Bayesian estimation of the Kullback-Leibler divergence for categorical systems using mixtures of Dirichlet priors

Francesco Camaglia,[1] Ilya Nemenman,[2] Thierry Mora,[1,*] and Aleksandra M. Walczak[1,*]
[1]*Laboratoire de physique de l'École normale supérieure, CNRS, PSL University, Sorbonne Université and Université de Paris, 75005 Paris, France*
[2]*Department of Physics, Department of Biology, and Initiative for Theory and Modeling of Living Systems, Emory University, Atlanta, Georgia 30322, USA*

(Received 7 July 2023; accepted 18 January 2024; published 13 February 2024)

In many applications in biology, engineering, and economics, identifying similarities and differences between distributions of data from complex processes requires comparing finite categorical samples of discrete counts. Statistical divergences quantify the difference between two distributions. However, their estimation is very difficult and empirical methods often fail, especially when the samples are small. We develop a Bayesian estimator of the Kullback-Leibler divergence between two probability distributions that makes use of a mixture of Dirichlet priors on the distributions being compared. We study the properties of the estimator on two examples: probabilities drawn from Dirichlet distributions and random strings of letters drawn from Markov chains. We extend the approach to the squared Hellinger divergence. Both estimators outperform other estimation techniques, with better results for data with a large number of categories and for higher values of divergences.

DOI: 10.1103/PhysRevE.109.024305

I. INTRODUCTION

Understanding of the structure and function of a large number of biological systems requires comparison between two probability distributions of their states or activities, generated under different conditions. For example, one may be interested in how the distribution of neural firing patterns underlying typical vocalizations in a song bird is different from patterns used to drive atypical, exploratory vocal behaviors [1]. One can similarly ask how different are the distributions of stimuli encoded by two different firing patterns; the difference then can be viewed as a measure of semantic similarity between these patterns [2]. In the context of immunology, one is often interested in information theoretic quantities to quantify diversity or to assess differences between distributions of immune receptors [3,4]. In these and similar examples, the Kullback-Leibler (KL) divergence D_{KL}, also known as relative entropy, is often used as a measure of dissimilarity. It is a nonsymmetric measure of the difference between two probability distributions with a wide range of applications in information theory [5]. While not a distance in the mathematical sense, it is often the choice measure of dissimilarity since it can be applied to categorical (nonordinal) data, when the usual statistical moments such as the mean and variance are not well defined. Indeed, like other "information theoretic quantities," the KL divergence is not associated to the category itself, but rather to the underlying probability distribution [6].

Estimation of information theoretic quantities is a hard problem, with a lot of attempts in the recent literature. Most of these have focused on the entropy and mutual information, but estimation of the KL divergence has also been investigated [7]. When faced with data without any knowledge of the true underlying distribution, empirical approaches (typically referred to as "naive" [8] or "plug-in" [9]) are often used. These methods approximate the true probabilities of events with their empirical frequencies, with an optional pseudocount. These types of estimators have been investigated thoroughly. The consensus is that, for all entropic quantities, these estimates are typically strongly biased [9–12]. To overcome this limitation, other approaches have been proposed to estimate the Shannon entropy (or the mutual information) of categorical data. These techniques include Bayesian methods [13,14], coverage adjusted methods [15] and bias corrected methods [10,11,16]. In the case of the KL divergence, the cross-entropy term, which diverges due to contributions where one distribution has samples and the other does not, makes it difficult to extend these methods in the absence of information about the joint distribution. The bias-corrected "Z-estimator" [7], proposed for KL divergence estimation, tackles these issues. However, it has a strong dependence on the sample size.

Here we propose a Bayesian estimator of the DKL for systems with finite number of categories using a mixture of symmetric Dirichlet priors [Dirichlet prior mixture (DPM)]. This approach is the generalization of the main idea from Ref. [13] that, to produce unbiased estimators, one needs to start with Bayesian priors that are (nearly) uniform not on the space of probability distributions, but directly on the quantity being estimated. Here we extend this idea beyond the estimation of entropy, for which it was first developed. We check that, for data distributed according to a Dirichlet prior, our new approach for estimation of the KL divergence consistently converges faster to the true value than other methods. We provide an algebraically equivalent expression

[*]These authors contributed equally to this work.

 ©2024 American Physical Society

for the Z-estimator (following Ref. [17]), which makes it applicable to large sample sizes. We also test the DPM technique on sequences generated by Markov chains, which are not typical within the DPM prior, obtaining better performance for datasets with many categories. We then focus our analysis on another measure of similarity between categorical distributions, the Hellinger divergence [18], which, unlike the DKL, is a well defined bounded distance between distributions. To show the generality of our approach, we also develop a DPM estimator for the squared Hellinger divergence. In computational tests, we show the DPM approach to be accurate for this quantity as well. Since no estimation method can be guaranteed to estimate entropic quantities without a bias for an arbitrary underlying probability distribution, we finish by discussing the method's reliability when applied to real experimental data, where the true values of the divergences are not known a priori.

II. RESULTS

A. Bayesian framework for the estimation of the divergence

Our goal is to derive an estimate of the Kullback-Leibler divergence between the distributions of categorical variables $\boldsymbol{t}$ and $\boldsymbol{q}$, $D_{\mathrm{KL}}(\boldsymbol{q}\|\boldsymbol{t})$. We consider a discrete set of K categories labeled with $i = 1, \ldots, K$. Examples of categorical variables include "words" defined as sequences of neuron firing patterns (spike counts in time windows), sets of coexisting ecological or molecular species or a sequence of amino acids or nucleotides. Each category i has a certain (unknown) probability q_i in the first condition, and t_i in the second condition. We observe this category n_i times in an experiment done in the first condition, and collect the data in the histogram $\boldsymbol{n} = \{n_i\}_{i=1}^K$, with $N \equiv \sum_i n_i$. An experiment in the second condition returns the counts $\boldsymbol{m} = \{m_i\}_{i=1}^K$, with $M \equiv \sum_i m_i$. We want to estimate the Kullback-Leibler divergence between $\boldsymbol{t}$ and $\boldsymbol{q}$ [5], defined as

$$D_{\mathrm{KL}}(\boldsymbol{q}\|\boldsymbol{t}) = H(\boldsymbol{q}\|\boldsymbol{t}) - S(\boldsymbol{q}) = \sum_{i=1}^{K} q_i \log \frac{q_i}{t_i}, \tag{1}$$

where we defined the cross-entropy between $\boldsymbol{t}$ and $\boldsymbol{q}$, $H(\boldsymbol{q}\|\boldsymbol{t}) = -\sum_i q_i \log t_i$, and the Shannon entropy, $S(\boldsymbol{q}) = -\sum_i q_i \log q_i$ [19].

Taking inspiration from Nemenman *et al.* [13], we choose to estimate the D_{KL} in a Bayesian framework. The approach is summarized in Fig. 1. We do not have access to the true probability distributions $\boldsymbol{t}$ and $\boldsymbol{q}$, only to the empirical histograms $\boldsymbol{n}$ and $\boldsymbol{m}$. The simple method consisting in approximating $q_i \approx n_i/N$ and likewise for $\boldsymbol{t}$ into Eq. (1) is known to work very poorly [11,12]. The issue comes from the presence of categories never observed in one sample, while they are present in the other, resulting in divergence of the logarithmic term. To go beyond that, we construct a prior of the true distributions $P_{\mathrm{prior}}(\boldsymbol{q},\boldsymbol{t})$ and weight the estimate of the divergence by posterior over $\boldsymbol{q}$ and $\boldsymbol{t}$:

$$\langle D_{\mathrm{KL}}(\boldsymbol{q},\boldsymbol{t})|\boldsymbol{n},\boldsymbol{m}\rangle = \int d\boldsymbol{q}d\boldsymbol{t}\; P_{\mathrm{post}}(\boldsymbol{q},\boldsymbol{t}) D_{\mathrm{KL}}(\boldsymbol{q}\|\boldsymbol{t}), \tag{2}$$

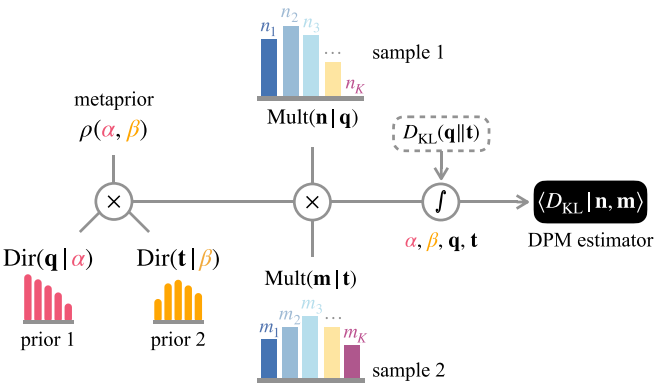

FIG. 1. Schematic representation of the Bayesian approach for the inference of the Kullback-Leibler divergence. Given two independent samples $\boldsymbol{n}$ and $\boldsymbol{m}$ of categorical data, we model the true distribution $\boldsymbol{q}$ and $\boldsymbol{t}$ as drawn from a mixture of Dirichlet distribution with unknown concentration parameters α and β. The DPM estimation of the D_{KL} is then obtained by averaging over all values of these parameters, weighted by the likelihood of the samples $\boldsymbol{n}$ and $\boldsymbol{m}$ (multinomial distributions).

where

$$P_{\mathrm{post}}(\boldsymbol{q},\boldsymbol{t}) = \frac{1}{Z} P_{\mathrm{prior}}(\boldsymbol{q},\boldsymbol{t}) P(\boldsymbol{n},\boldsymbol{m}|\boldsymbol{q},\boldsymbol{t}), \tag{3}$$

with $Z = P(\boldsymbol{n},\boldsymbol{m}) = \int d\boldsymbol{q}d\boldsymbol{t}\; P_{\mathrm{prior}}(\boldsymbol{q},\boldsymbol{t}) P(\boldsymbol{n},\boldsymbol{m}|\boldsymbol{q},\boldsymbol{t})$ a normalization.

The empirical observations $\boldsymbol{n}$ and $\boldsymbol{t}$ are assumed to be independent samples of $\boldsymbol{q}$ and $\boldsymbol{t}$ respectively, and are thus distributed according to a multinomial distribution:

$$P(\boldsymbol{n},\boldsymbol{m}|\boldsymbol{q},\boldsymbol{t}) = \mathrm{Mult}(\boldsymbol{n}|\boldsymbol{q})\mathrm{Mult}(\boldsymbol{m}|\boldsymbol{t}), \tag{4}$$

with

$$\mathrm{Mult}(\boldsymbol{n}|\boldsymbol{q}) \equiv \frac{N!}{\prod_i n_i!} \prod_{i=1}^{K} q_i^{n_i}. \tag{5}$$

A natural choice for the prior on $\boldsymbol{q}$ and $\boldsymbol{t}$ is the Dirichlet distribution, which is the conjugate prior of the multinomial distribution, and is defined as

$$\mathrm{Dir}(\boldsymbol{q}|\alpha) \equiv \frac{\delta\left(\sum_i q_i - 1\right)}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} q_i^{\alpha-1}, \tag{6}$$

where $B(\boldsymbol{x})$ is the multivariate β function:

$$B(\boldsymbol{x}) \equiv \frac{\prod_i \Gamma(x_i)}{\Gamma(\sum_i x_i)}, \tag{7}$$

where $\Gamma(x)$ is the Γ function. The parameter $\alpha \in (0,\infty)$ in Eq. (6) is the "concentration parameter," $\boldsymbol{\alpha} = \{\alpha\}_{i=1}^K$ and $\delta(x)$ is the Dirac's δ function imposing normalization. Rank plots associated to $\mathrm{Dir}(\boldsymbol{q}|\alpha)$ are shown in Fig. 3(a). For $\alpha \to \infty$, the prior tends to a uniform distribution $q_i = 1/K$. For small concentration parameters α, the distribution is peaked with weights given to just a few categories.

As noted in Ref. [13], entropies of distributions drawn from a Dirichlet with the same α all have similar entropies, strongly biasing the Shannon entropy estimate, especially in the large K limit. To reduce the bias, one then uses a mixture

of Dirichlet distributions at different α, allowing substantially different values of the entropy *a priori*. For a certain choice of the mixture distribution [the prior over α, $\rho(\alpha)$], one can achieve a nearly uniform *a priori* distribution of entropies and, consequently, a much smaller estimation bias [13,20]. We expect D_{KL} also to have very similar values for all distributions generated from the Dirichlet priors with fixed α and β. We then expect that a good estimator may be produced by using a mixture of Dirichlet distribution that allows to span different values of the expected D_{KL}:

$$P_{\mathrm{prior}}(\boldsymbol{q},\boldsymbol{t}) = \int_0^\infty \int_0^\infty d\alpha d\beta\ \rho(\alpha,\beta)\mathrm{Dir}(\boldsymbol{q}|\alpha)\mathrm{Dir}(\boldsymbol{t}|\beta), \quad (8)$$

where $\rho(\alpha,\beta)$ is a "hyper-prior," i.e., a prior over the hyper parameters α and β. Plugging this prior into Eqs. (2) and (3) gives

$$\langle D_{\mathrm{KL}}|\boldsymbol{n},\boldsymbol{m}\rangle = \frac{1}{Z}\int d\alpha d\beta\ P(\boldsymbol{n}|\alpha)P(\boldsymbol{m}|\beta)\rho(\alpha,\beta)\langle D_{\mathrm{KL}}|\boldsymbol{n},\boldsymbol{m};\alpha,\beta\rangle, \quad (9)$$

where the marginal likelihood $P(\boldsymbol{n}|\alpha)$ is equal to

$$P(\boldsymbol{n}|\alpha) = \int d\boldsymbol{q}\,\mathrm{Mult}(\boldsymbol{n}|\boldsymbol{q})\,\mathrm{Dir}(\boldsymbol{q}|\alpha) = \frac{B(\boldsymbol{n}+\boldsymbol{\alpha})N!}{B(\boldsymbol{\alpha})\prod_i n_i!} \quad (10)$$

and likewise for $P(\boldsymbol{m}|\beta)$. The normalization now reads

$$Z = \int d\alpha d\beta\ P(\boldsymbol{n}|\alpha)P(\boldsymbol{m}|\beta)\rho(\alpha,\beta). \quad (11)$$

The expected value of the D_{KL} inside the integral in Eq. (9) may be computed analytically (see Appendix A 1):

$$\begin{aligned}\langle D_{\mathrm{KL}}|\boldsymbol{n},\boldsymbol{m};\alpha,\beta\rangle &= \int d\boldsymbol{q}dt P(\boldsymbol{q},\boldsymbol{t}|\boldsymbol{m},\boldsymbol{n},\alpha,\beta)D_{\mathrm{KL}}(\boldsymbol{q}\|\boldsymbol{t})\\ &= \sum_{i=1}^{K}\frac{n_i+\alpha}{N+K\alpha}\{\Delta\psi(M+K\beta,m_i+\beta)\\ &\quad -\Delta\psi(N+K\alpha+1,n_i+\alpha+1)\}, \quad (12)\end{aligned}$$

where $\Delta\psi(z_1,z_2)=\psi(z_1)-\psi(z_2)$ is the difference of digamma functions ψ [i.e., polygamma function of order 0, see Eq. (A5)].

Similarly we can calculate $\langle D_{\mathrm{KL}}^2|\boldsymbol{n},\boldsymbol{m}\rangle$, which we can use to compute the posterior standard deviation of our method (Appendix A 1). For a given choice of $\rho(\alpha,\beta)$, the DPM estimate for D_{KL} in Eq. (9) can be computed numerically [same for D_{KL}^2 in Eq. (A27)], as described in detail in Appendix A 4. The code is available on github as specified in Appendix A 4 c.

We expect that, in the limit of large data ($N, M \gg K$), the integral of Eq. (9) will be dominated by the values of α and β that maximize the likelihoods $P(\boldsymbol{n}|\alpha)$ and $P(\boldsymbol{m}|\beta)$, regardless of the hyper-prior $\rho(\alpha,\beta)$. The dominant role of the likelihood $P(\boldsymbol{n}|\alpha)$ for increasing N was equivalently observed for the NSB entropy estimator [21]. By contrast, we expect the prior $\rho(\alpha,\beta)$ to play a role in the low-sampling regime, as can be seen from Fig. 2.

A simplified approach for the estimation of the D_{KL} would then be to provide a choice for the concentration parameters

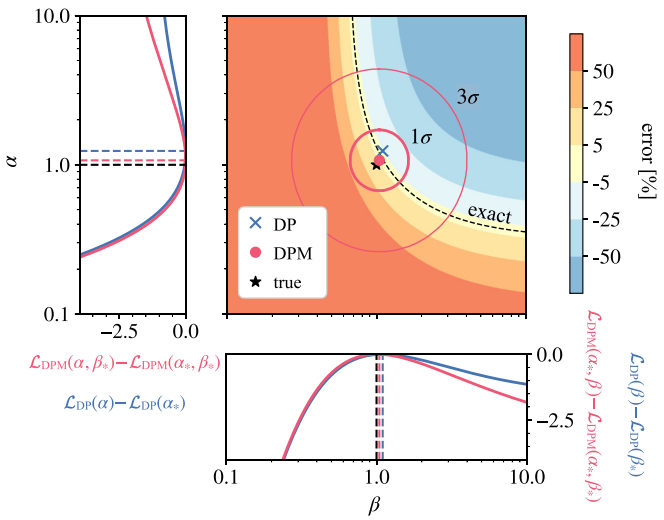

FIG. 2. Dependency on the concentration parameters α and β of the two main factors appearing in the average performed by the DPM estimator [Eq. (9)]: the posterior $\propto P(\boldsymbol{n}|\alpha)P(\boldsymbol{m}|\beta)\rho(\alpha,\beta)$, and the expected value $\langle D_{\mathrm{KL}}|\boldsymbol{n},\boldsymbol{m};\alpha,\beta\rangle$. For reasons of accuracy, integrals are numerically computed in logarithmic space, so that it is more informative to introduce the posterior density in $\log\alpha$ and $\log\beta$, so that we define: $\mathcal{L}_{\mathrm{DPM}}(\alpha,\beta)\equiv\log_{10}P(\boldsymbol{n}|\alpha)P(\boldsymbol{m}|\beta)\rho(\log\alpha,\log\beta)$. We compare it with the logarithm of the marginal likelihoods $\mathcal{L}_{\mathrm{DP}}(\alpha)\equiv\log_{10}P(\boldsymbol{n}|\alpha)$ and $\mathcal{L}_{\mathrm{DP}}(\beta)\equiv\log_{10}P(\boldsymbol{m}|\beta)$ in the left and bottom subplots. The central panel shows the relative error associated to $\langle D_{\mathrm{KL}}|\boldsymbol{n},\boldsymbol{m};\alpha,\beta\rangle$ as a function of the two concentration parameters α and β. The samples $\boldsymbol{n}$ and $\boldsymbol{m}$ of sizes $N=M=\frac{1}{4}K$ were generated from two distinct Dirichlet-multinomial processes with concentration parameters $\alpha_{\mathrm{true}}=1.0$ and $\beta_{\mathrm{true}}=1.0$ with $K=400$ (black star on the central panel). The dashed black line corresponds to 0 error. Blue cross: maximum of $\mathcal{L}_{\mathrm{DP}}(\alpha)+\mathcal{L}_{\mathrm{DP}}(\beta)$. Red circle: maximum of $\mathcal{L}_{\mathrm{DPM}}(\alpha,\beta)$. Red lines are standard deviations associated to $\mathcal{L}_{\mathrm{DPM}}$ around its maximum.

that maximizes the marginal likelihoods $P(\boldsymbol{n}|\alpha)$ [see Eq. (10)] and $P(\boldsymbol{m}|\beta)$ [22]. We refer to the application of Eq. (12) with such estimated values of α and β as the Dirichlet prior (DP) estimator.

B. Choosing the hyper-prior

To finalize the D_{KL} estimation, we need to choose a functional form for the hyper-prior $\rho(\alpha,\beta)$ in such a way that the resulting ensemble has an evenly distributed D_{KL}. In the limit of large numbers of categories ($K \gg 1$), both contributions of the D_{KL}, $S(\boldsymbol{q})$ and $H(\boldsymbol{q}\|t)$ are very peaked around their mean values, which can be computed analytically (Appendix A 1):

$$\mathcal{A}(\alpha)\equiv\langle S|\alpha\rangle = \Delta\psi(K\alpha+1,\alpha+1)\leqslant\log K \quad (13)$$

and

$$\mathcal{B}(\beta)\equiv\langle H|\alpha,\beta\rangle = \Delta\psi(K\beta,\beta)\geqslant\log K \quad (14)$$

(which only depends on β), at fixed concentration parameters. These mean values are shown in Fig. 3(b), and the corresponding $D_{\mathrm{KL}}=H-S$ in Fig. 3(c) as a function of α and β. We are interested in finding a hyper-prior such that the resulting prior over D_{KL} is not peaked. This results in the following inverse problem for finding the hyper-prior $\rho_z(z)$, where we denote

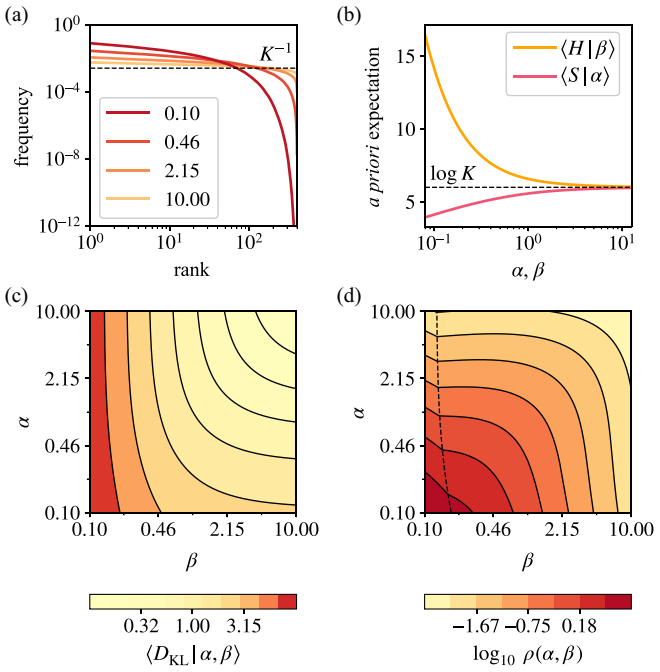

FIG. 3. (a) Average rank-frequency plots for probabilities drawn from a Dirichlet prior Dir($\boldsymbol{q}|\alpha$) for different choices of concentration parameters α. (b) Expected values of the cross-entropy $\langle H|\beta\rangle$ and the entropy $\langle S|\alpha\rangle$ under Dirichlet priors as functions of the concentration parameters. (c) Expected value of the D_{KL} divergence under Dirichlet priors $\langle D_{\mathrm{KL}}|\alpha,\beta\rangle$ as a function of the two concentration parameters. (d) Log-metaprior $\log_{10}\rho(\alpha,\beta)$ as a function of the two concentration parameters. Dashed black line represents the level $\langle D_{\mathrm{KL}}|\alpha,\beta\rangle = \log K$. $K = 20^2$ for all plots.

D_{KL} by z:

$$\rho_z(z) \approx \int_0^\infty d\alpha \int_0^\infty d\beta\ \rho(\alpha,\beta)\,\delta(\mathcal{B}(\beta) - \mathcal{A}(\alpha) - z), \quad (15)$$

with the choice $\rho_z(z)$ to be made. Because we have a one-dimensional target distribution $\rho_z(z)$, but a two-dimensional hyper-prior $\rho(\alpha,\beta)$, there are infinitely many solutions to this inverse problem. Without losing generality, we can make the change of variable from α and β to $\mathcal{A}$ and $\mathcal{B}$:

$$\rho_z(z) \approx \int_0^{\log K} d\mathcal{A} \int_{\log K}^{+\infty} d\mathcal{B}\ \rho_{AB}(\mathcal{A},\mathcal{B})\,\delta(\mathcal{B} - \mathcal{A} - z), \quad (16)$$

with

$$\rho(\alpha,\beta) = |\partial_\alpha \mathcal{A}|\big|\partial_\beta \mathcal{B}\big|\rho_{AB}(\mathcal{A}(\alpha),\mathcal{B}(\beta)). \quad (17)$$

Then a natural choice is to pick the Ansatz imposing that all values of $\mathcal{A}$ and $\mathcal{B}$ with the same D_{KL} are equiprobable:

$$\rho_{AB}(\mathcal{A},\mathcal{B}) = \phi(\mathcal{B} - \mathcal{A}). \quad (18)$$

Then $\phi(z)$ satisfies

$$\begin{aligned}\rho(z) &= \phi(z)\int_0^{\log K} d\mathcal{A}\ \theta(z + \mathcal{A} - \log K)\\ &= \phi(z)\{z\,\theta(\log K - z) + \log K\ \theta(z - \log K)\}, \end{aligned} \quad (19)$$

where $\theta(x) = 1$ if $x \geqslant 1$ and 0 otherwise (Heaviside function), or after inversion:

$$\phi(z) = \begin{cases} \rho(z)z^{-1} & z < \log K, \\ \rho(z)\dfrac{1}{\log K} & \text{otherwise.} \end{cases} \quad (20)$$

Equations (17), (18), and (20) give us the final form of the hyper-prior $\rho(\alpha,\beta)$. We are left with the choice of the distribution of the D_{KL}, $\rho(z)$. We pick a log-uniform (also known as "reciprocal") distribution, $\rho(z) \propto z^{-1}$ [23], allowing to evenly span over different orders of magnitude of the D_{KL}. The resulting hyper-prior is represented in Fig. 3(d).

C. Tests on synthetic Dirichlet samples

To assess the properties of the DPM estimator, we test it on data generated from distributions drawn from Dirichlets $\boldsymbol{q} \sim \mathrm{Dir}(\boldsymbol{q}|\alpha)$, $\boldsymbol{t} \sim \mathrm{Dir}(\boldsymbol{q}|\beta)$ [Eq. (6)], for various values of α and β. Having in mind applications to polypeptide sequences, we perform our tests for three different numbers of categories $K = 20^2$, 20^3, and 20^4, the numbers of all possible 2-mers, 3-mers, and 4-mers that can be produced with an alphabet of 20 letters (e.g., amino acids). For each choice of $\boldsymbol{q}$ and $\boldsymbol{t}$, samples $\boldsymbol{n}$ and $\boldsymbol{m}$ are generated from these distributions. This application may be viewed as a the consistency check for the estimator, since the estimator relied on the Dirichlet hypothesis, which is satisfied by the data.

We know that standard Bayesian consistency applies, ensuring that DPM (and DP) estimators converge to the true value in the limit of large samples. To understand how DPM estimator converges to the true value, we extract subsamples of increasing sizes $N = M$ from a larger sample of size 2×10^7. Figure 4 compares our D_{KL} estimate to several state-of-the-art estimators: the additive smoothing method with different values of the pseudocount (see below for details), the Z estimator, and the simplified version of our method, the DP estimator, obtained by fixing α and β to their maximum-likelihood values.

Additive smoothing estimators are defined as $D_{\mathrm{KL}}(\hat{\boldsymbol{q}}\|\hat{\boldsymbol{t}})$, with $\hat{q}_i = (n_i + a)/(N + Ka)$ and $\hat{t}_i = (m_i + b)/(M + Kb)$. We use four choices for the pseudocounts a and b, summarized in Table I. To avoid infinite values, in the case $b = 0$ we set to zero the terms for which $m_i = 0$.

It has been shown that naive estimators converge to the true value in the limit of large samples, but have an infinite bias due to low-probability categories [7]. The "Z-estimator" [7] was introduced to remove this bias asymptotically. Although its original definition was given as a series, one can show following Ref. [17] that its expression reduces to (Appendix A 2):

$$\widehat{D}_{\mathrm{KL}}^{(Z)} = \sum_{i=1}^{K} \frac{n_i}{N}[\Delta\psi(M+1, m_i+1) - \Delta\psi(N, n_i)], \quad (21)$$

where the first term in the sum corresponds to an estimator of $H(\boldsymbol{q}\|t)$, and the second term is the classic Schurmann-Grassberger estimator of the entropy $S(\boldsymbol{q})$ [16]. In Appendix A 2 we observe that $\langle D_{\mathrm{KL}}|\boldsymbol{n},\boldsymbol{m},\alpha,\beta\rangle \to \widehat{D}_{\mathrm{KL}}^{(Z)}$ in the limit $\alpha \to 0$, $\beta \to 1$, $N \gg K$ and $M \gg K$.

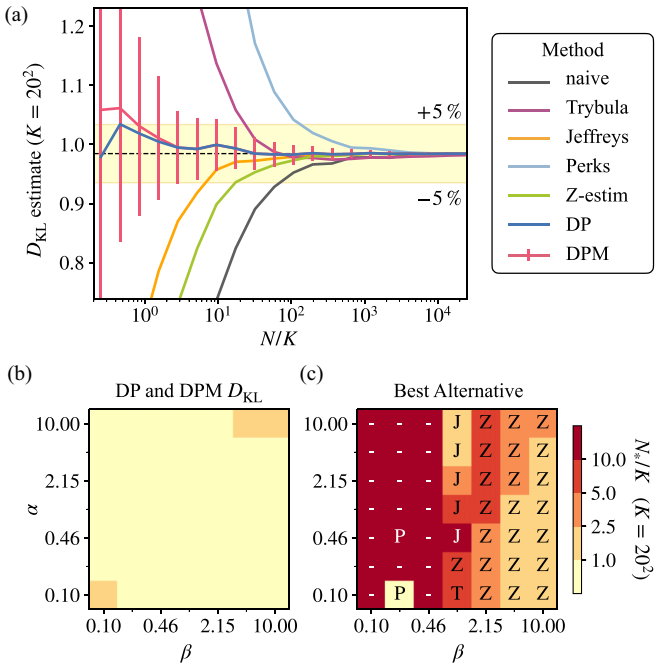

FIG. 4. Convergence of the D_{KL} estimates for increasing sample sizes. (a) We draw two independent histograms from Dirichlet-multinomial distributions with parameters α and β. We obtain subsamples of different sizes $N = M$ and we estimate the D_{KL} divergence for each of them. We compare the DP and DPM results to those obtained with the known alternative estimators [Table I and Eq. (21)], as a function of N/K. Here we plot the average over 100 repetitions for concentration parameters $\alpha = \beta = 1$ and $K = 20^2$. The highlighted region in yellow corresponds to an error of $\pm 5\%$ relative to the average true value, represented by the dashed black line. (b) Convergence of D_{KL} estimators for different (log-spaced) concentration parameters α, β. We plot the N_*/K score for the size at which the best between the DP and DPM estimators reach the true value up to a relative error of $\pm 5\%$ [highlighted region in panel (a)]. Lower N_*/K scores correspond to faster converge of the estimator. (c) The method with the best convergence score among the alternative methods is represented (first letter of its name). A dash symbol "-" indicates that no alternative has a score $N_*/K < 50$. The DP and DPM estimators shows faster convergence compared to all other methods for all parameters.

Comparing the convergence of the different estimators to the true D_{KL} value as a function of the subsample size N/K for $\alpha = \beta = 1$ and $K = 20^2$ [Fig. 4(a)], we see that the DPM performs better than other estimators. To assess how performance depends on the concentration parameters, we repeated this convergence analysis for different values of α and β. We measure convergence through N^*, defined as the sample size where the estimator get within 5% of the true value [Fig. 4(b)]. This measure of accuracy has the advantage to be applicable to all considered estimation methods.

Our estimator consistently performs well and compares favorably to other methods when data was generated from distributions drawn from symmetric Dirichlet. In most cases, the proposed DPM estimator converges faster than all other considered methods [Fig. 4(c)]. The better performance is striking also for larger numbers of categories, $K = 20^3$ and 20^4 (Fig. 7).

TABLE I. List of choices for the pseudocounts used to define alternative estimators of the D_{KL} [27]. $K^{(obs)} \leqslant K$ is the number of observed categories, for which $n_i > 0$ in each distinct sample.

Name	a	b	Reference
"Naive"	0	0	—
"Jeffreys"	0.5	0.5	[24]
"Trybula"	$\sqrt{N}/K$	$\sqrt{M}/K$	[25]
"Perks"	$1/K^{(obs)}$	$1/K^{(obs)}$	[26]

D. Tests on synthetic Markov chain sequences

To test the performance of DPM on a different synthetic system that does not satisfy the Dirichlet assumption, we generated L-long sequences (or "L-grams") from a Markov chain described by the transition matrix $\hat{W} \in \mathcal{M}_{20}$ with 20 states $\mu = 1, \ldots, 20$. We choose each transition probability $P(\mu \to \nu)$ from a uniform distribution in (0,1) and then impose that the transition matrix is a right stochastic matrix, $P(\mu \to \nu) = W_{\nu\mu}$ by normalizing to 1 each column of the transition matrix. An illustration where the states are the 20 amino acids is shown in Fig. 5(a). With this choice for the Markov transition matrix, all states communicate and are non absorbing. We verify there exists a stationary probability vector $\boldsymbol{\pi} = \{\pi_\mu\}_{\mu=1}^{L}$ that satisfies $\boldsymbol{\pi} = \hat{W}\boldsymbol{\pi}$. The number of categories is $K = 20^L$ and each category i corresponds to the L-gram $(x_1, \ldots, x_L)$ with the stationary probability q_i equal to $q_i = \pi_{x_1} W_{x_2 x_1} \ldots W_{x_L x_{L-1}}$.

We analytically compute the entropy associated to the stationary distribution $\boldsymbol{q}$ of L-grams to get

$$S^{(L)}(\boldsymbol{q}) = S(\boldsymbol{\pi}) - (L-1)\sum_{\mu\nu} W_{\nu\mu}\pi_\mu \log W_{\nu\mu}. \tag{22}$$

Typical values for the Shannon entropy of L-grams are shown in Fig. 5(b) along with the convergence curve of the NSB estimator. We assume that the L-grams of a second system are generated by a similar Markov process but with a transition matrix $\hat{V}$ and stationary probabilities of the $\boldsymbol{\sigma} = \{\sigma_\mu\}$ states. The cross-entropy between the $\boldsymbol{t}$ and $\boldsymbol{q}$ distributions reads

$$H^{(L)}(\boldsymbol{q}\|\boldsymbol{t}) = H(\boldsymbol{\pi}\|\boldsymbol{\sigma}) - (L-1)\sum_{\mu\nu} W_{\nu\mu}\pi_\mu \log V_{\nu\mu}. \tag{23}$$

Similar to the analysis in the previous section, we generate a large sample of L-grams from each distribution, with $N = M = 2\times 10^8$. We subsample this dataset at different sample sizes and estimate the D_{KL} and its standard deviation for $L = 2, 3, 4$. To study the average behavior, we divide the estimate by the expected result [Eq. (23)] and we average over 30 simulations.

We observe that, in the case of small numbers of categories [$K = 20^2$, Fig. 5(c) top panel], DPM (and DP) perform quite similar to the best alternative (Jeffreys), but with different sign biases. However, the DPM estimator performance greatly improves for larger K [Fig. 5(c) middle and bottom panels]. In all cases, the standard deviation associated to the DPM estimator [red bars in Fig. 5(c)] captures the spread across

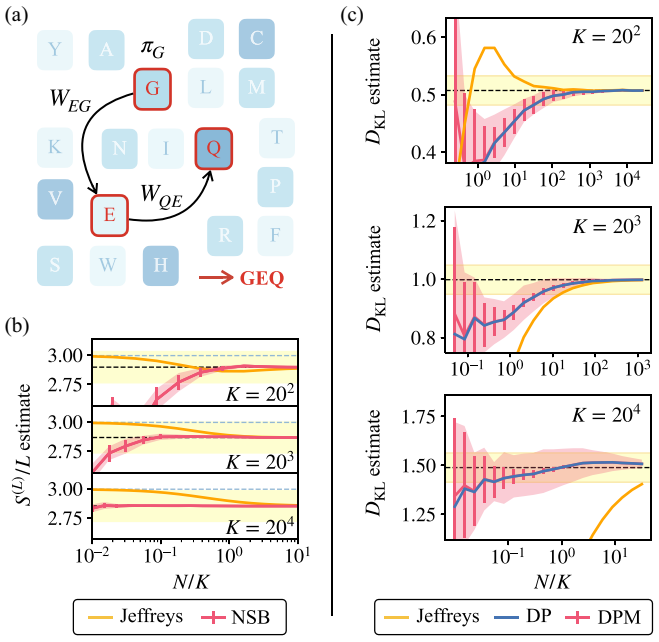

FIG. 5. (a) Schematic representation of the generation process of a $L = 3$-gram using a Markov chain with 20 states. (b) We generate a random Markov matrix and draw a sample of $2 \cdot 10^8$ independent L-grams, for $L = 2, 3, 4$. The Shannon entropy $S^{(L)}$ [Eq. (22)] estimated with the NSB method [13] for subsamples of size N, averaged over 30 repetitions normalized by the true value of the entropy and rescaled by the asymptotic value. The highlighted yellow region corresponds to a ±5% error range with respect to the average true value (dashed black line). (c) D_{KL} estimate and its standard deviation as a function of relative subsample size N/K, for $K = 20^2, 20^3, 20^4$. For $K = 20^2$, the DP and DPM estimators perform comparably to the best alternative ("Jeffreys" for all K), while they work better for larger K. The error bars represent the average posterior standard deviation of the D_{KL} estimate associated to the DPM method. The red shade is the standard deviation of the DPM D_{KL} estimates across the repetitions.

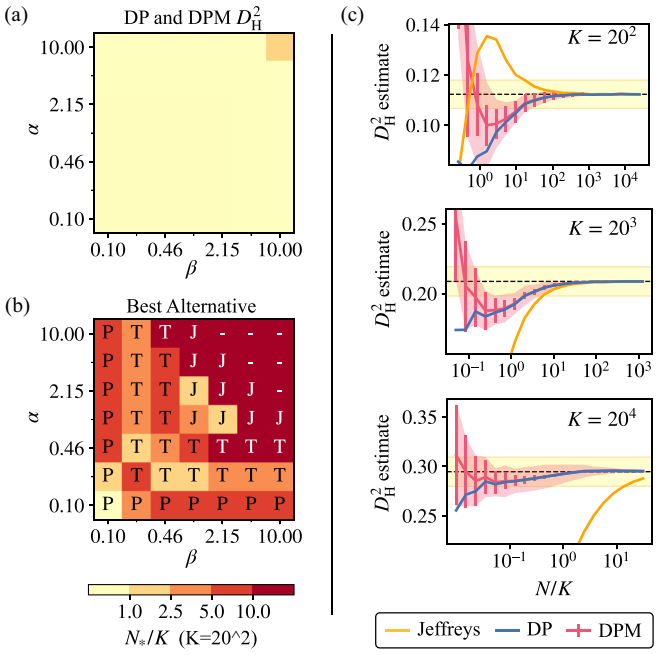

FIG. 6. Squared Hellinger divergence convergence. (a) Convergence score N_*/K of the DP and DPM D_H^2 estimators, tested on the same synthetic data as in Fig. 4(b). We consider 100 pairs of histograms drawn from Dirichlet-multinomial process with α and β concentration parameters. (b) We compare to the score of the best alternative method to the DP and DPM, chosen as pseudocount estimators with pseudocount given by Table I. These alternative methods perform worse for all values of the parameters. (c) Convergence of the DP and DPM D_H^2 estimators tested on the same Markov datasets as in Fig. 5(c) for different subsample sizes N. Each repetition is normalized by its true value, averaged and then rescaled by the asymptotic value (average of the true values). Red shade: the standard deviation across repetitions.

the different repetitions of the convergence curve [red shade in Fig. 5(c)].

E. Estimator for the Hellinger divergence

Last, we extend the DPM method to estimate the Hellinger divergence D_H between the discrete distributions $\boldsymbol{q}$ and $\boldsymbol{t}$ [18]. The Hellinger divergence is a symmetric statistical distance that satisfies the triangular inequality, making it a true distance in the mathematical sense [28]:

$$D_H(\boldsymbol{q},\boldsymbol{t})^2 = \frac{1}{2}\sum_{i=1}^{K}(\sqrt{q_i}-\sqrt{t_i})^2 = 1 - \sum_{i=1}^{K}\sqrt{q_i t_i}. \tag{24}$$

Following the same approach as for the Kullback-Leibler divergence (details in Appendix A 3), we obtain the DPM estimator for D_H^2:

$$\left\langle D_H^2 \middle| \boldsymbol{n},\boldsymbol{m}\right\rangle = \frac{1}{Z}\int d\alpha d\beta\ \rho_H(\alpha,\beta)P(\boldsymbol{n}|\alpha)P(\boldsymbol{m}|\beta)\left\langle D_H^2 \middle| \boldsymbol{n},\boldsymbol{m};\alpha,\beta\right\rangle, \tag{25}$$

with

$$\left\langle D_H^2 \middle| \boldsymbol{n},\boldsymbol{m};\alpha,\beta\right\rangle = 1 - \sum_{i=1}^{K}\frac{B\left(\frac{1}{2},N+K\alpha\right)}{B\left(\frac{1}{2},n_i+\alpha\right)}\frac{B\left(\frac{1}{2},M+K\beta\right)}{B\left(\frac{1}{2},m_i+\beta\right)}, \tag{26}$$

where $Z = \int d\alpha d\beta\ \rho_H(\alpha,\beta)P(\boldsymbol{n}|\alpha)P(\boldsymbol{m}|\beta)$ and $B(x_1,x_2) = \Gamma(x_1)\Gamma(x_1)/\Gamma(x_1+x_2)$ is the two-dimensional Beta function (see Appendix A 1).

We test the Hellinger divergence D_H estimator on the same synthetic datasets as in Fig. 4(b) (Fig. 6). For datasets generated with Dirichlet-multinomial distributed samples, the DPM outperforms all considered plug-in estimators $1 - \sum_{i=1}^{K}\sqrt{\hat{q}_i\hat{t}_i}$, with $\hat{q}_i$ and $\hat{t}_i$ defined as before with pseudocounts a, b chosen according to Table I [Fig. 6(a)]. As for the case of KL divergence, the performance improves for larger categories (Fig. 8). Tests on the synthetic Markovian L-grams (see previous paragraph) show the DPM estimator performs better for larger numbers of categories K, with comparable performance to the best alternative (Jeffreys) for $K = 20^2$ [Fig. 6(b)].

F. Limitations of the DPM method

The Bayesian framework developed so far implicitly assumes that the two distributions $\boldsymbol{q}$ and $\boldsymbol{t}$ are distinct. To test the impact of this assumption on the performance of the DP and DPM estimators, we estimate both D_{KL} and D_{H}^2 for independent samples drawn from the same distribution $\boldsymbol{q} = \boldsymbol{t} \sim \mathrm{Dir}(\boldsymbol{q}|\alpha)$. By definition $\alpha = \beta$. We show in Fig. 9 that both DP and DPM methods converge much more slowly towards the true value $D_{\mathrm{KL}} = 0$ for all α than the best alternative method (the Z-estimator). A faster convergence is observed for $D_{\mathrm{H}}^2 = 0$, with increased scores for larger α. These results are independent of the number of categories K.

III. DISCUSSION

Correctly estimating statistical divergences between two distributions is an open problem in the analysis of categorical systems. Alongside the entropy, divergences such as the Kullback-Leibler and the Hellinger distance, are an important tool in the analysis of categorical data [6].

We focused on categorical distributions with finite numbers of categories K (bounded domain), where K is a known quantity. We proposed a way (DPM) to extend the approach of Nemenman *et al.* [13] developed for Shannon entropy estimation, to Kullback-Leibler estimation. DPM introduces a mixture of symmetric Dirichlet priors with a log-uniform *a priori* expected divergence distribution [Eq. (20)]. We restricted our analysis to the case of the two finite samples of the same size N, although the method works for different sample sizes. We also propose a simplified estimator (DP), which assumes a Dirichlet prior with concentration parameter fixed to the maximum value of the likelihood. This estimator is faster to compute as it does not require to integrate over the concentration parameters.

We showed that the DPM method outperforms the tested empirical plug-in techniques in terms of D_{KL} estimation for synthetic data sampled from a Dirichlet-multinomial distribution with fixed concentration parameters. The estimation task gets harder for distributions with larger concentration parameters, i.e., closer to a uniform distribution, but easier for large numbers of categories K.

These convergence trends were confirmed by tests on sequences of L states generated by Markov chains. In this case, DPM compares well to the best plug-in estimator in the low sample size regime of $K = 20^2$ and outperforms it for $K \geqslant 20^3$. Similar results were obtained for the DPM estimate of the Hellinger divergence for both Dirichlet-multinomial and Markov chain datasets. To our knowledge, DPM estimator of the Hellinger divergence is the first attempt to extend the ideas of Ref. [13] and to build a uniform prior estimator for a non-entropy-like quantity.

Our tests were restricted to categorical systems with rank distributions having exponentially decaying tails. As previously discussed for the case of the NSB entropy estimator, the Dirichlet prior has major limitations in capturing the Shannon entropy if the system rank distribution is not decaying fast [14,20]. Many real systems exhibit long-tailed rank distributions that decay as power-laws [29], which are not well captured by a Dirichlet prior. Preliminary (unpublished) tests of the DPM method for such systems show poor performance. Similarly to the case of entropy estimates, we speculate that the limitations of this method are related to issues with the poor representation of long tails by Dirichlet priors. Introducing a Pitman-Yor prior [30] could overcome this problem, as has been shown for entropy estimation by Archer *et al.* [14], and offers a direction to generalize the applicability of the DPM method. Extending the Pitman-Yor prior to the case of statistical divergences would require to compute expected values over the probabilities of both systems, but to the best of our knowledge this is not possible because of the lack of an analytical expression for the Pitman-Yor distribution. Another difficulty may lie in the difficulty to encode correlations between the ranks of categories in the two distributions. Our priors assume that the two unknown distributions $\boldsymbol{q}$ and $\boldsymbol{t}$ are drawn independently. However, in real data they are generally correlated, which could have an impact on the quality of estimators when the distribution of frequencies becomes very broad.

In view of these complications, it is important to have practical criteria to ascertain if the output of the DPM estimator can be trusted for a specific dataset, or if it remains biased. Similar questions exist for estimation of many quantities, and specifically of entropic quantities, on categorial data since no estimator can be universally unbiased for them, and the decay of the bias with the sample size may be excruciatingly slow [9,10]. For entropy and mutual information, the standard approach is to observe if the empirical output drifts systematically as the sample size changes. One then declares the estimator trustworthy if the bias does not drift by more than the posterior standard deviation over about an order of magnitude change in the amount of data [8,31]. We expect this approach to transfer nearly verbatim to the D_{KL} and the Hellinger divergence context, easily detecting whether the DPM approach can be used for a specific dataset, or if other analysis methods should be sought.

ACKNOWLEDGMENTS

We thank Antonio C. Costa for helpful discussions. This work was partially supported by the European Research Council Consolidator Grant No. 724208 and ANR-19-CE45-0018 "RESP-REP" from the Agence Nationale de la Recherche. I.N. was supported in part by the Simons Foundation Investigator grant, Simons-Emory International Consortium on Motor Control, and by the U.S. NSF Grant No. 2209996.

APPENDIX

1. Mathematical relations

We first introduce mathematical relations and notations that are used for the computation of the DP and DPM estimators for D_{KL} and $D_{\mathrm{H}}{}^2$.

a. Wolpert-Wolf integrals

Given a vector $\boldsymbol{x} = \{x_i\}_{i=1}^K$, where $x_i \in (0, \infty)$ for all $i = 1, \ldots, K$, where K is a finite number of categories, the

Wolpert-Wolf [32] integral is a multivariate Beta function $B : \mathbb{R}_+^K \to \mathbb{R}_+$ in $\boldsymbol{x}$:

$$\int d\boldsymbol{q}\ \delta\left(\sum_{i=1}^{K} q_i - 1\right) \prod_{j=1}^{K} {q_j}^{x_j - 1} = \frac{\prod_{j=1}^{K} \Gamma(x_j)}{\Gamma(X)} = B(\boldsymbol{x}), \tag{A1}$$

where $X = \sum_i x_i$ and Γ is the Γ function:

$$\Gamma(x) \equiv \int_0^\infty dt\ e^{-t} t^{x-1}. \tag{A2}$$

All Bayesian calculations with multinomial likelihoods and multivariate Dirichlet priors involve the integral:

$$\begin{aligned} &\int d\boldsymbol{q}\ \delta\left(\sum_{i=1}^{K} q_i - 1\right) \prod_{j=1}^{K} f_j(q_j) \\ &\quad = \mathcal{L}^{-1}\left[\prod_{j=1}^{K} \mathcal{L}[f_j(q)](s)\right](q' = 1), \end{aligned} \tag{A3}$$

where the f_i are regular functions, $\mathcal{L}$ is the Laplace transform in q (which is a function of s) and $\mathcal{L}^{-1}$ is the inverse Laplace transform (which is a function of q').

b. Partial derivative operation

The "partial derivative operator" for the ith dimension $\partial_i = \frac{\partial}{\partial x_i}$ applied to the Beta function B returns

$$\begin{aligned} (\partial_i B)(\boldsymbol{x}) &= \int d\boldsymbol{q}\ \delta(\|\boldsymbol{q}\|_1 - 1) \prod_{j=1}^{K} {q_j}^{x_j - 1} \log q_i \\ &= B(\boldsymbol{x})[\psi(x_i) - \psi(X)], \end{aligned} \tag{A4}$$

where the function ψ is the polygamma function of order 0. The polygamma function of order ℓ is defined as

$$\psi_\ell(x) \equiv \frac{d^\ell}{dy^\ell} \log \Gamma(y)\bigg|_{y=x}. \tag{A5}$$

To simplify the calculations, we define the following quantities related to the partial derivative operation [Eq. (A4)]:

$$\Lambda_i(\boldsymbol{x}) \equiv \frac{(\partial_i B)(\boldsymbol{x})}{B(\boldsymbol{x})} = \Delta\psi(x_i, X), \tag{A6}$$

where we make use of the contraction $\Delta\psi(z_1, z_2) = \psi(z_1) - \psi(z_2)$. Iterating this derivation on the function B, one can express the second partial derivative as follows:

$$\begin{aligned} \Lambda_{ij}(\boldsymbol{x}) \equiv \frac{(\partial_i \partial_j B)(\boldsymbol{x})}{B(\boldsymbol{x})} &= \frac{\big(\partial_i \big(B \Lambda_j\big)\big)(\boldsymbol{x})}{B(\boldsymbol{x})} \\ &= \Lambda_i(\boldsymbol{x})\Lambda_j(\boldsymbol{x}) + \big(\partial_i \Lambda_j\big)(\boldsymbol{x}), \end{aligned} \tag{A7}$$

where the derivative $\partial_i \Lambda_j(\boldsymbol{x}) = \delta_{ij}\psi_1(x_i) - \psi_1(X)$ is a consequence of Eq. (A6), and δ_{ij} is the Kronecker δ.

c. Shift operation

We introduce the "shift operator" $e^{\lambda\partial_i}$ of parameter $\lambda \in \mathbb{R}$ for the ith dimension, with the condition $x_i + \lambda > 0$. The shift operator acts on the function B as follows:

$$\begin{aligned} (e^{\lambda\partial_i} B)(\boldsymbol{x}) &= B(\boldsymbol{x} + \lambda\widehat{i}) \\ &= \int d\boldsymbol{q}\ \delta\left(\sum_{i=1}^{K} q_i - 1\right) \prod_{j=1}^{K} {q_j}^{x_j - 1} {q_i}^{\lambda} \\ &= B(\boldsymbol{x}) \frac{B(\lambda, X)}{B(\lambda, x_i)}, \end{aligned} \tag{A8}$$

where $\widehat{i} = \{\delta_{ij}\}_{j=1}^K$ indicates the ith versor in the K-dimensional space of categories. The function $B(z_1, z_2)$ is the regular (two-dimensional) Beta function:

$$B(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1 + z_2)}. \tag{A9}$$

When $\lambda = n \in \mathbb{N}_+$, the shift simplifies to

$$(e^{n\partial_i} B)(\boldsymbol{x}) = B(\boldsymbol{x}) \prod_{n'=0}^{n-1} \frac{x_i + n'}{X + n'} \tag{A10}$$

as an immediate consequence of the recurrence relation $\Gamma(x+1) = x\Gamma(x)$. Similarly to the case of partial derivatives, we introduce a class of functions to deal with the shift:

$$\Omega_i(\boldsymbol{x}) \equiv \frac{(e^{\partial_i} B)(\boldsymbol{x})}{B(\boldsymbol{x})} = \frac{x_i}{X}, \tag{A11}$$

from which we compute two-dimensional shifts

$$\begin{aligned} \Omega_{ij}(\boldsymbol{x}) \equiv \frac{(e^{\partial_i} e^{\partial_j} B)(\boldsymbol{x})}{B(\boldsymbol{x})} &= \frac{e^{\partial_i}(B\Omega_j)(\boldsymbol{x})}{B(\boldsymbol{x})} \\ &= \Omega_i(\boldsymbol{x})\Omega_j(\boldsymbol{x} + \widehat{i}) = \frac{x_i}{X}\frac{(x_j + \delta_{ij})}{X + 1}. \end{aligned} \tag{A12}$$

d. Composed operations

Composing all these operations on the multivariate Beta function, we can obtain all the quantities presented in this work. We start computing the composition between shift [Eq. (A11)] and derivative operators [Eq. (A6)]:

$$\begin{aligned} \frac{(e^{\partial_i}\partial_j B)(\boldsymbol{x})}{B(\boldsymbol{x})} &= \frac{(e^{\partial_i}\Lambda_j B)(\boldsymbol{x})}{B(\boldsymbol{x})} \\ &= \frac{\Lambda_j(\boldsymbol{x} + \widehat{i})(e^{\partial_i} B)(\boldsymbol{x})}{B(\boldsymbol{x})} = \Omega_i(\boldsymbol{x})\Lambda_j(\boldsymbol{x} + \widehat{i}). \end{aligned} \tag{A13}$$

We point out that since the shift and derivative operators commute, the order of their application is not important. Using the same approach, we obtain the following results:

$$\frac{(e^{\partial_i} e^{\partial_j} \partial_k B)(\boldsymbol{x})}{B(\boldsymbol{x})} = \Omega_{ij}(\boldsymbol{x})\Lambda_k(\boldsymbol{x} + \widehat{i} + \widehat{j}), \tag{A14}$$

and

$$\frac{(e^{\partial_i} e^{\partial_j} \partial_k \partial_h B)(\boldsymbol{x})}{B(\boldsymbol{x})} = \Omega_{ij}(\boldsymbol{x})\Lambda_{kh}(\boldsymbol{x} + \widehat{i} + \widehat{j}). \tag{A15}$$

e. A priori and a posteriori expected values

The operations presented in the previous sections are used compute the posterior expected values $\langle F(\boldsymbol{q}, \boldsymbol{t}) | \boldsymbol{n}, \boldsymbol{m}; \alpha, \beta \rangle$ for

all the functions that can be expressed as

$$F(\boldsymbol{q},\boldsymbol{t}) = \sum_{i=1}^{K} f_i(\boldsymbol{q}) g_i(\boldsymbol{t}). \tag{A16}$$

Since the concentration parameters α, β are independent, for fixed concentration parameters the expected value of F factorizes

$$\langle F(\boldsymbol{q},\boldsymbol{t})|\boldsymbol{n},\boldsymbol{m};\alpha,\beta\rangle = \sum_{i=1}^{K} \langle f_i|\boldsymbol{n};\alpha\rangle\langle g_i|\boldsymbol{m};\beta\rangle, \tag{A17}$$

with

$$\begin{aligned}
&\langle f_i|\boldsymbol{n};\alpha\rangle \frac{B(\boldsymbol{n}+\boldsymbol{\alpha})N!}{B(\boldsymbol{\alpha})\prod_j n_j} \\
&= \int d\boldsymbol{q}\ \delta\left(\sum_{j=1}^{K} q_j - 1\right) \mathrm{Dir}(\boldsymbol{q}|\alpha)\mathrm{Mult}(\boldsymbol{n}|\boldsymbol{q}) f_i(\boldsymbol{q}) \\
&= \int d\boldsymbol{q}\ \delta\left(\sum_{j=1}^{K} q_j - 1\right) \frac{N!\prod_j {q_j}^{n_i+\alpha-1}}{B(\boldsymbol{\alpha})\prod_j n_j} f_i(\boldsymbol{q}).
\end{aligned} \tag{A18}$$

For all functions f_i that can be expressed in terms of partial derivative [Eq. (A4)] and/or shift operators [Eq. (A8)], a factor $B(\boldsymbol{n}+\boldsymbol{\alpha})$ appears and the expected value is obtained explicitly simplifying the constant factors. Specifically,

$$\langle q_i|\boldsymbol{n};\alpha\rangle = \frac{(e^{\partial_i}B)(\boldsymbol{n}+\boldsymbol{\alpha})}{B(\boldsymbol{n}+\boldsymbol{\alpha})}, \tag{A19}$$

$$\langle \log q_i|\boldsymbol{n};\alpha\rangle = \frac{(\partial_i B)(\boldsymbol{n}+\boldsymbol{\alpha})}{B(\boldsymbol{n}+\boldsymbol{\alpha})}, \tag{A20}$$

$$\langle q_i \log q_i|\boldsymbol{n};\alpha\rangle = \frac{(e^{\partial_i}\partial_i B)(\boldsymbol{n}+\boldsymbol{\alpha})}{B(\boldsymbol{n}+\boldsymbol{\alpha})}, \tag{A21}$$

$$\langle q_i q_j|\boldsymbol{n};\alpha\rangle = \frac{(e^{\partial_i}e^{\partial_j}B)(\boldsymbol{n}+\boldsymbol{\alpha})}{B(\boldsymbol{n}+\boldsymbol{\alpha})}, \tag{A22}$$

$$\langle q_i q_j \log q_i|\boldsymbol{n};\alpha\rangle = \frac{(e^{\partial_i}e^{\partial_j}\partial_i B)(\boldsymbol{n}+\boldsymbol{\alpha})}{B(\boldsymbol{n}+\boldsymbol{\alpha})}, \tag{A23}$$

and

$$\langle q_i q_j \log q_i \log q_j|\boldsymbol{n};\alpha\rangle = \frac{(e^{\partial_i}e^{\partial_j}\partial_i\partial_j B)(\boldsymbol{n}+\boldsymbol{\alpha})}{B(\boldsymbol{n}+\boldsymbol{\alpha})}. \tag{A24}$$

The *a priori* expected values are computed in the same way, noticing that $\langle f_j|\alpha\rangle = \langle f_j|\boldsymbol{n}=\boldsymbol{0};\alpha\rangle$.

f. KL divergence estimation

We can use the previous results to compute the *a posteriori* expected value for the D_{KL}. We start by computing the *a posteriori* expected value for the cross-entropy H which is given by

$$\begin{aligned}
\left\langle \sum_{i=1}^{K} H(\boldsymbol{q}\|\boldsymbol{t})|\boldsymbol{n},\boldsymbol{m},\alpha,\beta \right\rangle &= \sum_{i=1}^{K} \langle q_i|\boldsymbol{n},\alpha\rangle\langle \log t_i|\boldsymbol{m},\beta\rangle \\
&= \sum_{i}^{K} \frac{e^{\partial_i}B(\boldsymbol{n}+\boldsymbol{\alpha})}{B(\boldsymbol{n}+\boldsymbol{\alpha})} \frac{\partial_i B(\boldsymbol{m}+\boldsymbol{\beta})}{B(\boldsymbol{m}+\boldsymbol{\beta})} \\
&= \sum_{i}^{K} \frac{n_i+\alpha}{N+K\alpha} \Delta\psi(M+K\beta, m_i+\beta),
\end{aligned} \tag{A25}$$

where we took advantage of independence and used the relations Eqs. (A19) and (A20) to obtain the explicit expressions in Eqs. (A11) and (A6). Subtracting the *a posteriori* expected Shannon entropy $\langle S|\boldsymbol{n},\boldsymbol{m},\alpha,\beta\rangle = \sum_i \frac{n_i+\alpha}{N+K\alpha}\Delta\psi(N+K\alpha+1, n_i+\alpha+1)$, we finally obtain the KL expected value in Eq. (12):

$$\begin{aligned}
\langle D_{\mathrm{KL}}|\boldsymbol{n},\boldsymbol{m};\alpha,\beta\rangle = \sum_{i=1}^{K} \frac{n_i+\alpha}{N+K\alpha} \{&\Delta\psi(M+K\beta, m_i+\beta) \\
&- \Delta\psi(N+K\alpha+1, n_i+\alpha+1)\}.
\end{aligned} \tag{A26}$$

g. Squared KL divergence estimation

To compute the posterior standard deviation of the Kullback-Leibler divergence estimator, we calculate the expected value of the squared KL divergence:

$$\begin{aligned}
&\langle {D_{\mathrm{KL}}}^2|\boldsymbol{n},\boldsymbol{m}\rangle \\
&\quad = \int d\alpha d\beta\ P(\boldsymbol{n},\alpha)P(\boldsymbol{m}|,\beta)\rho(\alpha,\beta)\langle {D_{\mathrm{KL}}}^2|\boldsymbol{n},\boldsymbol{m};\alpha,\beta\rangle.
\end{aligned} \tag{A27}$$

Similarly to the case of D_{KL}, we can compute explicitly

$$\langle {D_{\mathrm{KL}}}^2|\boldsymbol{n},\boldsymbol{m};\alpha,\beta\rangle = \sum_{ij} \left\langle q_i q_j \log\frac{q_i}{t_i} \log\frac{q_j}{t_j} \middle| \boldsymbol{n},\boldsymbol{m};\alpha,\beta \right\rangle, \tag{A28}$$

which requires to rewrite

$$\begin{aligned}
&q_i q_j \log\frac{q_i}{t_i}\log\frac{q_j}{t_j} \\
&\quad = q_i q_j \log q_i \log q_j - 2 q_i q_j \log q_i \log t_j + q_i q_j \log t_i \log t_j.
\end{aligned} \tag{A29}$$

The explicit expression computed using Wolpert-Wolf properties [Eqs. (A6), (A7), (A12), (A14), (A15)] is

$$\begin{aligned}
&\langle q_i q_j \log\frac{q_i}{t_i}\log\frac{q_j}{t_j}|\boldsymbol{n},\boldsymbol{m};\alpha,\beta\rangle \\
&\quad = \frac{x_i(x_j+\delta_{ij})}{X(X+1)}\{\delta_{ij}\psi_1(x_i+2) - \psi_1(X+2) \\
&\qquad + \Delta\psi(x_i+1+\delta_{ij}, X+2)\cdot(i\leftrightarrow j) \\
&\qquad - 2\Delta\psi(x_i+1+\delta_{ij}, X+2)\Delta\psi(y_j, Y) \\
&\qquad + \delta_{ij}\psi_1(y_i) - \psi_1(Y) + \Delta\psi(y_i, Y)\cdot(i\leftrightarrow j)\},
\end{aligned} \tag{A30}$$

where we have introduced the following notation to contract the expression: $\boldsymbol{x} = \boldsymbol{n}+\boldsymbol{\alpha}, X = N+K\alpha, \boldsymbol{y} = \boldsymbol{m}+\boldsymbol{\beta}$ and $Y = M+K\beta$. The factor $(i\leftrightarrow j)$ means taking the term that it multiplies with inverted indexes i and j.

2. Zhang-Grabchak divergence estimator

In Ref. [7] Zhang and Grabchak proposed an estimator for the Kullback-Leibler divergence, defined as

$$\widehat{D}^{(z)}_{\mathrm{KL}} = \sum_{i=1}^{K} \frac{n_i}{N} \left\{ \sum_{v=1}^{M-m_i} \frac{1}{v} \prod_{s=1}^{v} \left(1 - \frac{m_i}{M-s+1}\right) - \sum_{v=1}^{N-n_i} \frac{1}{v} \prod_{s=1}^{v} \left(1 - \frac{n_i-1}{N-s}\right) \right\}, \tag{A31}$$

where v and s are dummy variables.

a. Expression of the Z-estimator

Schurmann [17] has shown that, in the entropy term of Eq. (A31), the summation in v of the ith element can actually be expressed in a more concise way as

$$\sum_{v=1}^{N-n_i} \frac{1}{v} \prod_{s=1}^{v} \left(1 - \frac{n_i-1}{N-s}\right) = \Delta\psi(N, n_i), \tag{A32}$$

times a factor n_i/N. The sum of these terms returns the Shurman-Grassberger entropy estimator $\widehat{S}_{\mathrm{SG}} = \sum_{i=1} \frac{n_i}{N} \Delta\psi(N, n_i)$ [16].

If we simply plug $N = M + 1$ and $n_i = m_i + 1$ in Eq. (A32), we can show that the analogous ith cross-entropy term in Eq. (A31) is equal to the following:

$$\sum_{v=1}^{M-m_i} \frac{1}{v} \prod_{s=1}^{v} \left(1 - \frac{m_i}{M-s+1}\right) = \Delta\psi(M+1, m_i+1). \tag{A33}$$

Finally, if we substitute Eqs. (A32) and (A33) in Eq. (A31), we obtain

$$\widehat{D}^{(z)}_{\mathrm{KL}} = \sum_{i=1}^{K} \frac{n_i}{N} [\Delta\psi(M+1, m_i+1) - \Delta\psi(N, n_i)], \tag{A34}$$

which is the expression in Eq. (21) of the main text.

b. Relation between the DP and the Z-estimator

The Z-estimator can be expressed as an *a posteriori* expected value of the D_{KL} at $\alpha = 0$ and $\beta = 1$, up to an additive constant. We start by showing the following relation

$$\begin{aligned} \lim_{\alpha\to 0} \langle S|\boldsymbol{n}, \alpha\rangle &= \sum_{i=1}^{K} \frac{n_i}{N} \Delta\psi(N+1, n_i+1) \\ &= \frac{1-K}{N} + \sum_{i=1}^{K} \frac{n_i}{N} \Delta\psi(N, n_i), \end{aligned} \tag{A35}$$

which makes use of the fact that $\psi(x+1) = \psi(x) + \frac{1}{x}$.

Considering now the cross-entropy term with $\beta = 1$, and performing the same limit as before, we observe that

$$\begin{aligned} \lim_{\alpha\to 0} \langle H|\boldsymbol{n}, \boldsymbol{m}, \alpha, \beta = 1\rangle &= \sum_{i=1}^{K} \frac{n_i}{N} \Delta\psi(M+K, m_i+1) \\ &= \Delta\psi(M+K, M+1) + \sum_{i=1}^{K} \frac{n_i}{N} \Delta\psi(M+1, m_i+1), \end{aligned} \tag{A36}$$

where we used the fact that $\Delta\psi(x, x) = \psi(x) - \psi(x) = 0$ to add the term $\Delta\psi(M+1, M+1)$ in the sum.

Recognizing the two terms in Eq. (A34) we subtract Eqs. (A35) and (A36) to obtain that

$$\begin{aligned} &\lim_{\alpha\to 0} \langle D_{\mathrm{KL}}|\boldsymbol{n}, \boldsymbol{m}, \alpha, \beta = 1\rangle \\ &\quad = \Delta\psi(M+K, M+1) + \frac{K-1}{N} + \widehat{D}^{(z)}_{\mathrm{KL}}. \end{aligned} \tag{A37}$$

3. DPM-squared Hellinger divergence estimator

We compute the DPM estimator for the squared Hellinger divergence D^2_{H} [Eq. (24)]. We do so by starting from the Bhattacharyya coefficient BC [33],

$$\mathrm{BC}(\boldsymbol{q}, \boldsymbol{t}) = \sum_{i=1}^{K} \sqrt{q_i}\sqrt{t_i} = 1 - {D_{\mathrm{H}}}^2(\boldsymbol{q}, \boldsymbol{t}). \tag{A38}$$

Its *a priori* expected value under the assumption of the prior

$$P_{\mathrm{prior}}(\boldsymbol{q}, \boldsymbol{t}) = p(\boldsymbol{q}, \boldsymbol{t}|\alpha, \beta) = \mathrm{Dir}(\boldsymbol{q}|\alpha)\mathrm{Dir}(\boldsymbol{t}|\beta) \tag{A39}$$

is equal to

$$\begin{aligned} \langle \mathrm{BC}|\alpha, \beta\rangle &= \sum_{i=1}^{K} \frac{(e^{\frac{1}{2}\partial_i} B)(\boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \frac{(e^{\frac{1}{2}\partial_i} B)(\boldsymbol{\beta})}{B(\boldsymbol{\beta})} \\ &= K \frac{B(\frac{1}{2}, K\alpha)}{B(\frac{1}{2}, \alpha)} \frac{B(\frac{1}{2}, K\beta)}{B(\frac{1}{2}, \beta)}, \end{aligned} \tag{A40}$$

where we used the shift property [Eq. (A8)] with parameter $\lambda = \frac{1}{2}$. Following the derivation of the D_{KL} in the main text, we choose a hyper-prior $\rho_{\mathrm{H}}(\alpha, \beta)$ to control the *a priori* squared Hellinger divergence $\langle {D_{\mathrm{H}}}^2|\alpha, \beta\rangle = 1 - \langle \mathrm{BC}|\alpha, \beta\rangle$:

$$\rho_{\mathrm{H}}(z) = \int d\alpha d\beta \; \rho_{\mathrm{H}}(\alpha, \beta)\, \delta\big(\big\langle D^2_{\mathrm{H}}\big|\alpha, \beta\big\rangle - z\big). \tag{A41}$$

We define $g(x) = \sqrt{K} B(\frac{1}{2}, Kx)/B(\frac{1}{2}, x)$, which is a function $g : \mathbb{R} \to [0, 1)$. Using a similar Ansatz of the one in the main text, we obtain $\rho_{\mathrm{H}}(\alpha, \beta) = |\partial_\alpha g(\alpha)||\partial_\beta g(\beta)|\phi(\langle D^2_{\mathrm{H}}|\alpha, \beta\rangle)$, where the condition in Eq. (A41) imposes

$$\phi(z) = \rho_{\mathrm{H}}(z) \frac{(1-z)^2}{z(2-z)}. \tag{A42}$$

We choose $\rho_{\mathrm{H}}(z)$ to be log-uniform.

Finally, knowing that the calculation for the posterior expected squared Hellinger divergence is analogous to the *a priori* expectation, we obtain the DPM-squared Hellinger estimator in Eqs. (25) and (26).

4. Numerical implementation

a. Computations with multiplicities

In the low sampling regime (sparse data), there is a limited number of values the counts can take, which means that many categories will see the same pairs of values $\boldsymbol{x_i} = (n_i, m_i)$. To reduce the computational cost associated to summation over the K categories, we introduce a set of "multiplicities" [10] contained in the vector $\nu_{\boldsymbol{x}}$, where each entry is the number of instances that appear n times in the first sample and m in the second. Since by construction the dimension of $\nu_{\boldsymbol{x}} \leqslant K$, we expressed all summation in terms of the multiplicities vector. Given a function of the two counts f, the sum over all categories is

$$\sum_{i=1}^{K} f(\boldsymbol{x_i}) = \sum_{\boldsymbol{x}} \nu_{\boldsymbol{x}} f(\boldsymbol{x}), \tag{A43}$$

where the last sum runs over the ensemble of distinct pairs of observed counts. In the case of double sums (e.g., for D_{KL}^2), one needs to re-express the function as

$$f(\boldsymbol{x_i}, \boldsymbol{x_j}) = \delta_{ij} f_{\parallel}(\boldsymbol{x_i}) + (1 - \delta_{ij}) f_{\perp}(\boldsymbol{x_i}, \boldsymbol{x_j}), \tag{A44}$$

where $f_{\parallel}$ and $f_{\perp}$ is the function f for $i = j$ and $i \neq j$. The summation over the terms in δ_{ij} is calculated as before, and the double summation is

$$\sum_{i,j} f_{\perp}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{\boldsymbol{x},\boldsymbol{x}'} \nu_{\boldsymbol{x}} \nu_{\boldsymbol{x}'} f_{\perp}(\boldsymbol{x}, \boldsymbol{x}'). \tag{A45}$$

These formulas allow us to exploit vectorial expressions in the numerical calculations.

b. Numerical integrations

Similar to Ref. [21], to compute numerically the quantities $\langle D_{\mathrm{KL}} | \boldsymbol{n}, \boldsymbol{m} \rangle$ [Eq. (9)] and $\langle {D_{\mathrm{KL}}}^2 | \boldsymbol{n}, \boldsymbol{m} \rangle$ [Eq. (A27)], we first seek for the maximum (α_*, β_*) of the quantity $\mathcal{L}(\alpha, \beta)$ (see Fig. 2 for further details). For accuracy, we perform this computation in logarithmic space of $\log \alpha$ and $\log \beta$. Rescaling $\mathcal{L}(\alpha, \beta)$ by its maximum, integrands are $\mathcal{O}(1)$ for $(\alpha, \beta) \sim (\alpha_*, \beta_*)$. To find the maximum of the log-evidence (minimum of the opposite), we use the "Limited-memory BFGS" optimization algorithm as coded in the function "minimize," module *optimize* of the Python package *scipy* (version 1.7.3).

We evaluate the integrals using the trapezoidal rule. From the Hessian at the maximum of the log-evidence, we compute the standard deviation in the α and the β-direction as if the posterior was Gaussian. We use this standard deviation to pick a range of parameters spanning 3 standard deviations on both sides of the maximum. We heuristically chose the number of bins within the ranges for the integration, to be equal to $10(\frac{K}{N})^2$ for α ($10(\frac{K}{M})^2$ for β).

c. Code availability

The software for the DP, DPM and alternative estimators of the Kullback-Leibler and the Hellinger divergence presented in this article are collected in a Python package which can be found in the repository [34]. In addition, the package provides a Python version for the NSB entropy estimator [13], and a NSB estimator for the Simpson index [35].

5. Supplementary figures

In this section, we provide supplementary figures for the main text. We explore the properties of the convergence of the DP and DPM estimators for larger numbers of categories K, where the samples are drawn from distinct Dirichlet distributions. For both the Kullback-Leibler (Fig. 7) and Hellinger (Fig. 8) divergences, the DP and DPM method show the best performances. We then restrict to the case of samples drawn from the same distribution, showing poor performance (Fig. 9).

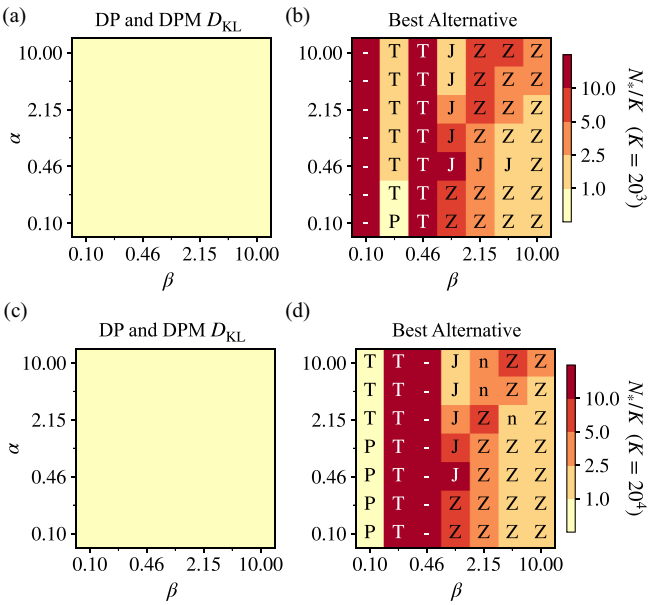

FIG. 7. The convergence of the D_{KL} estimates for different concentration parameters α, β. We use the same synthetic data of Fig. 4(b) to plot the best score N_*/K between DP and DPM for each combination of parameters. We choose as a score N_*/K, where N_* is the size $N = M$, at which the bias of the average estimate is smaller than 5%. The average is computed over 30 repetitions. (a) Case $K = 20^3$. (b) The first letter of the name of best alternative method or a symbol "-" if no method converges for $N_*/K < 50$. The DP and DPM always outperforms the best alternative in the tested cases. (c) DP and DPM convergence of Fig. 7(a) for the case $K = 20^4$. (d) Analogous of Fig. 7(b) for the case $K = 20^4$.

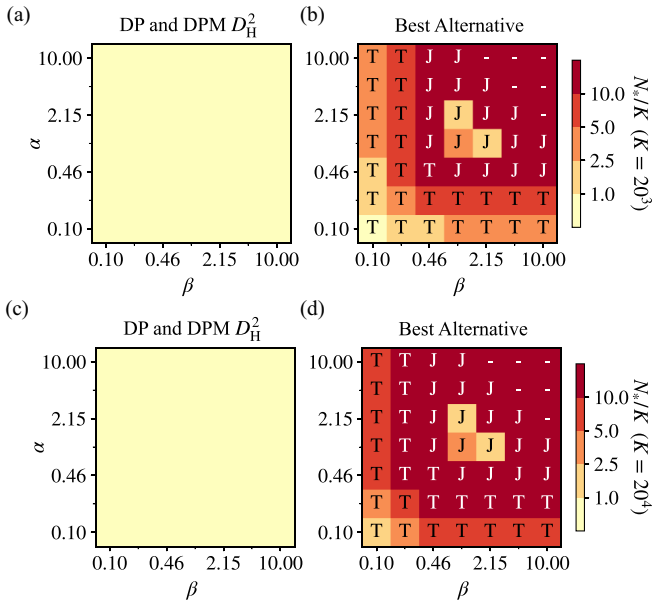

FIG. 8. The convergence of the D_{H}^2 estimates for different concentration parameters α, β. These figures use the same synthetic samples as in Fig. 6(a) and correspond to the case $K = 20^3$ and $K = 20^4$. Captions to panels (a)–(d) are analogous to Fig. 7.

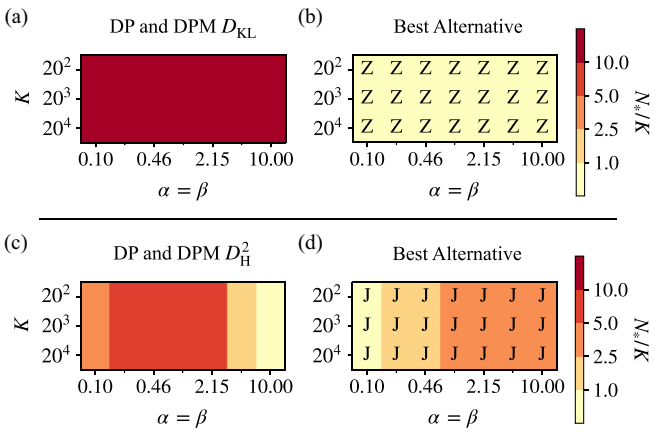

FIG. 9. The convergence of divergence estimates for samples drawn from the same Dirichlet distribution, i.e., $\boldsymbol{q} = \boldsymbol{t}$, at different concentration parameters $\alpha = \beta$. (a) The best between DP and DPM estimates converges to the true value $D_{\mathrm{KL}} = 0$ only for $N_*/K \in [10, 50]$. (b) For all concentration parameters the best alternative is the Z-estimator, which provides a faster convergence to the expected value. (c) Analogous results for the DP and DPM methods in the case of $D_{\mathrm{H}}{}^2 = 0$. (d) The best alternative is in all cases the Jeffreys estimator.

[1] D. G. Hernández, S. J. Sober, and I. Nemenman, Unsupervised Bayesian Ising approximation for decoding neural activity and other biological dictionaries, eLife **11**, e68192 (2022).

[2] E. Ganmor, R. Segev, and E. Schneidman, A thesaurus for a neural population code, eLife **4**, e06134 (2015).

[3] T. Mora and A. Walczak, Quantifying lymphocyte receptor diversity, in *Systems Immunology, An Introduction to Modeling Methods for Scientist*, edited by J. Das and C. Jayaprakash (CRC Press, Boca Raton, FL, 2019).

[4] F. Camaglia, A. Ryvkin, E. Greenstein, S. Reich-Zeliger, B. Chain, T. Mora, A. M. Walczak, and N. Friedman, Quantifying changes in the T cell receptor repertoire during thymic development, eLife **12**, e81622 (2023).

[5] S. Kullback and R. A. Leibler, On information and sufficiency, Ann. Math. Stat. **22**, 79 (1951).

[6] J. Zhang, Entropic statistics: Concept, estimation, and application in machine learning and knowledge extraction, Mach. Learn. Knowl. Extract. **4**, 865 (2022).

[7] Z. Zhang and M. Grabchak, Nonparametric estimation of Küllback-Leibler divergence, Neural Comput. **26**, 2570 (2014).

[8] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek, Entropy and information in neural spike trains, Phys. Rev. Lett. **80**, 197 (1998).

[9] A. Antos and I. Kontoyiannis, Convergence properties of functional estimates for discrete distributions, Random Struct. Alg. **19**, 163 (2001).

[10] L. Paninski, Estimation of entropy and mutual information, Neural Comput. **15**, 1191 (2003).

[11] Z. Zhang, Entropy estimation in Turing's perspective, Neural Comput. **24**, 1368 (2012).

[12] J. Jiao, K. Venkat, Y. Han, and T. Weissman, Maximum likelihood estimation of information measures, in *Proceedings of the IEEE International Symposium on Information Theory (ISIT'15)* (IEEE, Hong Kong, 2015), pp. 839–843.

[13] I. Nemenman, F. Shafee, and W. Bialek, Entropy and inference, revisited, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2001), Vol. 14.

[14] E. Archer, I. M. Park, and J. W. Pillow, Bayesian entropy estimation for countable discrete distributions, J. Mach. Learn. Res. **15**, 2833 (2014).

[15] A. Chao and T.-J. Shen, Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample, Environ. Ecol. Stat. **10**, 429 (2003).

[16] T. Schürmann, Bias analysis in entropy estimation, J. Phys. A: Math. Gen. **37**, L295 (2004).

[17] T. Schürmann, A note on entropy estimation, Neural Comput. **27**, 2097 (2015).

[18] E. Hellinger, Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen, J. Angew. Math. **1909**, 210 (1909).

[19] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379 (1948).

[20] D. G. Hernández, A. Roman, and I. Nemenman, Low-probability states, data statistics, and entropy estimation, Phys. Rev. E **108**, 014101 (2023).

[21] I. Nemenman, Coincidences and estimation of entropies of random variables with large cardinalities, Entropy **13**, 2013 (2011).

[22] A. Piga, L. Font-Pomarol, M. Sales-Pardo, and R. Guimerá, Bayesian estimation of information-theoretic metrics for sparsely sampled distributions, arXiv:2301.13647.

[23] R. W. Hamming, On the distribution of numbers, Bell Syst. Tech. J. **49**, 1609 (1970).

[24] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proc. R. Soc. A **186**, 453 (1946).

[25] S. Trybula, Some problems of simultaneous minimax estimation, Ann. Math. Stat. **29**, 245 (1958).

[26] W. Perks, Some observations on inverse probability including a new indifference rule, J. Inst. Actuaries **73**, 285 (1947).

[27] J. Hausser and K. Strimmer, Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks, J. Mach. Learn. Res. **10**, 1469 (2009).

[28] F. Liese and K. J. Miescke, Statistical models, in *Statistical Decision Theory: Estimation, Testing, and Selection*, Springer Series in Statistics (Springer, New York, NY, 2008), pp. 1–74.

[29] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley Press, Oxford, UK, 1949), pp. xi, 573.

[30] J. Pitman and M. Yor, The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator, Ann. Probab. **25**, 855 (1997).

[31] C. M. Holmes and I. Nemenman, Estimation of mutual information for real-valued data with error bars and controlled bias, Phys. Rev. E **100**, 022404 (2019).

[32] D. H. Wolpert and D. R. Wolf, Estimating functions of probability distributions from a finite set of samples, Phys. Rev. E **52**, 6841 (1995).

[33] A. Bhattacharyya, On a measure of divergence between two multinomial populations, Sankhya Indian J. Stat. **7**, 401 (1946).

[34] https://github.com/statbiophys/catede

[35] E. H. Simpson, Measurement of diversity, Nature (London) **163**, 688 (1949).