Amortised Inference in Structured Generative Models with Explaining Away

Anonymous Author(s)
Affiliation
Address
email

Abstract

A key goal of unsupervised learning is to go beyond density estimation and sample generation to reveal the structure inherent within observed data. Such structure can be expressed in the pattern of interactions between explanatory latent variables captured through a probabilistic graphical model. Although the learning of structured graphical models has a long history, much recent work in unsupervised modelling has instead emphasised flexible deep-network-based generation, either transforming independent latent generators to model complex data or assuming that distinct observed variables are derived from different latent nodes. Here, we extend the output of amortised variational inference to incorporate structured factors over multiple variables, able to capture the observation-induced posterior dependence between latents that results from "explaining away" and thus allow complex observations to depend on multiple nodes of a structured graph. We show that appropriately parameterised factors can be combined efficiently with variational message passing in elaborate graphical structures. We demonstrate the framework using synthetic data with a known generative process, We then fit a structured model to high-dimensional neural spiking time-series from the hippocampus of freely moving rodents, demonstrating model identifies latent signals that correlate with behavioural covariates.

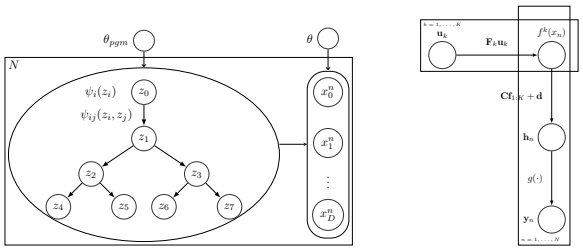
1 Introduction

A central goal of unsupervised learning is the extraction of low-dimensional latent structure that captures dependence between the component variables of high-dimensional observations. Recent developments in deep learning have led to a family of deep generative models (DGM) that have enjoyed remarkable success in density estimation and high-fidelity image/text generation [20, 41, 12, 13, 35]. However, much of the focus with DGMs has been on increasing the expressiveness and accuracy of generation, often overlooking the importance of uncovering the dependency structure inherent within the observed data.

Structured variational inference, by including conditional dependence between the set of latent variables, improves the expressiveness of the variational approximation for the latent posterior distribution. The structured variational autoencoder [SVAE; 17] connects the DGM framework with one element of such inference, combining a structured prior distribution with fully factored mean-field (and functionally conjugate) amortised likelihood potentials. However many generative models of interest involve observations that depend on more than one latent variable. This induces joint potentials in the likelihood, coupling latents even if they are independent *a priori*, a phenomenon sometimes known as "explaining away" [37]. Our goal here is to extend the amortised inference setting to capture such observation-induced dependence in the DGM setting.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

We develop the extension in two contexts. First we consider a simple discrete latent factor model with both *a priori* and *a posteriori* coupling, evaluating the utility of structured amortised inference in this case on synthetic data from a known generative process. Second, we extend Gaussian Process Factor Analysis (GPFA) models that are widely used in neuroscience, to incorporate deep network-based nonlinear mappings from the latent Gaussian process values to observations. The corresponding likelihood functions induce posterior "explaining-away" correlations between the processes. We develop "amortised explaining-away sparse variational GPFA" (AEA-svGPFA) to achieve scalable amortised inference of the full-covariance posterior distribution within the sparse inducing-point GP framework. We apply this model to spike-count data collected from a population of neurons in the hippocampal complex, and show that the unsupervised approach automatically captures dimensions underlying neural firing that reflect relevant behavioural correlates.



(a) (b)

Figure 1: **Structured PGM generative models** with **(a)** tree-structured latents (Section 3.2); **(b)** GPFA latents with sparse approximation (Section 3.3).

2 Background

2.1 Variational inference and structured variational autoencoders

We consider the variational inference (VI) problem of approximating the posterior distribution of the latent variables z given observations y. Consider a general generative process

$$z \sim p(z|\theta), \quad y \sim p(y|z;\gamma)$$

where θ and γ are the parameters of the prior and conditional likelihood distributions, respectively. Note that we assume a single θ for simplicity, but hyperpriors on θ can be incorporated in the context of Variational Bayes (VB). The posterior distribution, $p(z|y) \propto p(z|\theta)p(y|z,\gamma)$, is often analytically intractable. Hence it is common practice to seek a tractable variational approximation to the true posterior, $q(z)$, which is learned by minimising the KL-divergence with respect to the true posterior distributions, and is equivalent to the maximisation of the variational free energy [46]:

$$q^* = \operatorname*{argmin}_{q\in\mathcal{Q}} \mathrm{KL}[q(z)||p(z|y,\theta)] = \operatorname*{argmax}_{q\in\mathcal{Q}} \langle(\log p(y,z|\gamma,\theta) - \log q(z))\rangle_{q(z)}$$

The Variational Autoencoder (VAE) extends VI by modelling the parameters of the variational approximation using a recognition network that takes as input the observed values, and is able to optimise both generative and recognition network parameters using stochastic optimisation [20, 41]. Although the mean-field variational distribution is often utilised in both standard and amortised VI owing to its simplicity, the assumption of fully factorised latent distributions can be too restrictive to capture the complexity of the true posterior [43, 15]. Intuitively, increasing the flexibility of the class of variational distributions leads to a tighter free energy lower bound, and often leads to better empirical results in general [4].

A number of approaches have been proposed for improving the expressiveness of variational approximation [3, 45, 28, 26]. Here we specifically look at structured VAE [SVAE; 17]. Unlike standard VAEs, SVAE assumes the generative prior distribution, $p(z|\theta)$, to be specified by a structured probabilistic graphical model (PGM). Moreover, the amortised inference network outputs factors, $r(z|y,\phi)$, that approximate the generative likelihood function rather than the variational distribution itself. These are combined with the structured prior distribution to obtain the variational posterior.

$$q^*(z|y,\theta,\phi) \approx \operatorname*{argmax}_{q\in\mathcal{Q}} \left\langle \log\left(\frac{p(z|\theta)r(z|y,\gamma)}{q(z)}\right)\right\rangle_{q(z)} \propto r(z|y,\phi)p(z|\theta) \tag{1}$$

Johnson et al. [17] assumed that the amortised inference outputs fully factorised factor potentials (that are conjugate with $p(z|\theta)$). Comparing to the mean-field VI in standard VAE, Eq. 1 is able to carry the dependency structure in the prior distribution to the variational posterior distribution.

2.2 Gaussian Process Factor Analysis

Gaussian Process Factor Analysis (GPFA) is a widely used model in neural data analysis for inferring the dynamic latent structure underlying high-dimensional population spike trains [50]. Standard GPFA assumes the following generative model.

$$\begin{aligned} f^k(\cdot) &\sim \mathcal{GP}(m_\theta^k(\cdot), \kappa_\theta^k(\cdot,\cdot)), \text{ for } k = 1, \dots, K, \\ h_n(\cdot) &= \sum_{k=1}^{K} c_{nk} f^k(\cdot) + d_n, \text{ for } n = 1, \dots, N, \\ y_n(t) &= p(y_n(t)|g(h_n(t))), \text{ for } t = 1, \dots, T, \end{aligned} \tag{2}$$

where c_{nk} and d_n are parameters defining the affine mapping from the latent space to the neural space, and $g(\cdot)$ is some smooth function. We assume $m_\theta^k(\cdot) = 0$ unless stated otherwise. The temporal correlation in the high-dimensional neural space are fully captured by the latent temporal correlations modelled by the latent processes. The PGM describing the GPFA generative model is shown in 1b.

However, the applicability of GPFA models is limited by the poor scalability of exact posterior inference of Gaussian processes (GP), which scales cubically with the number of observations [39]. To this end, an extension of the sparse variational approximation for GPs is applied to the GPFA framework, leading to the svGPFA model [44, 1]. Details of svGPFA formulations are discussed in our discussion of AEA-svGPFA in Section 3.3.

3 Structured amortised inference and explaining away

3.1 Structured amortised inference with SVAE

We assume a structured prior distribution, specified by the factor potentials, $p(z|\theta) = \frac{1}{Z}\prod_{c\in C}\psi_c(z_c)$. Unlike the SVAE approach of enforcing a fully factorised form, we assume the recognition network outputs structured factor potentials over z, $r(z|y,\phi)\frac{1}{Z_r}\prod_{c\in C_r}\xi_c(z_c)$, approximating the likelihood function, $p(y|z,\theta)$, where C_r is the set of pre-specified factors of the amortised potentials and Z_r is the normalisation constant. Then the variational distribution is proportional to the product of the prior distribution and the amortised factor potentials (given partial optimisation, see Johnson et al. [17]).

$$q_{\text{AEA}}(z|x,\phi,\theta) \propto \prod_{c\in C}\psi_c(z_c)\prod_{c\in C_r}\xi_c(z_c) \tag{3}$$

We assume conjugate factors unless otherwise stated. Hence the analytical (approximate) form of $q_{AEA}(z|x,\phi,\theta)$ can be computed with (loopy, if necessary) belief propagation. In the most general form, we could let $r(z|y,\phi)$ be a single joint factor potential over the complete set of latent variables.

Similar to the SVAE formulation, the free energy objective takes the following form.

$$\mathcal{F}_{\text{AEA}}(\theta,\phi) = \langle \log p(z|\theta) + \log p(y|z,\gamma) - \log q_{\text{AEA}}(z|y,\phi,\theta)\rangle_{q_{\text{AEA}}(z|y,\phi,\theta)} \tag{4}$$

The training of the model follows standard VAE-style stochastic optimisation to update the parameters of the recognition and generative networks, as well as (optionally) the parameters of the prior PGM.

The following proposition justifies the utility of using structured amortised inference (the proof can be found in appendix C).

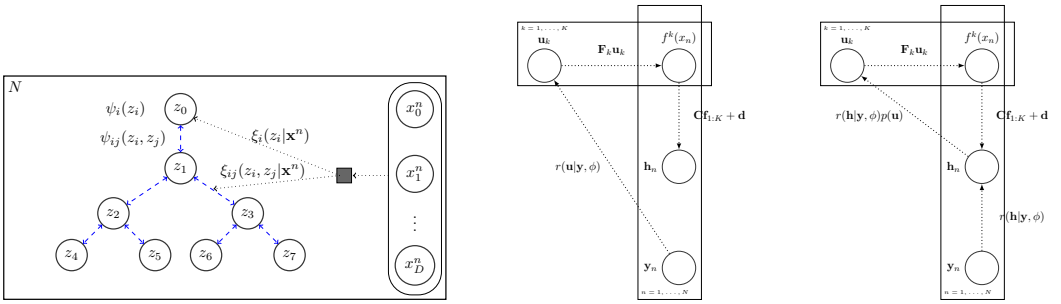
Proposition 3.1. *The AEA-SVAE objective function provides a tighter lower bound to the free energy than the SVAE objective function.*

$$\max_{q(z)} \mathcal{F}[q(z)] \geq \mathcal{F}_{AEA}(\theta,\phi) \geq \mathcal{F}_{SVAE}(\theta,\phi) \tag{5}$$

Below we describe how our framework can be instantiated with different latent PGM structure, the empirical evaluations of the discussed methods can be found in Section 5.

3.2 Tree-Structured Latent PGM

We instantiate our method with discrete tree-structured PGM (Figure 1a), which enables exact inference can be achieved given two iterations of message passing. A general tree-structured PGM



(a) (b) (c)

Figure 2: **Graphical models with structured inference**. **(a)** Tree-structured amortised potentials is combined with prior distribution (where blue dashed lines denote message passing steps); **(b)** Standard GPFA inference with sparse variational approximation; **(c)** Structured amortised inference of AEA-svGPFA model.

takes the following density function.

$$p(z) = \frac{1}{Z} \prod_i \psi_i(z_i) \prod_{(i,j)\in E} \psi_{ij}(z_i, z_j) \tag{6}$$

Consider an arbitrary non-linear generative likelihood function $p(y|z)$. Then the optimal approximation to $p(y|z)$ as a density function of z will in general contain non-singleton factors of z. We provide a analytical example illustrating this in appendix A.1.

The optimal variational distribution should have the same PGM structure as the posterior, which often contains joint factors over many (possibly all) latent variables. However, as the number of states grows exponentially with the number of latent variables, in practice it is usually not scalable to have amortised inference output a joint factor that approximates the exact posterior joint factor potentials. We hence seek a scalable alternative by enforcing the amortised factor potential to be of the same structure as the prior PGM; in this case, a tree-structured distribution with the same set of nodes and pairwise connections (Figure 2a).

$$r(z|y, \phi) = \frac{1}{Z_r} \prod_i \xi_i(z_i) \prod_{(i,j)\in E} \xi_{ij}(z_i, z_j)$$

Given these tree-structured amortised potentials, we can carry out exact inference with message passing to compute the variational distribution, then update the model parameters by maximising the free energy objective (Eq. 4) through stochastic optimisation. Further details of the tree-structured amortised inference can be found in appendix A.

3.3 Structured autoencoding nonlinear GPFA

Neural computations are thought to be implemented by the joint activity of large populations of neurons. Unsupervised methods have been developed to identify the dynamic covariates that underlie the population neuronal spiking trains, and relate these to behaviour or environmental variables so as to gain understanding of the functional and cognitive aspects of neural computations. One popular such model is GPFA [50]. In order to alleviate the scalability issue, sparse variational approximation inference for GPFAs with inducing points were proposed [44, 1, 10]. For $k = 1, \ldots, K$, the inducing points $\mathbf{u}_k$ represents the function values of the latent process f^k at the M_k inducing locations, $\mathbf{z}_k$. We then augment the GPFA generative model (Eq. 2) with the inclusion of the inducing points.

$$p(\mathbf{u}_k|\mathbf{z}_k) = \mathcal{N}(0, K^k_{\mathbf{z}_k\mathbf{z}_k}), \quad p(f_k(\cdot)|\mathbf{u}_k) = \mathcal{GP}(\mathbf{F}_k(\cdot)\mathbf{u}_k, \kappa^k_\theta(\cdot,\cdot) - \mathbf{F}_k(\cdot)K^k_{\mathbf{z}_k\mathbf{z}_k}\mathbf{F}^T_k(\cdot)), \tag{7}$$

where $\mathbf{F}_k(\cdot) = \kappa^k_\theta(\cdot, \mathbf{z}_k)(K^k_{\mathbf{z}_k\mathbf{z}_k})^{-1}$ is the linear operator that maps $\mathbf{u}_k$ to $f_k(\cdot)$; $\kappa^k_\theta(\cdot, \mathbf{z}_k)$ is a vector-valued function with $\kappa^k_\theta(x, \mathbf{z}_k) = [\kappa^k_\theta(x, z_{k1}), \kappa^k_\theta(x, z_{k2}), \ldots, \kappa^k_\theta(x, z_{kM_k})]$; and $K^k_{\mathbf{z}_k\mathbf{z}_k}$ is the Gram matrix of evaluation of the kernel function κ^k_θ at the inducing locations $\mathbf{z}_k$.

It is common practice to use the mean-field variational approximation of the joint posterior distribution of the latent process and the inducing points [44, 1], $q(f_{1:K}, \mathbf{u}_{1:K}) = \prod_{k=1}^K p(f_k\mathbf{u}_k)q(\mathbf{u}_k)$, where

$q(\mathbf{u}_k) = \mathcal{N}(\mathbf{m}_k, \mathbf{S}_k)$. The free energy objective for the svGPFA model given the mean-field variational approximation takes the following expression.

$$\mathcal{F} = \sum_n \langle \log p(\mathbf{y}_n|\mathbf{h}_n)\rangle_{q(\mathbf{h}_n)} - \sum_{k=1}^{K} \mathrm{KL}[q(\mathbf{u}_k)||p(\mathbf{u}_k|\mathbf{z}_k)]$$

where $q(\mathbf{h}_n) = \int d\mathbf{U} p(\mathbf{h}_n|\mathbf{U})q(\mathbf{U})$. It is easy to derive that $q(\mathbf{h}_n) \sim \mathcal{GP}(m_n^h(\cdot), \nu_n^h(\cdot,\cdot))$ is itself a GP with mean and kernel function given by the following expressions.

$$m_n^h(x) = \sum_{k=1}^{K} C_{nk}\mathbf{F}_k(x)\mathbf{m}_k + d_n, \quad \nu_n^h(x,x') = \sum_{k=1}^{K} C_{nk}^2(\kappa_k(x,x') + \mathbf{F}_k(x)(\mathbf{S}_k - K_{\mathbf{z}_k\mathbf{z}_k})\mathbf{F}_k(x')^T)$$

Hence the computational cost reduces from $\mathcal{O}(T^3)$ down to $\mathcal{O}(\bar{M}^3)$, where $\bar{M} = \frac{1}{K}\sum_{k=1}^{K} M_k$, and scales linearly with respect to the number of observations.

Previous work has extended GPFA to deal with non-conjugate observation likelihoods and point-wise nonlinearities [1, 10, 18]. However, adding a complex non-linear likelihood function introduces additional correlation between the latent processes given observations through "explaining away". Earlier work did not capture the cross-latent posterior correlations due to the mean-field variational approximation adopted. Here we describe a novel inference procedure that represents the full-covariance structure of the posterior distribution of the inducing points over all latent processes in a scalable manner.

Given the general expression of the GPFA generative model and the sparse approximation with inducing points, we have that for $n = 1, \ldots, N$.

$$p(\mathbf{h}_n|\mathbf{U}, x_n) = \mathcal{N}(\mathbf{C}\mathbf{F}_n\mathbf{U} + \mathbf{d}, \mathbf{C}(\mathbf{K}_n - \mathbf{F}_n\mathbf{K}_\mathbf{U}\mathbf{F}_n^T)\mathbf{C}^T), \tag{8}$$

where

$$\mathbf{F}_n = \begin{bmatrix} F_1(x_n) & & \\ & \ddots & \\ & & F_K(x_n) \end{bmatrix}, \quad \mathbf{K}_\mathbf{U} = \begin{bmatrix} \mathbf{K}_{\mathbf{z}_1\mathbf{z}_1} & & \\ & \ddots & \\ & & \mathbf{K}_{\mathbf{z}_K\mathbf{z}_K} \end{bmatrix}$$

with $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$ being the collection of all inducing points for all latents, and $\mathbf{K}_n = \mathrm{diag}[\kappa_\theta^k(x_n, x_n)]$. Assume that the amortised inference outputs a fully factorised Gaussian potential on $\mathbf{h}_n$, $r(\mathbf{h}_n|\mathbf{y}_n, \phi) = \mathcal{N}(\mu_n^h, \Psi_n^h)$. By the generative model described in Eq. 3, we can derive the variational distribution on $\mathbf{U}$ given the relationship $p(\mathbf{h}|\mathbf{U})$.

$$\begin{aligned} &q(\mathbf{U}) = \mathcal{N}(\mathbf{m}_\mathbf{U}, \mathbf{S}_\mathbf{U}) \propto p(\mathbf{U}) \prod_n \mathcal{N}(\mu_n^h, \Psi_n^h), \text{ with} \\ &\mathbf{S}_\mathbf{U}^{-1} = \mathbf{K}_\mathbf{U}^{-1} + \sum_n \mathbf{F}_n^T\mathbf{C}^T(\Psi_n^h)^{-1}\mathbf{C}\mathbf{F}_n, \quad \mathbf{m}_\mathbf{U} = \mathbf{S}_\mathbf{U}\left(\sum_n \mathbf{F}_n^T\mathbf{C}^T(\mathbf{\Psi}_n^h)^{-1}(\mu_n^h - \mathbf{d})\right) \end{aligned} \tag{9}$$

Hence giving

$$q(\mathbf{h}_n) \propto p(\mathbf{h}_n|\mathbf{U})q(\mathbf{U}) = \mathcal{N}(\mathbf{C}\mathbf{F}_n\mathbf{m}_\mathbf{U} + \mathbf{d}, \mathbf{C}(\mathbf{K}_n + \mathbf{F}_n(\mathbf{S}_\mathbf{U} - \mathbf{K}_\mathbf{U})\mathbf{F}_n^T)\mathbf{C}^T) = \mathcal{N}(\mathbf{m}_n^h, \mathbf{S}_n^h)$$

Then the complete (reparameterised) Monte Carlo estimate of the free energy objective is

$$\mathcal{F}(\theta, \gamma, \phi, \mathbf{C}, \mathbf{d}, \mathbf{Z}) = \sum_{n=1}^{N} \frac{1}{S}\sum_{s=1}^{S} \log p(\mathbf{y}_n|\mathbf{m}_n^h + \mathbf{L}_n^h\epsilon_s) - \mathrm{KL}[q(\mathbf{U})||p(\mathbf{U})] \tag{10}$$

where $\epsilon_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, θ is the set of kernel parameters, $\mathbf{L}_n^h$ is the lower-triangular matrix such that $\mathbf{S}_n^h = \mathbf{L}_n^h(\mathbf{L}_n^h)^T$, and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_K]$ is the set of all inducing locations.

We note that we have retained the scalability of amortised inference by outputting a diagonal amortised Gaussian potential on $\mathbf{h}$, and due to the relationship between the inducing points $\mathbf{U}$ and $\mathbf{h}$ ($\mathbf{F}$-projection and affine transformation, Eq. 8), we are able to compute a full-covariance variational Gaussian approximation for $\mathbf{U}$, allowing the computation of posterior correlations between latent processes induced by "explaining away". We term our model as Amortised Explaining Away svGPFA (AEA-svGPFA), which has the same generative model as svGPFA (Figure 1b) but replacing $g(\cdot)$ with some general non-linear mapping modelled by a neural network. The complete inference procedure for AEA-svGPFA described above is different from that of svGPFA (and that of other DGMs with GP latents [2]; Figure 2b), and is graphically illustrated in Figure 2c.

4 Related Works

Structured Deep Generative Models. Ample literature has studied the inclusion of latent graphical structures into DGM [22, 8, 17, 29, 27]. The prominent work by Johnson et al. [17] proposed SVAE, where the recognition network outputs fully factorised conjugate factor potentials with respect to the structured latent prior, and the variational distribution is computed through partial optimisation of the surrogate free energy given the recognition potentials. The model parameters are jointly optimised by maximising the free energy. In another closely related work, Lin et al. [29] proposed structured inference network, which approximates the variational distribution as the combination of the recognition potential and a separate structured latent distribution independent of the prior, the independent structured latent distribution can have non-conjugate factors to improve the expressiveness of the model. Our model can be interpreted as a strict generalisation of the SVAE framework, where we allow arbitrary structured amortised potential, broadening the flexibility of variational approximation given SVAE. However, the current framework cannot handle models with latent priors containing non-conjugate factors, which we leave as future work. In practice, we extend the recognition potential to contain factor potentials over multiple latent variables whereas Johnson et al. [17], Lin et al. [29] have only exhibited success with mean-field recognition potentials.

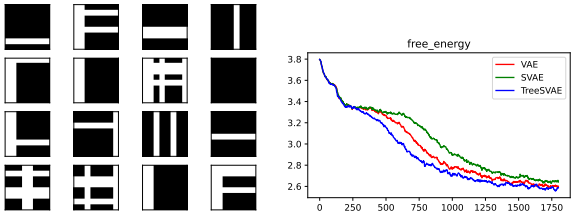
svGPFA and Extensions. High-dimensional time-series data, such as population neural firing pattern, usually lie on a low-dimensional manifold that can be characterised by the small number of dynamical covariates that fully explain the temporal correlation exhibited in the observations. Extraction of the latent covariates is particularly essential in analysing neural data, where the set of external measurable variables and internal mental states are potentially entangled and go through a highly non-linear process to influence the neuronal firing. Hence latent variable models such as GPFA are usually adopted for extracting the relevant latent covariates that contains non-trivial temporal correlations [50, 7, 10]. Recent advances have proposed sparse variational inference (sVI) and methods for dealing with non-conjugate likelihood function to promote the scalability and applicability of GPFA models [1, 10, 19]. There also exists prior works of combining DGM (e.g., VAEs) with GP latents [6, 5, 2]. However, AEA-svGPFA differs from all prior works as it is able to combine svGPFA with amortised inference to allow scalable variational inference of a "free-form" GP posterior (i.e., with full-covariance).

5 Results

5.1 Tree-Structured Amortised Inference with SVAE on "Bar" Dataset

We first evaluate tree-structured amortised inference on a synthetic dataset. We consider the "Bar" dataset of horizontal and vertical bars in $D \times D$ square-grids.

$$p(\mathbf{z}|\theta) = \mathcal{B}(z_0|p_0) \prod_{i>0} \mathcal{B}(z_i|\text{pa}(z_i)),$$
$$p(\mathbf{y}|\mathbf{z}, \mathbf{W}, \mathbf{b}) = \prod_{d=1}^{D^2} \mathcal{B}(y_d|\sigma(\mathbf{W}\mathbf{z} + \mathbf{d})_d) \tag{11}$$



(a) (b)

Figure 3: **Evaluation on the "Bar" dataset**. **(a)** Exemplary samples drawn from the generative model (Eq. 11); **(b)** Free energy of the three models (TreeSVAE, VAE and SVAE) over the training process on the standard "Bar" dataset. All evaluations are based on the averaging over 5 random seeds.

where $\text{pa}(z)$ denote the parent node of node z, $\sigma(\cdot) = \frac{1}{1+\exp(\cdot)}$ is the sigmoid function, and $\mathcal{B}(p)$ denotes the Bernoulli distribution with rate p. Each z_i denotes the activation of one bar in the square-grid (e.g., $z_1 = 1$ indicates the presence of the first horizontal bar). The generative parameters $\mathbf{W}$ and $\mathbf{b}$ are defined as

$$W_{ij} = \begin{cases} 2 \times \omega_i & \text{if pixel i is on the bar j} \\ 0 & \text{otherwise} \end{cases}, \quad b_i = -\omega_i, \text{ for } i = 1, \dots, D^2, \tag{12}$$

where $\{\omega_i\}_{i=1}^{D^2}$ are the temperature parameters that control the degree of randomness in the data generation. Exemplary samples from the "Bar" dataset if shown in Figure 3a ($D = 8$).

We apply our approach, SVAE with tree-structured amortised potentials (TreeSVAE), standard VAE, and standard SVAE, all with 16-dimensional latents (corresponding to the activation of 16 possible

bars), on the "Bar" dataset. The training curves of the free-energy objective is shown in Figure 3b. We observe that TreeSVAE consistently outperforms standard SVAE and VAE through the training process in terms of the free energy objective. Further details on the generation quality and dealing with mismatches in the PGM structures in can found in appendix D.1.

5.2 AEA-svGPFA

5.2.1 Synthetic and EEG Dataset

We firstly empirically evaluate whether AEA-svGPFA outperforms methods that do not capture the intra-latent posterior correlation. We use a synthetic dataset and a small-scale EEG dataset for comparison. Note that we assume Squared Exponential (SE) kernel, $\kappa(x, x') = \lambda \exp(-\frac{||x-x'||^2}{\tau^2})$, unless stated otherwise, where λ and τ are the scale and length-scale parameters, respectively.

Synthetic Dataset We generate from the GPFA generative model (Eq. 2), with $g(\cdot) = \sigma(\Phi(\cdot))$, where $\Phi(\cdot)$ represents the functional mapping through a fixed, randomly-initialised 2-layer MLP with ReLU hidden non-linearity, and σ is the sigmoid function.

EEG Dataset We follow the experimental procedure as described in Requeima et al. [40], and consider an EEG measurement dataset spanning 1 second at $256Hz$ frequency when images are shown to the patients [51]. Each datapoint consists of the voltage readings from 7 electrodes positioned on the patient's scalp. We assume all data are observed, contrary to the settings in [40, 2] (we include results and further discussion on partial observability in appendix D.2).

	synthetic dataset		**EEG datset**		**Population spike dataset**
Methods	SMSE	NLL	SMSE	NLL	SMSE
AEA-svGPFA	$\mathbf{0.31 \pm 0.02}$	$\mathbf{1.69 \pm 0.04}$	$\mathbf{0.27 \pm 0.05}$	$\mathbf{1.81 \pm 0.21}$	$\mathbf{0.47 \pm 0.18}$
SGPVAE [2]	0.36 ± 0.05	1.76 ± 0.07	0.35 ± 0.05	2.17 ± 0.17	0.55 ± 0.19

Table 1: Quantitative comparison of performance between AEA-svGPFA and SGPVAE [2] on synthetic dataset, EEG dataset and population spiking dataset (Section 5.2.2) with respect to standarised mean squared error (SMSE) and negative log-likelihood (NLL). All evaluations are averaged over 5 random seeds.

Our main baseline model is SGPVAE [2], it can be interpreted as a version of the SVAE model with independent GP latents. SGPVAE has the same generative model as the svGPFA generative model (Figure 1b), whilst having identity affine transformation between $\mathbf{f}_{1:K}$ and $\mathbf{h}$ (i.e., $\mathbf{C} = \mathbf{I}$ and $\mathbf{d} = \mathbf{0}$), and $g(\cdot)$ is a general non-linear function parameterised by a neural network. However, the non-trivial affine transformation in AEA-svGPFA (and svGPFA) can be easily modelled by the neural network likelihood function, hence the main difference between AEA-svGPFA and SGPVAE lie in the structured amortised inference step (Section 3.3). Given the evaluation comparison in Table 4, we see that AEA-svGPFA achieves stronger performance on both the synthetic and EEG dataset, with respect to both the SMSE and NLL metric. Hence, similar to the empirical findings with tree-structured amortised inference, we found that the inclusion of structured amortised inference always improve the performance. Moreover, the computational complexity of AEA-svGPFA is on the same order as SGPVAE, see appendix D.2 for further analysis and discussion.

5.2.2 Population Neuronal Firing Data

It is well-known that the hippocampal CA1 place cells and mEC grid cells are spatially modulated [34, 33], and previous work has also established speed and directional modulation of CA1 and mEC neurons [31, 42, 23]. Moreover, there are often non-trivial correlations between the behavioural covariates. For instance, one might reasonably expect correlation between speed and spatial location is an animal runs fastest in the central area of the environment and slower near the boundaries. Neither svGPFA or SGPVAE can capture the posterior correlations between latent variables given their mean-field assumptions, whereas AEA-svGPFA, given its free-form variational approximation with full covariance structure, is able to capture arbitrary correlation between the latents, and hence is an ideal model for unsupervised learning of the latent structure that correctly reflect the behavioural covariates underlying the neural firing pattern.

We use the single-cell spiking data of neurons in the hippocampal CA1 and medial entorhinal cortex (mEC) regions of rats recorded during exploration in a Z-shaped track reported by Ólafsdóttir et al.

[52]. The dataset contains 28 experiment sessions, each spanning 10 minutes. Exemplary neural firing patterns of the population are shown in Figure 14. Note that due to the nature of the task, we can interpret the spatial location of the rat being standard and "unfolded" distances along the track ("unfolded" meaning the sign of the distance value is dependent on the direction of movement). Other considered behavioural correlates are speed, direction of movement (inbound v.s. outbound), and head-direction.

We again use the same generative model in Eq. 2, but with Poisson observation likelihood.

$$\mathbf{y}(t) = \prod_{d=1}^{D} \text{Poisson}(y_d(t)|\lambda_{\text{NN}}(\mathbf{h})_d) \tag{13}$$

where $\lambda_{\text{NN}}(\cdot)$ is a neural-network mapping from the GPFA features, $\mathbf{h}$, to the rate of Poisson-distributed firing counts for each neuron over a bin-size of $100ms$.

We firstly apply AEA-svGPFA and SGPVAE to the dataset, and quantitative comparison indicates that AEA-svGPFA outperforms SGPVAE with respect to SMSE of prediction (Table 1).

Having trained the model, and given the firing pattern observations, $\mathbf{y}$, we could immediately compute the variational posterior latent distribution over the latent processes.

$$q(\mathbf{f}_n) \propto p(\mathbf{f}|\mathbf{U})q(\mathbf{U}) = \mathcal{N}(\mathbf{m}_n^f, \mathbf{S}_n^f), \quad \mathbf{m}_n^f = \mathbf{F}_n \mathbf{m}_{\mathbf{U}}, \quad \mathbf{S}_n^f = \mathbf{K}_n + \mathbf{F}_n(\mathbf{S}_{\mathbf{U}} - \mathbf{K}_{\mathbf{U}})\mathbf{F}_n^T \tag{14}$$

where $q(\mathbf{U})$ is computed as described in Eq. 9. Note that due to computational constraints and the stochastic mini-batch nature of the training process, the number of inducing points is much smaller than the total number of observations (64 v.s. 6000 in this case), hence causing temporal chunking artifacts in the posterior estimation. AEA-svGPFA can resolve the temporal chunking with an additional sVI step over a much larger time window (i.e., increasing the number of inducing points) with the weights of the recognition network fixed at the final state of the training process (see further discussion in Section D.2.2). We note that SGPVAE is unable to remove the chunking artifacts since it learns a recognition factor directly onto the inducing points, hence no further sVI steps can be performed with respect to a larger set of inducing locations.

We perform the two-dimensional Canonical Correlation Analysis (CCA) on the variational posterior means and the behavioural correlates over the entire trajectory [14]. Here we choose one session for in-depth investigation and mostly show the qualitative predictions and comparison, more qualitative and quantitative results across sessions can be found in Appendix D.2.2.

The heatmap of the correlation coefficients between the canonical correlates of the posterior means ($CCX\{1,2\}$) and each of the behavioural correlates is shown in Figure 4a. We observe an overall pattern of high-correlations, indicating the low-dimensional manifold parameterised by the conjunctive set of behavioural correlates can be accurately captured by the latent variables learned with AEA-svGPFA solely from neural spikes.

In order to verify that if the learned latent structure indeed contain decodable information with respect to the behavioural covariates, we extract the direction-modulated neurons predicted by the trained model (details in appendix D.2.2). In Figure 4b, we compare the direction-modulation of the such model-predicted neurons against that of the rest of the neurons, where the direction-modulation is defined as the correlation between raw single-cell spike counts and the direction values. We observe that the model-predicted direction-modulated neurons exhibits significantly greater direction-modulation than the other neurons (p-value=5.56×10^{-4}). For the neuron with the strongest predicted direction-modulation, we show the firing profiles of the neuron when the animal is travelling along both directions, Figure 4b shows that the model-predicted neuron has different firing fields corresponding to different travel direction, hence indeed exhibiting strong direction-modulation. Similar comparison holds for spatial and speed modulations as well, indicating AEA-svGPFA is able to learn the latent structure that coheres well with the ground-truth behavioural covariates, given solely the population spike trains, even for neurons that exhibit conjunctive coding mechanism.

We plot $CCX\{1,2\}$ against the (x, y) location in Figure 4d, where we observe clear clustered latent structure with respect to the direction of movement, indicating the latent dimensions learned by AEA-svGPFA exhibit disentanglement of direction from spatial location given simple linear projection, despite the conjunctive coding mechanism exhibited in the CA1 neurons. By comparing to the same plot with SGPVAE, we observe from Figure 15 that the spatial and direction variables are more intertwined, hence qualitatively indicating the superiority of AEA-svGPFA.

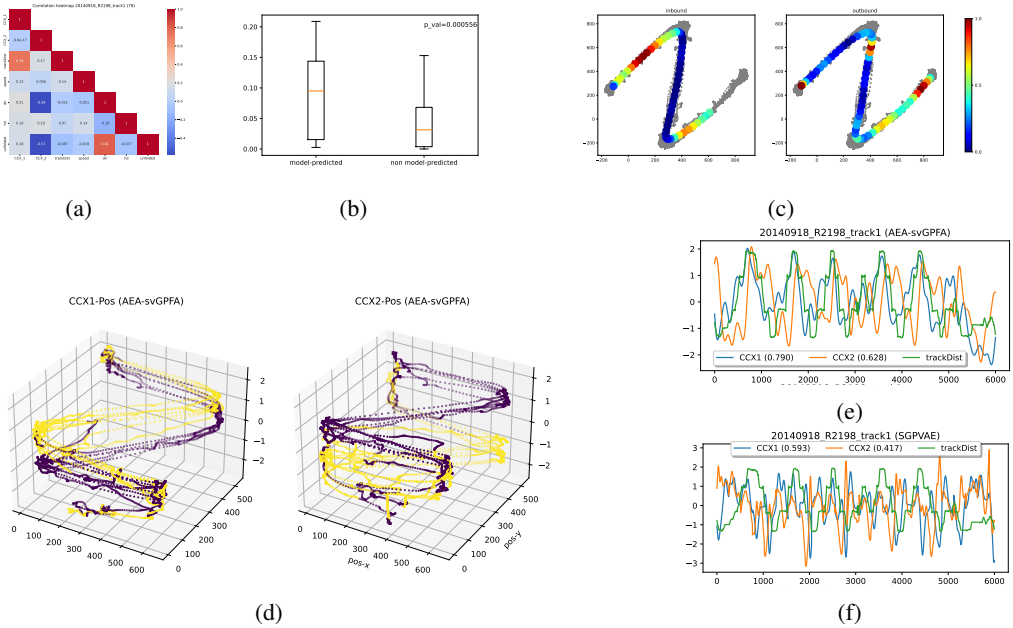


Figure 4: **Analysis of AEA-svGPFA posterior latent structure**. **(a)** Heatmap of the correlation between the CCs and the behavioural correlates for the selected session; **(b)** Directional-modulation comparison between the model-predicted direction-modulated neurons (appendix D.2.2) against other neurons; **(c)** Firing fields of the model-predicted direction-modulated neuron during inbound v.s. outbound movement; **(d)** Plot of $CCX\{1,2\}$ of AEA-svGPFA posterior latents against (x, y) position, color indicates direction of movement (yellow: inbound, magenta: outbound); **(e-f)** Temporal evolution of (standardised values of) $CCX\{1,2\}$ and distance along the track for AEA-svGPFA and SGPVAE (number in the legend is corr$(CCX\{1,2\}, CCY\{1,2\}$ where $CCY\{1,2\}$ are the canonical correlates for the bahavioural covariates.

Figure 4e shows the temporal evolution of the latent CCs for AEA-svGPFA and SGPVAE, as well as that of the distance along the track . We observe that $CCX1$ of AEA-svGPFA consistently tracks the dynamics of the distance, whereas neither $CCX1$ or $CCX2$ of SGPVAE could accurately represents the distance. Moreover, given the additional sVI step with larger time windows, $CCX\{1,2\}$ of AEA-svGPFA exhibits much less temporal chunking than that of SGPVAE.

6 Discussion

In this paper, we proposed structured amortised inference, which can be viewed as a strict generalisation of the SVAE framework [17], where we allow structured amortised factor potentials to capture the posterior correlations induced by the "explaining away" effect. We firstly instantiate our framework with tree-structured latent PGM, and demonstrate that the resulting model consistently yields higher free energy (lower negative free, Figure 5b) and better generation quality (appendix D.1).

We introduce AEA-svGPFA, a combination of the proposed structured amortised inference framework and VAEs with GPFA latents, where scalable free-form variational Gaussian approximation can be achieved. The inclusion of non-linear likelihoods both expand the expressiveness of the model and induces interesting posterior correlations that could inspire new findings. We show that AEA-svGPFA outperforms the baseline on all presented tasks, both quantitatively and qualitatively. By applying to population spiking data of hippocampal neurons, we show that AEA-svGPFA is able to identify latent signals that exhibit strong correlation with the behavioural covariates that are well-known to be conjunctively encoded by the recorded neurons. While we have focused on temporal correlation exhibited in time-series data, we note that the sparse amortised variational approximation presented here can also be straightforwardly applied to other GP-latent models, such as GP-LVM and KISS-GP [24, 49, 47]. We note that here we have only looked at single-cell spiking data, which could in principle be combined with the latent processes found by applying our model to EEG or calcium imaging datasets recorded during the same session, possibly with different likelihoods functions. Such joint analysis of latent structures found by analysing different datasets under the same behaviour could potentially discover new findings with respect to, e.g., replay, theta sweeps, etc.

References

[1] Adam, V., Hensman, J., and Sahani, M. Scalable transformed additive signal decomposition by non-conjugate gaussian process inference. In *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*, pp. 1–6. IEEE, 2016.

[2] Ashman, M., So, J., Tebbutt, W., Fortuin, V., Pearce, M., and Turner, R. E. Sparse gaussian process variational autoencoders. *arXiv preprint arXiv:2010.10177*, 2020.

[3] Blei, D. M. and Jordan, M. I. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[4] Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[5] Campbell, A. and Liò, P. tvgp-vae: Tensor-variate gaussian process prior variational autoencoder. *arXiv preprint arXiv:2006.04788*, 2020.

[6] Casale, F. P., Dalca, A., Saglietti, L., Listgarten, J., and Fusi, N. Gaussian process prior variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

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

[8] Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., and Shanahan, M. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[9] Dorta, G., Vicente, S., Agapito, L., Campbell, N. D., and Simpson, I. Training vaes under structured residuals. *arXiv preprint arXiv:1804.01050*, 2018.

[10] Duncker, L. and Sahani, M. Temporal alignment and latent gaussian process factor inference in population spike trains. *Advances in Neural Information Processing Systems*, 31, 2018.

[11] Ghahramani, Z. and Jordan, M. I. Factorial hidden markov models. *Machine learning*, 29(2): 245–273, 1997.

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] Gregor, K., Danihelka, I., Graves, A., Rezende, D., and Wierstra, D. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pp. 1462–1471. PMLR, 2015.

[14] Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[15] Hoffman, M. D. and Blei, D. M. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, pp. 361–369, 2015.

[16] Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[17] Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29:2946–2954, 2016.

[18] Keeley, S., Aoi, M., Yu, Y., Smith, S., and Pillow, J. W. Identifying signal and noise structure in neural population activity with gaussian process factor models. *Advances in Neural Information Processing Systems*, 33:13795–13805, 2020.

[19] Keeley, S., Zoltowski, D., Yu, Y., Smith, S., and Pillow, J. Efficient non-conjugate gaussian process factor models for spike count data using polynomial approximations. In *International Conference on Machine Learning*, pp. 5177–5186. PMLR, 2020.

[20] Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[21] Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM review*, 51(3): 455–500, 2009.

[22] Krishnan, R. G., Shalit, U., and Sontag, D. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

[23] Kropff, E., Carmichael, J. E., Moser, M.-B., and Moser, E. I. Speed cells in the medial entorhinal cortex. *Nature*, 523(7561):419–424, 2015.

[24] Lawrence, N. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16, 2003.

[25] LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[26] Lee, D. B., Min, D., Lee, S., and Hwang, S. J. Meta-gmvae: Mixture of gaussian vae for unsupervised meta-learning. In *International Conference on Learning Representations*, 2020.

[27] Li, C., Welling, M., Zhu, J., and Zhang, B. Graphical generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

[28] Li, Y., Turner, R. E., and Liu, Q. Approximate inference with amortised mcmc. *arXiv preprint arXiv:1702.08343*, 2017.

[29] Lin, W., Hubacher, N., and Khan, M. E. Variational message passing with structured inference networks. *arXiv preprint arXiv:1803.05589*, 2018.

[30] Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[31] McNaughton, B. L., Barnes, C. A., and O'Keefe, J. The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. *Experimental brain research*, 52(1):41–49, 1983.

[32] Morningstar, W., Vikram, S., Ham, C., Gallagher, A., and Dillon, J. Automatic differentiation variational inference with mixtures. In *International Conference on Artificial Intelligence and Statistics*, pp. 3250–3258. PMLR, 2021.

[33] Moser, E. I., Kropff, E., and Moser, M.-B. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31:69–89, 2008.

[34] O'keefe, J. and Nadel, L. *The hippocampus as a cognitive map*. Oxford university press, 1978.

[35] Oord, A. v. d., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., and Kavukcuoglu, K. Conditional image generation with pixelcnn decoders. *arXiv preprint arXiv:1606.05328*, 2016.

[36] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[37] Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[38] Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[39] Rasmussen, C. E. and Williams, C. K. I. *Gaussian processes for machine learning.* MIT Press, 2005.

[40] Requeima, J., Tebbutt, W., Bruinsma, W., and Turner, R. E. The gaussian process autoregressive regression model (gpar). In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1860–1869. PMLR, 2019.

[41] Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

[42] Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B. L., Witter, M. P., Moser, M.-B., and Moser, E. I. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312(5774):758–762, 2006.

[43] Saul, L. and Jordan, M. Exploiting tractable substructures in intractable networks. *Advances in neural information processing systems*, 8, 1995.

[44] Titsias, M. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pp. 567–574. PMLR, 2009.

[45] Tran, D., Ranganath, R., and Blei, D. M. The variational gaussian process. *arXiv preprint arXiv:1511.06499*, 2015.

[46] Wainwright, M. J. and Jordan, M. I. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[47] Wilson, A. and Nickisch, H. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International conference on machine learning*, pp. 1775–1784. PMLR, 2015.

[48] Winn, J., Bishop, C. M., and Jaakkola, T. Variational message passing. *Journal of Machine Learning Research*, 6(4), 2005.

[49] Wu, A., Roy, N. A., Keeley, S., and Pillow, J. W. Gaussian process based nonlinear latent structure discovery in multivariate spike train data. *Advances in neural information processing systems*, 30, 2017.

[50] Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S., Shenoy, K. V., and Sahani, M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Advances in neural information processing systems*, 21, 2008.

[51] Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W., and Litke, A. Event related potentials during object recognition tasks. *Brain research bulletin*, 38(6):531–538, 1995.

[52] Ólafsdóttir, H. F., Carpenter, F., and Barry, C. LFP and single-unit data published in Olafsdottir, Carpenter and Barry (2016) Nature Neuroscience 19: 792-794, October 2021. URL `https://doi.org/10.5281/zenodo.5566548`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]