On latent idealized models in symbolic datasets: unveiling signals in noisy sequencing data

Antony Pearson · Manuel E. Lladser

Received: 29 February 2020 / Revised: 19 June 2023 / Accepted: 25 June 2023 / Published online: 10 July 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
Data taking values on discrete sample spaces are the embodiment of modern biological research. “Omics” experiments based on high-throughput sequencing produce millions of symbolic outcomes in the form of reads (i.e., DNA sequences of a few dozens to a few hundred nucleotides). Unfortunately, these intrinsically non-numerical datasets often deviate dramatically from natural assumptions a practitioner might make, and the possible sources of this deviation are usually poorly characterized. This contrasts with numerical datasets where Gaussian-type errors are often well-justified. To overcome this hurdle, we introduce the notion of latent weight, which measures the largest expected fraction of samples from a probabilistic source that conform to a model in a class of idealized models. We examine various properties of latent weights, which we specialize to the class of exchangeable probability distributions. As proof of concept, we analyze DNA methylation data from the 22 human autosome pairs. Contrary to what is usually assumed in the literature, we provide strong evidence that highly specific methylation patterns are overrepresented at some genomic locations when latent weights are taken into account.

Keywords Categorical data · Contamination · DNA methylation · Exchangeability · Omics · Sequencing errors · Symbolic data

Mathematics Subject Classification 62P10 · 62G10 · 92-08 · 92-10 · 92-11 · 92C40

1 Introduction

Symbolic data is the epitome modern biological datasets due to the advent of high-throughput sequencing assays. These assays generate millions of comparatively short
DNA sequences of a few dozen to a few hundred nucleotides and allow scientists to assess various microscopic processes such as investigating the relative abundance of unculturable organisms in an environment (Lladser et al. 2011; Hampton and Lladser 2012), or pinpointing the location of the enzymes that are actively transcribing DNA into RNA along a genome (Lladser et al. 2017)—among many other possibilities (Wang et al. 2009; Core et al. 2008; Park 2009; Suzuki and Bird 2008). Unfortunately, these datasets are often very noisy, and it is often unclear how to describe the noise because the possible sources of corruption—which range from sequencing errors (Posfai and Roberts 1992; Ilie et al. 2010; Medvedev et al. 2011; Lou et al. 2013; Schmitt et al. 2014) to biological and reagent contamination (Salter et al. 2014; Stinson et al. 2019; National Center for Biotechnology Information 2020)—can be so intricate that there is little motivation for any specific and let alone universal representation of their overall effect. In contrast, Gaussian noise is often well-justified, phenomenologically or theoretically, with continuous numerical datasets.

There is a rich history of using mixtures to describe deviations from idealized continuous models (Newcomb 1886; Tukey 1960; Huber 1964, 1965). Such mixtures are called contamination models in the statistics literature and are usually of the form: $P = (1 - \epsilon) \cdot Q + \epsilon \cdot R$, where $P$ is the probabilistic source producing the data, $Q$ is an idealized Gaussian distribution, and $R$ is some “contaminating” probability distribution from which an observation is drawn with some small probability $\epsilon$. (In this context, $(1 - \epsilon)$ and $\epsilon$ are called the weights of $Q$ and $R$, respectively.) When the mixture is unspecified and $P$ ought to be estimated from data, a highly specific structure for $R$ is usually needed, for example, a Gaussian with known mean, to make $\epsilon$ and $Q$ identifiable (Punzo and McNicholas 2016).

In this manuscript, we address the problem of assessing the weight of an idealized mixture component in symbolic (i.e. categorical or discrete) but possibly corrupted datasets. Like prior work on continuous data, we decompose the source producing the data as a mixture. Unlike previous lines of work, we treat deviations from an idealized structure as incidental; in particular, we do not commit to any pre-specified form for it. To do so, we introduce the notion of latent weight with respect to a given idealized class of probabilistic models (e.g., the exchangeable probability distributions, which are the main focus of this manuscript). Broadly speaking, this is the largest weight a model in the idealized class can have as a component of the source producing the data. In particular, it describes the largest expected fraction of samples from a random sample which conform to a probabilistic model in the idealized class.

We argue that latent weights are always identifiable, and allow one to represent unstructured probabilistic models as mixtures with an idealized component; in particular, when this component carries a substantial (latent) weight, most observations can be attributed to it.

To fix ideas, consider binary random variables $X$ and $Y$ with joint probability distribution given by the matrix

$$P := \begin{pmatrix}
Y = 0 & Y = 1 \\
1/10 & 3/10 \\
1/10 & 1/2
\end{pmatrix} \quad \begin{array}{l}
x = 0 \\
x = 1
\end{array}. \quad (1)$$
Since \( X \sim \text{Bernoulli}(3/5) \) and \( Y \sim \text{Bernoulli}(4/5) \), it is straightforward to check that \( X \) and \( Y \) are not independent. Nevertheless, we can ask whether or not \( P \) can be represented as a mixture with a component with independent marginals. This is equivalent to determining if the latent weight of \( P \) with respect to the class \( Q \) of probability measures on \([0, 1]^2\) with independent marginals\(^1\) is strictly positive or not (see Definition 1). It turns out the largest weight one can give in a mixture decomposition of \( P \) to the component \( \text{Bernoulli}(3/5) \otimes \text{Bernoulli}(4/5) \)—i.e. the model that would treat \( X \) and \( Y \) as independent random variables—is \( 5/6 \approx 83\% \). Indeed:

\[
P = \frac{5}{6} \cdot \text{Bernoulli}(3/5) \otimes \text{Bernoulli}(4/5) + \frac{1}{6} \cdot \begin{pmatrix} 1/5 & 1/5 \\ 0 & 3/5 \end{pmatrix}.
\]

The latent weight of \( P \) w.r.t. \( Q \) must be therefore at least \( 5/6 \). In fact, a simple calculation reveals that

\[
P = \frac{24}{25} \cdot \text{Bernoulli}(5/8) \otimes \text{Bernoulli}(5/6) + \frac{1}{25} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},
\]

i.e., the latent weight of \( P \) with respect to \( Q \) is at least \( 24/25 = 96\% \). Further analysis would show that it is precisely this; in particular, up to a hidden event with \( 96\% \) probability, \( X \) and \( Y \) behave independently.

The above finding is noteworthy for many reasons. On one hand, perhaps unexpectedly, the marginal distributions of \( X \) and \( Y \) are not associated with the latent weight of \( P \) with respect to the class \( Q \). On the other hand, if we derived our beliefs about the independence of \( X \) and \( Y \) from a large sample from \( P \) (e.g., using a chi-squared test of independence), at moderate significance levels, we would typically reject the hypothesis that \( X \) and \( Y \) are independent, ultimately missing that for most samples this is not the case. In this regard, latent weights could be used as a proxy for the approximate correctness of a hypothesis, without being tied to the absolutes of truth or falsity in classical hypothesis testing approaches. Finally, if \( P \) were estimated from a large but corrupted sample, as is usually the case with modern biological datasets, the large weight of \( \text{Bernoulli}(5/8) \otimes \text{Bernoulli}(5/6) \) as a component of \( P \) would suggest modeling \( X \) and \( Y \) as independent—as opposed to a more complicated model with a spurious correlation induced by a seemingly \( 4\% \) deviation from the assumption of independence.

**Paper organization.** Since our primary interest is on symbolic data, we restrict ourselves to the setting of finite sample spaces. From a technical point of view, this helps to develop the theory but without trivializing it. Section 2 introduces and analyzes the notion of latent weight associated with a general class of probability models. Sections 3, 4, 5 are exclusively devoted to the class of exchangeable probability models—see (Pearson and Lladser 2020) for an alternative application of this theory. Section 3.1 introduces some upper bounds that may be useful for assessing an exchangeable latent weight when the sample space is too large relative to the number of observations from the model. Section 4 develops the statistical machinery necessary

---

\(^1\) In other words, in this example, \( Q \) denotes the set of product measures of the form \((\mu \otimes \nu)\), with \( \mu \) and \( \nu \) probability models supported on \([0, 1]\).
for inference of exchangeable latent weights when the probabilistic source (producing the data) is observed only indirectly through a random sample. Finally, Sect. 5 demonstrates the use of latent weights to assess the exchangeability of DNA methylation, a near-universal assumption in epigenomic analyses.

2 Latent weights

In what follows, \( \mathcal{P} \) denotes the set of all probability measures over certain finite non-empty sample space \( \Omega \). As such, \( \mathcal{P} \) is compact under any norm induced metric; in particular, the total variation norm. Recall that for \( P_1, P_2 \in \mathcal{P} \), this norm is defined as (Lindvall 1992):

\[
\| P_1 - P_2 \| := \max_{A \subset \Omega} | P_1(A) - P_2(A) | = \frac{1}{2} \sum_{\omega \in \Omega} | P_1(\omega) - P_2(\omega) | .
\]

In what remains of this manuscript, \( Q \subset \mathcal{P} \) denotes a closed non-empty subset of probability measures. In particular, \( Q \) is a compact subset of \( \mathcal{P} \). Instances like this include, for example, singletons (i.e., when \( Q \) consists of a unique probability measure), as well as models with independent marginals (when \( \Omega \) is a product space), among various other possibilities.

**Definition 1** Let \( P \in \mathcal{P} \). We define the (latent) weight of \( Q \) in \( P \) as the coefficient \( \lambda_Q(P) := \sup\{ \lambda \text{ such that } P \geq \lambda \cdot Q \text{ for some } Q \in Q \} \), where \( P \geq \lambda \cdot Q \) means that \( P(\omega) \geq \lambda \cdot Q(\omega) \), for all \( \omega \in \Omega \).

This definition resembles that given in Chestnut and Lladser (2010), Lladser and Chestnut (2013) but for the very different purpose of representing a long-lasting Markov chain by shorter-lived independent chains.

Clearly, \( 0 \leq \lambda_Q(P) \leq 1 \). Latent weights have various other properties which we now state. Our first result characterizes models in \( Q \) in terms of latent weights.

**Theorem 1** \( \lambda_Q(P) = 1 \) if and only if \( P \in Q \).

**Proof** If \( \lambda_Q(P) = 1 \) then there exists a sequence of real numbers \( (\lambda_n)_{n \geq 1} \) such that \( 1 - 1/n < \lambda_n \leq 1 \), and \( P \geq \lambda_n \cdot Q_n \) for some \( Q_n \in Q \). Without loss of generality assume that \( \lambda_n < 1 \). In particular, if we define \( R_n := (P - \lambda_n \cdot Q_n)/(1 - \lambda_n) \) then \( R_n \in \mathcal{P} \) and \( P = \lambda_n \cdot Q_n + (1 - \lambda_n) \cdot R_n \). As a result: \( (P - Q_n) = (1 - \lambda_n) \cdot (R_n - Q_n) \), which implies that \( \| P - Q_n \| = (1 - \lambda_n) \cdot \| R_n - Q_n \| < 1/n \). Hence \( P \in Q \) because \( Q \) is closed. The converse is immediate because \( \lambda_Q(Q) = 1 \) for all \( Q \in Q \).

The following result shows how to compute latent weights and interpret them as weights of a mixture in a general class of idealized models \( Q \). It also states that the convexity of this class is sufficient to exclude the remaining component from it. Examples of convex classes include exchangeable distributions and mixtures of any number of models in a parametric family of distributions (e.g., mixtures of Poisson distributions with arbitrary rates).
Theorem 2

\[ \lambda_Q(P) = \sup_{Q \in \mathcal{Q}} \min_{\omega \in \Omega} \frac{P(\omega)}{Q(\omega)}, \]

(2)

where any division by zero (including zero-over-zero) is to be interpreted as \(+\infty\), and this supremum is achieved; in particular, there is \( Q \in \mathcal{Q} \) such that \( P \geq \lambda_Q(P) \cdot Q \) and there is \( R \in \mathcal{P} \) such that

\[ P = \lambda_Q(P) \cdot Q + (1 - \lambda_Q(P)) \cdot R. \]

(3)

Further, if \( Q \) is convex (i.e., \( \rho \cdot Q_1 + (1 - \rho) \cdot Q_2 \in \mathcal{Q} \), for any \( Q_1, Q_2 \in \mathcal{Q} \) and \( 0 \leq \rho \leq 1 \)) and \( \lambda_Q(P) < 1 \) then \( \lambda_Q(R) = 0 \); in particular, \( R \notin \mathcal{Q} \).

**Proof** Fix \( P \in \mathcal{P} \), and define \( \lambda^* := \sup_{Q \in \mathcal{Q}} \min_{\omega \in \Omega} P(\omega)/Q(\omega) \). For each \( Q \in \mathcal{Q} \), note that \( P \geq \lambda \cdot Q \) if and only if \( \min_{\omega \in \Omega} P(\omega)/Q(\omega) \geq \lambda \); in particular, \( \lambda^* \geq \lambda_Q(P) \).

Define \( \epsilon := \lambda^* - \lambda_Q(P) \). If \( \epsilon > 0 \) then, from the definition of \( \lambda^* \), there would be \( Q \in \mathcal{Q} \) such that \( \min_{\omega \in \Omega} P(\omega)/Q(\omega) \geq \lambda^* - \epsilon/2 \). In particular, for all \( \omega \in \Omega \), \( P(\omega) \geq (\lambda^* - \epsilon/2)/Q(\omega) \), hence \( \lambda_Q(P) \geq \lambda^* - \epsilon/2 > \lambda_Q(P) \), a contradiction. \( \epsilon = 0 \) i.e. \( \lambda^* = \lambda_Q(P) \).

Next, note that for each \( Q \in \mathcal{Q} \):

\[ \min_{\omega \in \Omega} \frac{P(\omega)}{Q(\omega)} = \min_{\omega : Q(\omega) > 0} \frac{P(\omega)}{Q(\omega)}. \]

(4)

Further, if \( \lim_{n \to \infty} Q_n = Q \) in total variation distance then \( \lim_{n \to \infty} Q_n(\omega) = Q(\omega) \) uniformly for all \( \omega \in \Omega \). The transformation \( Q \rightarrow \min_{\omega \in \Omega} P(\omega)/Q(\omega) \) from \( \mathcal{Q} \) to \( \mathbb{R} \) is therefore continuous; in particular, because \( \mathcal{Q} \) is compact, the supremum in Eq. (2) is achieved, and there is \( Q \in \mathcal{Q} \) such that \( P \geq \lambda_Q(P) \cdot Q \). For brevity, define \( \lambda := \lambda_Q(P) \). Following an argument similar to the proof of Theorem 1, it follows that there is \( R \in \mathcal{P} \) such that \( P = \lambda \cdot Q + (1 - \lambda) \cdot R \). Likewise, if \( \alpha := \lambda_Q(R) \) then there is \( Q' \in \mathcal{Q} \) and \( R' \in \mathcal{P} \) such that \( R = \alpha \cdot Q' + (1 - \alpha) \cdot R' \). As a result, if \( Q \) is convex then \( P \geq \lambda \cdot Q + (1 - \lambda) \alpha \cdot Q' = (\lambda + (1 - \lambda)\alpha) \cdot Q'' \) for some \( Q'' \in \mathcal{Q} \). But then \( \lambda \geq \lambda + (1 - \lambda)\alpha \), which implies that \( (1 - \lambda)\alpha = 0 \). Thus, if \( \lambda < 1 \) then \( \alpha = 0 \) as claimed. 

We emphasize that the probability measure \( Q \) in Theorem 2 is not necessarily unique. For instance, if \( \mathcal{Q} \) is the space of probability measures over \( \Omega = \{0, 1\}^2 \) with i.i.d. marginals, and \( P \) is the uniform probability measure over the set \( \{(0, 0), (1, 1)\} \) then \( \lambda_Q(P) = 1/2 \), and \( \delta_{(0,0)} \) and \( \delta_{(1,1)} \), the point masses at \((0, 0)\) and \((1, 1)\), respectively, both achieve the supremum in Eq. (2). Likewise, since \( P = \delta_{(0,0)}/2 + \delta_{(1,1)}/2 \), this counterexample shows the importance of convexity to guarantee that the probability measure \( R \) in Eq. (3) does not belong to \( \mathcal{Q} \).

The identity in Eq. (3) implies that any probability model \( P \) over \( \Omega \) admits a mixture representation with a component in \( \mathcal{Q} \) of weight \( \lambda_Q(P) \). (This motivates the terminology of “latent weight.”) The weight of \( Q \) in \( P \) may be interpreted therefore as the largest expected fraction of observations from \( P \) that can be attributed to a single model in \( \mathcal{Q} \).
We also note that a large latent weight is indicative of closeness in total variation distance to \( Q \). In fact, if \( Q \) and \( R \) are as in Eq. (3) then (see proof of Theorem 1): 
\[
\|P - Q\| = (1 - \lambda Q(P)) \cdot \|R - Q\|.
\]
So, if \( \lambda Q(P) \) is close to 1 then \( P \) is close to \( Q \) in total variation distance. The converse is not necessarily true, however. For instance, consider \( \Omega = [0, 1]^d \), with finite \( d > 1 \), and let \( 0^d \) and \( 1^d \) denote the sequences of \( d \) zeros and ones, respectively. Define \( P(\Omega) := (2^d - 2)^{-1} \) for \( \Omega \in \Omega \setminus \{0^d, 1^d\} \).

In particular, \( P(0^d) = P(1^d) = 0 \), and \( \|P - \text{Uniform}([0, 1]^d)\| = 2^{2-d} \), i.e. \( P \) is very close in total variation distance to \( Q \). Nevertheless, the latent weight of \( P \) with respect to the class \( Q \) of i.i.d. distributions over \([0, 1]^d\) is 0. Indeed, for all \( \lambda > 0 \) and \( Q := \otimes_{k=1}^d \text{Bernoulli}(q) \), with \( 0 \leq q \leq 1 \), it is not possible to have \( P \geq \lambda \cdot Q \). Otherwise, because \( Q(0^d) = (1 - q)^d \) and \( Q(1^d) = q^d \), it would follow that \( q = 1 \) and \( q = 0 \) simultaneously.

### 3 Exchangeable weights

In what remains of the manuscript we focus on the class of exchangeable probabilistic models. Recall that a finite sequence of random variables \( X_1, \ldots, X_d \) is called exchangeable if for any permutation \( \sigma \) of \( (1, \ldots, d) \) the random vector \( (X_{\sigma(1)}, \ldots, X_{\sigma(d)}) \) has the same distribution as \( (X_1, \ldots, X_d) \).

Exchangeability is a common a priori assumption in Bayesian statistics (De Finetti 1937), permutation hypothesis testing (Good 2002), and coalescent theory (Kingman 1982). The class of exchangeable models contains all finite sequences of independent and identically distributed (i.i.d.) random variables, but is generally much larger. For instance, sampling colored balls from an urn without replacement produces a random sequence of colors which is exchangeable but not necessarily independent.

In what follows, \((X_1, \ldots, X_d)\), with \( d > 1 \) finite, is a random vector with probability distribution \( P \). Each \( X_i \) is assumed to take values in a certain finite set \( \mathcal{X} \) of cardinality \( k > 1 \). In addition, \( \mathcal{P} \) denotes the set of all probability models over \( \mathcal{X}^d \), and \( \mathcal{E} \subset \mathcal{P} \) the subset of exchangeable models, which is clearly closed. We refer to the latent weight of \( P \) with respect to \( \mathcal{E}, \lambda_{\mathcal{E}}(P) \), as the exchangeable weight of \( P \).

**Definition 2** For each \( x \in \mathcal{X}^d \), let \([x] \subset \mathcal{X}^d \) denote the set of all vectors of the form \((x_{\sigma(1)}, \ldots, x_{\sigma(d)})\), with \( \sigma \) a permutation of \((1, \ldots, d)\).

The set \( \mathcal{X}^d \) has \( \binom{k+d-1}{d} \) permutation-equivalence classes. (The number of permutation equivalence classes equals the number of ways to place \( d \) unlabelled balls in \( k \) labelled urns.) In fact, \( \mathcal{E} \) is a simplex, in particular, also a convex set, with \( \binom{k+d-1}{d} \) extreme points, each of which is a probability model having uniform mass over a single permutation-equivalence class (Diaconis 1977). In what follows, \([\mathcal{X}^d]\) denotes the set of permutation equivalence classes of \( \mathcal{X}^d \).

Next we show various properties of (latent) exchangeable weights, starting with an explicit formula, and a condition guaranteeing the uniqueness of the component associated with the exchangeable weight.
Theorem 3  For all $P \in \mathcal{P}$:

$$\lambda_{\mathcal{E}}(P) = \sum_{x \in \mathcal{X}^d} \min_{y \in [x]} P(y) = \sum_{z \in [\mathcal{X}^d]} |z| \cdot \min_{x \in z} P(x). \quad (5)$$

If $\lambda_{\mathcal{E}}(P) > 0$ then exactly one probability measure $Q \in \mathcal{E}$ is associated with the exchangeable weight of $P$: for each $x \in \mathcal{X}^d$, $Q(x) = \min_{y \in [x]} P(y)/\lambda_{\mathcal{E}}(P)$.

Proof Define $\lambda^* := \sum_{x \in \mathcal{X}^d} \min_{y \in [x]} P(y)$. Suppose $\beta \geq 0$ and $Q' \in \mathcal{E}$ are such that $P \geq \beta \cdot Q'$. Since exchangeability implies $Q'(y) = Q'(x)$ for all $y \in [x]$, it follows that

$$\beta \cdot Q'(x) \leq \min_{y \in [x]} P(y).$$

Moreover, because $Q'$ is a probability measure, we have that

$$\beta = \sum_{x \in \mathcal{X}^d} \beta \cdot Q'(x) \leq \sum_{x \in \mathcal{X}^d} \min_{y \in [x]} P(y) = \lambda^*,$$

which implies that $\lambda_{\mathcal{E}}(P) \leq \lambda^*$. But observe that $P \geq \lambda^* \cdot Q$, where $Q(x) := \min_{y \in [x]} P(y)/\lambda^*$. Since $Q$ is an exchangeable probability measure over $\mathcal{X}^d$, it follows that $\lambda_{\mathcal{E}}(P) \geq \lambda^*$, hence $\lambda_{\mathcal{E}}(P) = \lambda^*$. The second identity in Eq. (5) is now direct from this equality.

Finally, assume that $\lambda^* > 0$ and suppose that $S \in \mathcal{E}$ is such that $P \geq \lambda^* \cdot S$. We show using an argument by contradiction that $S \leq Q$. Indeed, if there were $x \in \mathcal{X}^d$ such that $S(x) > Q(x)$ then, because $Q$ and $S$ are exchangeable, the definition of $Q$ would imply that there is $y \in [x]$ such that $S(y) > Q(y)$ and $\lambda^* \cdot Q(y) = P(y)$. In particular, $P(y) < \lambda^* \cdot S(y)$, which is not possible. Thus $S \leq Q$, which implies that $S = Q$ as claimed.

Due to Theorem 2, whenever $P$ is not itself exchangeable i.e. $\lambda_{\mathcal{E}}(P) < 1$, $P$ also has a unique component $R := (P - \lambda \cdot Q)/(1 - \lambda)$, which we call the unexchangeable component, which admits no exchangeable component of its own. In this sense $P$ can be distilled entirely into its exchangeable and unexchangeable parts. This property allows one to combine an exchangeable probability model with one that is totally unexchangeable so that the resultant source has a desired exchangeable weight:

Corollary 1  If $P = \beta \cdot Q' + (1 - \beta) \cdot R'$, where $0 \leq \beta \leq 1$, $Q' \in \mathcal{E}$, and $R' \in \mathcal{P} \setminus \mathcal{E}$ is such that $\lambda_{\mathcal{E}}(R') = 0$, then $\lambda_{\mathcal{E}}(P) = \beta$.

Proof  Since $\lambda_{\mathcal{E}}(R') = 0$, the identity in Eq. (5) implies that for each $z \in [\mathcal{X}^d]$ there is $y \in z$ such that $R'(y) = 0$. In particular, because $Q'$ is exchangeable, reusing Eq. (5) we obtain that:

$$\lambda_{\mathcal{E}}(P) = \sum_{x \in \mathcal{X}^d} \left( \beta \cdot Q'(x) + (1 - \beta) \cdot \min_{y \in [x]} R'(y) \right) = \beta \cdot \sum_{x \in \mathcal{X}^d} Q'(x) = \beta.$$ 

\(\Box\)
3.1 Bounds on exchangeable weights

The following three results may be useful to estimate an upper bound on the exchangeable weight of a probabilistic source on $X^d$. Indeed, estimating $\lambda_\mathcal{E}(P)$ when $P$ is known only indirectly through data requires estimating $(k^d - 1)$ free parameters, which may be infeasible in practice. However, marginalizing $(X_1, \ldots, X_d)$ to a sub-vector of dimension $s < d$, one may significantly reduce the dimension of the estimation problem.

In what follows, for each non-empty $I \subset \{1, \ldots, d\}$, $P_I$ denotes the marginal distribution of $(X_i)_{i \in I}$. Clearly, $P_I$ is exchangeable when $P$ is exchangeable.

**Theorem 4** The exchangeable weight of a full $d$-dimensional joint distribution is a lower bound on the exchangeable weight of any marginal, i.e. $\lambda_\mathcal{E}(P) \leq \lambda_\mathcal{E}(P_I)$, for all $P \in \mathcal{P}$ and non-empty $I \subset \{1, \ldots, d\}$.

In particular, if $\lambda_\mathcal{E}(P_I)$ is small for some $I$ then so is $\lambda_\mathcal{E}(P)$; in which case, only very few of the data produced by $P$ could be attributed to an exchangeable source. We use this fact in the proof of concept Sect. 5.

**Proof** For each $x \in \mathcal{X}^{|I|}$ and $\alpha \in \mathcal{X}^{d-|I|}$, let $x \alpha$ be the vector $y \in \mathcal{X}^d$ such that $(y_i)_{i \in I} = x$ and $(y_i)_{i \notin I} = \alpha$. If $Q$ is the exchangeable probability measure given in Eq. (5) then

$$P_I(x) = \sum_{\alpha \in \mathcal{X}^{d-|I|}} P(x \alpha) \geq \sum_{\alpha \in \mathcal{X}^{d-|I|}} \lambda_\mathcal{E}(P) \cdot Q(x \alpha) = \lambda_\mathcal{E}(P) \cdot Q_I(x),$$

for each $x \in \mathcal{X}^{|I|}$. Hence, since $Q_I$ is exchangeable, $\lambda_\mathcal{E}(P) \leq \lambda_\mathcal{E}(P_I)$. □

The next result addresses the tightness of the inequality in Theorem 4, which is related to the notion of extendibility (Diaconis and Freedman 1980): for $1 \leq s \leq d$, an exchangeable probability model $\mu$ on $\mathcal{X}^s$ is called $d$-extendible if there is an exchangeable probability measure $\nu$ on $\mathcal{X}^d$ and $I \subset \{1, \ldots, d\}$ of cardinality $s$ such that $\mu = \nu_I$. Necessary and sufficient conditions for extendibility can be found in Gnedin (1996).

**Corollary 2** For any $P \in \mathcal{P}$ with $\lambda_\mathcal{E}(P) > 0$, and $I \subset \{1, \ldots, d\}$, if $\lambda_\mathcal{E}(P) = \lambda_\mathcal{E}(P_I)$ then the exchangeable component $P_I$ is $d$-extendible.

**Proof** Suppose that $\lambda_\mathcal{E}(P) = \lambda_\mathcal{E}(P_I) > 0$, with $|I| = s$. Let $Q$ and $\tilde{Q}$ denote the exchangeable components of $P$ and $P_I$, respectively. From the proof of Theorem 4, $P_I \geq \lambda_\mathcal{E}(P) \cdot Q_I = \lambda_\mathcal{E}(P_I) \cdot Q_I$; in particular, due to Theorem 3, $\tilde{Q} = Q_I$. □

The converse in the corollary is not necessarily true, however. For a counterexample, consider a probability model $P$ over $\{0, 1\}^3$ such that $P(1, 0, 1) = P(1, 1, 0) = P(1, 1, 1) = 1/3$; in particular, $\lambda_\mathcal{E}(P) = 1/3$. If $I = \{1, 2\}$ then $P_I(1, 0) = 1/3$ and $P_I(1, 1) = 2/3$, hence $\lambda_\mathcal{E}(P_I) = 2/3$, and $\tilde{Q} = \delta_{(1,1)}$, which is 3-extendible to $\delta_{(1,1,1)}$ yet $\lambda_\mathcal{E}(P) \neq \lambda_\mathcal{E}(P_I)$.

Finally, another way to bound from above the exchangeable weight of a probability measure on $X^d$ is to lump states in $X$, as described in the following result.

\[\text{Springer}\]
Theorem 5 Let $\mathcal{Y}$ be a finite set and $r : \mathcal{X} \to \mathcal{Y}$ a function, and define $\Phi : \mathcal{X}^d \to \mathcal{Y}^d$ as $\Phi(x) := (r(x_1), \ldots, r(x_d))$. Then, for each $P \in \mathcal{P}(\mathcal{X}^d)$, $\lambda_{\mathcal{E}}(P) \leq \lambda_{\mathcal{E}}(P \circ \Phi^{-1})$, where $P \circ \Phi^{-1}$ is the forward measure of $P$ by $\Phi$.

Proof If $Q$ denotes the exchangeable component of $P$ then

$$\lambda_{\mathcal{E}}(P \circ \Phi^{-1}) = \sum_{[y] \subset \mathcal{Y}^d} |[y]| \cdot \min_{z \in [y]} P(\Phi^{-1}(z))$$

$$\geq \sum_{[y] \subset \mathcal{Y}^d} |[y]| \cdot \min_{z \in [y]} \lambda_{\mathcal{E}}(P) \cdot Q(\Phi^{-1}(z))$$

$$= \lambda_{\mathcal{E}}(P) \cdot \sum_{[y] \subset \mathcal{Y}^d} |[y]| \cdot \min_{z \in [y]} Q(\Phi^{-1}(z)) = \lambda_{\mathcal{E}}(P),$$

where for the very last identity we have used that $Q$ is exchangeable; in particular, $\min_{z \in [y]} Q(\Phi^{-1}(z)) = Q(\Phi^{-1}(y))$, for each $y \in \mathcal{Y}^d$. $\square$

For a potential application of this result, imagine a researcher interested in the exchangeable weight of a length-$d$ DNA sequence produced by a particular source $P$, which may be observed only indirectly through data. This would require estimating $\Theta(4^d)$ free parameters, which may be unfeasible for the available number of samples from $P$; however, encoding DNA bases as purines and pyrimidines, the number of free parameters would be reduced by a factor of $2^d$. In particular, if the estimated exchangeable weight w.r.t. the lumping of DNA bases as purines and pyrimidines was small, the inaccessible exchangeable weight of the full DNA sequence should also be small.

4 Estimation of exchangeable weights

In this section, $X_1, \ldots, X_n$ denote $d$-dimensional i.i.d. samples from a probability measure $P$ defined over $\mathcal{X}^d$. Define $\lambda := \lambda_{\mathcal{E}}(P)$.

Let $\hat{P}_n := \sum_{i=1}^n \delta_{X_i} / n$ denote the empirical measure associated with the sample. A natural estimator of $\lambda$ is $\hat{\lambda}_n := \lambda_{\mathcal{E}}(\hat{P}_n)$. Since $\hat{P}_n$ is a maximum likelihood estimator (MLE) of $P$, $\hat{\lambda}_n$ is a MLE of $\lambda$. Moreover, as noted in the proof of Theorem 2, $\lambda_{\mathcal{E}}(\cdot)$ is continuous. In particular, since $\hat{P}_n \to P$ almost surely, $\hat{\lambda}_n \to \lambda$ also almost surely. Since $0 \leq \hat{\lambda}_n \leq 1$ for all $n \geq 1$, it follows that $\hat{\lambda}_n$ is an asymptotically unbiased estimator of $\lambda$. Nevertheless, because for each $z \in [\mathcal{X}^d]$ the transformation $P \to \min_{y \in \mathcal{Y}} P(y)$ is concave down, Eq. (5) implies that $\lambda_{\mathcal{E}}(\cdot)$ is concave down. Thus, by Jensen’s inequality, $E(\hat{\lambda}_n) \leq \lambda$, i.e. $\hat{\lambda}_n$ is a negatively biased estimator of $\lambda$.

For each $x \in \mathcal{X}^d$, consider the quantities

$$m_x := \min_{y \in [x]} P(y); \quad \sigma_x^2 := m_x(1 - m_x).$$
Since these quantities remain constant within each permutation equivalence class, we sometimes abuse the notation and write for \( z \in [\mathcal{X}^d] \): \( m_z \) and \( \sigma_z^2 \) to mean \( m_x \) and \( \sigma_x^2 \), with \( x \in z \), respectively.

Our next result characterizes implicitly the asymptotic distribution of \( \hat{\lambda}_n \).

**Theorem 6** If for each \( z \in [\mathcal{X}^d] \), \( C_z := \{ x \in z \text{ such that } P(x) = m_x \} \) and \( c_z := |C_z| \) then

\[
\lim_{n \to \infty} \sqrt{n}(\hat{\lambda}_n - \lambda) \overset{d}{=} \sum_{z \in [\mathcal{X}^d]} |z| \cdot \min_{x \in C_z} Z^{(x)},
\]

where \((Z^{(x)})_{z \in [\mathcal{X}^d], x \in C_z}\) is a normal random vector such that, for each \( z \in [\mathcal{X}^d] \), \((Z^{(x)})_{x \in C_z}\) is a \( c_z \)-dimensional zero-mean exchangeable normal random vector with variance-covariance matrix \( \Sigma_z \) such that \( \Sigma_z(x, x) = \sigma_x^2 \) and \( \Sigma_z(x, y) = -m_x^2 \), for all \( x, y \in C_z \) with \( x \neq y \), and for \( z_1, z_2 \in [\mathcal{X}^d] \) with \( z_1 \neq z_2 \), \( \text{cov}(Z^{(x)}_{z_1}, Z^{(y)}_{z_2}) = -m_{z_1}m_{z_2} \), for all \( x \in z_1 \) and \( y \in z_2 \).

**Proof** Our arguments follow closely those in Hall et al. (1993).

For the sake of notation, we remove the sub-index \( n \) from quantities defined in terms of \( \hat{P}_n \); in particular, we write \( \hat{P} \) instead of \( \hat{P}_n \) and \( \lambda \) instead of \( \hat{\lambda}_n \). For \( x \in \mathcal{X}^d \), define \( \hat{M}_x := \min_{y \in \mathcal{X}^d} \hat{P}(y) \) and \( \hat{m}_x := \min_{y \in \mathcal{X}^d} \hat{P}(y) \). According to the Law of Large Numbers, \( \hat{P} \to P \) almost surely; in particular:

\[
P\left( \text{there is } x \in \mathcal{X}^d \text{ such that } \hat{M}_x \neq \hat{m}_x \right) = o(1). \tag{7}
\]

Define \( \delta := \sqrt{n}(\lambda - \hat{\lambda}) \) and \( \Delta := \sqrt{n}\left(\sum_{[x] \subset \mathcal{X}^d} |[x]| \cdot \min_{y \in C_x} \hat{P}(y) - \lambda \right) \). That is, \( \Delta \) is proportional to the difference between \( \lambda \) and an estimator of it computed from only those \( \hat{P}(x) \) corresponding to outcomes \( x \in \mathcal{X}^d \) for which \( P(x) = m_x \).

Fix \( t \in \mathbb{R} \). By Eqs. (5) and (7),

\[
|P(\delta \leq t) - P(\Delta \leq t)| = o(1). \tag{8}
\]

Furthermore, due to the well-known Central Limit Theorem for the multinomial distribution:

\[
P(\Delta \leq t) = P\left( \sum_{[x] \subset \mathcal{X}^d} |[x]| \cdot \min_{y \in C_x} \sqrt{n} \cdot (\hat{P}(y) - m_x) \leq t \right) \to P\left( \sum_{[x] \subset \mathcal{X}^d} |[x]| \cdot \min_{y \in C_x} Z^{(y)}_{[x]} \leq t \right), \tag{9}
\]

where \((Z^{(x)})_{z \in [\mathcal{X}^d], x \in C_z}\) is a zero-mean normal random vector with variance-covariance matrix as described above. The theorem is now a direct consequence of Eqs. (8) and (9).

Let \( \mathcal{P}_U \subset \mathcal{P} \) denote the set of \( P \in \mathcal{P} \) such that \( c_x = 1 \) for each \( x \in \mathcal{X}^d \), i.e. for each \( z \in [\mathcal{X}^d] \) there is a unique \( y \in z \) such that \( P(y) = m_z \). The following result is
an almost direct consequence of the previous theorem. This result also follows, albeit less directly, from the multivariate delta method (van der Vaart 1998).

**Corollary 3** If \( P \in \mathcal{P}_U \) then

\[
\lim_{n \to \infty} \sqrt{n}(\hat{\lambda} - \lambda) \overset{d}{=} Z,
\]

where \( Z \) is a zero-mean normal random variable with variance

\[
V(Z) = \sum_{[x] \subseteq \mathcal{X}^d} ||[x]||^2 \sigma_x^2 - \sum_{[x] \neq [y]} ||[x]|| ||[y]|| m_x m_y.
\]

**Proof** Due to the hypothesis on \( P \),

\[
\sqrt{n}(\hat{\lambda} - \lambda) \overset{d}{\to} Z = \sum_{[x] \subseteq \mathcal{X}^d} ||[x]|| \cdot Z_{[x]},
\]

where \( (Z_{[x]})_{[x] \subseteq \mathcal{X}^d} \) is an \( \binom{k+d-1}{d} \)-dimensional exchangeable normal random vector such that \( E(Z_{[x]}) = 0 \), \( V(Z_{[x]}) = m_x(1 - m_x) \), and for \( [x] \neq [y] \), \( \text{cov}(Z_{[x]}, Z_{[y]}) = -m_x m_y \). In particular, \( Z \) has a normal distribution, from which the corollary follows.

\( \square \)

Let \( \mathbb{X} := (X_1, \ldots, X_n) \) denote an i.i.d. sample from \( P \in \mathcal{P} \), and \( \mathbb{X}^* := (X_1^*, \ldots, X_n^*) \) denote a single resample with replacement from \( \mathbb{X} \). Let \( \hat{\lambda}^* \) denote the exchangeable weight associated with the empirical measure \( \hat{P}^* := \sum_{i=1}^n \delta_{X_i^*}/n \). In the next theorem and corollary we characterize the asymptotic distribution of the bootstrap distribution estimator \( \sqrt{n}(\hat{\lambda}^* - \hat{\lambda}) \).

In what follows, \( Z = (Z_{z}^{(x)})_{z \in [\mathcal{X}^d], x \in C_z} \) denotes the normal random vector described in Theorem 6, which has dimension \( \sum_{z \in [\mathcal{X}^d]} c_z \).

**Theorem 7** Fix \( t \in \mathbb{R} \), and associate with each vector \( v := (v_z^{(x)})_{z \in [\mathcal{X}^d], x \in C_z} \), the function:

\[
\psi(v) := \mathbb{P}\left( \sum_{z \in [\mathcal{X}^d]} |z| \cdot \left[ \min_{x \in C_z} (Z_z^{(x)} + v_z^{(x)}) - \min_{x \in C_z} v_z^{(x)} \right] \leq t \right).
\]

Note that \( \psi(0) = \lim_{n \to \infty} \mathbb{P}(\sqrt{n}(\hat{\lambda} - \lambda) \leq t) \), the C.D.F. of the limiting random variable described in Theorem 6. The bootstrap estimator of \( \psi(0) \) is

\[
\hat{\psi} := \mathbb{P}\left( \sqrt{n}(\hat{\lambda}^* - \hat{\lambda}) \leq t \mid \mathbb{X} \right),
\]

and

\[
\hat{\psi} \overset{d}{\to} \psi(Y), \quad (10)
\]
where Y is an independent copy of Z.

The proof of this theorem resembles closely the arguments given in Hall et al. (1993). We first require the following lemma.

**Lemma 1** Define $\hat{\mathcal{M}}_*^x := \min_{y \in C_x} \hat{P}^*(y)$ and $\hat{m}^*_x := \min_{y \in [x]} \hat{P}^*(y)$, i.e. $\hat{m}^*_x$ is the minimum probability estimated from a bootstrap resample in each equivalence class, and $\mathcal{M}_x^*$ is the same, estimated only from $C_x$. Then

$$\mathbb{P}(\text{there is } x \in \mathcal{X}^d \text{ such that } \hat{\mathcal{M}}_*^x \neq \hat{m}^*_x | \mathcal{X}) = o_p(1).$$

**Proof** If $P \in \mathcal{E}$, $\hat{M}_x = \hat{m}_x$, and $\hat{M}_x^* = \hat{m}_x$ with probability 1 for each $x$, and the claim follows.

Instead, for $P \in \mathcal{P} \setminus \mathcal{E}$, define

$$\xi := \min_{[x] \subset \mathcal{X}^d \text{ s.t. } |[x]| < c_x} \left\{ \min_{y \in [x] \setminus C_x} p_y - m_x \right\},$$

that is, the smallest difference between $p_x$, for $x \notin C_x$, and $m_x$. Note that $\xi$ is positive. Similarly, define $\hat{\xi} := \min_{[x] \subset \mathcal{X}^d \text{ s.t. } |[x]| < c_x} \{\min_{y \in [x] \setminus C_x} \hat{P}_y - \hat{M}_x\}$. (If $\hat{\xi} < 0$, then $\hat{M}_x > \hat{m}_x$ for some $[x]$.) The Law of Large Numbers guarantees that $\mathbb{P}(\hat{\xi} \geq \xi / 2) = (1 - o(1)).$

Define $Y_n := \mathbb{P}(\text{there is } x \in \mathcal{X}^d \text{ such that } \hat{\mathcal{M}}_*^x \neq \hat{m}^*_x | \mathcal{X})$, a random variable taking values in $[0, 1]$. Fix $\epsilon > 0$. Then

$$\mathbb{P}(Y_n \geq \epsilon) = \mathbb{P}(Y_n \geq \epsilon | \hat{\xi} \geq \xi / 2) \cdot (1 - o(1)) + \mathbb{P}(Y_n \geq \epsilon | \hat{\xi} < \xi / 2) \cdot o(1).$$

A second application of the Law of Large Numbers now yields $\mathbb{P}(Y_n \geq \epsilon | \hat{\xi} \geq \xi / 2) \to 0$, proving the lemma.

**Proof of Theorem 7** Define

$$\delta^* := \sqrt{n}(\hat{\lambda}^* - \hat{\lambda})$$

$$\Delta^* := \sqrt{n} \left( \sum_{z \in [\mathcal{X}^d]} |z| \cdot (\hat{\mathcal{M}}_*^z - \hat{M}_z) \right).$$

Due to Lemma 1:

$$\mathbb{P}(\delta^* \leq t | \mathcal{X}) = \mathbb{P}(\Delta^* \leq t | \mathcal{X}) + o_p(1).$$

\(\square\) Springer
Then it follows:

\[
P(\Delta^* \leq t \mid X) = P\left( \sqrt{n} \sum_{z \in [\mathcal{X}^d]} |z| \cdot (\hat{M}_z^* - \hat{M}_z) \leq t \mid X \right)
\]

\[
= P\left( \sum_{z \in [\mathcal{X}^d]} |z| \cdot \left[ \min_{x \in C_z} \left\{ \sqrt{n}(\hat{p}_x^* - \hat{p}_x) + \sqrt{n}(\hat{p}_x - m_z) \right\} - \min_{x \in C_z} \sqrt{n}(\hat{p}_x - m_z) \right] \leq t \mid X \right)
\]

\[
= \psi\left( \left( \sqrt{n}[\hat{p}_x - m_z] \right)_{z \in [\mathcal{X}^d], x \in C_z} \right) + o_p(1).
\] (13)

The last equality is due to the fact that, for almost every sample sequence \(X_1, X_2, \ldots, \), \(\sqrt{n}(\hat{\lambda}^* - \hat{\lambda})\) has the same conditional limiting distribution given \(X\) that \(\sqrt{n}(\hat{\lambda} - \lambda)\) does unconditionally, because \(\hat{\lambda}\) has finite second moments and mean \(P\) (Bickel and Freedman 1981, Theorem 2.2). That is, for any \(s \in \mathbb{R}^{kd}\), \(P(\sqrt{n}(\hat{\lambda}^* - \hat{\lambda}) \leq s \mid X) \to \lim_{n \to \infty} P(\sqrt{n}(\hat{\lambda} - \lambda) \leq s)\) almost surely (and therefore in probability). For this reason, we can replace \(\sqrt{n}(\hat{\lambda}^* - \hat{\lambda})\mid X\) in line (12) by its limiting random vector, which introduces some random perturbation to the conditional C.D.F. which converges in probability to 0.

Finally, because \(\psi(v)\) is a continuous function of \(v\), we take Eq. (13) with Eq. (11) to find that

\[
P(\delta^* \leq t \mid X) = \psi\left( \left( \sqrt{n}[\hat{p}_x - m_z] \right)_{z \in [\mathcal{X}^d], x \in C_z} \right) + o_p(1) \xrightarrow{d} \psi\left( \left( Y_z^{(x)}(x) \right)_{z \in [\mathcal{X}^d], x \in C_z} \right),
\]

as claimed. \(\square\)

The above convergence is in the weak sense. However, by requiring \(P \in \mathcal{P}_U\), we can ensure convergence in probability of the bootstrap distribution estimator. The following corollary follows simply from the fact that when \(c_z = 1\) for each \(z \in [\mathcal{X}^d]\), \(\min_{x \in C_z}(Z_z^{(x)} - Y_z^{(x)}) = \min_{x \in C_z} Y_z^{(x)} = Z_z^{(x)}\).

**Corollary 4** When \(P \in \mathcal{P}_U\),

\[
P(\sqrt{n}(\hat{\lambda}^* - \hat{\lambda}) \leq t \mid X) \xrightarrow{P} P(\sqrt{n}(\hat{\lambda} - \lambda) \leq t),
\]

for each \(t \in \mathbb{R}\).

Let \(\mathcal{P}_N := \mathcal{P} \setminus \mathcal{P}_U\) denote the set of sources which have at least one permutation class \(z \in [\mathcal{X}^d]\) where \(c_z > 1\), i.e., sources for which at least one \(m_z\) is not uniquely achieved by \(y \in z\). When \(P \in \mathcal{P}_N\), a more explicit characterization of the distribution of \(Z\) seems very elusive, and in particular, is not Gaussian. To see why, observe that (Arellano-Valle and Genton 2008, Corollary 5) implies that \(\min_{x \in C_z} Z_z^{(x)}\) has
probability density function (p.d.f.):

\[ f_z(t) = \frac{c_z}{\sigma_z} \varphi \left( \frac{t}{\sigma_z} \right) \cdot \Phi_{c_z - 1} \left( \frac{t \sqrt{m_z}}{\sigma_z^2}, \ldots, \frac{t \sqrt{m_z}}{\sigma_z^2}; \rho x_{c_z - 1} + \sum_{x} \right), \]

where \( \varphi(\cdot) \) is the p.d.f. of a standard Normal random variable, \( \Phi_{k}(\cdot; \Sigma) \) is the cumulative distribution function (c.d.f.) of a zero-mean \( k \)-dimensional multivariate normal distribution with variance-covariance matrix \( \Sigma \), and \( I_k \) is the \( k \)-dimensional identity matrix.

Unfortunately, the above probability densities are not enough to describe the distribution of \( Z \) due to the correlation between the minima in Eq. (6). Furthermore, Theorem 7 does not hold when \( P \in P_N \), because whenever \( c_x > 1 \),

\[ \min_{x \in C_z} (Z_z(x) + Y_z(x)) - \min_{x \in C_z} Y_z(x) \neq \min_{x \in C_z} Z_z(x) \]

with probability 1. As a result, the weak convergence in Eq. (13) is to a version of the distribution of \( Z \), but with some Gaussian perturbation.

Luckily, the bootstrap estimator of \( Z \) when \( P \in P_U \) can be made consistent by choosing a resample size \( n_0 = o(n) \) such that \( n_0 \to \infty \). In this case, we would redefine \( \delta^* := \sqrt{n_0} (\hat{\lambda}_n^* - \hat{\lambda}) \) and \( \Delta^* := \sqrt{n_0} \left( \sum_{z \in [\lambda^d]} |z| \cdot (\hat{M}_z^* - M_z) \right) \), so that

\[ \hat{\psi} = \mathbb{P}(\delta^* \leq t \mid X) = \mathbb{P}(\Delta^* \leq t \mid X) + o_P(1) \]

\[ = \psi \left( \left( \sqrt{n_0}(\hat{p}_x - p_x)_{z \in [\lambda^d], x \in C_z} \right) \right) + o_P(1) \]

\[ \overset{p}{\to} \psi(0), \]

because \( \sqrt{n_0}(\hat{p}_x - p_x) \overset{p}{\to} 0 \).

Based on simulations, we have found that it is usually more accurate to use a full size-\( n \) resample for the purposes of correcting the bias of \( \hat{\lambda}_n \), even at moderate sample sizes and when \( P \in P_N \). Additionally, Monte Carlo bootstrap estimates of \( V(\hat{\lambda}) \) tend to be more accurate than the asymptotic formula given in Corollary 3. With very large samples, users may wish to try using a size \( n_0 := 2\sqrt{n} \) resample. Users might also explore ad hoc methods for combined estimators of \( \min_{y \in [\lambda]} p_y \) when there is strong reason to believe that \( P \in P_N \). In what follows in this manuscript, bias and variance of \( \hat{\lambda} \) are approximated from a Monte Carlo estimate of the bootstrap distribution

\[ \sqrt{n}(\hat{\lambda}_n^* - \hat{\lambda}) \mid X \overset{d}{\approx} \sqrt{n}(\hat{\lambda} - \lambda). \]

An explicit Berry-Esseen type bound on the error using the limiting normal distribution to approximate the sampling distribution of \( \hat{\lambda}_n \) remains elusive. The largest difficulty in estimating \( \hat{\lambda} \) is controlling negative bias, especially when \( P \in P_N \). Based on our empirical studies, the effect of sample size on bias depends on \( |\lambda^d| \) and the dimension \( d \) in a highly complicated way, and we cannot suggest a heuristic guideline for appropriate sample sizes based on \( |\lambda^d| \) alone. Therefore we recommend that practitioners simulate data from several test sources in \( \mathcal{E} \) to gauge a worst-case sampling distribution for a given sample size.
Fig. 1 Monte Carlo estimates of bias and standard deviation of \( \hat{\lambda} \) at various sample sizes and in various dimensions, with \( \Omega = \{0, 1\}^d \). We repeatedly simulated data of the given sample size from test source \( T \), with \( T(\{0\}^d) = T(\{1\}^d) = 0 \) and \( T(x) = \frac{1}{2^d-2} \) for each \( x \in \Omega \setminus \{0\}^d, \{1\}^d \). In this setting, \( \lambda(T) = 1 \).

We computed bias-corrected estimates and found the empirical bias and standard deviation of \( \hat{\lambda} \). In our experience, exchangeable weight estimation has the largest bias and standard deviation for a probabilistic source \( T \) over \( \mathcal{X}^d \) that assigns uniform mass everywhere—except on singletons (i.e., \( T(a, \ldots, a) = 0 \), for each \( a \in \mathcal{X} \)). In particular, \( \lambda_{e}(T) = 1 \).

To determine an appropriate sample size for estimating \( \lambda \), we identify the minimum sample size that achieves acceptably small bias and variance of \( \hat{\lambda} \) under the most pathological conditions, which should also achieve acceptably small bias and variance of \( \hat{\lambda}(P) \) for the true source \( P \). Therefore, we suggest the following heuristic: first, select several candidate sample sizes \( n_1, n_2, \ldots \). For each candidate sample size, repeatedly simulate \( n_i \) outcomes from the highly pathological test source described in the previous paragraph to get an empirical estimate of standard error and bias of \( \hat{\lambda} \). After selecting a sample size for which standard deviation and bias of \( \hat{\lambda}(T) \) appear acceptably small, collect samples of this size for every source in a coarse grid over \( \mathcal{P}(\mathcal{X}^d) \). This allows one to reliably obtain acceptable estimates of \( \lambda \) on any source \( P \in \mathcal{P} \).

In Fig. 1, we show the result of Monte Carlo estimates of bias and standard deviation of \( \hat{\lambda} \) at several sample sizes for the pathological test source with sample space \( \Omega = \{0, 1\}^d \) for \( d = 3, 4, \ldots, 10 \). As depicted in the figure, the standard deviation of \( \hat{\lambda} \) does not exceed 0.1, and the bias nears \(-0.2 \) or smaller, when \( n \gtrapprox 10 \cdot 2^d \). As expected, standard deviation of \( \hat{\lambda} \) is very small when \( n \gg 2^d \), as each permutation-equivalence class is likely to contain an unseen outcome, that is, \( \hat{\lambda} \approx 0 \) with high probability. A conservative guideline for an appropriate sample size might be \( n \geq 100 \cdot 2^d \) when the marginal sample spaces are binary, although bias and standard deviation appear to depend on the sample space in a complicated way that exceeds our current understanding.
When a practitioner is specifically searching for a source that has small exchangeable weight, Theorems 4 and 5 are useful for reducing the sample space. One can reduce the sample space by marginalizing or lumping the source in a number of ways so that the resulting sample spaces are small relative to sample size, and estimating $\lambda$ for each of the lumped and marginalized sources. The combinatorial number of ways to marginalize and lump the original source can be quite large, but some choices of lumpings or marginalizations may be more natural in context. The exchangeable weight of each simplified source is an upper bound, in particular, the minimum exchangeable weight of these sources. If this upper bound is small, one can conclude that the exchangeable weight of the original source, which was unfeasible to estimate, must also be small.

5 Proof of concept: DNA methylation

When a DNA sequence contains a cytosine residue (C) followed by a guanine residue (G) in the 5′-to-3′ sense, this dimer is referred to as a CpG. The cytosine in a CpG may or may not have methyl group bonded to it in the 5′ position of its pyrimidine ring. This methylation is regulated by reversible enzymatic processes and is known to modulate gene expression; increased methylation in gene promoters is associated with transcriptional silencing (Jones et al. 1998), and specific DNA methylation patterns have been linked to human disease (Robertson 2005). In particular, certain aberrant methylation patterns are a hallmark of some cancers (Jones 2012), and, as such, considerable effort has been expended to determine regions of DNA that have differential methylation under different cellular conditions.

One popular modern assay to assess DNA methylation is Whole-Genome Bisulfite Sequencing (WGBS) (Lister et al. 2008), a procedure in which unmethylated cytosines are chemically transformed into thymine (T) through treatment with a bisulfite catalyst. When bisulfite-treated DNA is then sequenced by high-throughput shotgun technology, methylated CpGs can be distinguished from unmethylated ones by the observation of a “CG” dimer versus a “TG” dimer, as depicted in Fig. 2.

When attempting to describe DNA methylation, it is routine to use a sliding window approach (Hansen et al. 2012; Akalin et al. 2012) wherein all observations of methylated and unmethylated CpGs are counted in the window, typically 1 Kb in length, to summarize local methylation. It is common to compare methylation between two different biological samples using, e.g., Fisher’s exact test (Akalin et al. 2012), concluding that a window is differentially methylated if the null hypothesis of equal distribution can be rejected. This approach assumes that in a single window the methylation status of each CpG contributes identically in its biological effect. To fix ideas, if we denote an unmethylated CpG as a ‘0’ and a methylated CpG as a ‘1’ and consider ten successive CpGs in different tissues, with the first tissue always producing the configuration ‘1111000000’ and the second always producing the configuration ‘0000111111’, the standard approach would be unable to identify this locus as differentially methylated. So current approaches for differential methylation implicitly assume that binary sequences representing methylation inside each window have an exchangeable distribution.
Fig. 2 Diagram of a typical WGBS experiment. The blue rectangle represents a segment of ssDNA, with the location of CpGs on that strand marked by black vertical bars. Black horizontal line segments represent reads mapped to a reference strand, and open and closed circles represent the partially-observed joint methylation status of several CpGs.

To evaluate this assumption, we examined WGBS data from 121 experimental replicates representing 77 unique biological samples, publicly available from ENCODE (Bernstein et al. 2012; Davis et al. 2017). These replicates include clinical tissue samples, cell lines, and primary cells. We selected replicates using single-end reads which were not flagged by ENCODE as having low coverage or insufficient read length. The list of the sample identifiers (ENCODE_IDs.xlsx) and processed datasets can be found on Figshare (Pearson and Lladser 2021).

Each replicate is associated with a BAM file generated by mapping reads to GRCh38 using Bismark (Krueger and Andrews 2011). We used the MethPipe methylation software suite (Song et al. 2013) to convert BAM files into MethPipe format and generated epiread files, an efficient format reporting the genomic index and methylation status of each CpG contained in a read.

To investigate the exchangeability of local DNA methylation we focused on sets of 3 successive CpGs, which we call “triplets.” Due to Theorem 4, the full joint distribution of methylation in a larger run of successive CpGs must be highly unexchangeable if any triplet within it has highly unexchangeable methylation. However, nearby runs of successive CpGs might have highly exchangeable methylation if they do not contain the unexchangeable triplet.

Within each biological replicate, we used the epireads file generated to extract data from “well-covered” triplets. After discarding reads which reported a CpG with ambiguous methylation status, we considered a triplet well-covered within a replicate only if all three CpGs are jointly covered by at least 100 reads from that sequencing run. Across all 121 replicates, we found 650,152 well-covered triplets, representing 75,212 unique loci in total. The sample size of well-covered triplets is summarized in Fig. 3.

For each autosome, within each replicate, we estimated the exchangeable weight of each well-covered triplet and corrected for estimator bias using a sample mean of \( N = 1000 \) full bootstrap resamples. That is, because \( \hat{\lambda} - \lambda \approx (\hat{\lambda}^* - \lambda) \) for each triplet, we take the Monte Carlo average \( \hat{E}(\hat{\lambda}^*) \) associated with \( N = 1000 \) resamples and subtract \( \hat{E}(\hat{\lambda}^* - \hat{\lambda}) \) from \( \hat{\lambda} \) to adjust for bias of \( \hat{\lambda} \). Although each estimate of a
triplet’s exchangeable weight $\hat{\lambda}$ lies in $[0, 1]$, the bias-adjusted estimate $\hat{\lambda} - (\hat{E}(\hat{\lambda}^*) - \hat{\lambda})$ may be larger than 1 or smaller than 0. Therefore we truncate these estimates to $[0, 1]$. Available on the Figshare repository (Pearson and Lladser 2021) are tab-separated value (tsv format) files containing processed triplets corresponding to each BAM file ID. Each row corresponds to a well-covered triplet, with columns corresponding to (1) chromosome number, (2) index of the first CpG in the triplet on the chromosome, (3) position of the first CpG in the triplet on the chromosome, (4) position of the nearest transcription start site, (5) distance between the triplet centroid and nearest transcription start site, (6) an estimate of the total variation distance to the class of exchangeable distributions, (7) an estimate of the exchangeable weight of the triplet (bias-corrected), (8) a bootstrap estimate of the standard deviation of $\hat{\lambda}$, (9–16) the counts of each of the 8 possible triplet configurations (ordered lexicographically, i.e. ‘000’, ‘001’, etc.), and (17–24) an estimate of the largest exchangeable component.

Estimates of triplet exchangeable weight are depicted in Fig. 4. As seen in the figure, in some chromosomal regions, particularly e.g. on chromosomes 6 and 13, there are triplets whose exchangeable weights are very small. In fact, some appear completely unexchangeable.

As seen in Table 1, triplet exchangeability does not appear strongly correlated with the genomic distance to the nearest promoter. Further, as seen in Fig. 5, within each chromosome, and within each dataset, the correlation between triplet exchangeability and distance from a promoter is usually small. There is a noticeable trend, however, that triplet exchangeable weight is more likely to be negatively correlated with distance from a promoter. Indeed, both a two-sided Wald test and a Spearman rank-order test of the null hypothesis that TSS proximity and estimated triplet exchangeable weight are uncorrelated give very small p-values ($p \ll 10^{-10}$). That is, despite the small
Fig. 4 Estimated exchangeable weights of well-covered triplets by chromosome. Dashed red lines denote the mean, and green plots the histograms associated with these weights. A high-resolution version of this figure can be found in the Supplementary Material.
Table 1  Correlation per chromosome between estimated exchangeable weight of well-covered triplets, and distance between their center and the nearest transcription start site (TSS)

| Chromosome | Correlation | Chromosome | Correlation |
|------------|-------------|------------|-------------|
| 1          | $-1.234 \times 10^{-1}$ | 12         | $-6.813 \times 10^{-2}$ |
| 2          | $+4.367 \times 10^{-2}$  | 13         | $-6.909 \times 10^{-2}$ |
| 3          | $+4.614 \times 10^{-2}$  | 14         | $-1.451 \times 10^{-2}$ |
| 4          | $-7.058 \times 10^{-2}$  | 15         | $-2.356 \times 10^{-1}$ |
| 5          | $-4.815 \times 10^{-1}$  | 16         | $-5.220 \times 10^{-2}$ |
| 6          | $-9.273 \times 10^{-2}$  | 17         | $-3.919 \times 10^{-1}$ |
| 7          | $-3.884 \times 10^{-1}$  | 18         | $-1.821 \times 10^{-1}$ |
| 8          | $-3.355 \times 10^{-1}$  | 19         | $-1.848 \times 10^{-1}$ |
| 9          | $-5.175 \times 10^{-2}$  | 20         | $+8.766 \times 10^{-2}$ |
| 10         | $-1.112 \times 10^{-1}$  | 21         | $-2.771 \times 10^{-2}$ |
| 11         | $-2.247 \times 10^{-1}$  | 22         | $+8.320 \times 10^{-3}$ |

Fig. 5  Top, plot of estimated exchangeable weight for each well-covered triplet versus its distance from the nearest TSS. Bottom, correlation between distance to the nearest TSS and exchangeable weight of each triplet, per chromosome (left) and per dataset (right).

magnitude of the effect, we can detect that triplets close to promoters tend to have more-exchangeable methylation configurations.

To confirm that these highly unexchangeable loci are not due to uncertainty in estimation of the exchangeable weight, we simulated data from the uniform distribution
We generated synthetic samples from the uniform distribution on \{(0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0)\} (red). The true distribution of these synthetic data has $\lambda = 1$, and is the worst case for negative bias in $\hat{\lambda}$. The sample sizes of the synthetic data are matched to those seen in the WGBS data. The averages of estimated exchangeable weights from real and synthetic data are given by the dashed green and red lines, respectively.

Over binary triplets except ‘000’ or ‘111’. Based on empirical study, this source is the worst case for estimating the exchangeable weight in terms of bias and standard error. Nevertheless, as seen in Fig. 6, the empirical distribution of estimated exchangeable weights of all triplets gives much greater probability mass near 0 than the corresponding sampling distribution of the synthetic data. That is, uncertainty from statistical estimation does not account for the apparent phenomenon of highly unexchangeable loci.

In general, it is impossible to disentangle data corruption which is caused by, e.g., sequencing errors or incomplete enzymatic conversion of unmethylated cytosines, from biological processes discriminating specific configurations of methylation. Under the assumption that contamination of the former kind is small, i.e. that we have a truly accurate picture of how methylation is configured in cells, we would expect triplets to have exchangeable weights close to one if overall methylation levels govern biological function. This might mean that in some cell types methylation far away from promoters (and likely far from CpG islands) is “locked in,” and specific patterns of methylation rather than overall methylation levels modulate biological function.

Identifying the biological reason for highly unexchangeable loci remains an open question, which may not have a universal answer. We conclude that there are some loci which are far from exchangeable—that is, some configurations of methylation are discriminated at these triplets. The identification of these loci opens opportunities for more high-resolution understanding of methylation patterns. In particular, these loci represent regions where very specific configurations of methylation may regulate function.
Supplementary Information  The online version contains supplementary material available at https://doi.org/10.1007/s00285-023-01961-1.

Acknowledgements  We thank a reviewer for their thorough reading and constructive remarks about our manuscript.

Funding  This work was partially supported by National Science Foundation Graduate Research Fellowship Program Grant No. 2016198773 (Pearson), and National Science Foundation IGERT Grant No. 1144807.

References

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13(10):R87

Arellano-Valle RB, Genton MG (2008) On the exact distribution of the maximum of absolutely continuous dependent random variables. Stat Probab Lett 78(1):27–35

Bernstein B, Birney E, Dunham I, Green E, Gunter C, Snyder M, ENCODE Project Consortium, Hubbard T (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74

Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. Ann Stat 9(6):1196–1217

Chestnut S, Lladser ME (2010) Occupancy distributions via Doeblin’s ergodicity coefficient. In: Proceedings of discrete mathematics and theoretical computer science, vol AM, pp 79–92

Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322(5909):1845–1848

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Stratton JS, Julanki O, Tanaka FY, Cherry JM (2017) The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 46(D1):D794–D801

De Finetti B (1937) La prévision: ses lois logiques, ses sources subjectives. Annales de l’institut Henri Poincaré 7(1):1–68

Diaconis P (1977) Finite forms of de Finetti’s theorem on exchangeability. Synthese 36(2):271–281

Diaconis P, Freedman D (1980) Finite exchangeable sequences. Ann Probab 8(4):745–764

Gnedin AV (1996) A class of exchangeable sequences. Stat Probab Lett 28(2):159–164

Good PI (2002) Extensions of the concept of exchangeability and their applications. J Mod Appl Stat Methods 1(2):34

Hall P, Härdle W, Simar L (1993) On the inconsistency of bootstrap distribution estimators. CORE Discussion Papers RP 1062, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE). https://EconPapers.repec.org/RePEc:cor:louvrp:1062

Hampton J, Lladser ME (2012) Estimation of distribution overlap of urn models. PLoS ONE 7(11):e42368

Hansen KD, Langmead B, Irizarry RA (2012) BSsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol 13(10):R83

Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35(1):73–101

Huber PJ (1965) A robust version of the probability ratio test. Ann Math Stat 36(6):1753–1758

Ilie L, Fazayeli F, Ilie S (2010) HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics 27(3):295–302

Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13(1):484–492

Jones PL, Veenstra GC, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J, Wolfe AP (1998) Methylated dna and mecp2 recruit histone deacetylase to repress transcription. Nat Genet 19(2):187

Kingman JFC (1982) On the genealogy of large populations. J Appl Probab 19(A):27–43

Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27(11):1571–1572

Lindvall T (1992) Lectures on the coupling method. Wiley series in probability and statistics—applied probability and statistics section. Wiley

Lister R, O’Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in arabidopsis. Cell 133(3):523–536

Springer
Lladser ME, Azofeifa JG, Allen MA, Dowell RD (2017) RNA Pol II transcription model and interpretation of GRO-seq data. J Math Biol 74(1-2):77–97
Lladser ME, Chestnut S (2013) Approximation of sojourn-times via maximal couplings: motif frequency distributions. J Math Biol 69
Lladser ME, Goeuet R, Reeder J (2011) Extrapolation of urn models via poissonization: accurate measurements of the microbial unknown. PLoS One 6(6)
Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc Natl Acad Sci 110(49):19872–19877. https://doi.org/10.1073/pnas.1319590110
Medvedev P, Scott E, Kakaradov B, Pevzner P (2011) Error correction of high-throughput sequencing datasets with non-uniform coverage. Bioinformatics 27(13):i137–i141
National Center for Biotechnology Information: Contamination in Sequence Databases. https://www.ncbi.nlm.nih.gov/tools/vecscreen/contam/. Accessed: 01-2020
Newcomb S (1886) A generalized theory of the combination of observations so as to obtain the best result. Am J Math 8(4):343–366
Park PJ (2009) Chip-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10(10):669
Pearson A, Lladser ME (2020) Hidden independence in unstructured probabilistic models. In: 31st International conference on probabilistic, combinatorial and asymptotic methods for the analysis of algorithms (AofA 2020), Leibniz International Proceedings in Informatics (LIPIcs), vol 159, pp 23:1–23:13. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany. https://doi.org/10.4230/LIPIcs.AofA.2020.23. https://drops.dagstuhl.de/opus/volltexte/2020/12053
Pearson A, Lladser ME (2021) Post-processed DNA methylation data. https://doi.org/10.6084/m9.figshare.16983499.v1
Posfai J, Roberts RJ (1992) Finding errors in DNA sequences. Proc Natl Acad Sci 89(10):4698–4702. https://doi.org/10.1073/pnas.89.10.4698
Punzo A, McNicholas PD (2016) Parsimonious mixtures of multivariate contaminated normal distributions. Biom J 58(6):1506–1537
Robertson KD (2005) DNA methylation and human disease. Nat Rev Genet 6(8):597
Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12
Schmitt MW, Fox EJ, Salk JJ (2014) Risks of double-counting in deep sequencing. Proc Natl Acad Sci 111(16):E1560–E1560
Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS ONE 8(12):e81148
Stinson LF, Keelan JA, Payne MS (2019) Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. Lett Appl Microbiol 68
Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9(6):465
Tukey JW (1960) A survey of sampling from contaminated distributions. Contrib Probab Stat (in: Olkin I et al., eds) pp 448–485
van der Vaart AW (1998) Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press
Wang Z, Gerstein M, Synder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.