Bernoulli **26**(2), 2020, 1535–1559
https://doi.org/10.3150/19-BEJ1173

Reliable clustering of Bernoulli mixture models

AMIR NAJAFI[1,*], SEYED ABOLFAZL MOTAHARI[1,†] and HAMID R. RABIEE[2]

[1]*Bioinformatics Research Lab (BRL), Computer Engineering Dept., Sharif University of Technology, Tehran, Iran. E-mail:* [*]*najafy@ce.sharif.edu;* [†]*motahari@sharif.edu*
[2]*Data Science and Machine Learning Lab (DML), Computer Engineering Dept., Sharif University of Technology, Tehran, Iran. E-mail: rabiee@sharif.edu*

A Bernoulli Mixture Model (BMM) is a finite mixture of random binary vectors with independent dimensions. The problem of clustering BMM data arises in a variety of real-world applications, ranging from population genetics to activity analysis in social networks. In this paper, we analyze the clusterability of BMMs from a theoretical perspective, when the number of clusters is unknown. In particular, we stipulate a set of conditions on the sample complexity and dimension of the model in order to guarantee the Probably Approximately Correct (PAC)-clusterability of a dataset. To the best of our knowledge, these findings are the first non-asymptotic bounds on the sample complexity of learning or clustering BMMs.

Keywords: high-dimensional statistics; mixture model analysis; PAC-learnability; sample complexity

1. Introduction

Demixing data samples from mixture models, also called model-based clustering, has long been studied by statisticians and computer scientists. Although plenty of promising algorithms have been introduced in this area, see [7,22,30,32], fewer efforts have been focused on deriving theoretical guarantees on reliable clustering of data samples. The aim of this paper is to elaborate on this shortcoming by deriving analytic guarantees on the clusterability of a particular case of interest: Bernoulli Mixture Models (BMM).

A Bernoulli Model (BM) refers to a random binary vector $\boldsymbol{X} = [X_1, \ldots, X_L] \in \{0, 1\}^L$ with independent random components, where L denotes the model dimension and each X_i is a Bernoulli random variable with success probability (or frequency) p_i, that is, $X_i \sim \mathrm{Bern}(p_i)$. Let us define $\boldsymbol{p} := [p_1, \ldots, p_L] \in [0, 1]^L$. Then, $\mathbb{P}_{\mathrm{BM}}(\boldsymbol{X}; \boldsymbol{p})$ denotes the probability distribution of a Bernoulli model with frequency vector $\boldsymbol{p}$:

$$\mathbb{P}_{\mathrm{BM}}(\boldsymbol{X}; \boldsymbol{p}) := \prod_{\ell=1}^{L} {p_\ell}^{X_\ell} (1 - p_\ell)^{1 - X_\ell}.$$

In this regard, a BMM is defined as a mixture of a finite number of Bernoulli models [24]. Mathematically speaking, the probability distribution of a BMM can be expressed as

$$\mathbb{P}_{\mathrm{BMM}}\big(\boldsymbol{X}; K, \boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(K)}, \boldsymbol{w}\big) = \sum_{k=1}^{K} w_k \mathbb{P}_{\mathrm{BM}}\big(\boldsymbol{X}; \boldsymbol{p}^{(k)}\big), \tag{1.1}$$

where $K \in \mathbb{N}$ denotes the number of mixture components (or clusters), $\{\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(K)}\}$ is the set of frequency vectors associated to mixture components, and $\boldsymbol{w} = (w_1, \ldots, w_K)$ is the mixture weight vector with $\sum_k w_k = 1$, and $w_k \geq 0$. Let $\boldsymbol{P}$ be a $K \times L$ frequency matrix with $\boldsymbol{p}^{(k)}$ as its kth row. For simplicity, we denote $\mathcal{B} = \mathcal{B}(K, \boldsymbol{P}, \boldsymbol{w})$ as the BMM with the above-mentioned parameters and specifications. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \{0, 1\}^L$ be n i.i.d. sample vectors drawn from $\mathcal{B}$. We define $\boldsymbol{\mathcal{X}}$ as the matrix $[\boldsymbol{X}_1 | \cdots | \boldsymbol{X}_n]^T \in \{0, 1\}^{n \times L}$ to represent a given dataset.

The problem that we have tackled in this paper is the clustering of rows of $\boldsymbol{\mathcal{X}}$, such that clusters with probability at least $1 - \zeta$ are approximately (up to an ϵ fraction of mis-clustering) correct, for arbitrarily small $\epsilon, \zeta > 0$. In order to guarantee the information-theoretic possibility of such clustering, that is, without considering the required computational cost, we establish novel bounds in the form of $n \geq \text{poly}(\epsilon^{-1}, \zeta^{-1})$ and $L > \text{poly}(\epsilon^{-1})$, where poly refers to a polynomial function. It should be noted that K, $\boldsymbol{P}$ and $\boldsymbol{w}$ are all assumed to be unknown. Statisticians have been studying BMMs for a long time [3,5,44,47]. However, PAC-learnability (or PAC-clusterability)[1] of BMMs in terms of the minimum required sample size and/or model dimension has remained an open problem. To the best of our knowledge, this paper is the first attempt toward this goal by deriving a set of non-asymptotic conditions under which reliable clustering is possible.

The paper is organized as follows: In Section 2, related works are discussed. Section 3 formally presents our main results, where proofs and further discussions are given in Section 4. Finally, conclusions are made in Section 5.

2. Related works

Employment of BMMs in order to model multi-dimensional categorical data goes back to [27], while more detailed mathematical and historical explanations can be found, for example, in [6,28, 29]. In two classic works [5] and [10], a series of heuristic measures have been introduced to assess the number of mixture components in a BMM; However, their performance is validated only through experimental investigations. Authors of these papers have *conjectured* that learnability is possible as long as independence holds between cluster parameters, while their studies lack a theoretical sufficiency analysis. From an algorithmic point of view, the Expectation–Maximization (EM) algorithm is the most widely used framework for statistical inference in BMMs (see [28] and [18]). In [17], a popular EM-based technique for unsupervised learning of finite mixture models (including BMMs) is introduced, which makes no assumption on the number of mixture components. Also, see [19] for another well-cited paper on model-based clustering of mixture model data. In [23], EM is employed for parameter initialization in a number of existing inference algorithms on BMMs. From a theoretical perspective, a set of statistical guarantees on the convergence of EM algorithm in mixture model problems has been recently given in [4], however, authors have mainly focused on Gaussian distributions rather than Bernoulli models. So far, theoretical analysis of Gaussian Mixture Models (GMM) have been more successful compared

[1] In this paper, the notion of PAC-learnability is used in the information-theoretic sense, and to address those learning tasks that can be learned with a polynomial sample complexity w.r.t. ϵ and ζ. This notation is consistent with that of [31]. For those cases where the learning algorithm has also a polynomial computational complexity, the term "efficiently PAC-learnable" has been used.

to many discrete mixture models [14,25]. This might be due to both the continuous nature, and also the more-favorable analytic form of Gaussian distributions [11]. Recently, nearly tight lower and upper-bounds on the sample complexity of learning GMM distributions have been derived [2].

Our work is also related to Bayesian non-parametric approaches in the sense that the number of clusters is open-ended, and will be inferred based on the observed data. Some good reviews on non-parametric approaches in statistics can be found, for example, in [20,34,41]. In particular, [43] has proposed a unified non-parametric framework for model-based clustering with the use of hierarchical Dirichlet mixtures. All of the studies on BMMs that we have been reviewed so far share a common property: at their best, they only prove convergence to a sub-optimal likelihood value, rather than providing guarantees on the accuracy of the final clustering/learning. Also, sample complexity lower-bounds, that is, minimum required sample size n or dimension L, for either reliable learning or clustering of BMMs is still an open problem.

From a geneticist's point of view, this paper basically builds upon the statistical model presented in [38]. Based on this model, the genetic sequence of an individual from a particular specie can be represented with a binary vector, where each dimension denotes the *absence* or *presence* of a certain genetic variant. In addition, for the majority of cases, dimensions can be assumed to be statistically independent from each other. According to [38], one can model the effect of population inhomogeneity (the presence of population mixtures in a biological dataset) via BMMs. More discussions on this issue can be found, for example, in [36,45] and [26]. In [15], authors have performed a simulation study based on [38], in order to assess the number of clusters in a given population. A number of software packages for computational population analysis can be found in [9,26,35,40], which mainly focus on binary datasets, the same configuration that we have considered in this paper. Our problem setting encompasses both models described in [38] and [16], since we make no restrictive assumptions, such as *independence*, on the latent frequencies. Recently, Genome-Wide Association Studies (GWAS) which involve associating human diseases to genetic variants have gained huge popularity. The role of population stratification in GWAS, an important application of BMMs in genetics research, is discussed in [45] and [39]. For more research on the employment of BMMs in GWA studies, see [33,37,48,49].

A critical issue that needs to be discussed here is the following fundamental question: when is a BMM guaranteed to be identifiable? We call a BMM identifiable whenever there is a unique set of parameters $(K, \boldsymbol{P}, \boldsymbol{w})$ that corresponds to its probability measure. Reliable clustering of BMMs seems meaningless if there exist more than one true generative models for the samples. Identifiability of BMMs has been addressed in [21], where authors have shown that BMMs cannot be *strictly* identifiable regardless of their dimension, meaning that there always exist some sets of parameters which result in the same probability distribution. However, it does not mean that for every setting $(K, \boldsymbol{P}, \boldsymbol{w})$ there must exist another parameter set to produce the same model. Motivated by this idea, in [8] authors have investigated practical identifiability of BMMs via computer simulations. In [1], it has been proved that for $L \geq 2\lceil \log_2 K \rceil + 1$, BMMs become *generically* identifiable meaning that sets of parameters with the same probability distributions have a zero Lebesgue measure in the space of parameters. Therefore, we can only focus on identifiable cases without any loss of generality for our results as long as the above condition holds. More precisely, the conditions that we stipulate in this paper to guarantee the clusterability of BMMs also satisfy the identifiability condition, since we guarantee a stronger property that encompasses identifiability.

3. Main result

In this section, we state our main result and explain its implications. Recall that the problem is to reliably cluster a set of i.i.d. samples which are drawn from a BMM with unknown parameters (the number of clusters is also assumed to be unknown). The samples are embedded as rows of the matrix $\boldsymbol{\mathcal{X}}$. In fact, we design an algorithm which outputs a vector $\boldsymbol{Z} \in \{1, 2, \ldots\}^n$ in which the ith element Z_i represents the cluster index of the data sample $\boldsymbol{X}_i$. Ultimately, we compare the output of the algorithm with the true clustering which is denoted by $\boldsymbol{Z}_T \in \{1, 2, \ldots, K\}^n$. In this regard, let us state two definitions in order to make the comparison mathematically concrete.

Definition 3.1 (ϵ-pureness). A selected row sub-matrix of $\boldsymbol{\mathcal{X}}$ is ϵ-pure if at least $1 - \epsilon$ fraction of its rows have the same index in $\boldsymbol{Z}_T$.

Definition 3.2 (ϵ-correctness). A clustering algorithm is ϵ-correct on $\boldsymbol{\mathcal{X}}$ if all the output clusters are ϵ-pure.

Obviously, for a reliable clustering to be feasible, mixture components of the underlying BMM need to be sufficiently far apart from each other. For instance, if a BMM contains two mixture components with exactly the same frequency vectors, no algorithm can index the samples correctly. Therefore, we impose a natural restriction on the parameters of the BMM which makes the clustering a feasible task.

Definition 3.3 ($(\mathcal{L}, \delta)$-separability). A frequency matrix $\boldsymbol{P} \in [0, 1]^{K \times L}$ is said to be $(\mathcal{L}, \delta)$-separable, if for each pair of rows of $\boldsymbol{P}$, say k and k' with $k \neq k'$, there exist at least $\mathcal{L} \leq L$ column indices $\{i_1, \ldots, i_{\mathcal{L}}\} \subseteq \{1, \ldots, L\}$ such that

$$|P_{k,i_\ell} - P_{k',i_\ell}| \geq \delta, \quad \ell = 1, \ldots, \mathcal{L}.$$

We may now present our main result in the form of the following theorem which provides a sufficient sample complexity for reliable clustering of BMMs.

Theorem 3.1 (Non-asymptotic bounds for clusterability of Bernoulli Mixture Models). *Let $\mathcal{B} = \mathcal{B}(K, \boldsymbol{P}, \boldsymbol{w})$ be a BMM with unknown parameters K, $\boldsymbol{P}$ and $\boldsymbol{w}$. However, $\boldsymbol{P}$ is assumed to be $(\mathcal{L}, \delta)$-separable for some $\mathcal{L} \leq L$ and $\delta > 0$, and there exists $0 < \alpha \leq 1$ such that $w_k \geq \alpha$ for all k. Parameters $\mathcal{L}$, δ and α are assumed to be known. Also, we obviously have $K \leq \lceil 1/\alpha \rceil$. Let $\boldsymbol{\mathcal{X}} = [\boldsymbol{X}_1 | \boldsymbol{X}_2 | \cdots | \boldsymbol{X}_n]^T$ be a dataset including n i.i.d. samples drawn from $\mathcal{B}$. Also, assume $\epsilon, \zeta > 0$, such that*

$$\mathcal{L} \geq \frac{B \log^3(1/\epsilon)}{\epsilon^{2+\frac{1-\alpha}{2(\alpha\delta)^2}}} \quad \text{and} \quad n \geq \frac{C \log^3(1/\epsilon)}{\epsilon^{2+\frac{1-\alpha}{2(\alpha\delta)^2}}} \log \frac{L}{\zeta},$$

where B and C are constants w.r.t. ϵ and ζ. Then, there exists a clustering algorithm $\mathscr{A} : \{0, 1\}^{n \times L} \to \mathbb{N}^n$, such that $\mathscr{A}$ clusters $\boldsymbol{\mathcal{X}}$ into at most $\lceil 1/\alpha \rceil$ clusters and is ϵ-correct on $\boldsymbol{\mathcal{X}}$ with probability at least $1 - \zeta$.

Proof of Theorem 3.1 with the mathematical formulation of the constants B and C are given in Section 4. Theorem 3.1 shows the feasibility of the reliable clustering of samples in $\boldsymbol{\mathcal{X}}$, such that clusters with probability at least $1-\zeta$ are ϵ-pure, for arbitrarily small $\epsilon, \zeta > 0$. On the other hand, the imposed conditions on sample size n and the number of *informative dimensions* $\mathcal{L}$ are $n \geq \text{poly}(\epsilon^{-1}, \zeta^{-1})$ and $\mathcal{L} > \text{poly}(\epsilon^{-1})$, respectively.

We have already discussed that assuming a minimum deviation among frequency vectors, such as $(\mathcal{L}, \delta)$-separability, is necessary for reliable clustering of data. Similarly, we also need to upper-bound the cluster number K, otherwise clustering becomes meaningless. For example, without any condition on K, one can always partition a dataset of size n into n distinct clusters where each clusters would be 0-pure and the clustering is ϵ-correct for any $\epsilon \geq 0$.

Once an ϵ-correct clustering is achieved for a dataset $\boldsymbol{\mathcal{X}}$, estimation of frequency matrix $\boldsymbol{P}$ and weight vector $\boldsymbol{w}$ becomes straightforward. In fact, by using Chernoff bound, it is easy to show that each entry of $\boldsymbol{P}$ and each component in $\boldsymbol{w}$ can be estimated with a maximum error of $\epsilon + O(n^{-1/2})$ with high probability. However, as mentioned earlier, we only focus on clustering in this paper and thus do not explain estimation of $\boldsymbol{P}$ and $\boldsymbol{w}$ in more details to avoid any distraction.

3.1. The algorithm

The proof of Theorem 3.1 is based on an algorithm which employs a *pureness check* measure. We propose a new variant of Total Correlation measure (also known as multivariate correlation or multi-information [42,46]) to reliably test whether a given clustering of the dataset $\boldsymbol{\mathcal{X}}$ does include any non ϵ-pure clusters or not. We call this new variant as Maximal Total Correlation (MTC). The core idea for employing such a measure is the following interesting property of BMMs: In a BMM, unlike a single Bernoulli model, different dimensions of the random binary vector are not statistically independent, and thus have positive Mutual Information (MI) w.r.t. each other [28]. Total correlation is a natural extension of mutual information which can handle more than two random variables [46].

Definition 3.4 (Total correlation). Assume $\boldsymbol{Q} \in \{0,1\}^{m \times d}$ to be a row/column sub-matrix of $\boldsymbol{\mathcal{X}}$ (with $m \leq n$ and $d \leq L$). Then, similar to [46] and [42], the empirical total correlation of $\boldsymbol{Q}$, denoted by $\mathcal{D}(\boldsymbol{Q})$, is defined as

$$\mathcal{D}(\boldsymbol{Q}) := \mathcal{D}_{\text{KL}}(\hat{\mathbb{P}}_{1,\boldsymbol{Q}} \parallel \hat{\mathbb{P}}_{2,\boldsymbol{Q}}), \tag{3.1}$$

where $\hat{\mathbb{P}}_{1,\boldsymbol{Q}}$ denotes the empirical probability distribution underlying the d-dimensional rows of $\boldsymbol{Q}$, while $\hat{\mathbb{P}}_{2,\boldsymbol{Q}}$ is defined as:

$$\hat{\mathbb{P}}_{2,\boldsymbol{Q}}(\boldsymbol{X}) := \prod_{\ell=1}^{d} \hat{p}_\ell^{X_\ell} (1-\hat{p}_\ell)^{1-X_\ell}, \quad \forall \boldsymbol{X} \in \{0,1\}^d, \tag{3.2}$$

with $\hat{p}_\ell$ being the empirical frequency of the ℓth column of $\boldsymbol{Q}$, that is,

$$\hat{p}_\ell := \frac{1}{n} \sum_{i=1}^{n} Q_{i,\ell}, \quad \ell = 1, 2, \ldots, d.$$

In fact, $\mathcal{D}(\boldsymbol{Q})$ is the Kullback–Leibler divergence between two distributions obtained under two separate assumptions. Under the first assumption, no restriction is imposed on the origin of the samples and the empirical distribution $\hat{\mathbb{P}}_{1,\boldsymbol{Q}}$ is simply an estimate of the true underlying distribution of the rows of $\boldsymbol{Q}$. Under the second assumption, samples are drawn from a single Bernoulli model and $\hat{\mathbb{P}}_{2,\boldsymbol{Q}}$ can be used as another estimate of the true distribution. Using the language of *types* from information theory, $\hat{\mathbb{P}}_{2,\boldsymbol{Q}}$ is a product of the marginal types along each dimension [13]. Therefore, if the second assumption does hold, then the two distributions become equal as n goes to infinity. In other words, based on the law of large numbers and also the fact that KL-divergence is continuous if its input arguments are discrete distributions, we have:

$$\lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q}) = \mathcal{D}_{\mathrm{KL}}\Big(\lim_{n\to\infty} \hat{\mathbb{P}}_{1,\boldsymbol{Q}} \,\Big\|\, \lim_{n\to\infty} \hat{\mathbb{P}}_{2,\boldsymbol{Q}}\Big) \overset{\text{a.s.}}{=} 0.$$

On the other hand, if $\boldsymbol{Q}$ consists of samples from a BMM with a sufficient level of contributions from different mixture components, then $\lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q})$ is proved to be strictly positive (Lemmas 4.1 and 4.3). Having defined $\mathcal{D}(\boldsymbol{Q})$, one needs to move forward and check whether a subset of samples (a row sub-matrix of $\boldsymbol{\mathcal{X}}$) is ϵ-pure or not. The Maximal Total Correlation (MTC) measure defined next is a tool to achieve this goal.

Definition 3.5 (Maximal Total Correlation). The Maximal Total Correlation (MTC) of $\boldsymbol{Y} \in \{0,1\}^{m\times L}$, a row sub-matrix of $\boldsymbol{\mathcal{X}}$, for a sub-dimension $d \le L$ is defined as

$$\mathcal{D}_{\max}(\boldsymbol{Y}; d) := \max_{\boldsymbol{Q}\in \mathrm{Col}(\boldsymbol{Y};d)} \mathcal{D}(\boldsymbol{Q}),$$

where maximization is taken over $\mathrm{Col}(\boldsymbol{Y}; d)$ which consists of all $\binom{L}{d}$ column sub-matrices of $\boldsymbol{Y}$ with size $m \times d$.

The MTC measure is the main tool used in our proposed clustering strategy which is presented in Algorithm 1. In fact, Algorithm 1 works by searching over all possible clusterings of the dataset $\boldsymbol{\mathcal{X}}$, which have the following two properties: (i) the number of clusters does not exceed $\lceil \frac{1}{\alpha} \rceil$, where α is a lower-bound on the probability of the smallest cluster in $\mathcal{B}$; and (ii) the smallest cluster has at least $\alpha n/2$ members. We start with a single cluster and then increase the number of clusters one by one. Given the conditions of Theorem 3.1, the true clustering $\boldsymbol{Z}_T$ would be in this search space with probability at least $1 - \zeta/3$.

For any given clustering, we check to see whether all the corresponding clusters are ϵ-pure or not. We can do this by evaluating the MTC over each clustered row sub-matrix of $\boldsymbol{\mathcal{X}}$. If $\mathcal{D}_{\max}$ has negligible values (smaller than a pre-defined threshold τ) in all the clusters, then the current clustering is accepted and the program terminates. We show that under the constraints of Theorem 3.1, the probability of accepting a clustering with even one non ϵ-pure cluster is less than $\zeta/3$. On the other hand, the algorithm eventually reaches the true clustering $\boldsymbol{Z}_T$ by assuming that we have not accepted any other candidates up to that point. Again, we show that the probability of rejecting the true clustering $\boldsymbol{Z}_T$ is no more than $\zeta/3$.

Therefore, one can deduce that with probability at least $1 - \zeta$, Algorithm 1 either outputs an ϵ-correct clustering on $\boldsymbol{\mathcal{X}}$ or the true clustering $\boldsymbol{Z}_T$ (which of course is ϵ-correct as well).

Algorithm 1 BMM clustering via Exhaustive Search

Inputs: Dataset $\boldsymbol{\mathcal{X}}$, and parameters $\mathcal{L}, \delta, \epsilon$ and α,

Set $d \leftarrow \frac{1-\alpha}{2(\alpha\delta)^2(1-\epsilon)}(1+\log\frac{1}{\alpha\epsilon})$ (Sub-matrix column size)
Set $\tau \leftarrow \frac{\epsilon}{2}(1+\log\frac{1}{\alpha\epsilon})$ (Pureness test threshold)
Set $\kappa \leftarrow 1$ (Cluster number)

while $\kappa < \lceil\frac{1}{\alpha}\rceil$ **do**
  for $\forall \boldsymbol{Z} \in \{1,\ldots,\kappa\}^n$, where the size of each cluster is at least $\alpha n/2$ **do**
    Set $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_\kappa \leftarrow$ The clustered row sub-matrices of $\boldsymbol{\mathcal{X}}$ based on $\boldsymbol{Z}$.
    if $\mathcal{D}_{\max}(\boldsymbol{Y}_k; d) \le \tau$ for $\forall k = 1,\ldots,\kappa$ **then**
      Set $\boldsymbol{Z}^* \leftarrow \boldsymbol{Z}$, and
      Terminate the program.
    end if
  end for
  Set $\kappa \leftarrow \kappa + 1$
end while

Output: $\boldsymbol{Z}^*$, a clustering of data in $\boldsymbol{\mathcal{X}}$.

The key property in our analysis that has made it possible for Algorithm 1 to work is that our proposed MTC measure can detect the impurity of data clusters with a decision error which decays exponentially w.r.t. $n \times \mathcal{L}$. Even though total correlation has been extensively used in the literature, proving the above-mentioned property and using this measure for clustering discrete mixture model data is novel.

3.2. Discussions

Both MTC and its parent, that is, total correlation, are very powerful in differentiating between pure and non-pure groups of samples. In this section, we elaborate on this fact. Assume a random vector $\boldsymbol{X} \in \{0,1\}^L$ with $\boldsymbol{X} \sim \mathcal{B}(K, \boldsymbol{P}, \boldsymbol{w})$. If we condition on $\boldsymbol{X}$ to be drawn from a particular mixture component of $\mathcal{B}$, say the kth one with $k \in \{1,2,\ldots,K\}$, then the probability distribution of $\boldsymbol{X}$ would be

$$\mathbb{P}(\boldsymbol{X}|k) = \prod_{\ell=1}^{L} \mathbb{P}(X_\ell|k). \tag{3.3}$$

However, according to (1.1), the distribution of $\boldsymbol{X}$ without this assumption is $\mathbb{P}(\boldsymbol{X}) = \sum_k w_k \mathbb{P}(\boldsymbol{X}|k)$. A more subtle comparison of $\mathbb{P}(\boldsymbol{X})$ and $\mathbb{P}(\boldsymbol{X}|k)$ simply reveals that in a mixture model, unlike the case of a single Bernoulli model, different dimensions of the vector are not necessarily independent from each other. This argument can be qualitatively justified as follows: a group of observed dimensions can convey information about the mixture component to which

$\boldsymbol{X}$ belongs, which then impacts the distribution of any other group of dimensions. However, this statistical dependency vanishes when $\boldsymbol{X}$ is known to be generated from a single Bernoulli model.

Based on the above argument, we have used the total correlation measure (defined in Definition 3.4), in order to quantify whether a selected row-subset of samples in $\boldsymbol{\mathcal{X}}$ are more likely to be drawn from a single Bernoulli model, or a mixture of various Bernoulli models with different parameters. When n is finite, our algorithm only considers a relatively small $m \times d$ sub-matrix of $\boldsymbol{\mathcal{X}}$. This is due to the fact that the number of data samples m which are required to make a reliable assessment of the ϵ-pureness of a sub-matrix grows exponentially with respect to d (see Lemmas 4.1 and 4.2 in the next section). This fact should not be surprising since reliable computation of total correlation is subject to having a relatively close estimation of a d-dimensional binary distribution. Hence large values of d are unsuitable for estimating $\mathcal{D}(\cdot)$. On the other hand, by choosing a small d, we are ignoring a huge amount of valuable information in the dataset. To exploit all the information embedded in $\boldsymbol{\mathcal{X}}$, the MTC is introduced. It computes numerous total correlations over various subsets of dimensions, and aggregates all these values to form a more informative measure.

4. Proof of Theorem 3.1

We first start by some lemmas which indicate the goodness of our proposed *purity check* measures. The following lemma shows that $\mathcal{D}(\boldsymbol{Q})$ deviates from zero with high probability whenever $\boldsymbol{Q}$ is generated by a BMM with $K \geq 2$ mixture components.

Lemma 4.1. *Assume $\mathcal{B}$ to be a BMM with $K \geq 2$ clusters, dimension d, frequency matrix $\boldsymbol{P} \in [0,1]^{K\times d}$ and cluster probability vector $\boldsymbol{w} = (w_1, \ldots, w_K)$. Let $\boldsymbol{P}$ be $(\mathcal{L}, \delta)$-separable for some $\mathcal{L} \leq d$ and $\delta > 0$. Also, assume there exists $\epsilon > 0$ such that $w_k \leq 1 - \epsilon$, $\forall k = 1, \ldots, K$. Consider $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ to be n i.i.d. samples drawn from $\mathcal{B}$, and let $\boldsymbol{Q} = [\boldsymbol{Q}_1 | \cdots | \boldsymbol{Q}_n]^T \in \{0,1\}^{n\times d}$. Then, if $\mathcal{L} > \frac{1+\log\frac{K}{\epsilon}}{(1-\epsilon)\delta^2}$, we have*

$$\mathbb{P}\big\{\mathcal{D}(\boldsymbol{Q}) \leq \tau\big\} \leq 2^{d+1} e^{-\beta n},$$

where $\tau := \frac{\epsilon}{2}(1 + \log\frac{K}{\epsilon})$ and $\beta := \frac{\tau^2}{d^4 2^{d+1}}$.

Proof of Lemma 4.1 is given in the Appendix. The assumption of $w_k \leq 1 - \epsilon$ for all k yields that when $n \to \infty$, the set of observations $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ would not be ϵ-pure, almost surely. Hence, for an asymptotically large non ϵ-pure set of observations, we have $\mathcal{D}(\boldsymbol{Q}) \overset{\text{a.s.}}{\geq} \tau$. On the other hand, we have already discussed that for a completely pure set, that is, when samples are drawn from a single Bernoulli model ($K = 1$), we have $\lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q}) \overset{\text{a.s.}}{=} 0$. Accordingly, the following lemma provides a concentration bound on $\mathcal{D}(\boldsymbol{Q})$ when rows of $\boldsymbol{Q}$ are drawn from a single Bernoulli model.

Lemma 4.2. *Let $\mathcal{B}$ be a single Bernoulli model ($K = 1$) with dimension d and an arbitrary frequency vector. Consider $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ to be n i.i.d. samples drawn from $\mathcal{B}$, and let*

$\boldsymbol{Q} = [\boldsymbol{Q}_1|\cdots|\boldsymbol{Q}_n]^T$. *Then, we have*

$$\mathbb{P}\big\{\mathcal{D}(\boldsymbol{Q}) \geq \tau\big\} \leq 2^{d+1} e^{-\beta d^2 n},$$

where τ *and* β *are the same as in Lemma* 4.1.

The proof for Lemma 4.2 is also given in the Appendix. And finally, the following lemma shows that the error probability in detecting an improper clustering of samples, that is, any clustering with at least one non ϵ-pure cluster, drops exponentially with respect to $n \times \mathcal{L}$.

Lemma 4.3. *Assume* $\mathcal{B}$ *to be a BMM with* K *clusters, dimension* L*, frequency matrix* $\boldsymbol{P}$ *and weight vector* $\boldsymbol{w} = \{w_1, \ldots, w_K\}$*. Consider* $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ *to be* n *i.i.d. samples drawn from* $\mathcal{B}$ *and let* $\boldsymbol{Y} = [\boldsymbol{Y}_1|\cdots|\boldsymbol{Y}_n]^T \in \{0,1\}^{n\times L}$*. For* $K \geq 2$*, assume* $\boldsymbol{P}$ *to be* $(\mathcal{L}, \delta)$*-separable for some* $\mathcal{L} \leq L$ *and* $\delta > 0$*. Also, assume there exists* $\epsilon > 0$ *such that* $w_k \leq 1 - \epsilon$*,* $\forall k$*. Assume* $\mathcal{L} > d := \frac{K(K-1)}{2(1-\epsilon)\delta^2}(1 + \log\frac{K}{\epsilon})$ *and let* $\tau := \frac{\epsilon}{2}(1 + \log\frac{K}{\epsilon})$*. Then,*

$$\mathbb{P}\big\{\mathcal{D}_{\max}(\boldsymbol{Y}; d) \leq \tau\big\} \leq 4^{\mathcal{L}} \exp\left(\frac{-\tau^2 n\mathcal{L}}{d^5 2^{d+1}}\right).$$

On the other hand, when $K = 1$ *we have*:

$$\mathbb{P}\big\{\mathcal{D}_{\max}(\boldsymbol{Y}; d) \geq \tau\big\} \leq \binom{L}{d} 2^{d+1} \exp\left(\frac{-\tau^2 n}{d^2 2^{d+1}}\right).$$

Based on Lemma 4.3, for sufficiently large n and $\mathcal{L}$, the probability of mis-detection between an ϵ-pure subset of samples in the dataset $\boldsymbol{\mathcal{X}}$ and a non ϵ-pure one is strictly bounded. In fact, Lemma 4.3 provides a mathematically rigor and reliable criterion to distinguish between a "good" and "bad" clustering of samples in a finite dataset.

The parameter α in Algorithm 1 is user-defined. For a BMM $\mathcal{B} = \mathcal{B}(K, \boldsymbol{P}, \boldsymbol{w})$, as long as we have $\min_k w_k \geq \alpha$, K cannot exceed $\lceil\frac{1}{\alpha}\rceil$. Given that the conditions in Theorem 3.1 are satisfied, with probability at least $1 - \zeta$ Algorithm 1 terminates before passing $\lceil\frac{1}{\alpha}\rceil$ clusters and as soon as it finds an ϵ-correct clustering of $\boldsymbol{\mathcal{X}}$. Otherwise, the algorithm just outputs a null clustering. In the following, we use the results from Lemma 4.3 to prove Theorem 3.1, which is also the mathematical analysis of Algorithm 1.

Proof of Theorem 3.1. Algorithm 1 checks all possible cluster numbers $\kappa \leq \lceil\frac{1}{\alpha}\rceil$, starting from $\kappa = 1$. Let us denote the number of clusterings that need to be checked before reaching the true latent clustering $\boldsymbol{Z}_T$ by N. Then, N obviously satisfies the following inequality:

$$N \leq 1^n + 2^n + \cdots + K^n \leq K^{n+1}.$$

In this regard, one can consider the following error events during the execution of Algorithm 1:

- $\mathcal{E}_1$: Accepting a non ϵ-correct clustering of dataset $\boldsymbol{\mathcal{X}}$, before reaching the true clustering $\boldsymbol{Z}_T$. Recall that a non ϵ-correct clustering denotes any clustering with at least one non ϵ-pure cluster.

- $\mathcal{E}_2$: Eventually reaching to the true clustering $\boldsymbol{Z}_T$, and denying it.
- $\mathcal{E}_3$: The smallest true cluster in the dataset $\boldsymbol{\mathcal{X}}$ has less than $\alpha n/2$ members.

Obviously, probability of the algorithm failure, denoted by P_E, can be upper-bounded as

$$P_E = \mathbb{P}\{\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3\} \leq \mathbb{P}\{\mathcal{E}_1\} + \mathbb{P}\{\mathcal{E}_2\} + \mathbb{P}\{\mathcal{E}_3\}.$$

In the following, we show that given the conditions of Theorem 3.1, we have $\mathbb{P}\{\mathcal{E}_i\} \leq \zeta/3$ for $i = 1, 2, 3$.

From Lemma 4.3, we know that the probabilities of accepting a non ϵ-pure clustering, and rejecting the correct one are both bounded and decrease exponentially w.r.t. n. In the following, we compute the corresponding error exponents for the particular parameter setting of Algorithm 1. Recall sub-dimension d and threshold τ as

$$d := \frac{1-\alpha}{2(\alpha\delta)^2(1-\epsilon)}\left(1 + \log\frac{1}{\alpha\epsilon}\right) = O\left(\log\frac{1}{\epsilon}\right), \tag{4.1}$$

and

$$\tau := \frac{\epsilon}{2}\left(1 + \log\frac{1}{\alpha\epsilon}\right) = O\left(\epsilon\log\frac{1}{\epsilon}\right). \tag{4.2}$$

As it becomes evident in the proceeding parts of the proof, we also need to compute the order of e^d w.r.t. ϵ. In this regard, one can write:

$$e^d = (e/\alpha)^{\frac{1-\alpha}{2(\alpha\delta)^2(1-\epsilon)}} \cdot (1/\epsilon)^{\frac{1-\alpha}{2\alpha^2\delta^2(1-\epsilon)}}.$$

The first term is $O(1)$ with respect to ϵ, when $\epsilon \to 0$. Therefore, we have

$$e^d = O\big((1/\epsilon)^{\frac{1-\alpha}{2(\alpha\delta)^2}}\big). \tag{4.3}$$

By using the union bound over all non ϵ-pure clusterings in the first N steps of the algorithm, we have the following inequality:

$$\mathbb{P}\{\mathcal{E}_1\} \leq N\mathbb{P}\big\{\mathcal{E}_1^{(1)}\big\}, \tag{4.4}$$

where $\mathcal{E}_1^{(1)}$ represents the error event corresponding to the acceptance of a single non ϵ-pure clustering. In (4.4), the factor N is an upper bound on the number of non ϵ-correct clusterings that need to be checked by Algorithm 1 before reaching the true clustering $\boldsymbol{Z}_T$. Moreover, it should be noted that for a non ϵ-correct clustering $\boldsymbol{Z} \in \{1, \ldots, \kappa\}^n$, at least one of the clusters is not ϵ-pure, and thus we need our MTC measure to detect at least one of such clusters.

Remember that all clusters are assumed to have at least $\alpha n/2$ samples, and also $N \leq K^{n+1}$. This way, by using Lemma 4.3 the following bound on the probability of error event $\mathcal{E}_1$ can be

attained:

$$\begin{aligned}\mathbb{P}\{\mathcal{E}_1\} &\le K^{n+1}\cdot 4^{\mathcal{L}}\cdot \exp\left(\frac{-\alpha\tau^2 n\mathcal{L}}{d^5 2^{d+2}}\right)\\ &= \exp\left(O(n)+O(\mathcal{L})-n\mathcal{L}\cdot\frac{\alpha\tau^2}{d^5 2^{d+2}}\right)\\ &\le \exp\left(O(n)+O(\mathcal{L})-n\mathcal{L}\cdot\frac{\alpha\tau^2}{d^5 e^{d+2}}\right)\\ &= \exp\left(O(n)+O(\mathcal{L})-n\mathcal{L}\cdot O\left(\frac{\epsilon^{2+\frac{1-\alpha}{2(\alpha\delta)^2}}}{\log^3(1/\epsilon)}\right)\right),\end{aligned} \tag{4.5}$$

where we have used the results of equations (4.1), (4.2) and (4.3). The dominant exponent in the r.h.s. of (4.5) corresponds to the $n\times\mathcal{L}$ term. In other words, by choosing sufficiently large n and $\mathcal{L}$, one can make $\mathbb{P}\{\mathcal{E}_1\}$ arbitrarily small, even though the union bound is over K^{n+1} events. Mathematically speaking, it can be confirmed that by choosing

$$\mathcal{L}\ge\frac{B\log^3(1/\epsilon)}{\epsilon^{2+\frac{1-\alpha}{2(\alpha\delta)^2}}},\qquad n\ge n^{(1)}_{\min}:=C^{(1)}\left(\frac{\log^3(1/\epsilon)}{\epsilon^{2+\frac{1-\alpha}{2(\alpha\delta)^2}}}+\log\frac{1}{\zeta}\right), \tag{4.6}$$

we can achieve $\mathbb{P}\{\mathcal{E}_1\}\le\zeta/3$ for arbitrary small $\epsilon,\zeta>0$, where coefficients B and $C^{(1)}$ do not depend on ϵ or ζ.

A similar argument can be used to obtain an upper-bound on $\mathbb{P}\{\mathcal{E}_2\}$. It should be noted that for $\mathcal{E}_2$ to occur, at least one of the true clusters in $\boldsymbol{Z}_T$ must have $\mathcal{D}_{\max}>\tau$. Since the number of clusters at that step of the algorithm is K, one can use the union bound over all K clusters each of which has at least $\alpha n/2$ members. Also, we aim to use the following inequality:

$$\log\binom{L}{d}\le d\log\frac{Le}{d}.$$

In this regard, by using the second inequality in Lemma 4.3, it can be shown that

$$\begin{aligned}\mathbb{P}\{\mathcal{E}_2\} &\le K\binom{L}{d}2^{d+1}\exp\left(\frac{-\alpha\tau^2 n}{d^2 2^{d+2}}\right)\\ &= \exp\left(O\left(\log\frac{1}{\epsilon}\left[1+\log\frac{L}{\log(1/\epsilon)}\right]\right)-nO\left(\epsilon^{2+\frac{1-\alpha}{2(\alpha\delta)^2}}\right)\right).\end{aligned} \tag{4.7}$$

Thus, by choosing

$$n\ge n^{(2)}_{\min}:=C^{(2)}\left(\frac{\log(1/\epsilon)\log L}{\epsilon^{2+\frac{1-\alpha}{2(\alpha\delta)^2}}}+\log\frac{1}{\zeta}\right), \tag{4.8}$$

we have $\mathbb{P}\{\mathcal{E}_2\} \leq \zeta/3$, where $C^{(2)}$ is a constant that does not depend on ϵ or ζ. Finally, according to Lemma A.3, by choosing $n \geq n^{(3)}_{\min}$ which is defined as

$$n^{(3)}_{\min} := C^{(3)} \log \frac{1}{\zeta},$$

one can guarantee that the probability of occurring $\mathcal{E}_3$ is less that $\zeta/3$, where again constant $C^{(3)}$ is independent of ϵ or ζ. Therefore, assuming that $\mathcal{L}$ satisfies the inequality in (4.6) and the sample size n satisfies

$$n \geq \frac{C \log^3(1/\epsilon)}{\epsilon^{2+\frac{1-\alpha}{2(\alpha\delta)^2}}} \log \frac{L}{\zeta} \geq \max\left\{n^{(1)}_{\min}, n^{(2)}_{\min}, n^{(3)}_{\min}\right\}$$

for some constant C, Algorithm 1 is guaranteed to output an ϵ-correct clustering on dataset $\boldsymbol{\mathcal{X}}$ with probability at least $1 - \zeta$. This completes the proof. □

Let us consider an asymptotic regime where dimension L is being increased while the number of informative dimensions $\mathcal{L}$ is kept fixed. Then, according to Theorem 3.1, the minimum required dataset size n should grow logarithmically w.r.t. L. This analytic observation makes sense since in finite n regimes, addition of more *non-informative* dimensions, that is, those dimensions that have the same frequency values between all or at least some of the clusters, only adds extra noise to the dataset $\boldsymbol{\mathcal{X}}$ and thus makes the clustering a more challenging task.

5. Conclusions

This paper aims to find the first sample complexity bounds on reliable clustering of Bernoulli Mixture Models, when the number of clusters is unknown. To this aim, we propose a novel variant of an existing measure in statistics, denoted by Maximal Total Correlation (MTC), and show it has interesting concentration properties. Based on this measure, we propose an algorithm that is capable of clustering the data with a maximum mis-clustering rate of ϵ with probability at least $1 - \zeta$ (for any $\epsilon, \zeta > 0$), as long as sample complexity n and the number of informative dimensions $\mathcal{L}$ grow polynomially w.r.t. ϵ and ζ. No restrictive assumptions have been made in our model, except those that are required for the meaningfulness of clustering, such as: existence of a non-zero difference among frequency vectors of different mixture components, and a minimum weight for each cluster in the model. As a result, our findings encapsulate many classes of BMM inference problems.

Our main result provides an estimator for parameters of the BMM, showing that the model is learnable under the right circumstances. However, the estimator in Algorithm 1 has exponential running time. This raises the natural question: Are BMMs efficiently learnable? In general, existence of an *efficient* (polynomial-time) algorithm for density estimation, clustering or parameter identification of many mixture models is unknown [2], including Gaussian mixture models. Therefore, obtaining an efficient method for clustering of BMMs is both theoretically and practically important. On the other hand, we are not aware of any sample complexity lower bounds

for clustering or learning of BMMs. Therefore, it is not clear whether the upper bounds in this paper are tight or not. Deriving lower bounds for Theorem 3.1 is also a good direction for future works in this area.

Appendix: Auxiliary lemmas and proofs

Proof of Lemma 4.1. Proof consists of two parts. First, we show $\mathcal{D}(\boldsymbol{Q})$ is almost surely greater than the positive threshold 2τ in the asymptotic case, that is,

$$\mathbb{P}\Big\{\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})\leq 2\tau\Big\}=0. \tag{A.1}$$

Second, we prove that the probability of $|\mathcal{D}(\boldsymbol{Q})-\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})|>\tau$ decays exponentially w.r.t. n, which complete the proof.

For the sake of simplicity in this proof, let $\mathbb{P}_{\boldsymbol{p}}$ represent the probability distribution of a Bernoulli model with frequency vector $\boldsymbol{p}\in[0,1]^d$. This way, one can write

$$\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})\overset{\text{a.s.}}{=}\mathcal{D}_{\mathrm{KL}}\Big(\sum_k w_k\mathbb{P}_{\boldsymbol{p}^{(k)}}\;\Big\|\;\mathbb{P}_{\bar{\boldsymbol{p}}}\Big), \tag{A.2}$$

where $\boldsymbol{p}^{(k)}$ denotes the kth row of frequency matrix $\boldsymbol{P}$, and $\bar{\boldsymbol{p}}:=\sum_k w_k\boldsymbol{p}^{(k)}$. Here, $\sum_k w_k\mathbb{P}_{\boldsymbol{p}^{(k)}}$ indicates a mixture of Bernoulli models (a BMM), while $\mathbb{P}_{\bar{\boldsymbol{p}}}$ denotes a single Bernoulli model whose frequency vector is the weighted average of the K frequency vectors in $\boldsymbol{P}$. Again, we have used the fact that KL-divergence is continuous when its inputs are discrete probability distributions.

It can be verified that when only one component of $\boldsymbol{w}$ is 1 and the rest are 0, the two probability distributions $\sum_k w_k\mathbb{P}_{\boldsymbol{p}^{(k)}}$ and $\mathbb{P}_{\bar{\boldsymbol{p}}}$ are equal and $\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})\overset{\text{a.s.}}{=}0$. However, if for some $\epsilon>0$ we have $w_k\leq 1-\epsilon$, $\forall k$, and $\mathcal{L}$ is *sufficiently* large, then we prove that $\sum_k w_k\mathbb{P}_{\boldsymbol{p}^{(k)}}$ cannot be consistently approximated by a single Bernoulli model with frequency vector $\sum_k w_k\boldsymbol{p}^{(k)}$.

Based on the definition of the Kullback–Liebler divergence, r.h.s. of (A.2) can be expanded as follows which helps us to find a proper lower-bound for $\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})$:

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}\Big(\sum_k w_k\mathbb{P}_{\boldsymbol{p}^{(k)}}\;\Big\|\;\mathbb{P}_{\bar{\boldsymbol{p}}}\Big)&=\sum_{\boldsymbol{X}\in\{0,1\}^d}\sum_k w_k\mathbb{P}_{\boldsymbol{p}^{(k)}}(\boldsymbol{X})\log\Big(\frac{\sum_u w_u\mathbb{P}_{\boldsymbol{p}^{(u)}}(\boldsymbol{X})}{\mathbb{P}_{\bar{\boldsymbol{p}}}(\boldsymbol{X})}\Big)\\
&\geq\sum_{\boldsymbol{X}\in\{0,1\}^d}\sum_k w_k\mathbb{P}_{\boldsymbol{p}^{(k)}}(\boldsymbol{X})\log\Big(\frac{w_k\mathbb{P}_{\boldsymbol{p}^{(k)}}(\boldsymbol{X})}{\mathbb{P}_{\bar{\boldsymbol{p}}(\boldsymbol{X})}}\Big)\\
&=\sum_k w_k\sum_{\ell=1}^{d}\Big(\sum_{X_\ell\in\{0,1\}}\mathbb{P}_{p_\ell^{(k)}}(X_\ell)\log\Big(\frac{\mathbb{P}_{p_\ell^{(k)}}(X_\ell)}{\mathbb{P}_{\bar{p}_\ell}(X_\ell)}\Big)\Big)-\mathbb{H}(\boldsymbol{w})\\
&=\sum_{\ell=1}^{d}\sum_{k=1}^{K}w_k\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{p_\ell^{(k)}}\,\|\,\mathbb{P}_{\bar{p}_\ell})-\mathbb{H}(\boldsymbol{w}),
\end{aligned} \tag{A.3}$$

where $\mathbb{H}(\boldsymbol{w}) := -\sum_k w_k \log w_k$ denotes the Shannon entropy of the discrete distribution $\boldsymbol{w}$. Moreover, it is easy to show that

$$\sum_{k=1}^{K} w_k \mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{p_\ell^{(k)}} \parallel \mathbb{P}_{\bar{p}_\ell}) = H\left(\sum_{k=1}^{K} w_k p_\ell^{(k)}\right) - \sum_{k=1}^{K} w_k H\left(p_\ell^{(k)}\right), \tag{A.4}$$

where, for the simplicity of notation, $H(p)$ for $0 \le p \le 1$ refers to $H(p) := \mathbb{H}(\mathrm{Bern}(p)) = -p\log p - (1-p)\log(1-p)$. Since $H(\cdot)$ is a strictly concave function, and considering the fact that $\mathbb{H}(\boldsymbol{w})$ is always upper-bounded by $\log K$ regardless of $\mathcal{L}$, one can conclude that the lower-bound for $\lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q})$ in (A.3) becomes strictly positive when i) $\mathcal{L}$ is sufficiently large, and ii) frequency vectors $\boldsymbol{p}^{(k)}$ for $k = 1, \ldots, K$ are sufficiently far apart from each other.

In order to simplify the lower-bound in (A.3), let us assume a random variable $\boldsymbol{A} \in [0, 1]$, and define $\boldsymbol{a} := \boldsymbol{A} - \mathbb{E}\boldsymbol{A}$. According to Taylor's theorem [12], one can write

$$\begin{aligned} H(\mathbb{E}\boldsymbol{A}) - \mathbb{E}H(\boldsymbol{A}) &= H(\mathbb{E}\boldsymbol{A}) - \mathbb{E}\left\{H(\mathbb{E}\boldsymbol{A}) + H'(\mathbb{E}\boldsymbol{A})\boldsymbol{a} + \frac{1}{2}H''(\mathbb{E}\boldsymbol{A} + \xi)\boldsymbol{a}^2\right\} \\ &\ge \frac{\mathbb{E}\boldsymbol{a}^2}{2} \inf_{0\le p\le 1} \left|H''(p)\right| = \inf_{0\le p\le 1} \frac{\mathbb{E}\boldsymbol{a}^2}{2p(1-p)} = 2\,\mathrm{var}(\boldsymbol{A}), \end{aligned} \tag{A.5}$$

where ξ is a random variable depending on $\boldsymbol{A}$, and $\mathrm{var}(\boldsymbol{A}) = \mathbb{E}\boldsymbol{a}^2$ denotes the variance of $\boldsymbol{A}$. Now, for $\ell = 1, \ldots, d$, let us define $\boldsymbol{A}_\ell$ as a random variable that takes the values $p_\ell^{(1)}, \ldots, p_\ell^{(K)}$ with probabilities $w_1, \ldots, w_K$, respectively. Using the inequality in (A.5), the lower-bound for $\lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q})$ can be written as

$$\begin{aligned} \lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q}) &\overset{\text{a.s.}}{\ge} \sum_{\ell=1}^{d} \big[H(\mathbb{E}\boldsymbol{A}_\ell) - \mathbb{E}H(\boldsymbol{A}_\ell)\big] - \mathbb{H}(\boldsymbol{w}) \ge 2\sum_{\ell=1}^{d} \mathrm{var}(\boldsymbol{A}_\ell) - \mathbb{H}(\boldsymbol{w}) \\ &= 2\sum_{\ell=1}^{d}\sum_{k=1}^{K} w_k \left(p_\ell^{(k)} - \sum_{u=1}^{K} w_u p_\ell^{(u)}\right)^2 - \mathbb{H}(\boldsymbol{w}). \end{aligned} \tag{A.6}$$

We have already assumed that $w_k \le 1 - \epsilon, \forall k$. Also, due to the $(\mathcal{L}, \delta)$-separability assumption, for all pairs of rows in $\boldsymbol{P}$, say i and j, there exists a subset of columns $\mathscr{C}_{i,j} \subseteq \{1, 2, \ldots, d\}$ where

$$\left|p_\ell^{(i)} - p_\ell^{(j)}\right| \ge \delta, \quad \ell \in \mathscr{C}_{i,j},$$

and $|\mathscr{C}_{i,j}| \ge \mathcal{L}$. This suggests that the values of $\mathrm{var}(\boldsymbol{A}_\ell)$, at least for $\ell \in \cup_{i,j}\mathcal{C}_{i,j}$, are greater than or equal to a positive function of ϵ and δ. On the other hand, the only negative term $-\mathbb{H}(\boldsymbol{w})$ is bounded and does not scale with $\mathcal{L}$. This suggests that for a large enough $\mathcal{L}$, the r.h.s. of (A.6) becomes strictly positive.

Lemma A.1 proves that the lower-bound in (A.6) can be further simplified as

$$\lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q}) \overset{\text{a.s.}}{\ge} 2\mathcal{L}\epsilon(1-\epsilon)\delta^2 - \mathbb{H}\big(\boldsymbol{w}^*\big) \ge 2\epsilon(1-\epsilon)\delta^2\left(\mathcal{L} - \frac{1 + \log\frac{K}{\epsilon}}{2(1-\epsilon)\delta^2}\right),$$

where $\boldsymbol{w}^*$ denotes a weight vector, or equivalently a discrete probability distribution supported on $\{1,\dots,K\}$, with $w_1^* = 1-\epsilon$ and $w_i^* = \epsilon/(K-1)$ for $i = 2,\dots,K$. Note that the second inequality directly results from

$$\mathbb{H}\big(\boldsymbol{w}^*\big) = (1-\epsilon)\log\frac{1}{1-\epsilon} + \epsilon\log\frac{K-1}{\epsilon} \le \epsilon\bigg(1+\log\frac{K}{\epsilon}\bigg).$$

Also, we have already assumed that $\mathcal{L} \ge \frac{1+\log\frac{K}{\epsilon}}{(1-\epsilon)\delta^2}$, which means the following relations hold:

$$\begin{aligned}\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q}) &\overset{\text{a.s.}}{\ge} 2\epsilon(1-\epsilon)\delta^2\bigg(\mathcal{L} - \frac{1+\log\frac{K}{\epsilon}}{2(1-\epsilon)\delta^2}\bigg)\\ &\ge 2\epsilon(1-\epsilon)\delta^2\bigg(\frac{1+\log\frac{K}{\epsilon}}{(1-\epsilon)\delta^2} - \frac{1+\log\frac{K}{\epsilon}}{2(1-\epsilon)\delta^2}\bigg)\\ &\ge \epsilon\bigg(1+\log\frac{K}{\epsilon}\bigg) = 2\tau,\end{aligned}$$

where the last equality is due to the definition of $\tau := \frac{\epsilon}{2}(1+\log\frac{K}{\epsilon})$ in the statement of Lemma 4.1. This way, the first part of the proof is complete.

So far, we have shown that $\mathcal{D}(\boldsymbol{Q})$ almost surely becomes greater than 2τ when n goes to infinity. However, $\mathcal{D}(\boldsymbol{Q})$ is supposed to be computed over a finite sample size of n, thus it is necessary to show that $|\mathcal{D}(\boldsymbol{Q}) - \lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})|$ concentrates around zero w.r.t. n. In fact, Lemma A.2 proves that the probability of the above error term exceeding τ decays exponentially with respect to n. Based on the result of Lemma A.2, the probability $\mathbb{P}\{\mathcal{D}(\boldsymbol{Q}) \le \tau\}$ can be upper-bounded as

$$\begin{aligned}\mathbb{P}\big\{\mathcal{D}(\boldsymbol{Q}) \le \tau\big\} &\le 2^{d+1}\exp\bigg(\frac{-n\epsilon^2(1-\epsilon)^2\delta^4}{d^4 2^{d+1}}\bigg(\mathcal{L} - \frac{1+\log\frac{K}{\epsilon}}{2(1-\epsilon)\delta^2}\bigg)^2\bigg)\\ &\le 2^{d+1}\exp\bigg(\frac{-n\tau^2}{d^4 2^{d+1}}\bigg),\end{aligned}$$

which completes the proof. □

Lemma A.1. *The lower-bound for* $\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})$ *in* (A.6)*, subject to* $(\mathcal{L},\delta)$*-separability of the frequency matrix* $\boldsymbol{P}$ *and assuming* $w_k \le 1-\epsilon, \forall k$*, is as follows*:

$$\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q}) \overset{\text{a.s.}}{\ge} 2\epsilon(1-\epsilon)\delta^2\bigg(\mathcal{L} - \frac{1+\log\frac{K}{\epsilon}}{2(1-\epsilon)\delta^2}\bigg).$$

Proof. Considering the assumption made in Lemma 4.1 with respect to the non ϵ-purity of $\boldsymbol{Q}$, let us define $\boldsymbol{W}(r)$ for $\epsilon \le r \le 1-1/K$ as

$$\boldsymbol{W}(r) := \bigg\{\boldsymbol{w}\in\mathbb{R}^K \,\Big|\, w_k \ge 0, \sum_k w_k = 1, \max_k w_k = 1-r\bigg\}.$$

Hence, according to (A.6) the lower-bound for $\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})$ (for a non ϵ-pure $\boldsymbol{Q}$) can be written as

$$\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})\overset{\text{a.s.}}{\geq}\inf_{r\in[\epsilon,1-1/K]}\left\{\inf_{\boldsymbol{w}\in\boldsymbol{W}(r)}2\sum_{\ell=1}^{d}\sum_{k=1}^{K}w_k\big(p_\ell^{(k)}-\bar{p}_\ell(\boldsymbol{w})\big)^2-\mathbb{H}(\boldsymbol{w})\right\},\tag{A.7}$$

with $\bar{p}_\ell(\boldsymbol{w}):=\sum_{k=1}^{K}w_kp_\ell^{(k)}$. In fact, (A.7) indicates minimization of the lower-bound over all asymptotically large non ϵ-pure matrices $\boldsymbol{Q}$. In order to further simplify the above lower-bound, minimization over $\boldsymbol{w}\in\boldsymbol{W}(r)$ can be carried out for the two terms in the r.h.s. of (A.7), in a separate manner. Mathematically speaking,

$$\begin{aligned}&\inf_{\boldsymbol{w}\in\boldsymbol{W}(r)}\left\{2\sum_{\ell=1}^{d}\sum_{k=1}^{K}w_k\big(p_\ell^{(k)}-\bar{p}_\ell(\boldsymbol{w})\big)^2-\mathbb{H}(\boldsymbol{w})\right\}\\&\quad\geq\inf_{\boldsymbol{w}\in\boldsymbol{W}(r)}2\sum_{\ell=1}^{d}\sum_{k=1}^{K}w_k\big(p_\ell^{(k)}-\bar{p}_\ell(\boldsymbol{w})\big)^2-\sup_{\boldsymbol{w}\in\boldsymbol{W}(r)}\mathbb{H}(\boldsymbol{w}).\end{aligned}$$

It should be reminded that for each weight vector (or equivalently, probability distributions) $\boldsymbol{w}\in\boldsymbol{W}(r)$, one of the components is exactly equal to $1-r$, and thus the rest of the components must sum to r. Therefore, the maximum Shannon entropy $\mathbb{H}(\boldsymbol{w})$ occurs when the latter $K-1$ components have an equal probability, that is, $r/(K-1)$, which indicates maximum possible randomness. In this regard, it is easy to see that

$$\sup_{\boldsymbol{w}\in\boldsymbol{W}(r)}\mathbb{H}(\boldsymbol{w})=(1-r)\log\frac{1}{1-r}+\sum_{k=2}^{K}\frac{r}{K-1}\log\frac{K-1}{r}\leq r\left(1+\log\frac{K}{r}\right),\tag{A.8}$$

which is also based on the fact that $(1-r)\log\frac{1}{1-r}\leq r$. On the other hand, for the first term in r.h.s of (A.7), the following lower-bound can be obtained:

$$\begin{aligned}&\inf_{\boldsymbol{w}\in\boldsymbol{W}(r)}2\sum_{\ell=1}^{d}\sum_{k=1}^{K}w_k\big(p_\ell^{(k)}-\bar{p}_\ell(\boldsymbol{w})\big)^2\\&\quad\geq\inf_{\boldsymbol{w}\in\boldsymbol{W}(r)}2\sum_{\ell=1}^{d}\inf_{p\in\mathbb{R}}\left\{\sum_{k=1}^{K}w_k\big(p_\ell^{(k)}-p\big)^2\right\}\\&\quad\geq2\sum_{\ell=1}^{d}\inf_{p\in\mathbb{R}}\min_{t=1,\ldots,K}\left\{(1-r)\big(p_\ell^{(t)}-p\big)^2+\min_{k|k\neq t}r\big(p_\ell^{(k)}-p\big)^2\right\}\\&\quad=2\sum_{\ell=1}^{d}\min_{k,t=1,\ldots,K|k\neq t}r(1-r)\big(p_\ell^{(k)}-p_\ell^{(t)}\big)^2.\end{aligned}$$

The last equality can be achieved by solving for $\inf_{p\in\mathbb{R}}$, analytically. Since for every pair (k,t), $k\neq t$ and at least $\mathcal{L}$ dimensions out of $\ell=1,2,\ldots,d$, the inequality $|p_\ell^{(k)}-p_\ell^{(t)}|\geq\delta$ holds, one can write

$$\inf_{\boldsymbol{w}\in\boldsymbol{W}(r)} 2\sum_{\ell=1}^{d}\sum_{k=1}^{K} w_k\big(p_\ell^{(k)}-\bar{p}_\ell(\boldsymbol{w})\big)^2 \geq 2r(1-r)\mathcal{L}\delta^2. \tag{A.9}$$

By combining the inequalities in (A.8) and (A.9), the following lower-bound can be achieved for $\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})$:

$$\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q}) \overset{\text{a.s.}}{\geq} \inf_{r\in[\epsilon,1-1/K]} 2r(1-r)\delta^2\bigg(\mathcal{L}-\frac{1+\log\frac{K}{r}}{2(1-r)\delta^2}\bigg). \tag{A.10}$$

The objective function in the r.h.s. of (A.10) is a monotonically increasing function w.r.t. r, when $\mathcal{L}>\frac{1+\log\frac{K}{\epsilon}}{2(1-\epsilon)\delta^2}$. This can be easily verified by taking derivatives w.r.t. r for a sufficiently small ϵ. Hence, the minimizer of $\inf_{r\in[\epsilon,1-1/K]}$ occurs when $r=\epsilon$, which completes the proof. □

Lemma A.2. *For $n,d\in\mathbb{N}$, assume the rows of $\boldsymbol{Q}\in\{0,1\}^{n\times d}$ to be n i.i.d. samples drawn from a BMM with an arbitrary parameter set. Then, the probability of observing a deviation error of $\varepsilon>0$ between $\mathcal{D}(\boldsymbol{Q})$ and the asymptotic measure $\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})$ can be upper-bounded as*

$$\mathcal{P}\bigg\{\Big|\mathcal{D}(\boldsymbol{Q})-\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})\Big|>\varepsilon\bigg\}\leq 2^{d+1}\exp\bigg(\frac{-n\varepsilon^2}{d^4 2^{d+1}}\bigg).$$

Proof. Let us define $\hat{\mathbb{P}}_{\boldsymbol{Q}}$ as the empirical measure underlying the rows of $\boldsymbol{Q}$ (same is $\hat{\mathbb{P}}_{1,\boldsymbol{Q}}$ in Definition 3.4). According to Definition 3.4, $\mathcal{D}(\boldsymbol{Q})$ only depends on $\hat{\mathbb{P}}_{\boldsymbol{Q}}$, and thus permutation of the rows of $\boldsymbol{Q}$ does not affect its value. In this regard, and for the sake of simplicity, let us define $g:\mathbb{R}^{2^d}\to\mathbb{R}$ such that $g(\hat{\mathbb{P}}_{\boldsymbol{Q}}):=\mathcal{D}(\boldsymbol{Q})$, that is, a function that maps the empirical distribution $\hat{\mathbb{P}}_{\boldsymbol{Q}}$ to $\mathcal{D}(\boldsymbol{Q})$. According to Definition 3.4, it can be readily verified that

$$\begin{aligned}
\mathcal{D}(\boldsymbol{Q}) &= \sum_{\boldsymbol{X}\in\{0,1\}^d}\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})\log\bigg(\frac{\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})}{\prod_{\ell=1}^{d}\hat{p}_\ell^{X_\ell}(1-\hat{p}_\ell)^{1-X_\ell}}\bigg)\\
&= \sum_{\boldsymbol{X}\in\{0,1\}^d}\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})\bigg[\log\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})-\sum_{\ell=1}^{d}\mathbf{1}_{X_\ell}\log\hat{p}_\ell+\mathbf{1}_{1-X_\ell}\log(1-\hat{p}_\ell)\bigg]\\
&= \sum_{\boldsymbol{X}\in\{0,1\}^d}\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})\bigg[\log\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})-\sum_{\ell=1}^{d}\log\bigg(\sum_{\boldsymbol{X}'\in\{0,1\}^d|X'_\ell=X_\ell}\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X}')\bigg)\bigg],
\end{aligned} \tag{A.11}$$

where $\mathbf{1}_X$ denotes the indicator function which returns 1 if $X = 1$ and zero otherwise. During the derivation of (A.11), we have used the following two facts for $\ell = 1, \ldots, d$:

$$\hat{p}_\ell = \sum_{\boldsymbol{X}' \in \{0,1\}^d | X'_\ell = 1} \hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X}'), \qquad 1 - \hat{p}_\ell = \sum_{\boldsymbol{X}' \in \{0,1\}^d | X'_\ell = 0} \hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X}').$$

Since g is a continuous function, the *law of large numbers* implies that

$$\lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q}) = g\Big(\lim_{n\to\infty} \hat{\mathbb{P}}_{\boldsymbol{Q}}\Big) \overset{\text{a.s.}}{=} g(\mathbb{P}_{\mathcal{B}}),$$

where $\mathbb{P}_{\mathcal{B}}$ represents the true distribution of the BMM $\mathcal{B}$ that underlies the rows of $\boldsymbol{Q}$. Obviously, unlike the empirical measure $\hat{\mathbb{P}}_{\boldsymbol{Q}}$, $\mathbb{P}_{\mathcal{B}}$ is a deterministic distribution which can be quantified based on the parameters of $\mathcal{B}$. In this regard, differential calculus implies the following relation:

$$\Big|\mathcal{D}(\boldsymbol{Q}) - \lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q})\Big| \overset{\text{a.s.}}{=} \big|g(\hat{\mathbb{P}}_{\boldsymbol{Q}}) - g(\mathbb{P}_{\mathcal{B}})\big| = \left|\int_{\mathscr{P}} \langle \nabla g | \mathrm{d}\mathscr{P} \rangle\right|,$$

where ∇ denotes the gradient operator, $\langle \cdot | \cdot \rangle$ denotes the inner product, and $\mathscr{P}$ is an arbitrary continuous path in $\mathbb{R}^{2^d}$ that starts from $\hat{\mathbb{P}}_{\boldsymbol{Q}}$ and ends in $\mathbb{P}_{\mathcal{B}}$.[2] Let us consider the following particular path $\mathscr{P}$: The union of 2^d sub-paths, where each sub-path is aligned to a distinct axis of $\mathbb{R}^{2^d}$. Thus, we move from $\hat{\mathbb{P}}_{\boldsymbol{Q}}$ to $\mathbb{P}_{\mathcal{B}}$ in 2^d steps, where at each step we only change one of the components and keep the rest fixed. Let us denote the above-mentioned 2^d sub-paths with $\mathscr{P}_{\boldsymbol{X}}$, $\boldsymbol{X} \in \{0,1\}^d$. In this regard, while moving along the axis that corresponds to a particular $\boldsymbol{X} \in \{0,1\}^d$, the term $\langle \nabla g | \mathrm{d}\mathscr{P} \rangle$ simply becomes $\nabla_{\boldsymbol{X}} g \, \mathrm{d}s$, where s denotes the length parameter of the sub-path associated to component $\boldsymbol{X}$ and $\nabla_{\boldsymbol{X}} g : \mathbb{R}^{2^d} \to \mathbb{R}$ denotes the component of the 2^d-dimensional gradient ∇g which corresponds to $\boldsymbol{X}$.

With the above specifications for $\mathscr{P}$, and using the *Mean Value Theorem (MVT)* [12], one can write:

$$\begin{aligned}\Big|\mathcal{D}(\boldsymbol{Q}) - \lim_{n\to\infty} \mathcal{D}(\boldsymbol{Q})\Big| &\overset{\text{a.s.}}{\leq} \sum_{\boldsymbol{X}\in\{0,1\}^d} \left|\int_{\mathscr{P}_{\boldsymbol{X}}} \nabla_{\boldsymbol{X}} g \, \mathrm{d}s\right| \\ &\overset{\text{(MVT)}}{\leq} \sum_{\boldsymbol{X}\in\{0,1\}^d} \Big(\sup_{\boldsymbol{v}\in\mathscr{P}} \big|\nabla_{\boldsymbol{X}} g(\boldsymbol{v})\big|\Big) \big|\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X}) - \mathbb{P}_{\mathcal{B}}(\boldsymbol{X})\big|.\end{aligned}$$

According to (A.11), it is easy to show that partial derivatives of g can be exactly computed at the true distribution $\mathbb{P}_{\mathcal{B}}$ through the following formula:

$$\nabla_{\boldsymbol{X}} g = \log\left(\frac{\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}{\prod_{\ell=1}^{d} \big(\sum_{\boldsymbol{X}' \in \{0,1\}^d | X'_\ell = X_\ell} \mathbb{P}_{\mathcal{B}}(\boldsymbol{X}')\big)}\right) - (d-1), \quad \forall \boldsymbol{X} \in \{0,1\}^d.$$

[2] In the proceeding relations, we also show that g is differentiable. Therefore, $\nabla g : \mathbb{R}^{2^d} \to \mathbb{R}^{2^d}$ exists.

Considering the fact that $\sum_{\boldsymbol{X}'\in\{0,1\}^d|X'_\ell=X_\ell}\mathbb{P}_{\mathcal{B}}(\boldsymbol{X}')\geq\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})$, the following upper-bound holds for the partial derivatives of g for all $\boldsymbol{X}\in\{0,1\}^d$ and sufficiently large n:

$$\sup_{\boldsymbol{\nu}\in\mathscr{P}}\left|\nabla_{\boldsymbol{X}}g(\boldsymbol{\nu})\right|\leq\left|\log\left(\frac{\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}{\prod_{\ell=1}^{d}\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}\right)\right|+(d-1)\leq d\left(\log\frac{1}{\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}+1\right).$$

So far, we have managed to upper-bound the estimation error in the current lemma by the following inequality:

$$\begin{aligned}\left|\mathcal{D}(\boldsymbol{Q})-\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})\right|&\overset{\text{a.s.}}{\leq}d\sum_{\boldsymbol{X}\in\{0,1\}^d}\left(\log\frac{1}{\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}+1\right)\left|\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})-\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})\right|\\&=d\sum_{\boldsymbol{X}\in\{0,1\}^d}\left(\log\frac{1}{\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}+1\right)\sigma_{\boldsymbol{X}}\left|\frac{\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})-\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}{\sigma_{\boldsymbol{X}}}\right|\\&\leq d\left(\max_{\boldsymbol{X}}\left|\frac{\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})-\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}{\sigma_{\boldsymbol{X}}}\right|\right)\\&\quad\times\sum_{\boldsymbol{X}\in\{0,1\}^d}\left(\log\frac{1}{\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}+1\right)\sigma_{\boldsymbol{X}},\end{aligned}\tag{A.12}$$

where $\sigma_{\boldsymbol{X}}:=\sqrt{\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})(1-\mathbb{P}_{\mathcal{B}}(\boldsymbol{X}))}$.

An important issue that should be noted is that for those cases where $\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})=0$ or 1, we have $\sigma_{\boldsymbol{X}}=0$. However, in such cases, the empirical probabilities always coincide with the true ones, and the corresponding error terms in the above summation become exactly zero. As a result, such cases are implicitly omitted from all the summations in (A.12).

It is easy to show that the summation over $\boldsymbol{X}\in\{0,1\}^d$ in the r.h.s. of (A.12) reaches its maximum when $\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})=2^{-d}$ for all $\boldsymbol{X}$, which means

$$\sum_{\boldsymbol{X}\in\{0,1\}^d}\left(\log\frac{1}{\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}+1\right)\sigma_{\boldsymbol{X}}\leq d2^{d/2}.$$

The only remaining part of the proof is to bound the difference between the true distribution $\mathbb{P}_{\mathcal{B}}$ and the empirical one $\hat{\mathbb{P}}_{\boldsymbol{Q}}$. Let us define the set of events $A_{\boldsymbol{X}}$, $\forall\boldsymbol{X}\in\{0,1\}^d$ as

$$A_{\boldsymbol{X}}:=\left|\frac{\hat{\mathbb{P}}_{\boldsymbol{Q}}(\boldsymbol{X})-\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}{\sigma_{\boldsymbol{X}}}\right|>\delta,\quad\text{where }\delta:=\frac{\varepsilon}{d^2 2^{d/2}}.$$

Based on the previous relations, it can be verified that if none of the events $A_{\boldsymbol{X}}$ occur, then we almost surely have $|\mathcal{D}(\boldsymbol{Q})-\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})|\leq\varepsilon$. Therefore, for $\varepsilon>0$, and using both the Union

Bound (UB) and Chernoff Bound (CB), one can show

$$\begin{aligned}\mathbb{P}\left\{\left|\mathcal{D}(\boldsymbol{Q}) - \lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})\right| > \varepsilon\right\} &\le \mathbb{P}\left\{\bigcup_{\boldsymbol{X}\in\{0,1\}^d} A_{\boldsymbol{X}}\right\} \overset{\text{UB}}{\le} \sum_{\boldsymbol{X}\in\{0,1\}^d}\mathbb{P}\{A_{\boldsymbol{X}}\}\\ &\overset{\text{CB}}{\le} \sum_{\boldsymbol{X}\in\{0,1\}^d} e^{-n\mathcal{D}_{\text{KL}}(\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})+\delta\sigma_{\boldsymbol{X}}\|\mathbb{P}_{\mathcal{B}}(\boldsymbol{X}))}\\ &\quad + \sum_{\boldsymbol{X}\in\{0,1\}^d} e^{-n\mathcal{D}_{\text{KL}}(\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})-\delta\sigma_{\boldsymbol{X}}\|\mathbb{P}_{\mathcal{B}}(\boldsymbol{X}))},\end{aligned}$$

where $\mathcal{D}_{\text{KL}}(\cdot\,\|\,\cdot)$ represents the Kullback–Leibler divergence, and by $\mathcal{D}_{\text{KL}}(x\,\|\,y)$ for $x, y \in [0, 1]$ we mean

$$x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}.$$

For $x \notin [0, 1]$, let us define $\mathcal{D}_{\text{KL}}(x\,\|\,y) := +\infty$.

KL divergence can be lower-bounded according to Chernoff's theorem [13]. In other words, we have

$$\begin{aligned}\mathbb{P}\left\{\left|\mathcal{D}(\boldsymbol{Q}) - \lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})\right| > \varepsilon\right\} &\le 2\cdot 2^d \cdot \max_{\boldsymbol{X}\in\{0,1\}^d}\max_{\theta\in\{-1,+1\}} e^{-n\mathcal{D}_{\text{KL}}(\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})+\theta\delta\sigma_{\boldsymbol{X}}\|\mathbb{P}_{\mathcal{B}}(\boldsymbol{X}))}\\ &\le 2^{d+1}\max_{\boldsymbol{X},\theta}\exp\left(\frac{-n\delta^2\theta^2\sigma_{\boldsymbol{X}}^2}{2\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})(1-\mathbb{P}_{\mathcal{B}}(\boldsymbol{X}))}\right).\end{aligned}$$

By substituting for δ and considering the definition of $\sigma_{\boldsymbol{X}}$, the probability of observing a deviation greater than ε in estimating $\lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})$ can be upper-bounded as

$$\mathcal{P}\left\{\left|\mathcal{D}(\boldsymbol{Q}) - \lim_{n\to\infty}\mathcal{D}(\boldsymbol{Q})\right| > \varepsilon\right\} \le 2^{d+1}\exp\left(\frac{-n\varepsilon^2}{d^4 2^{d+1}}\right),$$

which completes the proof. □

Proof of Lemma 4.2. The proof is highly similar to that of Lemma A.2. The main difference lies in the fact that when $K = 1$, that is, a single Bernoulli model, one can easily verify that for all $\boldsymbol{X} \in \{0, 1\}^d$, we have:

$$\nabla_{\boldsymbol{X}} g = \log\left(\frac{\mathbb{P}_{\mathcal{B}}(\boldsymbol{X})}{\prod_{\ell=1}^{d}(\sum_{\boldsymbol{X}'\in\{0,1\}^d|X'_\ell=X_\ell}\mathbb{P}_{\mathcal{B}}(\boldsymbol{X}'))}\right) - (d-1) = 1-d,$$

since in a single Bernoulli model, the probability distribution equals to the product of its marginals over each dimension. Therefore, we have $|\nabla_{\boldsymbol{X}} g| \le d$. Following the same steps as shown in the proof of Lemma A.2 gives us the claimed inequality and complete the proof. □

Proof of Lemma 4.3. Recall $\mathrm{Col}(\boldsymbol{Y}; d)$ as the set of all $\binom{L}{d}$ sub-matrices of $\boldsymbol{Y}$ with d columns. Then, the first inequality states that the probability of $\exists\, \boldsymbol{Q} \in \mathrm{Col}(\boldsymbol{Y}; d) \Rightarrow \mathcal{D}(\boldsymbol{Q}) < \tau$ is strictly bounded.

In the following, we show that by examining all $\binom{L}{d}$ sub-matrices in $\mathrm{Col}(\boldsymbol{Y}; d)$, one can find at least $h := \lfloor \mathcal{L}/d \rfloor$ disjoint column sub-matrices of $\boldsymbol{Y}$, denoted by $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_h$, such that the frequency sub-matrices that correspond to Q_is are guaranteed to be at least $(\lfloor \frac{2d}{K(K-1)} \rfloor, \delta)$-separable:

First, it should be noted that frequency matrix $\boldsymbol{P}$ is assumed to be $(\mathcal{L}, \delta)$-separable. Similar to the notation we used in the proof of Lemma 4.1, it can said that for all pairs of rows in $\boldsymbol{P}$, say i and j, there exists a subset of columns $\mathscr{C}_{i,j} \subseteq \{1, 2, \ldots, L\}$, where

$$\left| p_\ell^{(i)} - p_\ell^{(j)} \right| \geq \delta, \quad \ell \in \mathscr{C}_{i,j},$$

and $|\mathscr{C}_{i,j}| \geq \mathcal{L}$. For each $i = 1, \ldots, K-1$, let us take $\lfloor 2d/[K(K-1)] \rfloor$ arbitrarily chosen indices from each of the $K-i$ sets $\mathscr{C}_{i,j}$, $j > i$ and then put them in some new corresponding sets, denoted by $\mathscr{D}_{i,j}$, $j > i$. It should be noted that $\mathscr{D}_{i,j}$s may have non-empty overlaps. Let

$$\mathscr{D} := \bigcup_{j>i} \mathscr{D}_{i,j}.$$

Obviously, $\mathscr{D}$ cannot have more than d members, since its the union of $K(K-1)/2$ sets, each having $\lfloor 2d/[K(K-1)] \rfloor$ members. In many practical situations, *informative dimensions* are dispersed randomly and thus $\mathscr{D}_{i,j}$s can be chosen to have huge overlaps. However, we consider the worst case which assumes the overlaps are empty. Also, for the cases where $|\mathscr{D}| < d$, assume we add enough arbitrary indices to $\mathscr{D}$ until it has d members. In this regard, the indices in $\mathscr{D}$ correspond to a sub-matrix of frequency matrix $\boldsymbol{P}$ that is at least $(\lfloor 2d/[K(K-1)] \rfloor, \delta)$-separable.

On the other hand, we can repeat the above procedure for at least $h := \lfloor \mathcal{L}/d \rfloor$ times without choosing any dimension more than once. This results in at least h disjoint sub-matrices, called $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_h$, that possess the above-mentioned property. Since $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_h$ do not overlap with each other, they are statistically independent which then implies

$$\begin{aligned} \mathbb{P}\left\{ \mathcal{D}_{\max}(\boldsymbol{Y}, d) \leq \tau \right\} &\leq \mathbb{P}\left\{ \mathcal{D}(\boldsymbol{Q}_i) \leq \tau, \forall i \right\} \\ &= \prod_{i=1}^{h} \mathbb{P}\left\{ \mathcal{D}(\boldsymbol{Q}_i) \leq \tau \right\}. \end{aligned}$$

Using the upper-bound for each $\mathbb{P}\{\mathcal{D}(\boldsymbol{Q}_i) \leq \tau\}$ from Lemma 4.1 and approximating h with $\mathcal{L}/d$, one can simply prove the claimed inequality.

For the second inequality in the statement of Lemma 4.3, one can simply employ the union bound as follows:

$$\begin{aligned} \mathbb{P}\left\{ \mathcal{D}_{\max}(\boldsymbol{Y}; d) > \tau \right\} &\leq \mathbb{P}\left\{ \max_{\boldsymbol{Q} \in \mathrm{Col}(\boldsymbol{Y}; d)} \mathcal{D}(\boldsymbol{Q}) > \tau \right\} \\ &\leq \sum_{\boldsymbol{Q} \in \mathrm{Col}(\boldsymbol{Y}; d)} \mathbb{P}\left\{ \mathcal{D}(\boldsymbol{Q}) > \tau \right\}. \end{aligned} \tag{A.13}$$

Also, note that $\mathrm{Col}(\boldsymbol{Y}; d)$ includes $\binom{L}{d}$ members. Again, substitution of $\mathbb{P}\{\mathcal{D}(\boldsymbol{Q}) > \tau\}$ with the upper-bound derived in Lemma A.2 gives us the claimed inequality and completes the proof. □

Lemma A.3. *Consider $\mathcal{M} = \mathcal{M}(K, \boldsymbol{w})$ to be a multinomial distribution with K mutually exclusive outcomes and corresponding probability vector $\boldsymbol{w} = (w_1, \ldots, w_K)$. Assume there exists $\alpha > 0$ such that $\min_k w_k \geq \alpha$. Let $\boldsymbol{D} := \{X_1, \ldots, X_n\}$ to be n i.i.d. samples drawn from $\mathcal{M}$. For $\zeta > 0$, assume*

$$n \geq \frac{2}{\alpha^2} \log \frac{3K}{\zeta}.$$

Then, with probability at least $1 - \zeta/3$, the size of the smallest cluster in $\boldsymbol{D}$ is least $\alpha n/2$.

Proof. We denote the probability of the smallest cluster in $\boldsymbol{D}$ having less than $\alpha n/2$ members by P_E. Let $\mathcal{A}_1, \ldots, \mathcal{A}_K$ represent the following events: for $k = 1, \ldots, K$, $\mathcal{A}_k$ represents the event that the kth cluster in $\boldsymbol{D}$ (corresponding to probability component w_k) has less than $nw_k/2$ members. Then, the following holds according to union bound:

$$P_E \leq \sum_{k=1}^{K} \mathbb{P}\{\mathcal{A}_k\}. \tag{A.14}$$

For $k = 1, \ldots, K$, consider the binomial random variable Y_k with the following distribution:

$$\mathbb{P}(Y_k) := \begin{cases} w_k, & Y_k = 1, \\ 1 - w_k, & Y_k = 0, \end{cases} \tag{A.15}$$

with $\mathbb{E}Y_k = w_k$. Let $y_1, \ldots, y_n$ to be n i.i.d. samples of Y_k. Define $S_k := y_1 + \cdots + y_n$, while obviously we have $\mathbb{E}S_k = nw_k$. Using Hoeffding's inequality, one can easily verify the following chain of relations:

$$\begin{aligned} \mathbb{P}\{\mathcal{A}_k\} &= \mathbb{P}\{S_k < nw_k/2\} = \mathbb{P}\{S_k - \mathbb{E}S_k < -nw_k/2\} \\ &\leq \exp\left(\frac{-2n^2 w_k^2}{4\sum_{i=1}^{n} (\max Y_k - \min Y_k)^2}\right) \\ &= \exp\left(\frac{-nw_k^2}{2}\right). \end{aligned} \tag{A.16}$$

Recall that we have $w_k \geq \alpha$ for all $k = 1, \ldots, K$, thus one can write

$$P_E \leq \sum_{k=1}^{K} \exp\left(\frac{-nw_k^2}{2}\right) \leq K \exp\left(\frac{-n\alpha^2}{2}\right), \tag{A.17}$$

which given the condition on n in the lemma, results into $P_E \leq \zeta/3$ and completes the proof. □

References

[1] Allman, E.S., Matias, C. and Rhodes, J.A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554 https://doi.org/10.1214/09-AOS689

[2] Ashtiani, H., Ben-David, S., Harvey, N., Liaw, C., Mehrabian, A. and Plan, Y. (2018). Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes. In *Advances in Neural Information Processing Systems* 3412–3421.

[3] Baker, L.D. and McCallum, A.K. (1998). Distributional clustering of words for text classification. In *Proceedings of the* 21*st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 96–103. ACM.

[4] Balakrishnan, S., Wainwright, M.J. and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

[5] Biernacki, C., Celeux, G. and Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recogn. Lett.* **20** 267–272.

[6] Bishop, C.M. (2006). Pattern recognition and machine learning. *Mach. Learn.* **128** 1–58.

[7] Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Comput. Statist. Data Anal.* **71** 52–78. MR3131954 https://doi.org/10.1016/j.csda.2012.12.008

[8] Carreira-Perpinán, M.A. and Renals, S. (2000). Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Comput.* **12** 141–152.

[9] Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A. and Cresko, W.A. (2013). Stacks: An analysis tool set for population genomics. *Mol. Ecol.* **22** 3124–3140.

[10] Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *J. Classification* **13** 195–212. MR1421665 https://doi.org/10.1007/BF01246098

[11] Chan, S.-O., Diakonikolas, I., Servedio, R.A. and Sun, X. (2014). Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing* 604–613. ACM.

[12] Courant, R. (2011). *Differential and Integral Calculus. Vol. II. Wiley Classics Library*. New York: Wiley. MR1009559

[13] Cover, T.M. and Thomas, J.A. (2012). *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley Interscience. MR2239987

[14] Diakonikolas, I. (2016). Learning structured distributions. In *Handbook of Big Data. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 267–283. Boca Raton, FL: CRC Press. MR3674822

[15] Evanno, G., Regnaut, S. and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Mol. Ecol.* **14** 2611–2620.

[16] Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164** 1567–1587.

[17] Figueiredo, M.A.T. and Jain, A.K. (2002). Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 381–396.

[18] Fjellstad, O.-E. and Fossen, T.I. (2016). A generalized multivariate logistic model and EM algorithm based on the normal variance mean mixture representation. In *Statistical Signal Processing Workshop* (*SSP*) 1–5. IEEE.

[19] Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635 https://doi.org/10.1198/016214502760047131

[20] Gershman, S.J. and Blei, D.M. (2012). A tutorial on Bayesian nonparametric models. *J. Math. Psych.* **56** 1–12. MR2903470 https://doi.org/10.1016/j.jmp.2011.08.004

[21] Gyllenberg, M., Koski, T., Reilink, E. and Verlaan, M. (1994). Nonuniqueness in probabilistic numerical identification of bacteria. *J. Appl. Probab.* **31** 542–548. MR1274807 https://doi.org/10.2307/3215044

[22] Hollander, M., Wolfe, D.A. and Chicken, E. (2014). *Nonparametric Statistical Methods*, 3rd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR3221959

[23] Juan, A., García-Hernández, J. and Vidal, E. (2004). EM initialisation for Bernoulli mixture learning. In *Structural, Syntactic, and Statistical Pattern Recognition* 635–643.

[24] Juan, A. and Vidal, E. (2004). Bernoulli mixture models for binary images. In *Pattern Recognition*, 2004. *ICPR* 2004. *Proceedings of the* 17*th International Conference on* **3** 367–370. IEEE.

[25] Kalai, A.T., Moitra, A. and Valiant, G. (2016). Disentangling Gaussians. *Commun. ACM* **55** 113–120.

[26] Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. and Mayrose, I. (2015). Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* **15** 1179–1191.

[27] Lazarsfeld, P.F., Henry, N.W. and Anderson, T.W. (1968). *Latent Structure Analysis* **109**. Boston, MA: Houghton Mifflin.

[28] Li, C., Wang, B., Pavlu, V. and Aslam, J. (2016). Conditional Bernoulli mixtures for multi-label classification. In *Proceedings of the* 33*rd International Conference on Machine Learning* 2482–2491.

[29] McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. New York: Wiley Interscience. MR1789474 https://doi.org/10.1002/0471721182

[30] McNicholas, P.D. (2016). Model-based clustering. *J. Classification* **33** 331–373. MR3575621 https://doi.org/10.1007/s00357-016-9211-9

[31] Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2012). *Foundations of Machine Learning*. *Adaptive Computation and Machine Learning*. Cambridge, MA: MIT Press. MR3057769

[32] Müller, P., Quintana, F.A., Jara, A. and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. *Springer Series in Statistics*. Cham: Springer. MR3309338 https://doi.org/10.1007/978-3-319-18968-0

[33] Najafi, A., Janghorbani, S., Motahari, S.A. and Fatemizadeh, E. (2019). Statistical association mapping of population-structured genetic data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **16** 638–649.

[34] Orbanz, P. and Teh, Y.W. (2011). Bayesian nonparametric models. In *Encyclopedia of Machine Learning* 81–89. Springer.

[35] Peakall, R. and Smouse, P.E. (2006). GENALEX 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* **6** 288–295.

[36] Pella, J. and Masuda, M. (2006). The Gibbs and split merge sampler for population mixture analysis from genetic data with incomplete baselines. *Can. J. Fish. Aquat. Sci.* **63** 576–596.

[37] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

[38] Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.

[39] Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* **67** 170–181.

[40] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81** 559–575.

[41] Rousseau, J. (2016). On the frequentist properties of Bayesian nonparametric methods. *Annu. Rev. Stat. Appl.* **3** 211–231.
[42] Studenỳ, M. and Vejnarová, J. (1998). The multiinformation function as a tool for measuring stochastic dependence. In *Learning in Graphical Models* 261–297. Springer.
[43] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems* 1385–1392.
[44] Tiedeman, D. (1955). On the study of types. In *Symposium on Pattern Analysis* 1–14.
[45] Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* **90** 7–24.
[46] Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM J. Res. Develop.* **4** 66–82. MR0109755 https://doi.org/10.1147/rd.41.0066
[47] Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analysis. *Multivar. Behav. Res.* **5** 329–350.
[48] Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B. et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38** 203–208.
[49] Zhou, H., Blangero, J., Dyer, T.D., Chan, K.K., Lange, K. and Sobel, E.M. (2017). Fast genome-wide QTL association mapping on pedigree and population data. *Genet. Epidemiol.* **41** 174–186.

Received June 2019 and revised October 2019