Variational Inference for the Indian Buffet Process

Finale Doshi-Velez, Kurt T. Miller, Jurgen Van Gael, Yee Whye Teh

Computational and Biological Learning Laboratory
Department of Engineering
University of Cambridge

Technical Report CBL-2009-001
May 2009

Variational Inference for the Indian Buffet Process

Finale Doshi-Velez* Cambridge University
Kurt T. Miller* University of California, Berkeley
Jurgen Van Gael* Cambridge University
Yee Whye Teh Gatsby Unit

May 2009

Abstract

The Indian Buffet Process (IBP) is a nonparametric prior for latent feature models in which observations are influenced by a combination of hidden features. For example, images may be composed of several objects and sounds may consist of several notes. Latent feature models seek to infer these unobserved features from a set of observations; the IBP provides a principled prior in situations where the number of hidden features is unknown. Current inference methods for the IBP have all relied on sampling. While these methods are guaranteed to be accurate in the limit, samplers for the IBP tend to mix slowly in practice. We develop a deterministic variational method for inference in the IBP based on truncating to finite models, provide theoretical bounds on the truncation error, and evaluate our method in several data regimes. This technical report is a longer version of Doshi-Velez et al. (2009).

Keywords: Variational Inference, Indian Buffet Process, Nonparametric Bayes, Linear Gaussian, Independent Component Analysis

1 Introduction

Many unsupervised learning problems seek to identify a set of unobserved, co-occurring features from a set of observations. For example, given images composed of various objects, we may wish to identify the set of unique objects and determine which images contain which objects. Similarly, we may wish to extract a set of notes or chords from an audio file as well as when each note was played. In scenarios such as these, the number of latent features is often unknown a priori.

Unfortunately, most traditional machine learning approaches require the number of latent features as an input. To apply these traditional approaches to scenarios in which the number of latent features is unknown, we can use model selection to define and manage the trade-off between model complexity and model fit. In contrast, nonparametric Bayesian approaches treat the number of features as a random quantity to be determined as part of the posterior inference procedure.

The most common nonparametric prior for latent feature models is the Indian Buffet Process (IBP) (Griffiths and Ghahramani, 2005). The IBP is a prior on infinite binary matrices that allows us to simultaneously infer which features influence a set of observations and how many features there are. The form of the prior ensures that only a finite number of features will be present in any finite set of observations, but more features may appear as more observations are received. This property is both natural and desirable if we consider, for example, a set of

*Authors contributed equally.

images: any one image contains a finite number of objects, but, as we see more images, we expect to see objects not present in the previous images.

While an attractive model, the combinatorial nature of the IBP makes inference particularly challenging. Even if we limit ourselves to K features for N objects, there exist 2^{NK} possible feature assignments. In this combinatorial space, sampling-based inference procedures for the IBP often suffer because they assign specific values to the feature-assignment variables. Hard variable assignments give samplers less flexibility to move between optima, and the samplers may need large amounts of time to escape small optima and find regions with high probability mass. Unfortunately, all current inference procedures for the IBP rely on sampling. These approaches include Gibbs sampling (Griffiths and Ghahramani, 2005), which may be augmented with Metropolis split-merge proposals (Meeds et al., 2007), as well as slice sampling (Teh et al., 2007) and particle filtering (Wood and Griffiths, 2007).

Mean field variational methods, which approximate the true posterior via a simpler distribution, provide a deterministic alternative to sampling-based approaches. Inference involves using optimisation techniques to find a good approximate posterior. Our mean field approximation for the IBP maintains a separate probability for each feature-observation assignment. Optimising these probability values is also fraught with local optima, but using the soft variable assignments—that is, using probabilities instead of sampling hard assignments—gives the variational method a flexibility that samplers lack. In the early stages of the inference, this flexibility can help the variational method avoid small local optima.

Variational approximations have provided benefits for other nonparametric Bayesian models, including Dirichlet Processes (e.g. Blei and Jordan (2004), Kurihara et al. (2007a) and Kurihara et al. (2007b)) and Gaussian Processes (e.g. Winther (2000), Gibbs and MacKay (2000), and Snelson and Ghahramani (2006)). Of all the nonparametric Bayesian models studied so far, however, the IBP is the most combinatorial and is therefore in the most need of a more efficient inference algorithm.

We will show how our variational approximation is better in terms of both computation required and predictive likelihood than the sampling based approaches as we get more data and as the dimensionality of the data grows.

The remainder of this technical report is organised as follows:

- **Background.** Section 2 reviews the likelihood model that we will use throughout the technical report, the Indian Buffet Process (IBP), and the basic Gibbs sampler for the IBP. It also summarises the notation used in later sections. Appendix A details the Gibbs sampling equations used in our tests.

- **Variational Method.** Section 3 reviews the variational inference framework and outlines our specific approach which is based on a truncated representation of the IBP. Sections 4 and 5 contain all of the update equations required to implement the variational inference; full derivations are provided in Appendices C and D. Appendix E describes how similar updates can be derived for the infinite ICA model of Knowles and Ghahramani (2007).

- **Truncation Bound.** Section 6 derives bounds on the expected error due to our use of a truncated representation for the IBP; these bounds can serve as guidelines for what level of truncation may be appropriate. Extended derivations are left to Appendix F.

- **Results.** Section 7 demonstrates how our variational approach scales to high-dimensional data sets while maintaining good predictive performance.

2 Background

Section 2.1 describes the likelihood model that we will use throughout this report, Section 2.2 provides background information on the Indian Buffet Process and Section 2.3 summarises the notation used in the remainder of the report.

2.1 Data Model

Let $\boldsymbol{X}$ be an $N \times D$ matrix where each of the N rows contains a D-dimensional observation. In this report, we consider a likelihood model in which $\boldsymbol{X}$ can be approximated by $\boldsymbol{ZA}$ where $\boldsymbol{Z}$ is an $N \times K$ binary matrix and $\boldsymbol{A}$ is a $K \times D$ matrix. Each column of $\boldsymbol{Z}$ corresponds to a latent feature where $z_{nk} \equiv \boldsymbol{Z}(n,k) = 1$ if feature k is present in observation n and 0 otherwise. The parameters for feature k are stored in row k of $\boldsymbol{A}$. The observed data $\boldsymbol{X}$ is then given by $\boldsymbol{ZA} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is some measurement noise (see Figure 1). We assume that the noise is independent of $\boldsymbol{Z}$ and $\boldsymbol{A}$ and is uncorrelated across observations.

An important feature of this model is that only the nonzero columns of $\boldsymbol{Z}$ affect the likelihood. This will be important when we place a prior on the binary feature matrix $\boldsymbol{Z}$ in which $\boldsymbol{Z}$ has an infinite number of columns, but only a finite number of entries will be nonzero. Since the all-zero columns do not affect the likelihood, we will only need to reason about the nonzero columns.

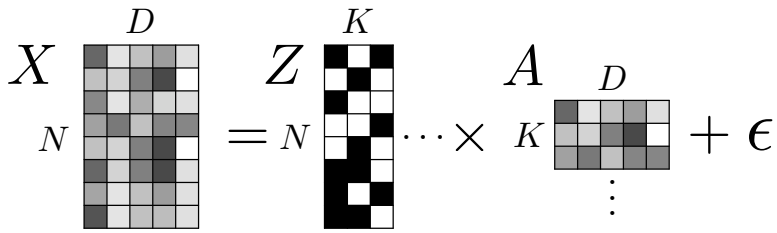


Figure 1: Our likelihood model posits that the data $\boldsymbol{X}$ is the product $\boldsymbol{ZA}$ plus some noise.

Given $\boldsymbol{X}$ and being in a Bayesian framework, we wish to find the posterior distribution of $\boldsymbol{Z}$ and $\boldsymbol{A}$. From Bayes rule,

$$p(\boldsymbol{Z}, \boldsymbol{A}|\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{A})p(\boldsymbol{Z})p(\boldsymbol{A})$$

where we have assumed that $\boldsymbol{Z}$ and $\boldsymbol{A}$ are independent a priori. The specific application will determine the likelihood function $p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{A})$ and the feature prior $p(\boldsymbol{A})$. In the linear-Gaussian model which will be the focus of this report, both the noise $\boldsymbol{\epsilon}$ and the features $\boldsymbol{A}$ have Gaussian priors. In Appendix E, we discuss a different likelihood model and show how to apply our variational approximation to this model.

We are left with placing a prior on $\boldsymbol{Z}$. Since we often do not know K, we desire a prior that allows the number of nonzero columns to be determined at inference time. The Indian Buffet Process is one option for such a prior.

2.2 Indian Buffet Process

The IBP places the following prior on $[\boldsymbol{Z}]$, a canonical form of $\boldsymbol{Z}$ that is invariant to the ordering of the features (see Griffiths and Ghahramani (2005) for details):

$$p([\boldsymbol{Z}]) = \frac{\alpha^K}{\prod_{h \in \{0,1\}^N \setminus \mathbf{0}} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K} \frac{(N - m_k)!(m_k - 1)!}{N!}. \tag{1}$$

Here, K is the number of nonzero columns in $\boldsymbol{Z}$, m_k is the number of ones in column k of $\boldsymbol{Z}$, H_N is the N^{th} harmonic number, and K_h is the number of occurrences of the nonzero binary vector h among the columns in $\boldsymbol{Z}$. The parameter α controls the expected number of features present in each observation. For any fixed setting of α, with probability one, a matrix $\boldsymbol{Z}$ drawn from this prior will have a finite number of nonzero entries.

Restaurant Construction. The following culinary metaphor is one way to sample a matrix $\boldsymbol{Z}$ from the prior described in Equation (1). Imagine the rows of $\boldsymbol{Z}$ correspond to customers and the columns correspond to dishes in an infinitely long (Indian) buffet. The customers choose their dishes as follows:

1. The first customer takes the first $\mathsf{Poisson}(\alpha)$ dishes.
2. The i^{th} customer then takes dishes that have been previously sampled with probability m_k/i, where m_k is the number of people who have already sampled dish k. The i^{th} customer also takes $\mathsf{Poisson}(\alpha/i)$ new dishes.

Then, z_{nk} is one if customer n tried the k^{th} dish and zero otherwise. The resulting matrix is infinitely exchangeable, meaning that the order in which the customers enter the buffet has no impact on the distribution of $\boldsymbol{Z}$ (up to permutations of the columns).

The Indian buffet metaphor leads directly to a Gibbs sampler for posterior inference. Bayes' rule states

$$p(z_{nk}|\boldsymbol{Z}_{-nk}, \boldsymbol{A}, \boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{Z})p(z_{nk}|\boldsymbol{Z}_{-nk}).$$

The likelihood term $p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{Z})$ is easily computed from the noise model while the prior term $p(z_{nk}|\boldsymbol{Z}_{-nk})$ is obtained by imagining that customer n was the last to enter the restaurant (this assumption is valid due to exchangeability). The prior term $p(z_{nk}|\boldsymbol{Z}_{-nk})$ is m_k/N for active features. New features are sampled by combining the likelihood model with the $\mathsf{Poisson}(\alpha/N)$ prior on the number of new dishes a customer will try.

If the prior on $\boldsymbol{A}$ is conjugate to the likelihood, we can marginalise out $\boldsymbol{A}$ from the likelihood $p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{A})$ and consider $p(\boldsymbol{X}|\boldsymbol{Z})$. This approach leads to a collapsed Gibbs sampler for the IBP. Marginalising out $\boldsymbol{A}$ gives the collapsed Gibbs sampler a level of flexibility that allows it to mix more quickly than an uncollapsed Gibbs sampler.

However, if the likelihood is not conjugate or if the dataset is large and high-dimensional, $p(\boldsymbol{X}|\boldsymbol{Z})$ may be much more expensive to compute than $p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{A})$. In these cases, the Gibbs sampler must also sample the feature matrix $\boldsymbol{A}$ based on its posterior distribution $p(\boldsymbol{A}|\boldsymbol{X}, \boldsymbol{Z})$. For further details and equations on Gibbs samplers for the IBP, please refer to Appendix A.

Stick-breaking Construction. The restaurant construction directly lends itself to a Gibbs sampler, but it does not easily lend itself to a variational approach. For the variational approach, we turn to an equivalent alternative construction of the IBP, the stick-breaking construction of Teh et al. (2007). To generate a matrix $\boldsymbol{Z}$ using the stick-breaking construction, we begin by assigning a parameter $\pi_k \in [0, 1]$ to each column of $\boldsymbol{Z}$. Given π_k, each z_{nk} in column k is sampled as an independent $\mathsf{Bernoulli}(\pi_k)$. Since each 'customer' samples a dish independently of the other customers, this representation makes it clear that the ordering of the customers does not impact the distribution.

The π_k themselves are generated by the following stick-breaking process. We first draw a sequence of independent random variables $v_1, v_2, \ldots$, each distributed $\mathsf{Beta}(\alpha, 1)$. We assign $\pi_1 = v_1$. For each subsequent k, we assign $\pi_k = v_k\pi_{k-1} = \prod_{i=1}^{k} v_i$, resulting in a decreasing sequence of probabilities π_k. In expectation, the probability of seeing feature k decreases exponentially with k. The parameter α affects how quickly these probabilities decrease. Larger values of α

mean that the values π_k decrease more slowly, which means that we expect to see more features in the data. For more details, see (Teh et al., 2007).

2.3 Notation

We now summarise the notation which we use throughout the technical report. Vectors or matrices of variables are bold face. A subscript of "$-i$" indicates all components except component i. A subscript "$\cdot$" indicates all components in a given dimension. For example, $\boldsymbol{Z}_{-nk}$ is the full $\boldsymbol{Z}$ matrix except the (n,k) entry, and $\boldsymbol{X}_{n\cdot}$ is the entire n^{th} row of $\boldsymbol{X}$. For probability distributions, a subscript indicates the parameters used to specify the distribution. For example, $q_{\boldsymbol{\tau}}(\boldsymbol{v}) = q(\boldsymbol{v};\boldsymbol{\tau})$.

Commonly recurring variables are:

- $\boldsymbol{X}$: The observations are stored in $\boldsymbol{X}$, an $N \times D$ matrix. The linear-Gaussian model posits $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{A} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a $N \times D$ matrix of independent elements, each with mean 0 and variance σ_n^2.
- $z_{nk}, \boldsymbol{Z}$: Each z_{nk} indicates whether feature k is present in observation n. Here, $n \in \{1 \dots N\}$ runs over the number of data-points and $k \in \{1 \dots \infty\}$ runs over the number of features. The matrix $\boldsymbol{Z}$ refers to the collection of all z_{nk}'s. It has dimensionality $N \times K$, where K is the finite number of nonzero features. All other $z_{nk}, k > K$ are assumed to be zero. We use α to denote the concentration parameter of the IBP.
- $\boldsymbol{\pi}$: The stick lengths (feature probabilities) are π_k.
- $\boldsymbol{\nu}$: The stick-breaking variables are ν_k.
- $\boldsymbol{A}$: The collection of Gaussian feature variables, a $K \times D$ matrix where each feature is represented by the vector $\boldsymbol{A}_{k\cdot}$. In the linear-Gaussian model, the prior states that the elements are $\boldsymbol{A}$ are independent with mean 0 and variance σ_A^2.

3 Variational Inference

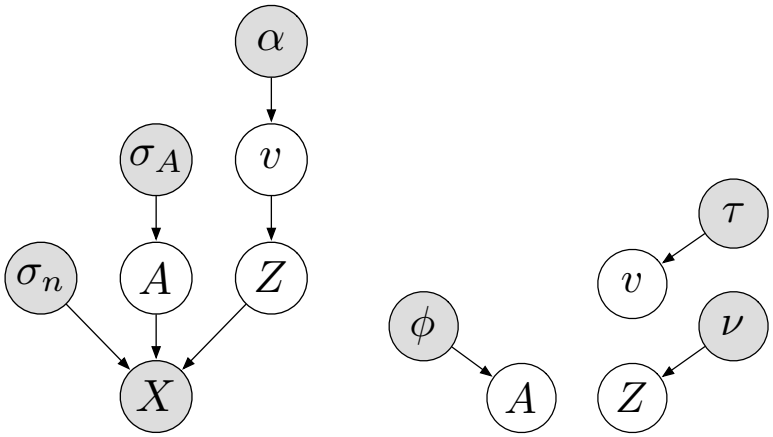
We will focus on variational inference procedures for the linear-Gaussian likelihood model (Griffiths and Ghahramani, 2005), in which $\boldsymbol{A}$ and $\boldsymbol{\epsilon}$ are Gaussian. These updates can be easily adapted to other exponential family likelihood models. As an example, we briefly discuss the variational procedure for the infinite ICA model (Knowles and Ghahramani, 2007) in Appendix E.

We denote the set of hidden variables in the IBP by $\boldsymbol{W} = \{\boldsymbol{\pi}, \boldsymbol{Z}, \boldsymbol{A}\}$ and the set of parameters by $\boldsymbol{\theta} = \{\alpha, \sigma_A^2, \sigma_n^2\}$. Computing the true log posterior

$$\log p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}) = \log p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) - \log p(\boldsymbol{X}|\boldsymbol{\theta})$$

is difficult due to the intractability of computing the log marginal probability $\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \log \int p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta}) d\boldsymbol{W}$.

Mean field variational methods approximate the true posterior with a *variational distribution* $q_{\boldsymbol{\Phi}}(\boldsymbol{W})$ from some tractable family of distributions Q (Beal, 2003; Wainwright and Jordan, 2008). Here, $\boldsymbol{\Phi}$ denotes the set of parameters used to describe the distribution q. Inference then reduces to performing an optimisation on the parameters $\boldsymbol{\Phi}$ to find the member $q \in Q$ that minimises the KL divergence $D(q_{\boldsymbol{\Phi}}(\boldsymbol{W})||p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}))$. Since the KL divergence $D(q||p)$ is nonnegative and equal to zero iff $p = q$, the unrestricted solution to our problem is to set



(a) True model $p(\boldsymbol{W})$ (b) Variational approximation $q(\boldsymbol{W})$

Figure 2: Graphical models for the true posterior distribution of $\boldsymbol{W}$ and for our infinite variational approximation.

$q_{\boldsymbol{\Phi}}(\boldsymbol{W}) = p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta})$. However, this general optimisation problem is intractable. We therefore restrict Q to a parameterised family of distributions for which this optimisation is tractable.

Specifically, we present two mean field variational approaches with two different families Q. In both models, we use a truncated model with truncation level K. A truncation level K means that $\boldsymbol{Z}$ (in our approximating distribution) is nonzero in at most K columns.

Our first approach minimises the KL-divergence $D(q||p_K)$ between the variational distribution and a finite approximation p_K to the IBP described in Section 4; we refer to this approach as the *finite variational* method. In this model, we let Q be a factorised family

$$q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{\pi}) q_{\boldsymbol{\phi}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z}) \tag{2}$$

where $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, and $\boldsymbol{\nu}$ are optimised to minimise $D(q||p_K)$. By minimising the KL-divergence with respect to p_K and not the true p, this first approach introduces an additional layer of approximation not present in our second approach. We do not show the graphical model for this approach, but it is equivalent to the one shown in Figure 2 if we replace v with π and let the true model be the truncated beta-Bernoulli.

Our second approach minimises the KL-divergence to the true IBP posterior $D(q||p)$. We call this approach the *infinite variational* method because, while our variational distribution is finite, its updates are based the true IBP posterior (which contains an infinite number of features). In this model, we work directly with the stick-breaking weights $\boldsymbol{v}$ instead of directly with $\boldsymbol{\pi}$. The family Q is then the factorised family

$$q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{v}) q_{\boldsymbol{\phi}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z})$$

where $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, and $\boldsymbol{\nu}$ are the variational parameters. The forms of the distributions q and the variational updates are specified in Section 5. The graphical model for this variational approximation is shown in Figure 2.

Inference in both approaches consists of optimising the parameters of the approximating distribution to most closely match the true posterior. This optimisation is equivalent to maximising a lower bound on the evidence since

$$\begin{aligned} \log p(\boldsymbol{X}|\boldsymbol{\theta}) &= \mathbb{E}_q[\log(p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})] + H[q] + D(q||p) \\ &\geq \mathbb{E}_q[\log(p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})] + H[q] \end{aligned} \tag{3}$$

where $H[q]$ is the entropy of distribution q, and therefore

$$\arg\min_{\tau,\phi,\nu} D(q||p) = \arg\max_{\tau,\phi,\nu} \mathbb{E}_q[\log(p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})] + H[q]. \tag{4}$$

This optimisation is not convex; in general, we can only hope to find variational parameters that are local optima.

To minimise $D(q||p)$, we cycle through each of the variational parameters, and for each one, perform a coordinate ascent that maximises the right side of Equation (4). In doing so, we also improve a lower bound on the log-likelihood of the data.

In Sections 4 and 5, we go over the finite and infinite approaches in detail assuming we have chosen our truncation level K. Section 6 presents a bound describing how close a truncated stick-breaking approximation of the IBP is to the true IBP, thereby giving a heuristic for how to choose the truncation level K in both models. This is the first such bound known for the IBP. Appendix B reviews the key concepts for variational inference with exponential family models.

4 The Finite Variational Approach

In this section, we introduce our *finite variational approach*, an approximate inference algorithm for an approximation to the IBP. Specifically, we assume that the IBP can be well approximated using the finite beta-Bernoulli model p_K introduced by (Griffiths and Ghahramani, 2005)

$$\begin{aligned}
\pi_k &\sim \mathsf{Beta}(\alpha/K, 1) && \text{for } k \in \{1 \dots K\},\\
z_{nk} &\sim \mathsf{Bernoulli}(\pi_k) && \text{for } k \in \{1 \dots K\}, n \in \{1 \dots N\},\\
\boldsymbol{A}_{k\cdot} &\sim \mathsf{Normal}(0, \sigma_A^2 I) && \text{for } k \in \{1 \dots K\},\\
\boldsymbol{X}_{n\cdot} &\sim \mathsf{Normal}(\boldsymbol{Z}_{n\cdot}\boldsymbol{A}, \sigma_n^2 I) && \text{for } n \in \{1 \dots N\},
\end{aligned}$$

where K is some finite (but large) truncation level. Griffiths and Ghahramani (2005) showed that as $K \to \infty$, this finite approximation converges in distribution to the IBP. Under the finite approximation, the joint probability of the data and latent variables is

$$p_K(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \left(p(\pi_k|\alpha) p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I) \prod_{n=1}^{N} p(z_{nk}|\pi_k) \right) \prod_{n=1}^{N} p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot}, \boldsymbol{A}, \sigma_n^2 I).$$

Even in this simplified finite approximation, working with the log posterior

$$\log p_K(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}) = \log p_K(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) - \log p_K(\boldsymbol{X}|\boldsymbol{\theta}),$$

is intractable. We will therefore use the following variational distribution as an approximation

$$q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{\pi}) q_{\boldsymbol{\phi}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z})$$

where we assume there are at most K nonzero columns of Z and

- $q_{\boldsymbol{\tau}_k}(\pi_k) = \mathsf{Beta}(\pi_k; \tau_{k1}, \tau_{k2})$,
- $q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) = \mathsf{Normal}(\boldsymbol{A}_{k\cdot}; \bar{\boldsymbol{\phi}}_k, \boldsymbol{\Phi}_k)$,
- $q_{\nu_{nk}}(z_{nk}) = \mathsf{Bernoulli}(z_{nk}; \nu_{nk})$.

Inference then involves optimising $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, and $\boldsymbol{\nu}$ to either minimise the KL divergence $D(q||p_K)$ or, equivalently, maximise the following lower bound on $p_K(\boldsymbol{X}|\boldsymbol{\theta})$:

$$\mathbb{E}_q[\log(p_K(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})] + H[q].$$

While variational inference with respect to the finite beta-Bernoulli model p_K is not the same as variational inference with respect to the true IBP posterior, the variational updates are significantly easier and, in the limit of large K, the finite beta-Bernoulli model is equivalent to the IBP.

4.1 Lower Bound on the Marginal Likelihood

We expand the lower bound in this section, leaving the full set of equations for Appendix C.1. Note that all expectations in this section are taken with respect to the variational distribution q. We therefore drop the use of E_q and instead use $E_{\boldsymbol{W}}$ to indicate which variables we are taking expectations over. Substituting expressions into Equation (3), our lower bound is

$$\begin{aligned}
\log p_K(\boldsymbol{X}|\boldsymbol{\theta}) &\geq \mathbb{E}_{\boldsymbol{W}}\left[\log p_K(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})\right] + H[q], \\
&= \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\pi}}\left[\log p_K(\pi_k|\alpha)\right] + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}}\left[\log p_K(z_{nk}|\pi_k)\right] \qquad (5) \\
&\quad + \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{A}}\left[\log p_K(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I)\right] + \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[\log p_K(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot}, \boldsymbol{A}, \sigma_n^2 I)\right] + H[q].
\end{aligned}$$

Evaluating these expectations are all straightforward exponential family calculations. Expanding each of these, the full lower bound is

$$\begin{aligned}
&\log p_K(\boldsymbol{X}|\boldsymbol{\theta}) \qquad (6) \\
&\geq \sum_{k=1}^{K}\left[\log\frac{\alpha}{K} + \left(\frac{\alpha}{K} - 1\right)\left(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2})\right)\right] \\
&\quad + \sum_{k=1}^{K}\sum_{n=1}^{N}\left[\nu_{nk}\psi(\tau_{k1}) + (1-\nu_{nk})\psi(\tau_{k2}) - \psi(\tau_{k1}+\tau_{k2})\right] \\
&\quad + \sum_{k=1}^{K}\left[\frac{-D}{2}\log(2\pi\sigma_A^2) - \frac{1}{2\sigma_A^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right] \\
&\quad + \sum_{n=1}^{N}\left[-\frac{D}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T - 2\sum_{k=1}^{K}\nu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T + 2\sum_{k<k'}\nu_{nk}\nu_{nk'}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_{k'}^T + \sum_{k=1}^{K}\nu_{nk}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right)\right] \\
&\quad + \sum_{k=1}^{K}\left[\log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) - (\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\right] \\
&\quad + \sum_{k=1}^{K}\left[\frac{1}{2}\log\left((2\pi e)^D|\Phi_k|\right)\right] + \sum_{k=1}^{K}\sum_{n=1}^{N}\left[-\nu_{nk}\log\nu_{nk} - (1-\nu_{nk})\log(1-\nu_{nk})\right].
\end{aligned}$$

where $\psi(\cdot)$ is the digamma function. Full derivations are left to Appendix C.1.

4.2 Parameter Updates

Our approach to variational inference involves cycling through each of the variational parameters and sequentially updating them using standard exponential family variational update

equations. High level details of this approach can be found in Appendix B, Blei and Jordan (2004), and Wainwright and Jordan (2008). Derivations of the following update equations are in Appendix C.2.

1. For $k = 1, \ldots, K$, we update the $\bar{\boldsymbol{\phi}}_k$ and $\boldsymbol{\Phi}_k$ in $\mathsf{Normal}(\boldsymbol{A}_{k\cdot}; \bar{\boldsymbol{\phi}}_k, \boldsymbol{\Phi}_k)$ as

$$\begin{aligned}
\boldsymbol{\Phi}_k &= \left(\frac{1}{\sigma_A^2} + \frac{\sum_{n=1}^N \nu_{nk}}{\sigma_n^2}\right)^{-1} I \\
\bar{\boldsymbol{\phi}}_k &= \left[\frac{1}{\sigma_n^2}\sum_{n=1}^N \nu_{nk}\left(\boldsymbol{X}_{n\cdot} - \left(\sum_{l:l\neq k} \nu_{nl}\bar{\boldsymbol{\phi}}_l\right)\right)\right]\left(\frac{1}{\sigma_A^2} + \frac{\sum_{n=1}^N \nu_{nk}}{\sigma_n^2}\right)^{-1}.
\end{aligned}$$

2. For $k = 1, \ldots, K$, $n = 1, \ldots, N$, update ν_{nk} in $\mathsf{Bernoulli}(z_{nk}; \nu_{nk})$ as

$$\nu_{nk} = \frac{1}{1 + e^{-\vartheta}}.$$

where

$$\vartheta = \psi(\tau_{k1}) - \psi(\tau_{k2}) - \frac{1}{2\sigma_n^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right) + \frac{1}{\sigma_n^2}\bar{\boldsymbol{\phi}}_k\left(\boldsymbol{X}_{n\cdot}^T - \left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right)$$

3. For $k = 1, \ldots, K$, we update the τ_{k1} and τ_{k2} in $\mathsf{Beta}(\pi_k; \tau_{k1}, \tau_{k2})$ as

$$\begin{aligned}
\tau_{k1} &= \frac{\alpha}{K} + \sum_{n=1}^N \nu_{nk}, \\
\tau_{k2} &= N + 1 - \sum_{n=1}^N \nu_{nk}.
\end{aligned}$$

5 The Infinite Variational Approach

In this section, we introduce the *infinite variational approach*, a method for doing approximate inference for the linear-Gaussian model with respect to a full IBP prior. The model for p is the full (untruncated) stick-breaking construction for the IBP:

$$\begin{aligned}
v_k &\sim \mathsf{Beta}(\alpha, 1) && \text{for } k \in \{1, \ldots, \infty\}, \\
\pi_k &= \prod_{i=1}^k v_i && \text{for } k \in \{1 \ldots \infty\}, \\
z_{nk} &\sim \mathsf{Bernoulli}(\pi_k) && \text{for } k \in \{1 \ldots \infty\}, n \in \{1 \ldots N\}, \\
\boldsymbol{A}_{k\cdot} &\sim \mathsf{Normal}(0, \sigma_A^2 I) && \text{for } k \in \{1 \ldots \infty\}, \\
\boldsymbol{X}_{n\cdot} &\sim \mathsf{Normal}(\boldsymbol{Z}_{n\cdot}\boldsymbol{A}, \sigma_n^2 I) && \text{for } n \in \{1 \ldots N\}.
\end{aligned}$$

The joint probability of the data and variables is

$$p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) = \prod_{k=1}^\infty \left(p(\pi_k|\alpha)p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I)\prod_{n=1}^N p(z_{nk}|\pi_k)\right)\prod_{n=1}^N p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot}, \boldsymbol{A}, \sigma_n^2 I).$$

Working with the log posterior

$$\log p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}) = \log p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) - \log p(\boldsymbol{X}|\boldsymbol{\theta}).$$

is again intractable, so we use a variational approximation. Similar to the approach used by Blei and Jordan (2004), our variational approach uses a truncated stick-breaking process that truncates the number of nonzero columns to a finite value K. In the truncated stick-breaking process, $\pi_k = \prod_{i=1}^{k} v_i$ for $k \leq K$ and zero otherwise.

The ordered feature probabilities $\{\pi_1 \dots \pi_K\}$ are dependent in this model, while the $\{v_1 \dots v_K\}$ are independent. We therefore use $\boldsymbol{v}$ instead of $\boldsymbol{\pi}$ as our hidden variable because it is simpler to work with and given $\boldsymbol{v}$, it is straightforward to calculate $\boldsymbol{\pi}$. Our mean field variational distribution is then:

$$q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{v}) q_{\boldsymbol{\phi}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z})$$

where

- $q_{\boldsymbol{\tau}_k}(v_k) = \mathsf{Beta}(v_k; \tau_{k1}, \tau_{k2})$,
- $q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) = \mathsf{Normal}(\boldsymbol{A}_{k\cdot}; \bar{\boldsymbol{\phi}}_k, \Phi_k)$,
- $q_{\nu_{nk}}(z_{nk}) = \mathsf{Bernoulli}(Z_{nk}; \nu_{nk})$.

As with the finite approach, inference involves optimising $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, and $\boldsymbol{\nu}$ to minimise the KL divergence $D(q||p)$, or equivalently to maximise the lower bound on $p(\boldsymbol{X}|\boldsymbol{\theta})$

$$\mathbb{E}_q[\log(p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})] + H[q].$$

Unfortunately, the update equations for this approximation are not as straightforward as in the finite approach.

5.1 Lower Bound on the Marginal Likelihood

As in the finite approach, we first derive an expression for the variational lower bound. However, parts of our model are no longer in the exponential family and require nontrivial computations. We expand upon these parts here, leaving the straightforward exponential family calculations to Appendix D.1.

The lower bound on $p(\boldsymbol{X}|\boldsymbol{\theta})$ can be decomposed as follows

$$\begin{aligned} \log p(\boldsymbol{X}|\boldsymbol{\theta}) \quad \geq \quad & \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{v}} \left[\log p(v_k|\alpha)\right] + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\log p(Z_{nk}|\boldsymbol{v})\right] \\ & + \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{A}} \left[\log p(\boldsymbol{A}_{k\cdot}|\sigma_A^2)\right] + \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}} \left[\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}, \boldsymbol{A}, \sigma_n^2)\right] + H[q], \end{aligned} \tag{7}$$

Except for the second term, all of the terms are exponential family calculations; evaluated they

come out to

$$
\begin{aligned}
&\log p(\boldsymbol{X}|\boldsymbol{\theta}) \qquad (8)\\
&\geq \sum_{k=1}^{K}\left[\log\alpha+(\alpha-1)\left(\psi(\tau_{k1})-\psi(\tau_{k1}+\tau_{k2})\right)\right]\\
&\quad+\sum_{k=1}^{K}\sum_{n=1}^{N}\left[\nu_{nk}\left(\sum_{m=1}^{k}\psi(\tau_{k2})-\psi(\tau_{k1}+\tau_{k2})\right)+(1-\nu_{nk})\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k}v_m\right)\right]\right]\\
&\quad+\sum_{k=1}^{K}\left[\frac{-D}{2}\log(2\pi\sigma_A^2)-\frac{1}{2\sigma_A^2}\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right]\\
&\quad+\sum_{n=1}^{N}\left[-\frac{D}{2}\log(2\pi\sigma_n^2)-\frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T-2\sum_{k=1}^{K}\nu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T+2\sum_{k<k'}\nu_{nk}\nu_{nk'}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_{k'}^T+\sum_{k=1}^{K}\nu_{nk}\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right)\right]\\
&\quad+\sum_{k=1}^{K}\left[\log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right)-(\tau_{k1}-1)\psi(\tau_{k1})-(\tau_{k2}-1)\psi(\tau_{k1})+(\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\right]\\
&\quad+\sum_{k=1}^{K}\frac{1}{2}\log\left((2\pi e)^D|\Phi_k|\right)+\sum_{k=1}^{K}\sum_{n=1}^{N}\left[-\nu_{nk}\log\nu_{nk}-(1-\nu_{nk})\log(1-\nu_{nk})\right]
\end{aligned}
$$

where $\psi(\cdot)$ is the digamma function, and we have left $\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k}v_m\right)\right]$, a byproduct of the expectation of $\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}}\left[\log p(Z_{nk}|\boldsymbol{v})\right]$, unevaluated. This expectation has no closed-form solution, so we instead lower bound it (and therefore lower bound the log posterior).

In this section, we present a multinomial approximation which leads to a computationally efficient lower bound and straightforward parameter updates.[1] An approach based on a Taylor series expansion is presented in Appendix D.1. Unlike the multinomial approximation, the Taylor approximation can be made arbitrarily precise; however, empirically we find that the multinomial bound is usually only 2-10% looser than a fifty-term Taylor series expansion—and about thirty times faster to compute. Also, parameter updates under the Taylor approximation do not have a closed form solution and must be numerically optimised. Thus, we recommend using the multinomial approximation and the corresponding parameter updates; the Taylor derivation is provided in the Appendix D.1 largely for reference.

To bound $\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k}v_m\right)\right]$ with the multinomial approximation, we introduce an auxiliary $k-$multinomial distribution $q_k=(q_{k1},q_{k2},\ldots,q_{kk})$ into the expectation and apply Jensen's inequality:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k}v_m\right)\right] &= \mathbb{E}_{\boldsymbol{v}}\left[\log\left(\sum_{y=1}^{k}(1-v_y)\prod_{m=1}^{y-1}v_m\right)\right]\\
&= \mathbb{E}_{\boldsymbol{v}}\left[\log\left(\sum_{y=1}^{k}q_{ky}\frac{(1-v_y)\prod_{m=1}^{y-1}v_m}{q_{ky}}\right)\right]\\
&\geq \mathbb{E}_y\mathbb{E}_{\boldsymbol{v}}\left[\log(1-v_y)+\sum_{m=1}^{y-1}\log v_m\right]+H[q_k]\\
&= \mathbb{E}_y\left[\psi\left(\tau_{y2}\right)+\left(\sum_{m=1}^{y-1}\psi(\tau_{m1})\right)-\left(\sum_{m=1}^{y}\psi(\tau_{m1}+\tau_{m2})\right)\right]+H[q_k].
\end{aligned}
$$

[1]Note that Jensen's inequality cannot be used here; the concavity of the log goes in the wrong direction.

Explicitly writing out q_k, we get

$$\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right] \geq \left(\sum_{m=1}^{k} q_{km}\psi(\tau_{m2})\right) + \left(\sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k} q_{kn}\right)\psi(\tau_{m1})\right) - \left(\sum_{m=1}^{k}\left(\sum_{n=m}^{k} q_{kn}\right)\psi(\tau_{m1}+\tau_{m2})\right) - \sum_{m=1}^{k} q_{km}\log q_{km}. \tag{9}$$

Since Equation (9) holds for any $q_{k1},\ldots,q_{kk}$ for all $1 \leq k \leq K$, we now optimise q_k to maximise the lower bound. Taking derivatives with respect to each q_{ki} and introducing λ as the Lagrangian variable to ensure that q_k is a distribution, we get

$$0 = \psi(\tau_{i2}) + \sum_{m=1}^{i-1}\psi(\tau_{m1}) - \sum_{m=1}^{i}\psi(\tau_{m1}+\tau_{m2}) - 1 - \log(q_{ki}) - \lambda$$

Solving for q_{ki}, we find

$$q_{ki} \propto \exp\left(\psi(\tau_{i2}) + \sum_{m=1}^{i-1}\psi(\tau_{m1}) - \sum_{m=1}^{i}\psi(\tau_{m1}+\tau_{m2})\right) \tag{10}$$

where the proportionality ensures that q_k a valid distribution. If we plug this multinomial lower bound back into Equation (8), we get a lower bound on $\log p(\boldsymbol{X}|\boldsymbol{\theta})$. We then optimise the remaining parameters to maximise the resulting lower bound as described in the next section.

The auxiliary distribution q_k is largely a computational tool, but it does have the following intuition. Since $\pi_k = \prod_{i=1}^{k} v_i$; we can imagine the event $z_{nk} = 1$ is equivalent to the event that a series of variables $u_i \sim \mathsf{Bernoulli}(v_i)$ all flip to one. If any of the u_i's equal zero, then the feature is off. The multinomial distribution q_{kj} can be thought of as a distribution over the event that the j^{th} variable u_j is the first u_i to equal 0.

5.2 Parameter Updates

The updates for the variational parameters for $\boldsymbol{A}$ and $\boldsymbol{Z}$ are still in the exponential family. For the parameters of $\boldsymbol{A}$, the updates are identical to those of the finite model. For the parameters of $\boldsymbol{Z}$, the updates are again similar to the finite model, except we must use an approximation for $\mathbb{E}_{\boldsymbol{v}}[\log(1-\prod_{i=1}^{k} v_i)]$.

The updates for the parameters for $\boldsymbol{v}$, however, strongly depend on how we approximate the term $\mathbb{E}_{\boldsymbol{v}}[\log(1-\prod_{i=1}^{k} v_i)]$. If we use the multinomial lower bound of Section 5.1, the updates have a nice closed form[2]. As in the finite approach, we sequentially update each of the variational parameters in turn. Derivations of the following update equations are in Appendix D.2.

1. For $k = 1,\ldots,K$, we update the $\bar{\boldsymbol{\phi}}_k$ and $\boldsymbol{\Phi}_k$ in $\mathsf{Normal}(\boldsymbol{A}_{k\cdot}; \bar{\boldsymbol{\phi}}_k, \boldsymbol{\Phi}_k)$ as

$$\begin{aligned}\boldsymbol{\Phi}_k &= \left(\frac{1}{\sigma_A^2} + \frac{\sum_{n=1}^{N}\nu_{nk}}{\sigma_n^2}\right)^{-1} I \\ \bar{\boldsymbol{\phi}}_k &= \left[\frac{1}{\sigma_n^2}\sum_{n=1}^{N}\nu_{nk}\left(\boldsymbol{X}_{n\cdot} - \left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l\right)\right)\right]\left(\frac{1}{\sigma_A^2} + \frac{\sum_{n=1}^{N}\nu_{nk}}{\sigma_n^2}\right)^{-1}.\end{aligned}$$

[2]Appendix D.2 describes an alternative approach that numerically optimises the variational lower bound based on our Taylor series approximation; however, we found the direct optimisation was less computationally efficient.

2. For $k = 1, \ldots, K$, $n = 1, \ldots, N$, update ν_{nk} in $\mathsf{Bernoulli}(z_{nk}; \nu_{nk})$ as

$$\nu_{nk} = \frac{1}{1 + e^{-\vartheta}}$$

where

$$\begin{aligned}\vartheta &= \sum_{i=1}^{k} (\psi(\tau_{i1}) - \psi(\tau_{i1} + \tau_{i2})) - \mathbb{E}_{\boldsymbol{v}}[\log(1 - \prod_{i=1}^{k} v_i)] \\ &\quad - \frac{1}{2\sigma_n^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k \bar{\boldsymbol{\phi}}_k^T\right) + \frac{1}{\sigma_n^2} \bar{\boldsymbol{\phi}}_k \left(\boldsymbol{X}_{n\cdot}^T - \left(\sum_{l:l\neq k} \nu_{nl} \bar{\boldsymbol{\phi}}_l^T\right)\right).\end{aligned}$$

We leave the term $\mathbb{E}_{\boldsymbol{v}}[\log(1 - \prod_{i=1}^{k} v_i)]$ unevaluated because the choice of how to approximate it does not change the form of the update. In practice, we use the multinomial lower bound for this term.

3. For $k = 1, \ldots, K$, we must update the τ_{k1} and τ_{k2} in $\mathsf{Beta}(v_k; \tau_{k1}, \tau_{k2})$. If we use the multinomial lower bound for $\mathbb{E}_{\boldsymbol{v}}[\log(1 - \prod_{i=1}^{k} v_i)]$, then we can first compute q_{ki} according to Equation (10) (remembering to normalise q_{ki}). Then the updates for τ_{k1} and τ_{k2} have the following closed form

$$\begin{aligned}\tau_{k1} &= \alpha + \sum_{m=k}^{K} \sum_{n=1}^{N} \nu_{nm} + \sum_{m=k+1}^{K} \left(N - \sum_{n=1}^{N} \nu_{nm}\right) \left(\sum_{i=k+1}^{m} q_{mi}\right) \\ \tau_{k2} &= 1 + \sum_{m=k}^{K} \left(N - \sum_{n=1}^{N} \nu_{nm}\right) q_{mk}.\end{aligned}$$

6 Bound for the Infinite Approximation

Both of our variational inference approaches require us to choose a truncation level K for our variational distribution. Building on results from Thibaux and Jordan (2007) and Teh et al. (2007), we present a bound on how close the marginal distribution of the data $\boldsymbol{X}$ using a truncated stick-breaking prior will be to the marginal distribution using the true IBP stick-breaking prior. The bound can serve as a rough guide for choosing K, though the results do not tell us how good our variational approximations will be.

Our development parallels a bound for the Dirichlet Process by Ishwaran and James (2001) and presents the first such truncation bound for the IBP. Let us denote the marginal distribution of observation $\boldsymbol{X}$ by $m_\infty(\boldsymbol{X})$ when we integrate $\boldsymbol{W}$ with respect to the true IBP stick-breaking prior $p(\boldsymbol{W}|\boldsymbol{\theta})$. Let $m_K(\boldsymbol{X})$ be the marginal distribution when $\boldsymbol{W}$ are integrated out with respect to the truncated stick-breaking prior with truncation level K as described at the beginning of Section 5. For consistency, we continue to use the notation from the linear-Gaussian model, but the derivation that follows is independent of the likelihood model.

Intuitively, the error in the truncation will depend on the probability that, given N observations, we observe more than K features in the data (otherwise the truncation should have no effect). Using the beta process representation for the IBP (Thibaux and Jordan, 2007) and using an analysis similar to the one in (Ishwaran and James, 2001), we can show that the difference

between the marginal distributions of $\boldsymbol{X}$ is at most

$$\begin{aligned}
\frac{1}{4}\int |m_K(\boldsymbol{X}) - m_\infty(\boldsymbol{X})| d\boldsymbol{X} &\leq \Pr(\exists k > K, n \text{ with } z_{nk} = 1) \\
&= 1 - \Pr\left(\text{all } z_{ik} = 0,\ i \in \{1, \ldots, N\}, k > K\right) \\
&= 1 - \mathbb{E}\left[\left(\prod_{i=K+1}^{\infty} (1 - \pi_i)\right)^N\right] \\
&\leq 1 - \left(\mathbb{E}\left[\prod_{i=K+1}^{\infty} (1 - \pi_i)\right]\right)^N . \qquad (11)
\end{aligned}$$

We begin the derivation of the formal truncation bound by noting that beta process construction for the IBP (Thibaux and Jordan, 2007) implies that the sequence of $\pi_1, \pi_2, \ldots$ may be modeled as a Poisson process on the unit interval $[0, 1]$ with rate $\mu(x) = \alpha x^{-1} dx$. It follows that the sequence of $\pi_{K+1}, \pi_{K+2}, \ldots$ may be modeled as a Poisson process on the interval $[0, \pi_K]$ with the same rate. The Levy-Khintchine formula (Applebaum, 2004) states that the moment generating function of a Poisson process X with rate μ can be written as

$$\mathbb{E}[\exp(tf(X))] = \exp\left(\int \left(\exp(tf(y)) - 1\right) \mu(y) dy\right).$$

where we use $f(X)$ to denote $\sum_{x \in X} f(x)$.

Returning to Equation (11), if we rewrite the final expectation as

$$\mathbb{E}\left[\left(\prod_{i=K+1}^{\infty} (1 - \pi_i)\right)\right] = \mathbb{E}\left[\exp\left(\sum_{i=K+1}^{\infty} \log(1 - \pi_i)\right)\right],$$

then we can apply the Levy-Khintchine formula to get

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\sum_{i=K+1}^{\infty} \log(1 - \pi_i)\right)\right] &= \mathbb{E}_{\pi_K}\left[\exp\left(\int_0^{\pi_K} \left(\exp(\log(1 - x)) - 1\right) \mu(x) dx\right)\right] \\
&= \mathbb{E}_{\pi_K}[\exp(-\alpha \pi_K)].
\end{aligned}$$

Finally, we apply Jensen's inequality, using the fact that π_K is the product of independent $\mathsf{Beta}(\alpha, 1)$ variables:

$$\begin{aligned}
\mathbb{E}_{\pi_K}[\exp(-\alpha \pi_K)] &\geq \exp\left(\mathbb{E}_{\pi_k}[-\alpha \pi_K]\right) \\
&= \exp\left(-\alpha \left(\frac{\alpha}{1 + \alpha}\right)^K\right).
\end{aligned}$$

Substituting this expression back into Equation (11) gives us the bound

$$\frac{1}{4}\int |m_K(X) - m_\infty(X)| dX \leq 1 - \exp\left(-N\alpha \left(\frac{\alpha}{1 + \alpha}\right)^K\right). \qquad (12)$$

Similar to truncation bound for the Dirichlet Process, the expected error increases as N and α – the factors that increase the expected number of features – increase. However, for fixed N and α, the bound decreases exponentially quickly as the truncation level K is increased.

Figure 3 shows our truncation bound and the true L_1 distance based on 1000 Monte Carlo simulations of an IBP matrix with $N = 30$ observations and $\alpha = 5$. As expected, the bound

decreases exponentially fast with the truncation level K. The bound is loose, however; in practice, we find that a heuristic approximation to the bound using a Taylor series expansion provides tighter estimates of the loss. Appendix F describes both this heuristic bound and other (principled) bounds that can be derived via other applications of Jensen's inequality. All of these provide guidance for how to choose K such that the truncated stick-breaking construction is close to the true stick-breaking construction, but do not tell us how close the truncated variational approximation is to the true posterior. For this, we use the bound merely as a heuristic.

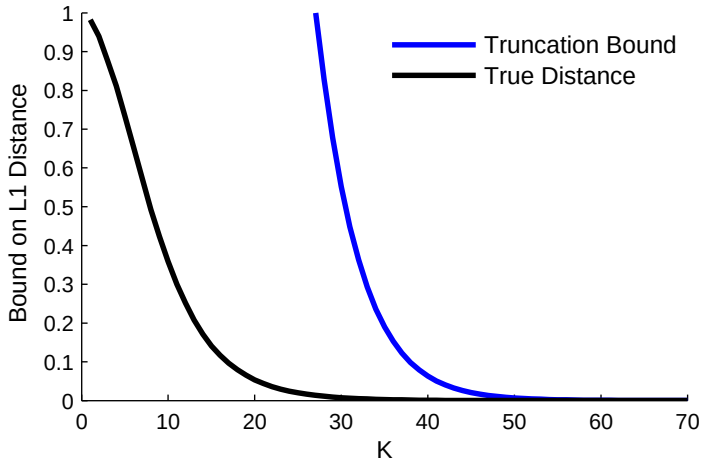


Figure 3: Truncation bound and true L_1 distance.

7 Experiments

We compared our variational approaches with both Gibbs sampling (Griffiths and Ghahramani, 2005) and particle filtering (Wood and Griffiths, 2007). As variational algorithms are only guaranteed to converge to a *local* optimum, we applied standard optimisation tricks to avoid small minima. Each run was given a number of random restarts and the hyperparameters for the noise and feature variance were annealed to smooth the posterior. We also experimented with several other techniques such as gradually introducing data and merging correlated features. The latter techniques proved less useful as the size and dimensionality of the datasets increased; they were not included in the final experiments.

We compared against two variants of the Gibbs sampling method introduced in Griffiths and Ghahramani (2005). The first is the collapsed Gibbs sampler described by (Griffiths and Ghahramani, 2005), the second is a partially-uncollapsed alternative in which instantiated features are explicitly represented and new features are integrated out. We compare against these two methods because the time complexity of the collapsed sampler scales cubicly, so as we get more data, the cost of this sampler becomes prohibitive. The uncollapsed sampler scales much better in terms of computation per iteration as we get more data, but in general does not explore the posterior as well as the collapsed sampler.

In contrast to the variational methods, the number of features present in the IBP matrix will adaptively grow or shrink in the samplers. To provide a fair comparison with the variational approaches, we also tested finite variants of the collapsed and uncollapsed Gibbs samplers in which the number of potential features was fixed. Details for these samplers are given in Appendix A. We also tested against the particle filter of Wood and Griffiths (2007). All sampling

methods were annealed and given an equal number of restarts as the variational methods.

Both the variational and Gibbs sampling algorithms were heavily optimised for efficient matrix computation so we could evaluate the algorithms both on their running times and the quality of the inference. For the particle filter, we used the implementation provided by Wood and Griffiths (2007). To measure the quality of these methods, we held out one third of the observations on the last half of the dataset. Once the inference was complete, we computed the predictive likelihood of the held out data (averaged over restarts).

7.1 Synthetic Data

The synthetic datasets consisted of $\boldsymbol{Z}$ and $\boldsymbol{A}$ matrices randomly generated from the truncated stick-breaking prior. Figure 4 shows the evolution of the test-likelihood over a thirty minute interval for a dataset with 500 observations of 500 dimensions and with 20 latent features. The error bars indicate the variation over the 5 random starts[3]. The finite uncollapsed Gibbs sampler (dotted green) rises quickly but consistently gets caught in lower optima and has higher variance than the variational approach. Examining the individual runs, we found the higher variance was not due to the Gibbs sampler mixing but due to each run getting stuck in widely varying local optima. The variational methods were slightly slower per iteration but soon found regions of higher predictive likelihoods. The remaining samplers were much slower per iteration, often failing to mix within the allotted interval.

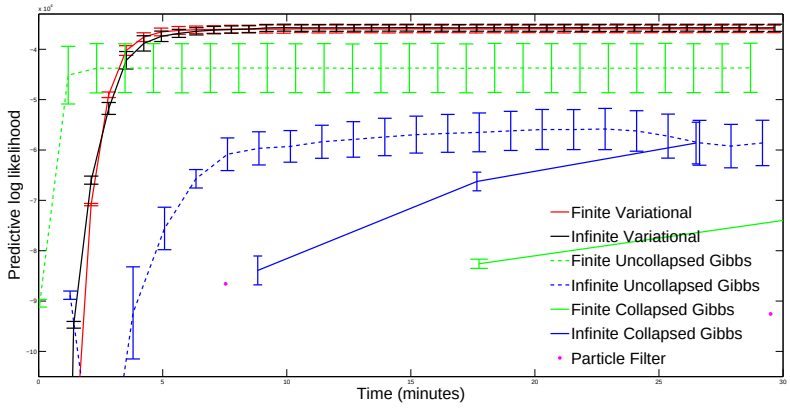


Figure 4: Evolution of test log-likelihoods over a thirty-minute interval for $N = 500$, $D = 500$, and $K = 20$. The finite uncollapsed Gibbs sampler has the fastest rise but gets caught in worse lower optima than the variational approach.

Figure 5 shows a similar plot for a smaller dataset with $N = 100$. Here, the variational approaches do not do as well as the collapsed Gibbs samplers at finding regions of high probability. However, it does as well or better than the uncollapsed samplers. We believe this is because in a smaller dataset, the collapsed Gibbs samplers mix fairly quickly and can therefore find regions of high probability mass. The uncollapsed samplers still get stuck in widely varying local optima as witnessed by the large variance again. The variational approach still converges quickly to optima that on average are better than the uncollapsed samplers, but are not as good the optima the collapsed samplers eventually find.

[3]The particle filter must be run to completion before making prediction, so we cannot test its predictive performance over time. We instead plot the test likelihood only at the end of the inference for particle filters with 10 and 50 particles (the two magenta points in the bottom half of the graph).

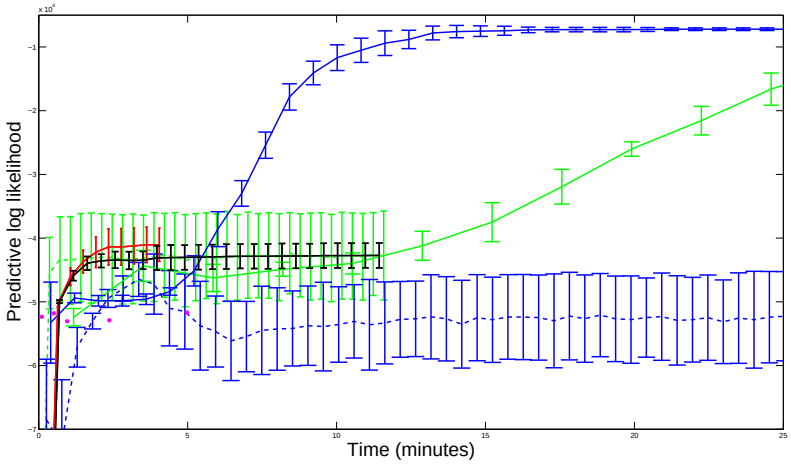


Figure 5: Evolution of test log-likelihoods over a thirty-minute interval for $N = 100$, $D = 500$, and $K = 25$. For smaller N, the Gibbs sampler does better at finding an optima of high probability mass.

Figures 6 and 7 show results from a systematic series of tests in which we tested all combinations of observation count $N = \{5, 10, 50, 100, 500, 1000\}$, dimensionality $D = \{5, 10, 50, 100, 500, 1000\}$, and truncation level $K = \{5, 10, 15, 20, 25\}$. Each of the samplers was run for 1000 iterations on three chains and the particle filter was run with 500 particles. For the variational methods, we used a stopping criterion that halted the optimisation when the variational lower bound between the current and previous iterations changed by a multiplicative factor of less than 10^{-6} and the annealing process had completed.

Figure 6 shows how the computation time scales with the truncation level. The variational approaches and the uncollapsed Gibbs sampler are consistently an order of magnitude faster than other algorithms. Figure 7 shows the interplay between dimensionality, computation time, and test log-likelihood for datasets of size $N = 5$ and $N = 1000$ respectively. For $N = 1000$, the collapsed Gibbs samplers and particle filter did not finish, so they do not appear on the plot. We chose $K = 20$ as a representative truncation level. Each line represents increasing dimensionality for a particular method (the large dot indicates $D = 5$, the subsequent dots correspond to $D = 10$, 50, etc.). The nearly vertical lines of the variational methods show that they are quite robust to increasing dimension. Moreover, as dimensionality and dataset size increase, the variational methods become increasingly faster than the samplers. By comparing the lines across the likelihood dimension, we see that for the very small dataset, the variational method often has a lower test log-likelihood than the samplers. In this regime, the samplers mix quickly and find regions of high likelihood.

The short summary of these results is that for smaller data sets (both smaller D and N), the collapsed samplers are preferable. They mix well and find configurations that have high predictive likelihood. However, as the size of the data grows (both D and N), these samplers do not scale well and their cost becomes prohibitive. In these regimes, our variational approximations are preferable over the uncollapsed samplers. Both converge to local optima quickly, but the variational approximations finds better local optima.

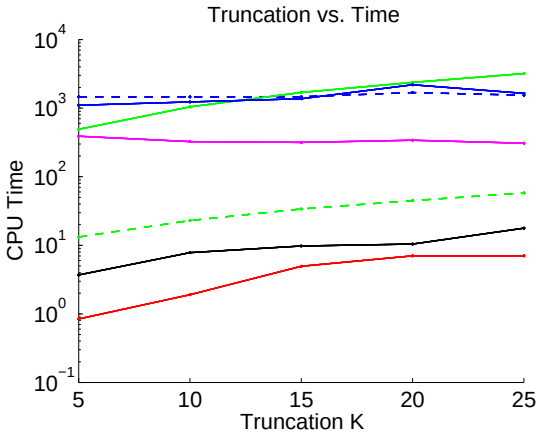


Figure 6: Time versus truncation (K). The variational approaches are generally orders of magnitude faster than the samplers (note log scale on the time axis).

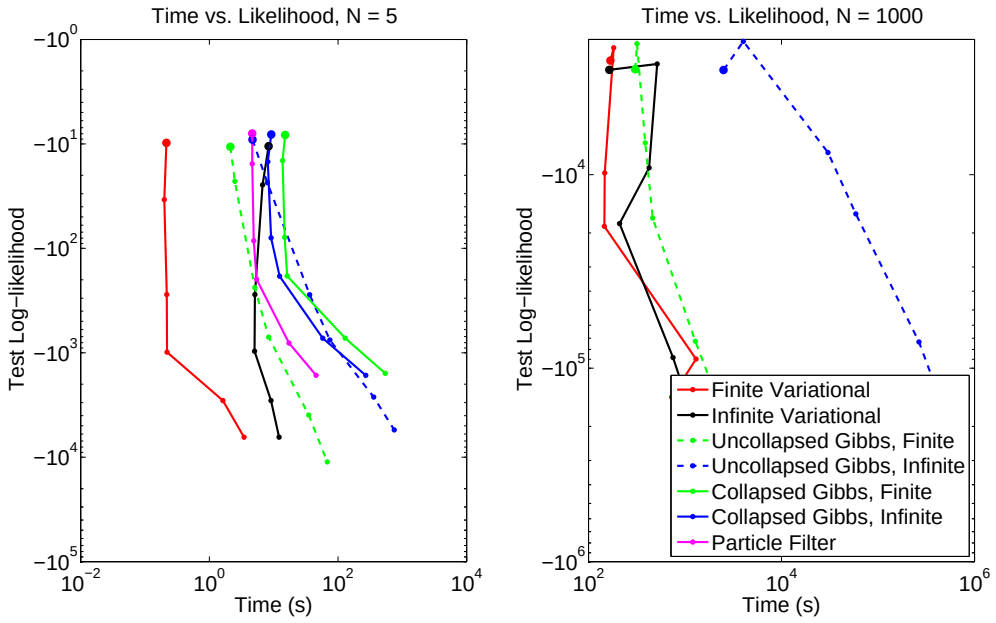


Figure 7: Time versus log-likelihood plot for $K = 20$. The larger dots correspond to $D = 5$ the smaller dots to $D = 10, 50, 100, 500, 1000$.

7.2 Real Data

To test the conclusions we arrived at based on synthetic data, we applied our variational method to two real-world datasets to see how it would fare with complex, noisy data not drawn from the IBP prior[4]. The first dataset, the Yale Faces (Georghiades et al., 2001) dataset, consisted of $N = 721$ 32x32 pixel frontal-face images ($D = 1024$) of 14 people with varying expressions and lighting conditions. We set σ_a and σ_n based on the variance of the data. The second dataset, a speech dataset, consisted of $N = 245$ observations sampled from a 10-microphone audio recording ($D = 10$) of 5 different speakers. We applied the iICA version of our inference algorithm, where a mixing matrix S modulated the effect of each speaker on the audio signals. The feature and noise variances were taken from an initial run of the Gibbs sampler where σ_n and σ_a were also sampled.

These two datasets were representative of the large N, D regime (Yale faces dataset) and the small N, D regime (speech dataset), so based on our synthetic results, we expect the variational approach to give us an advantage on the faces data, but not on the speech data.

Tables 1 and 2 show the results for each of the datasets. All Gibbs samplers were uncollapsed and run for only 200 iterations.[5] In the higher dimensional Yale dataset, the variational methods outperformed the uncollapsed Gibbs sampler. When started from a random position, the uncollapsed Gibbs samplers quickly became stuck in local optima. The variational method was able to find better local optima because it was initially very uncertain about which features were present in which data points; expressing this uncertainty explicitly through the variational parameters (instead of through a sequence of samples with hard assignments) allowed it the flexibility to improve upon its bad initial starting point.

Table 1: Running times in seconds and test log-likelihoods for the Yale Faces dataset.

Algorithm	**K**	**Time**	**Test Log-Likelihood** ($\times 10^6$)
Finite Gibbs	5	464.19	-2.250
	10	940.47	-2.246
	25	2973.7	-2.247
Finite Variational	5	163.24	-1.066
	10	767.1	-0.908
	25	10072	-0.746
Infinite Variational	5	176.62	-1.051
	10	632.53	-0.914
	25	19061	-0.750

The story for the speech dataset, however, is quite different. Here, the variational methods were not only slower than the samplers, but they also achieved lower test-likelihoods. The evaluation on the synthetic datasets points to a potential reason for the difference: the speech

[4]Note that our objective was compare inference techniques for the IBP on real data, not to demonstrate state of the art low-rank approximations.

[5]On the Yale dataset, we did not test the collapsed samplers because the finite collapsed Gibbs sampler required one hour per iteration with $K = 5$ and the infinite collapsed Gibbs sampler required fifty hours per sample. In the iICA model, the collapsed Gibbs sampler could not be run because the features $\boldsymbol{A}$ cannot be marginalised.

dataset is much simpler than the Yale dataset, consisting of 10 dimensions (vs. 1032 in the Yale dataset). In this regime, the Gibbs samplers perform well and the approximations made by the variational method become apparent. As the dimensionality grows, the samplers have more trouble mixing, but the variational methods are still able to find regions of high probability mass.

Table 2: Running times in seconds and test log-likelihoods for the speech dataset.

Algorithm	**K**	**Time**	**Test Log-Likelihood**
Finite Gibbs	2	56	-0.7444
	5	120	-0.4220
	9	201	-0.4205
Infinite Gibbs	na	186	-0.4257
Finite Variational	2	2477	-0.8455
	5	8129	-0.5082
	9	8539	-0.4551
Infinite Variational	2	2702	-0.8810
	5	6065	-0.5000
	9	8491	-0.5486

8 Summary

The combinatorial nature of the Indian Buffet Process poses specific challenges for sampling-based inference procedures. In this report, we derived a mean field variational inference procedure for the IBP. Whereas sampling methods work in the discrete space of binary matrices, the variational method allows for soft assignments of features because it approaches the inference problem as a continuous optimisation. We showed experimentally that for high dimensional problems, the soft assignments allow the variational methods to find much better optima than sampling-based approaches. On the other hand, we have shown that sampling based approaches are preferable over our variational approximation.

Acknowledgments

FD was supported by a Marshall scholarship. KTM was supported by contract DE-AC52-07NA27344 from the U.S. Department of Energy through Lawrence Livermore National Laboratory. JVG was supported by a Microsoft Research scholarship.

References

D. Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge University Press, 2004.

M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, UCL, 2003.

D. Blei and M. Jordan. Variational methods for the Dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

F. Doshi-Velez, K. T. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the indian buffet process. In *Proc. of the Conference on Artificial Intelligence and Statistics*, 2009.

A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6), 2001.

M. N. Gibbs and D. J. C. MacKay. Variational gaussian process classifiers. *IEEE-NN*, 11(6):1458, November 2000.

T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *TR 2005-001, Gatsby Computational Neuroscience Unit*, 2005.

Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for stick breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

D. Knowles and Z. Ghahramani. Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. *Lecture Notes in Computer Science*, 4666:381, 2007.

K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational Dirichlet process mixture models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 20, 2007a.

Kenichi Kurihara, Max Welling, and Nikos Vlassis. Accelerated variational dirichlet process mixtures. In *Advances in Neural Information Processing Systems 19*. 2007b.

Edward Meeds, Zoubin Ghahramani, Radford M. Neal, and Sam T. Roweis. Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems 19*, 2007.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*. MIT press, 2006.

Y. W. Teh, D. Gorur, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *Proceedings of the 11th Conference on Artificial Intelligence and Statistic*, 2007.

R. Thibaux and M. Jordan. Hierarchical beta processes and the indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2007.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Ole Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12, 2000.

Frank Wood and Thomas L. Griffiths. Particle filtering for nonparametric Bayesian matrix factorization. In *Advances in Neural Information Processing Systems 19*. 2007.

A Gibbs Sampling in the IBP

In this appendix, we briefly review the two Gibbs sampling methods that we compared against. Both of these samplers are for the linear-Gaussian likelihood model described in Section 2.1.

A.1 Collapsed Gibbs Sampler

The collapsed Gibbs sampler maintains samples over $\boldsymbol{Z}$ and integrates out $\boldsymbol{A}$, so we only need to specify the posterior distribution for $\boldsymbol{Z}$. If $\boldsymbol{A}$ is needed, then given $\boldsymbol{Z}$, it is easy to compute the posterior for $\boldsymbol{A}$ as described by Griffiths and Ghahramani (2005).

Sampling z_{nk} for Existing Features For existing (nonzero) features, the collapsed Gibbs sampler for the IBP resamples each element of the feature assignment matrix $\boldsymbol{Z}$ via the equation

$$p(z_{nk} = 1|\boldsymbol{Z}_{-nk}, \boldsymbol{X}) \propto \frac{m_{-n,k}}{N-1} p(\boldsymbol{X}|\boldsymbol{Z}) \tag{13}$$

where $m_{-n,k}$ is the number of observations not including z_{nk} containing feature k. The likelihood term $p(\boldsymbol{X}|\boldsymbol{Z})$ is given by

$$p(\boldsymbol{X}|\boldsymbol{Z}) = \frac{\exp(-\frac{1}{2\sigma_x^2}(\boldsymbol{X}^T(I - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \frac{\sigma_x^2}{\sigma_a^2}I)^{-1}\boldsymbol{Z}^T)\boldsymbol{X}))}{(2\pi)^{\frac{ND}{2}}\sigma_{\boldsymbol{X}}^{(N-K)D}\sigma_a^{KD}|\boldsymbol{Z}^T\boldsymbol{Z} + \frac{\sigma_x^2}{\sigma_a^2}I|^{\frac{D}{2}}}. \tag{14}$$

Sampling New Features There are an infinite number of remaining columns which contain all zeroes. For any particular z_{nk}, $k > K$, the probability that $z_{nk} = 1$ is zero. However, we can sample the number of columns that become nonzero, k_{new}, as a batch. (See (Griffiths and Ghahramani, 2005) for details.) The number of new features is sampled according to

$$p(k_{\text{new}}) \propto \mathsf{Poisson}\left(k_{\text{new}}; \frac{\alpha}{N}\right) p(\boldsymbol{X}|\boldsymbol{Z}_{\text{new}})$$

where $\boldsymbol{Z}_{\text{new}}$ is the feature-assignment matrix with k_{new} additional columns set to one for object n and zero otherwise. The term $p(\boldsymbol{X}|\boldsymbol{Z}_{\text{new}})$ can be computed from Equation (14). We compute these probabilities for $k_{\text{new}} = 0, \ldots, K_{\max}$ for some $K_{\max}$, normalise and sample from the resulting multinomial.

Modifications for the Finite Model If we are sampling from a finite model with K features and a finite beta-Bernoulli prior on $\boldsymbol{Z}$ as described in Section 4, then Equation (13) becomes

$$p(z_{nk} = 1|\boldsymbol{Z}_{-nk}, \boldsymbol{X}) \propto \frac{m_{-n,k} + \alpha/K}{N - 1 + \alpha} p(\boldsymbol{X}|\boldsymbol{Z}).$$

We never need to sample the number of new features since K is fixed.

A.2 Uncollapsed Gibbs Sampler

The disadvantage of the collapsed Gibbs sampler is that Equation (14) can be expensive to compute. The uncollapsed Gibbs sampler explicitly samples the feature matrix $\boldsymbol{A}$ and therefore does not need to evaluate Equation (14). Our samples are therfore over $\boldsymbol{Z}$ and $\boldsymbol{A}$.

Sampling z_{nk} for Existing Features The Gibbs sampling Equation for z_{nk} for existing features is now

$$p(z_{nk} = 1|\boldsymbol{Z}_{-nk}, \boldsymbol{A}, \boldsymbol{X}) \propto \frac{m_{-n,k}}{N-1} p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{A}) \tag{15}$$

where the likelihood term $p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{A})$ is given by

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{A}) = \frac{1}{(2\pi\sigma_n^2)^{ND/2}} \exp\left(-\frac{1}{2\sigma_n^2}\mathrm{tr}((\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{A})^\top(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{A}))\right).$$

Sampling A for Existing Features The posterior for resampling $\boldsymbol{A}$ given $\boldsymbol{Z}$ and $\boldsymbol{X}$ is

$$p(\boldsymbol{A}|\boldsymbol{X}, \boldsymbol{Z}) \sim \mathcal{N}\left(\left(\boldsymbol{Z}^\top\boldsymbol{Z} + \frac{\sigma_n^2}{\sigma_A^2}I\right)^{-1}\boldsymbol{Z}^\top\boldsymbol{X}, \sigma_n^2\left(\boldsymbol{Z}^\top\boldsymbol{Z} + \frac{\sigma_n^2}{\sigma_A^2}I\right)^{-1}\right).$$

Sampling New Features As in the collapsed sampler, we sample k_{new}, the number of new nonzero columns instead of sampling each of the infinite number of all-zero columns independently. As before, the probability of k_{new} is

$$p(k_{\text{new}}) \propto \mathsf{Poisson}\left(k_{\text{new}}; \frac{\alpha}{N}\right) p(\boldsymbol{X}|\boldsymbol{Z}_{\text{new}}, \boldsymbol{A}) \tag{16}$$

where $\boldsymbol{A}$ represents the initialised features. The likelihood $p(\boldsymbol{X}|\boldsymbol{Z}_{\text{new}}, \boldsymbol{A})$ is given by the integral $\int_{\boldsymbol{A}_{\text{new}}} p(\boldsymbol{X}|\boldsymbol{Z}_{\text{new}}, \boldsymbol{A}, \boldsymbol{A}_{\text{new}}) p(\boldsymbol{A}_{\text{new}})$.

If we want the sampler to be fully uncollapsed, one option for drawing k_{new} from the distribution in Equation 16 is to perform a Monte Carlo integration to evaluate $p(\boldsymbol{X}|\boldsymbol{Z}_{\text{new}}, \boldsymbol{A})$. For this, we first draw many pairs $(k_{\text{new}}, \boldsymbol{A}_{\text{new}})$ from their respective priors, assign a weight to each pair based on the data likelihood $p(\boldsymbol{X}|\boldsymbol{Z}_{\text{new}}, \boldsymbol{A}, \boldsymbol{A}_{\text{new}})$, and then we sample a pair $(k_{\text{new}}, \boldsymbol{A}_{\text{new}})$ based on the weights and take the k_{new} element of the pair as our k_{new}. The advantage of using importance sampling in this way is that the approach remains fully uncollapsed — no integrals need be evaluated. However, since the features of $\boldsymbol{A}$ are drawn from the prior, the fully uncollapsed approach is slow to mix.

Another option, if we allow ourselves to have a partially collapsed sampler, is to actually compute the integral in the likelihood in Equation (16) — that is, marginalise over the new features. This option results in a faster mixing sampler, and it was the option used in our tests. The equations below describe how to sample k_{new} when $\boldsymbol{A}_{\text{new}}$ is marginalised out. For notation, let $\boldsymbol{Z}_{\text{old}}$ be the current matrix $\boldsymbol{Z}$ and $\boldsymbol{A}_{\text{old}}$ be the current matrix $\boldsymbol{A}$. Similarly, let $\boldsymbol{Z}_{\text{new}}$ and $\boldsymbol{A}_{\text{new}}$ be the parts of $\boldsymbol{Z}$ and $\boldsymbol{A}$ that correspond to the k_{new} new features. Finally, let $\boldsymbol{Z}^*$ and $\boldsymbol{A}^*$ be the concatenation of the new and old matrices.

Using Bayes rule, we can write

$$p(k_{\text{new}}|X, \boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}) \quad \propto \quad p(X|\boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}, k_{\text{new}}) p(k_{\text{new}}) \tag{17}$$

where $p(k_{\text{new}})$ is $\mathsf{Poisson}(\alpha/N)$ and $p(\boldsymbol{X}|\boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}, k_{\text{new}})$ is the likelihood in which $\boldsymbol{A}_{\text{new}}$ has been marginalised out.

We must now specify $p(\boldsymbol{X}|\boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}, k_{\text{new}})$:

$$\begin{aligned}
&p(\boldsymbol{X}|\boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}, k_{\text{new}}) \\
&\quad = \quad \int p(\boldsymbol{X}|\boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}, \boldsymbol{A}_{\text{new}}, k_{\text{new}}) p(\boldsymbol{A}_{\text{new}}) d\boldsymbol{A}_{\text{new}} \\
&\quad = \quad \frac{1}{(2\pi\sigma_n^2)^{ND/2}} \frac{1}{(2\pi\sigma_A^2)^{k_{\text{new}}D/2}} \int \exp\left(-\frac{1}{2}\text{tr}\left(\frac{1}{\sigma_n^2}(\boldsymbol{X} - \boldsymbol{Z}^*\boldsymbol{A}^*)^\top(\boldsymbol{X} - \boldsymbol{Z}^*\boldsymbol{A}^*) + \frac{1}{\sigma_A^2}\boldsymbol{A}_{\text{new}}^\top\boldsymbol{A}_{\text{new}}\right)\right) d\boldsymbol{A}_{\text{new}}
\end{aligned}$$

where

$$(\boldsymbol{X} - \boldsymbol{Z}^*\boldsymbol{A}^*)^\top(\boldsymbol{X} - \boldsymbol{Z}^*\boldsymbol{A}^*) \quad = \quad \left(\boldsymbol{X} - \begin{bmatrix} \boldsymbol{Z}_{\text{old}} & \boldsymbol{Z}_{\text{new}} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{\text{old}} \\ \boldsymbol{A}_{\text{new}} \end{bmatrix}\right)^\top \left(\boldsymbol{X} - \begin{bmatrix} \boldsymbol{Z}_{\text{old}} & \boldsymbol{Z}_{\text{new}} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{\text{old}} \\ \boldsymbol{A}_{\text{new}} \end{bmatrix}\right)$$

Completing squares to integrate our $\boldsymbol{A}_{\text{new}}$, and dropping terms that do not depend on k_{new}, we get

$$\begin{aligned}
&p(\boldsymbol{X}|\boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}, k_{\text{new}}) \\
&\quad \propto \quad \frac{(\sigma_n/\sigma_A)^{k_{\text{new}}D}}{|1_{k_{\text{new}}\times k_{\text{new}}} + \frac{\sigma_n^2}{\sigma_A^2} I|^{D/2}} \\
&\qquad \times \exp\left\{\frac{1}{2\sigma_n^2}\text{tr}\left((\boldsymbol{X} - \boldsymbol{Z}_{\text{old}}\boldsymbol{A}_{\text{old}})^\top \boldsymbol{Z}_{\text{new}} \left(1_{k_{\text{new}}\times k_{\text{new}}} + \frac{\sigma_n^2}{\sigma_A^2} I\right)^{-1} \boldsymbol{Z}_{\text{new}}^\top (\boldsymbol{X} - \boldsymbol{Z}_{\text{old}}\boldsymbol{A}_{\text{old}})\right)\right\}.
\end{aligned}$$

We can therefore sample k_{new} according to Equation (17) where $p(\boldsymbol{X}|\boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}, k_{\text{new}})$ is specified above. Once we have sampled k_{new}, we need to sample the newly activated features $\boldsymbol{A}_{\text{new}}$. Based on the same calculations that give us $p(\boldsymbol{X}|\boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}, k_{\text{new}})$, we can sample $\boldsymbol{A}_{\text{new}}$ from the distribution

$$\begin{aligned} & p(\boldsymbol{A}_{\text{new}}|\boldsymbol{X}, \boldsymbol{Z}_{\text{new}}, \boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}) \\ & \quad \propto \quad p(\boldsymbol{X}|\boldsymbol{Z}_{\text{new}}, \boldsymbol{Z}_{\text{old}}, \boldsymbol{A}_{\text{old}}, \boldsymbol{A}_{\text{new}}) p(\boldsymbol{A}_{\text{new}}) \\ & \quad \sim \quad \mathcal{N}\left(\left(1_{k_{\text{new}} \times k_{\text{new}}} + \frac{\sigma_n^2}{\sigma_A^2} I \right)^{-1} \boldsymbol{Z}_{\text{new}}^\top \left(\boldsymbol{X} - \boldsymbol{Z}_{\text{old}} \boldsymbol{A}_{\text{old}} \right), \sigma_n^2 \left(1_{k_{\text{new}} \times k_{\text{new}}} + \frac{\sigma_n^2}{\sigma_A^2} I \right)^{-1} \right). \end{aligned}$$

Modifications for the Finite Model If we are sampling from a finite model with K features and a beta-Bernoulli prior on $\boldsymbol{Z}$, then Equation (15) becomes

$$p(z_{nk} = 1|\boldsymbol{Z}_{-nk}, \boldsymbol{X}) \propto \frac{m_{-n,k} + \alpha/K}{N - 1 + \alpha} p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{A}).$$

In addition, we never need to sample the number of new features since K is fixed.

B Variational Inference in Exponential Families

Recall that our goal is to find an approximating distribution $q \in Q$ with minimum KL divergence $D(q||p)$ to the true distribution p. Equation (4) rephrased this optimisation problem in terms of certain expectations and entropies:

$$\arg\min_{\tau,\phi,\nu} D(q||p) = \arg\max_{\tau,\phi,\nu} \mathbb{E}_q[\log(p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})] + H[q]. \tag{18}$$

In general, this optimisation can be quite difficult. However, when the conditional distribution and variational distribution are both in the exponential family, each step in the coordinate ascent has a closed form solution (Beal, 2003; Wainwright and Jordan, 2008). If we are updating the variational parameters ξ_i that correspond to W_i, then the optimal ξ_i are the solution to

$$\log q_{\xi_i}(W_i) = \mathbb{E}_{\boldsymbol{W}_{-i}}[\log p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})] + c \tag{19}$$

where the expectation is taken over all $\boldsymbol{W}$ except W_i according to the variational distribution. In the exponential family, this immediately gives us the updated values of the parameters ξ_i.

See (Beal, 2003; Wainwright and Jordan, 2008) for more details.

C Derivations for the Finite Variational Approach

This appendix derives the variational lower bound and the variational updates described in Section 4.

C.1 Variational Lower Bound

We derive expressions for each expectation in Equation (5):

1. For the feature probabilities, which are beta-distributed,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\pi}}\left[\log p(\pi_k|\alpha)\right] &= \mathbb{E}_{\boldsymbol{\pi}}\left[\log\left(\frac{\alpha}{K}\pi_k^{\alpha/K-1}\right)\right], \\ &= \log\frac{\alpha}{K} + \left(\frac{\alpha}{K} - 1\right)\mathbb{E}_{\boldsymbol{\pi}}\log(\pi_k), \\ &= \log\frac{\alpha}{K} + \left(\frac{\alpha}{K} - 1\right)\left(\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2})\right), \end{aligned}$$

where $\psi(\cdot)$ is the digamma function.

2. For the feature assignments, which are Bernoulli-distributed given the feature probabilities,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}}\left[\log p(z_{nk}|\pi_k)\right] &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}}\left[\log\left(\pi_k^{z_{nk}}(1-\pi_k)^{1-z_{nk}}\right)\right], \\
&= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}}\left[z_{nk}\log\pi_k + (1-z_{nk})\log(1-\pi_k)\right], \\
&= \mathbb{E}_{\boldsymbol{Z}}[z_{nk}]\mathbb{E}_{\boldsymbol{\pi}}[\log\pi_k] + (1-\mathbb{E}_{\boldsymbol{Z}}[z_{nk}])\mathbb{E}_{\boldsymbol{\pi}}[\log(1-\pi_k)], \\
&= \nu_{nk}\psi(\tau_{k1}) + (1-\nu_{nk})\psi(\tau_{k2}) - \psi(\tau_{k1}+\tau_{k2}).
\end{aligned}$$

3. For the features, which are Gaussian-distributed,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{A}}\left[\log p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I)\right] &= \mathbb{E}_{\boldsymbol{A}}\left[\log\left(\frac{1}{(2\pi\sigma_A^2)^{D/2}}\exp\left(-\frac{1}{2\sigma_A^2}\boldsymbol{A}_{k\cdot}^T\boldsymbol{A}_{k\cdot}\right)\right)\right], \\
&= \mathbb{E}_{\boldsymbol{A}}\left[\frac{-D}{2}\log(2\pi\sigma_A^2) - \frac{1}{2\sigma_A^2}\boldsymbol{A}_{k\cdot}^T\boldsymbol{A}_{k\cdot}\right], \\
&= \frac{-D}{2}\log(2\pi\sigma_A^2) - \frac{1}{2\sigma_A^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right).
\end{aligned}$$

4. For the likelihood, which is also Gaussian,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2 I)\right] \\
&= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[\log\left(\frac{1}{(2\pi\sigma_n^2)^{D/2}}\exp\left(-\frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\right)\left(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\right)^T\right)\right)\right], \\
&= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[-\frac{D}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\right)\left(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\right)^T\right], \\
&= -\frac{D}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T - 2\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{Z}_{n\cdot}]\mathbb{E}_{\boldsymbol{A}}[\boldsymbol{A}]\boldsymbol{X}_{n\cdot}^T + \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}[\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Z}_{n\cdot}^T]\right), \\
&= -\frac{D}{2}\log(2\pi\sigma_n^2) \\
&\quad -\frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T - 2\sum_{k=1}^K\nu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T + 2\sum_{k<k'}\nu_{nk}\nu_{nk'}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_{k'}^T + \sum_{k=1}^K\nu_{nk}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right),
\end{aligned}$$

where the final expectation is derived by

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}[\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Z}_{n\cdot}^T] &= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[\left(\sum_{k=1}^K z_{nk}\boldsymbol{A}_{k\cdot}\right)\left(\sum_{k=1}^K z_{nk}\boldsymbol{A}_{k\cdot}\right)^T\right], \\
&= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[\sum_{d=1}^D\left(\sum_{k=1}^K z_{nk}A_{kd}^2 + \sum_{k,k':k'\neq k} z_{nk}z_{nk'}\boldsymbol{A}_{kd}\boldsymbol{A}_{k'd}\right)\right], \\
&= \sum_{k=1}^K v_{nk}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right) + 2\sum_{k<k'}v_{nk}v_{nk'}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_{k'}^T.
\end{aligned}$$

5. Finally, for the entropy,

$$\begin{aligned}
H[q] &= -\mathbb{E}_q\log\left[\prod_{k=1}^K q_{\boldsymbol{\tau}_k}(\pi_k)\prod_{k=1}^K q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})\prod_{k=1}^K\prod_{n=1}^N q_{\nu_{nk}}(z_{nk})\right], \\
&= \sum_{k=1}^K\mathbb{E}_{\boldsymbol{\pi}}(-\log q_{\boldsymbol{\tau}_k}(\pi_k)) + \sum_{k=1}^K\mathbb{E}_{\boldsymbol{A}}(-\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})) + \sum_{k=1}^K\sum_{n=1}^N\mathbb{E}_{\boldsymbol{Z}}(-\log q_{\nu_{nk}}(z_{nk})),
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi}}(-\log q_{\boldsymbol{\tau}_k}(\pi_k)) &= \log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) \\
&\quad -(\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2}), \\
\mathbb{E}_{\boldsymbol{A}}(-\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})) &= \frac{1}{2}\log\left((2\pi e)^D|\Phi_k|\right), \\
\mathbb{E}_{\boldsymbol{Z}}(-\log q_{\nu_{nk}}(z_{nk})) &= -\nu_{nk}\log\nu_{nk} - (1-\nu_{nk})\log(1-\nu_{nk}).
\end{aligned}$$

Putting all the terms together gives us the variational lower bound in Equation (6).

C.2 Parameter Updates

To optimise the variational parameters, we can directly optimise Equation (6). However, since both our p_K and our variational approximation are in the exponential family, we can instead use Equation (19) from Appendix B to directly give us the update equations for each parameter given all the rest. We take the latter approach in this section to compute the update equations for the variational parameters in the finite model. Throughout this section, we let c be a constant independent of the variable of interest that may change from line to line.

1. For the feature distribution at the optimal $\bar{\boldsymbol{\phi}}_k$ and Φ_k

$$\begin{aligned}
&\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) \\
&= \mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z}}\left[\log p_K(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})\right] + c, \\
&= \mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z}}\left[\log p_K(\boldsymbol{A}_{k\cdot}|\sigma_A^2) + \sum_{n=1}^{N}\log p_K(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2)\right] + c, \\
&= -\frac{1}{2\sigma_A^2}\left(\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T\right) - \frac{1}{2\sigma_n^2}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z}}\left[\left(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\right)\left(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\right)^T\right] + c, \\
&= -\frac{1}{2}\left[\boldsymbol{A}_{k\cdot}\left(\frac{1}{\sigma_A^2}+\frac{\sum_{n=1}^{N}\nu_{nk}}{\sigma_n^2}\right)\boldsymbol{A}_{k\cdot}^T - 2\boldsymbol{A}_{k\cdot}\left(\frac{1}{\sigma_n^2}\sum_{n=1}^{N}\nu_{nk}\left(\boldsymbol{X}_{n\cdot}-\left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l\right)\right)\right)^T\right] + c.
\end{aligned}$$

   Completing the squares and using Equation (19) gives us that for the optimal $\bar{\phi}_k$ and $\boldsymbol{\Phi}_k$, we must have

$$\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) = -\frac{1}{2}\left(\boldsymbol{A}_{k\cdot}\boldsymbol{\Phi}_k^{-1}\boldsymbol{A}_{k\cdot}^T - 2\boldsymbol{A}_{k\cdot}\boldsymbol{\Phi}_k^{-1}\bar{\boldsymbol{\phi}}_k^T\right) + c,$$

   which gives us the updates

$$\begin{aligned}
\bar{\boldsymbol{\phi}}_k &= \left[\frac{1}{\sigma_n^2}\sum_{n=1}^{N}\nu_{nk}\left(\boldsymbol{X}_{n\cdot}-\left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l\right)\right)\right]\left(\frac{1}{\sigma_A^2}+\frac{\sum_{n=1}^{N}\nu_{nk}}{\sigma_n^2}\right)^{-1}, \\
\boldsymbol{\Phi}_k &= \left(\frac{1}{\sigma_A^2}+\frac{\sum_{n=1}^{N}\nu_{nk}}{\sigma_n^2}\right)^{-1} I.
\end{aligned}$$

2. For the feature state distribution at the optimal ν_{nk},

$$\begin{aligned}
\log q_{\nu_{nk}}(z_{nk}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{Z}_{-nk}}\left[\log p_K(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})\right] + c, \\
&= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{Z}_{-nk}}\left[\log p_K(z_{nk}|\pi_k) + \log p_K(X_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2)\right] + c,
\end{aligned}$$

where

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}_{-nk}}\left[\log p_K(z_{nk}|\pi_k)\right] \;=\; z_{nk}\Big[\psi(\tau_{k1})-\psi(\tau_{k2})\Big]+\psi(\tau_{k2})-\psi(\tau_{k1}+\tau_{k2}),$$

and

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}_{-nk}}\left[\log p_K(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2)\right]\\
&= \mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}_{-nk}}\left[-\frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\right)\left(\boldsymbol{X}_{n\cdot}-\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\right)^T\right]+c,\\
&= -\frac{1}{2\sigma_n^2}\mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}_{-nk}}\left[-2\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\boldsymbol{X}_{n\cdot}^T+\boldsymbol{Z}_{n\cdot}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Z}_{n\cdot}^T\right]+c,\\
&= -\frac{1}{2\sigma_n^2}\left[-2z_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T+z_{nk}\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)+2z_{nk}\bar{\boldsymbol{\phi}}_k\left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right]+c.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\log q_{\nu_{nk}}(z_{nk})\\
&= z_{nk}\left[\psi(\tau_{k1})-\psi(\tau_{k2})-\frac{1}{2\sigma_n^2}\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T-2\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T+2\bar{\boldsymbol{\phi}}_k\left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right)\right]+c.
\end{aligned}$$

From the canonical parameterisation of the Bernoulli distribution, we get that

$$\begin{aligned}
\log\frac{\nu_{nk}}{1-\nu_{nk}} &= \psi(\tau_{k1})-\psi(\tau_{k2})-\frac{1}{2\sigma_n^2}\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)+\frac{1}{\sigma_n^2}\bar{\boldsymbol{\phi}}_k\left(\boldsymbol{X}_{n\cdot}^T-\left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right),\\
&\equiv \vartheta.
\end{aligned}$$

Which gives us the update

$$\nu_{nk} \;=\; \frac{1}{1+e^{-\vartheta}}.$$

3. For the feature probabilities at the optimal τ_{k1} and τ_{k2},

$$\begin{aligned}
\log q_{\boldsymbol{\tau}_k}(\pi_k) &= \mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}}\left[\log p_K(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})\right]+c,\\
&= \mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}}\left[\log p_K(\pi_k|\alpha)+\sum_{n=1}^{N}\log p_K(z_{nk}|\pi_k)\right]+c,\\
&= \left(\frac{\alpha}{K}-1\right)\log\pi_k+\sum_{n=1}^{N}\left(\nu_{nk}\log\pi_k+(1-\nu_{nk})\log(1-\pi_k)\right)+c.
\end{aligned}$$

Hence the updates are

$$\begin{aligned}
\tau_{k1} &= \frac{\alpha}{K}+\sum_{n=1}^{N}\nu_{nk},\\
\tau_{k2} &= 1+\sum_{n=1}^{N}(1-\nu_{nk}).
\end{aligned}$$

D Derivations for the Infinite Variational Approach

This appendix derives the variational lower bound and the variational updates described in Section 5.

D.1 Variational Lower Bound

We derive expressions for each expectation in Equation (7):

1. Each stick is independent, and substituting the form of the beta prior we get

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v}}\left[\log p(v_k|\alpha)\right] &= \mathbb{E}_{\boldsymbol{v}}\left[\log\left(\alpha v_k^{\alpha-1}\right)\right], \\ &= \log\alpha + (\alpha-1)\,\mathbb{E}_{\boldsymbol{v}}\log(v_k), \\ &= \log\alpha + (\alpha-1)\left(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2})\right),\end{aligned}$$

   where $\psi(\cdot)$ is the digamma function.

2. For the feature assignments, which are Bernoulli-distributed given the feature probabilities, we first break the expectation into the following parts

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}}\left[\log p(z_{nk}|\boldsymbol{v})\right] &= \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}}\left[\log p(z_{nk}=1|\boldsymbol{v})^{z_{nk}} p(z_{nk}=0|\boldsymbol{v})^{1-z_{nk}}\right] \\ &= \mathbb{E}_{\boldsymbol{Z}}\left[z_{nk}\right]\mathbb{E}_{\boldsymbol{v}}\left[\log\prod_{m=1}^{k} v_m\right] + \mathbb{E}_{\boldsymbol{Z}}\left[1-z_{nk}\right]\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right] \\ &= \nu_{nk}\left(\sum_{m=1}^{k}\psi(\tau_{k2}) - \psi(\tau_{k1}+\tau_{k2})\right) + (1-\nu_{nk})\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right]\end{aligned}$$

   The second line follows from the definition of $\boldsymbol{v}$, while the third line follows from the properties of Bernoulli and beta distributions. In Section 5.1, we discussed how to compute a lower bound for $\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right]$ using a multinomial approximation since there is no closed form method to evaluate it. In practice, we use this lower bound for $\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right]$. The end of this subsection discusses an alternate approach that can give a closer lower bound, but is also much more computationally expensive. If we use a heuristic evaluation of this term that is not a strict lower bound, then we risk not having a definite means of evaluating convergence of our algorithm since we are no longer optimising a lower bound.

3. For the feature distribution, we simply apply the properties of expectations of Gaussians to get

$$\begin{aligned}\mathbb{E}_{\boldsymbol{A}}\left[\log p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I)\right] &= \mathbb{E}_{\boldsymbol{A}}\left[\log\left(\frac{1}{(2\pi\sigma_A^2)^{D/2}}\exp\left(-\frac{1}{2\sigma_A^2}\boldsymbol{A}_{k\cdot}^T\boldsymbol{A}_{k\cdot}\right)\right)\right], \\ &= \mathbb{E}_{\boldsymbol{A}}\left[\frac{-D}{2}\log(2\pi\sigma_A^2) - \frac{1}{2\sigma_A^2}\boldsymbol{A}_{k\cdot}^T\boldsymbol{A}_{k\cdot}\right], \\ &= \frac{-D}{2}\log(2\pi\sigma_A^2) - \frac{1}{2\sigma_A^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right).\end{aligned}$$

4. The likelihood for a particular observation is identical to the finite model, so we again have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}}\left[\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2 I)\right] \\
&= -\frac{D}{2}\log(2\pi\sigma_n^2) \\
&\quad -\frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T - 2\sum_{k=1}^{K}\nu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T + 2\sum_{k<k'}\nu_{nk}\nu_{nk'}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_{k'}^T + \sum_{k=1}^{K}\nu_{nk}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right).
\end{aligned}$$

5. The entropy can also be easily computed, since we have chosen exponential family distributions for our variational approximation:

$$\begin{aligned}
H[q] &= -\mathbb{E}_q \log\left[\prod_{k=1}^{K} q_{\boldsymbol{\tau}_k}(v_k)\prod_{k=1}^{K} q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})\prod_{k=1}^{K}\prod_{n=1}^{N} q_{\nu_{nk}}(z_{nk})\right], \\
&= \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{v}}(-\log q_{\boldsymbol{\tau}_k}(v_k)) + \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{A}}(-\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})) + \sum_{k=1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{Z}}(-\log q_{\nu_{nk}}(z_{nk})),
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{v}}(-\log q_{\boldsymbol{\tau}_k}(v_k)) &= \log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) \\
&\quad -(\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2}). \\
\mathbb{E}_{\boldsymbol{A}}(-\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})) &= \frac{1}{2}\log\left((2\pi e)^D|\Phi_k|\right). \\
\mathbb{E}_{\boldsymbol{Z}}(-\log q_{\nu_{nk}}(z_{nk})) &= -\nu_{nk}\log\nu_{nk} - (1-\nu_{nk})\log(1-\nu_{nk}).
\end{aligned}$$

Putting all the terms together gives us the variational lower bound in Equation (8).

Alternate Evaluation of $\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k}v_m\right)\right]$ We describe a Taylor series alternative to the multinomial lower bound from Section 5.1. As we noted before, the advantage of the Taylor series approximation is that we can make it arbitrarily accurate by including more terms. However, in practice, the multinomial approximation is nearly as accurate, computationally much faster, and leads to straightforward parameter updates.

Recall the Taylor series $\log(1-x) = -\sum_n^\infty \frac{x^n}{n}$ and that it converges for $x \in (-1,1)$. In our case, x corresponds to the product of probabilities, so the sum will converge unless *all* of the v_ms are one or any of them is zero. Since the distribution over the v_m are continuous densities on $[0,1]$, the series will converge with probability one.

Applying the Taylor expansion to our desired expectation, we obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k}v_m\right)\right] &= \mathbb{E}_{\boldsymbol{v}}\left[-\sum_{n=1}^{\infty}\frac{1}{n}\prod_{m=1}^{k}v_m^n\right] \\
&= -\sum_{n=1}^{\infty}\frac{1}{n}\prod_{m=1}^{k}\frac{\Gamma(\tau_{m1}+n)\Gamma(\tau_{m2}+\tau_{m1})}{\Gamma(\tau_{m1})\Gamma(\tau_{m2}+\tau_{m1}+n)} \\
&= -\sum_{n=1}^{\infty}\frac{1}{n}\prod_{m=1}^{k}\frac{(\tau_{m1})\cdots(\tau_{m1}+n-1)}{(\tau_{m2}+\tau_{m1})\cdots(\tau_{m2}+\tau_{m1}+n-1)}
\end{aligned}$$

where we have used the fact that the moments of $x \sim \mathsf{Beta}(\alpha, \beta)$ are given by

$$\mathbb{E}[x^n] = \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+n)}.$$

If we simply wish to approximate the variational lower bound, we could truncate the series after a certain number of terms. However, since all of the terms in the Taylor series are negative, truncating the series will not produce a lower bound. Thus, some extra work is required if we wish to preserve the lower bound.

To preserve the lower bound, we first note that if τ_{m2} were an integer, most terms in the numerator and denominator would cancel for $n > \tau_{m2}$. Let T be an integer greater than $\lfloor \max_{m\in 1,\ldots,k}(\tau_{m2}) \rfloor$, then we can write a lower bound for the series in the following form:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right] \\
&\geq \; -\sum_{n=1}^{T}\frac{1}{n}\prod_{m=1}^{k}\frac{(\tau_{m1})\cdots(\tau_{m1}+n-1)}{(\tau_{m1}+\tau_{m2})\cdots(\tau_{m1}+\tau_{m2}+n-1)} \\
&\quad -\prod_{m=1}^{k}((\tau_{m1})\cdots(\tau_{m1}+\lfloor\tau_{m2}\rfloor-1))\cdot\sum_{n=T+1}^{\infty}\frac{1}{n}\prod_{m=1}^{k}\frac{1}{(\tau_{m1}+n)\cdots(\tau_{m1}+\lfloor\tau_{m2}\rfloor+n-1)},
\end{aligned}$$

where we have factored the final term to make it clear that the second term has n only in the denominator. A fairly trivial upper bound on the second sum (and therefore lower bound on the expectation) is: $\zeta(1+\sum_{m=1}^{k}\lfloor\tau_{m2}\rfloor)$; a slightly better bound is $\zeta_H(1+\sum_{m=1}^{k}\lfloor\tau_{m2}\rfloor, T+1)$, where $\zeta_H(\cdot,\cdot)$ is the generalised or Hurwitz zeta function. The quality of the bound depends on the choice of T. For larger T, we have to compute more terms in the first summation, but the error introduced by the fact that the denominator of the second term is $(\tau_{m1}+n)$, not n, decreases. Empirically, we find that setting $T = \lceil 2\max_{m\in 1,\ldots,k}(\tau_{m2})\rceil$ results in very close approximations.

More precisely, we know that the Taylor series reaches the true value from above (since all of the terms in the series are negative) and that the value of the zeta function is a bound on the error. Thus, we can place the true expectation in an interval

$$\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right] \in -\sum_{n=1}^{T}\frac{1}{n}\prod_{m=1}^{k}\frac{(\tau_{m1})\cdots(\tau_{m1}+n-1)}{(\tau_{m1}+\tau_{m2})\cdots(\tau_{m1}+\tau_{m2}+n-1)} + [-\epsilon, 0],$$

where

$$\epsilon = \prod_{m=1}^{k}((\tau_{m1})\cdots(\tau_{m1}+\lfloor\tau_{m2}\rfloor-1))\zeta_H\left(1+\sum_{m=1}^{k}\lfloor\tau_{m2}\rfloor, T+1\right).$$

D.2 Parameter Updates

To optimise the parameters, we can directly optimise Equation (8). However, as in the finite case, when we are in the exponential family, we can sequentially update each of the parameters using Equation (19) from Appendix B. Regardless of how we compute the lower bound, the conditional updates for the features $\boldsymbol{A}$ and feature assignments $\boldsymbol{Z}$ remain within the exponential family, so we can use the exponential family updates. If we use the multinomial lower bound discussed in Section 5.1, then the updates for $\boldsymbol{\tau}$ will also be in the exponential family. However, the Taylor series approximation from Appendix D.1 takes us outside of the exponential family and therefore requires a numerical optimisation to update $\boldsymbol{\tau}$.

1. The updates for the features $\boldsymbol{A}$ are identical to the finite approximation; see Appendix C.2.

2. The updates for the variational distribution on $\boldsymbol{Z}$ are slightly different. For the $\boldsymbol{\nu}$ parameters,

$$\begin{aligned}\log q_{\nu_{nk}}(z_{nk}) &= \mathbb{E}_{\boldsymbol{v},\boldsymbol{A},\boldsymbol{Z}_{-nk}}\left[\log p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})\right] + c, \\ &= \mathbb{E}_{\boldsymbol{v},\boldsymbol{A},\boldsymbol{Z}_{-nk}}\left[\log p(z_{nk}|\boldsymbol{v}) + \log p(X_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2)\right] + c,\end{aligned}$$

where

$$\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}_{-nk}}\left[\log p(z_{nk}|\boldsymbol{v})\right] = z_{nk}\sum_{i=1}^{k}\left(\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})\right) + (1-z_{nk})\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{i=1}^{k} v_i\right)\right]$$

and as in Appendix C.2

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}_{-nk}}\left[\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2)\right] \\ &= -\frac{1}{2\sigma_n^2}\left[-2z_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T + z_{nk}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right) + 2z_{nk}\bar{\boldsymbol{\phi}}_k\left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right] + c.\end{aligned}$$

Therefore

$$\begin{aligned}\log q_{\nu_{nk}}(z_{nk}) = z_{nk}\Bigg[&\sum_{i=1}^{k}\left(\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})\right) - \mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{i=1}^{k} v_i\right)\right] \\ &-\frac{1}{2\sigma_n^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T - 2\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T + 2\bar{\boldsymbol{\phi}}_k\left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right)\Bigg] + c.\end{aligned}$$

From the canonical parameterisation of the Bernoulli distribution, we get that

$$\begin{aligned}\log\frac{\nu_{nk}}{1-\nu_{nk}} &= \sum_{i=1}^{k}\left(\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})\right) - \mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{i=1}^{k} v_i\right)\right] \\ &\quad -\frac{1}{2\sigma_n^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right) + \frac{1}{\sigma_n^2}\bar{\boldsymbol{\phi}}_k\left(\boldsymbol{X}_{n\cdot}^T - \left(\sum_{l:l\neq k}\nu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right) \\ &\equiv \vartheta,\end{aligned}$$

where the remaining expectation can be computed using either the multinomial approximation or the Taylor series. This gives us the update

$$\nu_{nk} = \frac{1}{1+e^{-\vartheta}}.$$

3. The updates for $\boldsymbol{\tau}$ depend on how we deal with the term $\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right]$. We first discuss the case of using a multinomial lower bound in which we have a closed form, exponential family update. We then discuss how numerical optimisation must be used for the Taylor series lower bound.

D.2.1 Multinomial Lower Bound.

When we use the multinomial bound, compute q_k and then hold q_k fixed, the terms in Equation (8) that contain τ_k are

$$\begin{aligned}\mathcal{L}_{\tau_k} &= \left[\alpha + \sum_{m=k}^{K}\sum_{n=1}^{N}\nu_{nm} + \sum_{m=k+1}^{K}\left(N - \sum_{n=1}^{N}\nu_{nm}\right)\left(\sum_{i=k+1}^{m} q_{mi}\right) - \tau_{k1}\right](\Psi(\tau_{k1}) - \Psi(\tau_{k1}+\tau_{k2})) \\ &\quad + \left[1 + \sum_{m=k}^{K}\left(N - \sum_{n=1}^{N}\nu_{nm}\right)q_{mk} - \tau_{k2}\right](\Psi(\tau_{k2}) - \Psi(\tau_{k1}+\tau_{k2})) + \ln\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right).\end{aligned}$$

We can then optimise this with respect to τ_k to find that the optimal values of τ_{k1} and τ_{k2} are

$$\begin{aligned}\tau_{k1} &= \alpha + \sum_{m=k}^{K}\sum_{n=1}^{N}\nu_{nm} + \sum_{m=k+1}^{K}\left(N - \sum_{n=1}^{N}\nu_{nm}\right)\left(\sum_{i=k+1}^{m} q_{mi}\right) \\ \tau_{k2} &= 1 + \sum_{m=k}^{K}\left(N - \sum_{n=1}^{N}\nu_{nm}\right)q_{mk}.\end{aligned}$$

These are equivalent to exponential family updates just like in the finite variational approximation.

D.2.2 Taylor Series Lower Bound.

When we use a Taylor series approximation (which we will then truncate as opposed to using a zeta function lower bound), we find that the terms in Equation (8) that contain τ_k are

$$\begin{aligned}\mathcal{L}_{\tau_k} &= (\alpha-1)(\Psi(\tau_{k1}) - \Psi(\tau_{k1}+\tau_{k2})) \\ &\quad + \sum_{m=k}^{K}\sum_{n=1}^{N}\nu_{nm}(\Psi(\tau_{k1}) - \Psi(\tau_{k1}+\tau_{k2})) \\ &\quad - \sum_{m=k}^{K}\sum_{n=1}^{N}(1-\nu_{nm})\sum_{r=1}^{\infty}\frac{1}{r}\frac{(\tau_{k1})\ldots(\tau_{k1}+r-1)}{(\tau_{k1}+\tau_{k2})\ldots(\tau_{k1}+\tau_{k2}+r-1)} \\ &\quad + \ln\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) - (\tau_{k1}-1)\Psi(\tau_{k1}) - (\tau_{k2}-1)\Psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\Psi(\tau_{k1}+\tau_{k2}).\end{aligned}$$

These is not a standard exponential family equation, so we must numerically optimise τ_k to increase the lower bound. The derivatives with respect to τ_{k1} are given by

$$\begin{aligned}&\frac{\partial\mathcal{L}_{\tau_k}}{\partial\tau_{k1}} = \\ &\quad (\alpha-1)(\Psi'(\tau_{k1}) - \Psi'(\tau_{k1}+\tau_{k2})) + \sum_{m=k}^{K}\sum_{n=1}^{N}\nu_{nm}(\Psi'(\tau_{k1}) - \Psi'(\tau_{k1}+\tau_{k2})) \\ &\quad - \sum_{m=k}^{K}\left(N - \sum_{n=1}^{N}\nu_{nm}\right)\sum_{r=1}^{\infty}\frac{1}{r}\left(\prod_{i=1}^{m}\frac{(\tau_{i1})\ldots(\tau_{i1}+r-1)}{(\tau_{i1}+\tau_{i2})\ldots(\tau_{i1}+\tau_{i2}+r-1)}\right)\sum_{j=1}^{r}\frac{\tau_{k2}}{(\tau_{k1}+j-1)(\tau_{k1}+\tau_{k2}+j-1)} \\ &\quad -(\tau_{k1}-1)\Psi'(\tau_{k1}) + (\tau_{k1}+\tau_{k2}-2)\Psi'(\tau_{k1}+\tau_{k2}).\end{aligned}$$

Similarly, the derivatives with respect to τ_{k2} are given by

$$\begin{aligned}\frac{\partial \mathcal{L}_{\tau_k}}{\partial \tau_{k2}} &= (\alpha-1)\left(-\Psi'(\tau_{k1}+\tau_{k2})\right) - \sum_{m=k}^{K}\sum_{n=1}^{N}\nu_{nm}\Psi'(\tau_{k1}+\tau_{k2}) \\ &\quad - \sum_{m=k}^{K}\left(N-\sum_{n=1}^{N}\nu_{nm}\right)\sum_{r=1}^{\infty}\frac{1}{r}\left(\prod_{i=1}^{m}\frac{(\tau_{i1})\dots(\tau_{i1}+r-1)}{(\tau_{i1}+\tau_{i2})\dots(\tau_{i1}+\tau_{i2}+r-1)}\right)\sum_{j=1}^{r}\frac{-1}{\tau_{k1}+\tau_{k2}+j-1} \\ &\quad -(\tau_{k2}-1)\Psi'(\tau_{k2})+(\tau_{k1}+\tau_{k2}-2)\Psi'(\tau_{k1}+\tau_{k2}).\end{aligned}$$

These can be computed for any particular parameter choices (and an arbitrary truncation level of the infinite sum). Also note that several computations can be reused across k, and others can be computed iteratively across r. We can plug these derivatives into an optimisation routine to get updates for τ_{k1} and τ_{k2}.

E Variational Inference for the iICA model

In this section we describe the variational approach to do approximate inference for the infinite Independent Component Analysis model. We refer to Knowles and Ghahramani (2007) for more details regarding the model, but to set the notation, we state the iICA model

$$\begin{aligned}
v_k &\sim \mathsf{Beta}(\alpha,1) && \text{for } k\in\{1,\dots,\infty\},\\
\pi_k &= \prod_{i=1}^{k} v_i && \text{for } k\in\{1\dots\infty\},\\
z_{nk} &\sim \mathsf{Bernoulli}(\pi_k) && \text{for } k\in\{1\dots\infty\},\\
s_{nk} &\sim \mathsf{Laplace}(1) && \text{for } k\in\{1\cdots K\}, n\in\{1\cdots N\},\\
\boldsymbol{A}_{k\cdot} &\sim \mathsf{Normal}(0,\sigma_A^2 I) && \text{for } k\in\{1\cdots\infty\},\\
\boldsymbol{X}_{n\cdot} &\sim \mathsf{Normal}((\boldsymbol{Z}_{n\cdot}\odot\boldsymbol{S}_{n\cdot})\boldsymbol{A},\sigma_n^2 I) && \text{for } n\in\{1\cdots N\},
\end{aligned}$$

where $\odot$ denotes pointwise multiplication between two vectors. In other words, we can write the joint probability of the data and latent variables as

$$p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{k=1}^{\infty}\left(p(\pi_k|\alpha)p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I)\prod_{n=1}^{N}p(z_{nk}|\pi_k)p(s_{nk})\right)\prod_{n=1}^{N}p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\boldsymbol{S},\sigma_n^2 I).$$

As in the linear-Gaussian model and as we will discuss in Section E.2, doing exact posterior inference on the latent variables $\boldsymbol{W}$ is intractable. We will propose two different approximation schemes just as in the linear-Gaussian model: first we introduce a finite approximation similar to the derivation in Section C and second we describe an infinite variational approximation to the iICA model analogous to Section D. Similar to our discussion of the linear-Gaussian model, we will refer to the distribution $p_K(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})$ as the finite beta-Bernoulli approximation of order K while $p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})$ refers to the exact iICA distribution defined above.

E.1 The Finite Variational Approach

A finite beta-Bernoulli approximation to the iICA model can be described as follows

$$\begin{aligned}
\pi_k &\sim \mathsf{Beta}(\alpha/K, 1) && \text{for } k \in \{1 \cdots K\},\\
z_{nk} &\sim \mathsf{Bernoulli}(\pi_k) && \text{for } k \in \{1 \cdots K\}, n \in \{1 \cdots N\},\\
s_{nk} &\sim \mathsf{Laplace}(1) && \text{for } k \in \{1 \cdots K\}, n \in \{1 \cdots N\},\\
\boldsymbol{A}_{k\cdot} &\sim \mathsf{Normal}(0, \sigma_A^2 I) && \text{for } k \in \{1 \cdots K\},\\
\boldsymbol{X}_{n\cdot} &\sim \mathsf{Normal}((\boldsymbol{Z}_{n\cdot} \odot \boldsymbol{S}_{n\cdot})\boldsymbol{A}, \sigma_n^2 I) && \text{for } n \in \{1 \cdots N\}.
\end{aligned}$$

where K is some finite (but large) truncation level. We refer to the set of hidden variables as $\boldsymbol{W} = \{\boldsymbol{\pi}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{S}\}$ and the set of parameters as $\boldsymbol{\theta} = \{\alpha, \sigma_A^2, \sigma_n^2\}$. Using this notation we can write the joint probability of the data and latent variables as

$$p_K(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \left(p(\pi_k|\alpha) p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I) \prod_{n=1}^{N} p(z_{nk}|\pi_k) p(s_{nk}) \right) \prod_{n=1}^{N} p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot}, \boldsymbol{A}, \boldsymbol{S}, \sigma_n^2 I).$$

We are interested in the posterior, or equivalently the log posterior, of the latent variables

$$\log p_K(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}) = \log p_K(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) - \log p_K(\boldsymbol{X}|\boldsymbol{\theta}). \tag{20}$$

For similar reasons to the linear-Gaussian model in Section 4, computing this quantity is intractable. Hence we use the following variational distribution as an approximation

$$q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{\pi}) q_{\boldsymbol{\phi}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z}) q_{\boldsymbol{\mu},\boldsymbol{\eta}}(\boldsymbol{S}).$$

where

- $q_{\boldsymbol{\tau}_k}(\pi_k) = \mathsf{Beta}(\pi_k; \tau_{k1}, \tau_{k2})$,
- $q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) = \mathsf{Normal}(\boldsymbol{A}_{k\cdot}; \bar{\boldsymbol{\phi}}_k, \boldsymbol{\Phi}_k)$,
- $q_{\nu_{nk}}(z_{nk}) = \mathsf{Bernoulli}(z_{nk}; \nu_{nk})$,
- $q_{\mu_{nk},\eta_{nk}}(s_{nk}) = \mathsf{Laplace}(s_{nk}; \mu_{nk}, \eta_{nk})$ where μ_{nk} is the mean and η_{nk} is the scale parameter of the Laplace distribution.

In contrast to the linear-Gaussian model, we now need to optimise the parameters $\boldsymbol{\mu}, \boldsymbol{\eta}$ in addition to $\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\nu}$ with the goal of minimizing KL divergence $D(q||p_K)$ or equivalently, maximise the following lower bound on $p_K(\boldsymbol{X}|\boldsymbol{\theta})$:

$$\mathbb{E}_q[\log(p_K(\boldsymbol{X}, \boldsymbol{W}|\theta)] + H[q].$$

As we discussed in the context of the linear-Gaussian model, inference with respect to this beta-Bernoulli model p_K is not the same as variational inference with respect to the true iICA model. The variational updates are significantly easier though and in the limit of large K, the finite beta-Bernoulli model is equivalent to the iICA model.

E.1.1 Variational Lower Bound

We expand the lower bound on $\log p_K(\boldsymbol{X}|\boldsymbol{\theta})$ into its components

$$\begin{aligned}\log p_K(\boldsymbol{X}|\boldsymbol{\theta}) &\geq \mathbb{E}_{\boldsymbol{W}}\left[\log p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})\right] + H[q], \qquad (21)\\ &= \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\pi}}\left[\log p(\pi_k|\alpha)\right] + \sum_{k=1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}}\left[\log p(z_{nk}|\pi_k)\right] + \sum_{k=1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{S}}\left[\log p(s_{nk})\right]\\ &\quad + \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{A}}\left[\log p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I)\right] + \sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A},\boldsymbol{S}}\left[\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\boldsymbol{S}_{n\cdot},\sigma_n^2 I)\right] + H[q],\end{aligned}$$

where the expectation are computed with respect to the variational distribution q. We derive expressions for each expectation in Equation (21):

1. The feature probabilities,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\pi}}\left[\log p(\pi_k|\alpha)\right] &= \mathbb{E}_{\boldsymbol{\pi}}\left[\log\left(\frac{\alpha}{K}\pi_k^{\alpha/K-1}\right)\right],\\ &= \log\frac{\alpha}{K} + \left(\frac{\alpha}{K}-1\right)\left(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2})\right),\end{aligned}$$

   where $\psi(\cdot)$ is the digamma function.

2. The signal distribution,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{S}}\left[\log p(s_{nk})\right] &= \mathbb{E}_{\boldsymbol{S}}\left[\log\left(\frac{1}{2}\exp\left(-|s_{nk}|\right)\right)\right],\\ &= -\log 2 - \left(|\mu_{nk}| + \eta_{nk}\exp\left(-\frac{|\mu_{nk}|}{\eta_{nk}}\right)\right).\end{aligned}$$

3. The feature state distribution,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}}\left[\log p(z_{nk}|\pi_k)\right] &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}}\left[\log\left(\pi_k^{z_{nk}}(1-\pi_k)^{1-z_{nk}}\right)\right],\\ &= \nu_{nk}\psi(\tau_{k1}) + (1-\nu_{nk})\psi(\tau_{k2}) - \psi(\tau_{k1}+\tau_{k2}).\end{aligned}$$

4. The feature distribution,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{A}}\left[\log p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I)\right] &= \mathbb{E}_{\boldsymbol{A}}\left[\log\left(\frac{1}{(2\pi\sigma_A^2)^{D/2}}\exp\left(-\frac{1}{2\sigma_A^2}\boldsymbol{A}_{k\cdot}^T\boldsymbol{A}_{k\cdot}\right)\right)\right],\\ &= \frac{-D}{2}\log(2\pi\sigma_A^2) - \frac{1}{2\sigma_A^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right).\end{aligned}$$

5. The likelihood,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{A},\boldsymbol{S}}\left[\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\boldsymbol{S},\sigma_n^2 I)\right]\\ &= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A},\boldsymbol{S}}\left[-\frac{D}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot} - (\boldsymbol{Z}_{n\cdot}\odot\boldsymbol{S}_{n\cdot})\boldsymbol{A}\right)\left(\boldsymbol{X}_{n\cdot} - (\boldsymbol{Z}_{n\cdot}\odot\boldsymbol{S}_{n\cdot})\boldsymbol{A}\right)^T\right],\\ &= -\frac{D}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\Bigg(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T - 2\sum_{k=1}^{K}\nu_{nk}\mu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T\\ &\qquad + 2\sum_{k<k'}\nu_{nk}\mu_{nk}\nu_{nk'}\mu_{nk'}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_{k'}^T + \sum_{k=1}^{K}\nu_{nk}(2\eta_{nk}^2+\mu_{nk}^2)\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\Bigg),\end{aligned}$$

where we use the fact that

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{S},\boldsymbol{A}}[(\boldsymbol{Z}_{n\cdot} \odot \boldsymbol{S}_{n\cdot})\boldsymbol{A}\boldsymbol{A}^T(\boldsymbol{Z}_{n\cdot} \odot \boldsymbol{S}_{n\cdot})^T] \\
&= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A},\boldsymbol{S}}\left[\left(\sum_{k=1}^{K}(z_{nk}s_{nk})\boldsymbol{A}_{k\cdot}\right)\left(\sum_{k=1}^{K}(z_{nk}s_{nk})\boldsymbol{A}_{k\cdot}\right)^T\right], \\
&= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A},\boldsymbol{S}}\left[\sum_{d=1}^{D}\left(\sum_{k=1}^{K}z_{nk}s_{nk}^2A_{kd}^2 + \sum_{k,k':k'\neq k}(z_{nk}s_{nk})(z_{nk'}s_{nk'})\boldsymbol{A}_{kd}\boldsymbol{A}_{k'd}\right)\right], \\
&= \sum_{k=1}^{K}\nu_{nk}(2\eta_{nk}^2+\mu_{nk}^2)\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right) + 2\sum_{k<k'}\nu_{nk}\mu_{nk}\nu_{nk'}\mu_{nk'}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_{k'}^T.
\end{aligned}
$$

6. Finally, for the entropy,

$$
\begin{aligned}
H[q] &= -\mathbb{E}_q \log\left[\prod_{k=1}^{K}q_{\boldsymbol{\tau}_k}(\pi_k)\prod_{k=1}^{K}q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})\prod_{k=1}^{K}\prod_{n=1}^{N}q_{\nu_{nk}}(z_{nk})\prod_{k=1}^{K}\prod_{n=1}^{N}q_{\mu_{nk},\eta_{nk}}(s_{nk})\right], \\
&= \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\pi}}(-\log q_{\boldsymbol{\tau}_k}(\pi_k)) + \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{A}}(-\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})) \\
&\quad + \sum_{k=1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{Z}}(-\log q_{\nu_{nk}}(z_{nk})) + \sum_{k=1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{S}}(-\log q_{\mu_{nk},\eta_{nk}}(s_{nk})),
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi}}(-\log q_{\boldsymbol{\tau}_k}(\pi_k)) &= \log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) - (\tau_{k1}-1)\psi(\tau_{k1}) \\
&\quad -(\tau_{k2}-1)\psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2}), \\
\mathbb{E}_{\boldsymbol{A}}(-\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot})) &= \frac{1}{2}\log\left((2\pi e)^D|\Phi_k|\right), \\
\mathbb{E}_{\boldsymbol{Z}}(-\log q_{\nu_{nk}}(z_{nk})) &= -\nu_{nk}\log\nu_{nk} - (1-\nu_{nk})\log(1-\nu_{nk}). \\
\mathbb{E}_{\boldsymbol{S}}(-\log q_{\mu_{nk},\eta_{nk}}(s_{nk})) &= \log(2e\eta_{nk}).
\end{aligned}
$$

Collecting all the above computations together in Equation (21) gives us the variational lower bound on $\log p_K(\boldsymbol{X}|\boldsymbol{\theta})$:

$$
\begin{aligned}
&\log p_K(\boldsymbol{X}|\boldsymbol{\theta}) \\
&\geq \sum_{k=1}^{K}\left[\log\frac{\alpha}{K}+\left(\frac{\alpha}{K}-1\right)\left(\psi(\tau_{k1})-\psi(\tau_{k1}+\tau_{k2})\right)\right] \\
&\quad -KN\log 2-\sum_{k=1}^{K}\sum_{n=1}^{N}\left(|\mu_{nk}|+\eta_{nk}\exp\left(-\frac{|\mu_{nk}|}{\eta_{nk}}\right)\right) \\
&\quad +\sum_{k=1}^{K}\sum_{n=1}^{N}\left[\nu_{nk}\psi(\tau_{k1})+(1-\nu_{nk})\psi(\tau_{k2})-\psi(\tau_{k1}+\tau_{k2})\right] \\
&\quad +\sum_{k=1}^{K}\left[\frac{-D}{2}\log(2\pi\sigma_A^2)-\frac{1}{2\sigma_A^2}\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right] \\
&\quad +\sum_{n=1}^{N}\left[-\frac{D}{2}\log(2\pi\sigma_n^2)-\frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T-2\sum_{k=1}^{K}\nu_{nk}\mu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T+2\sum_{k<k'}\nu_{nk}\mu_{nk}\nu_{nk'}\mu_{nk'}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_{k'}^T\right.\right. \\
&\qquad\qquad\left.\left.+\sum_{k=1}^{K}\nu_{nk}(2\eta_{nk}^2+\mu_{nk}^2)\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right)\right] \\
&\quad +\sum_{k=1}^{K}\left[\log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right)-(\tau_{k1}-1)\psi(\tau_{k1})-(\tau_{k2}-1)\psi(\tau_{k2})+(\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\right] \\
&\quad +\sum_{k=1}^{K}\left[\frac{1}{2}\log\left((2\pi e)^D|\Phi_k|\right)\right]+\sum_{k=1}^{K}\sum_{n=1}^{N}\left[-\nu_{nk}\log\nu_{nk}-(1-\nu_{nk})\log(1-\nu_{nk})+\log(2e\eta_{nk})\right].
\end{aligned}
$$

E.1.2 Parameter Updates

When we optimise the lower bound on $\log p_K(\boldsymbol{X}|\boldsymbol{\theta})$ we perform coordinate-wise gradient ascent by cycling through the variational parameters and update them in turn. For most parameter we will be able to use the standard exponential family variational update from Equation (19). For some parameters we will need to compute the gradient and perform a local gradient ascent step. Throughout this section, we let c be a constant independent of the variable of interest that may change from line to line.

1. For the feature distribution at the optimal $\bar{\boldsymbol{\phi}}_k$ and Φ_k,

$$
\begin{aligned}
&\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) \\
&\quad = \mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z},\boldsymbol{S}}\left[\log p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})\right]+c, \\
&\quad = \mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z},\boldsymbol{S}}\left[\log p(\boldsymbol{A}_{k\cdot}|\sigma_A^2)+\sum_{n=1}^{N}p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{S}_{n\cdot},\boldsymbol{A},\sigma_n^2)\right]+c, \\
&\quad = -\frac{1}{2\sigma_A^2}\left(\boldsymbol{A}_{k\cdot}\boldsymbol{A}_{k\cdot}^T\right)-\frac{1}{2\sigma_n^2}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{A}_{-k},\boldsymbol{Z},\boldsymbol{S}}\left[\left(\boldsymbol{X}_{n\cdot}-(\boldsymbol{Z}_{n\cdot}\odot\boldsymbol{S}_{n\cdot})\boldsymbol{A}\right)\left(\boldsymbol{X}_{n\cdot}-(\boldsymbol{Z}_{n\cdot}\odot\boldsymbol{S}_{n\cdot})\boldsymbol{A}\right)^T\right]+c, \\
&\quad = -\frac{1}{2}\left[\boldsymbol{A}_{k\cdot}\left(\frac{1}{\sigma_A^2}+\frac{\sum_{n=1}^{N}\nu_{nk}(2\eta_{nk}^2+\mu_{nk}^2)}{\sigma_n^2}\right)\boldsymbol{A}_{k\cdot}^T-2\boldsymbol{A}_{k\cdot}\left(\frac{1}{\sigma_n^2}\sum_{n=1}^{N}\nu_{nk}\mu_{nk}\left(\boldsymbol{X}_{n\cdot}-\left(\sum_{l:l\neq k}\nu_{nl}\mu_{nl}\bar{\boldsymbol{\phi}}_l\right)\right)\right)^T\right]+c
\end{aligned}
$$

Completing the square and using Equation (19) gives us that for the optimal parameter settings we must have

$$\begin{aligned}\log q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) &= -\frac{1}{2}\left(\boldsymbol{A}_{k\cdot}\boldsymbol{\Phi}_k^{-1}\boldsymbol{A}_{k\cdot}^T - 2\boldsymbol{A}_{k\cdot}\boldsymbol{\Phi}_k^{-1}\bar{\boldsymbol{\phi}}_k^T\right) + c,\\ &= -\frac{1}{2}\left(\boldsymbol{A}_{k\cdot} - \bar{\boldsymbol{\phi}}_k\right)\boldsymbol{\Phi}_k^{-1}\left(\boldsymbol{A}_{k\cdot} - \bar{\boldsymbol{\phi}}_k\right)^T + c.\end{aligned}$$

hence the parameter updates are

$$\begin{aligned}\bar{\boldsymbol{\phi}}_k &= \left[\frac{1}{\sigma_n^2}\sum_{n=1}^{N}\nu_{nk}\mu_{nk}\left(\boldsymbol{X}_{n\cdot} - \left(\sum_{l:l\neq k}\nu_{nl}\mu_{nl}\bar{\boldsymbol{\phi}}_l\right)\right)\right]\left(\frac{1}{\sigma_A^2} + \frac{\sum_{n=1}^{N}\nu_{nk}(2\eta_{nk}^2+\mu_{nk}^2)}{\sigma_n^2}\right)^{-1},\\ \boldsymbol{\Phi}_k &= \left(\frac{1}{\sigma_A^2} + \frac{\sum_{n=1}^{N}\nu_{nk}(2\eta_{nk}^2+\mu_{nk}^2)}{\sigma_n^2}\right)^{-1} I.\end{aligned}$$

2. For the feature state distribution at the optimal ν_{nk},

$$\begin{aligned}\log q_{\nu_{nk}}(z_{nk}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{Z}_{-nk},\boldsymbol{S}}\left[\log p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})\right] + c,\\ &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{Z}_{-nk},\boldsymbol{S}}\left[\log p(z_{nk}|\pi_k) + \log p(X_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\boldsymbol{S},\sigma_n^2)\right] + c,\end{aligned}$$

where

$$\mathbb{E}_{\boldsymbol{\pi}}\left[\log p(z_{nk}|\pi_k)\right] = z_{nk}\Big[\psi(\tau_{k1}) - \psi(\tau_{k2})\Big] + \psi(\tau_{k2}) - \psi(\tau_{k1}+\tau_{k2}),$$

and

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}_{-nk},\boldsymbol{S}}\left[\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2)\right]\\ &= \mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}_{-nk},\boldsymbol{S}}\left[-\frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot} - (\boldsymbol{Z}_{n\cdot}\odot\boldsymbol{S}_{n\cdot})\boldsymbol{A}\right)\left(\boldsymbol{X}_{n\cdot} - (\boldsymbol{Z}_{n\cdot}\odot\boldsymbol{S}_{n\cdot})\boldsymbol{A}\right)^T\right] + c,\\ &= -\frac{1}{2\sigma_n^2}\left[-2z_{nk}\mu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T + z_{nk}(2\eta_{nk}^2+\mu_{nk}^2)\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right) + 2z_{nk}\mu_{nk}\bar{\boldsymbol{\phi}}_k\left(\sum_{l:l\neq k}\nu_{nl}\mu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right] + c.\end{aligned}$$

Therefore

$$\begin{aligned}&\log q_{\nu_{nk}}(z_{nk})\\ &= z_{nk}\left[\psi(\tau_{k1}) - \psi(\tau_{k2}) - \frac{1}{2\sigma_n^2}\left((2\eta_{nk}^2+\mu_{nk}^2)(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T)\right.\right.\\ &\qquad \left.\left. -2\mu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T + 2\mu_{nk}\bar{\boldsymbol{\phi}}_k\left(\sum_{l:l\neq k}\nu_{nl}\mu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right)\right] + c.\end{aligned}$$

From the canonical parameterisation of the Bernoulli distribution, we get that

$$\begin{aligned}\log\frac{\nu_{nk}}{1-\nu_{nk}} &= \psi(\tau_{k1}) - \psi(\tau_{k2}) - \frac{2\eta_{nk}^2+\mu_{nk}^2}{2\sigma_n^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right) + \frac{\mu_{nk}}{\sigma_n^2}\bar{\boldsymbol{\phi}}_k\left(\boldsymbol{X}_{n\cdot}^T - \left(\sum_{l:l\neq k}\mu_{nl}\nu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right),\\ &\equiv \vartheta.\end{aligned}$$

which gives us the update

$$\nu_{nk} = \frac{1}{1+e^{-\vartheta}}.$$

3. The updates τ_{ki1} and τ_{k2} are only dependent on $\boldsymbol{Z}$ and $\boldsymbol{\pi}$ and hence are exactly the same as for the linear-Gaussian model. We refer to Section C.2 for details.

4. Finally, because the Laplace prior on the signal matrix $\boldsymbol{S}$ is not in the exponential family we cannot use the standard variational Bayes update formula as the posterior will not be in the exponential family anymore. Hence in order to optimise μ_{nk} and η_{nk} we perform a gradient ascent step on the variational lower bound. Let us denote the variational lower bound which we derived earlier with $\mathcal{L}$. The components of $\mathcal{L}$ that depend on μ_{nk} and η_{nk} are

$$-\tfrac{1}{2\sigma_n^2}\left(-2\nu_{nk}\mu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T + 2\left(\nu_{nk}\mu_{nk}\bar{\boldsymbol{\phi}}_k\right)\left(\textstyle\sum_{k':k'\neq k}\nu_{nk'}\mu_{nk'}\bar{\boldsymbol{\phi}}_{k'}^T\right) + \nu_{nk}(2\eta_{nk}^2+\mu_{nk}^2)\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right)$$
$$-|\mu_{nk}| - \eta_{nk}\exp\left(-\tfrac{|\mu_{nk}|}{\eta_{nk}}\right) + \log(2e\eta_{nk}).$$

It is now straightforward to compute the derivative of $\mathcal{L}$ with respect to the mean parameter of the Laplace distribution

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mu_{nk}} &= \frac{1}{\sigma_n^2}\left(\nu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T - \left(\nu_{nk}\bar{\boldsymbol{\phi}}_k\right)\left(\sum_{k\neq k'}\nu_{nk'}\mu_{nk'}\bar{\boldsymbol{\phi}}_{k'}^T\right) - \nu_{nk}\mu_{nk}\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right)\\ &\quad -\text{sign}(\mu_{nk})\left(1-\exp\left(-\frac{|\mu_{nk}|}{\eta_{nk}}\right)\right),\end{aligned}$$

and the derivative of $\mathcal{L}$ with respect to the scale of the Laplace distribution

$$\frac{\partial\mathcal{L}}{\partial\eta_{nk}} = -\frac{2\nu_{nk}\eta_{nk}}{\sigma_n^2}(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T) - \exp\left(-\frac{|\mu_{nk}|}{\eta_{nk}}\right) - \frac{|\mu_{nk}|}{\eta_{nk}}\exp\left(-\frac{|\mu_{nk}|}{\eta_{nk}}\right) + \frac{1}{\eta_{nk}}.$$

We can then numerically optimise μ_{nk} and η_{nk}.

E.2 The Infinite Variational Approach

We presented the iICA model at the start of Section E. Recall that our goal is to compute the log posterior

$$\log p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{\theta}) = \log p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta}) - \log p(\boldsymbol{X}|\boldsymbol{\theta}),$$

but this is intractable to compute. We use a variational approximation based on the truncated stick-breaking process as described in Section 5 and use the stick-breaking variables $\boldsymbol{v}$ instead of $\boldsymbol{\mu}$ as before. Our mean field variational distribution is now

$$q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{v})q_{\boldsymbol{\phi}}(\boldsymbol{A})q_{\boldsymbol{\nu}}(\boldsymbol{Z})q_{\boldsymbol{\mu},\boldsymbol{\eta}}(\boldsymbol{S}).$$

where

- $q_{\boldsymbol{\tau}_k}(v_k) = \mathsf{Beta}(v_k;\tau_{k1},\tau_{k2})$,
- $q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) = \mathsf{Normal}(\boldsymbol{A}_{k\cdot};\bar{\boldsymbol{\phi}}_k,\boldsymbol{\Phi}_k)$,
- $q_{\nu_{nk}}(z_{nk}) = \mathsf{Bernoulli}(z_{nk};\nu_{nk})$,
- $q_{\mu_{nk},\eta_{nk}}(s_{nk}) = \mathsf{Laplace}(s_{nk};\mu_{nk},\eta_{nk})$ where μ_{nk} is the mean and η_{nk} is the scale parameter of the Laplace distribution.

As with the finite approach, inference involves optimising $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, $\boldsymbol{\nu}$, $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ to minimise the KL divergence $D(q||p)$, or equivalently to maximise the lower bound on $p(\boldsymbol{X}|\boldsymbol{\theta})$

$$\mathbb{E}_q[\log(p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})] + H[q].$$

Although the update equations for this approximation are not as straightforward as in the finite approach, we can reuse many of the computations we did for the linear-Gaussian and beta-Bernoulli iICA approximation.

E.2.1 Variational Lower Bound

As in the finite approach, we first derive an expression for the variational lower bound. However, as in the linear-Gaussian model, parts of our model are no longer in the exponential family and require nontrivial computations. We expand the lower bound on $\log p(\boldsymbol{X}|\boldsymbol{\theta})$ into its components

$$\begin{aligned}
\log p(\boldsymbol{X}|\boldsymbol{\theta}) &\geq \mathbb{E}_{\boldsymbol{W}}\left[\log p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})\right] + H[q], \qquad (22)\\
&= \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\pi}}\left[\log p(v_k|\alpha)\right] + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{Z}}\left[\log p(z_{nk}|\pi_k)\right] + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{S}}\left[\log p(s_{nk})\right]\\
&\quad + \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{A}}\left[\log p(\boldsymbol{A}_{k\cdot}|\sigma_A^2 I)\right] + \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A},\boldsymbol{S}}\left[\log p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}_{n\cdot}, \boldsymbol{A}, \boldsymbol{S}_{n\cdot}, \sigma_n^2 I)\right] + H[q],
\end{aligned}$$

where the expectation are computed with respect to the variational distribution q. From Appendix E.1.1 we know how to compute all the expectations for the ICA likelihood; together with the theory from Section 5.1 and Appendix D.1 on the lower bound for the infinite variational approximation to the linear-Gaussian model, we can expand the expectation in Equation (21):

$$\begin{aligned}
&\log p(\boldsymbol{X}|\boldsymbol{\theta})\\
&\geq \sum_{k=1}^{K}\left[\log\alpha + (\alpha-1)\left(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2})\right)\right]\\
&\quad + \sum_{k=1}^{K}\sum_{n=1}^{N}\left[\nu_{nk}\left(\sum_{m=1}^{k}\psi(\tau_{k2}) - \psi(\tau_{k1}+\tau_{k2})\right) + (1-\nu_{nk})\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1 - \prod_{m=1}^{k} v_m\right)\right]\right]\\
&\quad - KN\log 2 - \sum_{k=1}^{K}\sum_{n=1}^{N}\left(|\mu_{nk}| + \eta_{nk}\exp\left(-\frac{|\mu_{nk}|}{\eta_{nk}}\right)\right)\\
&\quad + \sum_{k=1}^{K}\left[\frac{-D}{2}\log(2\pi\sigma_A^2) - \frac{1}{2\sigma_A^2}\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right]\\
&\quad + \sum_{n=1}^{N}\left[-\frac{D}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\left(\boldsymbol{X}_{n\cdot}\boldsymbol{X}_{n\cdot}^T - 2\sum_{k=1}^{K}\nu_{nk}\mu_{nk}\bar{\boldsymbol{\phi}}_k\boldsymbol{X}_{n\cdot}^T + 2\sum_{k<k'}\nu_{nk}\mu_{nk}\nu_{nk'}\mu_{nk'}\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_{k'}^T\right.\right.\\
&\qquad\qquad \left.\left. + \sum_{k=1}^{K}\nu_{nk}(2\eta_{nk}^2 + \mu_{nk}^2)\left(tr(\boldsymbol{\Phi}_k) + \bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)\right)\right]\\
&\quad + \sum_{k=1}^{K}\left[\log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) - (\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\right]\\
&\quad + \sum_{k=1}^{K}\left[\frac{1}{2}\log\left((2\pi e)^D|\Phi_k|\right)\right] + \sum_{k=1}^{K}\sum_{n=1}^{N}\left[-\nu_{nk}\log\nu_{nk} - (1-\nu_{nk})\log(1-\nu_{nk}) + \log(2e\eta_{nk})\right].
\end{aligned}$$

The term $\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right]$ we left unevaluated. It can be evaluated either with the multinomial approach from Section 5.1 (recommended) or the Taylor series approach from Appendix D.1

E.2.2 Parameter Updates

For the infinite variational approximation to the iICA model, we need to update the parameters $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$. For the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$, the updates are exactly the same as for the finite iICA approximation in Section C.2. For the parameters of $\boldsymbol{Z}$ we update ν_{nk} in $\mathsf{Bernoulli}(z_{nk};\nu_{nk})$ as

$$\nu_{nk} = \frac{1}{1+e^{-\vartheta}}$$

where

$$\begin{aligned}\vartheta &= \sum_{i=1}^{k}\left(\psi(\tau_{i1})-\psi(\tau_{i1}+\tau_{i2})\right)-\mathbb{E}_{\boldsymbol{v}}[\log(1-\prod_{i=1}^{k} v_i)] \\ &\quad -\frac{2\eta_{nk}^2+\mu_{nk}^2}{2\sigma_n^2}\left(tr(\boldsymbol{\Phi}_k)+\bar{\boldsymbol{\phi}}_k\bar{\boldsymbol{\phi}}_k^T\right)+\frac{\mu_{nk}}{\sigma_n^2}\bar{\boldsymbol{\phi}}_k\left(\boldsymbol{X}_{n\cdot}^T-\left(\sum_{l:l\neq k}\mu_{nl}\nu_{nl}\bar{\boldsymbol{\phi}}_l^T\right)\right).\end{aligned}$$

We leave the term $\mathbb{E}_{\boldsymbol{v}}[\log(1-\prod_{i=1}^{k} v_i)]$ unevaluated because the choice of how to approximate it does not change the form of the update.

Finally, to update τ_{k1} and τ_{k2} in $\mathsf{Beta}(v_k;\tau_{k1},\tau_{k2})$ we use the multinomial lower bound for $\mathbb{E}_{\boldsymbol{v}}[\log(1-\prod_{i=1}^{k} v_i)]$ and compute q_{ki} according to Equation (10). As in the linear-Gaussian model, the updates for τ_{k1} and τ_{k2} have the closed form

$$\begin{aligned}\tau_{k1} &= \alpha+\sum_{m=k}^{K}\sum_{n=1}^{N}\nu_{nm}+\sum_{m=k+1}^{K}\left(N-\sum_{n=1}^{N}\nu_{nm}\right)\left(\sum_{i=k+1}^{m} q_{mi}\right) \\ \tau_{k2} &= 1+\sum_{m=k}^{K}\left(N-\sum_{n=1}^{N}\nu_{nm}\right)q_{mk}.\end{aligned}$$

F Alternate Bounds for the Infinite Approximation

Recall that when giving a bound for how close $m_K(\boldsymbol{X})$ is to $m(\boldsymbol{X})$, we must bound

$$1-\mathbb{E}\left(\left[\prod_{i=K+1}^{\infty}(1-\pi_i)\right]^N\right).$$

We gave one approach to bounding this in Section 6. This approach applied Jensen's inequality as follows

$$1-\mathbb{E}\left(\left[\prod_{i=K+1}^{\infty}(1-\pi_i)\right]^N\right) \leq 1-\left(\mathbb{E}\left[\prod_{i=K+1}^{\infty}(1-\pi_i)\right]\right)^N$$

and then relied on the beta process representation from Thibaux and Jordan (2007) to bound the inner term.

Another approach to applying Jensen's inequality is to write

$$1 - \mathbb{E}\left(\left[\prod_{i=K+1}^{\infty}(1-\pi_i)\right]^N\right) \leq 1 - \exp\left(N\sum_{i=K+1}^{\infty}\mathbb{E}\log\left(1-\prod_{j=1}^{i}v_j\right)\right), \tag{23}$$

where v_j are the stick-breaking weights. This section derives a heuristic bound and a principled bound both based on this alternate application of Jensen's inequality.

To use Equation (23), we need to evaluate or bound $\sum_{i=K+1}^{\infty}\mathbb{E}\log\left(1-\prod_{j=1}^{i}v_j\right)$. We expand the expectation using the Taylor series approximation of Appendix D.1, noting that the expectation $\mathbb{E}[v^r]$ of a Beta$(\alpha,1)$ random variables is $\frac{\alpha}{\alpha+r}$:

$$\begin{aligned}
\sum_{i=K+1}^{\infty}\mathbb{E}\log\left(1-\prod_{j=1}^{i}v_j\right) &= -\sum_{i=K+1}^{\infty}\sum_{r=1}^{\infty}\frac{1}{r}\prod_{j=1}^{i}\frac{\alpha}{\alpha+r} \\
&= -\sum_{r=1}^{\infty}\frac{1}{r}\sum_{i=K+1}^{\infty}\left(\frac{\alpha}{\alpha+r}\right)^i \\
&= -\sum_{r=1}^{\infty}\frac{1}{r^2}\frac{\alpha^{K+1}}{(\alpha+r)^K}.
\end{aligned}$$

Substituting the expression above into the original bound, we get

$$1 - \mathbb{E}\left(\left[\prod_{i=K+1}^{\infty}(1-\pi_i)\right]^N\right) \leq 1 - \exp\left(-N\sum_{r=1}^{\infty}\frac{1}{r^2}\frac{\alpha^{K+1}}{(\alpha+r)^K}\right).$$

For any truncation level of the sum, we do not necessarily have a true bound, but we find empirically that it is very close to the true truncation bound.

To get a strict bound, we can write[6]

$$\begin{aligned}
\sum_{r=1}^{\infty}\frac{1}{r^2}\frac{\alpha^{K+1}}{(\alpha+r)^K} &\leq \frac{\alpha^{K+1}}{(\alpha+1)^K} + \int_1^{\infty}\frac{1}{r^2}\frac{\alpha^{K+1}}{(\alpha+r)^K} \\
&= \frac{\alpha^{K+1}}{(\alpha+1)^K} + \int_0^{1}\frac{\alpha^{K+1}}{(\alpha+\frac{1}{t})^K} \\
&= \frac{\alpha^{K+1}}{(\alpha+1)^K} + \frac{\alpha^{K+1}}{K+1}F(K,K+1;K+2;-a)
\end{aligned} \tag{24}$$

The first line applies the integral inequality, where we have included the first term to ensure that we have an upper bound. The last line substitutes $t = \frac{1}{r}$ into the integral and evaluates. Next, we apply the reflection law of hypergeometric functions. The reflection law states

$$\frac{1}{(1-z)^a}F\left(a,b;c;\frac{-z}{1-z}\right) = F(a,c-b;c;z).$$

[6]We thank Professor John Lewis (MIT) for his insights in deriving this bound.

Now we can simplify the hypergeometric function in Equation 24 by expanding it into its sum:

$$\begin{aligned}
\sum_{r=1}^{\infty} \frac{1}{r^2} \frac{\alpha^{K+1}}{(\alpha+r)^K} &\leq \frac{\alpha^{K+1}}{(\alpha+1)^K} + \frac{\alpha^{K+1}}{(K+1)(\alpha+1)^{K+1}} F(2, K+1; K+2; \frac{a}{a+1}) \\
&= \frac{\alpha^{K+1}}{(\alpha+1)^K} + \frac{\alpha^{K+1}}{(\alpha+1)^{K+1}} \sum_{j=0}^{\infty} \left(\frac{\alpha}{\alpha+1}\right)^j \frac{1+j}{K+1+j} \\
&\leq \frac{\alpha^{K+1}}{(\alpha+1)^K} + \frac{\alpha^{K+1}}{(\alpha+1)^{K+1}} \sum_{j=0}^{\infty} \left(\frac{\alpha}{\alpha+1}\right)^j \\
&= 2(\alpha+1)\left(\frac{\alpha}{\alpha+1}\right)^{K+1}
\end{aligned}$$

which we can plug into our original expression to get

$$1 - \exp\left(-N \sum_{r=1}^{\infty} \frac{1}{r^2} \frac{\alpha^{K+1}}{(\alpha+r)^K}\right) \leq 1 - \exp\left(-2N(\alpha+1)\left(\frac{\alpha}{\alpha+1}\right)^{K+1}\right).$$

We note this bound is very similar to the bound derived using the Levy-Khintchine approach from Section 6:

$$1 - \exp\left(-N\alpha\left(\frac{\alpha}{1+\alpha}\right)^K\right).$$