Latent Variable Modeling with Random Features

Gregory W. Gundersen* Princeton University ggundersen@princeton.edu

Michael Minyi Zhang* University of Hong Kong mzhang18@hku.hk

Barbara E. Engelhardt Princeton University bee@princeton.edu

Abstract

Gaussian process-based latent variable models are flexible and theoretically grounded tools for nonlinear dimension reduction, but generalizing to non-Gaussian data likelihoods within this nonlinear framework is statistically challenging. Here, we use random features to develop a family of nonlinear dimension reduction models that are easily extensible to non-Gaussian data likelihoods; we call these *random feature latent variable models* (RFLVMs). By approximating a nonlinear relationship between the latent space and the observations with a function that is linear with respect to random features, we induce closed-form gradients of the posterior distribution with respect to the latent variable. This allows the RFLVM framework to support computationally tractable nonlinear latent variable models for a variety of data likelihoods in the exponential family without specialized derivations. Our generalized RFLVMs produce results comparable with other state-of-the-art dimension reduction methods on diverse types of data, including neural spike train recordings, images, and text data.

1 INTRODUCTION

Many dimension reduction techniques, such as principal component analysis (Pearson, 1901; Tipping and Bishop, 1999) and factor analysis (Lawley and Maxwell, 1962), make two modeling assumptions: (1) the observations are Gaussian distributed, and (2) the latent structure is a linear function of the observations. However, for many applications, proper analysis requires us to break these assumptions. For example, in computational neuroscience, scientists collect firing rates for thousands of neurons simultaneously. These data are observed as counts, and neuroscientists believe that the biologically relevant latent structure is nonlinear with respect to the data (Cunningham and Byron, 2014).

To capture nonlinear relationships in latent variable models, one approach is to assume that the mapping between the latent manifold and observations is Gaussian process (GP)-distributed. A GP is a prior over the space of real-valued functions, and posterior inference is tractable when the GP prior is conjugate to the likelihood. This leads to the Gaussian process latent variable model (GPLVM, Lawrence, 2004).

The basic GPLVM model with a radial basis function (RBF) kernel has nice statistical properties that allow for exact, computationally tractable inference methods to be used when the number of observations is a reasonable size. Deviating from this basic model, however, leads to challenges with inference. In Poisson GPLVMs, for example, we cannot integrate out the GP-distributed functional map, and we no longer have closed form expressions for the gradient of the posterior with respect to the latent space. This renders maximum a posteriori (MAP) estimation difficult, leading to solutions at poor local optima (see Wu et al., 2017).

Random Fourier features (RFFs, Rahimi and Recht, 2008) were developed to avoid working with $N \times N$ dimensional matrices when fitting kernel machines. RFFs accelerate kernel machines by using a low-dimensional, randomized approximation of the inner product associated with a given shift-invariant kernel. For this approximation, RFFs induce a nonlinear map using a linear function of random features.

We propose to use RFFs to approximate the kernel function in a GPLVM to create a flexible, tractable, and modular framework for fitting GP-based latent variable models. In the context of GPLVMs, RFF approximations allow for closed-form gradients of the

* Denotes equal contribution.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

objective function with respect to the latent variable. Using RFFs solve a fundamental statistical problem with GPLVMs in non-Gaussian settings. With a Gaussian likelihood, we can obtain closed-form gradients by integrating out the GP-distributed maps, but this cannot be done in the non-Gaussian case. Our solution is to induce these closed-form gradients by making the data likelihood depend on the latent variables as a linear function of the random features. In addition, we can tractably explore the space of stationary covariance functions by using a Dirichlet process mixture prior for the spectral distribution of frequencies (BaNK, Oliva et al., 2016), leading to a flexible latent variable model.

This paper makes the following contributions to the space of nonlinear latent variable models: (1) we represent the nonlinear mapping in GPLVMs using a linear function of random Fourier features; (2) we leverage this representation to generalize GPLVMs to non-Gaussian likelihoods and derive a Markov chain Monte Carlo (MCMC) sampler for a wide variety of count-data likelihoods, such as the Poisson, binomial, negative binomial, and multinomial distributions; (3) we place a prior on the random features to allow data-driven exploration over the space of shift-invariant kernels, to avoid putting restrictions on the kernel's functional form. While implementing GPs with RFFs has been done before (Lázaro-Gredilla et al., 2007; Hensman et al., 2017; Cutajar et al., 2017), it has always been motivated by *scalability*. However, we motivate RFFs with statistical *tractability* of non-conjugate GPLVMs. Thus, our contributions focus on tractability and generality rather than scalability or state-of-the-art results for specialized models. We validate our approach on diverse simulated data sets, and show how results from RFLVMs compare with state-of-the-art methods on a variety of image, text, and scientific data sets. We release a Python library[1] with modular code for reproducing and building on our work.

2 RANDOM FEATURE LATENT VARIABLE MODELS

2.1 Random features for kernel machines

Here we briefly review random Fourier features (Rahimi and Recht, 2008) to motivate a randomized approximation of the GP-distributed maps in GPLVMs. Bochner's theorem (Bochner, 1959) states that any continuous shift-invariant kernel $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}')$ on $\mathbb{R}^D$ is positive definite if and only if $k(\mathbf{x} - \mathbf{x}')$ is the Fourier transform of a non-negative measure $p(\mathbf{w})$. If the kernel is properly scaled, the kernel's Fourier transform $p(\mathbf{w})$ is guaranteed to be a density. Let $h(\mathbf{x}) \triangleq \exp(i\mathbf{w}^\top \mathbf{x})$, and let $h(\mathbf{x})^*$ denote its complex conjugate. Observe that

$$\begin{aligned} k(\mathbf{x} - \mathbf{x}') &= \int_{\mathbb{R}^D} p(\mathbf{w}) \exp(i\mathbf{w}^\top(\mathbf{x} - \mathbf{x}'))\mathrm{d}\mathbf{w} \\ &= \mathbb{E}_{p(\mathbf{w})}[h(\mathbf{x})h(\mathbf{x}')^*]. \end{aligned} \tag{1}$$

So $h(\mathbf{x})h(\mathbf{x}')^*$ is an unbiased estimate of $k(\mathbf{x}-\mathbf{x}')$. If we drop the imaginary part for real-valued kernels, we can re-define $h(\mathbf{x}) \triangleq \cos(\mathbf{w}^\top \mathbf{x})$ by Euler's formula. Then we can use Monte Carlo integration to approximate Eq. 1 as $k(\mathbf{x}, \mathbf{x}') \approx \varphi_{\mathbf{W}}(\mathbf{x})^\top \varphi_{\mathbf{W}}(\mathbf{x}')$, where

$$\varphi_{\mathbf{W}}(\mathbf{x}) \triangleq \sqrt{\frac{2}{M}} \begin{bmatrix} \sin(\mathbf{w}_1^\top \mathbf{x}) \\ \cos(\mathbf{w}_1^\top \mathbf{x}) \\ \vdots \\ \sin(\mathbf{w}_{M/2}^\top \mathbf{x}) \\ \cos(\mathbf{w}_{M/2}^\top \mathbf{x}) \end{bmatrix}, \quad \mathbf{w}_m \overset{\text{iid}}{\sim} p(\mathbf{w}). \tag{2}$$

We draw $M/2$ samples from $p(\mathbf{w})$, and the definition in Eq. 2 doubles the number of RFFs to M. A representer theorem (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001) says that the optimal solution to the objective function of a kernel method, $f^*(\mathbf{x})$, is linear in pairwise evaluations of the kernel. Using this random projection, we can represent $f^*(\mathbf{x})$ as

$$\begin{aligned} f^*(\mathbf{x}) &= \sum_{n=1}^{N} \alpha_n k(\mathbf{x}_n, \mathbf{x}) = \sum_{n=1}^{N} \alpha_n \langle \phi(\mathbf{x}_n), \phi(\mathbf{x}) \rangle_{\mathcal{H}} \\ &\approx \sum_{n=1}^{N} \alpha_n \varphi_{\mathbf{W}}(\mathbf{x}_n)^\top \varphi_{\mathbf{W}}(\mathbf{x}) = \boldsymbol{\beta}^\top \varphi_{\mathbf{W}}(\mathbf{x}). \end{aligned} \tag{3}$$

In the second equality, the kernel trick implicitly lifts the data into a reproducing kernel Hilbert space $\mathcal{H}$ in which the optimal solution is linear with respect to the features. The randomized approximation of this inner product lets us replace expensive calculations involving the kernel with an M-dimensional inner product.

For example, the predictive mean in GP regression implicitly uses the representer theorem and kernel trick (Williams and Rasmussen, 2006). RFFs have been used to reduce the computational costs of fitting GP regression models from $\mathcal{O}(N^3)$ to $\mathcal{O}(NM^2)$ (Lázaro-Gredilla et al., 2010; Hensman et al., 2017). However, RFFs have not yet been used to make GPLVMs more computationally tractable.

2.2 Gaussian process latent variable models

Now we introduce the basic GPLVM framework (Lawrence, 2004). Let $\mathbf{Y}$ be an $N \times J$ matrix of N observations and J features, and let $\mathbf{X}$ be an $N \times D$ matrix of latent variables where $D \ll J$. If we take the mean function to be zero, and the observations $\mathbf{Y}$ to

[1] https://github.com/gwgundersen/rflvm

be Gaussian distributed, the GPLVM is:

$$\begin{aligned} \mathbf{y}_j &\sim \mathcal{N}_N(f_j(\mathbf{X}), \sigma_j^2\mathbf{I}), \\ f_j &\sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_X), \\ \mathbf{x}_n &\sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \end{aligned} \tag{4}$$

where $\mathbf{K}_X$ is an $N \times N$ covariance matrix defined by a positive definite kernel function $k(\mathbf{x}, \mathbf{x}')$ and $f_j(\mathbf{X}) = [f_j(\mathbf{x}_1) \dots f_j(\mathbf{x}_N)]^\top$. Due to conjugacy between the GP prior on f_j and Gaussian likelihood on $\mathbf{y}_j$, we can integrate out f_j in closed form. The resulting marginal likelihood for $\mathbf{y}_j$ is $\mathcal{N}_N(0, \mathbf{K}_X + \sigma_j^2\mathbf{I})$. We cannot find the optimal $\mathbf{X}$ analytically, but various approximations have been proposed. We can obtain a MAP estimate by integrating out the GP-distributed maps and then optimizing $\mathbf{X}$ with respect to the posterior using scaled conjugate gradients (Lawrence, 2004, 2005), where computation scales as $\mathcal{O}(N^3)$. To scale inference, we may use sparse inducing point methods where the computational complexity is $\mathcal{O}(NC^2)$, for $C \ll N$ inducing points (Lawrence, 2007).

Alternatively, we can introduce a variational Bayes approximation of the posterior and minimize the Kullback–Leibler divergence between the posterior and the variational approximation with the latent variables $\mathbf{X}$ marginalized out. However, integrating out $\mathbf{X}$ in the approximate marginal likelihood is only tractable when we assume that we have Gaussian observations and when we use an RBF kernel with automatic relevance determination, which limits its flexibility. This variational approach, called a *Bayesian GPLVM* (Titsias and Lawrence, 2010; Damianou et al., 2016), may be scaled using sparse inducing point methods.

2.3 Generative model for RFLVMs

The generative model of an RFLVM takes the form:

$$\begin{aligned} \mathbf{y}_j &\sim \mathcal{L}\big(g\big(\varphi_{\mathbf{W}}(\mathbf{X})\boldsymbol{\beta}_j\big), \boldsymbol{\theta}\big), & \boldsymbol{\theta} &\sim p(\boldsymbol{\theta}), \\ \boldsymbol{\beta}_j &\sim \mathcal{N}_M(\boldsymbol{\beta}_0, \mathbf{B}_0), & \mathbf{x}_n &\sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \\ \mathbf{w}_m &\sim \mathcal{N}_D(\boldsymbol{\mu}_{z_m}, \boldsymbol{\Sigma}_{z_m}), & z_m &\sim \text{CRP}(\alpha), \\ \alpha &\sim \text{Ga}(a_\alpha, b_\alpha), \\ (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &\sim \text{NIW}(\boldsymbol{\mu}_0, \nu_0, \lambda_0, \boldsymbol{\Psi}_0). \end{aligned} \tag{5}$$

Following exponential family notation, $\mathcal{L}(\cdot)$ is a likelihood function, $g(\cdot)$ is an invertible link function that maps the real numbers onto the likelihood parameters' support, and $\boldsymbol{\theta}$ are other likelihood-specific parameters. Following Wilson and Adams (2013) and Oliva et al. (2016), we assume $p(\mathbf{w})$ is a Dirichlet process mixture of Gaussians (DP-GMM, Ferguson, 1973; Antoniak, 1974). By sampling from the posterior of $\mathbf{w}$, we can explore the space of stationary kernels and estimate the kernel hyperparameters in a Bayesian way. We assign each $\mathbf{w}_m$ in $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_{M/2}]^\top$ to a mixture component with the variable z_m, which is distributed according to a Chinese restaurant process (CRP, Aldous, 1985) with concentration parameter α. This prior introduces additional random variables: the mixture means $\{\boldsymbol{\mu}_k\}_{k=1}^K$, and the mixture covariance matrices $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$ where K is the number of clusters in the current Gibbs sampling iteration.

The randomized map in Eq. 2 allows us to approximate the original GPLVM in Eq. 4 as

$$\begin{aligned} \mathbf{y}_j &\sim \mathcal{N}_N(\varphi_{\mathbf{W}}(\mathbf{X})\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I}), \\ \boldsymbol{\beta}_j &\sim \mathcal{N}_M(\mathbf{b}_0, \mathbf{B}_0), \\ \mathbf{x}_n &\sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}). \end{aligned} \tag{6}$$

We approximate $f_j(\mathbf{X})$ in Eq. 4 as $\varphi_{\mathbf{W}}(\mathbf{X})\boldsymbol{\beta}_j$, where $\varphi_{\mathbf{W}}(\mathbf{X}) = [\varphi_{\mathbf{W}}(\mathbf{x}_1) \dots \varphi_{\mathbf{W}}(\mathbf{x}_N)]^\top$. This is a Gaussian RFLVM when $\mathcal{L}(\cdot)$ is a Gaussian distribution and $g(\cdot)$ is the identity function. Because the prior distribution on the mapping weights $\boldsymbol{\beta}_j$ is Gaussian, the model is analogous to Bayesian linear regression given $\varphi_{\mathbf{W}}(\mathbf{X})$; if we integrate out $\boldsymbol{\beta}_j$, we recover a marginal likelihood that approximates the GPLVM's marginal likelihood.

We use this representation to generalize the RFLVM to other observation types in the exponential family. For example, a Poisson RFLVM takes the following form:

$$\begin{aligned} \mathbf{y}_j &\sim \text{Poisson}(\exp(\varphi_{\mathbf{W}}(\mathbf{X})\boldsymbol{\beta}_j)), \\ \boldsymbol{\beta}_j &\sim \mathcal{N}_M(\mathbf{b}_0, \mathbf{B}_0), \\ \mathbf{x}_n &\sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}). \end{aligned} \tag{7}$$

For distributions including the Bernoulli, binomial, and negative binomial, the functional form of the data likelihood is

$$\begin{aligned} &\mathcal{L}(\varphi_{\mathbf{W}}(\mathbf{X}), \boldsymbol{\beta}_j, a(\mathbf{y}_j), b(\mathbf{y}_j), c(\mathbf{y}_j)) \\ &= \prod_{n=1}^N c(y_{nj}) \frac{(\exp(\varphi_{\mathbf{W}}(\mathbf{x}_n)\boldsymbol{\beta}_j))^{a(y_{nj})}}{(1 + \exp(\varphi_{\mathbf{W}}(\mathbf{x}_n)\boldsymbol{\beta}_j))^{b(y_{nj})}}, \end{aligned} \tag{8}$$

for some functions of the data $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$. The general form of this logistic RFLVM is then:

$$\begin{aligned} \mathbf{y}_j &\sim \mathcal{L}(\varphi_{\mathbf{W}}(\mathbf{X}), \boldsymbol{\beta}_j, a(\mathbf{y}_j), b(\mathbf{y}_j), c(\mathbf{y}_j)), \\ \boldsymbol{\beta}_j &\sim \mathcal{N}_M(\mathbf{b}_0, \mathbf{B}_0), \\ \mathbf{x}_n &\sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}). \end{aligned} \tag{9}$$

For example, by setting $a(y_{nj}) = y_{nj}$, $b(y_{nj}) = y_{nj} + r_j$, and $c(y_{nj}) = \binom{y_{nj} + r_j - 1}{y_{nj}}$, we get the negative binomial RFLVM with feature-specific dispersion parameter r_j.

2.4 Inference for RFLVMs

We now present a general Gibbs sampling framework for all RFLVMs. A consequence of the linearization induced by the random features is that we can use all

available techniques for Gibbs sampling in linear models that are otherwise not possible for GPLVMs; and unlike other sampling methods such as HMC, Gibbs samplers do not require tuning parameters.

First, we write the Gibbs sampling steps to estimate the posterior of the covariance kernel. Next, we describe estimating the latent variable $\mathbf{X}$ by taking the MAP estimate. Then, we sample the data likelihood-specific parameters $\boldsymbol{\theta}$ and linear coefficients $\boldsymbol{\beta}_j$. Variables subscripted with zero, e.g., $\boldsymbol{\theta}_0$, denote hyperparameters. While the number of mixture components may change across sampling iterations, let K denote the number of components in the current Gibbs sampling step. We initialize all the parameters in our model by drawing from the prior, except for $\mathbf{X}$, which we initialize with PCA.

First, we sample z_m following Algorithm 8 from Neal (2000). We choose to use a sampling method that integrates out the Dirichlet process-distributed mixture weights because such samplers can better propose new features (Dubey et al., 2020) and therefore is more effective at exploring the posterior behavior of the covariance kernel. Let $n_k = \sum_\ell \delta(z_\ell = k)$, and let n_k^{-m} denote the same sum with z_m excluded. Then we sample the posterior of z_m from the following discrete distribution for $k = 1, 2, \ldots, K$:

$$p(z_m = k \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{W}, \alpha) = \begin{cases} \frac{n_k^{-m}}{M-1+\alpha}\mathcal{N}(\mathbf{w}_m \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) & n_k^{-m} > 0 \\ \frac{\alpha}{M-1+\alpha}\int \mathcal{N}(\mathbf{w}_m \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\mathrm{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\Sigma} & n_k^{-m} = 0. \end{cases} \tag{10}$$

Given assignments $\mathbf{z} = [z_1 \ldots z_{M/2}]^\top$ and RFFs $\mathbf{W}$, the posterior of $\boldsymbol{\Sigma}_k$ is inverse-Wishart distributed. Given $\boldsymbol{\Sigma}_k$, the posterior of $\boldsymbol{\mu}_k$ is normally distributed (Gelman et al., 2013):

$$\begin{aligned} \boldsymbol{\Sigma}_k &\sim \mathcal{W}^{-1}(\boldsymbol{\Psi}_k, \nu_k), \quad \boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{m}_k, \frac{1}{\lambda_k}\boldsymbol{\Sigma}_k). \\ \boldsymbol{\Psi}_k &= \boldsymbol{\Psi}_0 + \sum_{m:z_m=k}^{M/2} (\mathbf{w}_m - \bar{\mathbf{w}}^{(k)})(\mathbf{w}_m - \bar{\mathbf{w}}^{(k)})^\top \\ &\quad + \frac{\lambda_0 n_k}{\lambda_0 + n_k}(\mathbf{w}_m - \boldsymbol{\mu}_0)(\mathbf{w}_m - \boldsymbol{\mu}_0)^\top \\ \bar{\mathbf{w}}^{(k)} &= \frac{1}{n_k}\sum_{m:z_m=k}^{M} \mathbf{w}_m, \quad \nu_k = \nu_0 + n_k, \\ \mathbf{m}_k &= \frac{\lambda_0\boldsymbol{\mu}_0 + n_k\bar{\mathbf{w}}^k}{\lambda_0 + n_k}, \quad \lambda_k = \lambda_0 + n_k. \end{aligned} \tag{11}$$

We cannot sample from the full conditional distribution of $\mathbf{W}$, but prior work suggested a Metropolis–Hastings (MH) sampler using proposal distribution $q(\mathbf{W})$ set to the prior $p(\mathbf{W} \mid \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}_D(\boldsymbol{\mu}_{z_m}, \boldsymbol{\Sigma}_{z_m})$ (Eq. 5) and acceptance ratio $\rho_{\texttt{MH}}$ (Oliva et al., 2016):

$$\begin{aligned} \mathbf{w}_m^\star &\sim q(\mathbf{W}) \triangleq p(\mathbf{W} \mid \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \rho_{\texttt{MH}} &= \min\left\{1, \frac{p(\mathbf{Y} \mid \mathbf{X}, \mathbf{w}_m^\star, \boldsymbol{\theta})}{p(\mathbf{Y} \mid \mathbf{X}, \mathbf{w}_m, \boldsymbol{\theta})}\right\}. \end{aligned} \tag{12}$$

Finally, we sample the DP-GMM concentration parameter α (Escobar and West, 1995). We augment the model with variable η to make sampling α conditionally conjugate:

$$\begin{aligned} \eta &\sim \mathrm{Beta}(\alpha + 1, M), \\ \frac{\pi_\eta}{1-\pi_\eta} &= \frac{a_\alpha + K - 1}{M(b_\alpha - \log(\eta))}, \quad K = |\{k : n_k > 0\}|, \\ \alpha &\sim \pi_\eta \mathrm{Ga}(a_\alpha + K, b_\alpha - \log(h)) \\ &\quad + (1 - \pi_\eta)\mathrm{Ga}(a_\alpha + K - 1, b_\alpha - \log(\eta)). \end{aligned} \tag{13}$$

For the Gaussian RFLVM (Eq. 6), let $\mathbf{B}_0 = \sigma^{-2}\mathbf{S}_0$. We integrate out $\boldsymbol{\beta}_j$ and σ^{-2} in closed form to obtain a marginal likelihood,

$$p(\mathbf{y}_j \mid \mathbf{X}, \mathbf{W}) = \frac{1}{(2\pi)^{N/2}} \cdot \sqrt{\frac{|\mathbf{S}_0|}{|\mathbf{S}_N|}} \cdot \frac{b_0^{a_0}}{b_N^{a_N}} \cdot \frac{\Gamma(a_N)}{\Gamma(a_0)}, \tag{14}$$

where $\mathbf{S}_N = \varphi_\mathbf{W}(\mathbf{X})^\top\varphi_\mathbf{W}(\mathbf{X}) + \mathbf{S}_0$, $\boldsymbol{\beta}_N = \mathbf{S}_N^{-1}(\boldsymbol{\beta}_0^\top\mathbf{S}_0 + \varphi_\mathbf{W}(\mathbf{X})^\top\mathbf{y}_j)$, $a_N = a_0 + N/2$, and $b_N = b_0 + (1/2)(\mathbf{y}_j^\top\mathbf{y}_j + \boldsymbol{\beta}_0^\top\mathbf{S}_0\boldsymbol{\beta}_0 - \boldsymbol{\beta}_N^\top\mathbf{S}_N\boldsymbol{\beta}_N)$. See Appendix A or Minka (2000) for details. However, inference can be slow because marginalizing out $\boldsymbol{\beta}_j$ introduces dependencies between the latent variables, and the complexity becomes $\mathcal{O}(NM^2)$. Alternatively, we can Gibbs sample $\boldsymbol{\beta}_j$ and take the MAP estimate of $\mathbf{X}$ using the original log likelihood where the complexity is $\mathcal{O}(NM)$.

In the Poisson RFLVM (Eq. 7), we no longer have the option of marginalizing out $\boldsymbol{\beta}_j$. Instead, we take iterative MAP estimates of $\boldsymbol{\beta}_j$ and $\mathbf{X}$. Given $\varphi_\mathbf{W}(\mathbf{X})$, inference for $\boldsymbol{\beta}_j$ is analogous to Bayesian inference for a Poisson generalized linear model (GLM). In Secs. 3.1 and 3.2, we show that, by inducing closed-form gradients with respect to $\mathbf{X}$ through RFFs, this iterative MAP procedure produces results that are competitive with benchmarks on count data. For logistic RFLVMs (Eq. 9), we use Pólya-gamma augmentation (Polson et al., 2013) to make inference tractable. A random variable ω is Pólya-gamma distributed with parameters $b > 0$ and $c \in \mathbb{R}$, denoted $\omega \sim \mathrm{PG}(b, c)$, if

$$\omega \overset{d}{=} \frac{1}{2\pi^2}\sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}, \tag{15}$$

where $\overset{d}{=}$ denotes equality in distribution and $g_k \sim \mathrm{Ga}(b, 1)$ are independent gamma random variables.

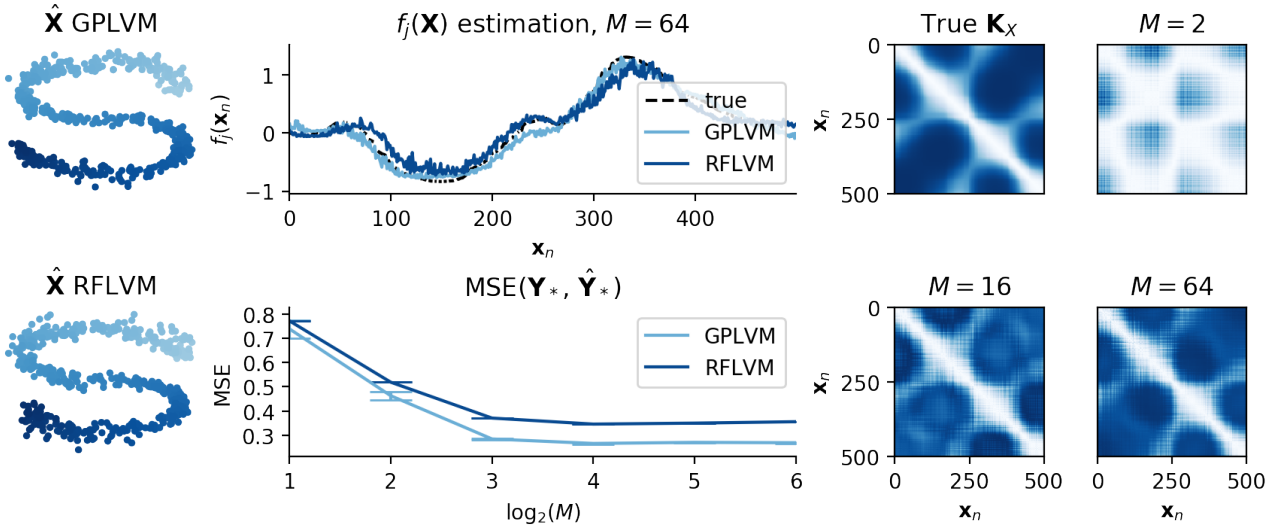

Figure 1: **Simulated data with Gaussian emissions.** (Left) Inferred latent variables for both a GPLVM and Gaussian RFLVM. (Upper middle) Comparison of estimated $f_j(\mathbf{X})$ for a single feature as estimated by GPLVM and RFLVM. (Lower middle) Comparison of MSE reconstruction error on held out $\mathbf{Y}_*$ for increasing M, where M is the number of inducing points for GPLVM and random Fourier features for RFLVM. (Right) Ground truth covariance matrix $\mathbf{K}_X$ compared with the RFLVM estimation for increasing M.

The identity critical for Pólya-gamma augmentation is

$$\frac{(e^{\psi_{nj}})^{a_{nj}}}{(1+e^{\psi_{nj}})^{b_{nj}}} = 2^{-b_{nj}} e^{\kappa_{nj}\psi_{nj}} \int_0^\infty e^{-\omega\psi_{nj}^2/2} p(\omega)\mathrm{d}\omega, \tag{16}$$

where $\kappa_{nj} = a_{nj} - b_{nj}/2$ and $p(\omega) = \mathrm{PG}(\omega \mid b_{nj}, 0)$. If we define $\psi_{nj} = \varphi_{\mathbf{W}}(\mathbf{x}_n)^\top \boldsymbol{\beta}_j$, then Eq. 16 allows us to rewrite the likelihood in Eq. 8 as proportional to a Gaussian. Furthermore, we can sample ω conditioned on ψ_{nj} as $p(\omega \mid \psi_{nj}) \sim \mathrm{PG}(b_{nj}, \psi_{nj})$. This enables convenient, closed-form Gibbs sampling steps of $\boldsymbol{\beta}_j$, conditioned on Pólya-gamma augmentation variables ω_{nj}:

$$\begin{aligned}
\omega_{nj} \mid \boldsymbol{\beta}_j &\sim \mathrm{PG}(b_{nj}, \varphi_{\mathbf{W}}(\mathbf{x}_n)^\top \boldsymbol{\beta}_j), \\
\boldsymbol{\beta}_j \mid \boldsymbol{\Omega}_j &\sim \mathcal{N}(\mathbf{m}_{\boldsymbol{\omega}_j}, \mathbf{V}_{\boldsymbol{\omega}_j}), \\
\mathbf{V}_{\boldsymbol{\omega}_j} &= (\varphi_{\mathbf{W}}(\mathbf{X})^\top \boldsymbol{\Omega}_j \varphi_{\mathbf{W}}(\mathbf{X}) + \mathbf{B}_0^{-1})^{-1}, \\
\mathbf{m}_{\boldsymbol{\omega}_j} &= \mathbf{V}_{\boldsymbol{\omega}_j}(\varphi_{\mathbf{W}}(\mathbf{X})^\top \boldsymbol{\kappa}_j + \mathbf{B}_0^{-1}\boldsymbol{\beta}_0),
\end{aligned} \tag{17}$$

where $\boldsymbol{\Omega}_j = \mathrm{diag}([\omega_{1j} \dots \omega_{Nj}])$ and $\boldsymbol{\kappa}_j = [\kappa_{1j} \dots \kappa_{Nj}]^\top$. This technique has been used to derive Gibbs samplers for binomial regression (Polson et al., 2013), negative binomial regression (Zhou et al., 2012), and correlated topic models (Chen et al., 2013; Linderman et al., 2015). Here, we use it to derive samplers for logistic RFLVMs.

RFLVMs are identifiable up to the rotation and scale of $\mathbf{X}$. As a result, MAP estimates of $\mathbf{X}$ between iterations are unaligned as they can be arbitrarily rescaled and rotated through inference. Thus, a point estimate of $\mathbf{X}$ that is a function of the Monte Carlo samples of $\mathbf{X}$, e.g., the expectation of $\mathbf{X}$ across the samples, will not be meaningful. To this end, we arbitrarily fix the rotation of $\mathbf{X}$ by taking the singular value decomposition (SVD) of the MAP estimate, $\hat{\mathbf{X}} = \mathbf{USV}^T$, and setting $\mathbf{X}$ to be the left singular vectors corresponding to the D largest singular values, $\mathbf{X} \triangleq [\mathbf{u}_1, \dots, \mathbf{u}_D]$ where $\mathrm{diag}(\mathbf{S}) = [s_1 \dots s_D]$ and $s_1 \geq s_2 \geq \dots \geq s_D$. Then, we rescale $\mathbf{X}$ so that the covariance of the latent space is the identity matrix. This has the effect of enforcing orthogonality, and does not allow heteroskedasticity in the latent dimensions. This operation is analogous to the covariance adjustment in parameter-expanded expectation–maximization (Liu et al., 1998) and has been used to fix the rotation in Bayesian factor analysis models (Ročková and George, 2016).

3 EXPERIMENTS

In our results, we refer to the Gaussian-distributed GPLVM using inducing point methods for inference as *GPLVM* (Titsias and Lawrence, 2010). We fit all GPLVM experiments using the `GPy` package (GPy, 2012). We refer to the Poisson-distributed GPLVM using a double Laplace approximation as *DLA-GPLVM* (Wu et al., 2017). DLA-GPLVM is designed to model multi-neuron spike train data, and the code[2] initializes the latent space using the output of a Poisson linear dynamical system (Macke et al., 2011), and places a GP prior on $\mathbf{X}$. To make the experiments

[2] https://github.com/waq1129/LMT

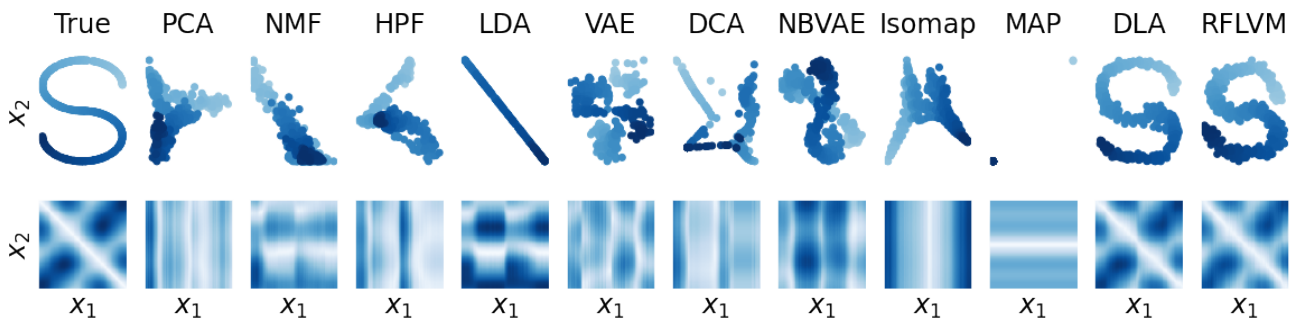

Figure 2: **Simulated data with Poisson emissions.** (Top) True latent variable $\mathbf{X}$ compared with inferred latent variables $\hat{\mathbf{X}}$ from benchmarks (see text for abbreviations) and a Poisson RFLVM. (Bottom) Distance matrices between true $\mathbf{X}$ and $\hat{\mathbf{X}}$ from the above benchmark (darker is farther away).

comparable for all GPLVM experiments, we initialize DLA-GPLVM with PCA and assume $\mathbf{x}_n \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I})$. We refer to our GPLVM with random Fourier features as *RFLVM* and explicitly state the assumed distribution. In Sec. 3.1, we use a Gaussian RFLVM with the linear coefficients $\{\boldsymbol{\beta}_j\}_{j=1}^J$ marginalized out (Eq. 14) for a fairer comparison with the GPLVM.

Since hyperparameter tuning our model on each dataset would be both time-consuming and unfair without also tuning the baselines, we fixed the hyperparameters across experiments. We used 2000 Gibbs sampling iterations with 1000 burn-in steps, $M = 100$, and $D = 2$. We initialized $K = 20$ and $\alpha = 1$. In Sec. 3.2, we used $D = 3$ and visualized $\hat{\mathbf{X}}$ after the best affine transformation onto 2-D rat positions following Wu et al. (2017). For computational reasons, MNIST and CIFAR-10 were subsampled (see Appendix D for details).

3.1 Simulated data

We first evaluate RFLVM on simulated data. We set $\mathbf{X}$ to be a 2-D S-shaped manifold, sampled functions $\mathbf{F} = \{f_j(\mathbf{X})\}_{j=1}^J$ from a Gaussian process with an RBF kernel, and then generated observations for Gaussian emissions (Eq. 4) and Poisson emissions (Eq. 7). For all simulations, we used $N = 500, J = 100$, and $D = 2$.

For these experiments, we computed the mean-squared error (MSE) between test set observations, $\mathbf{Y}_*$, and predicted observations $\hat{\mathbf{Y}}_*$, where we held out 20% of the observations for the test set. To evaluate our latent space results, we projected the estimated latent space, $\hat{\mathbf{X}}$, onto the hyperplane that minimizes the squared error with the ground truth, $\|\mathbf{X} - \hat{\mathbf{X}}\mathbf{A}\|_2^2$, and calculated the R^2 value between the true $\mathbf{X}$ and the projected latent space $\hat{\mathbf{X}}\mathbf{A}$. We evaluated our model's ability to estimate the GP outputs $f_j(\mathbf{X}) \approx \varphi_{\mathbf{W}}(\mathbf{X})\boldsymbol{\beta}_j$ by comparing the MSE between the estimated $\varphi_{\mathbf{W}}(\hat{\mathbf{X}})\boldsymbol{\beta}_j$ and the true generating $f_j(\mathbf{X})$. We computed the mean and standard deviation of the MSE and R^2 results by running each experiment five times.

We compared the performance of a Gaussian RFLVM to the GPLVM. We ran these experiments across multiple values of M, where M denotes the number of random features for the RFLVM and the number of inducing points for the GPLVM. Both models recovered the true latent variable $\mathbf{X}$ accurately and estimated the nonlinear maps, $\mathbf{F}$, well (Fig. 1, upper middle). Empirically, a GPLVM shows better performance for estimating $\mathbf{Y}_*$ than the RFLVM (Fig. 1, lower middle). We hypothesize that this is because Nyström's method has better generalization error bounds than RFFs when there is a large gap in the eigenspectrum (Yang et al., 2012), which is the case for $\mathbf{K}_X$. However, we see that the RFLVM approximates the true $\mathbf{K}_X$ given enough random features (Fig. 1, right), though perhaps less accurately than the GPLVM (Fig. 1, lower middle).

To demonstrate the utility of our model beyond Gaussian-distributed data, we compared results for simulated count data from a Poisson RFLVM with the following benchmarks: PCA, nonnegative matrix factorization (NMF, Lee and Seung, 1999), hierarchical Poisson factorization (HPF, Gopalan et al., 2015), latent Dirichlet allocation (LDA, Blei et al., 2003), variational autoencoder (VAE, Kingma and Welling, 2013), deep count autoencoder (DCA, Eraslan et al., 2019), negative binomial VAE (NBVAE, Zhao et al., 2020), and Isomap (Balasubramanian et al., 2002). Additionally, we compared results to our own naive implementation of the Poisson GPLVM that performs coordinate ascent on $\mathbf{X}$ and $\mathbf{F}$ by iteratively taking MAP estimates without using RFFs. We refer to this method as *MAP-GPLVM*. We found that the Poisson RFLVM infers a latent variable that is more similar to the true latent structure than other methods (Fig. 2). Linear methods such as PCA and NMF lack the flexibility to capture this nonlinear space, while nonlinear but Gaussian methods such as Isomap and VAEs recover smooth latent spaces that lack the original structure.

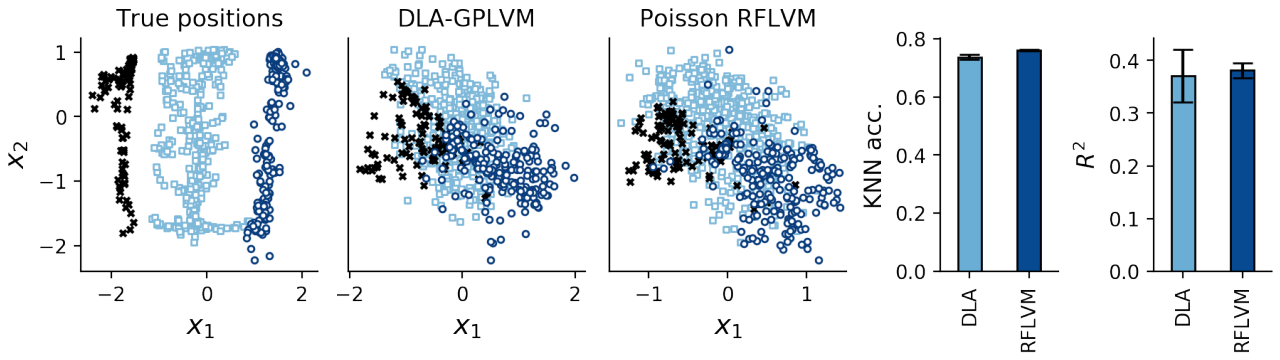

Figure 3: **Hippocampal place cells.** (Left three plots) Inferred latent space for the DLA-GPLVM and the Poisson RFLVM. The points are colored by three major regions of the true rat position in a W-shaped maze. (Right two plots) KNN accuracy using 5-fold cross validation and R^2 performance of the best affine transformation from $\hat{\mathbf{X}}$ onto the rat positions $\mathbf{X}$. Error bars computed using five trials.

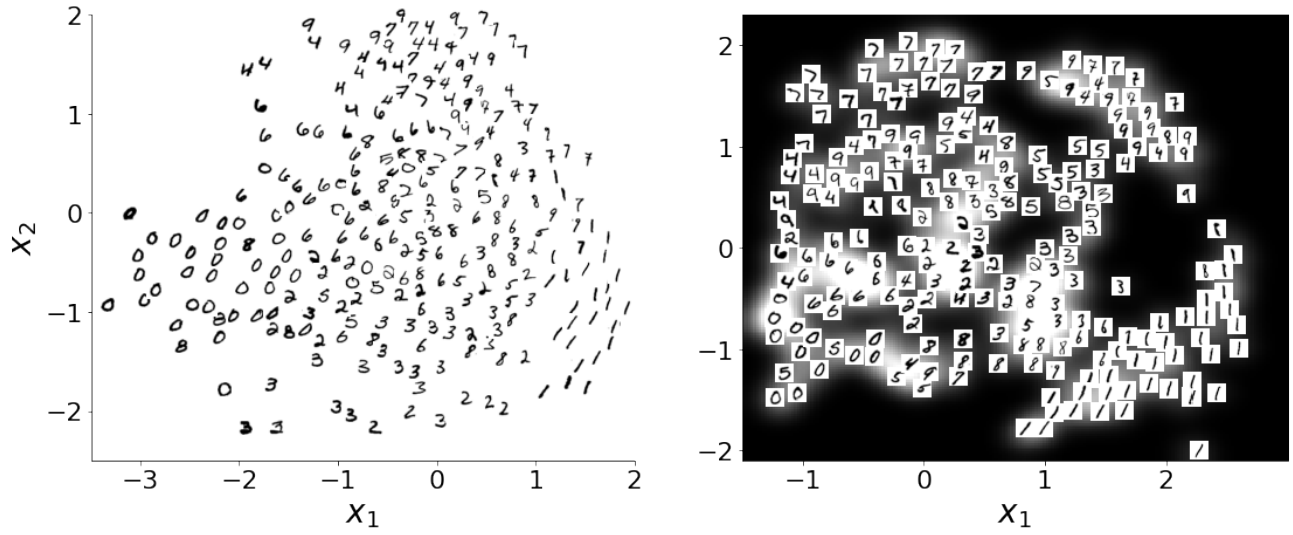

Figure 4: **MNIST digits.** Digits visualized in 2-D latent space inferred from DLA-GPLVM (left) and Poisson RFLVM (right). Following Lawrence (2004), we plotted images in a random order while not plotting any images that result in an overlap. The RFLVM's latent space is visualized as a histogram of 1000 draws after burn-in. The plotted points are the sample posterior mean.

The MAP-GPLVM appears to get stuck in poor local modes (see Wu et al., 2017) because we do not have gradients of the posterior in closed form. Both DLA-GPLVM and RFLVM, however, do have closed-form gradients and approximate the true manifold with similar R^2 and MSE values for $\hat{\mathbf{X}}$ and $\hat{f}_j(\mathbf{X})$ (not shown).

3.2 Hippocampal place cell data

Next, we checked whether a non-Gaussian RFLVM recovers an interpretable latent space when applied to a scientific problem. In particular, we use an RFLVM to model hippocampal place cell data (Wu et al., 2017). Place cells, a type of neuron, are activated when an animal enters a particular place in its environment. Here, $\mathbf{Y}$ is an $N \times J$ matrix of count-valued spikes where n indexes time and j indexes neurons. These data were jointly recorded while measuring the position of a rat in a W-shaped maze. We are interested in reconstructing the latent positions of the rat with $\mathbf{X}$.

We quantified goodness-of-fit of the latent space by assessing how well the RFLVM captures known structure, in the form of held-out sample labels, in the low-dimensional space. After estimating $\hat{\mathbf{X}}$, we performed K-nearest neighbors (KNN) classification on $\hat{\mathbf{X}}$ with $K = 1$. We ran this classification five times using 5-fold cross validation. We report the mean and standard deviation of KNN accuracy across five experiments.

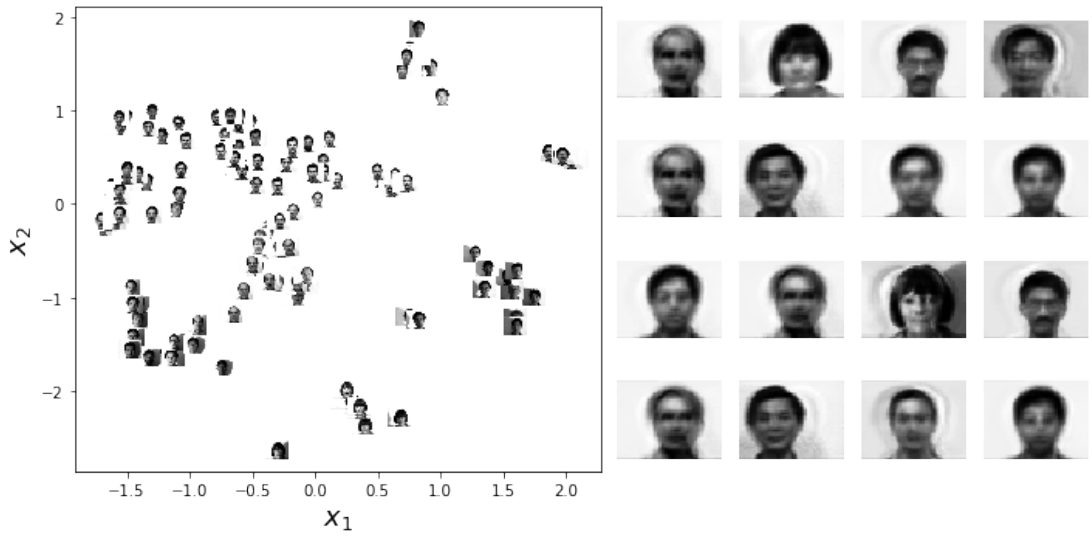

Figure 5: **Yale face data set.** Face data visualized in 2-D latent space using a Poisson RFLVM (left). Synthetic faces for the Yale dataset sampled from the posterior data generating process using a Poisson RFLVM (right).

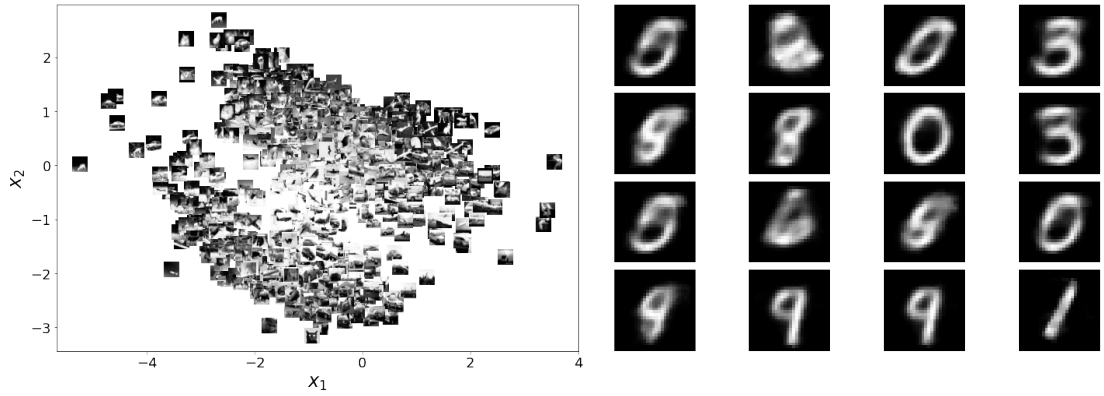

Figure 6: **CIFAR-10 and MNIST images.** CIFAR-10 image data set visualized in 2-D latent space using a Poisson RFLVM (left). Synthetic digits for MNIST sampled from the posterior data generating process using a Poisson RFLVM (right).

The Poisson RFLVM and DLA-GPLVM have similar performance in terms of how well they cluster samples in the latent space as measured by KNN accuracy using regions of the maze as labels. Furthermore, the models have similar performance in recovering the true rat positions $\mathbf{X}$, measured by R^2 performance (Fig. 3). While this clustering would not be impressive for many benchmark datasets such as MNIST, dimension reduction for large-scale neural recordings is an open problem (Cunningham and Byron, 2014; Linderman et al., 2016; Wu et al., 2017). These results suggest that our generalized RFLVM framework finds structure even in empirical, complex, non-Gaussian data and is competitive with models built for this specific task.

3.3 Text and image data

Finally, we examine whether an RFLVM captures the latent space of text, image, and empirical data sets. We hold out the labels and use them to evaluate the estimated latent space using the same KNN evaluation from Sec. 3.2. Across all eight data sets, the Poisson and negative binomial RFLVMs infer a low-dimensional latent variable $\hat{\mathbf{X}}$ that generally captures the latent structure as well as or better than linear methods like PCA and NMF (Lee and Seung, 1999). Moreover, adding nonlinearity but retaining a Gaussian data likelihood—as with real-valued models like Isomap (Tenenbaum et al., 2000), a variational autoencoder (VAE, Kingma and Welling, 2013), and the Gaussian RFLVM, or even using the Poisson-likelihood DLA-GPLVM—perform worse than the Poisson and

Table 1: Classification accuracy evaluated by fitting a KNN classifier ($K = 1$) with five-fold cross validation. Mean accuracy and standard deviation were computed by running each experiment five times.

	PCA	NMF	HPF	LDA	VAE	DCA
Bridges	0.8469 ± 0.0067	$\mathbf{0.8664 \pm 0.0164}$	0.7860 ± 0.0328	0.6747 ± 0.0412	0.8141 ± 0.0301	0.7093 ± 0.0317
CIFAR-10	0.2651 ± 0.0019	0.2450 ± 0.0028	0.2516 ± 0.0074	0.2248 ± 0.0040	0.2711 ± 0.0083	0.2538 ± 0.0178
Congress	0.5558 ± 0.0098	0.5263 ± 0.0108	0.6941 ± 0.0537	0.7354 ± 0.1018	0.6563 ± 0.0314	0.5917 ± 0.0674
MNIST	0.3794 ± 0.0146	0.2764 ± 0.0197	0.3382 ± 0.0370	0.2176 ± 0.0387	$\mathbf{0.6512 \pm 0.0228}$	0.1620 ± 0.0976
Montreal	0.6802 ± 0.0099	0.6878 ± 0.0207	0.6144 ± 0.1662	0.6238 ± 0.0271	0.6702 ± 0.0325	0.6601 ± 0.0997
Newsgroups	0.3896 ± 0.0043	0.3892 ± 0.0042	0.3921 ± 0.0122	0.3261 ± 0.0193	0.3926 ± 0.0113	0.4000 ± 0.0153
Spam	0.8454 ± 0.0037	0.8237 ± 0.0040	0.8719 ± 0.0353	0.8699 ± 0.0236	0.9028 ± 0.0128	0.8920 ± 0.0414
Yale	0.5442 ± 0.0129	0.4739 ± 0.0135	0.5200 ± 0.0071	0.3261 ± 0.0193	0.6327 ± 0.0209	0.2861 ± 0.0659
	NBVAE	Isomap	DLA-GPLVM	Poisson RFLVM	Neg. binom. RFLVM	Multinomial RFLVM
Bridges	0.7485 ± 0.0613	0.8375 ± 0.0240	0.8578 ± 0.0101	0.8440 ± 0.0165	$\mathbf{0.8664 \pm 0.0191}$	0.7984 ± 0.0102
CIFAR-10	0.2671 ± 0.0048	0.2716 ± 0.0056	0.2641 ± 0.0063	$\mathbf{0.2789 \pm 0.0080}$	0.2656 ± 0.0048	0.2652 ± 0.0024
Congress	$\mathbf{0.8541 \pm 0.0074}$	0.5239 ± 0.0178	0.7815 ± 0.0185	0.7673 ± 0.0109	0.8093 ± 0.0154	0.6516 ± 0.0385
MNIST	0.2918 ± 0.0174	0.4408 ± 0.0192	0.3820 ± 0.0121	0.6494 ± 0.0210	0.4463 ± 0.0313	0.3794 ± 0.0153
Montreal	0.7246 ± 0.0131	0.7049 ± 0.0098	0.2885 ± 0.0001	$\mathbf{0.8158 \pm 0.0210}$	0.7530 ± 0.0478	0.7555 ± 0.0784
Newsgroups	0.4079 ± 0.0080	0.4021 ± 0.0098	0.3687 ± 0.0077	$\mathbf{0.4144 \pm 0.0029}$	0.4045 ± 0.0044	0.4076 ± 0.0039
Spam	$\mathbf{0.9570 \pm 0.0045}$	0.8272 ± 0.0047	0.9521 ± 0.0069	0.9515 ± 0.0023	0.9443 ± 0.0035	0.9397 ± 0.0015
Yale	0.5261 ± 0.0346	0.5891 ± 0.0155	0.4788 ± 0.0991	$\mathbf{0.6894 \pm 0.0295}$	0.5394 ± 0.0117	0.5441 ± 0.0059

negative binomial RFLVMs (Tab. 1, Figs. 4, 5, 6). The point of these results is not that RFLVMs are the best method for every dataset, a spurious claim given "no free lunch" theorems (Wolpert and Macready, 1997), but rather that our framework allows for the easy implementation of a large number of practical non-conjugate GPLVMs. Thus, RFLVMs are useful when first performing nonlinear dimension reduction on non-Gaussian data. We posit that our improved performance is because the generating process from the latent space to the observations for these data sets is (in part) nonlinear, non-RBF, and integer-valued. See Appendix D.3 for wall-time experiments for the models in Table 1.

4 CONCLUSION

We presented a framework that uses random Fourier features to induce computational tractability between the latent variables and GP-distributed maps in Gaussian process latent variable models. Our approach allows the Gaussian model to be extended to arbitrary distributions, and we derived an RFLVM for Gaussian, Poisson and logistic distributions. We described distribution-specific inference techniques for each posterior sampling step. Our empirical results showed that each was competitive in downstream analyses with existing distribution-specific approaches on diverse data sets including synthetic, image, text, and multi-neuron spike train data. We are particularly interested in exploring extensions of our generalized RFLVM framework to more sophisticated models such as extending GP dynamic state-space models (Ko and Fox, 2011) to count data and neuroscience applications, which assume temporal structure in $\mathbf{X}$.

RFLVMs have a number of limitations that motivate future work. First, the latent variables are unidentifiable up to scale and rotation. Our rescaling procedure (Sec. 2.4) does not allow heteroscedastic dimensions and enforces orthogonality between the Gaussian latent variables. This prevents the use of more structured priors, such as a GP prior on $\mathbf{X}$, since any inferred structure is eliminated between iterations. We are interested in adopting constraints from factor analysis literature to address the identifiability issues without a restrictive rescaling procedure (Erosheva and Curtis, 2011; Millsap, 2001; Ghosh and Dunson, 2009). Second, label switching in mixture models is a well-studied challenge that is present in our model. Enforcing identifiability may improve inference and model interpretability (Stephens, 2000). In this work, we focused on distributions in the exponential family because this class is both ubiquitous and well-studied. However, we do not see obvious obstacles to extending our approach to data likelihoods outside the exponential family, as we only need closed-form gradients to learn the latent space. Finally, our model has a number of hyperparameters such as the latent dimension, the number of random Fourier features, and the number of Gibbs sampling iterations. Both simplifying the model and estimating these hyperparameters from data are two important directions to improve the usability of RFLVMs.

Acknowledgements

The authors would like to thank Anqi Wu and Jonathan Pillow for their comments. The contribution of Michael Zhang was partially funded by the HKU-URC Seed Fund for Basic Research for New Staff.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174.

Baker, S. G. (1994). The multinomial-Poisson transformation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(4):495–504.

Balasubramanian, M., Schwartz, E. L., Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2002). The Isomap algorithm and topological stability. *Science*, 295(5552):7–7.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Bochner, S. (1959). *Lectures on Fourier integrals*, volume 42. Princeton University Press.

Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013). Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*, pages 2445–2453.

Cunningham, J. P. and Byron, M. Y. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509.

Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M. (2017). Random feature expansions for deep Gaussian processes. In *International Conference on Machine Learning*, pages 884–893. PMLR.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. (2016). Variational inference for latent variables and uncertain inputs in Gaussian processes. *Journal of Machine Learning Research*, 17(1):1425–1486.

Dubey, K. A., Zhang, M. M., Xing, E., and Williamson, S. (2020). Distributed, partially collapsed MCMC for Bayesian nonparametrics. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 3685–3695. PMLR.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):1–14.

Erosheva, E. A. and Curtis, S. M. (2011). Dealing with rotational invariance in Bayesian confirmatory factor analysis. *Department of Statistics, University of Washington, Seattle, Washington, USA*.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica*, 78(1):35–71.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

Gopalan, P., Hofman, J. M., and Blei, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. In *UAI*, pages 326–335.

GPy (2012). GPy: A Gaussian process framework in Python. `http://github.com/SheffieldML/GPy`.

Hensman, J., Durrande, N., and Solin, A. (2017). Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18(1):5537–5588.

Holmes, C. C., Held, L., et al. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Ko, J. and Fox, D. (2011). Learning GP-BayesFilters via Gaussian process latent variable models. *Autonomous Robots*, 30(1):3–23.

Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229.

Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, pages 329–336.

Lawrence, N. D. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent

variable models. *Journal of Machine Learning Research*, 6(Nov):1783–1816.

Lawrence, N. D. (2007). Learning for larger datasets with the Gaussian process latent variable model. In *Artificial Intelligence and Statistics*, pages 243–250.

Lázaro-Gredilla, M., Quinonero-Candela, J., and Figueiras-Vidal, A. (2007). Sparse spectral sampling Gaussian processes. Technical report, Technical Report, Microsoft Research.

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11:1865–1881.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Linderman, S., Johnson, M. J., and Adams, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the Pólya-gamma augmentation. In *Advances in Neural Information Processing Systems*, pages 3456–3464.

Linderman, S. W., Johnson, M. J., Wilson, M. A., and Chen, Z. (2016). A Bayesian nonparametric approach for uncovering rat hippocampal population codes during spatial navigation. *Journal of Neuroscience Methods*, 263:36–47.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, 85(4):755–770.

Macke, J. H., Buesing, L., Cunningham, J. P., Byron, M. Y., Shenoy, K. V., and Sahani, M. (2011). Empirical models of spiking in neural populations. In *Advances in Neural Information Processing Systems*, pages 1350–1358.

Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling*, 8(1):1–17.

Minka, T. (2000). Bayesian linear regression. Technical report.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Oliva, J. B., Dubey, A., Wilson, A. G., Póczos, B., Schneider, J., and Xing, E. P. (2016). Bayesian nonparametric kernel-learning. In *Artificial Intelligence and Statistics*, pages 1078–1086.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184.

Ročková, V. and George, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Titsias, M. and Lawrence, N. D. (2010). Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851.

Williams, C. K. I. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT Press.

Wilson, A. and Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of International Conference on Machine Learning*, pages 1067–1075.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Wu, A., Roy, N. A., Keeley, S., and Pillow, J. W. (2017). Gaussian process based nonlinear latent structure discovery in multivariate spike train data. In *Advances in Neural Information Processing Systems*, pages 3496–3505.

Yang, T., Li, Y.-F., Mahdavi, M., Jin, R., and Zhou, Z.-H. (2012). Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems*, pages 476–484.

Zhao, H., Rai, P., Du, L., Buntine, W., Phung, D., and Zhou, M. (2020). Variational autoencoders for sparse and overdispersed discrete data. In *International Conference on Artificial Intelligence and Statistics*, pages 1684–1694. PMLR.

Zhou, M. and Carin, L. (2012). Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems*, pages 2546–2554.

Zhou, M., Li, L., Dunson, D., and Carin, L. (2012). Lognormal and gamma mixed negative binomial regression. In *Proceedings of International Conference on Machine Learning*, volume 2012, page 1343.