler divergence [24] of binary distributions with probabilities $p,q\in[0,1]$. We may now write

$$\mathrm{Prob}\left(N^{\alpha}_{\mu_\alpha}(\mathbf{A})\geqslant N(\tilde{A}(\mu_\alpha)+\epsilon)\right)\leqslant\mathrm{e}^{-ND(\tilde{A}(\mu_\alpha)+\epsilon\,||\tilde{A}(\mu_\alpha))}.\tag{28}$$

Following similar steps to bound the first term of (24) gives us also the inequality

$$\mathrm{Prob}\left(N^{\alpha}_{\mu_\alpha}(\mathbf{A})\leqslant N(\tilde{A}(\mu_\alpha)-\epsilon)\right)\leqslant\mathrm{e}^{-ND(\tilde{A}(\mu_\alpha)-\epsilon\,||\tilde{A}(\mu_\alpha))}.\tag{29}$$

In combination, our two bounds directly lead to

$$\begin{aligned}&\mathrm{Prob}\left(N^{\alpha}_{\mu_\alpha}(\mathbf{A})\notin(N(\tilde{A}(\mu_\alpha)-\epsilon),N(\tilde{A}(\mu_\alpha)+\epsilon))\right)\\&\leqslant2\,\mathrm{e}^{-N\min_{\sigma\in\{-1,1\}}D(\tilde{A}(\mu_\alpha)+\sigma\epsilon\,||\tilde{A}(\mu_\alpha))}.\end{aligned}\tag{30}$$

The probability for one or more of the events $N^{\alpha}_{\mu_\alpha}(\mathbf{A}) \notin (N(\tilde{A}(\mu_\alpha)-\epsilon), N(\tilde{A}(\mu_\alpha)+\epsilon))$ to occur (of which there are nK) can be bounded using Boole's inequality in combination with (30), as follows

$$\begin{aligned} &\text{Prob}\left(\cup_{\alpha,\mu_\alpha}\left\{N^{\alpha}_{\mu_\alpha}(\mathbf{A}) \notin (N(\tilde{A}(\mu_\alpha)-\epsilon), N(\tilde{A}(\mu_\alpha)+\epsilon))\right\}\right) \\ &\leqslant \sum_{\alpha=1}^{n}\sum_{\mu_\alpha=1}^{K} \text{Prob}\left(N^{\alpha}_{\mu_\alpha}(\mathbf{A}) \notin (N(\tilde{A}(\mu_\alpha)-\epsilon)), N(\tilde{A}(\mu_\alpha)+\epsilon))\right) \\ &\leqslant 2nK\,\mathrm{e}^{-N\min_{\alpha,\mu_\alpha}\min_{\sigma\in\{-1,1\}} D(\tilde{A}(\mu_\alpha)+\sigma\epsilon\,||\tilde{A}(\mu_\alpha))}. \end{aligned} \tag{31}$$

We conclude that for $N \to \infty$ the deviations of the random variables $N^{\alpha}_{\mu_\alpha}(\mathbf{A})$ from their averages $N\tilde{A}(\mu_\alpha)$ decay exponentially with N.

Let us next consider the entropy density

$$\begin{aligned} H(\mathbf{A})/N &= -\sum_{\mathbf{a}_i} P(\mathbf{a}_i)\log P(\mathbf{a}_i) = -\sum_{\nu,\boldsymbol{\mu}} \tilde{A}(\nu,\boldsymbol{\mu})\log\tilde{A}(\nu,\boldsymbol{\mu}) \\ &= -\sum_{\nu}\tilde{A}(\nu)\log\tilde{A}(\nu) - \sum_{\nu,\boldsymbol{\mu}}\tilde{A}(\nu)\tilde{A}(\boldsymbol{\mu}|\nu)\log\tilde{A}(\boldsymbol{\mu}|\nu). \end{aligned} \tag{32}$$

If we assume that

$$\tilde{A}(\boldsymbol{\mu}|\nu) = \prod_{\alpha=1}^{n}\tilde{A}(\mu_\alpha|\nu), \tag{33}$$

then

$$H(\mathbf{A})/N = -\sum_{\nu}\tilde{A}(\nu)\log\tilde{A}(\nu) - n\sum_{\nu,\mu}\tilde{A}(\nu)\tilde{A}(\mu|\nu)\log\tilde{A}(\mu|\nu). \tag{34}$$

The entropy of the distribution $q(\mathbf{C}|L)\left\{\prod_{\alpha=1}^{n}p(\mathbf{C}^{\alpha}|K)\right\}$, used in (19), is given by

$$\begin{aligned} H(p,q)/N &= -\frac{1}{N}\sum_{\mathbf{C}}\sum_{\{\mathbf{C}^{\alpha}\}} q(\mathbf{C}|L)\left[\prod_{\alpha=1}^{n}p(\mathbf{C}^{\alpha}|K)\right]\log\left\{q(\mathbf{C}|L)\left[\prod_{\alpha=1}^{n}p(\mathbf{C}^{\alpha}|K)\right]\right\} \\ &= H(q)/N + nH(p)/N, \end{aligned} \tag{35}$$

with $H(q) = -\sum_{\mathbf{C}} q(\mathbf{C}|L)\log q(\mathbf{C}|L)$ and $H(p) = -\sum_{\mathbf{C}} p(\mathbf{C}|K)\log p(\mathbf{C}|K)$. For the case of uniform distributions $q(\mathbf{C}|L) = 1/L!\mathcal{S}(N,L)$ and $p(\mathbf{C}|K) = 1/K!\mathcal{S}(N,K)$ the latter entropies are, respectively, $\log(L!\mathcal{S}(N,L))$ and $\log(K!\mathcal{S}(N,K))$. This gives us $H(p,q)/N = \log(L) + n\log(K)$ in the limit $N \to \infty$. Comparing this asymptotic result for $H(p,q)/N$ with $H(\mathbf{A})/N$ in (34), we see that the two expressions are equal for large N when $\tilde{A}(\nu) = 1/L$ and $\tilde{A}(\mu|\nu) = 1/K$. In this case, the distribution (19) apparently approaches the multinomial distribution (20). We expect this also to be true when the distribution $q(\mathbf{C}|L)$ is uniform, but subject to the constraints $\sum_{i=1}^{N} c_{i\nu} = N\tilde{A}(\nu)$.

4. Replica symmetric theory

4.1. Simplification of the saddle-point problem

Using the assumptions (21) and (33), we obtain a simplified expression for (15):

$$\begin{aligned}\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}] = \mathrm{i}\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}\int \mathrm{d}\mathbf{x}\, \hat{Q}_{\mu}^{\alpha}(\mathbf{x})Q_{\mu}^{\alpha}(\mathbf{x})\\ +\sum_{\nu,\boldsymbol{\mu}} A(\nu,\boldsymbol{\mu})\Big[\mathrm{i}\hat{A}(\nu,\boldsymbol{\mu})+\log\int \mathrm{d}\mathbf{x}\, q_{\nu}(\mathbf{x})\,\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\hat{Q}_{\mu_\alpha}^{\alpha}(\mathbf{x})}\Big]\\ +\beta\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}\frac{1}{N}\log\left\langle \mathrm{e}^{N\int \mathrm{d}\mathbf{x}\, Q_{\mu}^{\alpha}(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu})}\right\rangle_{\boldsymbol{\theta}_{\mu}}\\ +\log\Big[\sum_{\nu,\boldsymbol{\mu}}\tilde{A}(\nu)\mathrm{e}^{-\mathrm{i}\hat{A}(\nu,\boldsymbol{\mu})}\prod_{\alpha=1}^{n}\tilde{A}(\mu_\alpha|\nu)\Big].\end{aligned} \tag{36}$$

The extrema of this functional are seen to be the solutions of the following equations:

$$\hat{A}(\nu,\boldsymbol{\mu}) = \mathrm{i}\log\int \mathrm{d}\mathbf{x}\, q_{\nu}(\mathbf{x})\,\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\hat{Q}_{\mu_\alpha}^{\alpha}(\mathbf{x})} \tag{37}$$

$$A(\nu,\boldsymbol{\mu}) = \frac{\tilde{A}(\nu)\mathrm{e}^{-\mathrm{i}\hat{A}(\nu,\boldsymbol{\mu})}\prod_{\alpha=1}^{n}\tilde{A}(\mu_\alpha|\nu)}{\sum_{\tilde{\nu},\tilde{\boldsymbol{\mu}}}\tilde{A}(\tilde{\nu})\mathrm{e}^{-\mathrm{i}\hat{A}(\tilde{\nu},\tilde{\boldsymbol{\mu}})}\prod_{\alpha=1}^{n}\tilde{A}(\tilde{\mu}_\alpha|\tilde{\nu})} \tag{38}$$

$$Q_{\mu}^{\alpha}(\mathbf{x}) = \sum_{\nu,\boldsymbol{\mu}}\delta_{\mu;\mu_\alpha}A(\nu,\boldsymbol{\mu})\frac{q_{\nu}(\mathbf{x})\,\mathrm{e}^{-\mathrm{i}\sum_{\gamma=1}^{n}\hat{Q}_{\mu_\gamma}^{\gamma}(\mathbf{x})}}{\int \mathrm{d}\tilde{\mathbf{x}}\, q_{\nu}(\tilde{\mathbf{x}})\,\mathrm{e}^{-\mathrm{i}\sum_{\gamma=1}^{n}\hat{Q}_{\mu_\gamma}^{\gamma}(\tilde{\mathbf{x}})}} \tag{39}$$

$$\hat{Q}_{\mu}^{\alpha}(\mathbf{x}) = \mathrm{i}\beta\frac{\langle \mathrm{e}^{N\int \mathrm{d}\tilde{\mathbf{x}}\, Q_{\mu}^{\alpha}(\tilde{\mathbf{x}})\log P(\tilde{\mathbf{x}}|\boldsymbol{\theta})}\log P(\mathbf{x}|\boldsymbol{\theta})\rangle_{\boldsymbol{\theta}}}{\langle \mathrm{e}^{N\int \mathrm{d}\tilde{\mathbf{x}}\, Q_{\mu}^{\alpha}(\tilde{\mathbf{x}})\log P(\tilde{\mathbf{x}}|\boldsymbol{\theta})}\rangle_{\boldsymbol{\theta}}}. \tag{40}$$

For $N\to\infty$ we can evaluate the integrals in the last equation with the Laplace method [25], giving

$$\begin{aligned}\hat{Q}_{\mu}^{\alpha}(\mathbf{x}) &= \mathrm{i}\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu}^{\alpha})\\ \boldsymbol{\theta}_{\mu}^{\alpha} &= \mathrm{argmax}_{\boldsymbol{\theta}}\int \mathrm{d}\mathbf{x}\, Q_{\mu}^{\alpha}(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta}).\end{aligned} \tag{41}$$

Upon eliminating the conjugate order parameters $\{\hat{\mathbf{Q}},\hat{A}\}$ from our coupled equations and considering large N, we obtain after some straightforward manipulations the following expression for the nontrivial part of the average free energy (17),

$$f(\beta)-\phi(\beta) = -\lim_{n\to 0}\frac{1}{\beta n}\log\left\{\sum_{\nu}\tilde{A}(\nu)\int \mathrm{d}\mathbf{x}\, q_{\nu}(\mathbf{x})\prod_{\alpha=1}^{n}\left[\sum_{\mu=1}^{K}\tilde{A}(\mu|\nu)\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu}^{\alpha})}\right]\right\} \tag{42}$$

and the following closed equations for the remaining order parameters $\{\mathbf{Q},A\}$:

$$Q_{\mu}^{\alpha}(\mathbf{x}) = \sum_{\nu,\boldsymbol{\mu}}\delta_{\mu;\mu_\alpha}A(\nu,\boldsymbol{\mu})\frac{q_{\nu}(\mathbf{x})\,\mathrm{e}^{\sum_{\gamma=1}^{n}\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu_\gamma}^{\gamma})}}{\int \mathrm{d}\tilde{\mathbf{x}}\, q_{\nu}(\tilde{\mathbf{x}})\,\mathrm{e}^{\sum_{\gamma=1}^{n}\beta\log P(\tilde{\mathbf{x}}|\boldsymbol{\theta}_{\mu_\gamma}^{\gamma})}}, \tag{43}$$

$$A(\nu,\boldsymbol{\mu})=\frac{\tilde{A}(\nu)\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\Big[\prod_{\alpha=1}^n\tilde{A}(\mu_\alpha|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}^\alpha_{\mu_\alpha})}\Big]}{\sum_{\tilde{\nu}}\tilde{A}(\tilde{\nu})\int\mathrm{d}\mathbf{x}\,q_{\tilde{\nu}}(\mathbf{x})\Big[\prod_{\alpha=1}^n\sum_{\tilde{\mu}_\alpha}\tilde{A}(\tilde{\mu}_\alpha|\tilde{\nu})\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}^\alpha_{\tilde{\mu}_\alpha})}\Big]}.\tag{44}$$

In order to take the replica limit $n\to 0$ in (42)–(44) we will make the the 'replica symmetry' (RS) assumption [22], which here translates into $Q^\alpha_{\mu_\alpha}(\mathbf{x})=Q_{\mu_\alpha}(\mathbf{x})$. It then follows from (41), in turn, that $\boldsymbol{\theta}^\alpha_\mu=\boldsymbol{\theta}_{\mu_\alpha}$. The RS structure allows us to take the replica limit (see appendix B for details) and find the following equations:

$$\begin{aligned}Q_\mu(\mathbf{x})&=\sum_\nu\tilde{A}(\nu)\,q_\nu(\mathbf{x})\frac{\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)}}{\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}}\\ \boldsymbol{\theta}_\mu&=\mathrm{argmax}_{\boldsymbol{\theta}}\int\mathrm{d}\mathbf{x}\,Q_\mu(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta})\end{aligned}\tag{45}$$

$$\begin{aligned}A(\mu|\nu)&=\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\frac{\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)}}{\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}}\\ A(\nu)&=\tilde{A}(\nu)\end{aligned}\tag{46}$$

and the asymptotic form of the average free energy

$$f(\beta)=-\frac{1}{\beta}\int\mathrm{d}\mathbf{x}\sum_{\nu=1}^L\tilde{A}(\nu)q_\nu(\mathbf{x})\log\Big[\sum_{\mu=1}^K\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)}\Big]+\phi(\beta).\tag{47}$$

The physical meaning of the order parameters $Q_\mu(\mathbf{x})$ and $A(\mu|\nu)$ becomes clear if we define the following two densities

$$Q_\mu(\mathbf{x}|\mathbf{C},\mathbf{X})=\frac{1}{N}\sum_{i=1}^N c_{i\mu}\,\delta(\mathbf{x}-\mathbf{x}_i)\tag{48}$$

$$A(\nu,\mu|\mathbf{C},\mathbf{X})=\frac{1}{N}\sum_{i=1}^N c_{i\mu}\mathbb{1}\left[\mathbf{x}_i\sim q_\nu(\mathbf{x})\right].\tag{49}$$

If we sample $\mathbf{C}$ from the Gibbs–Boltzmann distribution

$$P_\beta(\mathbf{C}|\mathbf{X})=\frac{1}{Z_\beta(\mathbf{X})}P(\mathbf{C}|K)\mathrm{e}^{-\beta N\hat{F}_N(\mathbf{C},\mathbf{X})},\tag{50}$$

where $Z_\beta(\mathbf{X})=\sum_{\mathbf{C}}P(\mathbf{C}|K)\mathrm{e}^{-\beta N\hat{F}_N(\mathbf{C},\mathbf{X})}$ is the associated partition function, and with the conditional averages $\langle G(\mathbf{C})\rangle_{\mathbf{C}|\mathbf{X}}=\sum_{\mathbf{C}}P(\mathbf{C}|K)G(\mathbf{C})$, then one finds that

$$Q_\mu(\mathbf{x})=\lim_{N\to\infty}\Big\langle\langle Q_\mu(\mathbf{x}|\mathbf{C},\mathbf{X})\rangle_{\mathbf{C}|\mathbf{X}}\Big\rangle_{\mathbf{X}},\tag{51}$$

$$A(\nu,\mu)=\lim_{N\to\infty}\left\langle\langle A(\nu,\mu|\mathbf{C},\mathbf{X})\rangle_{\mathbf{C}|\mathbf{X}}\right\rangle_{\mathbf{X}},\tag{52}$$

(see appendix C for details). So, asymptotically, $Q_\mu(\mathbf{x})$ is the average distribution of data in cluster μ, and $A(\nu,\mu)$ is the average fraction of data originating from the distribution $q_\nu(\mathbf{x})$ that are allocated by the clustering process to cluster μ.

4.2. RS theory for $\beta\to\infty$

Let us study the behaviour of the RS order parameter equations (45)–(47) in the zero temperature limit $\beta\to\infty$. First, for the order parameter $Q_\mu(\mathbf{x})$, governed by the equation (45), and any test function a_μ we consider the sum

$$\begin{aligned}\sum_\mu Q_\mu(\mathbf{x})a_\mu&=\sum_\nu\tilde{A}(\nu)\,q_\nu(\mathbf{x})\,\frac{\sum_\mu\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)}a_\mu}{\sum_{\mu''}\tilde{A}(\mu''|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu''})}}\\&=\sum_\nu\tilde{A}(\nu)\,q_\nu(\mathbf{x})\,\frac{\sum_{\mu'}\tilde{A}(\mu'|\nu)\,\mathrm{e}^{-\beta(\max_{\tilde{\mu}}\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})-\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu'}))}a_{\mu'}}{\sum_{\mu''}\tilde{A}(\mu''|\nu)\,\mathrm{e}^{-\beta(\max_{\tilde{\mu}}\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})-\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu''}))}}\\&=\sum_\nu\tilde{A}(\nu)\,q_\nu(\mathbf{x})\,\frac{\sum_{\mu'}\tilde{A}(\mu'|\nu)\,\mathrm{e}^{-\beta\Delta_{\mu'}(\mathbf{x})}a_{\mu'}}{\sum_{\mu''}\tilde{A}(\mu''|\nu)\,\mathrm{e}^{-\beta\Delta_{\mu''}(\mathbf{x})}},\end{aligned}\tag{53}$$

where $\Delta_\mu(\mathbf{x})=\max_{\tilde{\mu}}\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})-\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)$. For $\beta\to\infty$ the average will tend to

$$\lim_{\beta\to\infty}\frac{\sum_{\mu'}\tilde{A}(\mu'|\nu)\,\mathrm{e}^{-\beta\Delta_{\mu'}(\mathbf{x})}\,a_{\mu'}}{\sum_{\mu''}\tilde{A}(\mu''|\nu)\,\mathrm{e}^{-\beta\Delta_{\mu''}(\mathbf{x})}}=\frac{\sum_{\mu'}\mathbb{1}\left[\Delta_{\mu'}(\mathbf{x})=0\right]\tilde{A}(\mu'|\nu)\,a_{\mu'}}{\sum_{\mu''}\mathbb{1}\left[\Delta_{\mu''}(\mathbf{x})=0\right]\tilde{A}(\mu''|\nu)}.\tag{54}$$

Hence for $\beta\to\infty$ we may write

$$Q_\mu(\mathbf{x})=\sum_\nu\tilde{A}(\nu)\,q_\nu(\mathbf{x})\,\frac{\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)}{\sum_{\mu'}\mathbb{1}\left[\Delta_{\mu'}(\mathbf{x})=0\right]\tilde{A}(\mu'|\nu)}.\tag{55}$$

Similarly, equation (46) for the order parameter $A(\mu|\nu)$ gives us

$$\sum_\mu A(\mu|\nu)a_\mu=\int\mathrm{d}\mathbf{x}\;q_\nu(\mathbf{x})\frac{\sum_\mu\tilde{A}(\mu|\nu)\,\mathrm{e}^{-\beta\Delta_\mu(\mathbf{x})}a_\mu}{\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\nu)\,\mathrm{e}^{-\beta\Delta_{\tilde{\mu}}(\mathbf{x})}},\tag{56}$$

so for $\beta\to\infty$ we may write, assuming the expectation and limit operators commute,

$$\sum_\mu A(\mu|\nu)a_\mu=\int\mathrm{d}\mathbf{x}\;q_\nu(\mathbf{x})\frac{\sum_\mu\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)\,a_\mu}{\sum_{\tilde{\mu}}\mathbb{1}\left[\Delta_{\tilde{\mu}}(\mathbf{x})=0\right]\tilde{A}(\tilde{\mu}|\nu)}.\tag{57}$$

We note that $A(\mu)=\int\mathrm{d}\mathbf{x}\;Q_\mu(\mathbf{x})$, as a consequence of the (48) and (49). Finally, taking $\beta\to\infty$ in the average free energy density (47) gives us

$$
\begin{aligned}
&\lim_{\beta\to\infty}\Big[f(\beta)-\phi(\beta)\Big]\\
&=-\lim_{\beta\to\infty}\frac{1}{\beta}\sum_{\nu}\tilde{A}(\nu)\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x})\log\Big[\mathrm{e}^{\beta\max_{\tilde{\mu}}\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}\sum_{\mu=1}^{K}\tilde{A}(\mu|\nu)\mathrm{e}^{-\beta\Delta_\mu(\mathbf{x})}\Big]\\
&=-\sum_{\nu}\tilde{A}(\nu)\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x})\max_{\mu}\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)\\
&\quad-\lim_{\beta\to\infty}\frac{1}{\beta}\sum_{\nu}\tilde{A}(\nu)\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x})\\
&\quad\times\log\left[\sum_{\mu=1}^{K}\tilde{A}(\mu|\nu)\Big(\mathbb{1}\left[\Delta_\mu(\mathbf{x})>0\right]\mathrm{e}^{-\beta\Delta_\mu(\mathbf{x})}+\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\Big)\right]\\
&=-\int \mathrm{d}\mathbf{x}\sum_{\nu}\tilde{A}(\nu)q_\nu(\mathbf{x})\max_{\mu}\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)\\
&\quad-\lim_{\beta\to\infty}\frac{1}{\beta}\sum_{\nu}\tilde{A}(\nu)\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x})\log\Big[\sum_{\mu=1}^{K}\tilde{A}(\mu|\nu)\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\Big]\\
&\quad-\lim_{\beta\to\infty}\frac{1}{\beta}\sum_{\nu}\tilde{A}(\nu)\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x})\log\left[1+\frac{\sum_{\mu=1}^{K}\mathbb{1}\left[\Delta_\mu(\mathbf{x})>0\right]\tilde{A}(\mu|\nu)\,\mathrm{e}^{-\beta\Delta_\mu(\mathbf{x})}}{\sum_{\mu=1}^{K}\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)}\right]\\
&=-\sum_{\nu}\tilde{A}(\nu)\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x})\max_{\mu}\log P(\mathbf{x}|\boldsymbol{\theta}_\mu).
\end{aligned}
\tag{58}
$$

The average energy $e(\beta)=\lim_{N\to\infty}\langle\langle\hat{F}_N(\mathbf{C},\mathbf{X})\rangle_{\mathbf{C}|\mathbf{X}}\rangle_{\mathbf{X}}$ is given by (see appendix D)

$$
e(\beta)=-\sum_{\mu=1}^{K}\int \mathrm{d}\mathbf{x}\, Q_\mu(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta}_\mu),
\tag{59}
$$

where $Q_\mu(\mathbf{x})$ is a solution of the equation (45). The latter reduces to (55) when $\beta\to\infty$, and hence in this limit we find

$$
e(\infty)=-\sum_{\mu=1}^{K}\sum_{\nu}\tilde{A}(\nu)\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)\frac{\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)}{\sum_{\mu'}\mathbb{1}\left[\Delta_{\mu'}(\mathbf{x})=0\right]\tilde{A}(\mu'|\nu)}.
\tag{60}
$$

It is trivial to show (and intuitive) that $e(\infty)=f(\infty)$. For finite β, the average free energy $f(\beta)-\phi_N$ and the energy $e(\beta)$, given by equations (47) and (59), can be used to compute the average entropy density of the Gibbs–Boltzmann distribution (50) via the Helmholtz free energy $f(\beta)=e(\beta)-\frac{1}{\beta}s(\beta)$,

$$
s(\beta)=-\lim_{N\to\infty}\frac{1}{N}\Big\langle\sum_{\mathbf{C}}P_\beta(\mathbf{C}|\mathbf{X})\log P_\beta(\mathbf{C}|\mathbf{X})\Big\rangle_{\mathbf{X}}.
\tag{61}
$$

From the Helmholtz free energy we immediately infer that $\lim_{\beta\to\infty}s(\beta)/\beta=0$.

4.3. RS theory for $\beta\to 0$

The RS theory simplifies considerably in the high temperature limit $\beta\to 0$. Here the order parameter $Q_\mu(\mathbf{x})$, which is governed by the equation (45), is given by

$$Q_\mu(\mathbf{x}) = \sum_\nu \tilde{A}(\nu)\tilde{A}(\mu|\nu)\, q_\nu(\mathbf{x}). \tag{62}$$

The fraction of data points originating from the distribution $q_\nu(\mathbf{x})$ assigned to cluster μ, $A(\mu,\nu)$, is $\tilde{A}(\nu)\tilde{A}(\mu|\nu)$ due to (46). Using this in (59) gives the average energy

$$e(0) = -\sum_{\nu=1}^{L} \tilde{A}(\nu) \sum_{\mu=1}^{K} \tilde{A}(\mu|\nu) \int \mathrm{d}\mathbf{x}\; q_\nu(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\theta}_\mu) \tag{63}$$

where $\boldsymbol{\theta}_\mu = \mathrm{argmax}_{\boldsymbol{\theta}} \int \mathrm{d}\mathbf{x}\; Q_\mu(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\theta})$. We note that (63) is equal to

$$F(\tilde{A}) = \sum_{\nu=1}^{L} \tilde{A}(\nu) \sum_{\mu=1}^{K} \tilde{A}(\mu|\nu) D(q_\nu||P_\mu) + \sum_{\nu=1}^{L} \tilde{A}(\nu) H(q_\nu), \tag{64}$$

where $H(q_\nu)$ is the differential entropy of $q_\nu(\mathbf{x})$, which is also the entropy function of the mean-field theory [21]. For finite N, the average energy $e_N(\beta) = \langle\langle \hat{F}_N(\mathbf{C}, \mathbf{X})\rangle_{\mathbf{C}}\rangle_{\mathbf{X}}$ is a monotonic non-increasing function of β. Also the limits $\lim_{\beta\to\infty} e_N(\beta)$ and $\lim_{\beta\to 0} e_N(\beta)$ exist. Thus $e_N(\infty) \leqslant e_N(0)$ for N finite and hence the average energy $e(\infty)$ is bounded from above by the mean-field entropy $F(\tilde{A})$, i.e. $e(\infty) \leqslant F(\tilde{A})$. For model distributions $P(\mathbf{x}|\boldsymbol{\theta}_\mu)$ with non-overlapping supports for different θ_μ, this upper bound can be optimised by replacing $F(\tilde{A})$ with $\min_{\tilde{A}} F(\tilde{A})$ and hence in this case

$$e(\infty) \leqslant \min_{\tilde{A}} F(\tilde{A}). \tag{65}$$

The minimum is computed over all prior parameters $\tilde{A}(\mu|\nu)$ satisfying the constraints $\tilde{A}(\mu|\nu) > 0$ and $\sum_{\mu\leqslant K} \tilde{A}(\mu|\nu) = 1$. Finally, we note that for $K = 1$, as a consequence of $Q_\mu(\mathbf{x}) = \sum_{\nu\leqslant L} \tilde{A}(\nu)\, q_\nu(\mathbf{x})$, we will have $e(\infty) = F(\tilde{A})$.

4.4. Recovery of true partitions

Equation (55) for $Q_\mu(\mathbf{x})$ can be used to derive the following expression for the distribution $\tilde{Q}_\mu(\mathbf{x}) = Q_\mu(\mathbf{x}) / \int \mathrm{d}\tilde{\mathbf{x}}\; Q_\mu(\tilde{\mathbf{x}})$ of data that are assigned to cluster μ:

$$\begin{aligned}
\tilde{Q}_\mu(\mathbf{x}) &= \frac{\sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x})\, \frac{\mathbb{1}[\Delta_\mu(\mathbf{x})=0]\tilde{A}(\mu|\nu)}{Z_\nu(\mathbf{x})}}{\int \mathrm{d}\tilde{\mathbf{x}}\; \sum_\nu \tilde{A}(\nu)\, q_\nu(\tilde{\mathbf{x}})\, \frac{\mathbb{1}[\Delta_\mu(\tilde{\mathbf{x}})=0]\tilde{A}(\mu|\nu)}{Z_\nu(\tilde{\mathbf{x}})}} \\
\Delta_\mu(\mathbf{x}) &= \max_{\tilde{\mu}} \log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}}) - \log P(\mathbf{x}|\boldsymbol{\theta}_\mu) \\
\boldsymbol{\theta}_\mu &= \mathrm{argmax}_{\boldsymbol{\theta}} \int \mathrm{d}\mathbf{x}\; \tilde{Q}_\mu(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\theta}),
\end{aligned} \tag{66}$$

where $Z_\nu(\mathbf{x}) = \sum_\mu \mathbb{1}\left[\Delta_\mu(\mathbf{x}) = 0\right] \tilde{A}(\mu|\nu)$. Suppose we knew the number of true clusters, i.e. $K = L$. If our clustering procedure was perfect we would then expect that each cluster holds data from at most one distribution, i.e. we expect $\tilde{Q}_\mu(\mathbf{x}) = q_\mu(\mathbf{x})$ to be a solution of the following equation

$$q_\mu(\mathbf{x}) = \frac{\sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x})\, \frac{\mathbb{1}[\Delta_\mu(\mathbf{x})=0]\tilde{A}(\mu|\nu)}{Z_\nu(\mathbf{x})}}{\int \mathrm{d}\tilde{\mathbf{x}}\; \sum_\nu \tilde{A}(\nu)\, q_\nu(\tilde{\mathbf{x}})\, \frac{\mathbb{1}[\Delta_\mu(\tilde{\mathbf{x}})=0]\tilde{A}(\mu|\nu)}{Z_\nu(\tilde{\mathbf{x}})}}. \tag{67}$$

This is certainly true if $\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)=\delta_{\nu;\mu}Z_\nu(\mathbf{x})$ for all $\mathbf{x}$ in the domain of $q_\mu(\mathbf{x})$. The latter condition implies that $\int \mathrm{d}\mathbf{x}\ q_\nu(\mathbf{x})\ \mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)Z_\nu^{-1}(\mathbf{x})=\delta_{\nu;\mu}$ which, by the definition of order parameter $A(\mu|\nu)$, is equivalent to $A(\mu|\nu)=\delta_{\nu;\mu}$, i.e. all data from the distribution $q_\nu(\mathbf{x})$ are in cluster μ. Thus if

$$\int \mathrm{d}\mathbf{x}\ q_\nu(\mathbf{x})\ \frac{\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)}{Z_\nu(\mathbf{x})}=\delta_{\nu;\mu} \tag{68}$$

holds for all pairs (ν,μ) in a bijective mapping of the set $[K]$ to itself, then $\tilde{Q}_\mu(\mathbf{x})=q_\mu(\mathbf{x})$ is a solution of equation (67). Let us define the set $S_P(\mathbf{x})=\{\mu\,|\,\Delta_\mu(\mathbf{x})=0\}$ and consider the average $\sum_\mu A(\mu|\nu)\mu$:

$$\begin{aligned}
&\int \mathrm{d}\mathbf{x}\ q_\nu(\mathbf{x})\ \frac{\sum_\mu \mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)\mu}{Z_\nu(\mathbf{x})}\\
&=\int \mathrm{d}\mathbf{x}\ \Big(\mathbb{1}\left[|S_P(\mathbf{x})|>1\right]+\mathbb{1}\left[|S_P(\mathbf{x})|=1\right]\Big)q_\nu(\mathbf{x})\frac{\sum_\mu \mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)\mu}{Z_\nu(\mathbf{x})}\\
&=\int \mathrm{d}\mathbf{x}\ q_\nu(\mathbf{x})\ \mathrm{argmax}_\mu \log P(\mathbf{x}|\boldsymbol{\theta}_\mu)\\
&\qquad+\int \mathrm{d}\mathbf{x}\ \mathbb{1}\left[|S_P(\mathbf{x})|>1\right]q_\nu(\mathbf{x})\frac{\sum_\mu \mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right]\tilde{A}(\mu|\nu)\mu}{Z_\nu(\mathbf{x})}.
\end{aligned} \tag{69}$$

We note that the second term is a contribution of sets that can be characterized as $\{\mathbf{x}\,|\,P(\mathbf{x}|\boldsymbol{\theta}_{\mu_1})=P(\mathbf{x}|\boldsymbol{\theta}_{\mu_2}),\ \mu_1<\mu_2\}$, for some (μ_1,μ_2). If we assume that this term is zero[8], then one of the consequences of (68) is equivalence of the two averages

$$\sum_\mu A(\mu|\nu)\mu=\nu=\int \mathrm{d}\mathbf{x}\ q_\nu(\mathbf{x})\mathrm{argmax}_\mu \log P(\mathbf{x}|\boldsymbol{\theta}_\mu), \tag{70}$$

and

$$\begin{aligned}
\int \mathrm{d}\mathbf{x}\ q_\nu(\mathbf{x})\mathrm{argmax}_\mu \log P(\mathbf{x}|\boldsymbol{\theta}_\mu)&=\int \mathrm{d}\mathbf{x}\ q_\nu(\mathbf{x})\mathrm{argmin}_\mu \log P^{-1}(\mathbf{x}|\boldsymbol{\theta}_\mu)\\
&=\mathrm{argmin}_\mu\int \mathrm{d}\mathbf{x}\ q_\nu(\mathbf{x})\log P^{-1}(\mathbf{x}|\boldsymbol{\theta}_\mu)\\
&=\mathrm{argmin}_\mu D(q_\nu||P_\mu)\ =\ \nu,
\end{aligned} \tag{71}$$

where $D(q_\nu||P_\mu)$ is the Kullback–Leibler distance between the distributions $q_\nu(\mathbf{x})$ and $P(\mathbf{x}|\boldsymbol{\theta}_\mu)$. Thus if (68) holds, then the results (70) and (71) show that the max and expectation operators commute. Using this property in the average energy (60) gives

$$\begin{aligned}
e(\infty)&=-\sum_\nu \tilde{A}(\nu)\int q_\nu(\mathbf{x})\max_\mu \log P(\mathbf{x}|\boldsymbol{\theta}_\mu)\mathrm{d}\mathbf{x}\\
&=\sum_\nu \tilde{A}(\nu)\min_\mu\int q_\nu(\mathbf{x})\log P^{-1}(\mathbf{x}|\boldsymbol{\theta}_\mu))\mathrm{d}\mathbf{x}\\
&=\sum_\nu \tilde{A}(\nu)\min_\mu D(q_\nu||P_\mu)+\sum_\nu \tilde{A}(\nu),H(q_\nu)
\end{aligned} \tag{72}$$

[8] This is certainly true for model distributions $P(\mathbf{x}|\boldsymbol{\theta}_\mu)$ with non-overlapping supports.

and in the distribution (55) it leads to the equation

$$\begin{aligned} Q_\mu(\mathbf{x}) &= \sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x})\, \delta_{\mu;\mathrm{argmax}_{\tilde{\mu}}\, \log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})} \\ \boldsymbol{\theta}_\mu &= \mathrm{argmax}_{\boldsymbol{\theta}} \int Q_\mu(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\theta})\mathrm{d}\mathbf{x}. \end{aligned} \tag{73}$$

We note that the above average energy and the MF (64) average energy are both bounded from below by the average entropy $\sum_\nu \tilde{A}(\nu) H(q_\nu)$. This bound is saturated when all $D(q_\nu||P_\mu)$ terms vanish, i.e. when the model matches the data exactly.

5. Implementation and application of the RS theory

5.1. Population dynamics algorithm

Equation (55) for the order parameter $Q_\mu(\mathbf{x})$ can be solved numerically by a population dynamics algorithm [5] which can be derived as follows. Firstly, we re-arrange the equation for $Q_\mu(\mathbf{x})$:

$$\begin{aligned} Q_\mu(\mathbf{x}) &= \sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x}) \frac{\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right] \tilde{A}(\mu|\nu)}{\sum_{\mu'} \mathbb{1}\left[\Delta_{\mu'}(\mathbf{x})=0\right] \tilde{A}(\mu'|\nu)} \\ &= \sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x}) \Big(\mathbb{1}\left[|S_P(\mathbf{x})|>1\right] + \mathbb{1}\left[|S_P(\mathbf{x})|=1\right] \Big) \frac{\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right] \tilde{A}(\mu|\nu)}{\sum_{\mu'} \mathbb{1}\left[\Delta_{\mu'}(\mathbf{x})=0\right] \tilde{A}(\mu'|\nu)} \\ &= \sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x})\, \mathbb{1}\left[|S_P(\mathbf{x})|=1\right] \mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right] + \cdots \\ &\quad \cdots + \sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x})\, \mathbb{1}\left[|S_P(\mathbf{x})|>1\right] \frac{\mathbb{1}\left[\Delta_\mu(\mathbf{x})=0\right] \tilde{A}(\mu|\nu)}{\sum_{\mu'} \mathbb{1}\left[\Delta_{\mu'}(\mathbf{x})=0\right] \tilde{A}(\mu'|\nu)}. \end{aligned} \tag{74}$$

Secondly, we note that the data distribution $\sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x})$ can be replaced by a large sample $\mathbf{X}$, i.e. by the data itself, via the empirical distribution $N^{-1}\sum_{i\leqslant N} \delta(\mathbf{x}-\mathbf{x}_i)$, which can be also written as $N^{-1}\sum_{\nu\leqslant L}\sum_{i_\nu\leqslant N_\nu} \delta(\mathbf{x}-\mathbf{x}_{i_\nu})$. Here N_ν, which satisfies $\lim_{N\to\infty} N(\nu)/N = \tilde{A}(\nu)$, is the number of data-points sampled from $q_\nu(\mathbf{x})$. Upon using both of these representations of $\sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x})$ in equation (74) we obtain

$$\begin{aligned} Q_\mu(\mathbf{x}) &= \frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{x}-\mathbf{x}_i) \mathbb{1}\left[|S_P(\mathbf{x}_i)|=1\right] \mathbb{1}\left[\Delta_\mu(\mathbf{x}_i)=0\right] + \cdots \\ &\quad \cdots + \frac{1}{N}\sum_{\nu=1}^{L}\sum_{i_\nu=1}^{N_\nu} \delta(\mathbf{x}-\mathbf{x}_{i_\nu}) \mathbb{1}\left[|S_P(\mathbf{x}_{i_\nu})|>1\right] \frac{\mathbb{1}\left[\Delta_\mu(\mathbf{x}_{i_\nu})=0\right] \tilde{A}(\mu|\nu)}{\sum_{\mu'} \mathbb{1}\left[\Delta_{\mu'}(\mathbf{x}_{i_\nu})=0\right] \tilde{A}(\mu'|\nu)}. \end{aligned} \tag{75}$$

Finally, it is very unlikely to find in $\mathbf{X}$, sampled from a distribution of continuous random variables $\sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x})$, data points which satisfy $|S_P(\mathbf{x})|>1$, so the second term in (75) is almost surely zero for any sample $\mathbf{X}$ of finite size. Thus

$$\begin{aligned} Q_\mu(\mathbf{x}) &= \frac{1}{N}\sum_{i=1}^{N}\delta(\mathbf{x}-\mathbf{x}_i)\mathbb{1}\left[|S_P(\mathbf{x}_i)|=1\right]\mathbb{1}\left[\Delta_\mu(\mathbf{x}_i)=0\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\delta(\mathbf{x}-\mathbf{x}_i)\delta_{\mu;\mathrm{argmax}_{\tilde{\mu}}\log P(\mathbf{x}_i|\boldsymbol{\theta}_{\tilde{\mu}})} \\ &= \frac{1}{N}\sum_{i=1}^{N}\delta_{\mu;\mu_i}\delta(\mathbf{x}-\mathbf{x}_i) \end{aligned} \tag{76}$$

where $\mu_i=\mathrm{argmax}_{\tilde{\mu}}\log P(\mathbf{x}_i|\boldsymbol{\theta}_{\tilde{\mu}})$. Using the above in equation (55), we obtain for $\mu\in[K]$ the following system of equations

$$\begin{aligned} Q_\mu(\mathbf{x}) &= \frac{1}{N}\sum_{i=1}^{N}\delta_{\mu,\mu_i}\delta(\mathbf{x}-\mathbf{x}_i) \\ \boldsymbol{\theta}_\mu &= \mathrm{argmax}_{\boldsymbol{\theta}}\int \mathrm{d}\mathbf{x}\; Q_\mu(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta}) \\ \mu_i &= \mathrm{argmax}_{\tilde{\mu}}\log P(\mathbf{x}_i|\boldsymbol{\theta}_{\tilde{\mu}}). \end{aligned} \tag{77}$$

This set can be solved numerically as follows. We create a 'population' of random variables $\{\mu_i : i\in[N]\}$ where $\mu_i\in[K]$ are at first sampled uniformly. We use this population to compute the parameters $\boldsymbol{\theta}_\mu$; The latter are then used to compute a new population $\{\mu_i\}$. The last two steps are repeated until one observes convergence of the energy $e(\infty)=-\sum_{\mu=1}^{K}\int\mathrm{d}\mathbf{x}\;Q_\mu(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)$. Finally, we note that using instead equation (73) as our starting point would lead us to the same population dynamics equations. Thus, for continuous data distributions $\sum_\nu\tilde{A}(\nu)\,q_\nu(\mathbf{x})$ represented by a large finite sample, the equations (55) and (73) are equal.

The population dynamics simplifies significantly if we assume that the distribution $p(\mathbf{x}|\boldsymbol{\theta})$ is the multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\mathbf{m},\boldsymbol{\Lambda}^{-1})=|2\pi\boldsymbol{\Lambda}^{-1}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\boldsymbol{\Lambda}(\mathbf{x}-\mathbf{m})} \tag{78}$$

with mean $\mathbf{m}$ and precision matrix (inverse covariance matrix) $\boldsymbol{\Lambda}$. The parameters $\boldsymbol{\theta}_\mu=(\mathbf{m}_\mu,\boldsymbol{\Lambda}_\mu^{-1})$ we can be estimated directly from the population via the equations

$$\begin{aligned} \mathbf{m}_\mu &= \frac{1}{\sum_{j=1}^{N}\delta_{\mu;\mu_j}}\sum_{i=1}^{N}\delta_{\mu;\mu_i}\mathbf{x}_i \\ \boldsymbol{\Lambda}_\mu^{-1} &= \frac{1}{\sum_{j=1}^{N}\delta_{\mu;\mu_j}}\sum_{i=1}^{N}\delta_{\mu;\mu_i}(\mathbf{x}_i-\mathbf{m}_\mu)(\mathbf{x}_i-\mathbf{m}_\mu)^T, \end{aligned} \tag{79}$$

where μ_i is given by

$$\begin{aligned} \mu_i &= \mathrm{argmax}_\mu\log\mathcal{N}(\mathbf{x}_i|\mathbf{m}_\mu,\boldsymbol{\Lambda}_\mu^{-1}) \\ &= \mathrm{argmax}_\mu -\frac{1}{2}\mathrm{Tr}\left\{\boldsymbol{\Lambda}_\mu(\mathbf{x}_i-\mathbf{m}_\mu)(\mathbf{x}_i-\mathbf{m}_\mu)^T\right\}+\frac{1}{2}\log|\boldsymbol{\Lambda}_\mu|-\frac{d}{2}\log 2\pi. \end{aligned} \tag{80}$$

5.2. Population dynamics algorithm for finite β

Also equation (45) can be solved via population dynamics. However, to replace the distribution of data $\sum_\nu \tilde{A}(\nu)\, q_\nu(\mathbf{x})$ with its empirical version $N^{-1}\sum_{i=1}^N \delta(\mathbf{x}-\mathbf{x}_i)$ we must assume that $\tilde{A}(\tilde{\mu}|\nu)=\tilde{A}(\tilde{\mu})$. For $\mu\in[K]$, this gives us the following equations:

$$\begin{aligned} Q_\mu(\mathbf{x}) &= \frac{1}{N}\sum_{i=1}^N \delta(\mathbf{x}-\mathbf{x}_i) w_i(\mu) \\ w_i(\mu) &= \frac{\tilde{A}(\mu)\,\mathrm{e}^{\beta\log P(\mathbf{x}_i|\boldsymbol{\theta}_\mu)}}{\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu})\,\mathrm{e}^{\beta\log P(\mathbf{x}_i|\boldsymbol{\theta}_{\tilde{\mu}})}} \\ \boldsymbol{\theta}_\mu &= \mathrm{argmax}_{\boldsymbol{\theta}} \int \mathrm{d}\mathbf{x}\, Q_\mu(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\theta}). \end{aligned} \tag{81}$$

They can be solved by creating a population $\{(w_i(1),\ldots,w_i(K)) : i\in[N]\}$ and using the above equations to update this population until convergence of the free energy

$$f(\beta) = -\frac{1}{\beta N}\sum_{i=1}^N \log\Big[\sum_{\mu=1}^K \tilde{A}(\mu)\,\mathrm{e}^{\beta\log P(\mathbf{x}_i|\boldsymbol{\theta}_\mu)}\Big]. \tag{82}$$

Finally, we note that both population dynamics algorithms derived in this subsection look somewhat similar to the expectation-maximisation (EM) algorithm, see e.g. [8]. Comparing the Gaussian EM, used for maximum likelihood inference of Gaussian mixtures, with (79) shows that the main difference is that EM uses the average $\langle\delta_{\mu;\mu_i}\rangle_{\mathrm{EM}}$, over some 'EM-measure', instead of the delta function $\delta_{\mu;\mu_i}$. Gaussian EM is hence an 'annealed' version of the population dynamics (79), but exactly how to relate the two algorithms in a more formal manner is not yet clear.

5.3. Numerical experiments

In the mean-field (MF) theory of Bayesian clustering in [21], the average entropy (64) (derived via a different route) was the central object. It was mainly used for the Gaussian data model $P(\mathbf{x}|\boldsymbol{\theta}_\mu)\equiv\mathcal{N}(\mathbf{x}|\mathbf{m}_\mu,\boldsymbol{\Lambda}_\mu^{-1})$, where it becomes the MF entropy

$$F(\tilde{A}) = \frac{1}{2}\sum_{\mu=1}^K \tilde{A}(\mu)\log\Big((2\pi\mathrm{e})^d\big|\boldsymbol{\Lambda}_\mu^{-1}(\tilde{A})\big|\Big), \tag{83}$$

where $\boldsymbol{\Lambda}_\mu^{-1}(\tilde{A})$ is the covariance matrix

$$\boldsymbol{\Lambda}_\mu^{-1}(\tilde{A}) = \sum_{\nu=1}^L \tilde{A}(\nu|\mu)\big\langle(\mathbf{x}-\mathbf{m}_\mu(\tilde{A}))(\mathbf{x}-\mathbf{m}_\mu(\tilde{A}))^T\big\rangle_\nu, \tag{84}$$

and $\mathbf{m}_\mu(\tilde{A})=\sum_{\nu=1}^L \tilde{A}(\nu|\mu)\langle\mathbf{x}\rangle_\nu$ is the mean. Here we use $\langle\cdots\rangle_\nu$ for the averages generated by $q_\nu(\mathbf{x})$. We note that (83) is also equal to

$$F(\tilde{A}) = \sum_{\mu,\nu}\tilde{A}(\nu,\mu)D(q_\nu||\mathcal{N}_\mu(\tilde{A})) + \sum_{\nu=1}^L \tilde{A}(\nu)H(q_\nu), \tag{85}$$

where $\mathcal{N}_\mu(\tilde{A}) \equiv \mathcal{N}\big(\mathbf{x}|\mathbf{m}_\mu(\tilde{A}), \mathbf{\Lambda}_\mu^{-1}(\tilde{A})\big)$. In addition, for the Gaussian model, the Laplace method, quite often used in statistics to approximate likelihoods [10], applied to the log-likelihood (3) for $N \to \infty$ gives the entropy

$$\hat{F}_N(\mathbf{C}, \mathbf{X}) = \frac{1}{2}\sum_{\mu=1}^{K}\frac{M_\mu(\mathbf{C})}{N}\log\Big((2\pi\mathrm{e})^d\big|\mathbf{\Lambda}_\mu^{-1}(\mathbf{C}, \mathbf{X})\big|\Big), \tag{86}$$

where $\mathbf{\Lambda}_\mu^{-1}(\mathbf{C}, \mathbf{X})$ is the empirical covariance of data in the cluster μ and $M_\mu(\mathbf{C}) = \sum_{i\leqslant N} c_{i\mu}$ is its size. This expression can be minimized for clustering, either by gradient descent [21] or any other algorithm. The MF (83) makes non-trivial predictions about $\hat{F}_N(\mathbf{C}, \mathbf{X})$, such as on structure of its local minima, etc, and correctly estimated $\hat{F}_N \equiv \min_{\mathbf{C}} \hat{F}_N(\mathbf{C}, \mathbf{X})$ for Gaussian data. However, it systematicaly overestimates $\hat{F}_N$ when $K > L$ and when the separations between clusters are small [21].

We expect the present replica theory, related to the MF theory via inequality $e(\infty) \leqslant F(\tilde{A})$, to be more accurate. To test this expectation, we generated samples from two isotropic Gaussian distributions $\mathcal{N}(\mathbf{m}_1, \mathbf{I})$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{I})$. Each sample $\mathbf{X}$, split equally between the distributions, is of size $N = 2000$ and dimension $d = 10$. We note that for any given N and d, there exists an $\epsilon > 0$ such that most of the $\mathbf{x}_i$ in sample $\mathbf{X}$ lie inside the two spheres centred at $\mathbf{m}_1$ and $\mathbf{m}_2$ and both of radius $\sqrt{d(1+\epsilon)}$[9]. The latter suggests that the Euclidean distance $\Delta = ||\mathbf{m}_1 - \mathbf{m}_2||$, measured relative to the natural scale $\sqrt{d}$, can be use as a measure of the degree of separation [26] between the 'clusters' centred at $\mathbf{m}_1$ and $\mathbf{m}_2$ (see figure 1).

We used gradient descent to find the low entropy states of (86) for our data. For each sample $\mathbf{X}$ we ran the algorithm from 10 different random initial states $\mathbf{C}\,(0)$, and computed $\hat{F}_N(\mathbf{C}\,(\infty), \mathbf{X})$. The latter was used to estimate $\hat{F}_N \equiv \min_{\mathbf{C}} \hat{F}_N(\mathbf{C}, \mathbf{X})$. For this data, the log-likelihood function $\hat{F}_N + \log(K)$ has a minimum at $K = 2$, i.e. when the number of assumed clusters K equals the number of true clusters L, so it can be used reliably to infer true number of clusters. However, this inference method no longer works when the separation Δ is too small (see figure 1), but the 'quality' of clustering, as measured by the purity $\rho_N(\mathbf{C}, \tilde{\mathbf{C}}) = \frac{1}{N}\sum_{\mu=1}^{K}\max_\nu \sum_{i=1}^{N} c_{i\mu}\tilde{c}_{i\nu}$ which compares [27] the clustering obtained by algorithm $\mathbf{C}$ with the true clustering $\tilde{\mathbf{C}}$, for $K = 2$, i.e. for the true number of clusters, is still reasonable[10] as can be seen in figure 1.

The predictions of the MF theory for $\hat{F}_N$, $\min_{\tilde{A}} F(\tilde{A})$, is $F_1 = \frac{1}{2}d\log(2\pi\mathrm{e}) + \frac{1}{2}\log[1 + (\Delta/2)^2]$ for $K = 1$, and $F_2 = \frac{1}{2}d\log(2\pi\mathrm{e})$ for $K = 2$. Thus $F_1 \geqslant F_2$, as required. Furthermore, if $\log(2) \geqslant \frac{1}{2}\log[1 + (\Delta/2)^2]$, which happens when $\Delta \leqslant 2\sqrt{3}$, then $F_2 + \log(K) \geqslant F_1$, so the MF theory is unable to recover the true number of clusters when the separation Δ is small. The numerical results for $\hat{F}_N + \log(K)$ are in qualitative agreement with the predicted values, but the MF predictions for $\hat{F}_N$ are indeed found to be inaccurate when the separation Δ is small, and wrong, $F_K \geqslant F_2$ by equation (85), when $K > 2$ (see figure 1).

To test the predictions of our replica theory we solve the Gaussian population dynamics equations (79) and (80) for the data with the same statistical properties as in the above gradient descent experiments, but with a population size $N = 20\,000$. We find that the average energy

[9] The probability of being outside a sphere is bounded from above by $N\mathrm{e}^{-dI(\epsilon)}$, where $I(\epsilon) = \big(\log(1+\epsilon)^{-1} + \epsilon\big)/2$ (see appendix E). A much tighter bound, given by $N\Gamma\left(d/2, d(1+\epsilon)/2\right)/\Gamma(d/2)$, uses that for $\mathbf{x}$ sampled from $\mathcal{N}(\mathbf{m}, \mathbf{I})$ the squared Euclidean distance $||\mathbf{x} - \mathbf{m}||^2$ follows the χ^2 distribution.

[10] We note that $0 < \rho_N \leqslant 1$ with $\rho_N = 1$ corresponding to a perfect recovery of true clusters and with $\rho_N \approx 1/L$ corresponding to a random (unbiased) assignment into clusters.

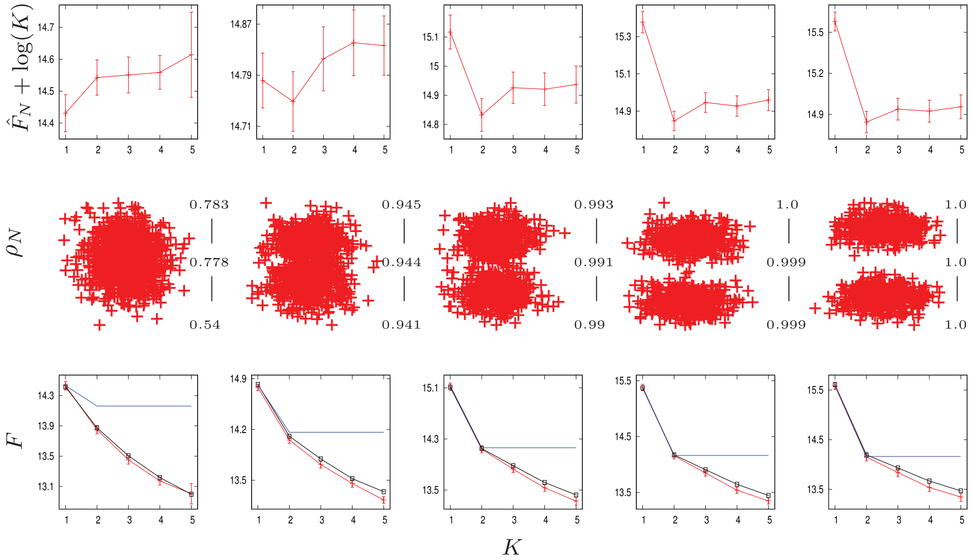

Figure 1. Bayesian clustering of data generated from Gaussian distributions $\mathcal{N}(\mathbf{m}_1, \mathbf{I})$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{I})$, with separation $\Delta = ||\mathbf{m}_1 - \mathbf{m}_2||$. The sample, split equally between the distributions, is of size $N = 2000$, and the data dimension is $d = 10$. The data was generated for $\Delta/\sqrt{d} \in \{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}\}$, from left to right. Middle: data projected into two dimensions for Δ increasing from left to right. The quality of the clustering, measured by the 'purity' ρ_N, obtained by the population dynamics clustering algorithm is increasing with Δ. ρ_N was measured for 10 random samples of data, but only the minimum, median and maximum values of ρ_N (numbers connected by lines) are shown. The size of each sample, split equally between the distributions, was $N = 20\,000$ and the clustering algorithm assumed the number of clusters to be $K = 2$. Top: $\hat{F}_N + \log(K)$ (red crosses connected by lines), with the log-likelihood $\hat{F}_N \equiv \min_{\mathbf{C}} \hat{F}_N(\mathbf{C}, \mathbf{X})$ computed by a gradient descent algorithm, shown as a function of the assumed number of clusters K. Symbols, connected by lines and with error bars, denote the average and ± one standard deviation, measured over 10 random samples of data. Bottom: the log-likelihood $\hat{F}_N$ (red crosses connected by lines) is compared with the results of the mean-field theory (blue line) and population dynamics (connected black squares). For $K \geqslant 2$ only the mean-field lower bound $\frac{d}{2}\log(2\pi \mathrm{e})$ is plotted.

$$e(\infty) = -\sum_{\mu \leqslant K} \int \mathrm{d}\mathbf{x}\, Q_\mu(\mathbf{x}) \log \mathcal{N}(\mathbf{x}|\mathbf{m}_\mu, \mathbf{\Lambda}_\mu^{-1}), \tag{87}$$

as computed by the population dynamics algorithm, is in good agreement with the value of $\hat{F}_N$ obtained by gradient descent minimization (see figure 1). The residual differences observed between $e(\infty)$ and $\hat{F}_N$ are finite size effects. Furthermore, we note that the numerical complexity of the population dynamics algorithm is consistent with the lower bound that is *linear* in N (on average), as follows from the complexity analysis in [21].

Finally, we compare the Gaussian variant of the population dynamics clustering algorithm with a popular software package [11] which uses EM algorithm to estimate the maximum $\mathcal{L}_N(\mathbf{X})$ of the log-likelihood

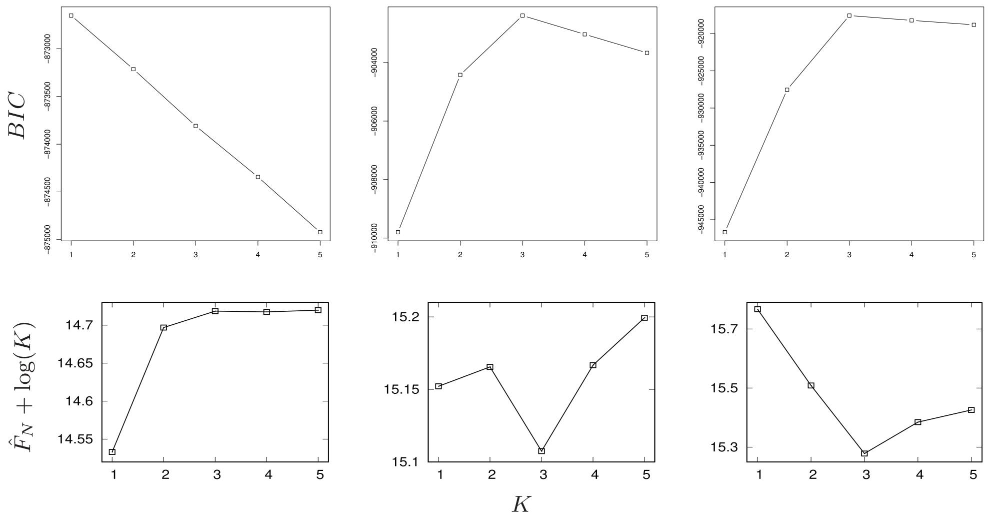

Figure 2. Inferring the number of clusters in data generated from Gaussian distributions $\mathcal{N}(\mathbf{m}_\mu, \mathbf{I})$ with separation $\Delta = ||\mathbf{m}_\mu - \mathbf{m}_\nu||$, where $(\mu, \nu) \in [3]$. The sample, split equally between the distributions, is of size $N = 3 \times 10^4$, and the data dimension is $d = 10$. The data was generated for $\Delta/\sqrt{d} \in \left\{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}\right\}$, but results shown here (from left to right) are only for $\Delta/\sqrt{d} \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$. Top: BIC $\equiv 2\mathcal{L}_N - n_{\mathcal{N}} \log(N)$, where $\mathcal{L}_N$ is the log-likelihood of GMM estimated by EM algorithm and $n_{\mathcal{N}}$ is the number of parameters, as a function of K. Bottom: $\hat{F}_N + \log(K)$, where $\hat{F}_N \equiv \min_{\mathbf{C}} \hat{F}_N(\mathbf{C}, \mathbf{X})$ is the log-likelihood function computed by the population dynamics algorithm, as a function of K.

$$\ell_N(\mathbf{X}) = \sum_{i=1}^{N} \log \left(\sum_{\mu=1}^{K} w(\mu)\, \mathcal{N}(\mathbf{x}_i | \mathbf{m}_\mu, \mathbf{\Sigma}_\mu) \right) \tag{88}$$

with respect to the parameters of the Gaussian mixture model (GMM) $\sum_{\mu \leqslant K} w(\mu)\, \mathcal{N}(\mathbf{x}_i | \mathbf{m}_\mu, \mathbf{\Sigma}_\mu)$, which are the means $\mathbf{m}_\mu$, the covariances $\mathbf{\Sigma}_\mu$ and the weights $w(\mu) \geqslant 0$, where $\sum_{\mu \leqslant K} w(\mu) = 1$. To this end we consider inferring number of clusters in the samples of a Gaussian data with more than $L = 2$ clusters, non-identity covariance matrices and a relatively large number of dimensions (see figures 2, 3 and 5). The software package uses the Bayesian Information Criterion (BIC) $2\mathcal{L}_N - n_{\mathcal{N}} \log(N)$, where $n_{\mathcal{N}}$ is the number of parameters used in GMM, and the population dynamics algorithm uses $\hat{F}_N + \log(K)$, with the log-likelihood $\hat{F}_N \equiv \min_{\mathbf{C}} \hat{F}_N(\mathbf{C}, \mathbf{X})$ estimated by the average energy $e(\infty)$, to infer the number of clusters in the data.

For uncorrelated data we observe in figure 2 that inference success in both methods is strongly affected by the degree of separation Δ of the clusters in the data, as measured by the Euclidean distance between the means of Gaussians. For small Δ the recovery of the true number $L = 3$ of clusters is not possible. A simple MF argument, similar to the one used for $L = 2$, predicts that this inference failure latter will happen when $\Delta \leqslant 2\sqrt{3}$, i.e. exactly as for $L = 2$. However, both algorithms are found to 'work' below this MF threshold (see figure 2) suggesting that the MF argument gives an upper bound. For correlated data, even when the separation parameter Δ is zero, the true number of clusters can still be recovered correctly by

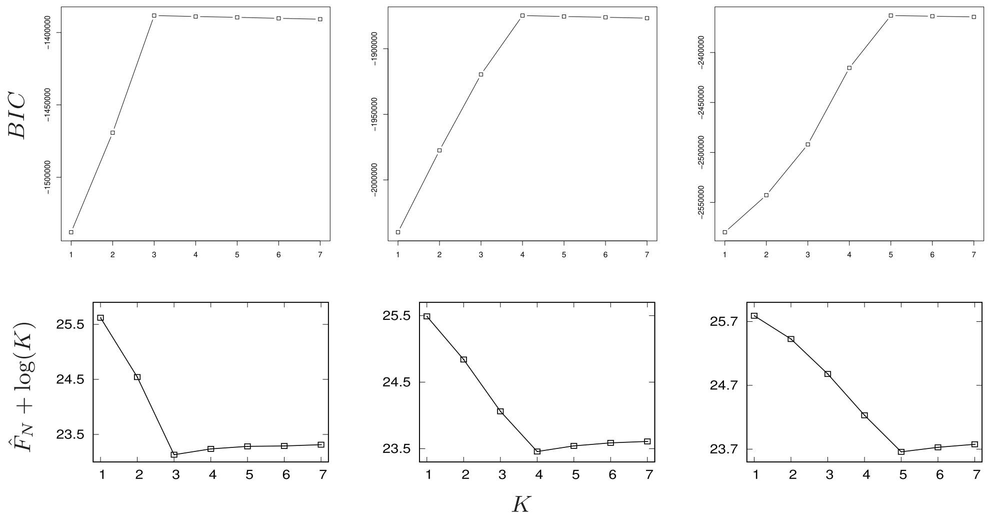

Figure 3. Inferring the number of clusters in data generated from Gaussian distributions $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mu)$ with (from left to right) $\mu \in [3]$, $\mu \in [4]$ and $\mu \in [5]$. The samples of dimension $d = 10$, split equally between the distributions, were, respectively, of the size $N = 3 \times 10^4$, $N = 4 \times 10^4$ and $N = 5 \times 10^4$. The covariance matrices $\mathbf{\Sigma}_\mu$ were sampled from the Wishart distribution with $d + 1$ degrees of freedom and precision matrix $\mathbf{I}$. Top: BIC $\equiv 2\mathcal{L}_N - n_{\mathcal{N}} \log(N)$, where $\mathcal{L}_N$ is the log-likelihood of GMM estimated by EM algorithm and $n_{\mathcal{N}}$ is the number of parameters, as a function of K. Bottom: $\hat{F}_N + \log(K)$, with $\hat{F}_N$ computed by the population dynamics algorithm, as a function of K.

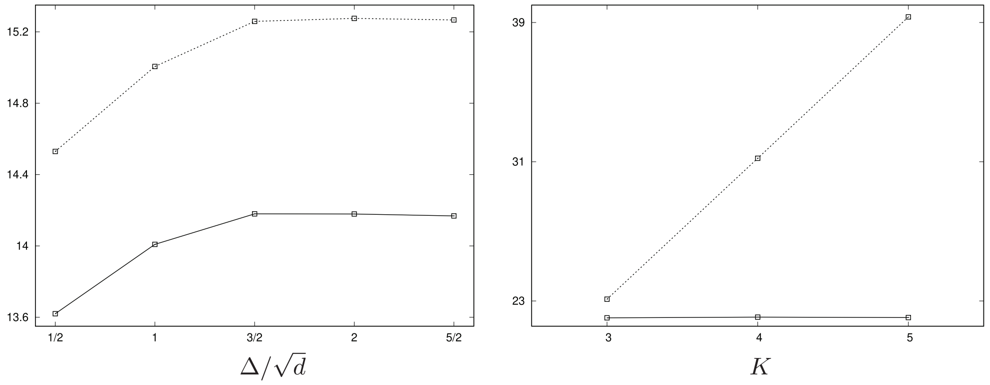

Figure 4. The log-likelihood densities $-\mathcal{L}_N/N$ (top dotted line) and $\hat{F}_N$ (bottom solid line) plotted as functions of, respectively, the cluster separation $\Delta/\sqrt{d}$ (computed at the inferred number of clusters) and inferred number of clusters K for the data described in figures 2 and 3.

both algorithms (see figure 3). In all numerical experiments described in figures 2 and 3 the log-likelihood density $-\mathcal{L}_N/N$, estimated by EM algorithm, is an upper bound for the log-likelihood density $\hat{F}_N$ computed by Gaussian population dynamics (see figure 4). This points, at least in the regime of finite dimension d and sample size $N \to \infty$, to a possible relation between these likelihood functions.

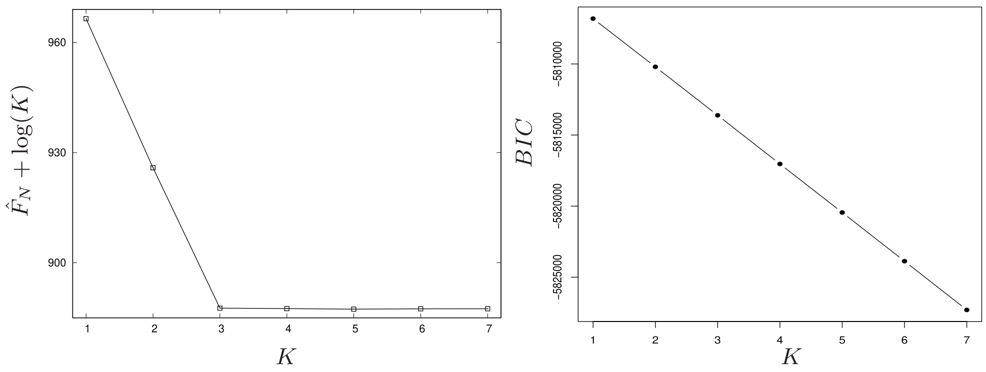

Figure 5. Inferring number of clusters in the data generated from Gaussian distributions $\mathcal{N}(\mathbf{m}_\mu, \mathbf{\Sigma}_\mu)$ with separation $\Delta/\sqrt{d} = ||\mathbf{m}_\mu - \mathbf{m}_\nu|| = 5.2$, where $(\mu, \nu) \in [3]$. The sample, split equally between the distributions, is of size $N = 3 \times 10^3$, and of dimension $d = 500$. The (diagonal) covariance matrices $\mathbf{\Sigma}_\mu$ were sampled from χ^2 distribution with 3 degrees of freedom. The maximum and minimum diagonal entries in these matrices is, respectively, 17.064 and 0.017, so $\Delta/\sqrt{d} = 5.2$ ensures that clusters in the sample are well separated (see appendix E). Left: $\hat{F}_N + \log(K)$, with $\hat{F}_N \equiv \min_{\mathbf{C}} \hat{F}_N(\mathbf{C}, \mathbf{X})$ computed by the population dynamics algorithm, as a function of K. Right: BIC $\equiv 2\mathcal{L}_N - n_{\mathcal{N}} \log(N)$, where $\mathcal{L}_N$ is the log-likelihood of GMM estimated by EM algorithm and $n_{\mathcal{N}}$ is the number of parameters, as a function of K.

In the high dimensional regime $d \to \infty$ and $N \to \infty$, with d/N finite, both algorithms fail to find the correct number of clusters (see figure 5), but they fail differently. The algorithm which uses Gaussian population dynamics, which was derived assuming finite d and $N \to \infty$, predicts more than $L = 3$ clusters in the data, and the algorithm which uses EM predicts only one cluster. However, the population dynamics 'almost' predicts the correct number $L = 3$ of clusters: the changes in the log-likelihood function $\hat{F}_N + \log(K)$ in the $K > 3$ regime are much smaller than in the $K \leqslant 3$ regime, as can be seen in figure 5. This behaviour is also observed for similarly generated data with the same sample size but with higher dimensions (not shown here), suggesting that taking into account the effect of the dimension d properly in the present theoretical framework could lead to improvements in inference.

6. Discussion

In this paper we use statistical mechanics to study model-based Bayesian clustering. The partitions of data are microscopic states, the negative log-likelihood of the data is the energy of these states, and the data act as disorder in the model. The optimal (MAP) partition corresponds to the minimal energy state, i.e. the ground state of this system. The latter can be obtained from the free energy via a low 'temperature' limit, so to investigate MAP inference we evaluate the free energy. We assume that in a very large system, i.e. for a large sample size, the free energy (density) is self-averaging. This allows us to focus on the disorder-averaged free energy, using the replica method. Following the prescription of the replica method we first compute the average for an integer n number of replicas, then we take the large system limit followed by the limit $n \to 0$. The latter is facilitated by assuming replica symmetry (RS) in the order parameter equation. The main order parameter in the theory is the (average) distribution of data in each cluster $\mu \in [K]$.

In the low temperature limit, the equations of the RS theory allow us to study the low energy states of the system. In this limit the average free energy and average energy are identical. We show that the true partitions of the data are recovered exactly when the assumed number of clusters K and the true number of clusters L are equal, and the model distributions $P(\mathbf{x}|\boldsymbol{\theta}_\mu)$ have non overlapping supports for different $\boldsymbol{\theta}_\mu$. The high temperature limit of the RS theory recovers the mean-field theory of [21]. In this latter limit, the average energy, which equals the MF entropy [21], is dominated by the prior. The MF entropy is an upper bound for the low temperature average energy, and can be optimised by selecting the prior. Our order parameter equation can be solved numerically using a population dynamics algorithm. Using this algorithm for the Gaussian data very accurately reproduces the results obtained by gradient descent, minimising the negative log-likelihood of data, algorithm even in the regime of a small separations between clusters and when $K > L$ where the MF theory gives incorrect predictions [21]. The zero temperature population dynamics algorithm can be used for MAP inference.

There are several interesting directions into which to extend the present work. Many current studies use the so-called Rand index [28], or the 'purity' [27], for measuring the dissimilarity between the true and inferred clusterings of data, but it would be also interesting to estimate the probability that the inferred clustering is 'wrong'. Another direction is to consider the high dimensional regime where $N \to \infty$ and $d \to \infty$, with d/N finite. We envisage that here the task of separating clusters may be 'easier' than in the lower dimensional $d/N \to 0$ regime, due to the 'blessing of dimensionality' phenomenon [29], according to which most data sampled from high-dimensional Gaussian distributions reside in the 'thin' shell of a sphere (see appendix E). Both the early study [30] and the more recent study [31] on Bayesian discriminant analysis indicate that the classification of data, a supervised inference problem closely related to clustering, becomes significantly easier in the high-dimensional regime. Alternatively, the high dimensional regime could also cause overfitting, and one may want to quantify this phenomena by using a more general information-theoretic measure of overfitting [3].

Acknowledgments

This work was supported by the Medical Research Council of the United Kingdom (Grant MR/L01257X/1).

Appendix A. Disorder average

In this appendix we study the average

$$\begin{aligned}
&\left\langle\left\langle \mathrm{e}^{-\beta N\sum_{\alpha=1}^{n}\hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})}\right\rangle_{\{\mathbf{C}^\alpha\}}\right\rangle_{\mathbf{X}}\\
&=\int \mathrm{d}\mathbf{x}_1\cdots\mathrm{d}\mathbf{x}_N\sum_{\mathbf{C}} q(\mathbf{C}|L)\Big\{\prod_{\nu=1}^{L}\prod_{i=1}^{N} q_\nu^{c_{i\nu}}(\mathbf{x}_i)\Big\}\left\langle \mathrm{e}^{-\beta N\sum_{\alpha=1}^{n}\hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})}\right\rangle_{\{\mathbf{C}^\alpha\}}\\
&=\int \mathrm{d}\mathbf{x}_1\cdots\mathrm{d}\mathbf{x}_N\left\langle\Big\{\prod_{\nu=1}^{L}\prod_{i=1}^{N} q_\nu^{c_{i\nu}}(\mathbf{x}_i)\Big\}\mathrm{e}^{-\beta N\sum_{\alpha=1}^{n}\hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})}\right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}}, \qquad \text{(A.1)}
\end{aligned}$$

where the average $\langle\cdots\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}}$ now refers to the distribution $\big\{\prod_{\alpha=1}^{n}P(\mathbf{C}^\alpha|K)\big\}\,q(\mathbf{C}|L)$. If we define the density

$$Q_\mu(\mathbf{x}|\mathbf{C}^\alpha,\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} c^\alpha_{i\mu}\,\delta(\mathbf{x}-\mathbf{x}_i), \tag{A.2}$$

then we may write

$$\begin{aligned} -N\sum_{\alpha=1}^{n}\hat{F}_N(\mathbf{C}^\alpha,\,\mathbf{X}) &= \sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}\log\left\langle \mathrm{e}^{\sum_{i=1}^{N} c^\alpha_{i\mu}\log P(\mathbf{x}_i|\boldsymbol{\theta}_\mu)}\right\rangle_{\boldsymbol{\theta}_\mu} \\ &= \sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}\log\left\langle \mathrm{e}^{N\int \mathrm{d}\mathbf{x}\; Q_\mu(\mathbf{x}|\mathbf{C}^\alpha,\mathbf{X})\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)}\right\rangle_{\boldsymbol{\theta}_\mu} \end{aligned} \tag{A.3}$$

and for (A.1) we obtain

$$\begin{aligned} &\int \mathrm{d}\mathbf{x}_1\cdots\mathrm{d}\mathbf{x}_N\,\left\langle\Big\{\prod_{\nu=1}^{L}\prod_{i=1}^{N} q_\nu^{c_{i\nu}}(\mathbf{x}_i)\Big\}\mathrm{e}^{\beta\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}\log\left\langle \mathrm{e}^{N\int Q_\mu(\mathbf{x}|\mathbf{C}^\alpha,\mathbf{X})\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)\mathrm{d}\mathbf{x}}\right\rangle_{\boldsymbol{\theta}_\mu}}\right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}} \\ &= \int \mathrm{d}\mathbf{x}_1\cdots\mathrm{d}\mathbf{x}_N\,\left\langle\Big\{\prod_{\nu=1}^{L}\prod_{i=1}^{N} q_\nu^{c_{i\nu}}(\mathbf{x}_i)\Big\}\right. \\ &\quad\times\prod_{\alpha=1}^{n}\prod_{\mu=1}^{K}\Big\{\prod_{\mathbf{x}}\int \mathrm{d}Q^\alpha_\mu(\mathbf{x})\;\delta\left[Q^\alpha_\mu(\mathbf{x})-Q_\mu(\mathbf{x}|\mathbf{C}^\alpha,\mathbf{X})\right]\Big\} \\ &\quad\times\left.\mathrm{e}^{\beta\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}\log\left\langle \mathrm{e}^{N\int Q^\alpha_\mu(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)\mathrm{d}\mathbf{x}}\right\rangle_{\boldsymbol{\theta}_\mu}}\right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}} \\ &= \int\Big\{\mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}}\Big\}\,\mathrm{e}^{\mathrm{i}N\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}\int\hat{Q}^\alpha_\mu(\mathbf{x})Q^\alpha_\mu(\mathbf{x})\mathrm{d}\mathbf{x}} \\ &\quad\times\mathrm{e}^{\beta\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}\log\langle \mathrm{e}^{N\int Q^\alpha_\mu(\mathbf{x})\log P(\mathbf{x}|\boldsymbol{\theta}_\mu)\mathrm{d}\mathbf{x}}\rangle_{\boldsymbol{\theta}_\mu}} \\ &\quad\times\left\langle\prod_{i=1}^{N}\int \mathrm{d}\mathbf{x}_i\Big\{\prod_{\nu=1}^{L} q_\nu^{c_{i\nu}}(\mathbf{x}_i)\Big\}\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}c^\alpha_{i\mu}\hat{Q}^\alpha_\mu(\mathbf{x}_i)}\right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}}. \end{aligned} \tag{A.4}$$

Using the properties of $\{c_{i\nu}\}$, the last line in the above expression can be rewritten as

$$\begin{aligned} &\left\langle\prod_{i=1}^{N}\int \mathrm{d}\mathbf{x}_i\Big\{\prod_{\nu=1}^{L} q_\nu^{c_{i\nu}}(\mathbf{x}_i)\Big\}\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}c^\alpha_{i\mu}\hat{Q}^\alpha_\mu(\mathbf{x}_i)}\right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}} \\ &= \left\langle\prod_{i=1}^{N}\Big\{\sum_{\nu=1}^{L}c_{i\nu}\int \mathrm{d}\mathbf{x}\;q_\nu(\mathbf{x})\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}c^\alpha_{i\mu}\hat{Q}^\alpha_\mu(\mathbf{x})}\Big\}\right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}} \\ &= \left\langle\mathrm{e}^{\sum_{i=1}^{N}\log\sum_{\nu=1}^{L}c_{i\nu}\int\mathrm{d}\mathbf{x}\;q_\nu(\mathbf{x})\exp\left[-\mathrm{i}\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K}c^\alpha_{i\mu}\hat{Q}^\alpha_\mu(\mathbf{x})\right]}\right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}}. \end{aligned} \tag{A.5}$$

Since $c_{i\nu}, c^\alpha_{i\nu} \in \{0,1\}$, subject to $\sum_{\nu=1}^{L} c_{i\nu} = \sum_{\mu=1}^{K} c^\alpha_{i\mu} = 1$, it follows that the vectors $\mathbf{c} = (c_1,\ldots,c_L)$, $\mathbf{c}_i = (c_{i1},\ldots,c_{iL})$, $\mathbf{c}^\alpha = (c^\alpha_1,\ldots,c^\alpha_K)$ and $\mathbf{c}^\alpha_i = (c^\alpha_{i1},\ldots,c^\alpha_{iK})$, will satisfy the identities $\mathbf{c}\cdot\mathbf{c}_i = \delta_{\mathbf{c},\mathbf{c}_i}$ and $\mathbf{c}^\alpha\cdot\mathbf{c}^\alpha_i = \delta_{\mathbf{c}^\alpha,\mathbf{c}^\alpha_i}$. Inserting $\sum_{\mathbf{c}}\mathbf{c}_i\cdot\mathbf{c} = 1$ and $\sum_{\mathbf{c}^\alpha}\mathbf{c}^\alpha_i\cdot\mathbf{c}^\alpha = 1$ into the exponential function in the average (A.5) now gives, with $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n) \in \{1,\ldots,K\}^n$:

$$
\begin{aligned}
&\sum_{i=1}^{N} \log \sum_{\nu=1}^{L} c_{i\nu} \int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x}) \mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K} c_{i\mu}^{\alpha}\hat{Q}_{\mu}^{\alpha}(\mathbf{x})} \\
&\quad = \sum_{\mathbf{c}} \sum_{\{\mathbf{c}^\alpha\}} \sum_{i=1}^{N} \mathbf{c}\cdot\mathbf{c}_i \prod_{\alpha=1}^{n} \mathbf{c}^{\alpha}\cdot\mathbf{c}_i^{\alpha} \log \sum_{\nu=1}^{L} c_\nu \int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x}) \mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\sum_{\mu=1}^{K} c_{\mu}^{\alpha}\hat{Q}_{\mu}^{\alpha}(\mathbf{x})} \\
&\quad = \sum_{\nu,\boldsymbol{\mu}} \sum_{i=1}^{N} c_{i\nu} \left\{ \prod_{\alpha=1}^{n} c_{i\mu_\alpha}^{\alpha} \right\} \sum_{\mathbf{c}} \sum_{\{\mathbf{c}^\alpha\}} c_\nu \left\{ \prod_{\alpha=1}^{n} c_{\mu_\alpha}^{\alpha} \right\} \\
&\qquad\qquad \times \log \sum_{\nu'=1}^{L} c_{\nu'} \int \mathrm{d}\mathbf{x}\, q_{\nu'}(\mathbf{x}) \mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\sum_{\mu'_\alpha=1}^{K} c_{\mu'_\alpha}^{\alpha}\hat{Q}_{\mu'_\alpha}^{\alpha}(\mathbf{x})} \\
&\quad = \sum_{\nu,\boldsymbol{\mu}} \sum_{i=1}^{N} c_{i\nu} \left\{ \prod_{\alpha=1}^{n} c_{i\mu_\alpha}^{\alpha} \right\} \log \int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x})\, \mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\hat{Q}_{\mu_\alpha}^{\alpha}(\mathbf{x})},
\end{aligned}
\tag{A.6}
$$

where we used the identities $\sum_{\mathbf{c}^\alpha} c_\mu^\alpha = 1$ for all (α,μ), and $\sum_{\mathbf{c}} c_\nu \log[\sum_{\nu'} c_{\nu'}\phi_{\nu'}] = \log\phi_\nu$ for all ν. Let us now define the density

$$
A(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\}) = \frac{1}{N}\sum_{i=1}^{N} c_{i\nu} \left\{ \prod_{\alpha=1}^{n} c_{i\mu_\alpha}^{\alpha} \right\},
\tag{A.7}
$$

where $NA(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\})$ is the number of data-points that are sampled from the distribution $q_\nu(\mathbf{x})$ and assigned to clusters $\mu_1,\ldots,\mu_n$ for the n replicas, respectively. Using this definition and (A.6) in equation (A.5) converts the latter expression into

$$
\begin{aligned}
&\left\langle \mathrm{e}^{N\sum_{\nu,\boldsymbol{\mu}} A(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\}) \log\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x}) \exp\left[-\mathrm{i}\sum_{\alpha=1}^{n}\hat{Q}_{\mu_\alpha}^{\alpha}(\mathbf{x})\right]} \right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}} \\
&= \left\langle \prod_{\nu,\boldsymbol{\mu}} \int \mathrm{d}A(\nu,\boldsymbol{\mu})\, \delta\left[A(\nu,\boldsymbol{\mu}) - A(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\})\right] \right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}} \\
&\quad \times \mathrm{e}^{N\sum_{\nu,\boldsymbol{\mu}} A(\nu,\boldsymbol{\mu}) \log\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x}) \exp[-\mathrm{i}\sum_{\alpha=1}^{n}\hat{Q}_{\mu_\alpha}^{\alpha}(\mathbf{x})]} \\
&= \int \{\mathrm{d}A\,\mathrm{d}\hat{A}\}\, \mathrm{e}^{N\tilde{\Psi}[\{\hat{\mathbf{Q}}\};\{A,\hat{A}\}]},
\end{aligned}
\tag{A.8}
$$

where

$$
\begin{aligned}
\tilde{\Psi}[\{\hat{\mathbf{Q}}\};\{A,\hat{A}\}] = {} & \sum_{\nu,\boldsymbol{\mu}} A(\nu,\boldsymbol{\mu}) \left[\mathrm{i}\hat{A}(\nu,\boldsymbol{\mu}) + \log\int \mathrm{d}\mathbf{x}\, q_\nu(\mathbf{x})\, \mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^{n}\hat{Q}_{\mu_\alpha}^{\alpha}(\mathbf{x})} \right] \\
& + \frac{1}{N}\log\left\langle \mathrm{e}^{-\mathrm{i}N\sum_{\nu,\boldsymbol{\mu}}\hat{A}(\nu,\boldsymbol{\mu})A(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\})} \right\rangle_{\{\mathbf{C}^\alpha\};\mathbf{C}}.
\end{aligned}
\tag{A.9}
$$

Finally, using (A.8) in the average (A.4) gives us the integral (14), as claimed.

Appendix B. Derivation of RS equations

The RS assumption implies that $Q_{\mu_\alpha}^{\alpha}(\mathbf{x}) = Q_{\mu_\alpha}(\mathbf{x})$, from which one deduces $\boldsymbol{\theta}_\mu^\alpha = \boldsymbol{\theta}_{\mu_\alpha}$ via (41). Insertion of these forms into the right-hand side of (43), using (44), leads to

$$\sum_{\nu,\boldsymbol{\mu}}\delta_{\mu;\mu_\alpha}A(\nu,\boldsymbol{\mu})\frac{q_\nu(\mathbf{x})\,\mathrm{e}^{\sum_{\gamma=1}^{n}\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu_\gamma})}}{\int q_\nu(\tilde{\mathbf{x}})\,\mathrm{e}^{\sum_{\gamma=1}^{n}\beta\log P(\tilde{\mathbf{x}}|\boldsymbol{\theta}_{\mu_\gamma})}\mathrm{d}\tilde{\mathbf{x}}}$$
$$=\sum_{\nu,\boldsymbol{\mu}}\delta_{\mu;\mu_\alpha}\frac{\tilde{A}(\nu)\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\Big[\prod_{\gamma=1}^{n}\tilde{A}(\mu_\gamma|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu_\gamma})}\Big]}{\sum_{\tilde{\nu}}\tilde{A}(\tilde{\nu})\int\mathrm{d}\mathbf{x}\,q_{\tilde{\nu}}(\mathbf{x})\Big[\prod_{\gamma=1}^{n}\sum_{\tilde{\mu}_\gamma}\tilde{A}(\tilde{\mu}_\gamma|\tilde{\nu})\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}_\gamma})}\Big]}$$
$$\times\frac{q_\nu(\mathbf{x})\,\mathrm{e}^{\sum_{\gamma=1}^{n}\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu_\gamma})}}{\int q_\nu(\tilde{\mathbf{x}})\,\mathrm{e}^{\sum_{\gamma=1}^{n}\beta\log P(\tilde{\mathbf{x}}|\boldsymbol{\theta}_{\mu_\gamma})}\mathrm{d}\tilde{\mathbf{x}}}$$
$$=\sum_{\nu}\tilde{A}(\nu)\frac{q_\nu(\mathbf{x})\,\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu})}\Big[\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}\Big]^{n-1}}{\sum_{\tilde{\nu}}\tilde{A}(\tilde{\nu})\int\mathrm{d}\mathbf{x}\,q_{\tilde{\nu}}(\mathbf{x})\Big[\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\tilde{\nu})\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}\Big]^{n}}. \tag{B.1}$$

We can now take the replica limit $n\to 0$, and obtain (45). Using the RS assumption in (44) gives us the following expression for the marginal $A(\nu)=\sum_{\boldsymbol{\mu}}A(\nu,\boldsymbol{\mu})$:

$$A(\nu)=\frac{\tilde{A}(\nu)\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\Big[\sum_{\mu}\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu})}\Big]^{n}}{\sum_{\tilde{\nu}}\tilde{A}(\tilde{\nu})\int\mathrm{d}\mathbf{x}\,q_{\tilde{\nu}}(\mathbf{x})\Big[\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\tilde{\nu})\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}\Big]^{n}}. \tag{B.2}$$

Hence $\lim_{n\to0}A(\nu)=\tilde{A}(\nu)$. The RS equation for the conditional $A(\boldsymbol{\mu}|\nu)$ becomes

$$A(\boldsymbol{\mu}|\nu)=\frac{\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\Big[\prod_{\alpha=1}^{n}\tilde{A}(\mu_\alpha|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu_\alpha})}\Big]}{\sum_{\tilde{\nu}}\tilde{A}(\tilde{\nu})\int\mathrm{d}\mathbf{x}\,q_{\tilde{\nu}}(\mathbf{x})\Big[\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\tilde{\nu})\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}\Big]^{n}}. \tag{B.3}$$

Its conditional marginal is

$$A(\mu|\nu)=\frac{\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu})}\Big[\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}\Big]^{n-1}}{\sum_{\tilde{\nu}}\tilde{A}(\tilde{\nu})\int\mathrm{d}\mathbf{x}\,q_{\tilde{\nu}}(\mathbf{x})\Big[\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\tilde{\nu})\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}\Big]^{n}}, \tag{B.4}$$

which for $n\to0$ becomes (46):

$$A(\mu|\nu)=\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\frac{\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu})}}{\sum_{\tilde{\mu}}\tilde{A}(\tilde{\mu}|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\tilde{\mu}})}}. \tag{B.5}$$

Finally, inserting $Q^{\alpha}_{\mu_\alpha}(\mathbf{x})=Q_{\mu_\alpha}(\mathbf{x})$ and $\boldsymbol{\theta}^{\alpha}_{\mu}=\boldsymbol{\theta}_{\mu_\alpha}$ into the nontrivial part of the average free energy (42) and taking the limit $n\to0$ gives equation (47):

$$\begin{aligned}f(\beta)-\phi(\beta)&=-\lim_{n\to0}\frac{1}{\beta n}\log\Big\{\sum_{\nu}\tilde{A}(\nu)\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\Big[\sum_{\mu=1}^{K}\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu})}\Big]^{n}\Big\}\\&=-\frac{1}{\beta}\sum_{\nu}\tilde{A}(\nu)\int\mathrm{d}\mathbf{x}\,q_\nu(\mathbf{x})\log\Big[\sum_{\mu=1}^{K}\tilde{A}(\mu|\nu)\,\mathrm{e}^{\beta\log P(\mathbf{x}|\boldsymbol{\theta}_{\mu})}\Big]\end{aligned} \tag{B.6}$$

Appendix C. Physical meaning of observables

Let us consider the following two averages:

$$Q_\mu(\mathbf{x}) = \Big\langle \langle Q_\mu(\mathbf{x}|\mathbf{C},\mathbf{X})\rangle_{\mathbf{C}|\mathbf{X}} \Big\rangle_{\mathbf{X}}, \tag{C.1}$$

$$A(\nu,\mu) = \Big\langle \langle A(\nu,\mu|\mathbf{C},\mathbf{X})\rangle_{\mathbf{C}|\mathbf{X}} \Big\rangle_{\mathbf{X}}, \tag{C.2}$$

in which $\langle\cdots\rangle_{\mathbf{C}|\mathbf{X}}$ is generated by the Gibbs–Boltzmann distribution (50) and the disorder average $\langle\cdots\rangle_{\mathbf{X}}$ by the distribution (8). Using the replica identity

$$\begin{aligned}\frac{\sum_{\mathbf{C}} W(\mathbf{C})F(\mathbf{C})}{\sum_{\mathbf{C}} W(\mathbf{C})} &= \lim_{n\to 0}\sum_{\mathbf{C}} W(\mathbf{C})F(\mathbf{C})\Big\{\sum_{\tilde{\mathbf{C}}} W(\tilde{\mathbf{C}})\Big\}^{n-1}\\ &= \lim_{n\to 0}\sum_{\mathbf{C}^1}\cdots\sum_{\mathbf{C}^n} F(\mathbf{C}^1)\prod_{\alpha=1}^{n} W(\mathbf{C}^\alpha)\end{aligned}\tag{C.3}$$

we may write for any test function $g(\mathbf{x})$

$$\begin{aligned}\int \mathrm{d}\mathbf{x}\; Q_\mu(\mathbf{x}) &= \Big\langle \sum_{\mathbf{C}^1}\cdots\sum_{\mathbf{C}^n}\int \mathrm{d}\mathbf{x}\; Q_\mu(\mathbf{x}|\mathbf{C}^1,\mathbf{X})g(\mathbf{x})\prod_{\alpha=1}^{n}\Big[P(\mathbf{C}^\alpha|K)\mathrm{e}^{-\beta N\hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})}\Big]\Big\rangle_{\mathbf{X}}\\ &= \Big\langle\Big\langle \mathrm{e}^{-\beta N\sum_{\alpha=1}^{n}\hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})}\int \mathrm{d}\mathbf{x}\; Q_\mu(\mathbf{x}|\mathbf{C}^1,\mathbf{X})g(\mathbf{x})\Big\rangle_{\mathbf{X}}\Big\rangle_{\{\mathbf{C}^\alpha\}}.\end{aligned}\tag{C.4}$$

Following the same steps we used in computing the disorder average in (13) we obtain

$$\begin{aligned}&\Big\langle\Big\langle \mathrm{e}^{-\beta N\sum_{\alpha=1}^{n}\hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})}\int Q_\mu(\mathbf{x}|\mathbf{C}^1,\mathbf{X})g(\mathbf{x})\,\mathrm{d}\mathbf{x}\Big\rangle_{\mathbf{X}}\Big\rangle_{\{\mathbf{C}^\alpha\}}\\ &= \int\{\mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}}\,\mathrm{d}A\,\mathrm{d}\hat{A}\}\,\mathrm{e}^{N\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}]}\int \mathrm{d}\mathbf{x}\; Q_\mu^1(\mathbf{x})g(\mathbf{x}),\end{aligned}\tag{C.5}$$

and for $n\to 0$, using $\int\{\mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}}\,\mathrm{d}A\,\mathrm{d}\hat{A}\}\,\mathrm{e}^{N\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}]}\int Q_\mu^1(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1$, this leads us for $N\to\infty$ to the desired asymptotic result

$$\begin{aligned}\lim_{N\to\infty}\int \mathrm{d}\mathbf{x}\; Q_\mu(\mathbf{x})g(\mathbf{x}) &= \lim_{N\to\infty}\frac{\int\{\mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}}\,\mathrm{d}A\,\mathrm{d}\hat{A}\}\,\mathrm{e}^{N\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}]}\int \mathrm{d}\mathbf{x}\; Q_\mu^1(\mathbf{x})g(\mathbf{x})}{\int\{\mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}}\,\mathrm{d}A\,\mathrm{d}\hat{A}\}\,\mathrm{e}^{N\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}]}}\\ &= \int \mathrm{d}\mathbf{x}\; Q_\mu^1(\mathbf{x})g(\mathbf{x}),\end{aligned}\tag{C.6}$$

where the distribution $Q_\mu^1(\mathbf{x})$ is the solution of equation (43). Thus, assuming that the replica symmetry assumption is correct, the physical meaning of the distribution in the our RS equation (45) is given by (51). Similarly we can work out

$$
\begin{aligned}
A(\nu,\mu) &= \left\langle \langle A(\nu,\mu|\mathbf{C},\mathbf{X})\rangle_{\mathbf{C}|\mathbf{X}} \right\rangle_{\mathbf{X}} \\
&= \int \mathrm{d}\mathbf{X}\, P(\mathbf{X}|L) \sum_{\mathbf{C}} P_\beta(\mathbf{C}|\mathbf{X}) A(\nu,\mu|\mathbf{C},\mathbf{X}) \\
&= \sum_{\tilde{\mathbf{C}}} q(\tilde{\mathbf{C}}|L) \int \mathrm{d}\mathbf{X}\, P(\mathbf{X}|\tilde{\mathbf{C}}) \sum_{\mathbf{C}} P_\beta(\mathbf{C}|\mathbf{X}) A(\nu,\mu|\mathbf{C},\mathbf{X}) \\
&= \sum_{\tilde{\mathbf{C}}} q(\tilde{\mathbf{C}}|L) \int \mathrm{d}\mathbf{X}\, P(\mathbf{X}|\tilde{\mathbf{C}}) \sum_{\mathbf{C}} P_\beta(\mathbf{C}|\mathbf{X}) \Big[\frac{1}{N}\sum_{i=1}^{N} c_{i\mu}\tilde{c}_{i\nu}\Big],
\end{aligned} \tag{C.7}
$$

where we used the definitions $\tilde{c}_{i\nu} = \mathbb{1}\left[\mathbf{x}_i \sim q_\nu(\mathbf{x})\right]$ and $P(\mathbf{X}|\mathbf{C}) = \prod_{\nu=1}^{L}\prod_{i=1}^{N} q_\nu^{c_{i\nu}}(\mathbf{x}_i)$. Substitution of the definition of $P_\beta(\mathbf{C}|\mathbf{X})$ allows us to work out the average further:

$$
\begin{aligned}
A(\nu,\mu) &= \sum_{\tilde{\mathbf{C}}} q(\tilde{\mathbf{C}}|L) \int \mathrm{d}\mathbf{X}\, P(\mathbf{X}|\tilde{\mathbf{C}}) \sum_{\mathbf{C}} \frac{P(\mathbf{C}|K)}{Z_\beta(\mathbf{X})} \mathrm{e}^{-\beta N \hat{F}_N(\mathbf{C},\mathbf{X})} \Big[\frac{1}{N}\sum_{i=1}^{N} c_{i\mu}\tilde{c}_{i\nu}\Big] \\
&= \lim_{n\to 0} \sum_{\tilde{\mathbf{C}}} q(\tilde{\mathbf{C}}|L) \int \mathrm{d}\mathbf{X}\, P(\mathbf{X}|\tilde{\mathbf{C}}) \sum_{\mathbf{C}} P(\mathbf{C}|K) \mathrm{e}^{-\beta N \hat{F}_N(\mathbf{C},\mathbf{X})} \\
&\qquad\qquad \times Z_\beta^{n-1}(\mathbf{X}) \Big[\frac{1}{N}\sum_{i=1}^{N} c_{i\mu}\tilde{c}_{i\nu}\Big] \\
&= \sum_{\mathbf{C}} q(\mathbf{C}|L) \sum_{\mathbf{C}^1} \cdots \sum_{\mathbf{C}^n} \Big[\prod_{\alpha=1}^{n} P(\mathbf{C}^\alpha|K)\Big] \int \mathrm{d}\mathbf{X}\, P(\mathbf{X}|\mathbf{C}) \\
&\qquad\qquad \times \mathrm{e}^{-\beta N \sum_{\alpha=1}^{n} \hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})} \Big[\frac{1}{N}\sum_{i=1}^{N} c_{i\nu} c_{i\mu}^{1}\Big] \\
&= \left\langle \left\langle \left\langle \mathrm{e}^{-\beta N \sum_{\alpha=1}^{n} \hat{F}_N(\mathbf{C}^\alpha,\mathbf{X})} \sum_{\boldsymbol{\mu}} \delta_{\mu;\mu_1} A(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\}) \right\rangle_{\mathbf{X}|\mathbf{C}} \right\rangle_{\mathbf{C}} \right\rangle_{\{\mathbf{C}^\alpha\}},
\end{aligned} \tag{C.8}
$$

in which $A(\nu,\boldsymbol{\mu}|\mathbf{C},\{\mathbf{C}^\alpha\})$ is defined in equation (A.7). The above expression can now be used, following the same steps as for the $Q_\mu(\mathbf{x})$ order parameter, to show that for $N \to \infty$ and $n \to 0$ the following will hold:

$$
\begin{aligned}
\lim_{N\to\infty} A(\nu,\mu) &= \lim_{N\to\infty} \frac{\int \{\mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}}\,\mathrm{d}A\,\mathrm{d}\hat{A}\}\, \mathrm{e}^{N\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}]} \sum_{\boldsymbol{\mu}} \delta_{\mu;\mu_1} A(\nu,\boldsymbol{\mu})}{\int \{\mathrm{d}\mathbf{Q}\,\mathrm{d}\hat{\mathbf{Q}}\,\mathrm{d}A\,\mathrm{d}\hat{A}\}\, \mathrm{e}^{N\Psi[\{\mathbf{Q},\hat{\mathbf{Q}}\};\{A,\hat{A}\}]}} \\
&= \sum_{\boldsymbol{\mu}} \delta_{\mu;\mu_1} A(\nu,\boldsymbol{\mu}),
\end{aligned} \tag{C.9}
$$

where $A(\nu,\boldsymbol{\mu})$ is the solution of equation (43). From this we deduce that (52) indeed gives the physical meaning of the RS expression (46).

Appendix D. Average energy

In this appendix we compute the average energy

$$
\begin{aligned}
e(\beta) &= \left\langle \left\langle \hat{F}_N(\mathbf{C}, \mathbf{X}) \right\rangle_{\mathbf{C}|\mathbf{X}} \right\rangle_{\mathbf{X}} \\
&= \lim_{n\to 0} \left\langle \sum_{\mathbf{C}} P(\mathbf{C}|K) \mathrm{e}^{-\beta N \hat{F}_N(\mathbf{C},\mathbf{X})} Z_\beta^{n-1}(\mathbf{X}) \hat{F}_N(\mathbf{C}, \mathbf{X}) \right\rangle_{\mathbf{X}},
\end{aligned} \tag{D.1}
$$

where iwe used the replica identity (C.3). Assuming initially that $n \in \mathbb{N}$ allows us to compute the average over $\mathbf{X}$ in the above expression as follows

$$
\begin{aligned}
&\left\langle \left\langle \mathrm{e}^{-\beta N \sum_{\alpha=1}^n \hat{F}_N(\mathbf{C}^\alpha, \mathbf{X})} \hat{F}_N(\mathbf{C}^1, \mathbf{X}) \right\rangle_{\mathbf{X}} \right\rangle_{\{\mathbf{C}^\alpha\}} \\
&= \left\langle \left\langle \mathrm{e}^{-\beta N \sum_{\alpha=1}^n \hat{F}_N(\mathbf{C}^\alpha, \mathbf{X})} \left[-\frac{1}{N} \sum_{\mu=1}^K \log \left\langle \mathrm{e}^{\sum_{i=1}^N c_{i\mu}^1 \log P(\mathbf{x}_i|\boldsymbol{\theta})} \right\rangle_{\boldsymbol{\theta}} \right] \right\rangle_{\mathbf{X}} \right\rangle_{\{\mathbf{C}^\alpha\}} \\
&= -\left\langle \left\langle \mathrm{e}^{-\beta N \sum_{\alpha=1}^n \hat{F}_N(\mathbf{C}^\alpha, \mathbf{X})} \left[\frac{1}{N} \sum_{\mu=1}^K \log \left\langle \mathrm{e}^{N \int \mathrm{d}\mathbf{x}\; Q_\mu(\mathbf{x}|\mathbf{C}^1, \mathbf{X}) \log P(\mathbf{x}|\boldsymbol{\theta})} \right\rangle_{\boldsymbol{\theta}} \right] \right\rangle_{\mathbf{X}} \right\rangle_{\{\mathbf{C}^\alpha\}}
\end{aligned} \tag{D.2}
$$

and, with the short-hand $\Psi[\ldots] = \Psi[\{\mathbf{Q}, \hat{\mathbf{Q}}\}; \{A, \hat{A}\}]$ and after taking the replica limit $n \to 0$ within the RS ansatz, we then arrive at equation (59):

$$
\begin{aligned}
&\lim_{n\to 0} \lim_{N\to\infty} \left\langle \left\langle \mathrm{e}^{-\beta N \sum_{\alpha=1}^n \hat{F}_N(\mathbf{C}^\alpha, \mathbf{X})} \hat{F}_N(\mathbf{C}^1, \mathbf{X}) \right\rangle_{\mathbf{X}} \right\rangle_{\{\mathbf{C}^\alpha\}} \\
&= -\lim_{N\to\infty} \frac{\int \{\mathrm{d}\mathbf{Q}\, \mathrm{d}\hat{\mathbf{Q}}\, \mathrm{d}A\, \mathrm{d}\hat{A}\}\, \mathrm{e}^{N\Psi[\ldots]} \left[\frac{1}{N} \sum_{\mu=1}^K \log \left\langle \mathrm{e}^{N \int \mathrm{d}\mathbf{x}\; Q_\mu^1(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\theta})} \right\rangle_{\boldsymbol{\theta}} \right]}{\int \{\mathrm{d}\mathbf{Q}\, \mathrm{d}\hat{\mathbf{Q}}\, \mathrm{d}A\, \mathrm{d}\hat{A}\}\, \mathrm{e}^{N\Psi[\ldots]}} \\
&= -\lim_{N\to\infty} \frac{1}{N} \sum_{\mu=1}^K \log \left\langle \mathrm{e}^{N \int \mathrm{d}\mathbf{x}\; Q_\mu^1(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\theta})} \right\rangle_{\boldsymbol{\theta}} \\
&= -\sum_{\mu=1}^K \max_{\boldsymbol{\theta}} \int \mathrm{d}\mathbf{x}\; Q_\mu^1(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\theta}).
\end{aligned} \tag{D.3}
$$

Appendix E. ‘Sphericity’ of normally distributed samples

Here we show that almost all points of any random sample from the d-dimensional Normal distribution $\mathcal{N}(\mathbf{x}|\mathbf{m}, \boldsymbol{\Sigma})$, with mean $\mathbf{m}$ and covariance $\boldsymbol{\Sigma}$, lie in the annulus $d(\lambda_{\max} - \epsilon) < ||\mathbf{x} - \mathbf{m}||^2 < d(\lambda_{\max} + \epsilon)$, where $||\cdots||$ is the Euclidean norm and $\lambda_{\max}$ is the maximum eigenvalue of $\boldsymbol{\Sigma}$, for sufficiently large d and $0 < \epsilon \ll 1$. If $\mathbf{x}$ is sampled from $\mathcal{N}(\mathbf{x}|\mathbf{m}, \boldsymbol{\Sigma})$, then $\langle ||\mathbf{x} - \mathbf{m}||^2 \rangle = \mathrm{Tr}(\boldsymbol{\Sigma})$. We want to bound the following probability:

$$
\begin{aligned}
&\mathrm{Prob}(||\mathbf{x} - \mathbf{m}||^2 \notin (\mathrm{Tr}(\boldsymbol{\Sigma}) - d\epsilon, \mathrm{Tr}(\boldsymbol{\Sigma}) + d\epsilon)) \\
&\quad = \mathrm{Prob}(||\mathbf{x} - \mathbf{m}||^2 \leqslant \mathrm{Tr}(\boldsymbol{\Sigma}) - d\epsilon) + \mathrm{Prob}(||\mathbf{x} - \mathbf{m}||^2 \geqslant \mathrm{Tr}(\boldsymbol{\Sigma}) + d\epsilon).
\end{aligned} \tag{E.1}
$$

Firstly, for sufficienty small positive α we can use the Markov inequality to obtain

$$\begin{aligned}&\mathrm{Prob}(||\mathbf{x}-\mathbf{m}||^2 \geqslant \mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon)\\ &\quad=\mathrm{Prob}(\mathrm{e}^{\frac{\alpha}{2}||\mathbf{x}-\mathbf{m}||^2}\geqslant \mathrm{e}^{\frac{\alpha}{2}(\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon)})\leqslant\left\langle \mathrm{e}^{\frac{\alpha}{2}||\mathbf{x}-\mathbf{m}||^2}\right\rangle \mathrm{e}^{-\frac{\alpha}{2}(\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon)}\\ &\quad=\mathrm{e}^{-\frac{1}{2}(\log|\mathbf{I}-\alpha\boldsymbol{\Sigma}|+\alpha(\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon))}.\end{aligned}\tag{E.2}$$

The last line, which assumes that $\boldsymbol{\Sigma}^{-1}-\alpha\mathbf{I}$ is positive definite, follows from (78). Denoting the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ by $\lambda_1,\ldots,\lambda_d$, we can bound $\log|\mathbf{I}-\alpha\boldsymbol{\Sigma}| = \sum_{\ell=1}^{d}\log(1-\alpha\lambda(\ell))$ from below by $d\log(1-\alpha\lambda_{\max})$, where $\lambda_{\max}=\max_\ell\lambda(\ell)$. Using this in (E.2) gives us the simpler inequality

$$\mathrm{Prob}(||\mathbf{x}-\mathbf{m}||^2 \geqslant \mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon)\leqslant \mathrm{e}^{-\frac{1}{2}(d\log(1-\alpha\lambda_{\max})+\alpha(\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon))}.\tag{E.3}$$

The function $d\log(1-\alpha\lambda_{\max})+\alpha(\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon)$ is found to have its maximum at $\alpha=(\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon-d\lambda_{\max})/(\lambda_{\max}(\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon))$, which allows us to optimise the upper bound in (E.3) and produce the inequality

$$\mathrm{Prob}(||\mathbf{x}-\mathbf{m}||^2 \geqslant \mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon)\leqslant \exp\Big[-\frac{d}{2}\Phi\Big(\frac{d\lambda_{\max}}{\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon}\Big)\Big],\tag{E.4}$$

where $\Phi(x)=\log(x)+x^{-1}-1$. We note that $\Phi(x)\geqslant 0$, by the inequality $\log(x)\geqslant 1-\frac{1}{x}$. Also, $\Phi(x)$ is monotonic increasing (decreasing) for $x>1$ ($x<1$), and is exactly zero when $x=1$. Secondly, we derive a similar bound for the second probability in (E.1):

$$\begin{aligned}&\mathrm{Prob}(||\mathbf{x}-\mathbf{m}||^2 \leqslant \mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon)\\ &\quad=\mathrm{Prob}(\mathrm{e}^{-\frac{\alpha}{2}||\mathbf{x}-\mathbf{m}||^2}\geqslant \mathrm{e}^{-\frac{\alpha}{2}(\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon)})\leqslant\left\langle \mathrm{e}^{-\frac{\alpha}{2}||\mathbf{x}-\mathbf{m}||^2}\right\rangle \mathrm{e}^{\frac{\alpha}{2}(\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon)}\\ &\quad=\mathrm{e}^{-\frac{1}{2}(\log|\mathbf{I}+\alpha\boldsymbol{\Sigma}|-\alpha(\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon))}.\end{aligned}\tag{E.5}$$

Now $\log|\mathbf{I}+\alpha\boldsymbol{\Sigma}|=\sum_{\ell=1}^{d}\log(1+\alpha\lambda(\ell))$ is bounded from below by $d\log(1+\alpha\lambda_{\min})$, where $\lambda_{\min}=\min_\ell\lambda(\ell)$. Using this in (E.5) gives us the inequality

$$\mathrm{Prob}(||\mathbf{x}-\mathbf{m}||^2 \leqslant \mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon)\leqslant \mathrm{e}^{-\frac{1}{2}(d\log(1+\alpha\lambda_{\min})-\alpha(\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon))}.\tag{E.6}$$

We note that the quantity $d\log(1+\alpha\lambda_{\min})-\alpha(\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon)$ takes its maximum for $\alpha=(\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon+d\lambda_{\min})/(\lambda_{\min}(\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon))$, which in (E.6), gives the new bound

$$\mathrm{Prob}(||\mathbf{x}-\mathbf{m}||^2 \leqslant \mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon)\leqslant \exp\Big[-\frac{d}{2}\Phi\Big(\frac{d\lambda_{\min}}{\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon}\Big)\Big].\tag{E.7}$$

By using the two inequalities (E.4) and (E.7) in (E.1), we obtain the inequality

$$\begin{aligned}&\mathrm{Prob}(||\mathbf{x}-\mathbf{m}||^2 \notin (\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon,\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon))\\ &\quad\leqslant 2\exp\Big[-\frac{d}{2}\min\Big\{\Phi\Big(\frac{d\lambda_{\min}}{\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon}\Big),\Phi\Big(\frac{d\lambda_{\max}}{\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon}\Big)\Big\}\Big].\end{aligned}\tag{E.8}$$

Moreover, since $\mathrm{Tr}(\boldsymbol{\Sigma})\leqslant d\lambda_{\max}$, we may also write

$$\begin{aligned}&\mathrm{Prob}(||\mathbf{x}-\mathbf{m}||^2 \notin (\mathrm{Tr}(\boldsymbol{\Sigma})-d\epsilon,\mathrm{Tr}(\boldsymbol{\Sigma})+d\epsilon))\\ &\quad\leqslant 2\exp\Big[-\frac{d}{2}\min\Big\{\Phi\Big(\frac{\lambda_{\min}}{\lambda_{\max}-\epsilon}\Big),\Phi\Big(\frac{\lambda_{\max}}{\lambda_{\max}+\epsilon}\Big)\Big\}\Big].\end{aligned}\tag{E.9}$$

The remaining extrema are given by

$$\epsilon \in (0,\epsilon_1):\ \min\Big\{\Phi\Big(\frac{\lambda_{\min}}{\lambda_{\max}-\epsilon}\Big),\Phi\Big(\frac{\lambda_{\max}}{\lambda_{\max}+\epsilon}\Big)\Big\} = \Phi\Big(\frac{\lambda_{\max}}{\lambda_{\max}+\epsilon}\Big) \tag{E.10}$$

$$\epsilon \in (\epsilon_1,\epsilon_2):\ \min\Big\{\Phi\Big(\frac{\lambda_{\min}}{\lambda_{\max}-\epsilon}\Big),\Phi\Big(\frac{\lambda_{\max}}{\lambda_{\max}+\epsilon}\Big)\Big\} = \Phi\Big(\frac{\lambda_{\max}}{\lambda_{\max}-\epsilon}\Big) \tag{E.11}$$

with

$$\epsilon_1 = \frac{\lambda_{\max}(\lambda_{\max}-\lambda_{\min})}{\lambda_{\max}+\lambda_{\min}},\quad \epsilon_2 = \lambda_{\max}-\lambda_{\min}. \tag{E.12}$$

Furthermore, when $\lambda_{\max}=\lambda_{\min}=\lambda$, i.e. $\mathbf{\Sigma}=\lambda\mathbf{I}$, one obtains

$$\epsilon \in (0,\lambda):\min\Big\{\Phi\Big(\frac{\lambda}{\lambda-\epsilon}\Big),\Phi\Big(\frac{\lambda}{\lambda+\epsilon}\Big)\Big\} = \Phi\Big(\frac{\lambda}{\lambda+\epsilon}\Big). \tag{E.13}$$

If, in contrast, we observe a sample $\mathbf{x}_1,\ldots,\mathbf{x}_N$ from $\mathcal{N}(\mathbf{x}|\mathbf{m},\mathbf{\Sigma})$, instead of a single vector $\mathbf{x}$, then the probability $\mathrm{Prob}(\cup_{i=1}^N\big\{||\mathbf{x}_i-\mathbf{m}||^2\notin(\mathrm{Tr}(\mathbf{\Sigma})-d\epsilon,\mathrm{Tr}(\mathbf{\Sigma})+d\epsilon)\big\})$ that at least one of the events $||\mathbf{x}_i-\mathbf{m}||^2\notin(\mathrm{Tr}(\mathbf{\Sigma})-d\epsilon,\mathrm{Tr}(\mathbf{\Sigma})+d\epsilon)$ occurs, can be bounded by combining Boole's inequality with inequalities (E.4) and (E.8):

$$\begin{aligned}&\mathrm{Prob}(\cup_{i=1}^N\big\{||\mathbf{x}_i-\mathbf{m}||^2\notin(\mathrm{Tr}(\mathbf{\Sigma})-d\epsilon,\mathrm{Tr}(\mathbf{\Sigma})+d\epsilon)\big\})\\ &\quad\leqslant\sum_{i=1}^N\mathrm{Prob}(||\mathbf{x}_i-\mathbf{m}||^2\notin(\mathrm{Tr}(\mathbf{\Sigma})-d\epsilon,\mathrm{Tr}(\mathbf{\Sigma})+d\epsilon))\\ &\quad\leqslant 2N\exp\Big[-\frac{d}{2}\min\Big\{\Phi\Big(\frac{d\lambda_{\min}}{\mathrm{Tr}(\mathbf{\Sigma})-d\epsilon}\Big),\Phi\Big(\frac{d\lambda_{\max}}{\mathrm{Tr}(\mathbf{\Sigma})+d\epsilon}\Big)\Big\}. \end{aligned}\tag{E.14}$$

Repeating similar steps to those followed earlier then gives for $\lambda_{\max}>\lambda_{\min}$:

$$\mathrm{Prob}(\cup_{i=1}^N\big\{||\mathbf{x}_i-\mathbf{m}||^2\notin(\mathrm{Tr}(\mathbf{\Sigma})-d\epsilon,\mathrm{Tr}(\mathbf{\Sigma})+d\epsilon)\big\})\leqslant 2N\exp\Big[-\frac{d}{2}\Phi\Big(\frac{\lambda_{\max}}{\lambda_{\max}+\epsilon}\Big)\Big] \tag{E.15}$$

provided $\epsilon\in(0,\lambda_{\max}(\lambda_{\max}-\lambda_{\min})/(\lambda_{\max}+\lambda_{\min}))$, whereas for $\mathbf{\Sigma}=\lambda\mathbf{I}$ we have

$$\mathrm{Prob}(\cup_{i=1}^N\big\{||\mathbf{x}_i-\mathbf{m}||^2\notin(\mathrm{Tr}(\mathbf{\Sigma})-d\epsilon,\mathrm{Tr}(\mathbf{\Sigma})+d\epsilon)\big\})\leqslant 2N\exp\Big[-\frac{d}{2}\Phi\Big(\frac{\lambda}{\lambda+\epsilon}\Big)\Big], \tag{E.16}$$

provided $\epsilon\in(0,\lambda)$. It is now clear that there is a function $d(\epsilon,\lambda_{\max},N)>0$ such that for $d>d(\epsilon,\lambda_{\max},N)$ almost all points of a sample from $\mathcal{N}(\mathbf{x}|\mathbf{m},\mathbf{\Sigma})$ lie in the annulus[11] $\sqrt{d(\lambda_{\max}-\epsilon)}<||\mathbf{x}-\mathbf{m}||<\sqrt{d(\lambda_{\max}+\epsilon)}$.

ORCID iDs

Alexander Mozeika https://orcid.org/0000-0003-1514-1650

[11] For small d, the bound in (E.15) is very loose, so it makes more sense to consider the probability that $\cup_{i\leqslant N}\{||\mathbf{x}_i-\mathbf{m}||^2\geqslant d(\lambda_{\max}+\epsilon)\}$, i.e. that at least one $\mathbf{x}_i$ in the sample $\mathbf{X}$ lies outside the ball $\mathcal{B}_{\sqrt{d(\lambda_{\max}+\epsilon)}}(\mathbf{m})$, given by $\mathrm{Prob}(\cup_{i\leqslant N}\{\mathbf{x}_i\notin\mathcal{B}_{\sqrt{d(\lambda_{\max}+\epsilon)}}(\mathbf{m})\})\leqslant N\exp[-\frac{d}{2}\Phi(\frac{\lambda_{\max}}{\lambda_{\max}+\epsilon})]$.

References

[1] Advani M and Ganguli S 2016 *Phys. Rev.* X **6** 031034

[2] Mozeika A, Dikmen O and Piili J 2014 *Phys. Rev.* E **90** 010101

[3] Coolen A C C, Barrett J E, Paga P and Perez-Vicente C J 2017 *J. Phys. A: Math. Theor.* **50** 375001

[4] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: an Introduction* (Oxford: Oxford University Press)

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

[6] de Souza R S, Dantas M L L, Costa-Duarte M V, Feigelson E D, Killedar M, Lablanche P Y, Vilalta R, Krone-Martins A, Beck R and Gieseke F 2017 *Mon. Not. R. Astron. Soc.* **472** 2808

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

[8] Bishop C M 2006 *Pattern Recognition and Machine Learning* (Berlin: Springer)

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

[10] Guihenneuc-Jouyaux C and Rousseau J 2005 *J. Comput. Graph. Stat.* **14** 75

[11] Scrucca L, Fop M, Murphy T B and Raftery A E 2016 *R. J.* **8** 205

[12] Barber D 2012 *Bayesian Reasoning and Machine Learning* (Cambridge: Cambridge University Press)

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

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

[15] Łuksza M, Lässig M and Berg J 2010 *Phys. Rev. Lett.* **105** 220601

[16] Watkin T L H and Nadal J-P 1994 *J. Phys. A: Math. Gen.* **27** 1899

[17] Barkai N and Sompolinsky H 1994 *Phys. Rev.* E **50** 1766

[18] Biehl M and Mietzner A 1994 *J. Phys. A: Math. Gen.* **27** 1885

[19] Lesieur T, De Bacco C, Banks J, Krzakala F, Moore C and Zdeborová L 2016 *54th Annual Allerton Conf. on Communication, Control, and Computing (Allerton)* p 601

[20] Corander J, Gyllenberg M and Koski T 2009 *Adv. Data Anal. Class.* **3** 3

[21] Mozeika A and Coolen A C C 2018 *Phys. Rev.* E **98** 042133

[22] Mézard M, Parisi G and Virasoro M 1987 *Spin Glass Theory and Beyond: an Introduction to the Replica Method and its Applications* (Singapore: World Scientific)

[23] Rennie B C and Dobson A J 1969 *J. Comb. Theory* **7** 116

[24] Cover T M and Thomas J A 2012 *Elements of Information Theory* (New York: Wiley)

[25] De Bruijn N G 1981 *Asymptotic Methods in Analysis* (New York: Dover)

[26] Dasgupta S 1999 *40th Annual Symp. on Foundations of Computer Science* p 634

[27] Manning C, Raghavan P and Schütze H 2008 *Introduction to Information Retrieval* (Cambridge: Cambridge University Press)

[28] Rand W M 1971 *J. Am. Stat. Assoc.* **66** 846

[29] Gorban A N and Tyukin I Y 2018 *Phil. Trans. R. Soc.* A **376** 20170237

[30] Barkai N, Seung H S and Sompolinsky H 1993 *Phys. Rev. Lett.* **70** 3167

[31] Shalabi A, Inoue M, Watkins J, De Rinaldis E and Coolen A C C 2018 *Stat. Methods Med. Res.* **27** 336