A generative nonparametric Bayesian model for whole genomes

Alan N. Amin*
Program in Systems, Synthetic and Quantitative Biology
Harvard Medical School
Boston, MA 02115
alanamin@fas.harvard.edu

Eli N. Weinstein*
Program in Biophysics
Harvard University
Cambridge, MA 02138
eweinstein@g.harvard.edu

Debora S. Marks
Department of Systems Biology
Harvard Medical School
Boston, MA 02115
debbie@hms.harvard.edu

Abstract

Generative probabilistic modeling of biological sequences has widespread existing and potential use across biology and biomedicine, particularly given advances in high-throughput sequencing, synthesis and editing. However, we still lack methods with nucleotide resolution that are tractable at the scale of whole genomes and that can achieve high predictive accuracy either in theory or practice. In this article we propose a new generative sequence model, the Bayesian embedded autoregressive (BEAR) model, which uses a parametric autoregressive model to specify a conjugate prior over a nonparametric Bayesian Markov model. We explore, theoretically and empirically, applications of BEAR models to a variety of statistical problems including density estimation, robust parameter estimation, goodness-of-fit tests, and two-sample tests. We prove rigorous asymptotic consistency results including nonparametric posterior concentration rates. We scale inference in BEAR models to datasets containing tens of billions of nucleotides. On genomic, transcriptomic, and metagenomic sequence data we show that BEAR models provide large increases in predictive performance as compared to parametric autoregressive models, among other results. BEAR models offer a flexible and scalable framework, with theoretical guarantees, for building and critiquing generative models at the whole genome scale.

1 Introduction

Measuring and making DNA is central to modern biology and biomedicine. Generative probabilistic modeling offers a framework for learning from sequencing data and forming experimentally testable predictions of unobserved or future sequences [18, 28, 54]. Existing approaches to genome modeling typically preprocess the data to build a matrix of genetic variants such as single nucleotide polymorphisms [23, 49]. However, most modes of sequence variation are more complex. Structural variation occurs widely within individuals (e.g. in cancer), between individuals (e.g. in domesticated plant populations) and between species (e.g. in the human microbiome), and methods for detecting and classifying structural

*These authors contributed equally.

Preprint. Under review.

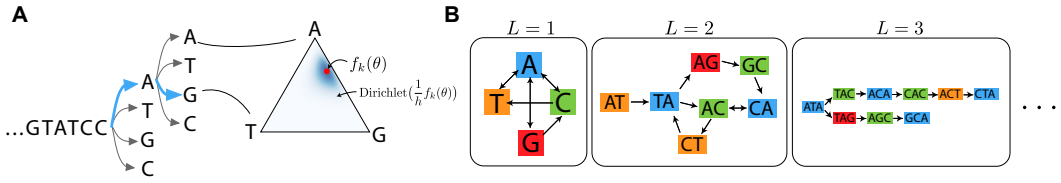

Figure 1: **Overview of the BEAR model.** (A) BEAR models employ a Dirichlet prior on Markov transition probabilities that is centered at the prediction of an AR model. (B) De Bruijn graphs showing BEAR transitions with non-zero probability under an example data-generating distribution. As the lag L increases, the model has higher resolution.

variants are heuristic and designed only for predefined types of sequence variation such as repeats [12, 41, 45, 58, 67]. Ideally, we would be able to directly model genome sequencing data and/or assembled genome sequences. However, building generative models that work with raw nucleotides, not matrices of alleles, raises the extreme statistical challenges of having enough *flexiblility* to account for genomic complexity, *interpretability* to reach scientific conclusions, and *scalability* to train on billions of nucleotides. Given the relevance of genetic analysis to human health, models should also possess strong *theoretical guarantees*.

Autoregressive (AR) models are a natural starting point for generative genome modeling, since they (1) have been successfully applied to protein sequences, as well as many other types of non-biological sequential data, (2) can be designed to have interpretable parameters, and (3) can be scaled to big datasets with very long sequences [56, 62]. However, since AR models are parametric models, they will in general suffer from misspecification; as we show empirically in Section 6, for genomic datasets misspecification can be a serious practical limitation not only for simple AR models but even for deep neural networks.

As an alternative strategy for building generative probabilistic models at the genome scale, we propose in Section 2 the nonparametric "Bayesian embedded autoregressive" (BEAR) model. BEAR models are Bayesian Markov models, with a prior on the lag and conjugate Dirichlet priors on the transition probabilities. The hyperparameters of the Dirichlet prior are controlled by an "embedded" AR model with parameters θ and an overall concentration hyperparameter h, both of which can be optimized via empirical Bayes. In Section 3 we show that BEAR models can capture arbitrary data-generating distributions, and establish asymptotic consistency guarantees and convergence rates for nonparametric density estimation. In Section 4, we show that the optimal h provides a diagnostic for whether or not the embedded AR model is misspecified and if so by how much, alerting the practitioner when the parameter estimates θ are untrustworthy. Besides estimation problems, BEAR models can also be used to construct goodness-of-fit tests and two-sample tests, thanks to their analytic marginal likelihoods, and we prove consistency results for these tests in Section 5. Finally we apply BEAR models at large scale, to genomic datasets with tens of billions of nucleotides, including whole genome, whole transcriptome, and metagenomic sequencing data; where comparable, BEAR models show greatly improved performance over AR models (Section 6).

Crucial to our the theoretical and empirical analysis is the statistical setting: we assume that the data $X_1, \ldots, X_N$ consists of finite but possibly variable length strings (with small alphabets) drawn i.i.d. from some underlying distribution p^*, and study the behavior of estimators and tests as $N \to \infty$. This setup differs from common theoretical analyses of sequence models outside of biology, which typically consider the limit as the length of an individual sequence goes to infinity [24]. In biology, however, we observe finite sequences recorded from many individual species, organisms, cells, molecules, etc. and want to generalize to unseen sequences, making $N \to \infty$ the appropriate large data limit.

2 Bayesian embedded autoregressive models

We first briefly review autoregressive (AR) models as applied to sequences of discrete characters. Let $f(\theta)$ denote an autoregressive function with parameter θ and let L denote

the lag of the autoregressive model; then the AR model generates data as

$$X_i|X_{i-L:i-1} \sim \text{Categorical}(f_{X_{i-L:i-1}}(\theta)), \tag{1}$$

where i indexes position in the sequence X and $X_{i-L:i-1}$ consists of the previous L letters in the sequence. Since sequence length as well as nucleotide or amino acid content is relevant to biological applications, we use a start symbol $\emptyset$ at the beginning and a stop symbol \$ at the end of each sequence; letters X_i are sampled sequentially starting from the start symbol and continuing until a stop symbol is drawn.

We propose the Bayesian embedded autoregressive (BEAR) model, a Bayesian Markov model that embeds an AR model into its prior. The BEAR model takes the form,

$$\begin{aligned} L \sim \pi(l) \qquad v_k \sim \text{Dirichlet}\big(\frac{1}{h} f_k(\theta)\big) \text{ for all } k \\ X_i|X_{i-L:i-1} \sim \text{Categorical}(v_{X_{i-L:i-1}}) \end{aligned} \tag{2}$$

where $\pi(l)$ is a prior on the lag with support up to infinity, $h > 0$ is a concentration hyperparameter, and k is a length L kmer. The BEAR model has three key properties (Fig. 1). First, the unrestricted transition parameter v and lag L allow the model to capture exact conditional distributions of p^* to arbitrarily high order: $p^*(X_i|X_{i-1})$ at $L = 1$, then $p^*(X_i|X_{i-2}, X_{i-1})$ at $L = 2$, etc.. This property allows the BEAR model to be used for nonparametric density estimation (Section 3). Second, in the limit where $h \to 0$, the BEAR model reduces to the embedded AR model (Eqn. 1). The optimal h provides a measurement of the amount of misspecification in the AR model (Section 4). Third, the choice of the conjugate Dirichlet prior allows the conditional marginals $p((X_n)_{n=1}^N|L, h, \theta)$ to be computed analytically, and (since L is one-dimensional) the total marginal likelihood $p((X_n)_{n=1}^N|h, \theta)$ to be estimated tractably. This allows BEAR models to be used for hypothesis testing (Section 5).

There are a variety of ways of performing inference in BEAR models, but for most applications we will focus on empirical Bayes methods that optimize point estimates of L, h and θ. Let $\#(k, b)$ denote the number of times the length L kmer k is seen followed by the letter or stop symbol b in the dataset $(X_n)_{n=1}^N$. Using a high-performance kmer counter optimized for nucleotide data, KMC, we can compute the count matrix $\#(\cdot, \cdot)$ for all observed kmers k in terabyte-scale datasets, even when the matrix does not fit in main memory (Section J.2) [35]. To optimize h and θ, we take advantage of the fact that the log conditional marginal likelihood can be written as a sum over observed kmers,

$$\log p((X_n)_{n=1}^N|L, h, \theta) = \sum_{k:\#k>0} \log \left[\frac{\Gamma(\sum_b \frac{1}{h} f_{kb}(\theta))}{\prod_b \Gamma(\frac{1}{h} f_{kb}(\theta))} \frac{\prod_b \Gamma(\frac{1}{h} f_{kb}(\theta) + \#(k, b))}{\Gamma(\sum_b \frac{1}{h} f_{kb}(\theta) + \#(k, b))} \right] \tag{3}$$

This decomposition lets us construct unbiased stochastic estimates of the gradient with respect to h and θ by subsampling rows of the count matrix (Section J.1). Empirical Bayes in the BEAR model therefore costs little extra time as compared to standard stochastic gradient-based optimization of the original AR model. Code is available at `https://github.com/debbiemarkslab/BEAR`.

Toy example We next briefly illustrate the properties and advantages of the BEAR model in simulation. We generated samples from an AR model in which $f_k(\theta)$ depends on k linearly as a function of both individual positions and pairwise interactions between positions, with the strength of the pairwise interaction weighted by a parameter β^* (Section I.1). We first fit (using maximum likelihood) a linear AR model that lacks pairwise terms and is thus misspecified when $\beta^* > 0$. When the AR model is misspecified, it does not asymptotically approach the true data-generating distribution p^* (Fig. 2A, gray). We next computed the posterior of a vanilla BEAR model without the embedded AR in its prior, instead using the Jeffreys prior $V_k \sim_{iid} \text{Dirichlet}(1/2, \ldots, 1/2)$. The vanilla BEAR model asymptotically approaches the true data generating distribution since it is a nonparametric model; however, it underperforms the AR model in the low data regime (Fig. 2A, black). Finally, we applied our empirical Bayes inference procedure to a BEAR model with the misspecified linear AR model embedded. The BEAR model performs just as well as the AR model in the low data regime, just as well as the vanilla model in the high data regime, and better than both at

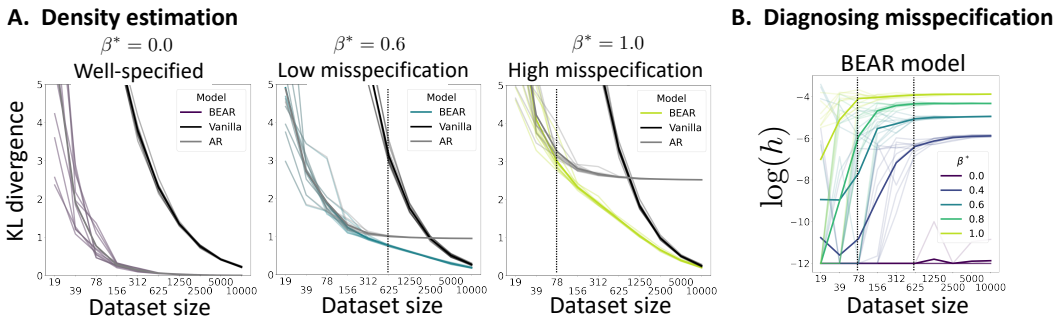

Figure 2: **BEAR models detect and avoid misspecification without sacrificing small dataset performance.** (A) Estimated KL divergence between simulated data-generating distribution p^* and model posterior predictive distribution, as a function of dataset size N. Five independent simulations were run; thin lines correspond to individual simulations, thick lines to the average across simulations. (B) The h misspecification diagnostic as a function of dataset size, for varying β^*. Dataset sizes at which h is close to convergence for $\beta^* = 0.6$ (right) and $\beta^* = 1.0$ (left) are marked with vertical lines.

intermediate values (Fig. 2A, orange). When the AR model is well-specified, the empirical Bayes estimates of AR parameters θ provided by the BEAR model match the maximum likelihood estimates of the AR model nearly exactly (Fig. S7). When the AR model is misspecified, however, the BEAR model provides a warning: the empirical Bayes estimate of h converges to a non-zero value, rather than zero (Fig. 2B). This warning emerges early: h converges well before the vanilla BEAR model starts outperforming the AR model.

Related Work The key idea behind BEAR models is to nonparametrically perturb a parametric model, following a similar strategy to the Polya tree method proposed by Berger and Guglielmi [6]. We too use independent Dirichlet priors centered at the parametric model's predictions and use conjugacy to construct goodness-of-fit tests. BEAR models extend these ideas from one-dimensional continuous data to sequences of discrete characters.

BEAR models are closely linked to key non-generative genome analysis methods. Assembly algorithms and variant callers often analyze paths in the de Bruijn graph of a sequence dataset; in the limit $h \to \infty$, samples from the posterior predictive distribution of the BEAR model, conditional on L, correspond to paths through the L-mer de Bruijn graph of the data [11, 30]. Comparisons between genomes and other sequences are often made on the basis of kmer counts; our two-sample test provides a generative perspective on this idea [3, 16, 67].

BEAR models are also connected to ideas in natural language processing, where kmers are referred to as ngrams. Under the vanilla BEAR model, the mean of the posterior predictive distribution conditional on L corresponds to an ngram additive smoothing model [9]. Comparisons between datasets using their ngram counts are also common in model evaluation metrics such as the BLEU score [47].

3 Density estimation

The density estimation problem is that of estimating p^* given data $(X_n)_{n=1}^N$. Density estimation is particularly crucial for biological sequence analysis due to its connections to fitness estimation [28]. State-of-the-art mutation effect prediction methods and clinical variant interpretation methods rely on density estimates of evolutionary sequence data, as do emerging techniques for protein design [19, 51, 54, 56]. Despite these applications, density estimation methods for biological sequences lack theoretical guarantees on their accuracy and are limited in their scale, being restricted to relatively short sequences [68]. Here, we show that the posterior distribution of the BEAR model is consistent and will concentrate on p^* as $N \to \infty$, regardless of what p^* actually is, so long as p^* generates finite length sequences almost surely (a.s.).

We first study the expressiveness of BEAR models. Let $\mathcal{M}_L$ be the set of Markov models p_v with transition probabilities v and lag L that generate finite length strings a.s.. Note

that $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots$. Define the union $\mathcal{M} = \cup_{L=1}^{\infty} \mathcal{M}_L$. We can compare $\mathcal{M}$ to the set of distributions over finite strings S, of which p^* is a member. In Section D we prove that,

Summary of Propositions 1-4 *Not all possible distributions over S are in $\mathcal{M}$. However, $\mathcal{M}$ is dense on the space of probability distributions over S with the total variation metric.*

The implication of this result is that although BEAR models cannot exactly match arbitrary data-generating distributions, they can approximate p^* arbitrarily well as L increases. This makes asymptotic consistency possible.

We now show that the posterior of the BEAR will in fact asymptotically concentrate on the true p^*, i.e. it is consistent. For tractability, we assume in this section that the prior is fixed (we do not use empirical Bayes). The result relies on the tools for understanding convergence rates of posteriors developed in Ghosal et al. [21]. The most important assumption is that p^* is subexponential, meaning that for some $t > 0$, $E_{p^*} \exp(t|X|) < \infty$ where $|X|$ is the sequence length. Let $\Pi(\cdot|(X_n)_{n=1}^N)$ denote the posterior over sequence distributions. Let $B(p^*, \delta)$ denote a ball of radius δ centered at p^*, using the Hellinger distance.

Summary of Theorem 35 *Given $M > 0$ large enough and $\epsilon \in (0, 1)$ small enough, we have $\Pi(B(p^*, MN^{-\frac{1}{2}\epsilon})|(X_n)_{n=1}^N) \to 1$ in probability.*

A proof is in Section H and simulations in Section I.2. This result states that the posterior distribution of the model converges to a delta function at the true distribution p^* regardless of what p^* is. It also provides a rate of convergence: in a parametric model, the uncertainty would shrink as $N^{-\frac{1}{2}}$, but here the rate is slower, $N^{-\frac{1}{2}\epsilon}$, a price paid for the nonparametric model's expressivity. The proof includes a variety of new theoretical constructions and algorithms that are used to approximate subexponential sequence distributions.

4 Robust parameter estimation

To derive a biological understanding of mutational processes, evolutionary history, functional constraints, etc. from sequence data, researchers must estimate model parameters (not just density). However, parameter estimates cannot in general be trusted when models are misspecified [31]. To reach robust scientific conclusions, therefore, parameter estimates should ideally come with a warning about whether or not the model is misspecified and some measurement of the degree of misspecification. Here, we study in BEAR models the asymptotic behavior of empirical Bayes estimates of the AR parameter θ, as well as the hyperparameter h, showing that h diagnoses misspecification in the embedded AR model.

Our analysis builds off the study of empirical Bayes consistency in Petrone et al. [48], which showed that empirical Bayes will, in general, maximize the prior probability of the true data-generating parameter value. Extending this theory to BEAR models is nontrivial, since in BEAR models the standard Laplace approximation to the marginal likelihood can fail. For theoretical tractability, as in many analyses of similar models, we fix L at some arbitrary and large value [27]. Define $p^{*(L)} = \text{argmin}_{p_v \in \mathcal{M}_L} \text{KL}(p^* \| p_v)$ as the closest model in $\mathcal{M}_L$ to p^*, and define v^* such that $p_{v^*} = p^{*(L)}$ (note $p^{*(L)} \to p^*$ as $L \to \infty$). We say that the AR model is misspecified "at resolution L" if f cannot approximate $p^{*(L)}$, i.e. if there does not exist some sequence of parameter values $\tilde{\theta}_N$ such that $p_{f(\tilde{\theta}_N)} \to p^{*(L)}$; otherwise, the AR model is well-specified at resolution L. Now we can study empirical Bayes estimates of h and θ, denoted h_N and θ_N.

Summary of Propositions 15-20 *Let $(h_N)_{N=1}^{\infty}$ and $(\theta_N)_{N=1}^{\infty}$ be sequences maximizing the BEAR marginal likelihood $p(\{X_n\}_{n=1}^N | L, h, \theta)$ for each N. If the model is well-specified at resolution L, then $h_N N^{1/4-\epsilon} \to 0$ for every $\epsilon > 0$ and $p_{f(\theta_N)} \to p^{*(L)}$ in distribution, with both sequences converging in probability. On the other hand, if the model is misspecified at resolution L, then h_N is eventually bounded below by some positive (non-zero) number a.s..*

Proofs are in Section F and simulations in Section I.1. The implication of this result is that when the AR model is well-specified, h_N converges to zero (at a rate that is a power of the dataset size) and θ_N converges to the parameter value θ^* at which the AR model matches the data (Corollary 16). On the other hand, when the AR model is misspecified, h_N does not converge to zero; heuristically, we find instead that h_N is approximately proportional to

a divergence between $p^{*(L)}$ and the AR model,

$$h_N \propto \sum_{k \in \mathrm{acc}_L(p^*)} \left(\mathrm{KL}(f_k(\theta_N) \| v_k^*) + \log(N) \sum_{b \notin \mathrm{supp}_L(p^*)|_k} f_{k,b}(\theta_N) \right) \tag{4}$$

where $\mathrm{acc}_L(p^*) = \{k \mid p^*(\#k > 0) > 0\}$ is the set of kmers with non-zero probability and $\mathrm{supp}_L(p^*)|_k = \{b \mid p^*(\#(k,b) > 0) > 0\}$ is the set of transitions from k with non-zero probability. In summary: when fitting a BEAR model by empirical Bayes, you get, along with a parameter estimate θ_N, a value h_N which tells you the amount (from zero to infinity) of misspecification in the AR model. If h_N is close to zero, you can trust the estimate θ_N.

5 Hypothesis testing

Goodness-of-fit test Building generative models based on natural sequences that are accurate enough to produce novel functional sequences is a major outstanding challenge. A crucial component of the problem is model evaluation: while relative model performance may be compared on the basis of likelihood, absolute performance – whether or not the model in fact provides an accurate description of the data – is usually addressed solely on the basis of limited numbers of summary statistics, such as average amino acid hydrophobicity or sequence length [54, 56]. Given a dataset $(X_n)_{n=1}^N \sim p^*$ i.i.d., a goodness-of-fit test asks whether or not the data distribution p^* matches a model distribution $\tilde{p}$. It takes into account all possible distributions p^* including those that differ from $\tilde{p}$ in a manner that cannot be captured by finitely many summary statistics. Our test compares the null hypothesis $\mathcal{H}_0 : p^* = \tilde{p}$ to the alternative $\mathcal{H}_1 : p^* \neq \tilde{p}$ using the Bayes factor $\mathrm{BF} = p((X_n)_{n=1}^N | h, \theta) / \tilde{p}(X_{1:n})$, where $p((X_n)_{n=1}^N | h, \theta) = \sum_L p((X_n)_{n=1}^N | L, h, \theta) \pi(L)$ is the marginal likelihood under the BEAR model. Note that practically, the sum over L is straightforward to approximate, and that the test can be computed in time linear in the amount of data.

We now prove consistency. As in comparable theoretical analyses of tests based on Polya trees, for theoretical tractability we truncate the prior, setting $\pi(L) = 0$ for L larger than some arbitrary $\tilde{L}$ but $\pi(L) > 0$ for $L \leq \tilde{L}$ [27]. We treat θ and $h > 0$ as fixed.

Summary of Proposition 21 *If $\tilde{p}$ is at least as close to p^* as $p^{*(L)}$ is, as measured by $\mathrm{KL}(p^* \| \cdot)$, then $\mathrm{BF} \to 0$ in probability as $N \to \infty$. On the other hand, if $p^{*(L)}$ is closer than $\tilde{p}$, then $\mathrm{BF} \to \infty$ in probability.* A proof is in Section G.1 and simulations in Section I.3.

An important practical limitation on nonparametric hypothesis testing is low power: since so many alternative distributions must be considered, the null hypothesis can rarely be rejected. However, Proposition 21 holds for the Bayes factor $\mathrm{BF}(L, h, \theta) = p((X_n)_{n=1}^N | L, h, \theta) / \tilde{p}((X_n)_{n=1}^N)$ with any choice of L, $h > 0$, and θ. Thus in practice to increase power we can maximize the value of $\mathrm{BF}(L, h, \theta)$ as a function of L, h, and/or θ (note that this approach is heuristic, since we have not proven the consistency of the maximized Bayes factor). Berger and Guglielmi [6] provide extensive methodological guidance on using analogous tests constructed with Polya trees. Based on their recommendations, we suggest first choosing θ such that $p_{f(\theta)}$ is as close as possible to $\tilde{p}$, then plotting the Bayes factor as a function of h and/or L to identify the maximum value and confirm that any conclusion is robust to changes in h and/or L. Another challenge in nonparametric hypothesis testing is that it can be difficult to understand how exactly a test reached its conclusion. To identify which sequences provided the most evidence for or against the null hypothesis, we can examine the Bayes factor for each individual sequence conditional on the rest of the dataset, in analogy to the witness function used in kernel-based tests [40, 59].

Two-sample test A two-sample test asks whether or not two datasets $(X_n)_{n=1}^N$ and $(X'_n)_{n=1}^{N'}$ are drawn from the same distribution. Efforts to compare different sequence datasets are widespread in biology: for instance, researchers often wish to determine whether two microbiome samples, taken under different conditions or at different timepoints, are the same up to sampling noise [41]. Two-sample tests can also be used to evaluate generative sequence models that lack tractable likelihoods (for which the goodness-of-fit test proposed above does not apply) such as energy-based models or implicit models like GANs and biophysical simulators [25, 39, 44]. Assume $(X_n)_{n=1}^N \sim p_1$ and $(X'_n)_{n=1}^{N'} \sim p_2$ iid. Our test compares the null hypothesis $\mathcal{H}_0 : p_1 = p_2$ to the alternative $\mathcal{H}_1 : p_1 \neq p_2$ using the Bayes factor

Table 1: **Heldout perplexity.** *Whole genome sequencing data*: YSD1: A Salmonella phage. *A. th.*: *Arabidopsis thaliana*, a plant (datasets represent different individuals). *Single cell RNA sequencing data*: PBMC: peripheral blood mononuclear cells, taken from a healthy donor. HL: Hodgkin's lymphoma tumor cells. GBM: glioblastoma tumor cells. *Metagenomic sequencing data*: HC: non-CD and non-UC controls. CD: Crohn's disease. UC: ulcerative colitis. *Full assembled genomes*: Bact.: Bacteria. *Models* Van.: Vanilla (Jeffreys prior). Lin.: Linear. CNN: convolutional neural network. Ref.: reference genome/transcriptome model.

Dataset	AR Lin.	AR CNN	AR Ref.	BEAR Van.	BEAR Lin.	BEAR CNN	BEAR Ref.
YSD1	3.953	3.873	1.266	1.165	**1.144**	**1.144**	1.145
A. th. 1	3.956	3.947	2.686	1.567	1.432	1.432	**1.411**
A. th. 2	3.953	3.949	1.982	1.650	1.463	1.462	**1.441**
A. th. 3	3.998	3.952	2.340	1.834	1.728	**1.727**	1.733
PBMC	3.991	3.974	2.097	1.402	**1.372**	**1.372**	1.374
HL	3.959	3.930	2.141	1.409	**1.378**	**1.378**	1.379
GBM	4.137	4.137	2.366	1.442	**1.406**	**1.406**	**1.406**
HC	3.966	3.946	-	1.652	1.465	**1.464**	-
CD	3.992	3.985	-	1.760	**1.524**	**1.524**	-
UC	3.989	3.986	-	1.644	**1.481**	**1.481**	-
Bact.	3.831	3.794	-	**3.774**	**3.774**	**3.774**	-

$\mathrm{BF} = p((X_n)_{n=1}^N|h,\theta)p((X'_n)_{n=1}^{N'}|h,\theta)/p((X_n)_{n=1}^N,(X'_n)_{n=1}^{N'}|h,\theta)$. As in the goodness-of-fit case, the test can be computed approximately in time linear in the amount of data, and the same advice on increasing power and identifying important sequences holds here too.

We next prove consistency, again truncating the prior at $\tilde{L}$ and fixing h and θ.

Summary of Proposition 22 *If* $p_1^{(\tilde{L})} = p_2^{(\tilde{L})}$, *then* $\mathrm{BF} \to 0$ *as* $N \to \infty$ *in probability. Otherwise, if* $p_1^{(\tilde{L})} \neq p_2^{(\tilde{L})}$, *then* $\mathrm{BF} \to \infty$ *in probability.* A proof is in Section G.2 and simulations in Section I.3.

6 Results

Predicting sequences We sought to evaluate BEAR models as compared to AR models on the task of predicting real nucleotide (nt) sequences. We considered eleven datasets of four different types: whole genome sequencing read data, single cell RNA sequencing read data (including from patient tumors), metagenomic sequencing read data (including from patient fecal samples) and full bacterial genomes from across the tree of life (Section K). Datasets ranged in total size from $\sim 10^7 - 10^{10}$ nt and in individual sequence length from $\sim 10^2 - 10^6$ nt (Table S1). 25% of data was randomly held out for testing, in the form of entire sequences (reads, genomes, etc., see Table S2); our goal was to evaluate BEAR models as density estimators, so we did not use masking (a common holdout strategy in natural language processing). We considered a linear AR model and a deep convolutional neural network (CNN) AR model with $> 10\times$ more parameters; we also designed a biologically-structured AR model which makes predictions based on a reference genome and a Jukes-Cantor mutation model (Section L.1). We then embedded each AR model to create a corresponding BEAR model. The BEAR models improve over the AR models in nucleotide prediction according to both perplexity (Table 1) and accuracy (Table S3) in all datasets, even when the model lag L is held fixed for comparison (Section L.3).

In 10 out of 11 datasets, BEAR models increase nucleotide prediction accuracy from near chance values of $30 - 35\%$ (in the case of the linear and CNN models) to $78 - 95\%$, bringing genome-scale models into the realm of potential practical use (Table S3). The training time for BEAR models is essentially identical to that of AR models, aside from the time required to build the transition count matrix, which need only be done once before training all models (Fig. S13). Remarkably, the optimal lag L chosen by empirical Bayes is often quite short, less than 20 nt (Table S4). The improvements offered by BEAR models that use

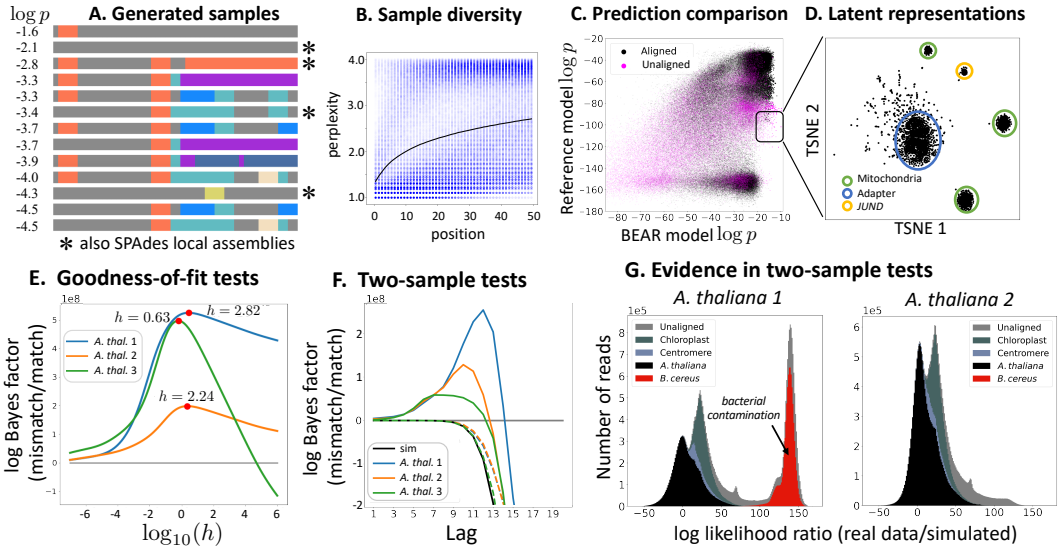

Figure 3: **Generation, visualization and testing.** (A) Sample extrapolations, colored to denote distinct paths through the L-mer de Bruijn graph. (B) Perplexity of the next Markov transition under the BEAR model, for each position of each sampled extrapolation, with the average across samples shown in black (Section M). (C) Log probability of each read in the HL dataset under the BEAR model and a model built from the reference transcriptome. Reads are colored by whether or not they map to the reference. (D) Latent representations of the reads highlighted in C, visualized using tSNE, with clusters annotated as likely coming from mitochondria, the sequencing adapter, or transcripts of the gene *JUND* (Section N). (E) Goodness-of-fit test Bayes factor as a function of hyperparameter h. (F) Two-sample test Bayes factor as a function of lag L. Black line compares simulated data to simulated data; dashed lines compare subsampled real data to subsampled real data; solid lines compare real data to simulated data. (G) Log probability of each read under the real data BEAR model minus the log probability under the simulated data BEAR model (Section O).

an embedded AR model over the vanilla BEAR model are modest for datasets of this size; however, sequencing experiments are often designed to collect enough data for downstream analyses. We found in an example that, if sequencing coverage was 3× instead of 100×, the improvement in prediction accuracy would have been greater than 10 percentage points instead of 0.1 (Section L.4; Fig. S14).

Measuring misspecification When conventional deep neural network methods fail to provide strong predictive performance, popular wisdom often ascribes the failure to too much model flexibility or not enough training data, especially in scientific applications. Examining the h misspecification diagnostic in the BEAR models described above, we see that this is not the case here (Table 2). The large values of h suggest that where the CNN fails it is not because of too much flexibility but because of too little: it suffers from misspecification. Meanwhile, the reference-based model has only two learned parameters, but is less misspecified than the CNN in all but one dataset. This too runs counter to popular wisdom in machine learning, which often assumes that when principled, low-flexibility scientific models outperform deep neural networks it is thanks to their low variance in the small data regime.

Table 2: **Diagnostic h.** Abbreviations as in Table 1.

Dataset	Lin.	CNN	Ref.
YSD1	5.528	5.461	4.183
A. th. 1	2.765	2.756	2.990
A. th. 2	2.643	2.633	2.326
A. th. 3	3.969	3.964	1.598
PBMC	4.167	4.145	3.762
HL	4.050	4.038	3.581
GBM	4.172	4.154	3.238
HC	4.668	4.651	-
CD	3.096	3.094	-
UC	3.843	3.835	-
Bact.	0.010	0.003	-

Generating samples BEAR models are generative and can be used to sample new sequences. We sampled extrapolations from the end of a read sequence recorded in a plant (*A. thaliana*) whole genome sequencing experiment, and compared to an alternative non-probabilistic

extrapolation method that is widely used in biology, local assembly (Fig. 3A; Section M). In this example the assembly algorithm SPAdes returns four possible assemblies, a relatively large number compared to other reads in the dataset (Fig. 3A stars) [5]. Samples from the BEAR model include these four possibilities, but also many more, some with higher probability. The distribution over possible nucleotide choices under the BEAR model is much wider than the number of assemblies would suggest: it has a perplexity of 1.4 per position (on average across samples) at the beginning of the extrapolation, and a perplexity of 2.7 at 50 nucleotides (Fig. 3B). These observations suggest that SPAdes, which does not provide a measurement of uncertainty, may not be capturing the full range of possible sequences.

Visualizing data Methods for learning local representations or features of biological sequences can be powerful tools for visualization and semisupervised learning [7]. One approach to extracting such representations is to learn a generative model $q(X_1, \ldots, X_{L+1})$ of kmers, for instance using a variational autoencoder. While such models are not autoregressive, the small size of the DNA alphabet makes it tractable to estimate the conditional $q(X_{L+1}|X_{1:L})$ by Bayes rule, and this conditional can then be embedded into a BEAR model. We applied this strategy to probabilistic PCA. We visualized in low dimensions the inferred latent representation for a model trained on a single cell RNA sequencing dataset (HL), and were able to assign annotations to clusters, including those containing unmapped reads (Fig. 3CD; Section N). The BEAR model however raises the warning that the model is misspecified ($h = 4.836$), suggesting there may be richer latent structure yet to discover.

Testing hypotheses The question of when and how microbiomes change is widespread, but has in the past relied on summary statistics of sequencing datasets [41]. Schreiber et al. [55] studied changes in patient urine microbiomes before and after kidney transplant, and performed both unbiased metagenomic sequencing and diagnostic quantitative polymerase chain reaction (qPCR) for a specific virus associated with complications (JC polyomavirus). They found evidence of donor-to-recipient viral transmission in 5 cases out of 14. We applied the BEAR two-sample test to patients' metagenomic sequencing data before and after transplantation, using the vanilla Jeffreys prior and integrating over lags, in order to detect changes; the test rejects the null hypothesis in all 5 cases where there was transmission, and accepts the null hypothesis in all but one of the remaining 9 cases (Table S6; Section O.1). These results show, in a small example, that the two-sample test has sufficient power to detect microbiome changes in real data, and can be consistent with more specific tests.

We next applied BEAR hypothesis tests to evaluate generative models. We evaluated the reference-based AR model described above using the BEAR goodness-of-fit test. The test identifies considerable evidence (log Bayes factor $> 10^8$) for misspecification in each *A. thaliana* whole genome sequencing dataset, and this conclusion is robust to a wide range of h values (Fig. 3E; Section O.2). Next, we evaluated a detailed simulation model (ART) that is intended to generate likely reads of a given reference genome [29]. The model lacks tractable likelihoods, so we use a two-sample test. When integrating over all lags, the test accepts the null hypothesis, but if we examine the test results for individual lags L to increase power, we see evidence of differences (Fig. 3F; Section O.2). To understand the source of these differences, we examined the conditional Bayes factor for individual reads, discovering clusters of reads that are poorly explained by the simulation model (Fig. 3G). One group mapped to chloroplasts, an organelle with its own genome that is variable in copy number; reads mapping to centromeres, an area of the plant genome for which the reference genome is considered unreliable, were also poorly explained by the simulation model. In one dataset we found a cluster of outliers that did not map to *A. thaliana* at all, and instead mapped to a common soil bacteria, *Bacillus cereus*, presumably a contaminant in the experiment (Fig. 3G, left). These results illustrate how BEAR hypothesis tests can be used not only for testing but also for detailed model criticism.

7 Discussion

In this article we proposed the nonparametric BEAR model, studied its theoretical properties, and developed algorithms and implementations for terabyte-scale inference. BEAR models substantially outperform standard AR models on a variety of datasets, and come with extensive theoretical guarantees, including for density estimation, misspecification detection,

and hypothesis testing. BEAR models are closely connected to non-probabilistic genome analysis methods, such as de Bruijn graph assembly, but provide an alternative that is uncertainty-aware. Note, however, that BEAR models do not explicitly account for paired-end read information, or other sources of long-distance information; this is an important area for future work. While there has been little previous empirical or theoretical work in the machine learning literature on generative models of full genomic, transcriptomic or metagenomic sequences, we hope BEAR models provide a useful starting point.

Acknowledgments and Disclosure of Funding

We thank Jean Disset for a small scale version of the kmer counting code and Rob Patro for crucial advice on large scale kmer counting. We thank Tessa Green, Chris Sander, and Elizabeth Wood for discussion and advice. We thank Winnie Wang for their illustrations used in the theory section of the supplementary materials. We thank members of the Marks Lab for discussion and ideas. E.N.W. is supported by the Fannie and John Hertz Foundation. D.S.M. is supported by the Chan Zuckerberg Initiative.

References

[1] 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. Cell, 166(2):481–491, July 2016.

[2] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and Others. Tensorflow: A system for large-scale machine learning. In 12th USENIX symposium on operating systems design and implementation (OSDI 16), pages 265–283. usenix.org, 2016.

[3] E. B. Alsop and J. Raymond. Resolving Prokaryotic Taxonomy without rRNA: Longer Oligonucleotide Word Lengths Improve Genome and Metagenome Taxonomic Classification. PLoS ONE, 8(7), 2013.

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. July 2016.

[5] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol., 19 (5):455–477, May 2012.

[6] J. O. Berger and A. Guglielmi. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. Journal of the American Statistical Association, 96(453):174–184, 2001.

[7] S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, and G. M. Church. Low-N protein engineering with data-efficient deep learning. Nat. Methods, 18(4):389–396, 2021.

[8] G. M. Boratyn, C. Camacho, P. S. Cooper, G. Coulouris, A. Fong, N. Ma, T. L. Madden, W. T. Matten, S. D. McGinnis, Y. Merezhuk, Y. Raytselis, E. W. Sayers, T. Tao, J. Ye, and I. Zaretskaya. BLAST: a more efficient report with usability improvements. Nucleic Acids Res., 41(Web Server issue):W29–33, July 2013.

[9] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. Comput. Speech Lang., 13(4):359–394, Oct. 1999.

[10] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11):1422–1423, June 2009.

[11] P. E. C. Compeau, P. A. Pevzner, and G. Tesler. How to apply de bruijn graphs to genome assembly. Nat. Biotechnol., 29(11):987–991, Nov. 2011.

[12] I. Cortés-Ciriano, J. J.-K. Lee, R. Xi, D. Jain, Y. L. Jung, L. Yang, D. Gordenin, L. J. Klimczak, C.-Z. Zhang, D. S. Pellman, PCAWG Structural Variation Working Group, P. J. Park, and PCAWG Consortium. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat. Genet., 52(3):331–341, Mar. 2020.

[13] S. Darmanis, S. A. Sloan, D. Croote, M. Mignardi, S. Chernikova, P. Samghababi, Y. Zhang, N. Neff, M. Kowarsky, C. Caneda, G. Li, S. D. Chang, I. D. Connolly, Y. Li, B. A. Barres, M. H. Gephart, and S. R. Quake. Single-Cell RNA-Seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. Cell Rep., 21 (5):1399–1410, Oct. 2017.

[14] A. P. Dawid. Posterior model probabilities. In P. S. Bandyopadhyay and M. R. Forster, editors, Philosophy of Statistics, volume 7, pages 607–630. North-Holland, Amsterdam, Jan. 2011.

[15] J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. Hoffman, and R. A. Saurous. TensorFlow distributions. Nov. 2017.

[16] V. B. Dubinkina, D. S. Ischenko, V. I. Ulyantsev, A. V. Tyakht, and D. G. Alexeev. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. BMC Bioinformatics, 17:38, Jan. 2016.

[17] R. A. Dunstan, D. Pickard, S. Dougan, D. Goulding, C. Cormie, J. Hardy, F. Li, R. Grinter, K. Harcourt, L. Yu, J. Song, F. Schreiber, J. Choudhary, S. Clare, F. Coulibaly, R. A. Strugnell, G. Dougan, and T. Lithgow. The flagellotropic bacteriophage YSD1 targets salmonella typhi with a chi-like protein tail fibre. Mol. Microbiol., 112(6):1831–1846, Dec. 2019.

[18] R. Durbin, S. Eddy, A. Krogh, and A. Mitchison. Biological Sequence Analysis. 1998.

[19] J. Frazer, P. Notin, M. Dias, A. Gomez, K. Brock, Y. Gal, and D. Marks. Large-scale clinical interpretation of genetic variants using evolutionary data and deep learning. Dec. 2020.

[20] S. Geman and C.-R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. The Annals of Statistics, 10(2):401–414, 1982.

[21] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. Ann. Stat., 28(2):500–531, 2000.

[22] J. Ghosh and R. Ramamoorthi. Bayesian Nonparametrics. 2003.

[23] P. Gopalan, W. Hao, D. M. Blei, and J. D. Storey. Scaling probabilistic models of genetic variation to millions of humans. Nat. Genet., 48(12):1587–1590, Dec. 2016.

[24] R. M. Gray. Entropy and Information Theory. Springer Science & Business Media, Jan. 2011.

[25] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. J. Mach. Learn. Res., 13:723–773, 2012.

[26] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. Nature, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

[27] C. C. Holmes, F. Caron, J. E. Griffin, and D. A. Stephens. Two-sample bayesian nonparametric hypothesis testing. Bayesian Anal., 10(2):297–320, June 2015.

[28] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Schärfe, M. Springer, C. Sander, and D. S. Marks. Mutation effects predicted from sequence co-variation. Nat. Biotechnol., 35(2):128–135, Feb. 2017.

[29] W. Huang, L. Li, J. R. Myers, and G. T. Marth. ART: a next-generation sequencing read simulator. Bioinformatics, 28(4):593–594, Feb. 2012.

[30] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean. De novo assembly and genotyping of variants using colored de bruijn graphs. Nat. Genet., 44(2):226–232, Jan. 2012.

[31] P. E. Jacob, L. M. Murray, C. C. Holmes, and C. P. Robert. Better together? statistical learning in models made of modules. Aug. 2017.

[32] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol., 37(8): 907–915, Aug. 2019.

[33] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[34] D. P. Kingma and M. Welling. Auto-Encoding variational bayes. Dec. 2013.

[35] M. Kokot, M. Dlugosz, and S. Deorowicz. KMC 3: counting and manipulating k-mer statistics. Bioinformatics, 33(17):2759–2761, Sept. 2017.

[36] W. Kool, H. van Hoof, and M. Welling. Stochastic beams and where to find them: The Gumbel-Top-k trick for sampling sequences without replacement. In International Conference on Machine Learning, pages 3499–3508. PMLR, 2019.

[37] A. Kucukelbir and D. M. Blei. Population empirical bayes. In Uncertainty in Artificial Intelligence, 2015.

[38] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. J. Mach. Learn. Res., 18(14):1–45, Jan. 2017.

[39] Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In International Conference on Machine Learning, pages 1718–1727. PMLR, 2015.

[40] J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In Advances in Neural Information Processing Systems, pages 829–837, 2015.

[41] J. Lloyd-Price, A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, A. B. Hall, A. Brady, H. H. Creasy, C. McCracken, M. G. Giglio, D. McDonald, E. A. Franzosa, R. Knight, O. White, and C. Huttenhower. Strains, functions and dynamics in the expanded human microbiome project. Nature, 550(7674):61–66, Oct. 2017.

[42] G. Marçais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27(6):764–770, Mar. 2011.

[43] J. W. Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. July 2019.

[44] S. Mohamed and B. Lakshminarayanan. Learning in implicit generative models. Oct. 2016.

[45] R. E. Mukamel, R. E. Handsaker, M. A. Sherman, A. R. Barton, Y. Zheng, S. A. McCarroll, and P.-R. Loh. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Jan. 2021.

[46] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., 44(D1):D733–45, Jan. 2016.

[47] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In Annual Meeting of the Association for Computational Linguistics, pages 311–318, 2002.

[48] S. Petrone, J. Rousseau, and C. Scricciolo. Bayes and empirical bayes: do they merge? Biometrika, 101(2):285–302, 2014.

[49] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. Genetics, 155(2):945–959, June 2000.

[50] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Proceedings of the 31st International Conference on Machine Learning, 2014.

[51] A. J. Riesselman, J. B. Ingraham, and D. S. Marks. Deep generative models of genetic variation capture the effects of mutations. Nat. Methods, 15(10):816–822, Oct. 2018.

[52] J. Rousseau. On the Frequentist Properties of Bayesian Nonparametric Methods. Annual Review of Statistics and Its Application, 3:211–231, 2016.

[53] J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. J. R. Stat. Soc. Series B Stat. Methodol., 73(5):689–710, Nov. 2011.

[54] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, and R. Ranganathan. An evolution-based model for designing chorismate mutase enzymes. Science, 369:440–445, 2020.

[55] P. W. Schreiber, V. Kufner, K. Hübel, S. Schmutz, O. Zagordi, A. Kaur, C. Bayard, M. Greiner, A. Zbinden, R. Capaul, J. Böni, H. H. Hirsch, T. F. Mueller, N. J. Mueller, A. Trkola, and M. Huber. Metagenomic virome sequencing in living donor and recipient kidney transplant pairs revealed JC polyomavirus transmission. Clin. Infect. Dis., 69 (6):987–994, Aug. 2019.

[56] J.-E. Shin, A. J. Riesselman, A. W. Kollasch, C. McMahon, E. Simon, C. Sander, A. Manglik, A. C. Kruse, and D. S. Marks. Protein design and variant prediction using autoregressive generative models. Nat. Commun., 12(1):2403, Apr. 2021.

[57] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol. ABySS: a parallel assembler for short read sequence data. Genome Res., 19(6):1117–1123, June 2009.

[58] P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M.-L. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Morsberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal, and P. J. Campbell. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell, 144(1):27–40, Jan. 2011.

[59] D. J. Sutherland, H. Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In International Conference on Learning Representations. arxiv.org, 2017.

[60] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. J. R. Stat. Soc. Series B Stat. Methodol., 61(3):611–622, Aug. 1999.

[61] D. Tran, M. Hoffman, D. Moore, C. Suter, S. Vasudevan, A. Radul, M. Johnson, and R. A. Saurous. Simple, distributed, and accelerated probabilistic programming. In Neural Information Processing Systems, 2018.

[62] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. Sept. 2016.

[63] L. van der Maaten. Visualizing data using t-SNE. J. Mach. Learn. Res., 9:2579–2605, 2008.

[64] A. W. van der Vaart. Asymptotic Statistics. 1998.

[65] R. Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science. 2020.

[66] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[67] Y. Voichek and D. Weigel. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nature Genetics, 52(5):534–540, 2020.

[68] E. N. Weinstein and D. S. Marks. A structured observation distribution for generative biological sequence prediction and forecasting. Feb. 2021.

A Broader impact statement

The social impact and bioethics of genetics is a complex and well-studied subject. Here, we briefly highlight some of the ways the methods presented in this article intersect with major issues. Health: statistical genetics methods have been crucial in diagnosing disease and dissecting its mechanisms, and we hope that the methods presented here will further this research. On the other hand, a detailed understanding of personal genetic information can be used as the basis for discrimination. Technology: as generative models, BEAR models may be unusually useful in designing new sequences such as therapeutics. Genetic engineering, however, is a dual-use technology. Culture: popular understanding of genetics is often bound up in the idea that genomes are relatively fixed and change only very slowly over time, which feeds into concerns over the "naturalness" of genetic modification and beliefs about the static nature of race and ethnicity. These public perceptions are likely partially a consequence of scientific models and methods that often simply ignore complex genetic variation and analyze individual genomic data by comparing it to "reference" individuals. BEAR models contribute theoretical and statistical methods for working with complex variants and without relying on references.

B Overview of supplemental material

Sections C-H are our theoretical results. Section I describes our simulation experiments. Section J details how we implemented scalable inference for BEAR models. Sections K-O provide details on our empirical results based on real data. The Datasets.xlsx file contains information on all the datasets, including links or accession numbers for public databases. Code is available at `https://github.com/debbiemarkslab/BEAR`.

C Theory Introduction

BEAR models can be used to address a variety of different estimation and testing problems, and the theoretical questions that arise in each case are related but distinct. One crucial, high-level distinction is between the “finite-lag case” (where we assume the model lag L is finite) and the “infinite-lag case” (where we allow the model lag L to approach infinity). In addressing nonparametric density estimation, it is crucial to consider the infinite lag case, since it is likely in practice that the true distribution can only be matched in the infinite L limit. On the other hand, when it comes to diagnosing misspecification or constructing

hypothesis tests, the finite lag case is more acceptable since it is likely in practice that any differences between the model and the data, or between two datasets, will be reflected in finite marginals of the data distribution. The finite lag case is complicated by the fact that it is likely that many kmer-to-base transitions have extremely low probability in practice; even on massive datasets, we observe many transitions with no counts whatsoever. To deal with this case, we develop theoretical tools to accommodate the possibility that some transitions truly have probability zero under the data generating distribution.

An essential and innovative aspect of our formalism is the focus on "subexponential" sequence distributions that obey an exponential moment bound on their length. Our choice to consider sequence distributions that have no upper bound on the lengths of sequences they produce separates our theory from the theory of distributions on finite sets. On the other hand, moment bound assumptions separate our theory from the theory of distributions on countable sets.

The theory will be organized as follows. Section D describes basic theoretical properties finite-lag Markov sequence models, including their expressiveness and subexponentiality. Subexponential sequence models will be introduced in general here. Section E demonstrates consistency of inference with a fixed lag and in model selection between lags. A connection is established between effective model dimensions and topologies of de Bruijn graphs. Section F describes the behavior of the model when inference proceeds by empirical Bayes. The parameter h is established as a descriptor of misspecification. Section G describes theoretical guarantees on the behavior of goodness-of-fit and two-sample tests. Finally, section H demonstrates consistency in the infinite lag case. Later sections depend on definitions and results established in previous sections with the exception that section H may be read immediately after reading the definitions at the top of section E.

C.1 Notation

We consider an alphabet $\mathcal{B}$ with more than one letter. Define $\tilde{\mathcal{B}} = \mathcal{B} \cup \{\$\}$ where \$ is interpreted as the stop symbol, i.e. \$ may only appear as the last letter of a sequence. Also define the set of strings of the alphabet $\mathcal{B}$ of length L that start with any number (including 0) of repeated $\emptyset$ symbols, $\mathcal{B}_L^o$. For a sequence X of letters in $\mathcal{B}$, possibly terminated by \$, we define $|X|$ as its length, including the stop symbol \$ but not any start symbols $\emptyset$. For two strings X, X' define $\#X'(X)$ the number of occurrences of X' as a substring in X and, if X is not terminated by \$, define (X, X') as the concatenated string. We also define the substring from index i to j (inclusive) of X as $X_{i:j}$.

Define the set S of all finite sequences terminated by a stop symbol and give it the discrete topology. Note that S is countable. Say p is a distribution of S. We will use $\mathbb{E}_p$, or $\mathbb{E}$ if there is an unambiguous data-generating distribution, to denote taking an expectation; for example, $\mathbb{E}_p \#X'$ is the expected number of occurrences of the substring X' in sequences drawn from p. For a sequence Y possibly not terminated by a stop symbol, we define $p(Y \dots) = p(\{X \in S \mid X_i = Y_i \ \forall i \leq |Y|\})$. We also define subexponential moment bounds, an assumption we will make great use of:

Definition 1 (Subexponential sequence distributions)**.** We say a distribution p on S is subexponential if for a $t > 0$, $\mathbb{E}_p \exp(t|X|) < \infty$.

For a random variable Z on a probability space with probability P, and a measurable set A in the sample space, we define

$$\mathbb{E}[Z; A] = \mathbb{E}[Z \mathbb{1}_A] = \mathbb{E}[Z|A]P(A)$$

where $\mathbb{1}_A$ is the random variable with $\mathbb{1}_A = 1$ on A and $\mathbb{1}_A = 0$ outside of A. As well, for two real sequences $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$, both possibly undefined for small n, we write $a_n \lesssim b_n$ to mean that there is a positive constant C such that eventually $a_n \leq C b_n$. We write $a_n \sim b_n$ when $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. We define $a \wedge b$ as the minimum of a and b, and $a \vee b$ as the maximum.

D Finite-lag Markov models

In this section we define finite-lag Markov models, and then study the expressiveness of the model class. After defining finite-lag Markov models, this section will concern itself with the expressiveness of the model class. We first show that while there are sequence distributions over S that are not finite-lag Markov models, the set of finite-lag Markov models is nevertheless dense in the space of distributions over S. We then show that finite-lag Markov models are subexponential.

The class of finite-lag Markov models is defined to be

$$\text{Parameters: lag } L, \text{transition probabilities } \{v_{k,b}\}_{k\in\mathcal{B}_L^o, b\in\tilde{\mathcal{B}}}$$
$$X_i = \emptyset \text{ for } i \le 0$$
$$X_{i+1} \sim \text{Categorical}(\{v_{X_{i-L+1:i},b}\}_{b\in\tilde{\mathcal{B}}})$$

stopping generation when a \$ symbol is drawn and with parameters picked so that $|X| < \infty$ a.s.. These models are equivalent to Markov processes on the set $\mathcal{B}_L^o \cup \{(X, \$) \mid X \in \mathcal{B}_{L-1}^o\}$. The requirement that generated sequences be finite length a.s. is equivalent to the requirement that for every $k \in \mathcal{B}_L^o$ that is Markov-accessible, there is another $k' \in \mathcal{B}_L^o$ that is Markov-accessible from k such that $v_{k',\$} > 0$. Call p_v a probability distribution generated this way with parameters L, v. Call the set of such probability distributions with lag L $\mathcal{M}_L$. Define the set of all finite lag Markov models $\mathcal{M} := \cup_{L=1}^{\infty}\mathcal{M}_L$ and note $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots$. Defining $\Delta_{\tilde{\mathcal{B}}}$ as the $|\tilde{\mathcal{B}}| - 1$-dimensional simplex with coordinates indexed by $\tilde{\mathcal{B}}$, $\mathcal{M}_L$ is parametrized by transition probabilities in $\Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}$. This parametrization is not defined everywhere on the boundary and is not injective as if an L-mer k is not Markov-accessible by a distribution p_v, the vector of probabilities v_k does not affect p_v's distribution. This parametrization is continuous in the sense of the topology described by proposition 2.

We first give some examples of simple sequence distributions that are not finite-lag Markov.

Proposition 1. *Not all possible distributions over S are in $\mathcal{M}$.*

Proof. Let $A \in \mathcal{B}$ and p^* be a distribution over finite sequences that puts probability a_i on the sequence $A \times i := A \ldots A$ of length i with $\sum_{i=0}^{\infty} a_i = 1$. Assume $p^* \in \mathcal{M}_L$ with transition probabilities $\{v_{k,b}\}_{k\in\mathcal{B}_L^o, b\in\tilde{\mathcal{B}}}$.

For $i \le L$, define $v_i := v_{(\emptyset\times(L-i),A\times i)}$, i.e. the vector of transition probabilities from the L-mer that is $L-i$ $\emptyset$ symbols followed by i A symbols. For $i \ge L$ call $v_i := v_L$.

Notice that for any i, the \$-component of the vector v_i is $p^*(|X| = i \mid |X| \ge i) = \frac{a_i}{S_i}$ where $S_i := \sum_{j=i}^{\infty} a_j$. Thus the A-component is $1 - \frac{a_i}{S_i} = \frac{S_{i+1}}{S_i}$. By the definition of the sequence $(v_i)_{i=1}^{\infty}$, it is constant for $i \ge L$. Call $\alpha := S_{L+1}/S_L = v_{L,A} = v_{i,A} = S_{i+1}/S_i$ for all $i \ge L$. Thus for all $i > L$, $a_i = S_i v_{i,\$} = \alpha^{i-L} S_L v_{L,\$} = \alpha^{i-L} a_L$. Thus the sequence a_i eventually decays exponentially and, as examples, it is impossible that $a_i \sim 1/i!$ or $a_i \sim 1/i^2$. □

Next we show that $\mathcal{M}$ is dense in the set of probability distributions on S. To speak of density, we review the topology and types of convergence on the set of distributions of S in this next proposition.

Proposition 2. *The topology of convergence in total variation, convergence in distribution, and pointwise convergence of the probability of each $X \in S$ are identical.*

Proof. Pointwise convergence of the probability of each $X \in S$ implies convergence in total variation by Scheffé's lemma. It is also known that the topology induced by the total variation metric is stronger than the topology of convergence in distribution. Finally, since for each $X \in S$, the set $\{X\}$ is open and closed, so that the Portmanteau lemma shows that convergence in distribution implies pointwise convergence. □

Lemma 3. *Say p is a distribution on S. There is a lag L Markov model, p_L, such that for all $X \in S$, if $|X| \le L$, $p_L(X) = p(X)$, and if $|X| > L$, $p_L(X) = p(X_{1:L})|\tilde{\mathcal{B}}|^{-(|X|-L)}$.*

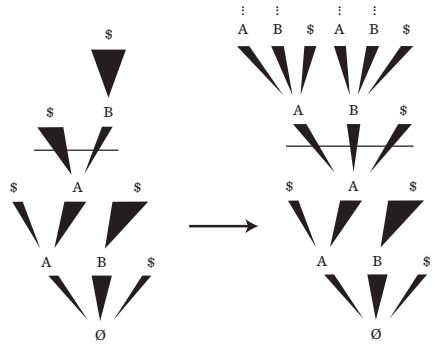

Figure S1: Example application of this construction to the distribution on the left. Transition probabilities for kmers smaller than $L = 2$ are those defined by the original distribution, while those for larger kmers are all 1/3. The thickness of each line denotes the probability of the transition.

Proof. For all $k \in \mathcal{B}_L^o, b \in \tilde{\mathcal{B}}$, if there is a start symbol $\emptyset$ in k, define $v_{k,b} = \frac{p((k,b)\ldots)}{p(k\ldots)}$, otherwise, define $v_{k,b} = |\tilde{\mathcal{B}}|^{-1}$. It is clear p_v satisfies the properties of p_L (Fig S1). □

Corollary 4. *$\mathcal{M}$ is dense in the set of distributions of S.*

Proof. Define p to be a distribution on S with finite support, $\{X_n\}_{n=1}^N$. Pick an $L > |X_n|$ for all n, so that with the definition of lemma 3, $p_L = p$ and thus $p \in \mathcal{M}_L$. Now note that any distribution on S can be approximated at finitely many points in S arbitrarily well by distributions with finite support. The result follows from proposition 2. □

Proposition 5. *Finite-lag Markov models are subexponential.*

Proof. Say $p \in \mathcal{M}_L$ for some L, with transition probabilities v. Every $k \in \mathcal{B}_L^o$ that is Markov-accessible by p has a $k' \in \mathcal{B}_L^o$ that is Markov-accessible from k in less than s_k transitions such that $v_{k',\$} > 0$. Thus, $\inf_i p(|X| \leq i + s_k \mid |X| > i, X_{i-L+1:i} = k) > 0$. Define $s = \max_{k \text{ accessible}} s_k$, $q = \inf_i p(|X| \leq i + s \mid |X| > i) > 0$. Now note, for any positive integer m, $p(|X| > ms) = \prod_{i=1}^m p(|X| > is \mid |X| > (i-1)s) \leq (1-q)^m$. For a random variable $Z \sim \text{Geom}(q)$,

$$\begin{aligned}
\mathbb{E}_p \exp(t|X|) &= \int_0^\infty dyp\left(|X| > s\left(\frac{1}{st}\log(y)\right)\right) \\
&\leq \int_0^\infty dyp\left(|X| > s\left(\lfloor\frac{1}{st}\log(y)\rfloor\right)\right) \\
&\leq \int_0^\infty dyP\left(Z > \left(\lfloor\frac{1}{st}\log(y)\rfloor\right)\right) \\
&\leq \int_0^\infty dyP\left(Z > \left(\frac{1}{st}\log(y) - 1\right)\right) \\
&= \mathbb{E}\exp\left(ts(Z+1)\right)
\end{aligned} \tag{5}$$

The integral is finite for some $t > 0$ as geometric random variables are sub-exponential. □

E Consistency in the finite L case

In this section we consider fitting to data BEAR models with fixed hyperparameters h and θ (that is, standard Bayesian Markov models). We first study the asymptotic behavior of the posterior over v, the transition probability parameter, conditional on a particular lag L. We prove a Wald-type consistency result, showing that the posterior concentrates on a neighborhood of the true data-generating parameter value v^*, if such a value exists;

when p^* is not in the model class $\mathcal{M}_L$, the posterior over v concentrates at the point v^* corresponding to the distribution $p_{v^*} \in \mathcal{M}_L$ closest in KL divergence to p^*. We next study the asymptotic behavior of the posterior over the lag L, building on the theory of nested model selection since L is a discrete variable. We show that the posterior concentrates at the true data-generating value L^* when such a lag exists (i.e. when there is some L^* such that $p^* \in \mathcal{M}_{L^*}$), and otherwise diverges. At a high level, neither of these results are surprising, and they would be expected to hold in general for well-behaved Bayesian models. The details of the model's asymptotic behavior, however, turn out to be somewhat unusual; as we shall see, the fact that some transitions from a particular kmer k to a base b have probability zero under the data-generating distribution p^* can complicate the normal story of Bayesian asymptotics.

To describe the possible kmer-base transitions, we define, for a distribution on S, p, and a lag L, the set of accessible kmers $\text{acc}_L(p) = \{k \in \mathcal{B}_L^o \mid p(\#k > 0) > 0\}$ and transitions $\text{supp}_L(p) = \{(k,b) \mid k \in \mathcal{B}_L^o, b \in \tilde{\mathcal{B}}, p(\#(k,b) > 0) > 0\}$. Define also, for any particular a $k \in \mathcal{B}_L^o$, the set of allowed transitions $\text{supp}_L(p)|_k := \{b \in \tilde{\mathcal{B}} \mid (k,b) \in \text{supp}_L(p)\}$. Define the restriction of the parameter space $\Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}$ to the support of p^*, $\tilde{\Delta}_L(p^*) = \prod_{k \in \text{acc}_L(p^*)} \Delta_{\text{supp}_L(p^*)|_k}$. If $v \in \Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}$, we will often use the abbreviation $\text{supp}(v) = \text{supp}_L(p_v)$ for convenience.

Say p^* is a distribution on S and L is a lag. Define the transition probabilities v^*, corresponding to the closest model in $\mathcal{M}_L$ to p^* (as measured by KL), as

$$v^* = \arg\min_{v \in \Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}} \text{KL}(p^* || p_v) = \arg\max_v \mathbb{E} \log p_v(X) = \arg\max_v \sum_{k,b} \mathbb{E}\left[\#(k,b)\right] \log v_{k,b}.$$

Unlike for many other statistical models studied in other contexts, here we can easily compute the closest model to the data-generating distribution: using Lagrange multipliers, one may see that for all $k \in \text{acc}_L(p^*)$, $v^*_{k,b} = \mathbb{E}\left[\#(k,b)\right] / \mathbb{E}\left[\#k\right]$. We then define $p^{*(L)} = p_{v^*}$ as the best approximation to p^* in $\mathcal{M}_L$. Note $\text{supp}(v^*) = \text{supp}_L(p^{*(L)}) = \text{supp}_L(p^*)$.

We now ask whether Bayesian inference on $\mathcal{M}_L$ is consistent, i.e., whether the posterior converges to a point mass at $p^{*(L)}$, even in the case where $\text{supp}_L(p^*)$ is not all of $\mathcal{B}_L^o \times \tilde{B}$. The result is a classic Wald-type argument, adapted from theorem 2.3 of Miller [43] and theorem 1.3.4 in Ghosh and Ramamoorthi [22]. The primary difficulty in the proof is that these previous theorems assume the true parameter value lies on the interior of the parameter space and rely on uniform convergence of the mean log likelihood in a neighborhood around the true value. In our case, we can have $v^*_{k,b} = 0$, so that the true parameter value lies on the boundary of its space $\Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}$ and the likelihood function diverges at this boundary point.

Theorem 6. *Say p^* is a distribution on S with $\mathbb{E}|X| < \infty$. Say Π is a prior on $\Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}$ that assigns probability 0 to the set of v with $p_v \notin \mathcal{M}_L$. Say $X_1, X_2, \cdots \sim p^*$ iid. Call $V = \{v \in \Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o} \mid p_v = p^{*(L)}\}$ and assume that it is not disjoint from the support of Π. Then for all open sets U containing V,*

$$\Pi(U | X_1, \ldots, X_N) \to 1$$

a.s.. As a probability distribution on the space of measures on S, $\Pi | X_1, \ldots, X_N \to \delta_{p^{*(L)}}$.

Proof. Define v^* as the transition probabilities of $p^{*(L)}$. Define $l_N(v) = -\frac{1}{N}\sum_{n=1}^N \log(p_v(X_n))$, which is continuous in v and $\nu^* = \min_{(k,b) \in \text{supp}(v^*)} v^*_{k,b}$. Note that

$$\mathbb{E} \log p^{*(L)}(X) = \mathbb{E} \sum_{i=1}^{|X|} \log v^*_{X_{i-L:i-1}, X_i} \geq \mathbb{E}|X| \log \nu^*.$$

First we show that the likelihood of the data is eventually small in a neighborhood of the boundary. Pick an $\eta_1 > 0$. Say $(k,b) \in \text{supp}(v^*) = \text{supp}_L(p^*)$ and define $q_{k,b} = p^*(\#(k,b) > 0)$ which is positive. Pick a positive

$$\nu_{k,b} < \exp\left(-q_{k,b}^{-1}(\eta_1 - \mathbb{E}|X| \log \nu^*)\right) \wedge v^*_{k,b}.$$

$$\begin{aligned}\mathbb{E}\sup_{v \text{ s.t. } v_{k,b}<\nu_{k,b}} l_1(v^*) - l_1(v) =& \mathbb{E}\left[\sup_{v \text{ s.t. } v_{k,b}<\nu_{k,b}} \log p_v(X)\right] - \mathbb{E}\left[\log p_{v^*}(X)\right] \\ \leq& q_{k,b}\log\nu_{k,b} + (-\log\nu^*)\mathbb{E}|X| < -\eta_1.\end{aligned} \tag{6}$$

Thus defining $U_1 = \{v \in \Delta_{\vec{\mathcal{B}}}^{\mathcal{B}_L^o} \mid \text{there exists } (k,b) \in \text{supp}(v^*) \text{ s.t. } v_{k,b} < \nu_{k,b}\}$, a.s., for large enough N, $l_N(v^*) - l_N(v) < -\eta_1$ for all $v \in U_1$ by the SLLN.

Call the complement of U_1 K. K is compact and for all $v \in K$, $\text{supp}(v^*) \subseteq \text{supp}(v)$. Note that V is compact and in the interior of K. Pick a positive ν_K which has, for every $(k,b) \in \text{supp}(v^*)$, $\nu_K < \nu_{k,b}$. Then

$$\mathbb{E}\sup_{v\in K}|l_1(v^*) - l_1(v)| \leq |\log(\nu_K \wedge \nu^*)|\mathbb{E}|X| < \infty.$$

Then by theorem 1.3.3 in Ghosh and Ramamoorthi [22], a.s., $l_N(v^*) - l_N(v)$ converges uniformly to $\text{KL}(p^*||p^{*(L)}) - \text{KL}(p^*||p_v) \leq 0$ on K (note, for the application of theorem 1.3.3 in Ghosh and Ramamoorthi [22], this quantity is well defined even if p_v is not a distribution over finite strings).

Now pick an open neighborhood U of V. By the continuity of $v \mapsto \text{KL}(p^*||p_v)$, since $K \setminus U$ is compact, $\inf_{v\in K\setminus U}\text{KL}(p^*||p_v) > \text{KL}(p^*||p^{*(L)})$ otherwise there would be a $v \in V \setminus K$. Thus we can pick a positive $\text{KL}(p^*||p^{*(L)}) + \eta_2 < \inf_{v\in K\setminus U}\text{KL}(p^*||p_v)$. Since $v \mapsto \text{KL}(p_{v^*}||p_v)$ is continuous and K is a neighborhood of V, there is an open $U_2 \subset K \cap U$ containing V such that one can pick an η_3 with $\sup_{v\in U_2}\text{KL}(p^*||p_v) - \text{KL}(p^*||p^{*(L)}) < \eta_3 < \eta_1 \wedge \eta_2$. Then a.s. eventually, $l_N(v^*) - l_N(v) < -\eta_2$ for all $v \in K \setminus U$ and $l_N(v^*) - l_N(v) > -\eta_3$ for all $v \in U_2$. Thus, a.s. for large enough N,

$$\begin{aligned}\Pi(U|X_1,\ldots,X_n) =& \frac{\int_U d\Pi e^{N(l_N(v^*)-l_N(v))}}{\int_U d\Pi e^{N(l_N(v^*)-l_N(v))} + \int_{K\setminus U} d\Pi e^{N(l_N(v^*)-l_N(v))} + \int_{U_1\setminus U} d\Pi e^{N(l_N(v^*)-l_N(v))}} \\ \geq& \left(1 + \frac{\int_{K\setminus U} d\Pi e^{N(l_N(v^*)-l_N(v))}}{\int_{U_2} d\Pi e^{N(l_N(v^*)-l_N(v))}} + \frac{\int_{U_1} d\Pi e^{N(l_N(v^*)-l_N(v))}}{\int_{U_2} d\Pi e^{N(l_N(v^*)-l_N(v))}}\right)^{-1} \\ \geq& \left(1 + \frac{\Pi(K\setminus U)e^{-N\eta_2}}{\Pi(U_2)e^{-N\eta_3}} + \frac{\Pi(U_1)e^{-N\eta_1}}{\Pi(U_2)e^{-N\eta_3}}\right)^{-1} \\ \to& 1.\end{aligned} \tag{7}$$

Finally, as a probability distribution on the space of measures on S, $\Pi|X_1,\ldots,X_n \to \delta_{p^{*(L)}}$. This follows from the fact that the prior and thus posterior probability of $p_v \notin \mathcal{M}_L$ is 0 and so one may push forward the measure from $\Delta_{|\vec{\mathcal{B}}|}^{\mathcal{B}_L^o}$ to the space of probability measures on S. The image of V is a point p_{v^*}. Since this mapping is continuous, it preserves the weak convergence of the measure, in this case to a point mass. □

Next we will study the posterior distribution of the BEAR model over the lag L, showing under general assumptions that the posterior concentrates on the true data-generating value L^* (when such a value exists). Our analysis builds off of standard asymptotic analyses of nested Bayesian model selection, since models with different lags are nested, i.e. $\mathcal{M}_L \subset \mathcal{M}_{L'}$ when $L' > L$. Typically, when a simpler model (e.g. $\mathcal{M}_L$) is nested inside a more complex model (e.g. $\mathcal{M}_{L'}$), and the data-generating distribution p^* is in the simpler model, the log Bayes factor comparing the two models will asymptotically prefer the simpler model and scale as $\frac{1}{2}(\dim' - \dim)\log N$ where $\dim'$ is the dimension of the parameter space in the more complex model and dim is the dimension in the simpler model [14]. This $O(\log N)$ term, which is independent of the prior, can be thought of as originating from the Laplace approximation to the marginal likelihood; it is the basis of such widely used model-selection techniques as the Bayesian information criterion.

Somewhat surprisingly, the fact that some transitions may have probability zero ($v^*_{k,b} = 0$) changes the asymptotic behavior of the log Bayes factor, in particular by altering the dimension factor $\dim' - \dim$. In effect, dimensions of the parameter space corresponding

to kmers that occur with probability zero under p^* do not contribute to the dimension count, while dimensions for which $v^*_{k,b} = 0$ do not count as full dimensions; this leads to the notion of an "effective model dimension", defined as $\dim^{\text{eff}}_L(p^*) := |\text{supp}_L(p^*)| - |\text{acc}_L(p^*)| + \sum_{k\in\text{acc}_L(p^*)}\sum_{b\notin\text{supp}_L(p^*)|_k}\alpha_{k,b}$ where $\alpha_{k,b}$ is the concentration of the Dirichlet prior. This effective dimension depends the data-generating distribution p^* and on the prior hyperparameters, not just on L. Note that the unusual asymptotic behavior of BEAR models does not just come from their Markov structure; even in the everyday example of a Dirichlet-Categorical model, if some outcomes of the Categorical distribution have probability exactly zero under the true data-generating distribution, the standard Laplace approximation does not hold, and the Dirichlet prior contributes $O(\log N)$ terms to the log marginal likelihood [53].

Theorem 7. *Say p^* is a distribution on S with $\mathbb{E}|X|^2 < \infty$ and say $X_1, X_2, \cdots \sim p^*$ iid. Given L, consider a $\text{Dirichlet}(\alpha_{k,b})_{b\in\tilde{\mathcal{B}}}$ prior on the simplex in $\Delta^{\mathcal{B}^o_L}_{\tilde{\mathcal{B}}}$ corresponding to the L-mer k. For all L, assume $\alpha_{k,b} > 0$ for $(k,b) \in \text{supp}_L(p^*)$ (otherwise $p((X_n)_{n=1}^N|\mathcal{M}_L)$ is eventually 0 a.s.).*

Define $\text{KL}(p^*||\mathcal{M}_L) := \inf_{p\in\mathcal{M}_L}\text{KL}(p^*||p)$. Given $L_1 \neq L_2$, if[2] *$\text{KL}(p^*||\mathcal{M}_{L_2}) > \text{KL}(p^*||\mathcal{M}_{L_1})$,*

$$\log\frac{p((X_n)_{n=1}^N|\mathcal{M}_{L_1})}{p((X_n)_{n=1}^N|\mathcal{M}_{L_2})} = N\left(\text{KL}(p^*||\mathcal{M}_{L_2}) - \text{KL}(p^*||\mathcal{M}_{L_1})\right) + O_p(\sqrt{N}). \tag{8}$$

Otherwise, if $p^* \in \mathcal{M}_{L_1}, \mathcal{M}_{L_2}$ and, defining, for a lag L, $dim^{\text{eff}}_L(p^*) := |\text{supp}_L(p^*)| - |\text{acc}_L(p^*)| + \sum_{k\in\text{acc}_L(p^*)}\sum_{b\notin\text{supp}_L(p^*)|_k}\alpha_{k,b}$,

$$\log\frac{p((X_n)_{n=1}^N|\mathcal{M}_{L_1})}{p((X_n)_{n=1}^N|\mathcal{M}_{L_2})} = \frac{1}{2}\left(dim^{\text{eff}}_{L_2}(p^*) - dim^{\text{eff}}_{L_1}(p^*)\right)\log N + O_p(1). \tag{9}$$

Proof. For a lag L, note $\dim(\tilde{\Delta}_L(p^*)) = \text{supp}_L(p^*) - \text{acc}_L(p^*)$. Put a $\text{Dirichlet}(\alpha_{k,b})_{b\in\text{supp}_L(p^*)|_k}$ prior on each $\Delta_{\text{supp}_L(p^*)|_k}$. Call $\tilde{\mathcal{M}}_L$ the set of probability distributions described by $\tilde{\Delta}_L(p^*)$. We will show that $\text{KL}(p^*||p.)$ is maximized in the interior of $\tilde{\Delta}_L(p^*)$ so that the asymptotics of the marginal likelihood $\left(p(X|\tilde{\mathcal{M}}_L)\right)$ are well understood. In $\Delta^{\mathcal{B}^o_L}_{|\tilde{\mathcal{B}}|}$ however, there are dimensions that correspond to k-mer - base transitions that are impossible under p^*. Using the symmetry of the Dirichlet prior, we can de-couple the asymptotics of these excess dimensions from the asymptotics of the much more "natural" space of $\tilde{\Delta}_L(p^*)$:

$$\begin{aligned}
\log\left(p((X_n)_{n=1}^N|\mathcal{M}_L)\right) &= \sum_{k\in\text{acc}_L(p^*)}\left(\log\frac{\Gamma(\sum_b\alpha_{k,b})}{\Gamma(\sum_b\alpha_{k,b}+\#k)} - \sum_{\text{supp}_L(p^*)|_k}\log\frac{\Gamma(\alpha_{k,b})}{\Gamma(\alpha_{k,b}+\#(k,b))}\right)\\
&= \log\left(p((X_n)_{n=1}^N|\tilde{\mathcal{M}}_L)\right)\\
&\quad + \sum_{k\in\text{acc}_L(p^*)}\left(\log\frac{\Gamma(\sum_b\alpha_{k,b})}{\Gamma(\sum_b\alpha_{k,b}+\#k)} - \log\frac{\Gamma(\sum'_b\alpha_{k,b})}{\Gamma(\sum'_b\alpha_{k,b}+\#k)}\right)
\end{aligned} \tag{10}$$

where $\sum'_b$ is a sum over the $b \in \text{supp}_L(p^*)|_k$, and where $\#k$ in this case is $\sum_{n=1}^N \#k(X_n)$ and $\#(k,b)$ is similar. We will deal with each of these terms in turn.

[2] We do not need to assume $\mathbb{E}\log p > -\infty$ as we may define in this case $\text{KL}(p^*||\mathcal{M}_{L_2}) - \text{KL}(p^*||\mathcal{M}_{L_1}) = -\mathbb{E}\log p^{*(L_2)}(X) + \mathbb{E}\log p^{*(L_1)}(X)$ which we will see is bounded by the moment bound assumption $\mathbb{E}|X|^2 < \infty$.

To analyze the first of these terms, we first check regularity conditions. For $v \in \tilde{\Delta}_L(p^*)$ and strings $X_1, \dots X_N$, define

$$l_N(v) = -\frac{1}{N} \log \prod_{n=1}^{N} p_v(X_n) = -\frac{1}{N} \sum_{(k,b) \in \text{supp}_L(p^*)} \#(k,b) \log v_{k,b}$$
$$l(v) = -\mathbb{E} \log p_v(X) = - \sum_{(k,b) \in \text{supp}_L(p^*)} \mathbb{E}[\#(k,b)] \log v_{k,b}.$$

Call v_n the minimizer of l_N and v^* the minimizer of l. Note v^* is also the minimizer of $v \mapsto \text{KL}(p^*||p_v)$ for $v \in \tilde{\Delta}_L(p^*)$ and has $v^*_{k,b} = \mathbb{E}\#(k,b)/\mathbb{E}\#k$. In particular $p_{v^*} = p^{*(L)}$ so that $\text{KL}(p^*||\mathcal{M}_L) = \text{KL}(p^*||\tilde{\mathcal{M}}_L)$. One may check that l_N is C^∞, and, by seeing that it is a sum of convex functions, convex. Calling D^m the m-th derivative operator (D^0 the identity), $\|\cdot\|$ some norm on $\mathbb{R}^{\dim(\tilde{\Delta}_L(p^*))^m}$, and E some set whose closure is in the interior of $\tilde{\Delta}_L(p^*)$

$$\mathbb{E} \sup_{v \in E} \|D^m l_N(v)\| \leq \sum_{(k,b) \in \text{supp}_L(p^*)} \mathbb{E}\left[\#(k,b)\right] \sup_{v \in E} \|D^m \log v_{k,b}\| < \infty$$

since E is relatively compact. Thus, by theorem 1.3.3 of Ghosh and Ramamoorthi [22], $D^m l_N \to \mathbb{E} D^m l_1 = D^m l$ locally uniformly where the last equality is by Leibniz's rule due to the local boundedness of all derivatives. In particular, $D^3 l_N$ are uniformly bounded across N on a neighborhood of v^* and, sending $E \nearrow \tilde{\Delta}_L(p^*)$, and noting l_N is a.s. eventually $-\infty$ on the boundary of $\tilde{\Delta}_L(p^*)$, we see $l_N \to l$ pointwise a.s..

As in the analysis of Dawid [14], we write

$$\log p((X_n)_{n=1}^N | \tilde{\mathcal{M}}_L) = \log \frac{p((X_n)_{n=1}^N | \tilde{\mathcal{M}}_L)}{p_{v_N}(X_n)_{n=1}^N} + \log \frac{p_{v_N}(X_n)_{n=1}^N}{p^{*(L)}(X_n)_{n=1}^N} + \log p^{*(L)}(X_n)_{n=1}^N.$$

The above paragraph demonstrates that we satisfy conditions (2) of theorem 3.2 of Miller [43] and thus we can write

$$\log \frac{p((X_n)_{n=1}^N | \tilde{\mathcal{M}}_L)}{p_{v_N}(X_n)_{n=1}^N} = -\frac{1}{2} \dim\left(\tilde{\Delta}_L(p^*)\right) \log N + O(1)$$

and say that $v_N \to v^*$. Now, using the mean value theorem,

$$\log \frac{p_{v_N}(X_n)_{n=1}^N}{p^{*(L)}(X_n)_{n=1}^N} = -n(l_N(v_N) - l_N(v^*)) = -(\sqrt{N}(v^* - v_N))^T D^2 l_N(v')(\sqrt{N}(v^* - v_N))$$

for some v'_N on the ray connecting v^* and v_N. Call $Z_N = \sqrt{D^2 l_N(v'_N)}(\sqrt{N}(v^* - v_N))$. By local uniform convergence, since $v_N \to v^*$, $D^2 l_N(v'_N) \to D^2 l(v^*)$. Satisfying the conditions on a neighborhood of v^*, since $v_N \to v^*$, by theorem 5.41 in van der Vaart [64], $\sqrt{N}(v^* - v_N)$ converges in distribution to $N(0, D^2 l(v^*)^{-1})$. Thus, by Slutsky's theorem, Z_n converges to $N(0, I)$, and by the continuous mapping theorem $\log \frac{p_{v_n}(X_n)_{n=1}^N}{p_{v_0}(X_n)_{n=1}^N} = Z_n^T Z_n$ converges in distribution to $\chi^2_{\dim(\tilde{\Delta}_L(p^*))}$; thus this term is $O_P(1)$. Recall from the remark in the last paragraph that $\text{KL}(p^*||\tilde{\mathcal{M}}_L) = \text{KL}(p^*||\mathcal{M}_L)$ for all L; note in particular $p^* \in \tilde{\mathcal{M}}_L$ if and only if $p^* \in \mathcal{M}_L$. Then finally, by the analysis of Dawid [14], since $\mathbb{E}[\log p^{*(L)}(X_n)_{n=1}^N]^2 \leq (\log(\min_{k,b} v_{k,b}^{-1}))^2 \mathbb{E}|X|^2 < \infty$, $\log p^{*(L)}(X_n)_{n=1}^N = \log p^*(X_n)_{n=1}^N$ if $p^* \in \tilde{\mathcal{M}}_L$ and

$$\log p^{*(L)}(X_n)_{n=1}^N = N\left[-\text{KL}(p^*||\mathcal{M}_L) + \mathbb{E} \log p(X)\right] + O_P(\sqrt{N})$$

otherwise.

By our analysis above we can say that given $L_1 \neq L_2$, if $\text{KL}(p^*||\mathcal{M}_{L_2}) > \text{KL}(p^*||\mathcal{M}_{L_1})$,

$$\log \frac{p((X_n)_{n=1}^N | \tilde{\mathcal{M}}_{L_1})}{p((X_n)_{n=1}^N | \tilde{\mathcal{M}}_{L_2})} = N\left(\text{KL}(p^*||\mathcal{M}_{L_2}) - \text{KL}(p^*||\mathcal{M}_{L_1})\right) + O_p(\sqrt{N}). \tag{11}$$

Otherwise, if $p^* \in \mathcal{M}_{L_1}, \mathcal{M}_{L_2}$,

$$\log \frac{p((X_n)_{n=1}^N | \tilde{\mathcal{M}}_{L_1})}{p((X_n)_{n=1}^N | \tilde{\mathcal{M}}_{L_2})} = \frac{1}{2}\left(\dim(\tilde{\Delta}_{L_2}(p^*)) - \dim(\tilde{\Delta}_{L_1}(p^*))\right) \log N + O_p(1). \tag{12}$$

Moving to the second term, for a $k \in \text{supp}(v^*)$, by Stirling's formula,

$$\begin{aligned}
&\left(\log \frac{\Gamma(\sum_b \alpha_{k,b})}{\Gamma(\sum_b \alpha_{k,b} + \#k)} + \log \frac{\Gamma(\sum_b' \alpha_{k,b} + \#k)}{\Gamma(\sum_b' \alpha_{k,b})}\right) \\
&= \left(\sum_b \alpha_{k,b} - \frac{1}{2}\right) \log\left(\sum_b \alpha_{k,b}\right) - \sum_b \alpha_{k,b} \\
&\quad - \left(\#k + \sum_b \alpha_{k,b} - \frac{1}{2}\right) \log\left(\#k + \sum_b \alpha_{k,b}\right) + \left(\#k + \sum_b \alpha_{k,b}\right) \\
&\quad - \left(\sum_b{}' \alpha_{k,b} - \frac{1}{2}\right) \log\left(\sum_b{}' \alpha_{k,b}\right) + \sum_b{}' \alpha_{k,b} \\
&\quad + \left(\#k + \sum_b{}' \alpha_{k,b} - \frac{1}{2}\right) \log\left(\#k + \sum_b{}' \alpha_{k,b}\right) - \left(\#k + \sum_b{}' \alpha_{k,b}\right) \\
&\quad + O(1) \\
&= \left(\#k + \sum_b{}' \alpha_{k,b} - \frac{1}{2}\right) \log\left(\frac{\sum_b' \alpha_{k,b} + \#k}{\sum_b \alpha_{k,b} + \#k}\right) \\
&\quad - \left(\sum_b \alpha_{k,b} - \sum_b{}' \alpha_{k,b}\right) \log\left(\sum_b \alpha_{k,b} + \#k\right) + O(1) \\
&= \left(\#k + \sum_b{}' \alpha_{k,b} - \frac{1}{2}\right) O\left(\frac{1}{\#k}\right) \\
&\quad - \left(\sum_b \alpha_{k,b} - \sum_b{}' \alpha_{k,b}\right) \log \#k + O(1) \\
&= -\left(\sum_{b \notin \text{supp}(p^*)|_k} \alpha_{k,b}\right) \log \#k + O(1)
\end{aligned} \tag{13}$$

Now note $\log \#k = \log N + \log\left(\frac{1}{N}\#k\right) = \log N + O(1)$ by the strong law of large numbers. Putting this together with 11, 12, 10, and 13 gives the result. □

So far, we've studied pairwise comparisons between models with different lags; we now study the posterior over lags. We start with the case where there is no true data-generating lag, i.e. $p^* \notin \mathcal{M}$. In this case, we can apply theorem 7 to show that the posterior over lags diverges to infinity.

Corollary 8. *Let $\pi(L)$ denote a prior over lags, with $\pi(L) > 0$ for all L. Choose for each lag a Dirichlet prior on the simplex $\Delta_{\mathcal{B}}^{\mathcal{B}_L^o}$ that satisfies the conditions of Theorem 7. If p^* is subexponential but $p^* \notin \mathcal{M}$, the posterior diverges in the sense that for any choice of lag $\tilde{L}$, we have $\Pi(L > \tilde{L} | (X_n)_{n=1}^N) \to 1$ a.s..*

Proof. It is shown in the proof of theorem 23 that as $L \to \infty$, we have $\text{KL}(p^* \| \mathcal{M}_L) \to 0$. Say $\tilde{L}$ is a lag, so, since $p^* \notin \mathcal{M}_{\tilde{L}}$, there exists some $\tilde{L}' > \tilde{L}$ such that $\text{KL}(p^* \| \mathcal{M}_{\tilde{L}'}) < \text{KL}(p^* \| \mathcal{M}_{\tilde{L}}) \leq \text{KL}(p^* \| \mathcal{M}_L)$ for all $L \leq \tilde{L}$. Note we have

$$\Pi(L \leq \tilde{L} | (X_n)_{n=1}^N) \leq \frac{\sum_{L \leq \tilde{L}} p((X_n)_{n=1}^N | \mathcal{M}_{L'})}{\sum_{L \leq \tilde{L}} p((X_n)_{n=1}^N | \mathcal{M}_{L'}) + p((X_n)_{n=1}^N | \mathcal{M}_{\tilde{L}'})}.$$

There are only finitely many L' less than or equal to $\tilde{L}$, so we can apply theorem 7 and the conclusion follows. □

We now consider the case where $p^* \in \mathcal{M}$. Pick L^* to be the minimum lag such that $p^* \in \mathcal{M}_{L^*}$. We will need to assume, for theoretical tractability, that the prior over lags has finite support. Then we can establish sufficient conditions for the posterior to concentrate on the true value L^*.

Lemma 9. *Let $\pi(L)$ be a prior over lags with $\pi(L) > 0$ for all L less than some $\tilde{L} \geq L^*$, and with $\pi(L) = 0$ for all $L > \tilde{L}$. Then $\Pi(L^*|(X_n)_{n=1}^N) \to 1$ in probability if $(\dim_L^{\text{eff}}(p^*))_{L \geq L^*}$ is non-decreasing and $\dim_{L^*+1}^{\text{eff}}(p^*) > \dim_{L^*}^{\text{eff}}(p^*)$.*

Proof. Apply theorem 7. □

If transition probabilities $v^*_{k,b}$ were always non-zero, the effective dimension of the model would simply be the dimension of the parameter space $\Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}$, and thus the dimension would always increase with increasing lag, making lag selection consistent. Allowing for $v^*_{k,b} = 0$ makes the situation more complicated, since in fact the effective dimension may not increase with increasing lag. If this is indeed the case, the posterior will no longer be guaranteed to determine the true L^* from data, even asymptotically. In order to describe how the effective dimension in fact scales with the lag, we will introduce the notion of a distribution's de Bruijn graph: for a distribution p on S, the L-mer de Bruijn graph is the directed graph with nodes $\text{acc}_L(p)$ and a directed edge connecting L-mers $k \to k'$ if $k' = (k_{2:L}, b)$ for a $b \in \text{supp}_L(p)|_k$. (De Bruijn graphs are a common data analysis tool in biological sequence analysis, where they are typically constructed from an empirical distribution over observed sequences; here, we are in effect studying the asymptotic de Bruijn graph, i.e. the de Bruijn graph that we would have if an infinite amount of data were observed.) Call a de Bruijn graph a tree if every node has at most one parent (since sequences must start and end with start and stop symbols, there cannot be a loop where each kmer has just one parent). The next two results show that we can only consistently infer the true lag if the the L^*-mer de Bruijn graph of p^* is not a tree.

Proposition 10. *Say $p^* \in \mathcal{M}_{L^*}$ and for each L, consider a $\text{Dirichlet}(\alpha_{k,b})_{b \in \tilde{\mathcal{B}}}$ prior on the simplex in $\Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}$ corresponding to the L-mer k. Say for $L \geq L^*$, for all L-mers k and bases b, $\alpha_{k,b} = \alpha_{k_{L-L^*+1:L},b}$ (i.e. the prior concentration depends only on the last L^* letters of the L-mer). There exists a $\tilde{L}$ (possibly infinity) such that for all $L \geq L^*$, the L-mer de Bruijn graph is a tree if and only if $L > \tilde{L}$. Then $(\dim_L^{\text{eff}}(p^*))_{L \geq L^*}$ is a non-decreasing sequence, strictly increasing until $\tilde{L}$, and constant past $\tilde{L}$.*

Proof. Call v^* the transition coefficients of p^*. Say $L > L^*$, $k \in \text{acc}_L(p^*)$. Call $k' \in \text{acc}_{L^*}(p^*)$ the last L^* letters of k. If for some $b \in \tilde{\mathcal{B}}$, $p^*(\#(k,b) > 0) > 0$ then clearly $p^*(\#(k',b) > 0)$ thus $\text{supp}_L(p^*)|_k \subseteq \text{supp}_{L^*}(p^*)|_{k'}$. On the other hand, say $b \in \text{supp}_{L^*}(p^*)|_{k'} = \text{supp}(v^*)|_{k'}$ and Y is a string, not terminated with \$, and with its last L characters equal to k and $p^*(Y \dots)$. $p^*((Y,b) \dots \mid Y \dots) = v^*_{k',b} > 0$ so, $p^*(\#(k,b) > 0) > 0$. Thus $\text{supp}_L(p^*)|_k = \text{supp}_{L^*}(p^*)|_{k'}$.

Now write

$$\dim_L^{\text{eff}}(p^*) = \sum_{k \in \text{acc}_L(p^*)} \sum_{b \in \text{supp}_L(p^*)|_k} \left[\mathbb{1}_{b \in \text{supp}_L(p^*)|_k} + \mathbb{1}_{b \notin \text{supp}_L(p^*)|_k} \alpha_{k,b}\right] - 1$$

where, for a statement A, $\mathbb{1}_A = 1$ if A is true and $\mathbb{1}_A = 0$ if A is false. Thus, since in this case $\text{supp}_L(p^*)|_k = \text{supp}_{L^*}(p^*)|_{k'}$, and by the assumption on the prior coefficients,

$$\begin{aligned} \dim_L^{\text{eff}}(p^*) = &\sum_{k' \in \text{acc}_{L^*}(p^*)} |\{k \in \text{acc}_{L^*}(p^*) \mid k_{L-L^*+1:L} = k'\}| \\ &\times \left(\sum_{b \in \text{supp}_L(p^*)|'_k} \left[\mathbb{1}_{b \in \text{supp}_L(p^*)|_{k'}} + \mathbb{1}_{b \notin \text{supp}_L(p^*)|_{k'}} \alpha_{k',b}\right] - 1 \right). \end{aligned} \tag{14}$$

Since for each $k' \in \text{acc}_{L^*}(p^*)$ there is a $k \in \text{acc}_L(p^*)$ that has its last L^* letters equal to k', $\dim_L^{\text{eff}}(p^*) \geq \dim_{L^*}^{\text{eff}}(p^*)$. Since $p^* \in \mathcal{M}_L$ for all $L \geq L^*$ the argument may be repeated for all pairs $L_1 > L_2 \geq L^*$ to conclude $(\dim_L^{\text{eff}}(p^*))_{L \geq L^*}$ is non-decreasing.

Note if for $L' > L$, $\dim^{\text{eff}}_{L'}(p^*) = \dim^{\text{eff}}_{L}(p^*)$ then for all $k' \in \text{acc}_L(p^*)$ there is a unique $k \in \text{acc}_{L'}(p^*)$ with its last L letters equal to k. Thus if $X_1, X_2 \in S$ with $p^*(X_1), p^*(X_2) > 0$ and X_1, X_2 end in the same last L letters (not including \$), then X_1, X_2 end in the same last L' letters. Looking at positions $|X_j| - L' : |X_j| - L' + L - 1$, one can also conclude that X_1, X_2 end in the same last $L' + (L' - L)$ letters. Continuing, one may conclude $X_1 = X_2$. It can be seen that this is equivalent to the L-mer de Bruijn of p^* being a tree. On the other hand it is not difficult to see that if the L-mer de Bruijn of p^* is a tree then $\dim^{\text{eff}}_{L'}(p^*) = \dim^{\text{eff}}_{L}(p^*)$ for all $L' > L$. □

Corollary 11. *Say $p^* \in \mathcal{M}$ and L^* is the minimum lag such that $p^* \in \mathcal{M}_{L^*}$. Let $\pi(L)$ be a prior over lags with $\pi(L) > 0$ for all L less than some $\tilde{L} \geq L^*$, and with $\pi(L) = 0$ for all $L > \tilde{L}$. For each L, consider a* $\text{Dirichlet}(\alpha_{k,b})_{b \in \tilde{\mathcal{B}}}$ *prior on the simplex in $\Delta^{\mathcal{B}^o_L}_{\tilde{\mathcal{B}}}$ corresponding to the L-mer k. Assume that for $L \geq L^*$, for all L-mers k and bases b, $\alpha_{k,b} = \alpha_{k_{L-L^*+1:L},b}$. Then lag selection is consistent if and only if the L^*-mer de Bruijn graph of p^* is not a tree.*

Remark 1. If $p^*(X) > 0$ for infinitely many $X \in S$, as is the case if the transition coefficients of p^* are all positive or there is a cycle in the L^*-mer de Bruijn graph of p^*, then no L-mer de Bruijn graph of p^* is a tree as sequences with $p(X) > 0$ cannot be identified by their last L letters. As another example, pick a particular sequence $X \in S$ and say X' is one letter away from X. For a $0 < q < 1$, define $p = q\delta_X + (1-q)\delta_{X'}$. Pick L^* the smallest lag such that $p^* \in \mathcal{M}_{L^*}$. Then the L^*-mer de Bruijn graph splits into two paths at the position where X and X' differ. These paths may rejoin after L^* nodes. Thus the L^*-mer de Bruijn graph is a tree if and only if the position at which X and X' differ is less than L^* letters away from the end symbol \$.

F Misspecification detection

In this section, we turn from studying the parameter v and lag L in the BEAR model to studying the hyperparameters h and θ. Intuitively, we expect the empirical Bayes estimate of h to behave as a diagnostic of misspecification, since h controls the extent to which the prior predictive distribution of the BEAR model is concentrated at the embedded AR model. Here we make this idea rigorous by examining the asymptotic behavior of the empirical Bayes estimates of h and θ.

We first briefly introduce the setup and some notation. We will assume p^* is subexponential. We will work with fixed lag L, though the results can be straightforwardly extended to the case of a prior over a finite number of lags. The function $f : \Theta \mapsto \Delta^{\mathcal{B}^o_L}_{\tilde{\mathcal{B}}}$ defines an autoregressive model, with parameter space Θ some set. For any $h > 0, \theta \in \Theta$, define a prior $\pi(\cdot|h,\theta)$ on $\Delta^{\mathcal{B}^o_L}_{\tilde{\mathcal{B}}}$ consisting of independent $\text{Dirichlet}(\frac{1}{h} f_{k,b}(\theta))_{b \in \tilde{\mathcal{B}}}$ priors on each simplex corresponding to $k \in \mathcal{B}^o_L$. Define $m((X_n)_{n=1}^N | h, \theta)$ to be the marginal likelihood of the data $(X_n)_{n=1}^N$ under the prior $\pi(\cdot|h,\theta)$, that is $m((X_n)_{n=1}^N|h,\theta) = \int p_v((X_n)_{n=1}^N)\pi(v|h,\theta)$. For our purposes we may assume $f_{k,b}(\theta) > 0$ for all $(k,b) \in \text{supp}_L(p^*)$; if this is not the case for some θ then the marginal likelihood at θ, for any choice of h, is a.s. eventually 0. We will study maximum marginal likelihood/empirical Bayes estimates $(h_N, \theta_N) = \text{argmax}_{h,\theta} m((X_n)_{n=1}^N|h,\theta)$.

Our starting point is the analysis of empirical Bayes presented in Petrone et al. [48]. Here is the (very heuristic) intuition behind their result: the Laplace approximation to the marginal likelihood is proportional to the probability of the true data-generating parameter under the prior, so asymptotically we expect $m((X_n)_{n=1}^N|h,\theta) \propto \pi(v^*|h,\theta)$. Then, roughly speaking, the empirical Bayes estimate will be $(h_N, \theta_N) \approx \text{argmax}_{h,\theta}\pi(v^*|h,\theta)$; in other words, the empirical Bayes estimate should asymptotically maximize the probability of the true parameter parameter value under the prior. Petrone et al. [48] give conditions under which this is indeed true, but BEAR models fail to meet them. There are two major problems: (1) in the limit as $h \to 0$, the prior converges to a point mass, making the Laplace approximation invalid (the "degenerate" case mentioned by Petrone et al. [48]) and (2) when some transitions have probability zero, $v^*_{k,b} = 0$, the standard Laplace approximation does not hold regardless of the value of h. Our analysis in this section adjusts for both these issues, and also provides more detailed insight such as convergence rates and intuitive approximations for the optimal h.

In analyzing extremum estimators, such as the maximum marginal likelihood estimator used in empirical Bayes, uniform convergence results are particularly powerful. Ideally, we might try to establish a Laplace-like approximation to the marginal likelihood that holds uniformly for all h and θ, but this is unavailable because of the degeneracy at $h = 0$. Our strategy will be to first demonstrate a uniform Laplace approximation over all h, θ with some caveats: (1) we ignore transitions that are not possible under p^* and analyze their contribution to the likelihood later; (2) if $h \to 0$ we assume it does not decrease too fast; and (3) we assume similar control over the prior density at the "true" transition probabilities v^*. In proposition 13 we prove that (3) must indeed hold for when h_N, θ_N are the maximizers of the marginal likelihood.

For any $v \in \tilde{\Delta}_L(p^*)$, define the negative average log likelihood $l_N(v) = -\frac{1}{N} \log p_v(X_n)_{n=1}^N$, and let $v_N \in \tilde{\Delta}_L(p^*)$ be the (a.s. eventually unique) maximizer of l_N. Define a prior $\tilde{\pi}(\cdot|h, \theta)$ on $\tilde{\Delta}_L(p^*)$ consisting of independent Dirichlet$(\frac{1}{h} f_{k,b}(\theta))_{b \in \text{supp}_L(p^*)|_k}$ priors on each simplex corresponding to $k \in \text{acc}_L(p^*)$ (for a scalar α, Dirichlet(α) is just defined as the point mass on the 0-dimensional simplex $\{1\}$). Let $\tilde{m}((X_n)_{n=1}^N|h, \theta)$ denote the marginal likelihood under the prior $\tilde{\pi}(\cdot|h, \theta)$ and define

$$\log r_N(h, \theta) = \sum_{k \in \text{acc}_L(p^*)} \left(\log \frac{\Gamma(\sum_b \frac{1}{h} f_{k,b}(\theta))}{\Gamma(\sum_b \frac{1}{h} f_{k,b}(\theta) + \#k)} - \log \frac{\Gamma(\sum_b' \frac{1}{h} f_{k,b}(\theta))}{\Gamma(\sum_b' \frac{1}{h} f_{k,b}(\theta) + \#k)} \right)$$

where $\sum_b'$ is a sum over the $b \in \text{supp}_L(p^*)|_k$. So, as shown in theorem 7, $\log m((X_n)_{n=1}^N|h, \theta) = \log \tilde{m}((X_n)_{n=1}^N|h, \theta) + \log r_N(h, \theta)$. define $B(v, \eta)$ to be the ball of radius η around v in some norm; finally, define $B_{\text{KL}}(\eta) = \{v \in \tilde{\Delta}_L(p^*) \mid \mathbb{E} \log \frac{p^{*(L)}(X)}{p_v(X)} < \eta\}$ and, for convenience $B(\eta) = B(v^*, \eta)$, for any $\eta > 0$.

Theorem 12. *With probability* 1, *for any sequence* $(h_N)_N$ *and* $(\theta_N)_N$, *possibly dependent on the data, if* $h_N N^{1/4-\epsilon} \to \infty$ *for an* $1/4 > \epsilon > 0$ *and* $\liminf(\log \tilde{\pi}(v^*|h_N, \theta_N))/\sqrt{N} \neq -\infty$, *then*

$$\left| \log \tilde{m}((X_n)_{n=1}^N|h_N, \theta_N) - \left(-N l_N(v_N) - \frac{1}{2} \dim \tilde{\Delta}_L(p^*) \log N + \log \tilde{\pi}(v^*|h_N, \theta_N) + C_{v^*} \right) \right| \to 0$$

for a fixed C_{v^*} *dependent only on* v^*.

Proof. First note, calling $e_{k,b}$ the indicator vector at position k, b for some $k \in \text{acc}_L(p^*), b, b' \in \text{supp}_L(p^*)|_k$, the directional derivatives with respect to v

$$D_{e_{k,b} - e_{k,b'}} \log \tilde{\pi}(v|h, \theta) = \frac{\frac{1}{h} f_{k,b}(\theta) - 1}{v_{k,b}} - \frac{\frac{1}{h} f_{k,b'}(\theta) - 1}{v_{k,b'}}$$

are bounded by J/h, for some $J > 0$ in a neighborhood of v^* for all θ.

For an $\eta > 0$, define the KL ball

$$\hat{B}_{\text{KL}}(\eta) = \{v \in \tilde{\Delta}_L(p^*) \mid v_{k,b} \geq v^*_{k,b}(1 - \eta/\mathbb{E}|X|) \ \forall k, b\}.$$

Note if $v \in \hat{B}_{\text{KL}}(\eta)$, then the KL divergence is bounded,

$$\mathbb{E} \log \frac{p^{*(L)}(X)}{p_v(X)} \leq (\mathbb{E}|X|) \sup_{k,b} \log \frac{v^*_{k,b}}{v_{k,b}} \leq \eta$$

so $v \in B_{\text{KL}}(\eta)$. Note

$$(w_{k,b})_{(k,b) \in \text{supp}_L(p^*)} \mapsto \left(\frac{\eta}{\mathbb{E}|X|} w_{k,b} + v^*_{k,b} \left(1 - \frac{\eta}{\mathbb{E}|X|} \right) \right)_{(k,b) \in \text{supp}_L(p^*)}$$

is a diffeomorphism from $\tilde{\Delta}_L(p^*)$ to $\hat{B}_{\text{KL}}$ so by the change of variables theorem the volume of $\hat{B}_{\text{KL}}$ is $(\eta/\mathbb{E}|X|)^{\dim \tilde{\Delta}_L(p^*)}$ (which comes from the factor multiplying $w_{k,b}$) times the volume of $\tilde{\Delta}_L(p^*)$. Finally note that by an application of the triangle inequality, $\hat{B}_{\text{KL}}(\eta) \subset B(2\eta \text{diam}(\tilde{\Delta}_L(p^*))/\mathbb{E}|X|)$.

Define the information matrix at $\tilde{v}^*$, $\mathcal{I} = \mathbb{E}[D^2 l_1(\tilde{v}^*)]$, and an $\epsilon' > 0$ less than the smallest eigenvalue of $\mathcal{I}$ ($\mathcal{I}$ is positive definite by the strict convexity of l_0 described in theorem 7). Also pick an $\epsilon'' < \frac{1}{8}\epsilon'$ such that $\hat{B}_{\text{KL}}(\epsilon''\eta^2) \subset B(\eta)$ for all small η. Now define a sequence $\eta_N = N^{-(1/4-\epsilon)}$ noting $\eta_N/h_N \to 0$. Let $|\mathcal{I}|$ denote the determinant of the information matrix.

$$
\begin{aligned}
&\left| \log \tilde{m}((X_n)_{n=1}^N | h_N, \theta_N) - \left(-N l_N(v_N) - \frac{1}{2} \dim \tilde{\Delta}_L(p^*) \log(2\pi N) - \frac{1}{2} \log |\mathcal{I}| + \log \tilde{\pi}(v^* | h_N, \theta_N) \right) \right| \\
&\leq \left| \log \left(\int_{\tilde{\Delta}_L(p^*)} e^{-N l_N(v)} \tilde{\pi}(v | h_N, \theta_N) \right) - \log \left(\int_{B(\eta_N)} e^{-N l_N(v)} \tilde{\pi}(v | h_N, \theta_N) \right) \right| \\
&\quad + \left| \log \left(\int_{B(\eta_N)} e^{-N l_N(v)} \tilde{\pi}(v | h_N, \theta_N) \right) - \log \left(\int_{B(\eta_N)} e^{-N l_N(v)} \tilde{\pi}(v^* | h_N, \theta_N) \right) \right| \\
&\quad + \left| \log \left(\int_{B(\eta_N)} e^{-N l_N(v)} \tilde{\pi}(v^* | h_N, \theta_N) \right) - \log \left(\int_{B(v_N, \eta_N)} e^{-N l_N(v)} \tilde{\pi}(v^* | h_N, \theta_N) \right) \right| \\
&\quad + \Bigg| \log \left(\int_{B(v_N, \eta_N)} e^{-N l_N(v)} \tilde{\pi}(v^* | h_N, \theta_N) \right) \\
&\qquad - \left(-N l_N(v_N) - \frac{1}{2} \dim \hat{\Delta}_L(p^*) \log(2\pi N) - \frac{1}{2} \log |\mathcal{I}| + \log \tilde{\pi}(v^* | h_N, \theta_N) \right) \Bigg| \\
&\leq \log \left(1 + \left(\int_{\tilde{\Delta}_L(p^*) \setminus B(\eta_N)} e^{N l_N(v_N) - N l_N(v)} \tilde{\pi}(v | h_N, \theta_N) \right) / \left(\int_{B(\eta_N)} e^{N l_N(v_N) - N l_N(v)} \tilde{\pi}(v | h_N, \theta_N) \right) \right) \\
&\quad \left| \log \left(\left(\int_{B(\eta_N)} e^{-N l_N(v)} \frac{\tilde{\pi}(v | h_N, \theta_N)}{\tilde{\pi}(v^* | h_N, \theta_N)} \right) / \left(\int_{B(\eta_N)} e^{-N l_N(v)} \right) \right) \right| \\
&\quad + \log \left(\left(\int_{B(v_N, \eta_N + \|v_N - v^*\|)} e^{-N l_N(v)} \right) / \left(\int_{B(v_N, \eta_N - \|v_N - v^*\|)} e^{-N l_N(v)} \right) \right) \\
&\quad + \left| \log \left((2\pi)^{-\frac{1}{2} \dim \hat{\Delta}_l(p^*)} |\mathcal{I}|^{-1/2} \int_{\|y\| < \eta_N \sqrt{N}} e^{N(l_N(v_N) - l_N(v_N + y/\sqrt{N}))} \right) \right| \\
&\leq \exp \left(N \sup_{\|v^* - v\| > \eta_N} (l_N(v^*) - l_N(v)) \right) / \left(\int_{\hat{B}_{\text{KL}}(\epsilon'' \eta_N^2)} e^{N l_N(v^*) - N l_N(v)} \tilde{\pi}(v | h_N, \theta_N) \right) \\
&\quad + \sup_{v \in B(\eta_N)} |\log \tilde{\pi}(v | h_N, \theta_N) - \log \tilde{\pi}(v^* | h_N, \theta_N)| \\
&\quad + \left(\int_{B(v_N, \eta_N + \|v_N - v^*\|) \setminus B(v_N, \eta_N - \|v_N - v^*\|)} e^{-N l_N(v)} \right) / \left(\int_{B(v_N, \eta_N - \|v_N - v^*\|)} e^{-N l_N(v)} \right) \\
&\quad + \left| \log \left((2\pi)^{-\frac{1}{2} \dim \hat{\Delta}_l(p^*)} |\mathcal{I}|^{-1/2} \int_{\|y\| < \eta_N \sqrt{N}} e^{N(l_N(v_N) - l_N(v_N + y/\sqrt{N}))} \right) \right|.
\end{aligned}
\tag{15}
$$

The third line in this inequality follows since $B(v_N, \eta_N - \|v_N - v^*\|) \subseteq B(v_N, \eta_N) \cap B(\eta_N)$ and $B(v_N, \eta_N) \cup B(\eta_N) \subseteq B(v_N, \eta_N + \|v_N - v^*\|)$. First note that the second term is bounded by $J\eta_N/h_N$ and thus vanishes a.s.. We will show the rest of these terms also vanish a.s..

To analyze the last term, we will use a simplified proof of a Laplace approximation. First note, given the regularity conditions established in the proof of theorem 7, a.s. $v_N \to v^*$, and $D^2 l_N \to D^2 E l_N$ locally uniformly. Thus, for each y, since $\eta_N \sqrt{N} \to \infty$, and and $\eta_N \to 0$ (so that if $\|y\| < \eta_N \sqrt{N}$ then $y/\sqrt{N} \leq \eta_N \to 0$), a.s.

$$
\mathbb{1}_{\|y\| < \eta_N \sqrt{N}} e^{N(l_N(v_N) - l_N(v_N + y/\sqrt{N}))} = \mathbb{1}_{\|y\| < \eta_N \sqrt{N}} e^{-\frac{1}{2} y^T D^2 l_N(v'_N) y} \to e^{-\frac{1}{2} y^T \mathcal{I} y},
$$

where v'_N is on a ray connecting v_N to $v_N + y/\sqrt{N}$. As well, eventually,

$$\mathbb{1}_{\|y\|<\eta_N\sqrt{N}} e^{N(l_N(v_N)-l_N(v_N+y/\sqrt{N}))} = \mathbb{1}_{\|y\|<\eta_N\sqrt{N}} e^{-\frac{1}{2}y^T D^2 l_N(v'_N)y} \leq e^{-\frac{1}{4}y^T \mathcal{I} y}.$$

The right hand side is integrable and takes the form of a Gaussian pdf. Thus, integrating the Gaussian pdf, the last term of equation 15 goes to 0 a.s. by the dominated convergence theorem.

To analyze the third term of equation 15, recall from the proof of 7 that l_N is convex, so, the value of $-Nl_N$ is less on the annulus $B(v_N, \eta_N + \|v_N - v^*\|) \setminus B(v_N, \eta_N - \|v_N - v^*\|)$ than on $B(v_N, \eta_N - \|v_N - v^*\|)$. Thus, to demonstrate that this term vanishes, it suffices to show that $\|v_N - v^*\|/\eta_N \to 0$ a.s.. Recall from the proof of 7 that we showed that a.s. $v_N \to v^*$ and $D^2 l_N$ converges to $\mathbb{E}D^2 l_1$ uniformly in a neighborhood of v^*. Thus, eventually, recalling the definition of ϵ' as less than the minimal eigenvalue of $\mathcal{I}$, and defining $t \mapsto v_t$ as a linear path from v_N to v^*,

$$\|Dl_N(v^*)\| = \|Dl_N(v^*) - Dl_N(v_N)\| = \left\| \left(\int_0^1 dt D^2 l_N(v_t) \right) (v^* - v_N) \right\| \geq \frac{1}{2}\epsilon' \|v^* - v_N\|.$$

On the other hand, defining $e_{k,b}$ as above, $|D_{e_{k,b}-e_{k,b'}} l_1(v^*)| \leq |X|/\inf_{k,b} v^*_{k,b}$ and so, $D_{e_{k,b}-e_{k,b'}} l_1(v^*)$ is subexponential. Recalling $\mathbb{E}Dl_1(v^*) = D\mathbb{E}l_1(v^*) = 0$, using Bernstein's inequality (theorem 2.8.1 in Vershynin [65]),

$$p^*(|D_{e_{k,b}-e_{k,b'}} l_N(v^*)| > \eta_N^2) \leq C\exp(-C'N\eta_N^4) \leq C\exp(-C'N^{4\epsilon}).$$

Since $\sum_{N=1}^{\infty} C\exp(-C'N^{4\epsilon}) \lesssim \int_0^{\infty} dx \exp\left(-C'x^{4\epsilon}\right) < \infty$, by the Borel-Cantelli lemma, a.s. eventually, $\|Dl_N(v^*)\| \leq C\eta_N^2$ for some $C > 0$. Finally, since $\eta_N \to 0$, we have $\|v_N - v^*\|/\eta_N \to 0$ a.s..

To analyze the first term of equation 15 first note that for small enough η_N, recalling that $\mathbb{E}l_N$ is convex with maximum at v^*, and by the definition of ϵ', we can Taylor expand around v^* and find

$$\sup_{\|v^*-v\|>\eta_N} (\mathbb{E}l_N(v^*) - \mathbb{E}l_N(v)) = \sup_{\|v^*-v\|=\eta_N} (\mathbb{E}l_N(v^*) - \mathbb{E}l_N(v)) \leq -1/2\epsilon'\eta_N^2.$$

We will also show below that a.s. eventually, for all v away from the boundary (i.e. outside a fixed neighborhood of the boundary), $|l_N(v) - \mathbb{E}l_N(v)| < \frac{1}{16}\epsilon'\eta_N^2$. For now, assume that this is the case. So, a.s. eventually, $\sup_{\|v^*-v\|>\eta_N}(l_N(v^*) - l_N(v)) < -3/8\epsilon'\eta_N^2$, by the triangle inequality. Having bounded the numerator, we now turn to the denominator. Note that by equi-continuity, since $J\eta_N/h_N$ is eventually less than $\log 2$, $\tilde{\pi}(v|h_N, \theta_N) \geq \frac{1}{2}\tilde{\pi}(v^*|h_N, \theta_N)$ for all $v \in B(\eta_N)$. As well, again, by a triangle inequality, a.s. eventually, for all $v \in B_{\mathrm{KL}}(\epsilon''\eta_N^2)$, $l_N(v^*) - l_N(v) \geq -\epsilon''\eta_N^2 - \frac{1}{8}\epsilon'\eta_N^2 \geq -\frac{1}{4}\epsilon'\eta_N^2$. Recall that the volume of $\hat{B}_{\mathrm{KL}}(\epsilon''\eta_N^2)$ is equal to $C(C'\eta_N^2)^{\dim \hat{\Delta}_L(p^*)}$ for some $C, C' > 0$. Then the first term of equation 15 is bounded above by

$$2C\exp\left(-\frac{1}{8}\epsilon' N\eta_N^2 + 2\dim \tilde{\Delta}_L(p^*)\log\left(\eta_N^{-1}\right) - \log\tilde{\pi}(v^*|h_N, \theta_N)\right)$$

for some $C > 0$. This expression goes to 0 as $\log\tilde{\pi}(v^*|h_N, \theta_N)/\sqrt{N}$ is bounded below and thus $\liminf \log\tilde{\pi}(v^*|h_N, \theta_N)/N^{1/2+2\epsilon} = 0$.

We now show that a.s. eventually, for all v away from the boundary, $|l_N(v) - \mathbb{E}l_N(v)| < \frac{1}{16}\epsilon'\eta_N^2$. First write

$$D_{e_{k,b}-e_{k,b'}} l_N(v) = \frac{1}{N}\#(k,b)v_{k,b}^{-1} - \frac{1}{N}\#(k,b')v_{k,b'}^{-1}$$

which is almost surely eventually bounded by the strong law of large numbers for all v away from the boundary of $\tilde{\Delta}_L(p^*)$. The derivatives of $\mathbb{E}l_N$ with respect to v are similarly bounded away from the boundary; say the derivatives of both functions are eventually bounded by J'. Also note that the random variables $|l_1(v)(X)| \leq C''|X|$ are uniformly sub-exponential for all v away from the boundary. The covering number of $\tilde{\Delta}_L(p^*)$ by balls

of radius $\frac{1}{64}J'^{-1}\epsilon'\eta_N^2$ is $\lesssim \eta_N^{-2\dim \hat{\Delta}_L(p^*)}$. Say $(v_i)_i$ are centers of the balls of such a covering. By uniform sub-exponentiality and Bernstein's inequality (theorem 2.8.1 in **?**]), for small enough η_N, $P(|l_N(v_i) - \mathbb{E}l_N(v_i)| > \frac{1}{32}\epsilon'\eta_N^2) \lesssim \exp(-CN\eta_N^4) = \exp(-CN^{4\epsilon})$ for some $C > 0$. Now, for some $C, C' > 0$,

$$\begin{aligned}
&\sum_{N=0}^{\infty} P(\text{there is a } v_i \text{ such that } |l_N(v_i) - \mathbb{E}l_N(v_i)| > \frac{1}{32}\epsilon'\eta_N^2) \\
&\qquad \leq \sum_{N=0}^{\infty}\sum_i P(|l_N(v_i) - \mathbb{E}l_N(v_i)| > \frac{1}{32}\epsilon'\eta_N^2) \\
&\qquad \lesssim \sum_{N=0}^{\infty} \exp\left(-CN^{4\epsilon} - 2\dim \hat{\Delta}_L(p^*)\log\eta_N\right) \qquad (16)\\
&\qquad \lesssim \sum_{N=0}^{\infty} \exp\left(-C'N^{4\epsilon}\right) \\
&\qquad \lesssim \int_0^{\infty} dx \exp\left(-C'x^{4\epsilon}\right) < \infty.
\end{aligned}$$

By the Borel-Cantelli lemma, $|l_N(v_i) - \mathbb{E}l_N(v_i)| \leq \frac{1}{32}\epsilon'\eta_N^2$ for all i a.s. eventually. Thus, eventually, by the triangle inequality and the a.s. eventual boundedness of the derivatives of l_N and $\mathbb{E}l_N$, $|l_N(v) - \mathbb{E}l_N(v)| \leq \frac{1}{16}\epsilon'\eta_N^2$ for all v away from the boundary a.s. eventually. □

We now focus on the behavior of not just any sequence of h_N, θ_N, but rather specifically on h_N, θ_N which maximize the marginal likelihood.[3] The next two results both use a proof by contradiction strategy that relies on the following logic.

Remark 2. Fix h, θ. We showed in theorem 7 that $\log r_N(h,\theta) = O(\log N)$ a.s. and we can conclude from theorem 12 that $\log \tilde{m}((X_n)_{n=1}^N \mid h, \theta) = -Nl_N(v_N) - O(\log(N))$. Thus, $m((X_n)_{n=1}^N \mid h, \theta) = -Nl_N(v_N) - O(\log(N))$. On the other hand, for any h', θ', $\log r_N(h', \theta') \leq 0$ and $\log \tilde{m}((X_n)_{n=1}^N \mid h', \theta') \leq -Nl_N(v_N)$. Thus for the maximizers of m, h_N, θ_N, it is a contradiction if $\log r_N(h_N, \theta_N) \lesssim -N^\beta$ or $\log \tilde{m}((X_n)_{n=1}^N \mid h_N, \theta_N) \leq -Nl_N(v_N) - CN^\beta$ for any $\beta > 0$: say $\log \tilde{m}(h_N, \theta_N) \leq -Nl_N(v_N) - N^\beta$. Then, for some $C > 0$, $-C\log(N) \leq m((X_n)_{n=1}^N |h, \theta) + Nl_N(v_N) \leq m((X_n)_{n=1}^N |h_N, \theta_N) + Nl_N(v_N) \leq \log \tilde{m}((X_n)_{n=1}^N |h_N, \theta_N) + Nl_N(v_N) \leq -CN^\beta$, a contradiction. On the other hand, say $\log r_N(h_N, \theta_N) \lesssim -N^\beta$. Then $-C\log(N) \leq m((X_n)_{n=1}^N |h, \theta) + Nl_N(v_N) \leq m((X_n)_{n=1}^N |h_N, \theta_N) + Nl_N(v_N) \leq \log r_N(h_N, \theta_N) \leq -C'N^\beta$, also a contradiction.

Proposition 13. *Say $(h_N)_N$ and $(\theta_N)_N$ are sequences maximizing $\log m((X_n)_{n=1}^N | h_N, \theta_N)$ for each N. Then a.s. there is no subsequence $(h_{N_j})_j$ and $(\theta_{N_j})_j$ such that for some $\epsilon > 0$, $h_{N_j}N_j^{1/4-\epsilon} \to \infty$ and for some $\beta > 0$, $\lim \log \tilde{\pi}(v^*|h_{N_j}, \theta_{N_j})/N_j^\beta < 0$.*

Proof. Assume the opposite. Define $(v_N)_N$ and pick $(\eta_N)_N, \epsilon'$ as in theorem 12 such that a.s. eventually, for all v away from the boundary, $|l_N(v) - \mathbb{E}l_N(v)| < \frac{1}{16}\epsilon'\eta_N^2$, $\eta_{N_j}/h_{N_j} \to 0$, and $\inf_{\|v^*-v\|>\eta_N} \mathbb{E}l_N(v) \geq \mathbb{E}l_N(v_N) + \frac{1}{2}\epsilon'\eta_N^2$. Then, eventually,

$$\begin{aligned}
\int_{B(\eta_{N_j})^C} e^{-N_j l_{N_j}(v)}\tilde{\pi}(v|h_{N_j}, \theta_{N_j}) &\leq \exp\left(-N_j \inf_{\|v^*-v\|>\epsilon} l_{N_j}(v)\right) \\
&\leq \exp\left(-N_j(l_{N_j}(v_{N_j}) + \frac{3}{8}\epsilon'\eta_{N_j}^2)\right) \qquad (17)\\
&\leq \exp\left(-N_j l_{N_j}(v_{N_j}) - \frac{3}{8}\epsilon' {N_j}^{1/4}\right).
\end{aligned}$$

[3] It is not crucial that maximizers of the marginal likelihood exist for any of the result below: the results below hold assuming only that h_N, θ_N are approximate maximizers, i.e. $\log m((X_n)_{n=1}^N \mid h_N, \theta_N) = \sup_{h,\theta} \log m((X_n)_{n=1}^N \mid h, \theta) + o(1)$ or in slightly altered form swapping the $o(1)$ for $o_P(1)$.

where $B(\eta_{N_j})^C$ denotes the complement of $B(\eta_{N_j})$. On the other hand, by equi-continuity of the prior density, since η_{N_j}/h_{N_j} becomes small, for some $C > 0$

$$\begin{aligned}\int_{B(\eta_{N_j})} e^{-N_j l_{N_j}(v)}\tilde{\pi}(v|h_{N_j},\theta_{N_j}) &\lesssim \exp\left(-N_j l_{N_j}(v_{N_j}) + \log\tilde{\pi}(v^*|h_{N_j},\theta_{N_j}) + \dim\tilde{\Delta}_L(p^*)\log(\eta_{N_j})\right)\\ &\leq \exp\left(-N_j l_{N_j}(v_{N_j}) - C{N_j}^{\beta} + O(\log N_j))\right)\end{aligned} \tag{18}$$

for some $C > 0$. By remark 2, this completes the proof. □

We have so far explored what happens to the marginal likelihood when h_N does not converge quickly to 0, showing that it satisfies a Laplace-like approximation in this case. Next we show that h_N will in fact converge to zero quickly only if the estimated autoregressive model $f(\theta_N)$ converges to the optimal parameter value v^*.

For a sequence $(\theta_N)_N$ define, for $k \in \mathrm{acc}_L(p^*)$, $\sigma_{N,k} = \sum_{b\in\mathrm{supp}_L(p^*)|_k} f_{k,b}(\theta_N)$ and $\lambda_{N,k} = 1 - \sigma_{N,k}$.

Proposition 14. *Say $(h_N)_N$ and $(\theta_N)_N$ are sequences maximizing $\log m(\{X_n\}_{n=1}^N|h_N,\theta_N)$. Then a.s., $\limsup h_{N_j}N_j^{\beta} < \infty$ for some $\beta > 0$ along a subsequence $(N_j)_j$ only if $f_{k,b}(\theta_{N_j}) \to v^*_{k,b}$ for all $k, b \in \mathrm{supp}_L(p^*)$.*

Proof. Take a subsequence such that: $h_{N_j} \to 0$; $h_{N_j}N_j^{\beta}$ and $h_{N_j}N_j$ both converge, the latter possibly to ∞; $f_{k,b}(\theta_{N_j})$ converges for all k, b; and $f_{k,b}(\theta_{N_j})/h_{N_j}$ converges, possibly to ∞, for all k, b. Note since $[0,\infty]$ is compact, every subsequence with $\limsup h_{N_j}N_j^{\beta} < \infty$ has a further subsequence with these properties. Thus it will be sufficient to show that $f_{k,b}(\theta_{N_j}) \to v^*_{k,b}$ for all $k \in \mathrm{acc}_L(p^*), b \in \tilde{\mathcal{B}}$. Now define $\lambda_k = \lim\lambda_{N_j,k}$ and σ_k similarly for all $k \in \mathrm{acc}_L(p^*)$.

The proof will proceed in two parts. First we will show that if $\lambda_k \neq 0$ for some $k \in \mathrm{acc}(p^*)$, then $\log r_{N_j}(h_{N_j},\theta_{N_j}) \lesssim -N_j^{\beta'}$ for some $\beta' > 0$. This is a contradiction by remark 2 so that $\lambda_k = 0$ and $\sigma_k = 1$ for all k. Then we will show that if $f_{k,b}(\theta_{N_j}) \not\to v^*_{k,b}$ for any $k, b \in \mathrm{supp}_L(p^*)$, eventually $\sup_{v\in B(\eta)}\log\tilde{\pi}(v|h_{N_j},\theta_{N_j}) \lesssim -N_j^{\beta''}\left(\|f(\theta_{N_j}) - v^*\| - \eta\right)^2$. for some $\beta'' > 0$ for small η. Assume this is the case for now. By similar logic to that in equation 17 of proposition 13, for small fixed η, it can be seen that for some $\beta''', C, C' > 0$,

$$\log\int_{B(\eta)^C} e^{-N_j l_{N_j}(v)}\tilde{\pi}(v|h_{N_j},\theta_{N_j}) \leq -N_j l_{N_j}(v_{N_j}) - CN^{\beta'''}.$$

As well,

$$\begin{aligned}\log\int_{B(\eta)} e^{-N_j l_{N_j}(v)}\tilde{\pi}(v|h_{N_j},\theta_{N_j}) &\leq -N_j l_{N_j}(v_{N_j}) + \sup_{\|v^*-v\|<\eta}\log\tilde{\pi}(v \mid h_{N_j},\theta_{N_j})\\ &\leq -N_j l_{N_j}(v_{N_j}) - C'{N_j}^{\beta''}.\end{aligned}$$

using the fact that $\log\tilde{\pi}(v|h_{N_j},\theta_{N_j}) \lesssim -{N_j}^{\beta''}$. This is also a contradiction by remark 2 and the statement of the theorem follows.

Part one: Assume that for some k', $\lambda_{k'} > 0$. Performing the Stirling approximation on the terms of $\log r_{N_j}$ depends on the behavior of $\sigma_{N_j,k}/h_{N_j}$. Based on the properties of the subsequence we chose, this quantity converges. If it converges to a number greater than or equal to 1 we can perform the usual Stirling approximation with $O(1)$ error. On the other hand, if $\sigma_{N_j,k}/h_{N_j}$ has limit less than 1, using the properties of the Gamma function we write

$$\begin{aligned}\log\Gamma\left(\frac{\sigma_{N_j,k}}{h_{N_j}}\right) &= -\log\left(\frac{\sigma_{N_j,k}}{h_{N_j}}\right) + \log\Gamma\left(1 + \frac{\sigma_{N_j,k}}{h_{N_j}}\right)\\ &= \left(\frac{\sigma_{N_j,k}}{h_{N_j}} - 1\right)\log\left(\frac{\sigma_{N_j,k}}{h_{N_j}}\right) - \frac{\sigma_{N_j,k}}{h_{N_j}} + O(1)\end{aligned} \tag{19}$$

where additional $O(1)$ terms were added explicitly in the second line so that the approximation is similar in form to the usual Stirling approximation with the exception of a 1 in the first term instead of 1/2. Define $\gamma_k = 1/2$ if the limit of $\frac{\sigma_{N_j,k}}{h_{N_j}}$ is greater than or equal to 1 and 1 otherwise. Finally recall that $h_{N_j} \to 0$ and write

$$
\begin{aligned}
\log r_{N_j}(h_{N_j}, \theta_{N_j}) &= \sum_{k \in \mathrm{acc}_L(p^*)} \left[\log \frac{\Gamma\left(\frac{1}{h_{N_j}}\right)}{\Gamma\left(\frac{1}{h_{N_j}} + \#k\right)} - \log \frac{\Gamma\left(\frac{\sigma_{N_j,k}}{h_{N_j}}\right)}{\Gamma\left(\frac{\sigma_{N_j,k}}{h_{N_j}} + \#k\right)} \right] \\
&= \sum_{k \in \mathrm{acc}_L(p^*)} \left[\left(\frac{1}{h_{N_j}} - \frac{1}{2}\right) \log\left(\frac{1}{h_{N_j}}\right) \right. \\
&\quad - \left(\frac{1}{h_{N_j}} + \#k - \frac{1}{2}\right) \log\left(\frac{1}{h_{N_j}} + \#k\right) \\
&\quad - \left(\frac{\sigma_{N_j,k}}{h_{N_j}} - \gamma_k\right) \log\left(\frac{\sigma_{N_j,k}}{h_{N_j}}\right) \\
&\quad \left. + \left(\frac{\sigma_{N_j,k}}{h_{N_j}} + \#k - \frac{1}{2}\right) \log\left(\frac{\sigma_{N_j,k}}{h_{N_j}} + \#k\right) \right] + O(1) \\
&= \sum_{k \in \mathrm{acc}_L(p^*)} \left[-\frac{\lambda_{N_j,k}}{h_{N_j}} \log(1 + h_{N_j} \#k) \right. \\
&\quad - \left(\frac{\sigma_{N_j,k}}{h_{N_j}} - \frac{1}{2}\right) \log\left(\sigma_{N_j,k}\right) \\
&\quad \left. + \left(\frac{\sigma_{N_j,k}}{h_{N_j}} + \#k - \frac{1}{2}\right) \log\left(\frac{\sigma_{N_j,k} + \#k h_{N_j}}{1 + \#k h_{N_j}}\right) \right] \\
&\quad + \sum_{k \in \mathrm{acc}_L(p^*)} (\gamma_k - 1/2) \log\left(\frac{\sigma_{N_j,k}}{h_{N_j}}\right) + O(1) \\
&= \sum_{k \in \mathrm{acc}_L(p^*)} \frac{1}{h_{N_j}} \left[-\lambda_{N_j,k} \log(1 + \#k h_{N_j}) - \sigma_{N_j,k} \log(\sigma_{N_j,k}) \right. \\
&\quad \left. + \left(\sigma_{N_j,k} + \#k h_{N_j}\right) \log\left(\frac{\sigma_{N_j,k} + \#k h_{N_j}}{1 + \#k h_{N_j}}\right) \right] \\
&\quad + \sum_{\sigma_{N_j,k}/h_{N_j} \to 0} (\gamma_k - 1/2) \log\left(\frac{\sigma_{N_j,k}}{h_{N_j}}\right) + O(1) \\
&\leq \sum_{k \in \mathrm{acc}_L(p^*)} \frac{1}{h_{N_j}} \left[-\lambda_{N_j,k} \log(1 + \#k h_{N_j}) - \sigma_{N_j,k} \log(\sigma_{N_j,k}) \right. \\
&\quad \left. + \left(\sigma_{N_j,k} + \#k h_{N_j}\right) \log\left(\frac{\sigma_{N_j,k} + \#k h_{N_j}}{1 + \#k h_{N_j}}\right) \right] + O(1)
\end{aligned}
\tag{20}
$$

The function

$$
x \mapsto -\lambda_{N_j,k} \log(1 + x) - \sigma_{N_j,k} \log(\sigma_{N_j,k}) + \left(\sigma_{N_j,k} + x\right) \log\left(\frac{\sigma_{N_j,k} + x}{1 + x}\right)
$$

has intercept 0, and derivative $\log\left(\frac{\sigma_{N_j,k}+x}{1+x}\right)$, and is thus convex since the derivative is increasing (Fig S2). As $x \to \infty$, the function is $-\lambda_{N_j,k} \log x + O(1)$ while the function has tangent $x \mapsto x \log \sigma_{N_j,k}$ at $x = 0$. In our case, we evaluate at $x = h_{N_j} N_j$, which, based on the chosen subsequence, is either bounded or goes to infinity. First assume $h_{N_j} N_j$ is bounded, say by M, and recall that we assumed $\lambda_{k'} > 0$ for some k', so $\sigma_{k'} < 1$. Then, because the

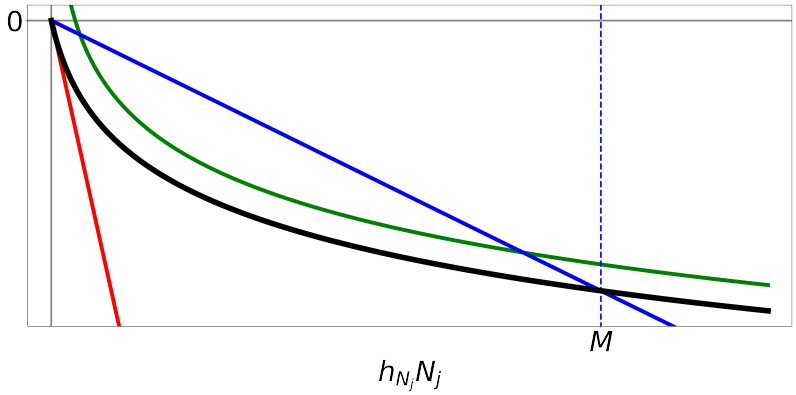

Figure S2: Graph of the function evaluated at $h_{N_j}N_j$ in black when $\sigma_{N_j,k} < 1$. The red line shows the tangent at 0 with slope $\log(\sigma_{N_j,k}) < 0$. The blue line shows that in this case, where $\sigma_{N_j,k} < 1$, the function may be dominated by some line for all values less than M. The green line shows that as $h_{N_j}N_j \to \infty$, the function is $-\lambda_{N_j,k}\log(h_{N_j}N_j) + O(1)$.

function is decreasing and eventually has negative derivative at 0, we can eventually bound it on $[0, M]$ by a line with negative slope and intercept 0 (Fig S2), so eventually, for some $C, C' > 0$,

$$\log r_{N_l}(h_{N_j}, \theta_{n_l}) \leq -C\frac{1}{h_{N_j}}N_j h_{N_j} + C' \lesssim -N_j.$$

Otherwise $h_{N_j}N_j \to \infty$ so, by the above remark about the limits of the function as $x \to \infty$,

$$\log r_{N_l}(h_{N_j}, \theta_{n_l}) \leq -\frac{1}{2h_{N_j}}\log(h_{N_j}N_j)\sum_{k\in \mathrm{acc}_L(p^*)}\lambda_{N_j,k} + C$$

for some $C > 0$ eventually. Recalling that $h_{N_j}N_j^\beta$ is eventually bounded above, and by assumption $\log(h_{N_j}N_j) \to \infty$,

$$\log r_{N_l}(h_{N_j}, \theta_{N_l}) \lesssim -N_j^\beta\frac{\log(h_{N_j}N_j)}{h_{N_j}N_j^\beta}\max_k \lambda_k \lesssim -N_j^\beta \max_k \lambda_k.$$

This completes part one of the proof.

Part two: Assume $\|f_{k,b}(\theta_{N_j}) - \tilde{v}_k\| \not\to 0$. We will perform the same technique to allow a Stirling approximation of the prior: define $\gamma_{k,b} = 1/2$ if the limit of $f_{k,b}(\theta_{N_j})/h_{N_j}$ is greater than or equal to 1 and 1 otherwise. Then, for all $v \in \tilde{\Delta}_L(p^*)$ away from the boundary,

recalling that we showed in part 1 $\sigma_{N_j,k} \to 1$ for all k, if $\frac{f_k(\theta_{N_j})}{\sigma_{N_j,k}} \neq v_k$ for some k,

$$\begin{aligned}
\log \tilde{\pi}(v|h_{N_j}, \theta_{N_j}) =& \sum_k \log \Gamma\left(\frac{\sigma_{N_j,k}}{h_{N_j}}\right) \\
&- \sum_{b \in \mathrm{supp}_L(p^*)|_k} \left[\log \Gamma\left(\frac{1}{h_{N_j}} f_{k,b}(\theta_{N_j})\right) - \frac{1}{h_{N_j}} f_{k,b}(\theta_{N_j}) \log v_{k,b}\right] + O(1) \\
=& \sum_k \left(\frac{\sigma_{N_j,k}}{h_{N_j}} - 1/2\right) \log\left(\frac{\sigma_{N_j,k}}{h_{N_j}}\right) \\
&- \sum_{b \in \mathrm{supp}_L(p^*)|_k} \left[\left(\frac{1}{h_{N_j}} f_{k,b}(\theta_{N_j}) - \gamma_{k,b}\right) \log\left(\frac{1}{h_{N_j}} f_{k,b}(\theta_{N_j})\right)\right. \\
&\qquad \left. - \frac{1}{h_{N_j}} f_{k,b}(\theta_{N_j}) \log v_{k,b}\right] + O(1) \\
=& \frac{1}{2} \dim \tilde{\Delta}_L(p^*) \log\left(\frac{1}{h_{N_j}}\right) - \frac{1}{h_{N_j}} \sum_k \sigma_{N_j,k} \mathrm{KL}\left(\frac{f_k(\theta_{N_j})}{\sigma_{N_j,k}} \middle\| v_k\right) \\
&+ \sum_{k,b \text{ s.t. } \gamma_{k,b}=1} \left(\gamma_{k,b} - \frac{1}{2}\right) \log\left(\frac{1}{h_{N_j}} f_{k,b}(\theta_{N_j})\right) + O(1) \\
\lesssim& - \frac{1}{h_{N_j}} \sum_k \mathrm{KL}\left(\frac{f_k(\theta_{N_j})}{\sigma_{N_j,k}} \middle\| v_k\right).
\end{aligned} \tag{21}$$

Now note $h_{N_j} \lesssim N^{-\beta}$ and for any norm $|\cdot|$, by Pinsker's inequality,

$$\sum_k \mathrm{KL}\left(\frac{f_k(\theta_{N_j})}{\sigma_{N_j,k}} \middle\| v_k\right) \gtrsim \sum_k \left\| \frac{f_k(\theta_{N_j})}{\sigma_{N_j,k}} - v_k \right\|^2.$$

One may check that $\left(\sum_k \|\cdot\|^2\right)^{1/2}$ is also a norm and $\sigma_{N_j,k} \to 1$ for all k, so

$$\sum_k \mathrm{KL}\left(\frac{f_k(\theta_{N_j})}{\sigma_{N_j,k}} \middle\| v_k\right) \gtrsim \left\| f(\theta_{N_j}) - v \right\|^2 + o(1)$$

for any norm $\|\cdot\|$. Now note if $\eta < \|f(\theta_{N_j}) - v^*\|$,

$$\sup_{v \in B(\eta)} \log \tilde{\pi}(v|h_{N_j}, \theta_{N_j}) \lesssim -N_j^\beta \left(\|f(\theta_{N_j}) - v^*\| - \eta\right)^2.$$

This concludes part two. □

We now have the tools to determine the behavior of h_N and $f(\theta_N)$ in the well and misspecified cases.

F.1 The well-specified case

We now examine the asymptotic behavior of empirical Bayes inference for the BEAR model in the well-specified case, or, more precisely, when the model is well-specified "at resolution L", in the sense that there are $\tilde{\theta}_N$ such that for all $k, b \in \mathrm{supp}_L(p^*)$, $f_{k,b}(\tilde{\theta}_N) \to v^*_{k,b}$ (we say the model is misspecified at resolution L otherwise). We first show that the misspecification diagnostic is guaranteed to converge to zero ($h_N \to 0$), correctly indicating that the model is well-specified, and that the embedded AR model converges to the true transition probabilities ($f(\theta_N) \to v^*$). We also give a bound on the rate for the convergence of h_N, a power of the dataset size. We then establish additional weak conditions under which θ_N also converges to the true value θ^*.

Proposition 15. *Say the model is well-specified and $(h_N)_N$ and $(\theta_N)_N$ are sequences maximizing $\log m(\{X_n\}_{n=1}^N | h_N, \theta_N)$. Then $h_N N^{1/4-\epsilon} \to 0$ for every $\epsilon > 0$ and $f_{k,b}(\theta_N) \to v^*_{k,b}$ for all $k, b \in \text{supp}_L(p^*)$ with both sequences converging in probability.*

Proof. If U is a neighborhood of v^* and $\beta > 0$, proposition 14 shows that

$$p^*(h_N < N^{-\beta}, f(\theta_N) \notin U) \to 0$$

(otherwise $p^*(h_N < N^{-\beta}, f(\theta_N) \notin U \text{ for infinitely many } N) > 0$). We show below that $p^*(h_N \geq N^{-1/4+\epsilon}) \to 0$ for any $\epsilon > 0$ and it will thus follow that we also get $f(\theta) \to v^*$ in probability.

Proposition 13 shows that

$$p^*(h_N \geq N^{-1/4+\epsilon}, \log \tilde{\pi}(v^* | h_N, \theta_N) < -\sqrt{N}) \to 0$$

as $h_N \leq N^{-1/4+\epsilon}$ if and only if $h_N N^{1/4-\epsilon/2} \geq N^{\epsilon/2}$. Thus it is sufficient to show that

$$p^*(h_N \geq N^{-1/4+\epsilon}, \log \tilde{\pi}(v^* | h_N, \theta_N) \geq -\sqrt{N}) \to 0.$$

On this set, we may apply theorem 12, but we will need to control $\log \tilde{\pi}(v^* | h, \theta)$.

For any h, θ, defining $\gamma_{k,b} = 1$ if $\frac{1}{h} f_{k,b}(\theta) < 1$ and $1/2$ otherwise, and $\hat{\gamma}_k = 1$ if $\frac{\sigma_k}{h} < 1$ (where recall $\sigma_k = \sum_{b \in \text{supp}_L(p^*)|_k} f_{k,b}(\theta)$) and $1/2$ otherwise, by the same derivation as equation 21,

$$\begin{aligned} \log \tilde{\pi}(v^* | h, \theta) =& \frac{1}{2} \dim \tilde{\Delta}_L(p^*) \log\left(\frac{1}{h}\right) - \frac{1}{h} \sum_k \sigma_k \text{KL}\left(\frac{f_k(\theta)}{\sigma_{N_j,k}} \middle\| v^*_k\right) \\ &+ \sum_{k,b \text{ s.t. } \gamma_{k,b}=1} \left(\gamma_{k,b} - \frac{1}{2}\right) \log\left(\frac{1}{h} f_{k,b}(\theta)\right) \\ &- \sum_{k, \text{ s.t. } \hat{\gamma}_k = 1} \left(\hat{\gamma}_k - \frac{1}{2}\right) \log\left(\frac{\sigma_k}{h}\right) + O(1) \end{aligned} \tag{22}$$

where $O(1)$ is uniform over h or θ. Since $\hat{\gamma}_k = 1$ only if $\gamma_{k,b} = 1$ for all $b \in \text{supp}_L(p^*)|_k$, by the concavity of the log function, the sum of these last two terms is negative. Thus,

$$\log \tilde{\pi}(v^* | h, \theta) \leq \frac{1}{2} \dim \tilde{\Delta}_L(p^*) \log\left(\frac{1}{h}\right) + C \tag{23}$$

for all h, θ for some $C > 0$.

Now we derive a lower bound for $\tilde{m}((X_n)_{n=1}^N | h_N, \theta_N)$. Pick $\tilde{\theta}_j$ such that for all $k, b \in \text{supp}_L(p^*)$, $f_{k,b}(\tilde{\theta}_j) \to v^*_{k,b}$. Thus, $\tilde{\pi}(\cdot | h, \tilde{\theta}_j) \to \prod_{k \in \text{acc}_L(p^*)} \text{Dirichlet}(\frac{1}{h} v^*_{k,b})_{b \in \text{supp}_L(p^*)|_k}$ for any $h > 0$ in distribution. And as $h \to 0$, we also have $\prod_{k \in \text{acc}_L(p^*)} \text{Dirichlet}(\frac{1}{h} v^*_{k,b})_{b \in \text{supp}_L(p^*)|_k} \to \delta_{v^*}$. So, pick a sequence $\tilde{\theta}'_j, \tilde{h}_j$ such that $\tilde{\pi}(\cdot | \tilde{h}_j, \tilde{\theta}'_j) \to \delta_{v^*}$ in distribution.[4] Then $\log m((X_n)_{n=1}^N | \tilde{h}_j, \tilde{\theta}'_j) \to -N l_N(v^*)$. Thus, $\log m((X_n)_{n=1}^N | h_N, \theta_N) \geq -N l_N(v^*)$. Also recall that from the proof of theorem 7 that, defining $Z_N = N l_N(v_n) - N l_N(v^*)$, Z_N converges in distribution (to a chi-squared distribution). Since $\log r_N \leq 0$ we can write

$$\log \tilde{m}((X_n)_{n=1}^N | h_N, \theta_N) \geq -N l_N(v_N) + Z_N. \tag{24}$$

Now, when both $h_N \geq N^{-1/4+\epsilon}, \log \tilde{\pi}(v^* | h_N, \theta_N) \geq -\sqrt{N}$, applying theorem 12, we've shown that with probability going to 1, for some fixed $C > 0$,

$$\log \tilde{m}((X_n)_{n=1}^N | h_N, \theta_N) \leq -N l_N(v_N) - \frac{1}{2} \dim \hat{\Delta}_L(p^*) \log N + \frac{1}{2} \dim \tilde{\Delta}_L(p^*) \log\left(\frac{1}{h}\right) + C.$$

[4] Since $\tilde{\Delta}_L(p^*)$ is compact, the set of polynomials with rational coefficients, $(g_i)_{i=1}^\infty$ is dense in the space of continuous functions under the infinite norm. Pick $\tilde{h}_j$ to have $|g_i(v^*) - \int g_i d \prod_{k \in \text{acc}_L(p^*)} \text{Dirichlet}(\frac{1}{h} v^*_{k,b})_{b \in \text{supp}_L(p^*)|_k}| < 1/j$ for all $i \leq j$ and then $\tilde{\theta}'_j$ to have $|\int g_i d \prod_{k \in \text{acc}_L(p^*)} \text{Dirichlet}(\frac{1}{h} v^*_{k,b})_{b \in \text{supp}_L(p^*)|_k} - \int g_i d\tilde{\pi}(\cdot | h_j, \tilde{\theta}'_j)| < 1/j$ for all $i \leq j$.

Thus, as $h_N \geq N^{-1/4+\epsilon}$,

$$\begin{aligned} -\frac{1}{4}\dim \hat{\Delta}_L(p^*)\log N + C \geq & -\frac{1}{2}\dim \hat{\Delta}_L(p^*)\log N + (1/4-\epsilon)\frac{1}{2}\dim \tilde{\Delta}_L(p^*)\log N + C \\ & \geq \log \tilde{m}((X_n)_{n=1}^N | h_N, \theta_N) + N l_N(v_N) \\ & \geq Z_N. \end{aligned} \tag{25}$$

Since Z_N converges in distribution, this occurs with vanishing probability. □

We have thus far discussed the asymptotic behavior of h_N and $f(\theta_N)$. To draw conclusions about θ_N itself, we need to place some assumptions on the autoregressive function f. Here we provide an example of such assumptions, drawn from the theory of M-estimators, which say in essence that f must have an isolated peak at θ^*. These assumptions are enough to guarantee that the empirical Bayes estimate of the AR model parameter θ converges to the true value θ^*.

Corollary 16. *Say $\theta^* \in \Theta$ and d is a metric on Θ such that $f_{k,b}(\theta^*) = v^*_{k,b}$ for all $k, b \in \text{supp}_L(p^*)$ and for all $\delta > 0$,*

$$0 < \inf_{d(\theta,\theta^*)>\delta} \|f(\theta) - v^*\|.$$

Then $\theta_N \to \theta^*$ in probability.

Proof. Since by proposition 15 we have $\|f(\theta_N) - v^*\| = o_P(1)$, we may apply theorem 5.7 of van der Vaart [64] to get the result. □

Taking a step back, a perhaps surprising aspect of these results is the weak conditions on f. Were we, instead of trying to diagnose misspecification in the AR model, simply trying to analyze uncertainty in the AR model's parameter estimate, we might proceed by putting a prior on θ and performing Bayesian inference for the AR model. In this case, to guarantee asymptotic normality and well-calibrated frequentist coverage, we would in general need strong conditions on f, such as bounded third derivatives [43]. Intuitively, the task of diagnosing misspecification might seem to be harder than describing parameter uncertainty, but our conditions on f in this section and the next are in fact much weaker, involving no restrictions on the derivatives of f whatsoever.

F.2 The misspecified case

We now consider the case where the AR model is misspecified at resolution L. In this case, we can rewrite the marginal likelihood of the BEAR model (using propositions 13 and 14 to apply theorem 12) as

$$\log m((X_n)_{n=1}^N | h_N, \theta_N) = -N l_N(v_N) - \frac{1}{2}\dim \tilde{\Delta}_L(p^*)\log N + C_{v^*} - \mathcal{L}_N(h_N, \theta_N) + o(1)$$

where we define $\mathcal{L}_N(h_N, \theta_N) = -\log \tilde{\pi}(v^* | h, \theta) - r_N(h, \theta)$.[5] This expression for the marginal likelihood takes the form of a modified Laplace approximation where, instead of the original prior π evaluated at the true parameter value, we have the prior over the support of the data, $\tilde{\pi}(v^*|h,\theta)$, as well as the additional term r_N, which is $O(\log N)$ rather than $O(1)$ and depends on the concentration of the prior outside the support of the data. Instead of the standard empirical Bayes behavior described by Petrone et al. [48], wherein the prior probability of the true parameters is maximized, we instead heuristically expect that the objective function $\mathcal{L}_N(h,\theta)$ is minimized. The following result makes this intuition formal, showing that h_N and θ_N indeed behavior similarly to the minimizers of $\mathcal{L}_N$.

Corollary 17. *If the model is misspecified at resolution L, a.s. $\mathcal{L}_N(h_N, \theta_N) = \sup_{h,\theta} \mathcal{L}_N(h,\theta) - o(1)$.*

[5] $\mathcal{L}_N$ is stochastic due to r_N, but since $h_N N^\beta \to \infty$ for any $\beta > 0$, using the expansion in equation 20, one may show that the $\#k$ in r_N can be replaced with $N\mathbb{E}\#k$ incurring only a penalty of $O_P(N^{-1/2+\epsilon})$.

Proof. Say $\hat{h}_N, \hat{\theta}_N$ are sequences such that $\mathcal{L}_N(\hat{h}_N, \hat{\theta}_N) = \sup_{h,\theta} \mathcal{L}_N(h,\theta) - o(1)$. For fixed h, θ, we have $\mathcal{L}_N(h,\theta) = O(\log N)$. Thus, for any $\beta > 0$ we clearly have $\liminf(\log \tilde{\pi}(v^* \mid \hat{h}_N, \hat{\theta}_N))/N^\beta \geq 0$ and since we are in the misspecified case, following the logic of proposition 14, equation 20 may be used to see that we also have $\hat{h}_N N^\beta \to \infty$. Thus theorem 12 may be applied to $\hat{h}, \hat{\theta}$ and a comparison of the Laplace approximations of $m((X_n)_{n=1}^N | \hat{h}_N, \hat{\theta}_N)$ and $m((X_n)_{n=1}^N | h_N, \theta_N)$ gives the result. □

We next examine in greater detail the behavior of the misspecification diagnostic h_N, along with the AR parameter estimate θ_N. There are two cases to consider. First, if the support of the AR model matches the support of the data-generating distribution (that is, $\text{supp}(f(\theta)) = \text{supp}_L(p^*)$ for all θ), then $r_N = 0$ and $\mathcal{L}_N = -\log \tilde{\pi}(v^* | h, \theta)$; we thus recover the standard empirical Bayes behavior of Petrone et al. [48], with h_N and θ_N asymptotically maximizing the prior probability of the true parameter value. In this case we find that h_N converges to a finite positive value. The second case to consider is when $\text{supp}_L(p^*) \subsetneq \text{supp}(f(\theta))$. Here, we have $r_N \neq 0$, and in particular $r_N(h,\theta) \approx -\frac{1}{h}\log(N) \sum_k \lambda_k(\theta)$. In this case we find that $h_N \to \infty$. Thus, in either case, $h_N \not\to 0$, and so h_N will correctly diagnose misspecification in the AR model.

Corollary 18. *If the model is misspecified at resolution L but $\text{supp}(f(\theta)) = \text{supp}_L(p^*)$ for all θ, h_N is eventually bounded above and below.*

Proof. Recall from proposition 14 that if $h \to 0$, $\log \tilde{\pi}(v^* | h, \theta) \leq -C\frac{1}{h} \inf_\theta \|f(\theta) - v^*\|$ for some $C > 0$. This expression diverges to $-\infty$ as $h \to 0$. We also showed in proposition 15 that $\log \tilde{\pi}(v^* | h, \theta) \leq \frac{1}{2} \dim \tilde{\Delta}_L(p^*) \log(1/h) + C$ for some $C > 0$. This expression also diverges as $h \to \infty$. Combining these two observations along with corollary 17 we get the result. □

To say something about θ_N, due to corollary 17, we may use the theory of extremum estimators we can apply theorem 5.7 of van der Vaart [64], replacing limits in probability with a.s. limits to get

Corollary 19. *Say the model is misspecified at resolution L but $\text{supp}(f(\theta)) = \text{supp}_L(p^*)$ for all θ. Say also that $\theta^* \in \Theta$, $h^* > 0$ and d is a metric on Θ such that for every $\delta > 0$,*

$$\log \tilde{\pi}(v^* | h^*, \theta^*) > \sup_{|h-h^*| \vee d(\theta,\theta^*) > \delta} \log \tilde{\pi}(v^* | h, \theta).$$

Then $\theta_N \to \theta^*$ and $h_N \to h^*$ a.s..

Now we consider the case where the support do not match, i.e. $\inf_\theta \max_k \lambda_k(\theta) > 0$, where $\lambda_k(\theta) = \sum_{b \notin \text{supp}_L(p^*)|_k} f_{k,b}(\theta)$.

Proposition 20. *If the model is misspecified at resolution L, $\text{supp}_L(p^*) \subsetneq \text{supp}(f(\theta))$ for all θ, and $\inf_\theta \max_k \lambda_k(\theta) > 0$, then $h_N \to \infty$.*

Proof. We first show h_N is a.s. bounded below. Since $h_N N^\beta \to \infty$ for all $\beta > 0$, if $h_{N_j} \to 0$ for some subsequence, we showed in proposition 14 that a.s. $\log r_{N_j}(h_{N_j}, \theta_{N_j}) \leq -C\frac{\log(h_{N_j} N_j)}{2h_{N_j}} \inf_\theta \max_k \lambda_{k,\theta} + C' \leq -C''\frac{\log(N_j)}{2h_{N_j}} \inf_\theta \max_k \lambda_{k,\theta} + C'$ for some $C, C', C'' > 0$. In particular, $\log r_{N_j} \lesssim -O(\log(N))$ but $\log r_{N_j} \not\sim -O(\log(N))$ if $h_{N_j} \to 0$. Thus, since $\log r_N(h,\theta) \geq -C\log(N)$ for fixed h, θ, for some $C > 0$ dependent on h, θ and $\tilde{\pi}$ also diverges as $h \to 0$, the assumption that h_N maximizes the marginal likelihood is contradicted. Thus, $h_N \not\to 0$. In particular, we showed in proposition 15 (equation 23) that $\log \tilde{\pi}(v^* | h, \theta) \leq \frac{1}{2} \dim \tilde{\Delta}_L(p^*) \log(1/h) + C$ for some $C > 0$ so we get that $\log \tilde{\pi}(v^* | h_N, \theta_N)$ is bounded above a.s..

Assume h_N is bounded above; we will show that this leads to a contradiction. Define $\gamma_{N,k} = 1/2$ if $\sigma_k(\theta_N)/h_N \geq 1$ and $\gamma_{N,k} = 1$ otherwise. Define $\hat{\gamma}_{N,k}$ similarly for $1/h_N$ alone. We next perform the same trick as in proposition 14, expanding $\Gamma(\frac{1}{h_N})$ in the form of a

Stirling approximation, to analyze r_N further. Noting that $\log(h_N N) \to \infty$, we have a.s.,

$$\begin{aligned}
\log r_N(h_N, \theta_N) = \sum_{k \in \mathrm{acc}_L(p^*)} \frac{1}{h_N} \Bigg[& -\lambda_k(\theta_N) \log(1 + \#k h_N) - \sigma_{N,k} \log(\sigma_k(\theta_N)) \\
& + (\sigma_k(\theta_N) + \#k h_N) \log\left(\frac{\sigma_k(\theta_N) + \#k h_N}{1 + \#k h_N}\right) \Bigg] \\
& + \sum_{k \in \mathrm{acc}_L(p^*)} (\gamma_{N,k} - 1/2) \log\left(\frac{\sigma_k(\theta_N)}{h_N}\right) \\
& - \sum_{k \in \mathrm{acc}_L(p^*)} (\hat{\gamma}_{N,k} - 1/2) \log\left(\frac{1}{h_N}\right) + O(1) \qquad (26) \\
= & -\frac{\log(h_N N)}{h_N} \sum_{k \in \mathrm{acc}_L(p^*)} \Big[\lambda_k(\theta_N) + o(1) \Big] \\
& + \sum_{k \in \mathrm{acc}_L(p^*)} (\gamma_{N,k} - 1/2) \log\left(\frac{\sigma_k(\theta_N)}{h_N}\right) \\
& - \sum_{k \in \mathrm{acc}_L(p^*)} (\hat{\gamma}_{N,k} - 1/2) \log\left(\frac{1}{h_N}\right) + O(1).
\end{aligned}$$

Note $\hat{\gamma}_{N,k} = 1$ only if $\gamma_{N,k} = 1$ so that the the sum of these last two terms is negative. So, since h_N is bounded above, $\log r_N(h_N, \theta_N) \leq -C \log(N) \inf_\theta \max_k \lambda_k(\theta)$ for some $C > 0$. Thus, since we also have that $\log \tilde{\pi}(v^* | h_N, \theta_N)$ is bounded above a.s., we get that $\mathcal{L}_N(h_N, \theta_N) \gtrsim \log(N)$ a.s.. On the other hand, with fixed θ, if $\hat{h}_N \to \infty$ (so we still have $\log(h_N N) \to \infty$), then

$$\log r_N(\hat{h}_N, \theta) = -\frac{\log(N)}{\hat{h}_N} \sum_{k \in \mathrm{acc}_L(p^*)} \Big[\lambda_k(\theta) + o(1) \Big] + \frac{1}{2} \sum_{k \in \mathrm{acc}_L(p^*)} \log\left(\sigma_k(\theta)\right) + O(1)$$

which is $-o(\log N)$, where we wrote $\log(\hat{h}_N)/\hat{h}_N = o(1)$. Now pick $\hat{h}_N$ increasing slowly so that $\mathcal{L}_N(\hat{h}_N, \theta) = o(\log(N))$. This is eventually less than $\mathcal{L}_N(h_N, \theta_N)$, a contradiction. Thus, $h_N \to \infty$. □

We can also study the behavior of θ_N in this mismatched supports case, using again the theory extremum estimators. We briefly outline the strategy, omitting details. Further analysis of equations 22 and 26 gives an objective, as $h \to \infty$, [6]

$$\mathcal{L}(h, \theta) = -\frac{\log(N)}{h} \left(\sum_k \lambda_k + o(1) \right) - \dim \tilde{\Delta}_L(p^*) \log h + (1 + o(1)) \sum_{k,b \in \mathrm{supp}_L(p^*)} \log(f_{k,b}(\theta)) + C + o(1)$$

for some fixed $C > 0$. Careful analysis of the $o(1)$ terms shows that h approaches $\frac{\log(N) \sum_k \lambda_k}{\dim \tilde{\Delta}_L(p^*)}$. Plugging this value of h in, the objective becomes

$$\mathcal{L}(h, \theta) = -\dim \tilde{\Delta}_L(p^*) \log \sum_k \lambda_k + \sum_{k,b \in \mathrm{supp}_L(p^*)} \log(f_{k,b}(\theta)) + C_N + o(1)$$

for some constant C_N dependent only on N and p^*. One can then see that θ_N is an M-estimator of $\dim \tilde{\Delta}_L(p^*) \log \sum_k \lambda_k + \sum_{k,b \in \mathrm{supp}_L(p^*)} \log(f_{k,b}(\theta))$ and apply a similar analysis as in corollary 19.

So far we have seen that $h_N \not\to 0$ when the AR model is misspecified at resolution L, but exactly what value will h_N take and what can it tell us about the amount of misspecification?

[6] Note that the KL term in $\tilde{\pi}$ can be dominated by $\sum_{k,b \in \mathrm{supp}_L(p^*)} \log(f_{k,b}(\theta))$.

Here we analyze the objective $\mathcal{L}_N$ heuristically to address these questions. From the expansions in proposition 14, we can write, for reasonable values of h, θ, assuming not too much misspecification,

$$\log \tilde{\pi}(v^*|h,\theta) \approx \frac{1}{2} \dim \tilde{\Delta}_L(p^*) \log\left(\frac{1}{h}\right) - \frac{1}{h} \sum_{k \in \mathrm{acc}_L(p^*)} \mathrm{KL}(f_k(\theta) \| v_k^*)$$

$$\log r_N(h,\theta) \approx -\frac{\log(N)}{h} \sum_{k \in \mathrm{acc}_L(p^*)} \lambda_k(\theta).$$

We see, then, that θ_N and h_N depend on an unconventional but valid divergence between the AR model and $p^{*(L)}$: the sum of the KL divergence between the AR model transition probabilities (from kmers that occur with non-zero probability) and the true transition probabilities, plus a penalty proportional to $\log(N)$ when the support of the AR model does not match the support of p^*. We can thus interpret h_N not only as a diagnostic of misspecification, but also as a measurement of the *amount* of misspecification, and make comparisons between different AR models on the basis of their h_N values.

G Hypothesis testing

In this section we use the results of the above sections to develop goodness-of-fit and two sample tests.

G.1 Goodness-of-fit test

Say p^* is a distribution on S with $\mathbb{E}|X|^2 < \infty$ and say $X_1, X_2, \cdots \sim p^*$ iid. Say $\tilde{p}$ is another distribution on S with $\mathbb{E} \log^2 \tilde{p}(X) < \infty$ where the expectation is with respect to p^*. We are interested in testing whether or not $p^* = \tilde{p}$, so we will consider the Bayes factor

$$\mathrm{BF}_L = \frac{\tilde{p}(X_n)_{n=1}^N}{p((X_n)_{n=1}^N \mid \mathcal{M}_L)}.$$

This test asks whether or not $\tilde{p}$ approximates p^* at least as well as the optimal model in $\mathcal{M}_L$. We can use it in particular to test whether $\tilde{p}$ matches the data-generating distribution p^* at resolution L, that is, whether $\tilde{p}$ matches $p^{*(L)}$.

Proposition 21. *Given L, consider a* $\mathrm{Dirichlet}(\alpha_{k,b})_{b \in \tilde{\mathcal{B}}}$ *prior on the simplex in* $\Delta^{\mathcal{B}_L^o}_{\tilde{\mathcal{B}}}$ *corresponding to the L-mer k. For all L, assume $\alpha_{k,b} > 0$ for $(k,b) \in supp_L(p^*)$ (otherwise $p((X_n)_{n=1}^N | \mathcal{M}_L)$ is eventually 0 a.s.). Then if $\tilde{p} \neq p^{*(L)}$,*

$$\log BF_L = N(\mathrm{KL}(p^*||p^{*(L)}) - \mathrm{KL}(p^*||\tilde{p})) + O_P(\sqrt{N}),$$

which goes to ∞ in probability if $\mathrm{KL}(p^*||p^{*(L)}) > \mathrm{KL}(p^*||\tilde{p})$ and to $-\infty$ in probability if $\mathrm{KL}(p^*||p^{*(L)}) < \mathrm{KL}(p^*||\tilde{p})$. If $\tilde{p} = p^{*(L)}$

$$\log BF_L = \frac{1}{2} \dim_L^{eff}(p^*) \log N + O_P(1),$$

which goes to ∞ in probability.

Proof. Note that as shown in the proof of theorem 7, $\mathrm{KL}(p^*||\mathcal{M}_L) = \mathrm{KL}(p^*||p^{*(L)})$, and

$$\log p((X_n)_{n=1}^N | \mathcal{M}_L) = \log p^{*(L)}((X_n)_{n=1}^N) - \frac{1}{2} \dim_L^{eff}(p^*) \log N + O_P(1). \tag{27}$$

As well, $\tilde{p}(X_n)_{n=1}^N = N\mathbb{E} \log \tilde{p}(X) + O_P(\sqrt{N})$ and a similar expression can be written for $p^{*(L)}$. These two facts prove the result. □

Remark 3. One may also consider a Bayes factor that integrates over many L:

$$\mathrm{BF} = \frac{\tilde{p}(X_n)_{n=1}^N}{\sum_{L=1}^{\tilde{L}} \pi(L) p((X_n)_{n=1}^N \mid \mathcal{M}_L)} = \left(\sum_{L=1}^{\tilde{L}} \pi(L) \mathrm{BF}_L^{-1} \right)^{-1}$$

for a prior π with $\pi(\tilde{L}) > 0$. By proposition 21, this Bayes factor goes to 0 if $\mathrm{KL}(p^*\|p^{*(\tilde{L})}) < \mathrm{KL}(p^*\|\tilde{p})$ and goes to ∞ if $\mathrm{KL}(p^*\|p^{*(\tilde{L})}) > \mathrm{KL}(p^*\|\tilde{p})$ or $\tilde{p} = p^{*(\tilde{L})}$ (this later condition is implied by $\tilde{p} = p^{*(L)}$ for some $L \leq \tilde{L}$ and $\mathrm{KL}(p^*\|p^{*(L)}) = \mathrm{KL}(p^*\|\tilde{p})$). Thus this Bayes factor has the same asymptotics as $\mathrm{BF}_{\tilde{L}}$.

G.2 Two-sample test

To set up the two-sample testing problem, consider two distributions p_1 and p_2 on S such that $\mathbb{E}_{p_j}|X|^2 < \infty$ for $j \in \{1, 2\}$. We will assume that the two groups of datapoints are sampled together according to a mixture model with observed labels. That is, let $j_1, j_2, \ldots$ be observed Bernoulli iid random variables indicating the group, with $j_n = 1$ with probability β and $j_n = 2$ with probability $1 - \beta$ for a $0 < \beta < 1$. Then, let $X_n \sim p_{j_n}$ independently. The pooled dataset thus follows the generative process $X_1, X_2, \cdots \sim p^* = \beta p_1 + (1 - \beta)p_2$ iid. We are interested in whether or not $p_1 \neq p_2$. To make this question theoretically tractable, we will fix the lag L, and attempt only to discern whether $p_1^{(L)} \neq p_2^{(L)}$ where $p_j^{(L)}$ is the best approximation to p_j in $\mathcal{M}_L$ (as defined in section E). In other words, we attempt to distinguish between p_1 and p_2 only up to a "resolution", in analogy to Holmes et al. [27]. We thus consider the Bayes factor

$$\begin{aligned}\mathrm{BF}_L &= \frac{p((X_n)_{n=1}^N|(j_n)_{n=1}^N, p_1 = p_2 \text{ and } p_1, p_2 \in \mathcal{M}_L)}{p((X_n)_{n=1}^N|(j_n)_{n=1}^N, p_1 \neq p_2 \text{ and } p_1, p_2 \in \mathcal{M}_L)} \\ &= \frac{p((X_n)_{n=1}^N|\mathcal{M}_L)}{p((X_n)_{n\leq N, j_n=1}|\mathcal{M}_L)p((X_n)_{n\leq N, j_n=2}|\mathcal{M}_L)}.\end{aligned} \tag{28}$$

In the subsequent remark, we also extend the theory to Bayes factors that integrate over all L up to some fixed maximum.

Consider independent $\mathrm{Dirichlet}(\alpha_{k,b})_{b\in\tilde{\mathcal{B}}}$ priors on the simplexes in $\Delta_{|\tilde{\mathcal{B}}|}^{\mathcal{B}_L^o}$ corresponding to the L-mers k. Assume $\alpha_{k,b} > 0$ for $(k, b) \in \mathrm{supp}_L(p^*) = \mathrm{supp}_L(p_1) \cup \mathrm{supp}_L(p_2)$.

Proposition 22. *If* $p_1^{(L)} \neq p_2^{(L)}$,

$$\begin{aligned}\log BF_L &= N\left[\beta\mathbb{E}_{p_1}\log\frac{p^{*(L)}(X)}{p_1^{(L)}(X)} + (1-\beta)\mathbb{E}_{p_2}\log\frac{p^{*(L)}(X)}{p_2^{(L)}(X)}\right] + O_P(\sqrt{N}) \\ &\to -\infty \text{ as } N \to \infty.\end{aligned} \tag{29}$$

Otherwise $p_1^{(L)} = p_2^{(L)}$ *and*

$$\begin{aligned}\log BF_L &= \frac{1}{2}\dim_L^{eff}(p^*)\log N + O_P(1) \\ &\to \infty \text{ as } N \to \infty.\end{aligned} \tag{30}$$

Proof. First note that as shown in the proof of theorem 7, noting $|\{n|j_n = j\}|/N = O_P(1)$,

$$\begin{aligned}\log p((X_n)_{n=1}^N|\mathcal{M}_L) &= \log p^{*(L)}((X_n)_{n=1}^N) - \frac{1}{2}\dim_L^{eff}(p^*)\log N + O_P(1) \\ \log p((X_n)_{n\leq N, j_n=j}|\mathcal{M}_L) &= \log p_j^{(L)}((X_n)_{n\leq N, j_n=j}) - \frac{1}{2}\dim_L^{eff}(p_j)\log N + O_P(1)\end{aligned} \tag{31}$$

for $j \in \{1, 2\}$. As well, $\log p^{*(L)}((X_n)_{n=1}^N) = N\mathbb{E}\log p^{*(L)}(X) + O_P(\sqrt{N})$ by our assumption on the moments $\mathbb{E}_{p_j}|X|^2 < \infty$ and similar expressions exist for p_1 and p_2. Finally note that

$$\begin{aligned}\arg\min_{v\in\Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}} \mathrm{KL}(p^*||p_v) &= \arg\max \mathbb{E}_{p^*}\log p_v(X) \\ &= \arg\max \beta\mathbb{E}_{p_1}\log p_v(X) + (1-\beta)\mathbb{E}_{p_2}\log p_v(X).\end{aligned} \tag{32}$$

Thus, if $p_1^{(L)} = p_2^{(L)}$ then $p_1^{(L)} = p_2^{(L)} = p^{*(L)}$.

First assume $p_1^{(L)} \neq p_2^{(L)}$. So, we have

$$\begin{aligned}\log \mathrm{BF}_L &= N\mathbb{E}_{p^*} \log p^{*(L)} - \beta N \mathbb{E}_{p_1} \log p_1^{(L)}(X) - (1-\beta) N \mathbb{E}_{p_2} \log p_2^{(L)}(X) + O_P(\sqrt{N}) \\ &= N\left[\beta \mathbb{E}_{p_1} \log \frac{p^{*(L)}}{p_1^{(L)}} + (1-\beta)\mathbb{E}_{p_2} \log \frac{p^{*(L)}}{p_2^{(L)}}\right] + O_P(\sqrt{N}).\end{aligned} \tag{33}$$

Note $\mathbb{E}_{p_1} \log \frac{p^{*(L)}}{p_1^{(L)}} = \mathrm{KL}(p_1||p_1^{(L)}) - \mathrm{KL}(p_1||p^{*(L)}) \leq 0$ by the definition of $p_1^{(L)}$. Since $p_1^{(L)} \neq p_2^{(L)}$, at least one of $\mathbb{E}_{p_1} \log \frac{p^{*(L)}}{p_1^{(L)}}, \mathbb{E}_{p_2} \log \frac{p^{*(L)}}{p_2^{(L)}}$ must be negative and so $\log \mathrm{BF}_L \to -\infty$.

Now say $p_0^{(L)} = p^{*(L)} = p_1^{(L)}$. In this case,

$$\log \mathrm{BF}_L = \frac{1}{2} \dim_L^{eff}(p^*) \log N + O_P(1).$$

Clearly $\log \mathrm{BF}_L \to \infty$.

□

Remark 4. One may also consider a Bayes factor that integrates over many lags:

$$\mathrm{BF} = \frac{\sum_{L=1}^{\tilde{L}} \pi(L) p((X_n)_{n=1}^N | \mathcal{M}_L)}{\left(\sum_{L=1}^{\tilde{L}} \pi(L) p((X_n)_{n \leq N, j_n = 1} | \mathcal{M}_L)\right)\left(\sum_{L=1}^{\tilde{L}} \pi(L) p((X_n)_{n \leq N, j_n = 2} | \mathcal{M}_L)\right)}.$$

By theorem 7, for all three sums, eventually either (a) assuming the condition for consistency in corollary 11 the term corresponding to the smallest L such that $p^* \in \mathcal{M}_L$ will dominate, if $p^* \in \mathcal{M}_{\tilde{L}}$, or (b) the term corresponding to $\tilde{L}$ will dominate, if $p^* \notin \mathcal{M}_{\tilde{L}}$. Thus, by analysis similar to that of proposition 22, in any case, we have equation 29 with L replaced by $\tilde{L}$, so that the Bayes factor goes to 0 if $p_1^{(\tilde{L})} \neq p_2^{(\tilde{L})}$. If, on the other hand, we have $p_1^{(\tilde{L})} = p_2^{(\tilde{L})}$, then there are two cases: $p_1 = p_2 \in \mathcal{M}_{L^*}$ for some $L^* \leq \tilde{L}$ (and L^* is picked to be the smallest such lag), or $p_1, p_2 \notin \mathcal{M}_{\tilde{L}}$. In the first case, $p^* \in \mathcal{M}_{L^*}$ so the asymptotics of BF are identical to that of BF_{L^*} and we can refer to proposition 22 to see that the Bayes factor goes to ∞. In the second case, we may still have $p^* \in \mathcal{M}_{L^*}$ for some minimal $L^* \leq \tilde{L}$; if p^* is not a Markov model with lag $\leq \tilde{L}$, call $L^* = \tilde{L}$. In this case, by the analysis of proposition 22,

$$\log \mathrm{BF} = \left(\dim_{\tilde{L}}^{\mathrm{eff}}(p^*) - \frac{1}{2}\dim_{L^*}^{\mathrm{eff}}(p^*)\right) \log N + O_P(1) \to \infty.$$

Thus the asymptotics of this integrated Bayes factor are identical to that of $\mathrm{BF}_{\tilde{L}}$.

H Consistency in the infinite L case

So far we have only studied consistency in the finite lag L case, that is, our results only show that we can approximate p^* up to some finite resolution L (corresponding to the largest available lag). In this section, we develop frequentist and Bayesian consistency results for the fully nonparametric model, that is, we allow for priors with support over all lags L up to infinity, and show that we can approximate p^* itself even if $p^* \notin \mathcal{M}$. The Bayesian consistency result is our main result, and the most practically useful, but the frequentist result is a natural first step toward the Bayesian result, and an opportunity to develop novel constructions (such as the projection algorithm in section H.2) useful in proving the Bayesian result.

H.1 Frequentist consistency

We first show that maximum likelihood estimation is consistent, using the method of sieves described in Geman and Hwang [20]. The idea is to increase the size of the model class with the amount of data N slowly enough to avoid over-fitting. We define the model class considered for N data points first with the lag L, but also by restricting transition

probabilities to be bounded below by a ν: In particular, when there are N datapoints, the model class we consider, or the N-th "sieve", is $\mathcal{S}_N = \{v \in \Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_{L_N}^o} \mid \forall k, b, v_{k,b} \geq \nu_N\}$ where $(\nu_N)_{N=1}^{\infty}$, $(L_N)_{N=1}^{\infty}$ are sequences with $L_N \to \infty$, $\nu_N \to 0$.

Theorem 23. *Say $X_1, X_2, \cdots \sim p^*$ iid where p^* is a subexponential distribution on S. Say p_{v_N} is a maximum likelihood distribution with $v_N \in \mathcal{S}_N$ given $(X_n)_{n=1}^N$. $p_{v_N} \to p^*$ and $\text{KL}(p^*||p_{v_N}) \to 0$ a.s. if for some $\epsilon > 0$,*

$$\frac{|supp_{L_N}(p^*)|(\log\left(\nu_N^{-1}\right))^{1+\epsilon}}{N} \to 0. \tag{34}$$

Proof. The proof follows that of theorem 3 of Geman and Hwang [20].

First note that $\mathcal{S}_N$ is compact and the likelihood function is continuous so a maximum likelihood v_N always exists. This satisfies condition C1 of theorem 2 of Geman and Hwang [20].

Next, to satisfy condition C2 (b) of theorem 2 of Geman and Hwang [20] we show that there are $\tilde{v}_N \in \mathcal{S}_N$ such that $\text{KL}(p^*||p_{\tilde{v}_N}) \to 0$. First, for each L, pick a distribution p^L on S such that for all $|X| \leq L$, $p^L(X) > 0$ and $\text{KL}(p^*||p^L) \to 0$ as $L \to \infty$ (for example, pick $p^L(|X| > L) = p^*(|X| > L)$, $p^L(\cdot||X| > L) = p^*(\cdot||X| > L)$ and $p^L(\cdot||X| \leq L)$ positive with $\text{KL}(p^*(\cdot||X| \leq L)||p^L(\cdot||X| \leq L)) < 1/L$). p_L^L as defined in proposition 3 is a lag L Markov model with positive transition probabilities. Thus, for large N, its transition probabilities are in $\mathcal{S}_N$. Now notice,

$$\begin{aligned}
\text{KL}(p^*||p_L^L) =& \mathbb{E}\left[\log\left(\frac{p^*(X)}{p_L^L(X)}\right); |X| \leq L\right] + \mathbb{E}\left[\log\left(\frac{p^*(X)}{p_L^L(X)}\right); |X| > L\right] \\
=& \mathbb{E}\left[\log\left(\frac{p^*(X)}{p^L(X)}\right); |X| \leq L\right] \\
&+ \mathbb{E}\left[\log\left(\frac{p^*(X)}{p^L(X_{1:L}\ldots)|\tilde{\mathcal{B}}|^{-(|X|-L)}}\right); |X| > L\right] \\
\leq& \mathbb{E}\left[\log\left(\frac{p^*(X)}{p^L(X)}\right); |X| \leq L\right] \\
&+ \mathbb{E}\left[\log\left(\frac{p^*(X)}{p^L(X)|\tilde{\mathcal{B}}|^{-(|X|-L)}}\right); |X| > L\right] \\
=& \text{KL}(p^*||p^L) + \left(\log|\tilde{\mathcal{B}}|\right)\mathbb{E}\left[|X| - L; |X| > L\right] \\
\to& 0 \text{ as } L \to \infty \text{ as } \mathbb{E}|X| < \infty.
\end{aligned} \tag{35}$$

Now we can pick $\tilde{v}_N \in \mathcal{S}_n$ such that $\text{KL}(p^*||p_{\tilde{v}_N}) \to 0$.

That $\text{KL}(p^*||p_N) \to 0$ implies $p_N \to p$ for distributions p_N on S follows from Pinsker's inequality. This satisfies condition C2 (a) of theorem 2 of Geman and Hwang [20]. However, note that the proof of theorem 2 of Geman and Hwang [20] also shows that if v_N is an MLE in S_N and the conditions of the theorem hold, then $\text{KL}(p^*||p_{v_N}) \to 0$ a.s..

Finally, we define a partition of each S_N that satisfies conditions i-iii of theorem 2 of Geman and Hwang [20] to get the result. Pick a sequence $\rho_N \to 0$ with $\log\left(\nu_N^{-1}\right) > (\log(1+\rho_N))^{-1}$ eventually. Call $\mathbb{N}$ the set of positive integers and for a $\zeta \in \mathbb{N}^{\text{supp}_{L_N}(p^*)}$, define

$$\hat{\mathcal{O}}_N(\zeta) := \{v \in \mathcal{S}_N \mid \forall (k,b) \in \text{supp}_{L_N}(p^*),\ (1+\rho_n)^{\zeta_{k,b}}\nu_N > v_{k,b} \geq (1+\rho_N)^{\zeta_{k,b}-1}\nu_N\}$$

so that $\cup_{\zeta \in \mathbb{N}^{\text{supp}_{L_N}(p^*)}} \hat{\mathcal{O}}_N(\zeta) = \mathcal{S}_N$ (Fig. S3). Call $\gamma_N = \left(\frac{\log\left(\nu_N^{-1}\right)}{\log(1+\rho_N)} + 1\right)$ and note $(1+\rho_N)^{\gamma_N - 1}\nu_N = 1$. Thus the number of choices of ζ that give non-empty sets, call this $\#\hat{\mathcal{O}}_N$, is bounded above by $\gamma_N^{|\text{supp}_{L_N}(p^*)|}$. Now notice eventually

$$\gamma_N = \frac{\log\left(\nu_N^{-1}\right)}{\log(1+\rho_N)} + 1 \geq \left(\log\left(\nu_N^{-1}\right)\right)^2 + 1 \geq \left(\log\left(\nu_N^{-1}\right)\right)^4$$

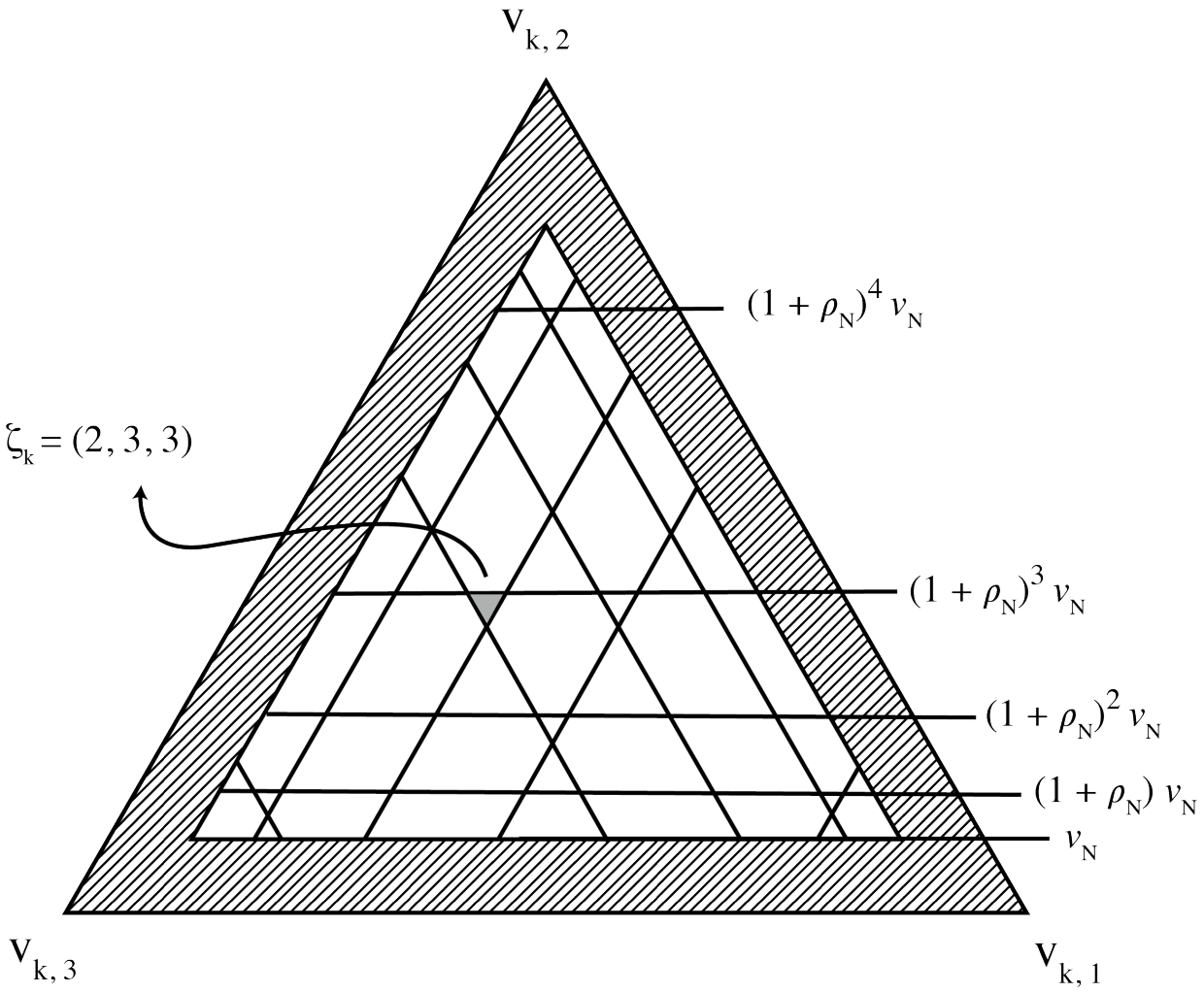

Figure S3: Sieves $\mathcal{S}_N$ are broken up into subsets $\hat{\mathcal{O}}_N(\zeta)$, each a Cartesian product of subsets of $\Delta_{\tilde{\mathcal{B}}}$, and these subsets in turn are indexed by ζ_k for each k. Here we illustrate one such subset of $\Delta_{\tilde{\mathcal{B}}}$, when $|\tilde{\mathcal{B}}| = 3$ and $\text{supp}_{L_N}(p^*)|_k = \tilde{\mathcal{B}}$. The region included in $\hat{\mathcal{O}}_N(\zeta)$ when $\zeta_k = (2,3,3)$ is shown in solid gray, while all other possible subsets for different values of ζ_k are shown in white. The region adjacent to the border of the simplex (hatched lines) corresponds to those transition vectors that have components less than ν_N and are therefore not part of the sieve $\mathcal{S}_N$.

so that $\#\hat{\mathcal{O}}_N \leq \exp\left(4\left(\log\log\left(\nu_n^{-1}\right)\right)|\mathrm{supp}_{L_n}(p^*)|\right)$.

Say $\eta > 0$ and, picking a $\zeta \in \mathbb{N}^{\mathrm{supp}_{L_N}(p^*)}$, define

$$\mathcal{O}_N(\zeta) = \{v \in \hat{\mathcal{O}}_N(\zeta) |\ \mathrm{KL}(p^*||p_{\tilde{v}_N}) - \mathrm{KL}(p^*||p_v) = \mathbb{E}\log\left(\frac{p_v(X)}{p_{\tilde{v}_N(X)}}\right) \leq -\eta\}$$

$$\phi_\zeta(t) = \mathbb{E}\exp\left(t\log\left(\frac{\sup_{v\in\mathcal{O}_N(\zeta)} p_v(X)}{p_{\tilde{v}_N}(X)}\right)\right).$$

Note $\phi_\zeta(t) \leq \mathbb{E}\exp(t|X|(\log\left(\nu_N^{-1}\right)))$ which is finite for small enough t by assumption. ϕ_ζ and the bound $\mathbb{E}\exp(t|X|(\log\left(\nu_n^{*-1}\right)))$ are partition functions for exponential families so, since they are finite for small t, they are C^∞ with derivatives obtained by exchanging differentiation and integration for small t by theorem 4.5 of van der Vaart [64]. In particular, for $t < C_{p^*}/(\log\left(\nu_N^{-1}\right))$ for some C_{p^*} that depends on p^*, defining another constant that depends on p^*, $C'_{p^*} < \infty$,

$$\begin{aligned}\phi''_\zeta(t) =& \mathbb{E}\left[\left(\log\left(\frac{\sup_{v\in\mathcal{O}_N(\zeta)} p_v(X)}{p_{\tilde{v}_N}(X)}\right)\right)^2 \exp\left(t\log\left(\frac{\sup_{v\in\mathcal{O}_N(\zeta)} p_v(X)}{p_{\tilde{v}_N}(X)}\right)\right)\right] \\ \leq& (\log\left(\nu_N^{-1}\right))^2\mathbb{E}\left[|X|^2\exp\left(t|X|(\log\left(\nu_N^{-1}\right))\right)\right] \\ \leq& C'_{p^*}(\log\left(\nu_N^{-1}\right))^2.\end{aligned} \tag{36}$$

As well, for any $v_1, v_2 \in \hat{\mathcal{O}}_N(\zeta)$, for all $(k,b) \in \mathrm{supp}_{L_N}(p^*)$, $|\log\left(v_{1,k,b}/v_{2,k,b}\right)| < \log(1+\rho_N) < \rho_N$. Thus, for all $v \in \mathcal{O}_N(\zeta)$, since if $p^*(X) > 0$ then all L_N-mer-base transitions in X are in $\mathrm{supp}_{L_N}(p^*)$, $\mathbb{E}\log\left(\frac{\sup_{v\in\mathcal{O}_N(\zeta)} p_v(X)}{p_v(X)}\right) < \rho_N\mathbb{E}|X|$. So, defining $C''_{p^*} = \mathbb{E}|X|$,

$$\phi'_\zeta(0) = \mathbb{E}\log\left(\frac{\sup_{v\in\mathcal{O}_N(\zeta)} p_v(X)}{p_v(X)}\right) + \mathbb{E}\log\left(\frac{p_v(X)}{p_{\tilde{v}_N(X)}}\right) < \rho_N C''_{p^*} - \eta. \tag{37}$$

Putting things together we get, for small t,

$$\phi_\zeta(t) \leq 1 + t(\rho_N C''_{p^*} - \eta) + \frac{1}{2}t^2 C'_{p^*}(\log\left(\nu_N^{-1}\right))^2. \tag{38}$$

Picking $t = 2(\log\left(\nu_N^{-1}\right))^{-(1+\epsilon)}$ for some $\epsilon > 0$ gives, for large enough N, for any ζ, $\phi_\zeta(t) \leq 1 - \eta/(\log\left(\nu_N^{-1}\right))^{1+\epsilon} \leq \exp(-\eta/(\log\left(\nu_n^{-1}\right))^{1+\epsilon})$. Finally note that

$$\frac{(\log\left(\nu_N^{-1}\right))^{1+\epsilon}}{N^{1-\epsilon'}} = \left(\frac{(\log\left(\nu_N^{-1}\right))^{(1+\epsilon)/(1-\epsilon')}}{N}\right)^{1-\epsilon'} \to 0$$

by equation 34 if ϵ, ϵ' are small enough. Now write, for large N' and positive constants ϵ'', C, C', C'',

$$\begin{aligned}\sum_{N>N'}^{\infty}\left(\#\hat{\mathcal{O}}_N\right)\left(\sup_\zeta \inf_{t>0}\phi_\zeta(t)\right)^N \leq& \sum_N \exp\left(4\left(\log\log\left(\nu_N^{-1}\right)\right)|\mathrm{supp}_{L_N}(p^*)| - \frac{N\eta}{\left(\log\left(\nu_N^{-1}\right)\right)^{1+\epsilon}}\right) \\ \leq& \sum_N \exp\left(-\frac{N\eta}{\left(\log\left(\nu_N^{-1}\right)\right)^{1+\epsilon}}\left(1 - C\frac{|\mathrm{supp}_{L_N}(p^*)|\left(\log\left(\nu_N^{-1}\right)\right)^{1+\epsilon''}}{N}\right)\right) \\ \leq& \sum_N \exp\left(-N^{\epsilon'}C'\right) \\ \leq& \int_0^\infty dx\exp\left(-C''x^{\epsilon'}\right) \\ =& \epsilon'^{-1}C''^{-1/\epsilon'}\int_0^\infty dx x^{1/\epsilon'-1}\exp\left(-x\right) \\ =& \epsilon'^{-1}C''^{-1/\epsilon'}\Gamma(1/\epsilon') \\ <& \infty\end{aligned} \tag{39}$$

using the assumptions of the theorem and replacing ϵ by ϵ'' to absorb $\log\log\left(v_n^{-1}\right)$ (note one can make $\epsilon, \epsilon', \epsilon''$ as close to 1 as desired). This shows that all conditions of theorem 2 of Geman and Hwang [20] are satisfied. □

Remark 5. To pick viable $(L_N)_N, (\nu_N)_N$, note $|\mathrm{supp}_{L_N}(p^*)| \leq |\tilde{\mathcal{B}}||\mathcal{B}^o_{L_N}|$, so, since

$$|\mathcal{B}^o_N| = \sum_{l=0}^{L_n} |\mathcal{B}|^l = \frac{|\mathcal{B}|^{L_N+1}-1}{|\mathcal{B}|-1} \leq |\mathcal{B}|^{L_N+1},$$

we have $|\mathrm{supp}_{L_N}(p)| \lesssim |\mathcal{B}|^{L_N}$. Thus, as an example, for $c_1, c_2 > 0$ such that $1 > c_1 + c_2$, $L_N = \lceil c_1 \log N / \log|\mathcal{B}| \rceil$ and $\nu_N = e^{-N^{c_2}}$ satisfy condition 34. We can see that without any *a priori* knowledge of $|\mathrm{supp}_{L_N}(p^*)|$ we are forced to pick a very slow growing sequence $(L_N)_N$, and thus it is likely that are model class is too conservative for p^* whose support have cardinality far from the upper bound. By adapting L_N to the content of the data in addition to its cardinality, the Bayesian approach described in section H.3 does not suffer from this conceptual issue.

H.2 The projection algorithm

Fix L and ν for this section and define $\mathcal{S} = \{v \in \Delta^{\mathcal{B}^o_L}_{\tilde{\mathcal{B}}} \mid v_{k,b} \geq \nu \ \forall\ k, b\}$. Given data $X_1, \ldots, X_N$, any maximum likelihood estimate (MLE) in $\mathcal{M}_L$, v, has, for every L-mer k that is seen in the data, $v_{k,b} = \#(k,b) / \left(\sum_{b' \in \tilde{\mathcal{B}}} \#(k,b')\right)$ where $\#(k,b)$ is the number of times k is seen in the data immediately preceding b. If $\bar{v}$ is a MLE in $\mathcal{S}$, it will be shown that for each L-mer k that is seen in the data, $(\bar{v}_{k,b})_{b\in\tilde{\mathcal{B}}}$ is equal to a "projection" of $(v_{k,b})_{b\in\tilde{\mathcal{B}}}$ onto the smaller simplex $\{v_k \in \Delta_{|\tilde{\mathcal{B}}|} \mid v_{k,b} \geq \nu\ \forall b\}$. This projection is defined in algorithm 2, and the rest of this section will be devoted to its properties, including continuity, bounds, and proof of the above statement in proposition 28. Some of these bounds will be used to prove the consistency of nonparametric Bayesian inference in section H.3. For ease of exposition, we will first present a conceptually simpler version of the projection algorithm, algorithm 1.

Algorithm 1 PROJECTION ALGORITHM I

Input : Non-negative numbers $(u_b)_{b\in\tilde{\mathcal{B}}}$, with $\sum_{b\in\tilde{\mathcal{B}}} u_b > 0$, and a positive number $\nu \leq 1/|\tilde{\mathcal{B}}|$.
Output: $(\bar{u}_b)_{b\in\tilde{\mathcal{B}}}$ such that $\sum_{b\in\tilde{\mathcal{B}}} \bar{u}_b = 1$ and $\bar{u}_b \geq \nu$ for all b.

$\bar{u}^{(0)}_b \leftarrow u_b / (\sum_{b'\in\tilde{\mathcal{B}}} u_{b'})$
$B^{(0)} \leftarrow |\{b \mid \bar{u}^{(0)}_b \leq \nu\}|$
$i \leftarrow 1$
while there exists a b with $\bar{u}^{(i-1)}_b < \nu$ **do**
    for $b \in \tilde{\mathcal{B}}$ **do**
        if $\bar{u}^{(i-1)}_b \leq \nu$ **then**
            $\bar{u}^{(i-1)}_b \leftarrow \nu$
        else
            $\bar{u}^{(i-1)}_b \leftarrow \left(1 - B^{(i-1)}\nu\right) u_b / (\sum_{b' \mid \bar{u}^{(i-1)}_{b'} > \nu} u_{b'})$.
    $B^{(i)} \leftarrow |\{b' \mid \bar{u}^{(i)}_{b'} \leq \nu\}|$
    $i \leftarrow i + 1$
for $b \in \tilde{\mathcal{B}}$ **do**
    $\bar{u}_b \leftarrow \bar{u}^{(i-1)}_b$

Proposition 24. *Say algorithm 1 is applied to non-negative numbers $(u_b)_{b\in\tilde{\mathcal{B}}}$ with $\sum_b u_b > 0$. Define $(\bar{u}_b)_b$, $((\bar{u}^{(i)}_b)_b)_i$ and $(B^{(i)})_i$ as in the algorithm. Say the algorithm terminates at step I.*

1) For all i, $\sum_{b=1}^{|\tilde{\mathcal{B}}|} \bar{u}^{(i)}_b = 1$.

2) If $(u_b)_b$ are scaled by a positive constant, the output $(\bar{u}_b)_b$ remains the same.

3) Say $(\bar{v}^{(i)}_b)_b$ is the i-th iteration of algorithm 1 with input $(\bar{u}^{(j)}_b)_b$. $(\bar{v}^{(i)}_b)_b = (\bar{u}^{(j+i)}_b)_b$.

4) $I < |\tilde{\mathcal{B}}|$. *The algorithm remains unchanged if the while loop were replaced by "**for** $i = 1, \dots, |\tilde{\mathcal{B}}| - 1$ **do**".*

5) $\bar{u}_b \geq (1 - (|\tilde{\mathcal{B}}| - 1)\nu)\bar{u}_b^{(0)}$.

Proof. Results 1 and 2 are clear. For 3, note that if both $\bar{u}_b^{(j)}$ and $\bar{u}_{b'}^{(j)}$ are greater than ν, then $\bar{u}_b^{(j)}/\bar{u}_{b'}^{(j)} = u_b/u_{b'}$. Thus, if $\bar{u}_b^{(j)} > \nu$,

$$\bar{u}_b^{(j+1)} = \left(1 - B^{(j)}\nu\right)\bar{u}_b^{(j)} / \left(\sum_{b' \mid \bar{u}_{b'}^{(j)} > \nu} \bar{u}_{b'}^{(j)}\right) = \bar{v}_b^{(1)}.$$

Similar logic may be used to show $(\bar{v}_b^{(2)})_b = (\bar{u}_b^{(j+2)})_b$ and so on.

To see 4, notice that for every $i \leq I$, at least one b has $\bar{u}_b^{(i)} = \nu$ while $\bar{u}_b^{(i-1)} < \nu$. Thus, $(\hat{B}^{(i)})_{i=0}^{I} := (|\{b' \mid \bar{u}_{b'}^{(i)} \leq \nu\}|)_{i=0}^{I}$ is a strictly increasing sequence. $\hat{B}^{(i)} \leq |\tilde{\mathcal{B}}|$ as $\nu \leq 1/|\tilde{\mathcal{B}}|$. If $\hat{B}^{(I)} = |\tilde{\mathcal{B}}|$ then $\nu = 1/|\tilde{\mathcal{B}}|$ and, by property 1, $\hat{B}^{(i)} \neq |\tilde{\mathcal{B}}| - 1$ for every i. In any case, the sequence $(\hat{B}^{(i)})_{i=0}^{I}$ may take on at most $|\tilde{\mathcal{B}}|$ values (including 0) and thus $I < |\tilde{\mathcal{B}}|$. The second statement of 4 follows from the fact that for all b, $\bar{u}_b \geq \nu$ and thus would remain unaltered by the procedure in the while statement.

Finally, for 5, first say $\bar{u}_b > \nu$ and note that $B^{(I-1)} < |\tilde{\mathcal{B}}|$ (otherwise the algorithm is terminated or property 1 is violated).

$$\bar{u}_b = (1 - B^{(I-1)}\nu)\frac{u_b}{\left(\sum_{b' \mid \bar{u}_{b'}^{(I-1)} > \nu} u_{b'}\right)} \geq (1 - B^{(I-1)}\nu)\frac{u_b}{\left(\sum_{b'} u_{b'}\right)} \geq (1 - (|\tilde{\mathcal{B}}| - 1)\nu)\bar{u}_b^{(0)}.$$

Now say $\bar{u}_b = \nu$. Call i' the first step such that $\bar{u}_b^{(i')} = \nu$. If $i' = 0$ or $i' = 1$ then $\bar{u}_b = \nu \geq (1 - (|\tilde{\mathcal{B}}| - 1)\nu)\nu \geq (1 - (|\tilde{\mathcal{B}}| - 1)\nu)\bar{u}_b^{(0)}$. Finally, if $i' > 1$ then

$$\bar{u}_b = \nu \geq \bar{u}_b^{(i'-1)} = \left(1 - B^{(i'-2)}\nu\right)u_b / \left(\sum_{b' \mid \bar{u}_{b'}^{(i'-2)} > \nu} u_{b'}\right) \geq \left(1 - (|\tilde{\mathcal{B}}| - 1)\nu\right)\bar{u}_b^{(0)}.$$

Thus in all cases $\bar{u}_b \geq (1 - (|\tilde{\mathcal{B}}| - 1)\nu)\bar{u}_b^{(0)}$. □

We now turn to the main projection algorithm.

Algorithm 2 PROJECTION ALGORITHM II

Input : Non-negative numbers $(u_b)_{b=1}^{|\tilde{\mathcal{B}}|}$, with $\sum_{b=1}^{|\tilde{\mathcal{B}}|} u_b > 0$, and a positive number $\nu \leq 1/|\tilde{\mathcal{B}}|$.

Output: $(\bar{u}_b)_{b=1}^{|\tilde{\mathcal{B}}|}$ such that $\sum_{b=1}^{|\tilde{\mathcal{B}}|} \bar{u}_b = 1$ and $\bar{u}_b \geq \nu$ for all b.

$\bar{u}_b^{(0)} \leftarrow u_b/(\sum_{b'=1}^{|\tilde{\mathcal{B}}|} u_{b'})$
$\mathcal{C}^{(0)} \leftarrow \emptyset$
$i \leftarrow 1$
while there is a $b \notin \mathcal{C}^{(i-1)}$ with $\bar{u}_b^{(i-1)} \leq \nu$ **do**
    Pick $b^{(i-1)} \in \{b \mid \bar{u}_b^{(i-1)} \leq \nu\} \setminus \mathcal{C}^{(i-1)}$
    $\mathcal{C}^{(i)} \leftarrow \mathcal{C}^{(i-1)} \cup \{b^{(i-1)}\}$
    for $b = 1, \dots, |\tilde{\mathcal{B}}|$ **do**
        if $b \in \mathcal{C}^{(i)}$ **then**
            $\bar{u}_b^{(i)} \leftarrow \nu$
        else
            $\bar{u}_b^{(i)} \leftarrow (1 - i\nu)\, u_b / \left(\sum_{b' \notin \mathcal{C}^{(i)}} u_{b'}\right)$
    $i \leftarrow i + 1$
for $b = 1, \dots, |\tilde{\mathcal{B}}|$ **do**
    $\bar{u}_b \leftarrow \bar{u}_b^{(i-1)}$

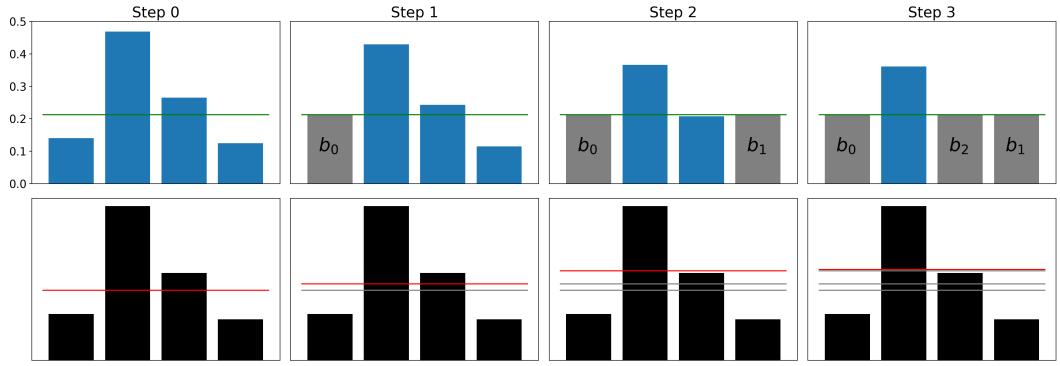

Figure S4: Example application of algorithm 2. $(\bar{u}_b^{(i)})_b$ at the end of each step of the algorithm is shown on the top row with ν in green and those elements in $\mathcal{C}^{(i)}$ in grey. $(u_b)_{b=1}^{|\tilde{\mathcal{B}}|}$ is shown as black bars in the plots in the bottom row with $u^{\mathcal{C}^{(i)}}$ shown as a red line. $u^{\mathcal{C}^{(j)}}$ for previous steps $j < i$ are also shown on the bottom row as grey lines. The scale of the inputs $(u_b)_{b=1}^{|\tilde{\mathcal{B}}|}$ is of no consequence for the algorithm.

An example run of algorithm 2 is visualized in figure S4 (top row). Clearly this algorithm returns $\bar{u}_b = \nu$ if $\nu = 1/|\tilde{\mathcal{B}}|$ and all the following results are trivial. Thus below we will assume $\nu < 1/|\tilde{\mathcal{B}}|$.

Remark 6. We will first consider an alternative representation of the algorithm.

Given a $\mathcal{C} \subset \tilde{\mathcal{B}}$, call

$$u^{\mathcal{C}} = \nu \frac{\sum_{b \notin \mathcal{C}} u_b}{1 - |\mathcal{C}|\nu}$$

and if $u^{\mathcal{C}} > 0$, define

$$\bar{u}_b^{\mathcal{C}} := (1 - |\mathcal{C}|\nu) u_b / \left(\sum_{b' \notin \mathcal{C}} u_{b'} \right) = \nu u_b / u^{\mathcal{C}}$$

for $b \notin \mathcal{C}$ and $\bar{u}_b^{\mathcal{C}} = \nu$ for $b \in \mathcal{C}$; so one gets $\bar{u}_b^{(\tilde{i})} = \bar{u}_b^{\mathcal{C}^{(\tilde{i})}}$ at each iteration $\tilde{i}$. If $b \notin \mathcal{C}$, $\bar{u}_b^{\mathcal{C}} \leq \nu$ if and only if $u_b \leq u^{\mathcal{C}}$.

Say $b \notin \mathcal{C}$ and call $\mathcal{C}' := \{b\} \cup \mathcal{C}$.

$$u^{\mathcal{C}'} - u^{\mathcal{C}} = \nu \frac{\nu \left(\sum_{b' \notin \mathcal{C}} u_{b'} \right) - u_b \left(1 - |\mathcal{C}|\nu\right)}{\left(1 - |\mathcal{C}'|\nu\right)\left(1 - |\mathcal{C}|\nu\right)} = \frac{\nu}{1 - |\mathcal{C}'|\nu} \left(u^{\mathcal{C}} - u_b \right).$$

Thus $u^{\mathcal{C}'} \geq u^{\mathcal{C}}$ if and only if $u_b \leq u^{\mathcal{C}}$ with equality if and only if $u_b = u^{\mathcal{C}}$.

We can see that at iteration i the next $b^{(i-1)}$ is chosen from $\{b \mid \bar{u}_b^{(i-1)} \leq \nu\} \setminus \mathcal{C}^{(i-1)} = \{b \mid u_b \leq u^{\mathcal{C}^{(i-1)}}\} \setminus \mathcal{C}^{(i-1)}$, i.e. from those b with u_b below the threshold $u^{\mathcal{C}^{(i-1)}}$. Thus, $u^{\mathcal{C}^{(0)}} \leq u^{\mathcal{C}^{(1)}} \leq \ldots$. This is reflected in figure S4 (bottom row).

By induction (or from inspection of figure S4), one may show that all the elements b of $\mathcal{C}^{(i)}$ must have u_b below the threshold $u^{\mathcal{C}^{(i-1)}}$ and the algorithm is complete only when all b with u_b below the threshold $u^{\mathcal{C}^{(i)}}$ are inside $\mathcal{C}^{(i)}$. In other words, for $i < I$ (where I is the final iteration) we have $\mathcal{C}^{(i)} \subsetneq \{b \mid u_b \leq u^{\mathcal{C}^{(i)}}\}$, and $\mathcal{C}^{(I)} = \{b \mid u_b \leq u^{\mathcal{C}^{(I)}}\}$.

The important points from the above remark are summarized as:

Lemma 25. *1) Given a $\mathcal{C} \subset \tilde{\mathcal{B}}$, say $b \notin \mathcal{C}$ and call $\mathcal{C}' := \{b\} \cup \mathcal{C}$. $u^{\mathcal{C}'} \geq u^{\mathcal{C}}$ if and only if $u_b \leq u^{\mathcal{C}}$ with equality if and only if $u_b = u^{\mathcal{C}}$.*

2) $u^{\mathcal{C}^{(0)}} \leq u^{\mathcal{C}^{(1)}} \leq \ldots$.

3) If the algorithm ends on step I, $\mathcal{C}^{(i)} \subseteq \{b \mid u_b \leq u^{\mathcal{C}^{(i-1)}}\}$ for all $i \leq I$, $\mathcal{C}^{(i)} \subsetneq \{b \mid u_b \leq u^{\mathcal{C}^{(i)}}\}$ for $i < I$, and $\mathcal{C}^{(I)} = \{b \mid u_b \leq u^{\mathcal{C}^{(I)}}\}$.

Proposition 26. *Say algorithm 2 is applied to non-negative numbers $(u_b)_{b\in\tilde{\mathcal{B}}}$ with $\sum_b u_b > 0$. Define $(\bar{u}_b)_b$, $((\bar{u}_b^{(i)})_b)_i$ and $(\mathcal{C}^{(i)})_i$ as in the algorithm. Say the algorithm terminates at step I.*

1) The output of the algorithm is the same regardless of the choice of $(b_0, b_1, \dots)$.

2) The output of the algorithm is the same as that of algorithm 1.

3) we can replace lines 4 and 5 of algorithm 2 with

*4: **while** there is a $b \notin \mathcal{C}^{(i-1)}$ with $\bar{u}_b^{(i-1)} < \nu$ **do***

*5: **Pick** $b^{(i-1)} \in \{b \mid \bar{u}_b^{(i-1)} < \nu\} \setminus \mathcal{C}^{(i-1)}$*

and receive the same output. With this adjustment, $I < |\tilde{\mathcal{B}}|$.

4) Say $b \notin \mathcal{C}^{(i)}$. $\bar{u}_b^{(i-1)} - \bar{u}_b^{(i)} \leq |\tilde{\mathcal{B}}|(\nu - \bar{u}_{b^{(i-1)}}^{(i-1)})$ so that $\bar{u}_b^{(i-1)}$ is close to $\bar{u}_b^{(i)}$ if $\bar{u}_{b^{(i-1)}}^{(i-1)}$ is close to ν.

Proof. 1) Say the choices $(b^{(0)}, \dots, b^{(I)})$ were made when running the algorithm. Consider a different sequence of choices $(b'^{(0)}, \dots, b'^{(I')})$ to produce $\mathcal{C}'^{(I')}$. Note that by lemma 25, $\mathcal{C} := \mathcal{C}^{(I)} = \{b \mid u_b \leq u^{\mathcal{C}^{(I)}}\}$ and $\mathcal{C}' := \mathcal{C}'^{(I')} = \{b \mid u_b \leq u^{\mathcal{C}'^{(I')}}\}$. Without loss of generality assume $\mathcal{C} \supsetneq \mathcal{C}'$ so $u^{\mathcal{C}} > u^{\mathcal{C}'}$. We will show that this leads to a contradiction. Pick the smallest $i \leq I$ such that $u^{\mathcal{C}^{(i)}} > u^{\mathcal{C}'}$. Then $u^{\mathcal{C}^{(i-1)}} \leq u^{\mathcal{C}'}$, so by lemma 25, $\mathcal{C}^{(i)} \subseteq \mathcal{C}'$.

Pick an enumeration $(\tilde{b}_1, \dots, \tilde{b}_J) = (\mathcal{C}' \setminus \mathcal{C}^{(i)})$. $u_{\tilde{b}_1} \leq u^{\mathcal{C}'} \leq u^{\mathcal{C}^{(i)}}$ so $u^{\mathcal{C}^{(i)}\cup\{\tilde{b}_1\}} \geq u^{\mathcal{C}^{(i)}}$. By induction, one may show that $u^{\mathcal{C}'} = u^{\mathcal{C}^{(i)}\cup(\mathcal{C}'\setminus\mathcal{C}^{(i)})} \geq u^{\mathcal{C}^{(i)}\cup\{\tilde{b}_1,\dots,\tilde{b}_{J-1}\}} \geq \dots \geq u^{\mathcal{C}^{(i)}\cup\{\tilde{b}_1\}} \geq u^{\mathcal{C}^{(i)}}$. This contradicts the choice of i above. Thus, $\mathcal{C} = \mathcal{C}'$ and $I = |\mathcal{C}| = |\mathcal{C}'| = I'$. Moreover, since the final output $(\bar{u}_b)_{b=1}^{|\tilde{\mathcal{B}}|}$ of the algorithm can be defined purely in terms of the final set $\mathcal{C}^{(i)}$, the output must be identical among runs of the algorithm.

2) Consider choosing $(b^{(0)}, \dots, b^{(i)})$ as such: first pick $\{b^{(0)}, \dots, b^{(i_1-1)}\} = \{b \mid \bar{u}_b^{(0)} \leq \nu\}$, which we know can be done since by lemma 25, $\bar{u}_b^{(0)} \leq \nu$ if and only if $u_b \leq u^{\mathcal{C}^{(i_1)}}$ and $u^{\mathcal{C}^{(i_1)}} \leq u^{\mathcal{C}^{(i_1+1)}} \leq \dots$. This is equivalent to one step of the while loop of algorithm 1. Then choose $\{b^{(i_1)}, \dots, b^{(i_2-1)}\} = \{b \mid \bar{u}_b^{(i_1)} \leq \nu\} \setminus \mathcal{C}^{(i_1)}$, which we can do by similar logic. This is equivalent to the second step of the while loop of algorithm 1. Continuing the construction in the same way, by conclusion (1) above, we get that the outputs of algorithms 1 and 2 are identical.

3) Note, by lemma 25, picking a $b^{(i-1)}$ with $\bar{u}_{b^{(i-1)}}^{(i-1)} = \nu$ gives $u^{\mathcal{C}^{(i)}} = u^{\mathcal{C}^{(i-1)}}$ and $\bar{u}_b^{(i)} = \bar{u}_b^{(i-1)}$ for all b. Say $(b_0, \dots, b_i)$, $i < I$ are selected in the algorithm such that $\bar{u}_{b_j}^{(j)} < \nu$ for each $j \leq i$ and all $b \in \{b \mid \bar{u}_b^{(i)} \leq \nu\} \setminus \mathcal{C}^{(i)}$ have $\bar{u}_b^{(i)} = \nu$, then $(\bar{u}_{b'}^{(i+1)})_{b'} = (\bar{u}_{b'}^{(i)})_{b'}$ and all $b \in \{b \mid \bar{u}_b^{(i+1)} \leq \nu\} \setminus \mathcal{C}^{(i+1)}$ have $\bar{u}_b^{(i)} = \nu$. Continuing by induction demonstrates property (3). That $I < |\tilde{\mathcal{B}}|$ follows by the same logic as conclusion (4) in proposition 24 on algorithm 1.

4) Say $b \notin \mathcal{C}^{(i)}$,

$$\begin{aligned}
\bar{u}_b^{(i-1)} - \bar{u}_b^{(i)} =& \nu u_b \left(1/u^{\mathcal{C}^{(i-1)}} - 1/u^{\mathcal{C}^{(i)}}\right) \\
=& \frac{u_b}{\sum_{b'\notin\mathcal{C}^{(i)}} u_{b'}} \frac{\nu(\sum_{b'\notin\mathcal{C}^{(i-1)}} u_{b'}) - u_{b^{(i-1)}}(1-(i-1)\nu)}{\sum_{b'\notin\mathcal{C}^{(i-1)}} u_{b'}} \\
=& \bar{u}_b^{(i)}(1-i\nu)^{-1}(\nu - \bar{u}_{b^{(i-1)}}^{(i-1)}) \\
\leq& |\tilde{\mathcal{B}}|(\nu - \bar{u}_{b^{(i-1)}}^{(i-1)})
\end{aligned} \tag{40}$$

with the last inequality since $i \leq |\tilde{\mathcal{B}}| - 1$ and $\nu \leq 1/|\tilde{\mathcal{B}}|$.

□

Next we show that the projection defined by algorithm 26 is continuous.

Lemma 27. *Say $0 < \nu \leq 1/|\tilde{\mathcal{B}}|$ and $((u_{j,b})_{b=1}^{|\tilde{\mathcal{B}}|})_{j=1}^{\infty}$ is a sequence of sets of non-negative numbers, each with at least one positive element, with $u_{j,b} \to u_b$ for each b as $j \to \infty$, where $(u_b)_{b=1}^{|\tilde{\mathcal{B}}|}$ is set of non-negative numbers with at least one positive element. Apply algorithm 1 or 2 to each set $((u_{j,b})_{b=1}^{|\tilde{\mathcal{B}}|})_{j=1}^{\infty}$ to get $((\bar{u}_{j,b})_{b=1}^{|\tilde{\mathcal{B}}|})_{j=1}^{\infty}$ and to $(u_b)_{b=1}^{|\tilde{\mathcal{B}}|}$ to get $(\bar{u}_b)_{b=1}^{|\tilde{\mathcal{B}}|}$. Then $\bar{u}_{j,b} \to \bar{u}_b$ for all b.*

Proof. Define $\bar{u}_{j,b}^{(i)}$ as in the steps of algorithm 2, with $b^{(0)}, b^{(1)}, \ldots$ to be defined below. Say $\bar{u}_{b^{(0)}}^{(0)} < \nu$. Eventually, $\bar{u}_{j,b^{(0)}}^{(0)} < \nu$ and thus it becomes possible to pick $b^{(0)}$ in the first step of the algorithm for all large enough j. Then, we get $\bar{u}_{j,b^{(0)}}^{(1)} = \nu = \bar{u}_{b^{(0)}}^{(1)}$. For $b \neq b^{(0)}$, $\bar{u}_{j,b}^{(1)} = \nu u_{j,b}/u_j^{\mathcal{C}^{(1)}}$ as defined as part of lemma 25. $u^{\mathcal{C}^{(1)}}$ is a continuous function of $(u_b)_{b=1}^{|\tilde{\mathcal{B}}|}$ so that $\bar{u}_{j,b}^{(1)} \to \bar{u}_b^{(1)}$ for all b. Using the same logic, for large enough j, we may pick an $b^{(1)}$ with $\bar{u}_{b^{(1)}}^{(1)} < \nu$ and see $\bar{u}_{j,b}^{(2)} \to \bar{u}_b^{(2)}$ for all b. We may continue as such until the algorithm terminates for $(u_b)_{b=1}^{|\tilde{\mathcal{B}}|}$ by property (3) in proposition 26. Thus, for some i, we have that $\bar{u}_{j,b}^{(i)} \to \bar{u}_b$ for all b.

Note each $(u_{j,b}^{(i)})_{b=1}^{|\tilde{\mathcal{B}}|}$ may require another $|\tilde{\mathcal{B}}| - i - 1$ steps for the algorithm to complete. For large enough j, we have the implication $\bar{u}_b > \nu \implies \bar{u}_{j,b}^{(i)} > \nu$ for all b so that if for a b, $\bar{u}_{j,b}^{(i)} < \nu$, then $\bar{u}_{j,b}^{(i)} \to \bar{u}_b = \nu$. Applying property (4) in proposition 26 to each of the remaining steps of the algorithm applied to $(u_{j,b})_{b=1}^{|\tilde{\mathcal{B}}|}$ for high enough j, considering $\bar{u}_{j,b}^{(i)} \to \bar{u}_b$ for all b, we can see that $\bar{u}_{j,b} \to \bar{u}_b$ for all b. □

Finally, we can show that the projection algorithms 1 and 2 indeed return the MLE on the sieve $\mathcal{S}$, given observed kmer transition counts.

Proposition 28. *Given data $X_1, \ldots, X_N$, a lag L, and a positive number $\nu < 1/|\tilde{\mathcal{B}}|$, say $\bar{v}$ is an MLE in $\mathcal{S} := \{v \in \Delta_{|\tilde{\mathcal{B}}|}^{\mathcal{B}_L^o} | \ \forall k, b, v_{k,b} \geq \nu\}$. For every L-mer k that has been seen in the data, $(\bar{v}_{k,b})_{b \in \tilde{\mathcal{B}}}$ is equal to the output of algorithm 1 or 2 applied to $(\#(k,b))_{b \in \tilde{\mathcal{B}}}$ where $\#(k,b)$ is the number of times k is seen in the data immediately preceding b.*

Proof. The likelihood of the data under a $p_v \in \mathcal{M}_L$ is

$$\sum_k \sum_b \#(k,b) \log(v_{k,b}).$$

Thus, the MLE in $\mathcal{S}$ can be found by finding, for each k with $\#k > 0$,

$$\text{argmax}_{v_k \in \Delta^{(0)}} \sum_b \#(k,b) \log(v_{k,b})$$

where $\Delta^{(0)} := \{v_k \in \Delta_{\tilde{\mathcal{B}}} \mid \text{for all } b, v_{k,b} \geq \nu\}$.

Say k has been seen in the data, so the MLE on $\Delta_{\tilde{\mathcal{B}}}$, $v_k^{(0)}$, is unique and satisfies $v_{k,b}^{(0)} \propto \#(k,b)$. Call $\hat{v}_k$ an MLE on $\Delta^{(0)}$. Say $v_k^{(0)} \notin \Delta^{(0)}$ so that for some b, $v_{k,b}^{(0)} < \nu_n$. By the uniqueness of the MLE, the likelihood of the data under $v_k^{(0)}$ must be strictly greater than under $\hat{v}_k$. Connecting $\hat{v}_k$ and $v_k^{(0)}$ by a line, considering the concavity of the log likelihood function, the likelihood must be decreasing from $v_k^{(0)}$ to $\hat{v}_k$. As the likelihood function is analytic and not constant on the line, it must be strictly decreasing. Thus the line cannot intersect $\Delta^{(0)}$

except at $\hat{v}_k$. For every b, $\lambda\hat{v}_{k,b} + (1-\lambda)v_{k,b}^{(0)} \geq \nu$ for all $\lambda \in [0,1]$ if $v_{k,b}^{(0)} \geq \nu$; for all $\lambda \in [c,1]$ for a $c < 1$ if $v_{k,b}^{(0)} < \nu_n$ and $\hat{v}_{k,b} > \nu$; and only for $\lambda = 1$ if $v_{k,b}^{(0)} < \nu_n$ and $\hat{v}_{k,b} = \nu$. Therefore, for some $b^{(0)}$ such that $v_{k,b^{(0)}}^{(0)} < \nu$ we have $\hat{v}_{k,b^{(0)}} = \nu$.

Call $v_k^{(1)}$ the MLE on $\{v_k \in \Delta_{\vec{\mathcal{B}}} \mid v_{k,b^{(0)}} = \nu\}$. Using Lagrange multipliers again, one may see that

$$v_{k,b}^{(1)} = (1-\nu)\frac{(k,b)}{\sum_{b \neq b^{(1)}} \#(k,b)}$$

for $b \neq b^{(0)}$. Note that $v_k^{(1)}$ is the result of one step of applying algorithm 2 to $v_k^{(0)}$ using $b^{(0)}$. Call $\Delta^{(1)} := \{v_k \in \Delta_{\vec{\mathcal{B}}} \mid \text{for all } b, v_{k,b} \geq \nu \text{ and } v_{k,b^{(1)}} = \nu\}$ so $\hat{v}_k \in \Delta^{(1)}$. One may perform the same analysis as above to see that if for some b, $v_{k,b}^{(1)} < \nu$, then there is a $b^{(1)}$ such that $v_{k,b^{(1)}}^{(1)} < \nu$ and $\hat{v}_{k,b^{(1)}} = \nu$.

We may then construct $v_k^{(2)}, v_k^{(3)}, \ldots$ by applying algorithm 2, picking $b^{(i)}$. Defining $\Delta^{(i)}$ in analogy to $\Delta^{(0)}$ and $\Delta^{(1)}$, the algorithm stops at step i when $v_k^{(i)} \in \Delta^{(i)}$ and $\hat{v}_k = v_k^{(i)} = \bar{v}_k$. That $\bar{v}_k$ is unique follows from property (2) in remark 26. □

H.3 Bayesian consistency

In this section we take a Bayesian approach to inferring a subexponential p^* from data $X_1, X_2, \cdots \sim p^*$ iid. We put a prior on L, with support over all $L > 0$, to construct a nonparametric Bayesian model and then study the consistency and concentration rate of its posterior. Recall that the Bernstein von-Mises theorem states that given some regularity conditions, for a Bayesian parametric model, the posterior concentrates in a neighborhood centered at the data-generating distribution, with radius proportional to $1/\sqrt{N}$. For nonparametric models in general, and (as we shall see) the BEAR model in particular, the concentration rate of the posterior can be strictly slower than $\sqrt{N}$ [21, 52].

In order to guarantee consistency and derive a concentration rate, we will, instead of placing a prior directly on L, place a prior on sieves constructed similarly to those in section H.1. In particular, define for all L, $\nu' > 0$ and $\nu > 0$ the sieve

$$\mathcal{S}(\nu', \nu, L) = \{v \in \Delta_{\vec{\mathcal{B}}}^{\mathcal{B}_L^o} \mid \forall k,\ v_{k,\$} \geq \nu \text{ and } v_{k,b} \geq \nu' \forall b \in \mathcal{B}\}$$

where ν is a lower bound on the stop transition probability and ν' is a lower bound on all other transitions. In particular, we will define a prior over the sieves that depends on how well a distribution from each sieve can match p^*. Define the sieve approximation mismatch

$$\xi(\nu', \nu, L) = \min_{v \in \mathcal{S}(\nu', \nu, L)} \mathbb{E} \log^2 \left(\frac{p^*(X)}{p_v(X)}\right).$$

In the next section, we will show that we can guarantee ξ is sufficiently small by using the fact that p^* is subexponential. Here, we define the prior.

We may now define our prior:

Condition 29. *Assume, for monotonic sequences $(\nu_m)_m, (L_m)_m$, and a distribution on the natural numbers π,*

$$\log\left(\nu_m^{-1}\right) \sim m^{c_1}$$
$$|\mathcal{B}|^{L_m} \sim m^{c_2}$$
$$\xi(\nu_m, \nu_m, L_m) \lesssim m^{-c_3}$$
$$\log \pi(m) \sim -m^{\omega}$$

with $c_1, c_2, c_3 > 0$ and $1 > c_1 + c_2$. c_3 must obey the following condition: calling $\delta = 1 - \frac{1-(c_1+c_2)}{c_3/2}$, $\delta > 0$ and $(1-\delta)^{-1}(c_1+c_2) \geq \omega > c_1 + c_2$. Consider positive numbers $(\alpha_{k,b})_{L \geq 1, k \in \mathcal{B}_L^o, b \in \vec{\mathcal{B}}}$ such that $\sup \alpha_{k,b} < \infty$ and $\inf \alpha_{k,b} > 0$. Consider a prior Π on the

disjoint union $\sqcup_{m=1}^{\infty} S(0, \nu_m, L_m)$ *that factorizes as such:*

$$\Pi(p_v) = \pi(m) \prod_{k \in \mathcal{B}^o_{L_m}} \Pi_k(v_k) \text{ if } p_v \in \mathcal{S}_m.$$

where for a $k \in \mathcal{B}^o_{L_m}$, Π_k *is a restricted and renormalized Dirichlet*$(\alpha_{k,b})_{b \in \tilde{\mathcal{B}}}$ *prior on the simplex in* $\mathcal{S}(0, \nu_m, L_m)$ *corresponding to transition coefficients out of* k.

Note as well the difference between the sieve we approximate p^* with $(S(\nu_m, \nu_m, L_m))$ and the one our prior is defined over $(S(0, \nu_m, L_m))$. It is best to consider the constraints on c_1, c_2, ω with the fact that c_3 is limited in the values it may take on by how well p^* can be approximated by finite lag Markov models. Our main result will be the consistency of the posterior under this prior and the calculation of its concentration rate.

Remark 7. Using the techniques in section H.2, we can see that the maximum *a posteriori* estimate on each sieve $\mathcal{S}(0, \nu_m, L_m)$ has, for every k that has been seen in the data,

$$v_{k,b} \propto \#(k, b) + \alpha_{k,b}$$

if $\frac{\#(k,\$)+\alpha_{k,\$}}{\sum_{b' \neq \$}(\#(k,b')+\alpha_{k,b'})} \geq \nu_m$; otherwise, $v_{k,\$} = \nu_m$ but we still have $v_{k,b} \propto \#(k, b) + \alpha_{k,b}$ for $b \in \mathcal{B}$. One may then compare the densities of the maximum *a posteriori* estimators in each sieve across L to get the maximum *a posteriori* estimator of the entire posterior.

We now discuss two interpretations of this prior. On the one hand, $\Pi = \sum_{L=1}^{\infty} \sum_{m \mid L_m = L} \pi_{\text{lag}}(m) \Pi(\cdot \mid \mathcal{S}(0, \nu_m, L_m))$ and thus, since $\mathcal{S}(0, \nu_m, L_m) \subset \mathcal{M}_{L_m}$, and the fact that multiple m correspond to the same L_m, the prior can be interpreted as similar to putting a prior on the lag, with the standard Dirichlet priors on each $\mathcal{M}_L$, but with the prior having a "staircase" shape for very small stopping probabilities. On the other hand, we have carefully chosen the values of ν_m and L_m in order to balance the size of $\mathcal{S}_m$ against the amount of information about p^* received from m datapoints. How this works will become clear in the proof of theorem 35.

In section H.3.1 we will show that there exists a c_3 such that $\xi(\nu_m, \nu_m, L_m) \lesssim m^{-c_3}$, i.e., p^* may be efficiently approximated by the sieves. Then we will derive our main result with the concentration rate in section H.3.2. Finally we describe how to use this result in practice on real data in section H.3.3. Throughout we will consider a data generating distribution p^* and all expectations will be with respect to the data generating distribution unless otherwise stated.

H.3.1 Approximating subexponential sequence distributions

In this section we will be interested in finding an asymptotic upper bound for $\xi(\nu_m, \nu_m, L_m)$ of the form m^{-c_3}, thus showing that a prior as in Condition 29 exists (proposition 32). The result relies on the assumption that p^* is subexponential; our main consistency result (theorem 35) would only require $\mathbb{E}|X|^2 < \infty$ if Condition 29 were somehow otherwise satisfied. In its essence, this section is about constructing approximations to subexponential sequence distributions, with control not only over the expected log ratio of p^* and the approximating distribution p – the KL divergence, $\mathbb{E} \log(p^*(X)/p(X))$ – but also over the variance of this log ratio – i.e. control of $\mathbb{E} \log^2 (p^*(X)/p(X))$. We will make use of lemma 3 but need another construction and technical lemma.

Note that if p^* is a distribution on S and $X \in S$,

$$p^*(X) = p^*(X_1 \dots) p^*(X_{1:2} \dots | X_1 \dots) \dots p^*(X_{1:|X|} | X_{1:|X|-1} \dots)$$

where, recall, for a sequence Y, possibly not terminated by \$, $p^*(Y \dots) = p^*(\{X \in S \mid X_i = Y_i \ \forall i \leq |Y|\})$. Thus a probability distribution on S may be described by its infinite-lag transition probabilities $p^*((Y, b) \dots | Y \dots)$ for sequences Y not terminated by \$ and $b \in \tilde{\mathcal{B}}$, ignoring those Y with $p^*(Y \dots) = 0$. Infinite-lag transition probabilities were considered in the construction of p^*_L in proposition 3. Below we will be interested in constructing another distribution from p by projecting, for some L, the transition probabilities at each Y with $|Y| < L$ onto $\{v \in \Delta_{|\tilde{\mathcal{B}}|} \mid v_b \geq \nu^* \ \forall \ b\}$. This first lemma will be used to guarantee the existence of this distribution.

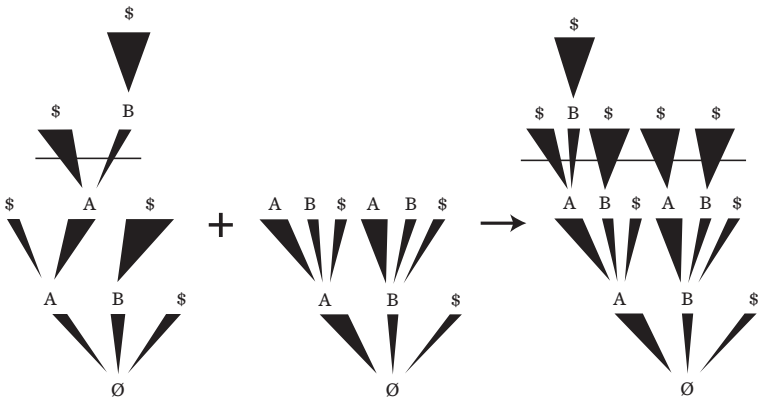

Figure S5: Example application of this construction to the distribution p^* on the left, with the v represented in the center. Transition probabilities for kmers smaller than $L = 2$ are those defined by v while those after are those of the original distribution. Thickness of lines denote probability of particular transition.

Lemma 30. *Say p^* is a probability distribution on S. Given a lag L and positive numbers $((v_{X,b})_{b\in\tilde{\mathcal{B}}})_{l\in\{0,\dots,L-1\},X\in\mathcal{B}^l}$ with $\sum_b v_{X,b} = 1$ for all X, there is a p^{*L} such that for all sequences Y not terminated by \$,*

$$p^{*L}((Y,b)\dots|Y\dots) = \begin{cases} v_{Y,b} & \text{if } |Y| < L \\ p^*(Yb\dots|Y\dots) & \text{if } |Y| \geq L \text{ and } p^*(Y\dots) > 0. \end{cases} \tag{41}$$

Proof. For $X \in S, |X| \leq L$ define

$$p^{*L}(X) = \prod_{i=1}^{|X|} v_{X_{1:i-1},X_i}.$$

For $Y \in \mathcal{B}^L$ with $p(Y\dots) = 0$, define

$$p^{*L}((Y,\$)) = \prod_{i=1}^{L} v_{Y_{1:i-1},Y_i}$$

and $p^{*L}(X) = 0$ for $X \in S$ with $X_{1:L} = Y$ and $X_{L+1} \neq \$$. Finally, if $p^*(Y\dots) > 0$ define, for all $X \in S$ with $X_1 \cdots X_L = Y$,

$$p^{*L}(X) = \left(\prod_{i=1}^{L} v_{Y_{1:i-1},Y_i}\right) p^*(X|Y\dots).$$

It is not difficult to check that p^{*L} is well defined and satisfies the requirements in the statement (Fig. S5). □

Finally, we write a technical lemma:

Lemma 31. *There exists a positive constant C such that for any p^* and p that are distributions over S,*

$$\mathbb{E}_{p^*} \log^2\left(\frac{p^*(X)}{p(X)}\right) \leq \mathbb{E}_{p^*}\left[\log^2\left(\frac{p^*(X)}{p(X)}\right); p^*(X) > p(X)\right] + C\mathrm{KL}(p^*||p)^{1/2}.$$

Proof. $x \mapsto (\log x)^2$ is differentiable with derivative $2x^{-1}\log x$. The derivative is bounded above on $[1,\infty)$, say by C. Thus, for all $x \geq 1$, $(\log x)^2 \leq (\log 1)^2 + C(x-1) = C(x-1)$. Now,

$$\begin{aligned}\mathbb{E}_{p^*}\left[\log^2\left(\frac{p(X)}{p^*(X)}\right); p^*(X) \leq p(X)\right] \leq & C\mathbb{E}_{p^*}\left[\left(\frac{p(X)}{p^*(X)} - 1\right); p^*(X) \leq p(X)\right] \\ = & C\left(p(p(X) > p^*(X)) - p^*(p(X) > p^*(X))\right) \\ \leq & C\|p^* - p\|_{\mathrm{TV}} \\ \lesssim & \mathrm{KL}(p^*||p)^{1/2}.\end{aligned} \tag{42}$$

□

Proposition 32. *If* $\mathbb{E}\exp(t|X|) < \infty$ *for some* $t > 0$ *then* $\xi(\nu_m, \nu_m, L_m) \lesssim m^{-\frac{c_2}{\log|\tilde{\mathcal{B}}|}t}$.

Proof. To approximate p^* with a distribution in $\mathcal{S}(\nu_m, \nu_m, L_m)$ we will use the construction in lemma 3, however we must make sure that the transition probabilities are not less than ν_m. To do so, for sequences X without \$, with $|X| < L_m$, define $(v_{X,b})_{b\in\tilde{\mathcal{B}}}$ to be the output of the application of algorithm 1 or 2 to $(p^*((X,b)\ldots|X\ldots))_{b\in\tilde{\mathcal{B}}}$ if $p^*(X\ldots) > 0$. For X with $p^*(X\ldots) = 0$, make any choice of $(v_{X,b})_b$ with $v_{X,b} \geq \nu_m$ for all b. Thus, for all X, b, $v_{X,b} \geq \nu_m$. Now, by lemma 30, there is a distribution p^{*L_m} with the same infinite-lag transition probabilities as p^* for $|X| \geq L_m$ and infinite-lag transition probabilities $(v_{X,b})_{b\in\tilde{\mathcal{B}}}$ for $|X| < L_m$. Finally perform the construction in lemma 3 to p^{*L_m} to produce a $p^{*L_m}_{L_m} \in \mathcal{S}(\nu_m, \nu_m, L_m)$.

By lemma 31

$$\mathbb{E}\log^2\left(\frac{p^*(X)}{p^{*L_m}_{L_m}(X)}\right) \lesssim \mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{*L_m}_{L_m}(X)}\right); p^*(X) > p^{*L_m}_{L_m}(X)\right] + \left[\mathbb{E}\log\left(\frac{p^*(X)}{p^{*L_m}_{L_m}(X)}\right)\right]^{1/2}.$$

To achieve our result, we will show the first of these terms is $\lesssim m^{-\frac{c_2}{\log|\tilde{\mathcal{B}}|}t}$ and one may use a similar proof to make the same deduction about the second term.

First we will split the term into two that represent the "distance" from p^* to p^{*L_m} and that from p^{*L_m} to $p^{*L_m}_{L_m}$:

$$\begin{aligned}&\mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{L_m}_{L_m}(X)}\right); p^{L_m}_{L_m}(X) < p^*(X)\right] \\ =&\mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{L_m}(X)}\right); |X| \leq L_m, p^{L_m}_{L_m}(X) < p^*(X)\right] \\ &+ \mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{L_m}((X_1,\ldots,X_{L_m})\ldots)|\tilde{\mathcal{B}}|^{-(|X|-L_m)}}\right); |X| > L_m, p^{L_m}_{L_m}(X) < p^*(X)\right] \\ \leq&\mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{L_m}(X)}\right); |X| \leq L_m, p^{L_m}_{L_m}(X) < p^*(X)\right] \\ &+ \mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{L_m}(X)|\tilde{\mathcal{B}}|^{-(|X|-L_m)}}\right); |X| > L_m, p^{L_m}_{L_m}(X) < p^*(X)\right] \\ \leq&\mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{L_m}(X)}\right); |X| \leq L_m, p^{L_m}_{L_m}(X) < p^*(X)\right] \\ &+ 4\mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{L_m}(X)}\right); |X| > L_m, p^{L_m}_{L_m}(X) < p^*(X)\right] \\ &+ 4\log^2\left(|\tilde{\mathcal{B}}|\right)\mathbb{E}\left[(|X| - L_m); |X| > L_m\right] \\ \lesssim&\mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{L_m}(X)}\right)\right] + \mathbb{E}\left[(|X| - L_m)^2; |X| > L_m\right].\end{aligned} \tag{43}$$

Now we will show each of these two terms $\lesssim m^{-\frac{c_2}{\log|\tilde{\mathcal{B}}|}t}$ in turn.

We will first consider $\mathbb{E}\left[(|X|-L_m)^2; |X|>L_m\right]$.

$$\begin{aligned} p^*((|X|-L_m)^2 > l) =& p^*(e^{t|X|} > e^{t(\sqrt{l}+L_m)}) \\ \leq& e^{-tL_m}\mathbb{E}\left[e^{t|X|}\right]e^{-t\sqrt{l}} \end{aligned} \tag{44}$$

by Markov's inequality, so

$$\begin{aligned} \mathbb{E}\left[(|X|-L_m)^2; |X|>L_m\right] &= \int_{L_m}^{\infty} p^*((|X|-L_m)^2 > l)dl \\ &\leq e^{-tL_m}\mathbb{E}\left[e^{t|X|}\right]\int_{L_m}^{\infty} e^{-t\sqrt{l}}dl \\ &\leq e^{-tL_m}\mathbb{E}\left[e^{t|X|}\right]2t^{-2}(t\sqrt{L_m}+1)e^{-t\sqrt{L_m}} \\ &= \exp\Bigg(-tL_m - t\sqrt{L_m} - 2\log t \\ &\quad + \log\left(t\sqrt{L_m}+1\right) + \text{const.}\Bigg) \\ &\lesssim \exp\left(-tL_m\right) \\ &\sim m^{-\frac{c_2}{\log|\tilde{\mathcal{B}}|}t} \end{aligned} \tag{45}$$

as desired.

For the other term in equation 43, again by lemma 31,

$$\mathbb{E}\log^2\left(\frac{p^*(X)}{p^{*L_m}(X)}\right) \lesssim \mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{*L_m}(X)}\right); p^*(X) > p^{*L_m}(X)\right] + \left[\mathbb{E}\log\left(\frac{p^*(X)}{p^{*L_m}(X)}\right)\right]^{1/2}.$$

In this case, we will show that the first of these terms is $\lesssim e^{-Cm^{c_1}}$ for some positive constant C, and by a similar proof one may show the same for the second. This will complete the proof of part 2.

If $p^*(X) > p^{L_m}(X) \geq 0$, by the definition of p^{*L_m},

$$\frac{p^*(X)}{p^{*L_m}(X)} = \prod_{i=1}^{L_m \vee |X|} \frac{p^*(X_{1:i}\ldots|X_{1:i-1}\ldots)}{v_{X_{1:i-1},X_i}} \leq \left(1-(|\tilde{\mathcal{B}}|-1)\nu_m\right)^{L_m}$$

with the inequality by property (5) in proposition 24. Thus,

$$\mathbb{E}\left[\log^2\left(\frac{p^*(X)}{p^{*L_m}(X)}\right); p^*(X) > p^{*L_m}(X)\right] \lesssim L_m^2\nu_m^2 \lesssim \log^2(m)e^{-2Cm^{c_1}} \lesssim e^{-C'm^{c_1}}$$

for two positive constants C, C'. □

H.3.2 Consistency and rate

The proof of theorem 35 relies on a consequence of theorem 2.1 of Ghosal et al. [21], which is stated in a simplified form herein as theorem 33. Intuitively, the key challenge in establishing nonparametric consistency is that the size of the space of probability measures $\mathscr{P}$ (infinite dimensional) may overwhelm the evidence provided by the data, leading to a posterior that is too spread out. To establish consistency, theorem 2.1 of Ghosal et al. [21] requires that the prior over probability measures is sufficiently large on a neighborhood of p^* (denoted $\mathfrak{B}_\eta$), and sufficiently small on the complement of an effectively parametric (finite dimensional) subset of $\mathscr{P}$ (denoted $\mathscr{P}_N$).

Theorem 33. *Say $\mathscr{P}$ is a set of probability measures, $p^* \in \mathscr{P}$. $X_1, \ldots, X_N \sim p^*$ iid, d is the Hellinger distance, Π is a distribution on $\mathscr{P}$, $(\eta_N)_{N=1}^{\infty}$ is a sequence of positive numbers such that $\eta_N \to 0$ and $N\eta_N^2 \to \infty$, and $(\mathscr{P}_N)_{N=1}^{\infty}$ are a sequence of subsets of $\mathscr{P}$. Define, for positive η,*

$$\mathfrak{B}_\eta = \{p \in \mathscr{P} \mid \text{KL}(p^*||p) < \eta^2, \text{Var}[\log(p^*(X)/p(X))] < \eta^2\}.$$

Then if

i) $\log \mathcal{N}(\eta_N/2, \mathscr{P}_N, d) \lesssim N\eta_N^2$

ii) $\log \Pi(\mathfrak{B}_{\eta_N}) \gtrsim -N\eta_N^2$

iii) For an $\epsilon > 0$, $\Pi(\mathscr{P} \setminus \mathscr{P}_N)\Pi(\mathfrak{B}_{\eta_N})^{-1} e^{(1+\epsilon)N\eta_N^2} \to 0$

Then for large enough M,

$$\Pi(B(p^*, M\eta_n)|X_1, \ldots, X_N) \to 1$$

in probability, where $B(p^*, \delta)$ *is a Hellinger ball of radius* δ *centered at* p^*

Proof. For some C,

$$CN\eta_N^2 \geq \log \mathcal{N}(\eta_N/2, \mathcal{S}_N, d) \geq \log \mathcal{D}(\eta_N, \mathcal{S}_N, d).$$

Defining $\eta'_N = \sqrt{C}\eta_N$, condition 2.2 in theorem 2.1 of Ghosal et al. [21] is satisfied for the sequence $(\eta'_N)_{N=1}^{\infty}$. Note condition 2.4 is also satisfied by the above condition ii.

Note by lemma 8.1 in Ghosal et al. [21]

$$\begin{aligned} D_N = &\int \prod_{n=1}^{N} \frac{p(X_n)}{p^*(X_n)} d\Pi(p) \\ \geq &\Pi(\mathfrak{B}_{\eta'_N}) \left(\frac{1}{\Pi(\mathfrak{B}_{\eta'_N})} \int_{\mathfrak{B}_{\eta'_N}} \prod_{n=1}^{N} \frac{p(X_n)}{p^*(X_n)} d\Pi(p) \right) \\ \geq &\Pi(\mathfrak{B}_{\eta'_N}) e^{-(1+\epsilon)N\eta_N'^2} \end{aligned} \tag{46}$$

with probability $1 - \left(\epsilon^2 N\eta_N'^2\right)^{-1} \to 1$. Call the set where this occurs A. As in the proof of theorem 2.1 of Ghosal et al. [21], for large enough M, C', we may then use condition i to write

$$\begin{aligned} 1 - \mathbb{E}_{p^*}\left[\Pi(B(p^*, M\eta'_N)|X_1, \ldots, X_N)\right] \leq &2e^{-C'N\epsilon'_N} + (1 - p^*(A)) \\ &+ \mathbb{E}_{p^*}\left[D_N^{-1} \left(\Pi(\mathscr{P} \setminus \mathscr{P}_N) + e^{-C'NM^2\epsilon_N'^2} \right); A \right]. \end{aligned} \tag{47}$$

By conditions ii and iii, this last term $\to 0$ for large enough M. Finally, write $M\eta'_N = \left(M\sqrt{C}\right)\eta_N$ to get the result in terms of η_N.

□

To work with sieves without restrictions on transition probabilities to $b \in \mathcal{B}$ we need the following technical lemma.

Lemma 34. *Assume for positive numbers* $(\alpha_{k,b})_{L>0, k\in\mathcal{B}_L^o, b\in\tilde{\mathcal{B}}}$, $\sup_{L>0,k\in\mathcal{B}_L^o,b\in\tilde{\mathcal{B}}} \alpha_{k,b} < \infty$ *and* $\inf_{L>0,k\in\mathcal{B}_L^o,b\in\tilde{\mathcal{B}}} \alpha_{k,b} > 0$. *Consider independent* $\text{Dirichlet}(\alpha_{k,b})_{b\in\tilde{\mathcal{B}}}$ *priors on each simplex of* $\Delta_{\tilde{\mathcal{B}}}^{\mathcal{B}_L^o}$ *indexed by* $k \in \mathcal{B}_L^o$. *Call the joint distribution* Π. *Then, for some* $C, \epsilon > 0$, *for all* $\nu > \nu'$ *small,* L,

$$\log \frac{\Pi(\mathcal{S}(\nu', \nu, L))}{\Pi(\mathcal{S}(0, \nu, L))} \geq -C|\mathcal{B}|^L \nu'^{\epsilon}$$

.

Proof. Define $\alpha^{\wedge} \leq \inf_{L,k,b} \alpha_{k,b}$. Let $Z_k \sim \text{Dirichlet}(\alpha_{k,b})_b$ for some k. As a property of the Dirichlet distribution,

$$\left(Z_{k,\$}, \frac{\sum_{b'\in\mathcal{B}} Z_{k,b'}}{\sum_{b'\in\tilde{\mathcal{B}}} Z_{k,b'}} \right) \perp\!\!\!\perp \left(\frac{Z_{k,b'}}{\sum_{b'\in\mathcal{B}} Z_{k,b'}} \right)_{b'\in\mathcal{B}}$$

Call this later variable Y_k, and note $Y_k \sim \text{Dirichlet}(\alpha_{k,b})_{b\in\mathcal{B}}$. Now for any $b \in \mathcal{B}$, $v < \nu$, since $(Y_{k,b}, \sum_{b'\neq b} Y_{b'}) \sim \text{Beta}(\alpha_{k,b}, \sum_{b'\neq b} \alpha_{k,b'})$,

$$\begin{aligned} P(Y_{k,b} < \nu'/(1-v)) =& \frac{\Gamma(\sum_{b'\in\tilde{\mathcal{B}}} \alpha_{k,b'})}{\Gamma(\alpha_{k,b})\Gamma(\sum_{b'\neq b} \alpha_{k,b'})} \int_0^{\nu'/(1-v)} x_b^{\alpha_{k,b}-1}(1-x_b)^{(\sum_{b'\neq b}\alpha_{k,b'})-1} \\ =& O(1)\int_0^{\nu'/(1-v)} x_b^{\alpha_{k,b}-1} \\ =& O((\nu'/(1-v))^{\alpha_{k,b'}}). \end{aligned} \tag{48}$$

Thus, using a union bound, for some C, regardless of the choice of k,

$$P(Y_{k,b} < \nu'/(1-v) \text{ for some } b \in \mathcal{B}) \leq C(\nu'/(1-v))^{\alpha^\wedge}.$$

Thus, for some $C' > 0$, calling $F_{k,\$}$ the density of $Z_{k,b}$, noting $P(Z_{k,\$} > \nu) = O(1)$ for small ν,

$$\begin{aligned} P(Z_{k,b} < \nu' \text{ for some } b \in \mathcal{B} \mid Z_{k,\$} > \nu) \lesssim& \int_\nu^1 P(Y_{k,b} < \nu'/(1-v) \text{ for some } b \in \mathcal{B}) dF_{k,\$}(v) \\ \lesssim& \nu'^{\alpha^\wedge} \int_\nu^1 dv v^{\alpha_{k,\$}-1}(1-v)^{\sum_{b\in\mathcal{B}}\alpha_{k,b}-1-\alpha^\wedge}. \end{aligned} \tag{49}$$

The integral is equal to the probability of a $(Beta)(\alpha_{k,\$}, \sum_{b\in\mathcal{B}}\alpha_{k,b} - \alpha^\wedge)$ distribution being greater than ν and is thus $O(1)$. For small enough ν, ν', for some $C' > 0$,

$$\begin{aligned} \log \frac{\Pi(\mathcal{S}(\nu',\nu,L))}{\Pi(\mathcal{S}(0,\nu,L))} =& \prod_{k\in\mathcal{B}_L^o} \log P(Z_{k,b} \geq \nu' \text{ for all } b \in \mathcal{B} \mid Z_{k,\$} > \nu) \\ \geq& \log\left((1 - C\nu'^{\alpha^\wedge})^{|\mathcal{B}_L^o|}\right) \\ \geq& -C'|\mathcal{B}_L^o|\nu'^{\alpha^\wedge} \\ \geq& -C''|\mathcal{B}|^L\nu'^{\alpha^\wedge}. \end{aligned} \tag{50}$$

□

We can now prove the main result, establishing posterior consistency and the posterior convergence rate. We show that the prior in condition 29 satisfies the conditions of 33. In particular, we use sieves $\mathcal{S}$ to define the effectively parametric subset $\mathscr{P}_N$ of the infinite dimensional space of probability measures $\mathscr{P}$, and then condition 29 controls the prior probability over the $\mathfrak{B}_{\eta_N}$ and $\mathscr{P}_N$.

Theorem 35. *Assume p^* is sub-exponential and thus we can choose a prior as in condition 29. For any large enough M,*

$$\Pi(B(p^*, MN^{-\frac{1}{2}(1-(c_1+c_2))})|X_1,\ldots,X_N) \to 1$$

in probability where $B(p^*,\delta)$ is a Hellinger ball of radius δ centered at p^*.

Proof. The proof will proceed by checking the conditions of theorem 33. First define a monotonic sequence $(\nu'_m)_{m=1}^\infty$ with $\log \nu_N'^{-1} \sim N^\omega$, $\xi_N = \xi(\nu_N, \nu_N, L_N)$, $\mathscr{P}$ the set of distributions on S, and

$$\mathscr{P}_N = \{p_v \mid v \in \cup_{n=1}^N \mathcal{S}(\nu'_n, \nu_n, L_n)\} = \{p_v \mid v \in S(\nu'_N, \nu_N, L_N)\}.$$

Throughout we will use $\eta_N = N^{-\frac{1}{2}(1-(c_1+c_2))}$ and so checking the conditions of theorem 33 will demonstrate a posterior concentration rate of $\frac{1}{2}(1-(c_1+c_2))$.

First we will check condition i. Define, for $\zeta \in \mathbb{N}^{\mathcal{B}_{L_N}^o \times \tilde{\mathcal{B}}}$, $\rho_N > 0$,

$$\hat{\mathcal{O}}_N(\zeta) = \{v \in \mathcal{S}(\nu'_N, \nu_N, L_N) |\ \forall (k,b),\ (1+\rho_N)^{\zeta_{k,b}}\nu_N^b > v_{k,b} \geq (1+\rho_N)^{\zeta_{k,b}-1}\nu_N^b\}$$

(where $\nu_N^b = \nu'_N$ if $b \neq \$$ and equal to ν_N otherwise) so that $\cup_\zeta \hat{\mathcal{O}}_N(\zeta) = \mathcal{S}(\nu'_N, \nu_N, L_N)$ (Fig. S3).

Note that for $v_1, v_2 \in \hat{\mathcal{O}}_N(\zeta)$, $\mathrm{KL}(p_{v_1}||p_{v_2}) \leq \log(1+\rho_N)\mathbb{E}_{v_1}|X| \leq \rho_N \nu_N^{-1}$ the last inequality as $p(|X| > L||X| \geq L) \geq \nu_N$ where the last inequality comes from $p(|X| = L||X| \geq L) \geq \nu_N$ and a geometric sum (this is where a distinction between ν_N and ν'_N is necessary). Defining d as the Hellinger metric,

$$d(p_{v_1}, p_{v_2}) \leq \frac{1}{\sqrt{2}}\|p_{v_1} - p_{v_2}\|_1^{1/2} \leq \mathrm{KL}(p_{v_1}||p_{v_2})^{1/4} \leq (\rho_N \nu_N^{-1})^{1/4}$$

so picking $\rho_N = \nu_N(\eta_N/2)^4$, for $v_1, v_2 \in \hat{\mathcal{O}}_N(\zeta)$, $d(p_{v_1}, p_{v_2}) \leq \eta_N/2$. Call $\gamma^b = \left(\frac{\log\left(\left(\nu_N^b\right)^{-1}\right)}{\log(1+\rho_N)} + 1\right)$ and note $(1+\rho_N)^{\gamma^b - 1}\nu_N^b = 1$. Thus the number of choices of $\zeta \in \mathbb{N}^{\mathcal{B}^o_{L_N} \times \tilde{\mathcal{B}}}$ that give non-empty $\hat{\mathcal{O}}_N(\zeta)$, is bounded above by $\prod_{b\in\tilde{\mathcal{B}}} \left(\gamma^b\right)^{|\mathcal{B}^o_{L_N}|}$. Note also that since $\rho_N \to 0$, $\gamma^b \lesssim \frac{\log\left(\left(\nu_N^b\right)^{-1}\right)}{\rho_N}$. Now we can establish condition i of theorem 33:

$$\begin{aligned}
\log \mathcal{N}(\eta_N/2, \mathcal{S}_N, d) \leq& \log \#\{\zeta \mid \hat{\mathcal{O}}_N(\zeta) \neq \emptyset\} \\
\leq& |\mathcal{B}^o_{L_N}| \sum_b \log\left(\gamma^b\right) \\
\lesssim& |\mathcal{B}|^{L_N} \sum_b \left(\log\log\left(\left(\nu_N^b\right)^{-1}\right) - \log(\nu_N(\eta_N/2)^4)\right) \\
\lesssim& |\mathcal{B}|^{L_N}\left(\log(\nu_N^{-1}) + \log(N)\right) \\
\lesssim& N^{c_1+c_2} \\
\lesssim& N\eta_N^2.
\end{aligned} \tag{51}$$

Now we will demonstrate condition ii. Define, as in theorem 33,

$$\begin{aligned}
\mathfrak{B}_\eta =& \{p \in \mathcal{M} \mid \mathrm{KL}(p^*||p) < \eta^2, \mathrm{Var}[\log(p^*(X)/p(X))] < \eta^2\} \\
\supseteq& \{p \in \mathcal{M} \mid \mathbb{E}\log^2(p^*(X)/p(X)) < \eta^4 \wedge 1\}
\end{aligned} \tag{52}$$

since $\mathrm{Var}[\log(p^*(X)/p(X))] \vee \mathrm{KL}(p^*||p)^2 \leq \mathbb{E}\log^2(p^*(X)/p(X))$.

Fix N. First we will delineate a volume in $\mathcal{S}(\nu_m, \nu_m, L_m)$ for any $m > 0$ that is within $\mathfrak{B}_{\eta_N}$. Using the definition of ξ, we can label a $v_m^* \in \mathcal{S}(\nu_m, \nu_m, L_m)$ such that $\mathbb{E}[\log(p^*(X)/p_{v_m^*}(X))^2] \leq 2\xi_m$. Note that if there exists a $v \in \mathcal{S}(\nu_m, \nu_m, L_m)$ such that for some $\rho_m > 0$ and all k, b, $(1+\rho_m) \geq \frac{v_{k,b}}{v^*_{m,k,b}} \geq (1+\rho_m)^{-1}$ then

$$\begin{aligned}
\mathbb{E}[\log(p^*(X)/p_v(X))^2] \leq& 8\xi_m + 4\mathbb{E}\log^2(p_{v_m^*}(X)/p_v(X)) \\
\leq& 8\xi_m + 4\log^2(1+\rho_m)\mathbb{E}|X|^2.
\end{aligned} \tag{53}$$

Now pick, for large enough m,

$$\rho_m = \sqrt{\frac{\eta_N^4 - 8\xi_m}{4\mathbb{E}|X|^2}} \leq \exp\left(\sqrt{\frac{\eta_N^4 - 8\xi_m}{4\mathbb{E}|X|^2}}\right) - 1$$

so that if $(1+\rho_m) \geq \frac{v_{k,b}}{v^*_{m,k,b}} \geq (1+\rho_m)^{-1}$ for all k, b, then $p_v \in \mathfrak{B}_{\eta_m}$.

Fixing k, the probability under a Dirichlet$(\alpha_{k,b})_b$ distribution of $W_{m,k} = \{v_k \mid (1+\rho_m) \geq \frac{v_{k,b}}{v^*_{m,k,b}} \geq (1+\rho_m)^{-1} \ \forall\ b\}$ (depicted in Fig. S6(A)) is, considering the case where v_m^* is on one of the corners of the simplex $\{v_k \mid v_{k,b} \geq \nu_m\}$, at least

$$V_{m,k} = \left(C_1 \nu_m^{(\sum_b (\alpha_b \wedge 1 - 1))}\right)\left(C_2(\nu_m\rho_m)^{|\tilde{\mathcal{B}}|-1}\right)$$

where the first term is a lower bound on the density and the second on the volume of $W_{m,k}$ and C_1, C_2 are constants depending on $|\tilde{\mathcal{B}}|$. $C_1 > 0$ as $\inf_{k,b} \alpha_{k,b} > 0$. As well, one

may check that the volume is minimized should $v^*_{m,k,b} = \nu_m$ for all but one b; in this case, the volume forms a particular diamond-like shape with side-lengths scaled as $\nu_m\rho_m$ and dimensionality $|\tilde{\mathcal{B}}| - 1$ (Fig. S6(B)), (if $v^*_{m,k,b} = 1 - (|\tilde{\mathcal{B}}| - 1)\nu_m$, then the condition $v_{k,b} \geq (1+\rho_m)^{-1} v^*_{m,k,b} \gtrsim \rho_m$ does not affect the $W_{m,k}$ for large m as $\nu_m \to 0$) (Fig. S6).

Now we will lower bound the probability of $\mathfrak{B}_{\eta_N}$ by the probability of the above defined volume for a particular m, m_N. Call $\delta = 1 - \frac{1-(c_1+c_2)}{c_3/2} > 0$ and define

$$m_N = \left\lceil \left(\frac{\eta_N^4}{16C} \right)^{-1/c_3} \right\rceil \lesssim N^{1-\delta}$$

so that $8\xi_{m_N} \leq \frac{1}{2}\eta_N^4$ for all $m \geq m_N$, and $m_N \to \infty$. Now,

$$\begin{aligned}
\log(\Pi(\mathfrak{B}_{\eta_N})) \geq & \log \left(\pi(m_N) \prod_{k \in \mathcal{B}^o_{L_{m_N}}} V_{m_N,k} \right) \\
\gtrsim & \log(\pi(m_N)) + \left(|\mathcal{B}^o_{L_{m_N}}| - \sum_{k,b} \alpha_{k,b} \wedge 1 \right) \log(\nu^{-1}_{m_N}) \\
& - |\mathcal{B}^o_{L_{m_N}}|(|\tilde{\mathcal{B}}| - 1) \log(\rho^{-1}_{m_N}) \\
\gtrsim & \log(\pi(m_N)) - |\mathcal{B}|^{L_{m_N}} \log(\nu^{-1}_{m_N}) - |\mathcal{B}|^{L_{m_N}} \log(\rho^{-1}_{m_N}).
\end{aligned} \tag{54}$$

For the first term, due to condition 29, $(c_1 + c_2) > (1-\delta)\omega > (1-\delta)(c_1+c_2)$, so,

$$\log \pi(m_N) \sim -m_N^\omega \gtrsim -N^{(1-\delta)\omega} \gtrsim -N^{c_1+c_2}.$$

The second term has

$$|\mathcal{B}|^{L_{m_N}} \log(\nu^{-1}_{m_N}) \lesssim {m_N}^{(c_1+c_2)} \lesssim N^{(1-\delta)(c_1+c_2)}.$$

Finally, for the third, note that since $8\xi_{m_N} \leq \frac{1}{2}\eta_N^4$,

$$\log(\rho^{-1}_{m_N}) \lesssim -\log(\eta_N^4 - 8\xi_{m_N}) \lesssim -\log(\eta_N^4) \lesssim -\log(N).$$

Thus,

$$\log(\Pi(\mathfrak{B}_{\eta_N})) \gtrsim -N^{(1-\delta)(1+\omega)c_2} \gtrsim -N^{(c_1+c_2)} = -N\eta_N^2.$$

Finally, for condition iii, note

$$\Pi(\mathscr{P} \setminus \mathscr{P}_N) = \pi(m > N) + \sum_{m=1}^{N} \pi(m)(1 - \Pi(\mathcal{S}(\nu'_N, \nu_m, L_m) \mid \mathcal{S}(0, \nu_m, L_m))).$$

From lemma 34, we have, for $C, C', \epsilon > 0$, the second term is dominated by

$$\begin{aligned}
\sum_{m=1}^{N} \pi(m) \log \frac{\Pi(\mathcal{S}(0,\nu_m,L_m))}{\Pi(\mathcal{S}(\nu'_N,\nu_m,L_m))} \lesssim & \sum_{m=1}^{N} |\mathcal{B}|^{L_m} \nu'^{\epsilon}_N \\
\lesssim & \nu'^{\epsilon}_N L_N |\mathcal{B}|^{L_N} \\
\lesssim & \exp(-2\epsilon C N^\omega)
\end{aligned} \tag{55}$$

for some $C > 0$. On the other hand, since one may check that $\pi(m+1)/\pi(m) < 1/2$ for all L, we have $\pi(m > N) \leq \pi(N)$. Thus,

$$\log \Pi(\mathscr{P} \setminus \mathscr{P}_N) \lesssim -N^\omega.$$

Now we may write, for any $\epsilon > 0$, since $\omega > c_1 + c_2$

$$\log(\log \Pi(\mathscr{P} \setminus \mathscr{P}_N) e^{(1+\epsilon)N\eta_N^2} \Pi(\mathfrak{B}_{\eta_N})^{-1}) \lesssim -N^\omega + N^{c_1+c_2} + N^{(1-\delta)\omega} \to -\infty.$$

□

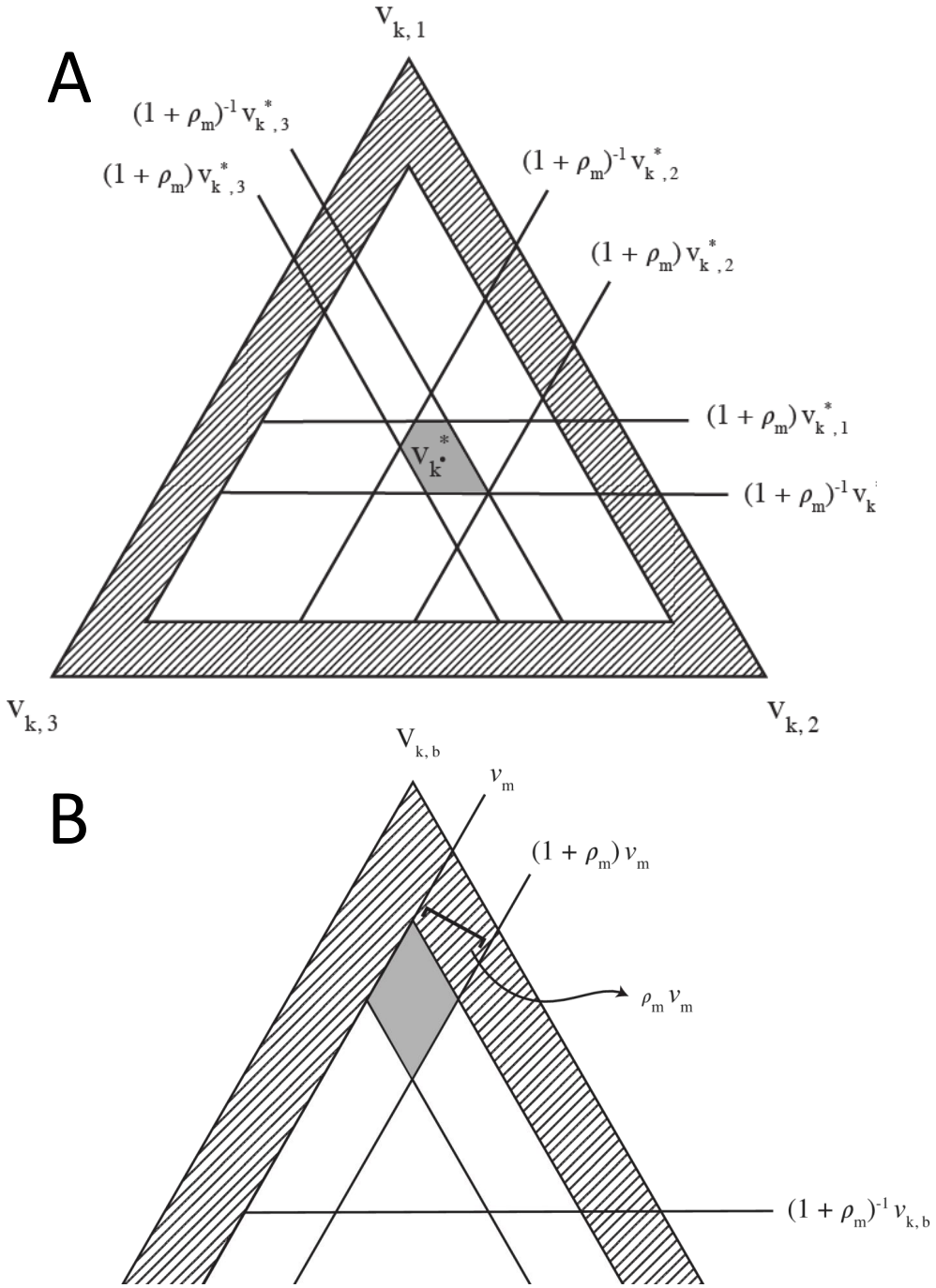

Figure S6: (A) Example of a set $W_{m,k}$ (solid gray) where $(1+\rho_m) \geq \frac{v_{k,b}}{v^*_{m,k,b}} \geq (1+\rho_m)^{-1}\ \forall\ b$ on $\Delta_{\tilde{\mathcal{B}}}$ for a particular k and m when $|\mathcal{B}| = 3$. (B) Depiction of minimum volume possible. The dashed region represents those transition probabilities that have components less than ν_m.

H.3.3 Use in practice

Theorem 35 reveals that the choice of prior controls a kind of bias-variance tradeoff in the model's posterior. In particular, from condition 29 we have

$$c_3 > 2(1 - (c_1 + c_2)) \tag{56}$$

Decreasing the prior hyperparameters c_1 and c_2 decreases the width of the posterior distribution (which plays the role of variance). However, reducing c_1 and c_2 forces down c_3 (by the definition of ξ), and this reduces the weight that the prior places on larger sieves that can match the data distribution better (i.e. sieves with lower $\xi(\nu, \nu, L)$ values), consequently increasing the model's bias. When c_1 and c_2 become low enough, the bias becomes overwhelming, equation 56 is violated, and consistency is no longer guaranteed.

In practice it is often sensible heuristically to set $\nu_m = 0$. In the case, for instance, of short-read sequencing data, there's relatively little correlation between the letters of the read and where it terminates. The probability of stopping is thus often similar across different kmers, even when comparing among kmers of different length. As the posterior concentrates at a roughly constant stopping probability, even a low one, ν_m quickly becomes irrelevant as it decays to zero exponentially. When $\nu_m = 0$, the prior simplifies: it can be written as a distribution over lags $\pi(L)$ times independent Dirichlet priors on each $\mathcal{M}_L$ for $L \in \{1, 2, \ldots\}$. The prior over lags takes the form

$$\log \pi(\{m \mid L_m = L\}) \sim -|\mathcal{B}|^{\frac{\omega}{c_2} L}.$$

Since $\omega > c_2$, we may write $\frac{\omega}{c_2}$ as $1 + c$ for a small $c > 0$.

I Toy models

In this section we describe in depth our simulation experiments.

I.1 Finite lag models

This subsection describes experiments conducted to study in practice the finite lag consistency results described in Sections E and F, and includes details on the results presented in Section 2 and Figure 2.

I.1.1 Setup

To simulate data, we used an AR model with parameters $\theta = (A, B)$ defined by the function,

$$f_k(A, B) = \text{softmax}\left((1 - \beta^*) \sum_{l=1}^{L} \sum_{b' \in \mathcal{B}^o} A_{b,l,b'} k_{l,b'} + \beta^* \sum_{l,l'=1}^{L} \sum_{b', b'' \in \mathcal{B}^o} B_{b,l,l',b',b''} k_{l,b'} k_{l',b''} \right)_{b \in \tilde{\mathcal{B}}} \tag{57}$$

where $\mathcal{B}^o = \mathcal{B} \cup \{\emptyset\}$ and $k_{l,b}$ is 1 if $k_l = b$ and 0 otherwise. The AR model thus takes the form of a multi-output logistic regression, with β^* controlling the contribution of the pairwise interaction terms. In each independent simulation, rows of the matrix A were sampled following,

$$(A_{b,l,b'})_{b \neq \$} \sim (5/L)(\text{Categorical}),\ A_{\$,l,b'} = -1.5/L.$$

for each l, b', where (Categorical) denotes a one-hot encoded sample from a Categorical distribution with uniform probabilities. The matrix B was generated similarly,

$$(B_{b,l,l',b',b''})_{b \neq \$} \sim (5/L^2)(\text{Categorical}),\ B_{\$,l,l',b',b''} = -1.5/L^2.$$

for each l, l', b', b''. Simulations were repeated five times for each β^* value. We set $L = 5$.

We trained the θ parameter in the AR models using maximum likelihood and the h, θ hyperparameters in the BEAR model using empirical Bayes. In both cases, we trained without mini-batching, using 1000 steps of the Adam optimizer with a training rate of 0.05 [33].

To approximate the KL divergence and total variation distance between the models and the data, 2,000 independent sequences were sampled from the data-generating distribution p^* and used to calculate averages of $\log\left(p^*(X)/p(X)\right)$ and $\frac{1}{2}\left|1 - p(X)/p^*(X)\right|$ respectively, where p is either the maximum likelihood estimator (for the AR models) or the posterior predictive (for the BEAR models, estimated using the maximum *a posteriori* value). (Note that the total variation distance is equal to half the L^1 distance since the set of sequences is countable.)

The parameter A is not identifiable, so to compare between the value of A inferred by the models and the true data-generating value, we transformed A to a canonical representation. Define $\tilde{A}_{b,l,b'} = A_{b,l,b'} - A_{\$,l,b'}$ and define the canonical representation

$$A^{\text{can}}_{b,l,b'} = \tilde{A}_{b,l,b'} - \frac{1}{|\mathcal{B}^o|}\left(\sum_{b''}\tilde{A}_{b,l,b''} - \frac{1}{L}\sum_{l',b''}\tilde{A}_{b,l',b''}\right).$$

Proposition 36. *Two linear AR matrices A, A' define the same linear AR model of lag L if and only if $A^{can} = A'^{can}$.*

Proof. Define the vector space

$$V = \{v \in \mathbb{R}^{L\times\mathcal{B}^o} \mid \forall i, j, \sum_{b'} v_{i,b'} = \sum_{b'} v_{j,b'}\}.$$

One hot encodings of sequences of length L are contained in V. As well, it can be seen that V is spanned by the vectors $(e_{i,b} - e_{i,b'})_{1\leq i\leq L, b\neq b'\in\mathcal{B}^o}$ (where $e_{i,b}$ is the indicator of position i, b) and the vector consisting of ones in each entry. This basis of V is made up of linear combinations of one hot encodings of sequences of length L and thus the span of one hot encodings of sequences of length L is V. The orthogonal complement of V is spanned by $(e_i - e_1)_{1<i\leq L}$ where e_i is 1 at position j, b if $j = i$ and 0 otherwise. The transformation

$$v \mapsto \left(v_{i,b} - \frac{1}{|\mathcal{B}^o|}\left(\sum_{b''} v_{i,b''} - \frac{1}{L}\sum_{i',b''} v_{i',b''}\right)\right)_{1\leq i\leq L, b\neq b'\in\mathcal{B}^o}$$

preserves V and annihilates the orthogonal complement of V and is thus the orthogonal projection onto V, P_V.

Thanks to the softmax in Equation 57, two linear AR matrices A and A' define the same linear AR model if there is a constant C such that for all sequences k of length L and $b \in \tilde{\mathcal{B}}$,

$$\sum_{l=1}^{L}\sum_{b'\in\mathcal{B}^o} A_{b,l,b'}k_{l,b'} = \sum_{l=1}^{L}\sum_{b'\in\mathcal{B}^o} A'_{b,l,b'}k_{l,b'} + C.$$

This is equivalent to the condition

$$\sum_{l=1}^{L}\sum_{b'\in\mathcal{B}^o} \tilde{A}_{b,l,b'}k_{l,b'} = \sum_{l=1}^{L}\sum_{b'\in\mathcal{B}^o} \tilde{A}'_{b,l,b'}k_{l,b'}$$

for all k, b and thus to the condition

$$P_V\tilde{A}_b = P_V\tilde{A}'_b$$

for all b. □

I.1.2 Results

We first fixed L at the same value as the simulation data, to study the effect of the structured prior in the BEAR model. Figure 2A shows the convergence in KL of each model as the dataset size increases, and Figure S8 the convergence in total variation distance. Figure 2B shows the convergence of the hyperparameter h in the BEAR model. In Figure S7, we compare the parameter A inferred with the AR model to the true data-generating value

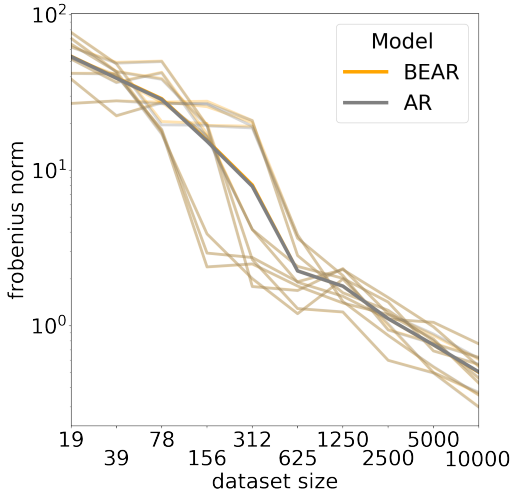

Figure S7: Frobenius norm between the canonical representation (Section I.1) of the AR model parameters θ inferred by fitting an AR model with maximum likelihood and those inferred by fitting the BEAR model with empirical Bayes, in the well-specified ($\beta^* = 0$) case. Thick lines show the average across five independent simulations (small lines). Note that the differences between the two models are indistinguishable relative to the variation across datasets and the variation as dataset size increases.

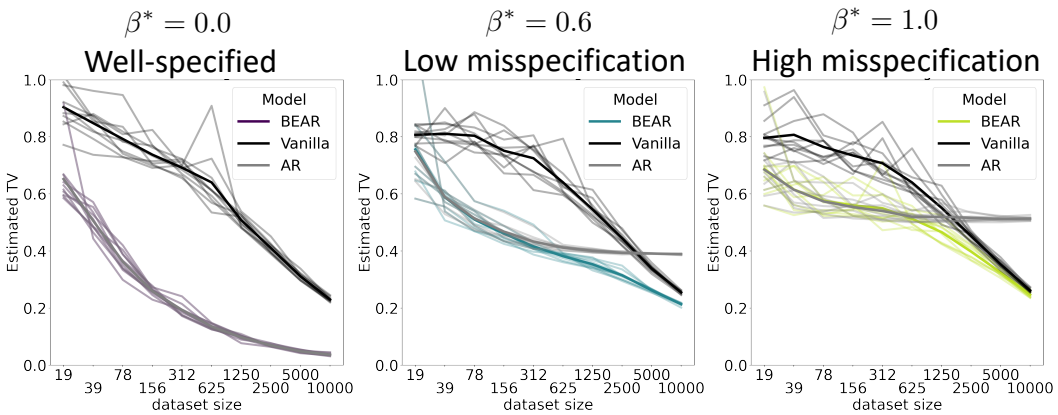

Figure S8: As in Figure 2A, except using the total variation distance in place of the KL norm.

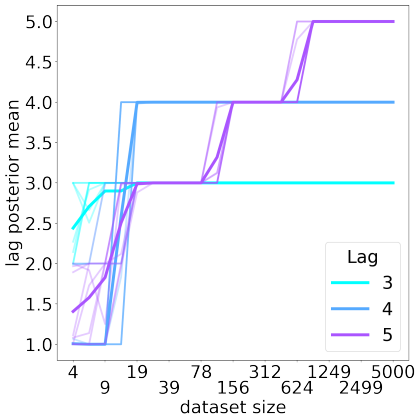

Figure S9: Mean of the BEAR model posterior over lags, as a function of dataset size. Thick lines show the average across five independent simulations (thin lines).

using the Frobenius norm of the canonical representation of each; likewise for the parameter A inferred with the BEAR model. In this well-specified case, we see that the BEAR model parameter estimate converges just as quickly as the AR model.

Next we considered inference of L. We simulated data from models with different L values ($L \in \{3, 4, 5\}$) and $\beta^* = 0$. We computed the expected value of L under the posterior with a uniform prior on lags from 1 to 8. Figure S9 shows that the inferred lag converges to the true data-generating value.

I.2 Infinite lag models

This subsection describes experiments conducted to study the infinite lag (nonparametric) consistency results of Section H in practice.

I.2.1 Setup

To generate from a distribution that was not a finite lag AR model, we chose the first letter in each sequence X uniformly from the alphabet $\mathcal{B}$, then sampled the rest of the sequence following,

$$p(X_i = b | X_1, \ldots, X_{i-1}) \propto \sum_{l=1}^{i-1} l^{-2} \sum_{b' \in \mathcal{B}^o} A_{b,l,b'} X_{i-l,b'}.$$

In each independent simulation, the parameter A was sampled as $A_{b,l,b'} \sim \text{Bernoulli}(0.2)$ for each l, b and $b' \neq \$$, and as $A_{b,l,b'} \sim (0.2)(\text{Bernoulli}(0.2))$ for each l, b and $b' = \$$.

Following Section H.3.3, we set $\nu_m = 0$ and used the prior on lags $\pi(L) \propto \exp(-4^{(1+c)L})$. We used a Jeffreys prior ($\alpha_{k,b} = 1/2$ for all k, b) and took the maximum *a posteriori* value of L and v. We also considered the maximum likelihood estimator of L (i.e. with the prior dropped). To approximate the KL divergence and the total variation distance, we used 30,000 samples; the training procedure was otherwise the same as in Section I.1.

I.2.2 Results

We examined the convergence of the posterior predictive distribution of the BEAR model for different values of the prior hyperparameter c. In all cases we see convergence to p^* in both total variation and KL (Figure S10AB). Decreasing c produces a longer-tailed prior, making the maximum *a posteriori* value of L diverge more quickly with dataset size (Figure S10C). In this example, decreasing c yields faster convergence to p^*. Using the maximum likelihood value of L (equivalent to an improper uniform prior) yields even faster convergence to p^*. As discussed in Section H.3.3, lower c corresponds to larger c_2, and so is expected to yield lower posterior variance but larger bias; in this simulation, the reduction in bias clearly

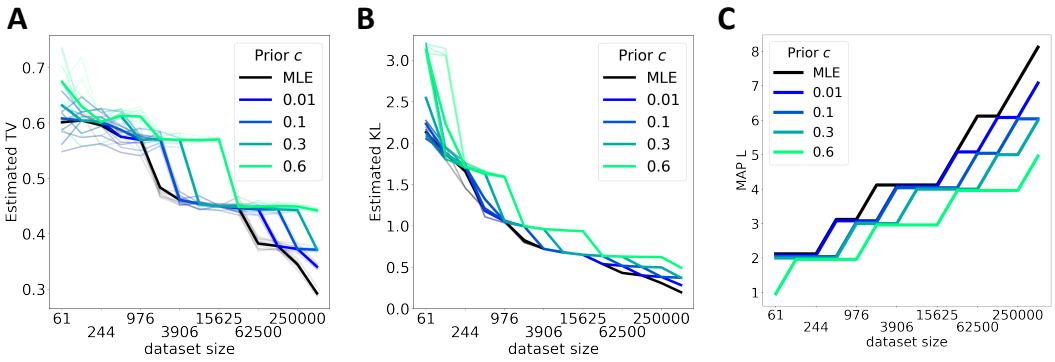

Figure S10: Convergence in total variation (A) and KL (B) between data-generating distribution and model. Thick lines indicate averages across five individual simulations (thin lines). (C) Maximum *a posteriori* estimator of the lag L in an individual example simulation.

contributes more to accurate density estimation. This may be because the data-generating distribution is close enough to a finite-lag Markov model that the asymptotics of the BEAR model behave similarly to the finite-lag case.

I.3 Hypothesis testing

This subsection describes experiments conducted to study the hypothesis testing consistency results of Section G in practice.

I.3.1 Setup

We used the same setup as in Section I.1.1, including the same training and divergence estimation procedures, and sampled datasets from a linear AR model with different values of β^*.

In the goodness-of-fit test, we set $\tilde{p}$ (the model we aimed to test) to a linear AR model with the true, data-generating value of the parameter A but $\beta^* = 0$. We embedded the same linear model, with the same value of A and β^*, in the BEAR model to compute a Bayes factor. Here we set $h = 10^{-3}$, and fixed L at the data-generating value, $L = 5$.

In the two-sample test, instead of comparing to $\tilde{p}$ directly, we compared to samples drawn from $\tilde{p}$. Here we used a Jeffreys prior rather than embed a more complex AR model. We explored both fixing $L = 5$ and using a truncated uniform prior $\pi(L) = 1/8$ for L from 1 to 8 (to evaluate both forms of the consistency results in Section G).

I.3.2 Results

We first examined the consistency of the goodness-of-fit test, using the Bayes factor $\text{BF} = p((X_n)_{n=1}^N)/\tilde{p}((X_n)_{n=1}^N)$ which compares the probability of the data under the BEAR model to the probability under the model of interest $\tilde{p}$. Figure S11A shows the Bayes factor diverge to $+\infty$ when the data does not match the model ($\beta^* > 0$), but diverge to $-\infty$ when the data does match the model ($\beta^* = 0$). We also explored the Bayes factor as function of h, holding the amount of data fixed at $N = 2500$ (Figure S11B). In the limit $h \to 0$, the BEAR model reduces to its embedded AR model $\tilde{p}$, and so the Bayes factor converges to 0. On the other hand, in the limit $h \to \infty$, the BEAR model becomes diffuse and the Bayes factor diverges to negative infinity (accepting the null hypothesis). Intermediate values of h in effect "center" the test at the model $\tilde{p}$ we aim to evaluate, increasing its power to detect differences between the data and the model [6].

We next examined the consistency of the two-sample test, using the Bayes factor $\text{BF} = p((X_n)_{n=1}^N)p((X'_n)_{n=1}^{N'})/p((X_n)_{n=1}^N, (X'_n)_{n=1}^{N'})$, which compares the probability of the two samples being drawn from separate distributions to the probability of their being drawn from the same distribution. Both when using the Bayes factor computed with fixed lag $L = 5$,

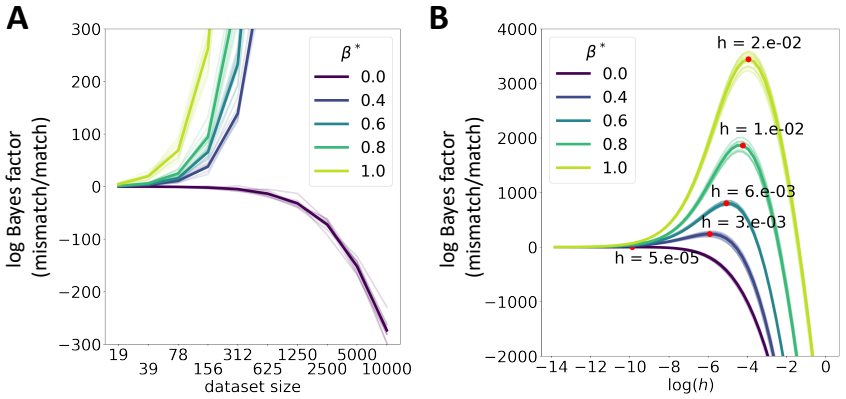

Figure S11: (A) Log Bayes factor for the BEAR goodness-of-fit test. (B) Log Bayes factor as a function of the hyperparameter h, with peaks identified by red points. In both subfigures, thick lines are averages across five simulations (thin lines).

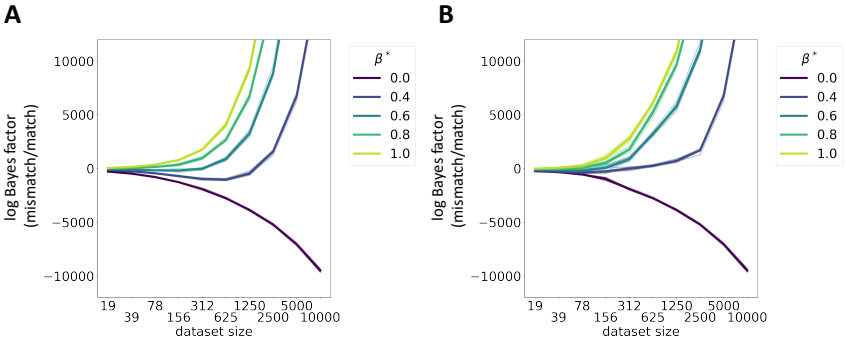

Figure S12: (A) Log Bayes factor for the BEAR two-sample test, using fixed L. (B) Log Bayes factor for the BEAR two-sample test, marginalizing over a truncated prior on L. In both subfigures, thick lines are averages across five simulations (thin lines). Dataset size is the size of each individual dataset that the two-sample test compares, not their pooled size.

and when using the Bayes factor computed by marginalizing over a truncated uniform prior on L, we find consistency, with the Bayes factor diverging to $+\infty$ when $\beta^* > 0$ and to $-\infty$ when $\beta^* = 0$ (Figure S12).

J Scalable inference

In this section we describe how BEAR models were trained at large scale on real data.

J.1 Stochastic gradient estimates

Let $\mathcal{S}$ be a set of length L kmers k in $\hat{\mathcal{B}}_L$ chosen uniformly at random (a minibatch). Then, we can form an unbiased stochastic gradient estimate of the marginal likelihood as

$$\nabla_{h,\theta} \log p(X_{1:n}|L,h,\theta) \approx \frac{|\hat{\mathcal{B}}_L|}{|\mathcal{S}|} \sum_{k \in \mathcal{S}} \nabla_{h,\theta} \log \left[\frac{\Gamma(\sum_b \frac{1}{h} f_{kb}(\theta))}{\prod_b \Gamma(\frac{1}{h} f_{kb}(\theta))} \frac{\prod_b \Gamma(\frac{1}{h} f_{kb}(\theta) + \#(k,b))}{\Gamma(\sum_b \frac{1}{h} f_{kb}(\theta) + \#(k,b))} \right].$$

Note also that it is straightforward to parallelize the training algorithm by sending individual minibatches to individual processors at each step, then compiling the results.

Table S1: **Dataset sizes** In nucleotides (nt). Dataset abbreviations as in Table 1.

Dataset	Total nt	Max. sequence length (nt)
YSD1	151,691,700	150
A. th. 1	3,238,613,507	100
A. th. 2	2,485,960,312	100
A. th. 3	6,831,756,793	100
PBMC	34,935,800,234	91
HL	24,185,778,348	91
GBM	21,506,001,361	65
HC	2,283,930,547	202
CD	1,052,405,190	202
UC	956,179,237	202
Bact.	1,388,421,381	6,358,077

J.2 Extracting summary statistics

KMC counts kmers in large sequence datasets, outputting a list of kmers k and counts $\#k$ that is typically too large to fit in memory. However, our inference procedure requires full count vectors $\#(k, \cdot)$. We take advantage of the lexicographical ordering of KMC's output to merge kmer counts into count vectors in a (single pass) streaming algorithm. We also take advantage of the lexicographical ordering to construct count vectors $\#(k, \cdot)$ for all lags L given just KMC's output for the largest lag L, thus reducing the number of times KMC needs to be run; this too is done using a single pass streaming algorithm. In order to quickly evaluate models by heldout marginal likelihood, it is convenient to store together the counts $\#(k, \cdot)$ associated with both the training and testing datasets. We accomplish this by merging the KMC output for different datasets as part of the same single pass streaming algorithm. This dataset merging is also useful in training the reference-based models proposed in Section L.1, and we merge reference genome counts with sequencing dataset counts in the same way.

J.3 Code availability

Code for implementing BEAR models is available at `https://github.com/debbiemarkslab/BEAR`, along with documentation (including a tutorial for getting started and reproducing basic results); it is available under an MIT license. The models are implemented using TensorFlow and TensorFlow Probability, available under an Apache License [2, 15]. The code also uses NumPy [26], SciPy [66], and BioPython [10] (all BSD 3-Clause licenses). KMC is available under a GNU GPL 3 license.

K Datasets

Here we briefly describe each data type and dataset used in evaluating BEAR models, along with some motivation for each. NCBI accession numbers and links for each dataset can be found in the supplementary table Datasets.xlsx. Dataset sizes are listed in Table S1. All data is publicly available for research use. Patient data was anonymized by the creators of each dataset, and further details on ethical oversight and patient consent can be found in the cited links and papers.

K.1 Whole genome sequencing

Whole genome sequencing is a standard technique for measuring genome sequences. It is often the starting point for running a genome assembly algorithm or variant caller, which aims to infer (non-probabilistically) the underlying genome from the read data. Directly modeling sequencing reads can be interesting, however, since (a) there are typically portions

of the genome that are difficult to reliably assemble, such as centromeres and telomeres, (b) there may not be enough data to reliably detect variants via standard variant callers or assembly, and (c) although the experiment may be directed towards a particular organism's genome other DNA may still be present.

- **YSD1** This is a bacteriophage found in the waterways of the United Kingdom which infects *Salmonella*. It was chosen as an example of a relatively small genome sequencing experiment (phage genomes are short). The sequencing experiment was reported in Dunstan et al. [17].
- **A. th.** Arabidopsis thaliana is a small flowering plant, used as a model organism in plant research. Structural variants are extremely complicated in plants, making traditional variant-calling methods challenging, and kmer-based analysis approaches are of considerable ongoing interest in the literature (see e.g. Voichek and Weigel [67]). The datasets are from the 1001 Genomes Consortium, `https://1001genomes.org/` [1].

K.2 Single cell RNA sequencing

Single cell RNA sequencing is an increasingly ubiquitous technique for characterizing the transcriptional state of cells. It is used to discover new cell types, track development and disease, as a readout in cellular engineering efforts, and more. Most analysis techniques coarse-grain the data by just counting transcripts or isoforms. Statistical modeling of reads at the nucleotide level may lead to new insight into the joint distribution of sequences and their expression levels, accounting for such phenomena as somatic variation and RNA editing. Single cell RNA sequencing is increasingly used as a method for understanding tumors and their microenvironment; cancer involves both genome mutations as well as transcriptional changes.

- **PBMC** Samples of peripheral blood mononuclear cells are easy to collect from humans, making this a standard type of single cell RNA sequencing dataset. These cells were taken from a healthy donor. The dataset is from 10x Genomics, using its v3 technology.
- **HL** These cells come from a human dissociated lymph node tumor, from a 19-year-old male Hodgkin's lymphoma patient. The dataset is from 10x Genomics, using its v3 technology.
- **GBM** These cells were taken from a patient with glioblastoma, the most common primary brain cancer in adults, and include both tumor and peripheral cells. The dataset was reported in [13] and uses a distinct technology from 10x Genomics methods.

K.3 Metagenomics

Metagenomics is an increasingly ubiquitous technique for characterizing microbiomes, including human and environmental microbiomes. Analysis often proceeds by local assembly, annotation of genes or taxa, etc. Statistical modeling of reads at the nucleotide level avoids this coarse graining and can enable detection and analysis of changes in the microbiome outside known genomic elements.

All three of the metagenomics datasets analyzed in the prediction experiments are from [41], a study of inflammatory bowel disease (IBD) as part of the Integrative Human Microbiome Project, and were taken from stool samples. IBD affects more than 3.5 million people worldwide.

- **HC** This dataset was collected from a control patient without IBD.
- **CD** This dataset was collected from a patient with Crohn's disease, a form of IBD involving relapsing and remitting inflammation of the gastrointestinal tract.
- **UC** This dataset was collected from a patient with ulcerative colitis, a form of IBD involving relapsing and remitting inflammation of the colon.

We also examined metagenomics datasets from a study of kidney transplants [55]. Viral transmission from donor to recipient has been associated with complications and increases the risk of allograft failure. Schreiber et al. [55] performed metagenomic sequencing on patient urine samples before and after transplant to assess viral transmission. Further description of this dataset can be found in Section O.

K.4 Full assembled genomes

Comparisons between distant species are challenging due to complex and large scale genomic changes over evolutionary time. However, generative probabilistic models of protein sequences separated by billions of years of evolution has yielded direct insight into their functional constraints, as well as improved understanding of the large scale evolution of life on earth [28, 51]. As a first step towards extending these ideas to whole genomes, we analyzed diverse bacterial genomes from across the tree of life.

- **Bact.** We examined reference bacterial genomes available in RefSeq [46]. Genomes were selected to be taxonomically diverse, representing different genera and families from across the kingdom of Bacteria; the NCBI accessions are listed in Datasets.xlsx.

L Prediction experiments details

Here we provide details on the results reported in the **Predicting sequences** and **Measuring misspecification** subsections of the results (Section 6).

L.1 Model architectures

- **Linear** The linear model is the same as that used in the toy experiments,

$$f_k(A) = \text{softmax}\left(\sum_{l=1}^{L}\sum_{b'\in\mathcal{B}^o} A_{b,l,b'}k_{l,b'}\right)_{b\in\tilde{\mathcal{B}}}. \tag{58}$$

- **CNN** We use a four layer convolutional neural network with the architecture: input $\mapsto$ convolution $\mapsto$ elu $\mapsto$ elu $\mapsto$ softmax $\mapsto$ output, where the convolution is one-dimensional and the elu layers are exponential linear units. Layer normalization was used before each of the elu nonlinearities [4]. Exact details on the model architecture can be found in the supplementary code (Section J.3, function `make_ar_func_cnn` in `ar_funcs.py`).
- **Reference-based** Biologists often make use of a reference genome – a canonical example sequence that is intended to be representative of a species – in analyzing genome sequencing data; reference transcriptomes are used similarly in RNA sequencing analysis, etc.. Reads are aligned to the reference in order to infer the portion of the underlying genome or transcriptome that the read originated from. We built on this basic idea to design an AR model that uses a reference sequence to make predictions. In particular, let $\#_{\text{ref}}(k,b)$ denote the number of times the length $L+1$ kmer (k,b) occurs in the reference sequence(s). One way to form a prediction is by normalizing these counts for each lag, i.e. $f_{k,b} = \#_{\text{ref}}(k,b)/\sum_{b'}\#_{\text{ref}}(k,b')$. We go a step further by (1) accounting for possible mutational or sequencing noise using a Jukes-Cantor mutation model, and (2) accounting for short reads by learning the stop symbol probability. Our complete model is

$$f_{k,b}(\nu,\tau) = (1-\nu)\left[e^{-\tau}\frac{\#_{\text{ref}}(k,b)}{\sum_{b'\neq\$}\#_{\text{ref}}(k,b')} + (1-e^{-\tau})\frac{1}{|\mathcal{B}|}\right] + \nu\mathbb{I}(b=\$) \tag{59}$$

where $\tau\in[0,\infty)$ is the (scalar) Jukes-Cantor time parameter, $\nu\in[0,1]$, and $\mathbb{I}(\cdot)$ is the indicator function that takes value 1 when the expression is true and 0 otherwise. The reference sequences for each dataset are listed in the supplementary table Datasets.xlsx. In analyzing human single cell RNAseq data we pooled multiple reference transcriptomes. We included the reverse complement of each sequence as well as the original sequence when constructing the reference kmer transition counts.

Table S2: **Training parameters** Train-test splits and Adam optimization parameters. Dataset abbreviations as in Table 1. Accum. steps stands for accumulation steps, the number of steps gradients were accumulated over. Paired end reads were treated as separate and split into train and test sets independently.

Dataset	Train/test split	Epochs	Learning rate	Accum. steps
YSD1	3:1 on reads	500	0.01	10
A. th. 1	3:1 on reads	15	0.02	100
A. th. 2	3:1 on reads	15	0.02	100
A. th. 3	3:1 on reads	3	0.02	100
PBMC	3:1 on reads	3	0.02	100
HL	3:1 on reads	5	0.02	100
GBM	55:23 on cells	4	0.02	100
HC	3:1 on reads	10	0.02	100
CD	3:1 on reads	10	0.02	100
UC	3:1 on reads	10	0.02	100
Bact.	500:166 on genomes	2000	0.01	1

L.2 Training

The maximum marginal likelihood lag L was chosen for the vanilla BEAR model (with prior concentration parameter $\alpha_{k,b} = 0.5$ for all k, b). We found in general that the posterior was strongly peaked at a particular lag (Figure S15). All other models (both BEAR and AR) were run with this same lag (that is, we did not integrate over all lags in the BEAR model). Using a fixed lag L as a comparison point provides a controlled study of the effects of switching from an AR model of transition probabilities to the BEAR model's AR-structured prior, and choosing L based on the vanilla BEAR model ensures that the comparison to the vanilla BEAR model is conservative.

The kmer count summary statistics were shuffled once before training (in chunks, due to the large size dataset size), and visited in the same order across epochs. Training was initialized only once; preliminary experiments suggested that training was robust to changes in the random seed. Gradient updates were computed in parallel across two GPUs, at double precision. The minibatch size was 250,000. Gradients were accumulated across minibatches to reduce variance (that is, the gradients from multiple minibatches were added together), and optimization was performed using Adam [33]. Models were trained to convergence. Detailed training hyperparameters are displayed in Table S2. The CNN models used 30 filters of width 8, except in the case of YSD1 where the filter width was reduced to 5 (for both BEAR and AR models); other neural network architecture hyperparameters are given in the supplementary code (function `make_ar_func_cnn` in `ar_funcs.py`). Experiments were run on an internal cluster (Tesla K80, Tesla M40 and Tesla V100 GPUs).

L.3 Evaluation

Accuracy was evaluated based on the maximum likelihood prediction (in the case of AR models) and the maximum *a posteriori* prediction (in the case of BEAR models). Ties in prediction probabilities were resolved uniformly at random.

The perplexity was calculated based on the heldout test dataset as

$$\exp\left[-\frac{\log p((X_n)_{n=1}^{N_{\text{test}}})}{\sum_{n=1}^{N_{\text{test}}} |X_n|}\right] \tag{60}$$

where $p((X_n)_{n=1}^{N_{\text{test}}})$ is the probability of the heldout data conditional on the maximum likelihood parameter value (in the case of AR models) or the marginal probability of the heldout data under the posterior predictive distribution (in the case of BEAR models).

Table S3: **Predictive accuracy.** *Whole genome sequencing data* YSD1: A Salmonella phage. *A. th.*: *Arabidopsis thaliana*, a plant (datasets represent different individuals). *Single cell RNA sequencing data* PBMC: peripheral blood mononuclear cells, taken from a healthy donor. HL: Hodgkin's lymphoma tumor cells. GBM: glioblastoma tumor cells. *Metagenomic sequencing data* HC: healthy (non-CD and non-UC) controls. CD: Crohn's disease. UC: ulcerative colitis. *Full assembled genomes* Bact.: Bacteria. *Models* Van: Vanilla (constant). Lin: Linear. CNN: convolutional neural network. Ref: reference genome/transcriptome model (only applicable to datasets with a reference).

Dataset	AR Lin.	AR CNN	AR Ref.	BEAR Van.	BEAR Lin.	BEAR CNN	BEAR Ref.
YSD1	33.73%	35.86%	90.8%	94.69%	**94.75%**	**94.75%**	94.71%
A. th. 1	35.47%	35.59%	53.81%	86.03%	86.32%	86.34%	**86.50%**
A. th. 2	35.32%	35.61%	70.41%	85.36%	85.71%	**85.77%**	85.66%
A. th. 3	34.94%	35.41%	60.94%	76.46%	78.51%	**78.52%**	77.13%
PBMC	34.36%	34.76%	67.39%	87.83%	**88.16%**	**88.16%**	87.99%
HL	34.67%	35.59%	67.17%	87.68%	**87.96%**	**87.96%**	87.82%
GBM	30.71%	30.9%	61.3%	78.99%	80.44%	80.42%	**81.43%**
HC	32.98%	33.54%	–	83.86%	85.03%	**85.06%**	–
CD	32.13%	32.32%	–	81.72%	83.30%	**83.32%**	–
UC	32.27%	32.23%	–	82.71%	84.26%	**84.27%**	–
Bact.	33.89%	34.78%	-	35.27%	**35.28%**	**35.28%**	-

L.4 Further performance results

The maximum marginal likelihood lag L (under the vanilla BEAR model) for each dataset is reported in S4. Interestingly, the optimal lags are intermediate between the large kmer lengths (e.g. more than 30) often used for non-probabilistic assembly algorithms (e.g. [57]) and the small kmer lengths (e.g. less than 10) often used as features in clustering or classification algorithms (e.g. [3]). The marginal likelihood was in general strongly peaked at a particular value (Figure S15). Increasing the lag generally led to slightly better performance in terms of both perplexity and accuracy for the non-vanilla BEAR models and the AR models, but (unsurprisingly) worse performance for the vanilla BEAR model; the increases in AR model performance were far from enough to make up the difference with BEAR models (Table S5).

Plots of training loss versus wall clock time for an AR model and the corresponding BEAR model (with the same fixed lag L) are shown in Figure S13; the loss for each is normalized by the minimum and maximum values to be comparable (the BEAR model substantially outperforms the AR model). The BEAR model converges at least as fast as the AR model.

To evaluate performance as a function of dataset size, we subsampled reads uniformly at random without replacement from the YSD1 dataset, and retrained the models on the smaller datasets (Figure S14). The original dataset had $\sim 1000\times$ coverage of the bacteriophage genome, meaning that on average 1000 reads were observed overlapping each position in the genome. Note that the vanilla BEAR model performance falls off substantially relative to the BEAR model below $\sim 3\times$ coverage (in the case of the reference model) (Figure S14BD)

M Generation details

Here we provide details on the results reported in the **Generating samples** subsection of the results (Section 6).

The CNN BEAR model was trained on the full (combined train/test data) *Arabidopsis thaliana* 1 dataset, with $L = 17$, using identical training parameters as in the performance experiments (Table S2). 50 bases were generated on the end of reads using the maximum *a posteriori* value of v, and conditional on a stop symbol not occurring, i.e. following the

Table S4: **Maximum marginal likelihood lag L.** Maximum marginal likelihood lag L for the vanilla BEAR model. Dataset abbreviations as in Table 1.

Dataset	L
YSD1	13
A. th. 1	17
A. th. 2	17
A. th. 3	18
PBMC	18
HL	17
GBM	17
HC	16
CD	16
UC	16
Bact.	9

Table S5: **Performance with increasing lag L.** The symbol † indicates the maximum marginal likelihood lag L for the vanilla BEAR model. Dataset abbreviations as in Table 1.

Perplexity

Dataset	Lag	AR Lin.	AR CNN	AR Ref.	BEAR Van.	BEAR Lin.	BEAR CNN	BEAR Ref.
YSD1	13†	3.953	3.873	1.266	1.165	**1.144**	**1.144**	1.145
YSD1	20	3.937	3.855	1.352	1.177	**1.138**	**1.138**	**1.138**
Bact.	9†	3.831	3.794	-	**3.774**	**3.774**	**3.774**	-
Bact.	12	3.807	3.772	-	3.776	3.741	**3.738**	-

Accuracy

Dataset	Lag	AR Lin.	AR CNN	AR Ref.	BEAR Van.	BEAR Lin.	BEAR CNN	BEAR Ref.
YSD1	13†	33.73%	35.86%	90.8%	94.69%	**94.75%**	**94.75%**	94.71%
YSD1	20	34.19%	36.3%	87.21%	94.88%	94.97%	**94.98%**	94.91%
Bact.	9†	33.89%	34.78%	-	35.27%	**35.28%**	**35.28%**	–
Bact.	12	34.42%	35.13%	–	35.54%	35.86%	**35.93%**	–

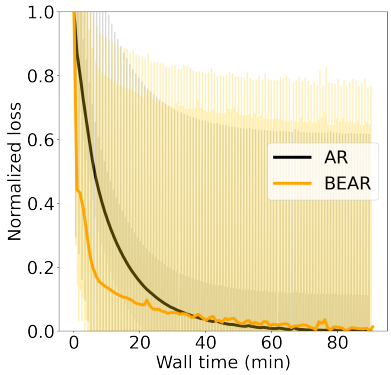

Figure S13: Relative loss (normalized to be between 0 and 1 based on minimum and maximum values) as a function of wall time for a CNN AR model versus the corresponding BEAR model on the YSD1 dataset ($L = 20$).

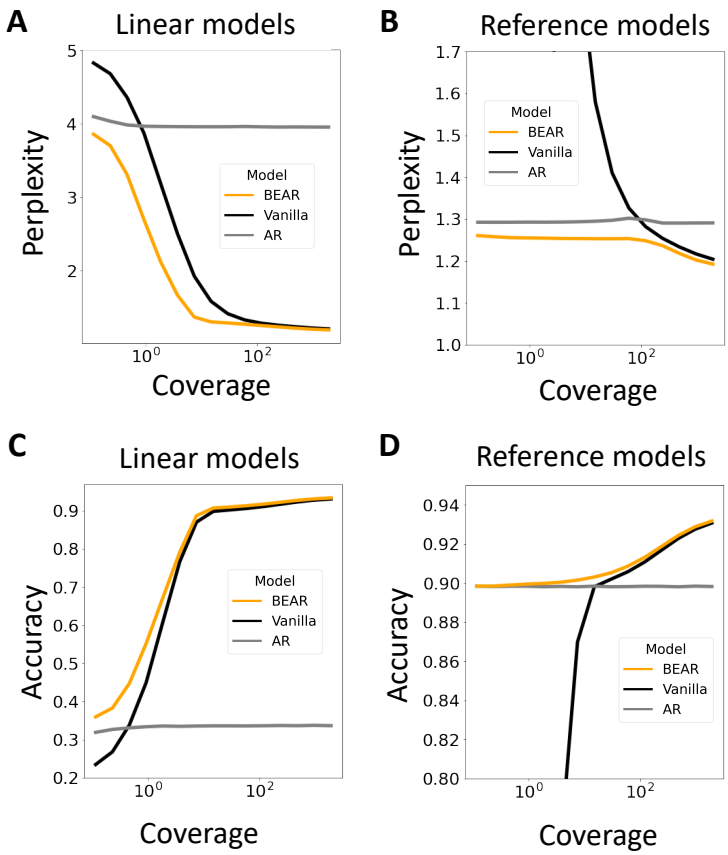

Figure S14: Perplexity (AB) and accuracy (CD) of AR and BEAR models as a function of total dataset size, measured in terms of coverage (coverage is the expected number of reads from each position in the genome; it is linearly proportional to the total number of reads). Subfigures A and C show results for the the linear AR model (and its BEAR embedding), and B and D for the reference-based AR model (and its BEAR embedding). The lag was held fixed in all cases.

distribution

$$p_{\text{extr}}(X_i = b|k = (X_{i-L}, ..., X_{i-1})) = \frac{f_{k,b}(\theta)/h + \#(k,b)}{\sum_{b' \neq \$} f_{k,b'}(\theta)/h + \#(k,b')} \tag{61}$$

for $b \neq \$$ and $p(X_i = \$|k) = 0$, where recall $\#(k,b)$ is the number of times b is seen succeeding k in the data, and θ and h are the learned hyperparameters. The values of $\#(k,b)$ are retrieved from the dataset efficiently using the Jellyfish kmer indexing package [42]. 50 extrapolations each of length 50 were sampled without replacement using the stochastic beam search method proposed by Kool et al. [36].

We performed local assembly using SPAdes, starting from the last 17 bases of the read, and recorded the portion of each scaffold returned by SPAdes that extended in the direction of extrapolation. We used the `--careful` flag in SPAdes, following Voichek and Weigel [67].

The colors in Figure 3A correspond to unique paths through the 17-mer de Bruijn graph. Figure 3B plots the per nucleotide perplexity of the sampled extrapolations, i.e.

$$\exp\left(-\sum_b p_{\text{extr}}(b|k = (X_{n,i-L}, ..., X_{n,i-1})) \log p_{\text{extr}}(b|k = (X_{n,i-L}, ..., X_{n,i-1}))\right)$$

where n indexes the sampled extrapolation and i the position in the sample.

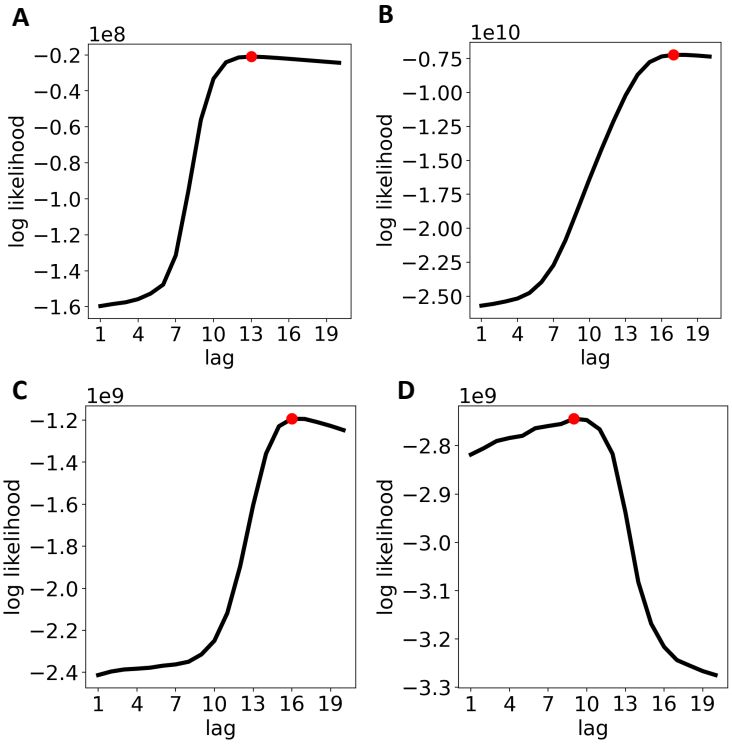

Figure S15: Marginal log likelihood under the vanilla BEAR model as a function of lag L for the bacteriophage YSD1 (A), glioblastoma GBM (B), control metagenomic HC (C) and bacteria Bact. (D) datasets. Note the large scale (upper left) of each plot.

N Visualization details

Here we provide details on the results reported in the **Visualizing data** subsection of the results (Section 6).

N.1 Latent representation model

As a local latent representation model, we used a categorical probabilistic principal component analysis (pPCA) model, with automatic relevance determination [38, 60]. We trained on kmers (k_t, b_t) of length $L+1 = 18$ and used $D = 20$ latent dimensions. The complete model was,

$$\begin{aligned} \kappa_d &\sim \text{Exponential for } d \in \{1, \ldots, D\} \\ W_d &\sim \text{Normal}(0_{L+1,|\mathcal{B}|}, 1/\kappa_d) \text{ for } d \in \{1, \ldots, D\} \\ W_0 &\sim \text{Normal}(0_{L+1,|\mathcal{B}|}, 1) \\ z_t &\sim \text{Normal}(0_D, 1) \\ (k_t, b_t) &\sim \text{Categorical}\left(\text{softmax}(W \cdot z_t + W_0)\right) \end{aligned} \tag{62}$$

where $t \in \{1, \ldots, T\}$ runs over all length $L+1$ kmers in the dataset, $0_{L+1,|\mathcal{B}|}$ is an $L+1 \times |\mathcal{B}|$ matrix of zeros, and 0_D is a length D vector of zeros. Here the local variable z_t provides a representation associated with the kmer (k_t, b_t), the global parameter W controls the factors of variation, and κ determines the relevance of each factor through the variance of the prior on W. We trained this latent representation model, and embedded it into a BEAR model, in three stages.

Stage 1 First, we performed stochastic variational inference to learn the parameters of the model [34, 37, 50]. In particular, we used normally distributed mean field posterior approximations $q(W), q(z|k, b)$, and a deterministic approximation to κ, and optimized the

evidence lower bound (ELBO)

$$\mathbb{E}_{W\sim q(W)}\Bigg[\sum_{k,b}\#(k,b)\left(\mathbb{E}_{z\sim q(z|k,b)}\log p\left(k,b|W,z\right)-\mathrm{KL}(q(z|k,b)||p(z))\right) + \log p\left(W|\kappa\right)-\mathrm{KL}(q(W)||p(W))\Bigg]+\log p(\kappa) \tag{63}$$

where $\#(k,b)$ denotes the number of kmers (k,b) seen in the data and the sum runs over all $k\in\mathcal{B}_L^o$, $b\in\tilde{\mathcal{B}}$. For the local latent variable z, we use a guide (recognition network) $q(z|k,b)=\mathrm{Normal}(\mu(k,b),\sigma(k,b))$ where $\mu(k,b)$ and $\sigma(k,b)$ are each small CNNs. Gradients with respect to the variational approximation parameters were taken using automatic differentiation and the reparameterization trick (elliptical standardization), with one sample for the Monte Carlo approximation at each step.

Stage 2 Once the pPCA model was trained, we approximated its conditional distribution. In particular, we obtained a variational approximation to $p(z|k,(k_t,b_t)_{t=1}^T)$, namely $q(z|k)$, by optimizing the evidence lower bound

$$\mathbb{E}_{W\sim q(W)}\left[\sum_{k}\#k\left(\mathbb{E}_{z\sim q(z|k)}\log p\left(k|W,z\right)-\mathrm{KL}(q(z|k)||p(z))\right)\right]. \tag{64}$$

Note that $q(W)$ was held fixed, at the value learned in stage 1. $q(z|k)$ was parameterized analogously to $q(z|k,b)$. Now we can approximate the conditional distribution of the pPCA model as

$$p(b|k)\approx\mathbb{E}_{W\sim q(W)}\mathbb{E}_{z\sim q(z|k)}p(b|W,z).$$

This defines an AR model.

Stage 3 Finally, we embedded the conditional pPCA AR model into a BEAR model and optimized h via empirical Bayes (note that here we are not using empirical Bayes to train the BEAR model's embedded AR parameters θ, but instead embedding a pretrained AR model). Since the variational distribution $q(W)$ was highly concentrated at a single point, we used a computationally convenient approximation to the marginal likelihood of the BEAR model, moving the expectation over the global parameters outside the log marginal likelihood:

$$\mathbb{E}_{W\sim q(W)}\left[\sum_{k}\log\mathrm{DirichletCategorical}\left(\#(k,\cdot)|\frac{1}{h}\mathbb{E}_{z\sim q(z|k)}p(b|W,z)\right)\right]$$

where $\mathrm{DirichletCategorical}\left(\#(k,\cdot)|\alpha_k\right)$ denotes the probability of the count vector $\#(k,\cdot)$ under a Dirichlet-Categorical distribution with concentration vector α_k.

Training protocol and hyperparameters The entire variational inference and embedding procedure was implemented using the Edward2 [61] probabilistic programming language with a TensorFlow [2] back-end. We applied the method to the Hodgkin's lymphoma single cell RNAseq described in section K, using the same train/test split as for the performance results in Section L. Optimization was performed with Adam with a batch size of $125,000$. Gradients were accumulated over 200 steps. The three stages of training described above were repeated iteratively four times until each converged. In each iteration, the first two stages were trained for 5 epochs, and we used a decaying learning rate across iterations $\{0.02,0.02,0.01,0.005\}$; the third stage was trained for 100 batches with a constant learning rate of 0.1 across all iterations.

Inference results At the end of training, the BEAR model had a perplexity of 4.276 on heldout data.

N.2 Visualization and annotation

We next sought to understand in greater depth what the BEAR model had learned in the lymphoma dataset.

Reference model We first aimed to understand how the model's predictions differed from predictions based on the reference transcriptome. On the full dataset (combined train/test)

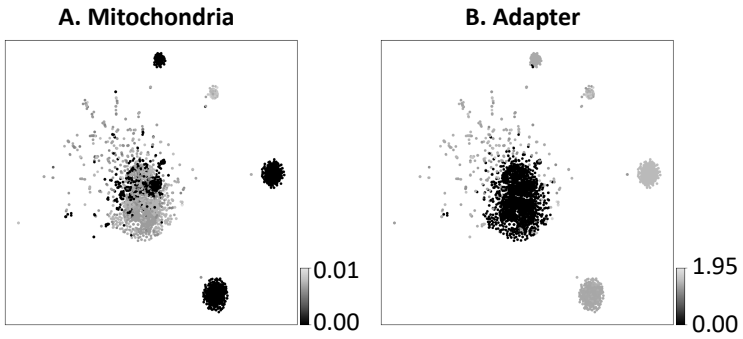

Figure S16: tSNE visualization of a cluster of single cell RNAseq reads colored by (A) latent embedding distance to the mitochondrial reference genome and (B) latent embedding distance from the sequencing adapter.

we compared the log probability of each read under the pPCA BEAR model to the log probability of each read under a vanilla BEAR model trained on the reference transcriptome (Figure 3C; see Datasets.xlsx for details on the reference transcriptome). We found a substantial disparity between the two model's predictions, with a number of reads having high probability under the BEAR model but low probability according to the reference model.

Alignments Single cell RNAseq analysis often begins by aligning reads to the reference transcriptome; reads that do not align are typically discarded from further analysis. We performed alignments on the read dataset with hisat2 [32] using parameters `-reorder -no-hd -n-ceil L,0,0.001 -no-sq -k 1 -p 4` and with the default hisat2 *Homo sapiens* GRCh38 genome index with transcripts and SNPs, available at `https://genome-idx.s3.amazonaws.com/hisat/grch38_snptran.tar.gz`. Whether or not each read was successfully aligned is indicated in Figure 3C. We observe that many of the reads with low probability under both the pPCA BEAR model and the reference model are unaligned. We also observed a cluster with a large number of unaligned reads, with high probability under the pPCA BEAR model and relatively low probability under the reference model. We focused on a subset of this cluster with particularly high probabilities under the pPCA BEAR model for follow-up visualization (black box in Figure 3C).

Visualization The pPCA model provides a latent embedding of kmers in a $D = 20$ dimensional continuous space. We sought to visualize the representation of each sequence's kmers in a low dimensional space. To compare two sequences X, X', we defined a measure of dissimilarity,

$$\inf_{i,i'} \mathrm{KL}(q(z|X_{i-L:i})||q(z|X'_{i'-L:i'})) + \mathrm{KL}(q(z|X'_{i'-L:i'})||q(z|X_{i-L:i})).$$

where $i > L$ and $i' > L$ index positions in X and X' respectively. This dissimilarity measure was used to define a distance matrix over reads in the Hodgkin's lymphoma dataset, which was passed to tSNE [63] to obtain a low-dimensional visualization (Figure 3D).

Annotation Observing the clusters in Figure 3D, we sought to determine where the reads in each cluster likely originated from, and, by implication, what the reference transcriptome model had trouble explaining in the data. We started by using NCBI's BLAST tool [8] to search for likely sources, and found hits against the mitochondrial genome and the transcript of the gene *JUND*, part of the AP-1 early response transcription factor. We found that the mitochondrial reads are from a nonreference haplotype, which explains why the reference model gave them low probability. The low likelihood of the *JUND* reads under the reference was due to a TG repeat region in the 3' UTR; similar repeats are present in many variations in different transcripts, thus the particular kmer-base transitions in this case become less likely. We also observed that many reads were chimeric, consisting of fusions of sequences from various parts of the transcriptome with some portion of the sequence CTGTCTCTTAT-ACACATCTCTGAACGGGCTGGCAAGGCAGACCG. The prefix CTGTCTCTTATACA-CATCT is a standard Illumina Nextera adapter sequence `https://support-docs.illumina.`

`com/SHARE/AdapterSeq/illumina-adapter-sequences.pdf`, and the remainder of the sequence is presumably part of the primer. The adapter is an experimental artifact (presumably left in the read data due to inaccurate read trimming and quality control), and so is not part of the reference human transcriptome.

We used the same dissimilarity measure as above to compare reads to the mitochondria reference genome (Datasets.xlsx) and to the adapter sequence CTGTCTCTTATACACATCTCT-GAACGGGCTGGCAAGGCAGACCG (Figure S16). (The distance to each of these sequences was taken to be the minimum of the distance to the forward and reverse complements.) Figure S16, along with the BLAST results for *JUND*, were the basis for the annotations in Figure 3D.

O Hypothesis tests details

Here we provide details on the results reported in the **Testing hypotheses** subsection of the results (Section 6).

O.1 Kidney transplant metagenomics

The Schreiber et al. [55] data is available for public download, as detailed in Datasets.xlsx. The read data was pre-sorted into viral and non-viral reads, but we pooled each of these to reconstruct the full sequencing experiment. We compared the day zero timepoint, i.e. before transplant, to the 4-6 week timepoint, i.e. after transplant, for each patient for which samples from both were available (note this did not include all patients in the study). We used the BEAR two-sample test, with the Jeffreys prior on v, and a truncated uniform prior over lags $1 \leq L \leq 20$. We cross referenced our two-sample test results with whether Schreiber et al. [55] determined there to be likely JC polyomavirus (JCPyV) transmission.

The results are shown in Table S6, and suggest that JCPyV transmission is associated with an overall shift in the patient microbiome at the sequence level. Patients indicated with an asterisk were diagnosed as having JCPyV before receiving the transplant, and thus the determination of whether the transplant transmitted JCPyV is less certain; for patient wdk036, phylogenetic analysis suggested that the transplant did transmit JCPyV, while for jns976 phylogenetic analysis suggested that it did not. Although the two-sample test results show close correlation with whether or not there was transmission, we caveat them by noting that for very small lags the Bayes factor rejects the null hypothesis for all patients; the question of the most "biologically relevant" prior on the lag L is an open question.

O.2 *A. thaliana* hypothesis tests

Goodness-of-fit test We trained reference-based AR models (described in Section L.1) via maximum likelihood on each *A. thaliana* sequencing dataset (the full dataset, with train/test subsets combined). We used $L = 17$ in the AR model for all three datasets (corresponding the vanilla BEAR maximum marginal likelihood lag for two datasets, see Table S4). We embedded each trained AR model into a BEAR model to construct a goodness-of-fit test (i.e. we used the learned $f(\theta)$). We fixed $L = 17$ in the BEAR model (i.e. a deterministic prior over L) to determine if there was misspecification at the same resolution as the AR model. Figure 3E plots the Bayes factor as a function of h.

Two-sample tests We simulated sequencing reads based on the *A. thaliana* reference genome (Datasets.xlsx) using the ART Illumina [29] simulator with parameters `-ss HS20 -p -l 100 -m 200 -s 10 -f 30`. We simulated roughly the same number of reads as was in each real dataset. We examined the Bayes factor $\mathrm{BF}(L) = p((X_n)_{n=1}^N|L)p((X'_n)_{n=1}^{N'}|L)/p((X_n)_{n=1}^N, (X'_n)_{n=1}^{N'}|L)$, computed using vanilla BEAR models for each term (Figure 3F). As control experiments, we cut each dataset (and the simulated data) in half, and compared each of these halves to each other using the same two-sample test; as shown by the dotted lines in Figure 3F, the two-sample test correctly accepts the null hypothesis in these cases.

Table S6: BEAR two-sample test results, performed on patient metagenome samples from before and after kidney transplant. Bayes factors that reject the null hypothesis are colored red, for easy comparison with whether or not JC polyomavirus (JCPyV) transmission was detected. Asterisks * indicate patients that were already infected with JCPyV before the transplant occurred.

Patient id	JCPyV transmission	log Bayes factor
ume111	True	110407
vpi912	False	234361
iwv346	False	-955252
pqg516	False	-504784
tvy653	True	70223
bgk952	False	-357457
wdk036*	True	3152401
jns976*	False	-199006
aag951	True	242877
qfv506	False	-155391
qnx429	True	369129
poo581	False	-290382
xph346	False	-254856
mek642	False	-348120

Individual log likelihood ratio To understand in detail the differences between the real and simulated data, we computed the conditional individual Bayes factor $\log p(X_n|(X_n)_{n=1}^N) - \log p(X_n|(X'_n)_{n=1}^{N'})$ where $(X_n)_{n=1}^N$ is the real data and $(X'_n)_{n=1}^{N'}$ the simulated data. We approximated the log likelihood using the maximum *a posteriori* value of the transition parameter v under the vanilla BEAR model, and fixed $L = 17$. Computing this likelihood efficiently for each read requires retrieving counts $\#(k, \cdot)$ for each kmer k in the read, which we accomplished using the Jellyfish kmer indexing package [42]. Histograms of the log likelihood ratio of each read X_n in two of the *A. thaliana* datasets are shown in Figure 3G (gray).

Annotation Observing the distinct peaks in Figure 3G, we sought to determine where the reads in each originated from. We discovered that many reads in the outlier peak from *A. thaliana* 1 matched *Bacillus cereus*, using NCBI's BLAST tool [8]. To annotate the clusters further, we aligned the reads to reference sequences for centromeres, chloroplasts, and *B. cereus*, as well as (if the read did not align to one of these) the reference *A. thaliana* genome (reference sequences are listed in Datasets.xlsx). Alignments were performed using `hisat2` on paired end read data using parameters `-reorder -no-hd -n-ceil L,0,0.001 -no-sq -k 1 -p 4` to facilitate subsequent analysis and remove reads with ambiguous bases. The alignment to the centromere included the parameter `-mp 1,1` to allow lower quality alignments. Histograms of the set of reads that align to each reference are shown (stacked on top of one another, not overlayed) in Figure 3G.