Variational Log-Gaussian Point-Process Methods for Grid Cells

Michael Everett Rule[1], Prannoy Chaudhuri-Vayalambrone[2], Marino Krstulovic[2], Marius Bauza[3], Julia Krupic[2], Timothy O'Leary[1]

[1]Engineering Department, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, United Kingdom. [2]Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge, United Kingdom. [3]Sainsbury Wellcome Centre, University College London, London, United Kingdom.

Abstract

We present practical solutions to applying Gaussian-process methods to calculate spatial statistics for grid cells in large environments. Gaussian processes are a data efficient approach to inferring neural tuning as a function of time, space, and other variables. We discuss how to design appropriate kernels for grid cells, and show that a variational Bayesian approach to log-Gaussian Poisson models can be calculated quickly. This class of models has closed-form expressions for the evidence lower-bound, and can be estimated rapidly for certain parameterizations of the posterior covariance. We provide an implementation that operates in a low-rank spatial frequency subspace for further acceleration. We demonstrate these methods on a recording from grid cells in a large environment.

Introduction

Grid cells in the hippocampal formation modulate their firing rates as a periodic function of location (Hafting et al., 2005; Rowland et al., 2016). Some grid cells are also modulated by head direction (Sargolini et al., 2006; "conjunctive cells"), and recent studies have found subtle dependence on head direction and (Gerlei et al., 2020) and landmarks (Keinath et al., 2018; Krupic et al., 2018) even in non-conjunctive cells. Exploring these relationships requires efficient statistical estimators to compare changes in the spatial dependence of grid-cell activity across conditions.

Standard approaches to spatial statistics have limitations. Grid-cell firing-rate maps are often estimated using a Gaussian kernel-density smoother (e.g. Hafting et al., 2005; Langston et al., 2010; Brandon et al., 2011; Killian et al., 2012). Naïve smoothing approaches remain noisy when data are limited, do not provide a quantification of uncertainty, cannot adapt to inhomogeneous spatial sampling, and cannot take advantage of the center-surround structure of grid-cell firing. Conversely, approaches based on spatial autocorrelations (e.g. Hafting et al., 2005, many others) reduce noise by averaging over space, but cannot be applied to single grid fields.

Gaussian-Process (GP) estimators are a promising solution to these challenges. They offer a principled, Bayesian approach to estimating firing-rate maps. They incorporate assumptions to improve statistical efficiency, and provide a posterior distribution that quantifies uncertainty.

However, open challenges remain in applying existing algorithms to exploratory analysis of large grid-cell datasets. Bayesian priors suitable for grid cells have not been described in the literature, and existing implementations are either limited to specific kernels or are too computationally intensive for large datasets. We resolve both of these issues, and illustrate some practical benefits of Gaussian-process methods compared to non-Bayesian estimators.

We will briefly review Gaussian process methods in neuroscience to contextualize our results, then present (1) a tutorial on applying Gaussian processes to grid-cell data; (2) a technical review of approximate inference algorithms; (3) an application of these methods on an example dataset.

Background

Gaussian processes generalize the multivariate normal distribution to a distribution over functions (MacKay, 1998; Rasmussen, 2003; Keeley and Pillow, 2018). They are a natural candidate for describing neuronal curves as a function of continuous variables, and have emerged as the gold-standard for analyzing neuronal activity in the low-data regime. Many algorithms have been developed for capturing the relationship between neural activity and other variables, or for inferring latent neural states (Yu et al., 2009; Rad and Paninski, 2010; Park et al., 2014; Frigola et al., 2014; Wu et al., 2017; Zhao and Park, 2017; Duncker and Sahani, 2018; Brandman et al., 2018; Rule et al., 2019; Keeley et al., 2020; Jensen et al., 2020, 2021).

Formally, a Gaussian-process distribution is specified by its mean function $\mu(\boldsymbol{x})$ and two-point covariance function $\Sigma(\boldsymbol{x}, \boldsymbol{x}')$, which are analogous to the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ of the multivariate normal distribution (see MacKay, 1998; Rasmussen, 2003; Keeley and Pillow, 2018 for a more thorough introduction). In computation,

however, Gaussian processes are almost always represented in terms of a finite-dimensional approximation. We will use the finite-dimensional notation $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, with the understanding that this represents a particular finite-dimensional projection of our Gaussian process model.

Previous works have described Gaussian process methods for place and grid cells (e.g. Rad and Paninski, 2010; Wu et al., 2017). However, we encountered practical challenges when applying these methods to grid cells in large arenas. Computational efficiency is paramount for exploratory analyses of large datasets. While scalable solutions exist, the fastest methods require spatial covariance priors that can be described in terms of nearest-neighbor interactions (Rad and Paninski, 2010; Cseke et al., 2016). This is not ideal for grid cells, which can display spatial correlations between response fields separated by several centimeters. Previous works have developed ways of approximating the Gaussian process covariances that support fast calculations, while remaining expressive (Jensen et al., 2021). We will elaborate upon these ideas, with a particular focus on grid cells, and introduce some new numerical approaches.

Specifically, the new contributions of this manuscript are (1) Tools for designing GP priors that take advantage of the local spatial topography of grid cells; (2) Efficient and expressive variational Bayesian methods; (3) Numerical approaches for achieving good performance on consumer-grade hardware; (4) A Python package containing a reference implementation and example application to grid-cell data.

Results

We will first review log-Gaussian Poisson models of neural spiking in the context of inferring a grid-cell firing-rate map. These combine a Gaussian prior on (log) firing rate with a Poisson likelihood for spikes. We review numerical approaches for finding the mode of the Bayesian posterior in this scenario, and discuss strategies for designing suitable priors for grid cells.

We then explore variational Bayesian approaches for finding a good Gaussian-process approximation to our posterior log-rate map. The the gradients needed to optimize this variational posterior take on an especially simple form for log-Gaussian Poisson models. We argue that a diagonal update to the precision (inverse covariance) matrix is an ideal approximation for the posterior covariance in log-Gaussian Poisson models. Finally, we discuss how to optimize kernel hyperparameters summarize confidence in the estimated peak locations.

An example experiment

Figure 1 illustrates a typical spatial-navigation experiment in which a rat foraged in a 2 m × 1 m environment (Figure 1a). Spike counts y_t from a grid cell in entorhinal cortex, along with position $\boldsymbol{x}_t = \{x_{1;t}, x_{2;t}\}^\top$, were recorded in 20 ms bins, yielding time series $\boldsymbol{X} = \{\boldsymbol{x}_1, .., \boldsymbol{x}_T\}^\top$ and $\boldsymbol{y} = \{y_1, .., y_T\}^\top$ with T samples. Throughout this manuscript, we will denote scalars as lower-case letters "x", column vectors as bold lower-case letters "$\boldsymbol{x}$", and matrices as bold capital letters "$\boldsymbol{X}$".

The resulting spatial data consists of a map of the number of times the rat visited each location, and the number of spikes observed during each visit. These can be summed on a spatial grid to form occupancy and spike-count histograms, which can be combined to yield a firing-rate histogram (Figure 1b). Here, we bin data on an $M = L \times L$ spatial grid (In Figure 1, we used $L = 128$). We refer to individual bins by a single index m ranging from 1 to M. We denote the tallies of visits to each bin as $\boldsymbol{n} = \{n_1, .., n_m\}^\top$ and the tallies of spikes in each bin as $\boldsymbol{k} = \{k_1, .., k_m\}^\top$.

Estimating a smoothed log-rate map

Our approach will follow variational inference for Gaussian-process generalized linear models as outlined in Challis and Barber (2013). We consider "latent" Gaussian processes, whose values are observed indirectly through a firing-rate nonlinearity and neuronal spiking activity. We will model the log-firing-rate $z(\boldsymbol{x})$ as distributed according to a Gaussian process, with spiking events then modelled as conditionally Poisson. This model is sometimes called a log-Gaussian Cox process, after David Cox (Cox, 1955), and it captures both correlations and over-dispersion in the covariance structure of $z(\boldsymbol{x})$.

We model spike counts within a small time bin Δt as Poisson distributed with firing rate $\lambda(\boldsymbol{x}) = \exp[z]$:

$$\int_t^{t+\Delta t} y(t)\, dt \sim \text{Poisson}[\int_t^{t+\Delta t} \lambda(t)\, dt] \tag{1}$$

We will see later that this choice of an exponential firing-rate nonlinearity $\lambda = \exp(z)$ is useful for obtaining closed-form solutions in variational inference. For small Δt (1–2 spikes/sample), a locally-constant model of the firing-rate suffices, $\int_t^{t+\Delta t} \lambda(t)\, dt \sim \lambda(t) \cdot \Delta t$. For simplicity, we will choose time coordinates such that $\Delta t = 1$ and omit it going forward. The log-likelihood of observing spike count y given rate λ is then:

$$\ln \Pr(y|z) = y \ln(\lambda) - (\lambda) - \ln(y!). \tag{2}$$

The overall likelihood of all spiking observations $\boldsymbol{y}$ depends on the log-firing-rate map $z(\boldsymbol{x})$, and the animal's trajectory over time $\boldsymbol{X}$. We assume that the spiking observations are independent conditioned on the latent variables $\boldsymbol{z}$, so that the likelihood of the overall dataset $\Pr(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{X})$ factors as

$$\begin{aligned} \ln \Pr(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{X}) &= \textstyle\sum_{..T}^{t=1} \ln \Pr(y_t|z(\boldsymbol{x}_t)) \\ &= \textstyle\sum_{..T}^{t=1} [y_t z(\boldsymbol{x}_t) - e^{z(\boldsymbol{x}_t)} + \ln(y_t!)]. \end{aligned} \tag{3}$$

For numerical implementations, we model the function $z(\boldsymbol{x}_t)$ as a vector $\boldsymbol{z} = \{z_1, .., z_M\}$, where each z_m reflects

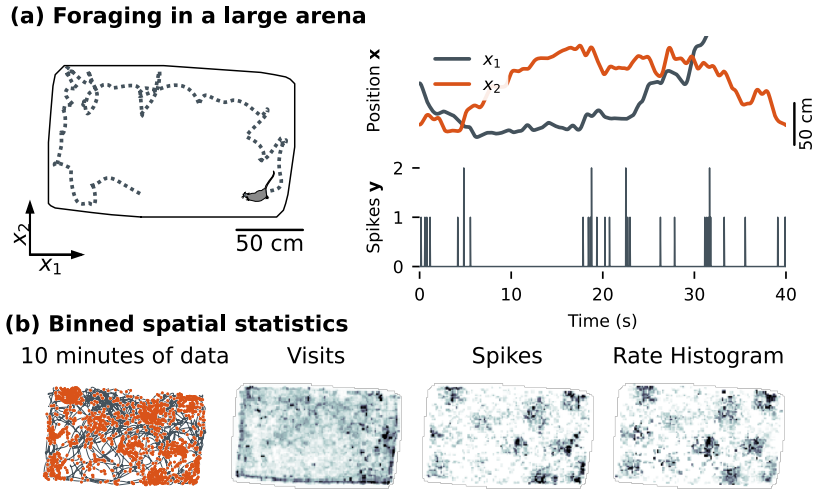

Figure 1: **An example experiment.** **(a)** In this experiment, a rat foraged in a 3×3 meters-squared open environment (left). The rat's position over time "$\boldsymbol{x}$" (right, top), as well as spike counts "$\boldsymbol{y}$" from a single neuron in entorhinal cortex (right, bottom) were recorded (data from Krupic et al., 2018). **(b)** A firing-rate histogram (right, $\boldsymbol{k}/\boldsymbol{n}$) can be estimated by dividing the total number of spikes tallied at each location "$\boldsymbol{k}$" (left) by the number of visits to each location "$\boldsymbol{n}$" (middle). (Color scales are not quantitative.)

the value of $z(\boldsymbol{x}_m)$ at one of M spatial locations. To make the notation easier to read in these derivations, we will interpret each location z_m as a piecewise-constant model of the firing-rate map in a small region of the environment with value $z(\boldsymbol{x}) \approx z_m$ if $\boldsymbol{x} \in b_m$ (in practice we use linearly interpolated binning for improved resolution; see Methods: *Binning of spatial data*).

We can combine time points that fall in the same spatial bin into a single aggregate observation, since these share the same log-rate variable z_m (making our approach a form of pseudo-point method; Quinonero-Candela and Rasmussen, 2005). For each bin, the counts "n_m" and "k_m" tally the total number of visits and spikes, respectively:

$$\begin{aligned}\ln \Pr(\boldsymbol{y}|z, \boldsymbol{X}) &= \textstyle\sum_{..M}^{m=1} \sum_{\boldsymbol{x}_t \in b_m}^{t \text{ s.t.}} [y_t z_m - e^{z_m} + \ln(y_t!)] \\ &= \textstyle\sum_{..M}^{m=1} [k_m z_m - n_m e^{z_m}] + \text{constant}\end{aligned} \tag{4}$$

Since $\ln(y!)$ does not influence the gradient of (4) with respect to z, we can ignore it when optimizing z.

Having combined data from repeated visits to the same location, we now refer to the observations as $\boldsymbol{k} = \{k_1, .., k_M\}$. The number of visits to all locations "$\boldsymbol{n}$" is treated as a known constant. The likelihood in Equation (4) can then be written in vector notation as:

$$\ln \Pr(\boldsymbol{k}|z) = z^\top \boldsymbol{k} - \boldsymbol{n}^\top \boldsymbol{\lambda}(z) + \text{constant}. \tag{5}$$

This is the log-likelihood of observations $\boldsymbol{k}$ given z. To simplify the notation going forward, we will leave the dependence of some quantities on z implicit. For example, we abbreviate $\boldsymbol{\lambda}(z)$ as $\boldsymbol{\lambda}$ when unambiguous.

This observation model has the same form as a Point-Process Generalized-Linear Model (PP-GLM; e.g. Paninski, 2004; Truccolo et al., 2005; Truccolo, 2016). However, adjusting z to maximize (5) alone will lead to overfitting. Instead, one can obtain a smoothed map by taking a Bayesian approach.

We can encode constraints like smoothness or periodicity in our choice of the prior $\Pr(z)$. We use a multivariate Gaussian prior $z \sim \mathcal{N}(\boldsymbol{\mu}_z, \Sigma_z)$, which has the log-probability density

$$\ln \Pr(z) = -\tfrac{1}{2} \left\{ \ln |2\pi\Sigma_z| + (z - \boldsymbol{\mu}_z)^\top \Sigma_z^{-1} (z - \boldsymbol{\mu}_z) \right\}. \tag{6}$$

Summing the log-likelihood (5) and log-prior (6) yields an expression for the log-posterior of z (up to constant terms):

$$\begin{aligned}\ln \Pr(z|\boldsymbol{k}) = -\tfrac{1}{2} &\left\{ \ln |2\pi\Sigma_z| + (z - \boldsymbol{\mu}_z)^\top \Sigma_z^{-1} (z - \boldsymbol{\mu}_z) \right\} \\ &+ z^\top \boldsymbol{k} - \boldsymbol{n}^\top \boldsymbol{\lambda} + \text{constant}.\end{aligned} \tag{7}$$

When the dimension of z is large, estimating (7) via sampling or evaluating it on a grid is infeasible. Instead, we approximate the posterior as a multivariate Gaussian distribution.

Covariance kernels for grid cells

Throughout this manuscript, we will assume that the prior covariance between two points $\Sigma_z(\boldsymbol{x}_1, \boldsymbol{x}_2)$ depends only on the distance between them. In this case, the prior covariance takes the form of a convolution kernel. Since we evaluate our rate map on a regular square grid, and since the prior covariance is a convolution, the matrix Σ_z^{-1} is a circulant matrix and products like $\Sigma_z^{-1}(z - \boldsymbol{\mu}_z)$ can be computed using the Fast Fourier Transform (FFT) in $O(M \log(M))$ time.

How should one select an appropriate prior covariance Σ_z? In Gaussian process regression, the covariance kernel describes how correlated (or anticorrelated) two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the inferred rate map are expected to be, as a function of the distance between them $|\boldsymbol{x}_i - \boldsymbol{x}_j|$:

$$[\Sigma_z]_{ij} = \mathcal{K}\left(|\boldsymbol{x}_i - \boldsymbol{x}_j|\right). \tag{8}$$

It is common to select a kernel that decays as distance increases, although we will use kernels that oscillate to analyze grid cells. For any collection of spatial locations, the Σ_z induced by the kernel needs to be a valid covariance

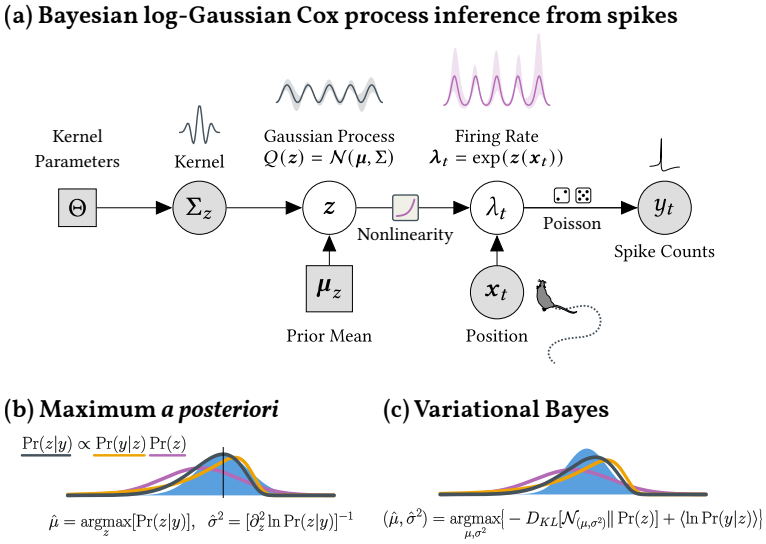

Figure 2: **A Bayesian model for firing-rate maps with spiking observations. (a)** A graphical diagram of the inference procedure. The prior mean and kernel are set externally. A log-Gaussian process parameterizes the inferred firing-rate map. Spiking observations are explained in terms of spatial tuning to location. **(b)** The posterior distribution over the log-rate map z is difficult to calculate directly. The maximum *a posteriori* (Methods: *Finding the posterior mode*) estimator approximates $\Pr(z)$ as Gaussian, with mean equal to the posterior mode, and covariance taken from the curvature at this mode (Methods: *Connection to the Laplace Approximation*). **(c)** Variational Bayesian inference finds a multivariate Gaussian model for the posterior on z by maximizing a lower-bound on the model likelihood. This can be more accurate when the posterior is skewed, and the same lower bound can be used to select hyperparameters.

matrix. That is, Σ_z should be positive definite: It should be symmetric, real-valued, and have all positive eigenvalues.

Grid cells exhibit periodic structure in their firing-rate maps. For our inference procedure to be sensitive to this, our kernel needs a periodic structure. However, not all radially-symmetric functions $\mathcal{K}(|\Delta \boldsymbol{x}|)$ yield a positive-definite kernel. Defining the kernel in terms of the autocorrelation of a known function can guarantee that it will be positive definite.

To find an appropriate kernel for grid cells, we consider an idealized hexagonal grid. A hexagonal map with period P can be defined as the sum of three cosine plane waves, rotated at $\pi/3$ radians from each-other:

$$\psi(\boldsymbol{x}) = \textstyle\sum_{i=0..2} \cos\left\{\frac{2\pi}{P}\left[x_1 \cos\left(\frac{\pi}{3}i\right) - x_2 \sin\left(\frac{\pi}{3}i\right)\right]\right\}, \quad \text{where } \boldsymbol{x} = \{x_1, x_2\} \in \mathbb{R}^2. \tag{9}$$

The autocorrelation function of this grid can serve as a suitable (positive definite) covariance kernel. This can be calculated by taking the average product of (9), evaluated at all pairs of points separated by a given distance $\Delta \boldsymbol{x} = \boldsymbol{x}_1 - \boldsymbol{x}_0$:

$$\mathcal{G}(\Delta \boldsymbol{x}) = \langle \psi(\boldsymbol{x}_0)\psi(\boldsymbol{x}_0 + \Delta \boldsymbol{x}) \rangle_{\boldsymbol{x}_0}.$$

The ideal grid kernel $\mathcal{G}(\Delta \boldsymbol{x})$ presumes a particular orientation of the grid map. To remove this assumption, we work with the radial autocorrelation "$\mathcal{R}(r)$", which averages over all possible orientations $\varphi \in [0, 2\pi)$ for a given radius $r = |\Delta \boldsymbol{x}|$:

$$\mathcal{R}(r) = \frac{1}{2\pi r}\int_0^{2\pi} \mathcal{G}\left[\Delta \boldsymbol{x}(r, \varphi)\right] d\varphi, \text{ where } \Delta \boldsymbol{x}(r, \varphi) = \Delta r \cdot \{\cos(\varphi), \sin(\varphi)\}^\top. \tag{10}$$

This integral evaluates to the zeroth-order Bessel function of the first kind,

$$\mathcal{R}(r) = J_0\left(\frac{2\pi}{P} r\right). \tag{11}$$

In the 2D Fourier domain with $r = |\Delta \boldsymbol{x}|$, the Fourier transform of (11) consists of a ring of all spatial-frequency components $\xi = \rho e^{i\omega}$ that match the spatial frequency $\rho = 1/P$ of the grid (Fig. 2d).

The kernel in (10) also assumes perfect long-range order. To be more conservative, we limit the kernel to interactions between adjacent grid fields by applying a circular window $\mathcal{W}(\Delta \boldsymbol{x})$ with radius $k_3 P/(2\pi)$, where $k_3 \approx 8.65$ is the third zero of J_0:

$$\mathcal{W}(\Delta \boldsymbol{x}) = \Theta\left(|\Delta \boldsymbol{x}| - \frac{P}{2\pi} k_3\right), \tag{12}$$

where $\Theta(\cdot)$ is the Heaviside step function. We also remove high spatial frequencies from the kernel by convolving it with a 2D Gaussian-blur kernel "$\mathcal{K}_{\sigma_f}$" with radius $\sigma_f = P/\pi$:

$$\mathcal{K}_{\sigma_f}(\Delta \boldsymbol{x}) = \frac{1}{P}\exp(-|\Delta \boldsymbol{x}|^2/\sigma_f^2). \tag{13}$$

Finally, we introduce a scale parameter "σ_0^2" and constant offset "c". These control the height of the kernel and the the variance of the mean (log) firing rate, respectively. The overall kernel "$\mathcal{K}_{\boldsymbol{\theta}}(\Delta \boldsymbol{x})$" is controlled by hyperparameters $\boldsymbol{\theta} = (P, \sigma_0^2, c)$, and is parameterized as:

$$\mathcal{K}_{\boldsymbol{\theta}}(\Delta \boldsymbol{x}) = \sigma_0^2\left[\mathcal{K}_{\sigma_f} * (\mathcal{W} \circ \mathcal{R})\right](\Delta x) + c \tag{14}$$

where $*$ denotes convolution and $\circ$ is pointwise (function) multiplication. We discuss hyperparameter selection in Section *Optimizing kernel hyperparameters*. Other parameterizations are possible, for example, one may want to parameterize the interaction radius, or allow for anisotropy in the grid.

Generally, one can construct suitable covariance kernels by computing the autocorrelation of a prototype firing-rate map, averaging to achieve any desired symmetries, and applying desired spatial or spectral windowing. If the

kernel is defined as a convolution over a regular grid, these operations can be computed quickly using the FFT. Since any linear combination of positive-definite kernels is also positive definite, complicated kernels can be constructed out of simple primitives.

Variational Inference

One can optimize the log-posterior (7) in $\boldsymbol{z}$ to obtain a smoothed firing-rate map. This is known as the Maximum *a posteriori* (MAP) estimator (Fig. 2b; Methods: *Finding the posterior mode*). The MAP estimator allows us to specify prior assumptions (e.g. smoothness and periodicity) by selecting the appropriate prior covariance Σ_z. However, it is important to assess our confidence in the resulting rate map, and to have a formal way of checking whether our prior is reasonable. Variational Bayesian methods provide a formal way to approximate posterior uncertainty, in the form of a Gaussian-process covariance function.

In variational Bayesian inference, we approximate the true posterior with a simpler distribution "$Q_{\boldsymbol{\phi}}(\boldsymbol{z})$" defined by some parameters $\boldsymbol{\phi}$. We use a multivariate Gaussian approximation here, so $\boldsymbol{\phi} = (\boldsymbol{\mu}, \Sigma)$ and $Q_{\boldsymbol{\phi}}$ has the log-probability density

$$\ln Q_{\boldsymbol{\phi}}(\boldsymbol{z}) = -\tfrac{1}{2}\{\ln|2\pi\Sigma| - (\boldsymbol{z}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\}. \tag{15}$$

Variational inference selects $\boldsymbol{\phi}$ by maximizing a quantity called the evidence lower bound. This is equivalent to simultaneously minimizing the Kullback-Leibler divergence "D_{KL}" from the prior to the posterior, while maximizing the expected log-likelihood (5) under $Q_{\boldsymbol{\phi}}$:

$$\boldsymbol{\phi} \leftarrow \operatorname*{argmax}_{\boldsymbol{\phi}} \left\{ -D_{\mathrm{KL}}\left[Q_{\boldsymbol{\phi}} \| \Pr(\boldsymbol{z})\right] + \langle \ln \Pr(\boldsymbol{k}|\boldsymbol{z}) \rangle \right\}, \tag{16}$$

where $\langle\cdot\rangle$ denotes expectation with respect to $Q_{\boldsymbol{\phi}}$.

The first term in (16) reflects the information gained by revising our estimates of $\boldsymbol{z}$ compared to our prior beliefs. Since both $Q_{\boldsymbol{\phi}}(\boldsymbol{z})$ and $\Pr(\boldsymbol{z})$ are multivariate Gaussian, this term has the closed form:

$$\begin{aligned} D_{\mathrm{KL}}\left[Q_{\boldsymbol{\phi}} \| \Pr(\boldsymbol{z})\right] = \tfrac{1}{2}\big\{&(\boldsymbol{\mu}-\boldsymbol{\mu}_z)^\top \Sigma_z^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_z) \\ &+ \operatorname{tr}(\Sigma_z^{-1}\Sigma) + \ln|\Sigma^{-1}\Sigma_z| - M\big\}. \end{aligned} \tag{17}$$

The second term in (16) is the expectation of our Poisson observation model (5) with respect to $Q_{\boldsymbol{\phi}}$:

$$\langle \ln \Pr(\boldsymbol{k}|\boldsymbol{z}) \rangle = \boldsymbol{\mu}^\top \boldsymbol{k} - \boldsymbol{n}^\top \langle \boldsymbol{\lambda} \rangle + \text{constant}. \tag{18}$$

We can write the overall objective "$\mathcal{L}$" to be maximized as:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\phi}) = -\tfrac{1}{2}\big\{&(\boldsymbol{\mu}-\boldsymbol{\mu}_z)^\top \Sigma_z^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_z) \\ &+ \operatorname{tr}(\Sigma_z^{-1}\Sigma) + \ln|\Sigma^{-1}\Sigma_z|\big\} \\ &+ \boldsymbol{\mu}^\top \boldsymbol{k} - \boldsymbol{n}^\top \langle \boldsymbol{\lambda} \rangle + \text{constant}. \end{aligned} \tag{19}$$

The term "$\operatorname{tr}(\Sigma_z^{-1}\Sigma)$" encourages the posterior covariance to be close to the prior, and the term "$-\ln|\Sigma|$" encourages the posterior to have large entropy.

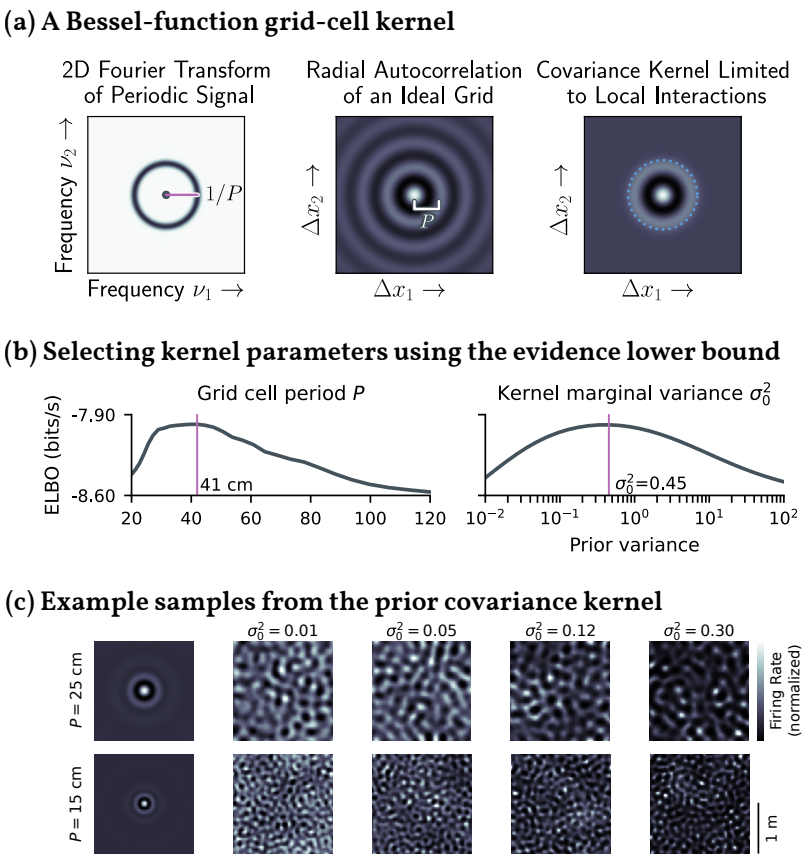

Figure 3: **Incorporating spatial-scale priors to infer grid-cell maps. (a)** For grid cells, we design a prior covariance kernel that is concentrated at the spatial frequencies contained in the periodic grid (left). The radial autocorrelation for an ideal rate map is shown in the center panel. To be more conservative, we window this kernel to allow only local interactions (right). **(b)** We can compare the Evidence Lower Bound ("ELBO", normalized here to units of bits per second) after optimizing the variational posterior for different kernels. This allows us to optimize kernel hyperparameters and infer properties of the grid-field map. The loss functions shown here are from fitting the dataset in Figure 1. **(c)** Example samples from the grid-cell prior. The period P controls the grid scale, and the prior variance σ_0^2 controls the variance of the log-rate, which affects the sharpness of the grid "fields". The radial prior lacks an orientation bias, hence the irregularity of these samples compared to true grid-cell maps.

An convenient property of the log-Gaussian-Poisson model is that the expected firing rate $\langle \boldsymbol{\lambda} \rangle$ required to calculate (19) has a simple closed form (c.f. Rule and Sanguinetti, 2018). Since we have assumed a multivariate Gaussian distribution for $\boldsymbol{z}$, and since $\boldsymbol{\lambda} = \exp(\boldsymbol{z})$, the firing-rate $\boldsymbol{\lambda}$ is log-normally distributed. The expectation $\langle \boldsymbol{\lambda} \rangle$ is the mean of this log-normal distribution, and has the closed-form expression

$$\langle \boldsymbol{\lambda} \rangle = \exp(\boldsymbol{\mu} + \tfrac{1}{2}\,\mathrm{diag}[\Sigma]). \tag{20}$$

To simplify notation, we define "$\bar{\boldsymbol{\lambda}}$" as the expected rate (with dependence on $\boldsymbol{\mu}$ and Σ implicit), corrected for the number of visits in each location, i.e. $\bar{\boldsymbol{\lambda}} = \boldsymbol{n} \circ \langle \boldsymbol{\lambda}(\boldsymbol{z}) \rangle$. We discuss numerical approaches for calculating (20) briefly in the next section, and in more detail in Methods: *Calculating the expected firing rate.*

Optimizing the variational posterior

With these preliminaries out of the way, we can now consider the derivatives of (19) in terms of $\boldsymbol{\mu}$ and Σ. These can be computed using modern automatic differentiation tools (e.g. Jax; Bradbury et al., 2018). However, substantial speedups are possible by considering the analytic forms of the derivatives, and identifying simpler ways to calculate them. The gradient and Hessian of (19) with respect to $\boldsymbol{\mu}$ are

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} \mathcal{L} &= -\Sigma_z^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_z) + \boldsymbol{k} - \bar{\boldsymbol{\lambda}} \\ \text{and} \quad \nabla_{\boldsymbol{\mu}} \nabla_{\boldsymbol{\mu}}^{\top} \mathcal{L} &= -\Sigma_z^{-1} - \mathrm{diag}[\bar{\boldsymbol{\lambda}}], \end{aligned} \tag{21}$$

respectively. The derivative of (19) in Σ is more involved. We derive this in Methods: *Derivatives*, Equations (36)-(38):

$$\partial_{\Sigma}[\mathcal{L}] = \tfrac{1}{2}\{\Sigma^{-1} - \Sigma_z^{-1} - \mathrm{diag}[\bar{\boldsymbol{\lambda}}]\}. \tag{22}$$

Optimizing the full $M \times M$ posterior covariance is impractical. Typically, one chooses a simpler parameterization. Combinations of low-rank factorizations and Toeplitz or circulant matrices are common (Jensen et al., 2021).

In our case, an *exact* low-dimensional parameterization of the variational posterior covariance is available (Seeger, 1999; Challis and Barber, 2013 Eq. (10)). Note that the stationary point of (22) occurs when $\Sigma^{-1} = \Sigma_z^{-1} + \mathrm{diag}[\bar{\boldsymbol{\lambda}}]$. This means that *all* variational posterior covariance matrices can be parameterized by a diagonal update $\mathrm{diag}[\bar{\boldsymbol{\lambda}}]$ to the prior precision matrix Σ_z^{-1}. We parameterize this update by the vector $\boldsymbol{q} = \{q_1, .., q_M\}$, and seek a self-consistent solution $\boldsymbol{q} = \bar{\boldsymbol{\lambda}}$:

$$\Sigma^{-1} = \Sigma_z^{-1} + \mathrm{diag}[\boldsymbol{q}]. \tag{23}$$

This models the posterior precision as a sum of the prior precision, plus information provided by the observations at each location.

We can obtain the gradient of $\mathcal{L}$ in $\boldsymbol{q}$ from (22) and (23) using the chain rule. After some calculation (Methods: *Derivatives*, Equations (38)-(40)), one finds:

$$\nabla_{\boldsymbol{q}} \mathcal{L} = \tfrac{1}{2}\,\mathrm{diag}\left\{\Sigma\,\mathrm{diag}[\bar{\boldsymbol{\lambda}} - \boldsymbol{q}]\,\Sigma\right\}. \tag{24}$$

This gradient is zero when $\boldsymbol{q} = \bar{\boldsymbol{\lambda}}$. If Σ is full rank, this zero is unique, and one may optimize $\boldsymbol{q}$ by ascending the much simpler gradient $\bar{\boldsymbol{\lambda}} - \boldsymbol{q}$, which has the same fixed point.

Overall, we can maximize our evidence lower bound (19) by alternatively updating $\boldsymbol{\mu}$ and $\boldsymbol{q}$. Updates to $\boldsymbol{\mu}$ are similar to finding the MAP estimator. For fixed $\boldsymbol{\mu}$, we use an iterative procedure that amounts to setting $\boldsymbol{q} \leftarrow \bar{\boldsymbol{\lambda}}$ repeatedly (Methods: *Iteratively estimating* $\boldsymbol{q}$).

There is one remaining difficulty to address. Calculating the expected firing rate (20) requires computing $\mathrm{diag}[\Sigma]$. These are the marginal variances of the firing-rate at each location. For the parameterization in Equation (23), one must compute

$$\mathrm{diag}[\Sigma] = \mathrm{diag}\left\{(\Sigma_z^{-1} + \mathrm{diag}[\boldsymbol{q}])^{-1}\right\}. \tag{25}$$

We calculate this using a low-rank approximation of the posterior covariance in Fourier space (Methods: *Working in a low-rank subspace*).

We summarize all steps of this iterative procedure in pseudocode in Algorithm 1. The key takeaways regarding the numeric implementation are this: Firstly, the posterior mean can be optimized readily using Newton-Raphson iteration, in much the same way as one might estimate the posterior mode for a log-Gaussian-Poisson generalized linear model. Secondly, the ideal parameterization of the variational posterior covariance takes the form of a diagonal update to the precision matrix, which reflects the amount of information available at each spatial location. This can be updated by a relatively straightforward fixed-point iteration reminiscent of the the Laplace approximation (Methods: *Laplace Approximation*). This perspective therefore provides an accessible introduction to Gaussian-process methods to practitioners familiar with regularized generalized linear models, and intuition for designing efficient algorithms.

Optimizing kernel hyperparameters

The prior covariance kernel in Equation (14) depends on three unknown hyperparameters: period "P", scale "σ_0^2", and mean offset "c". The variational Bayesian framework provides a principled way to optimize these.

To evaluate the quality of hyperparameters $\boldsymbol{\theta} = (P, \sigma_0^2, c)$, one can first optimize the variational posterior using the kernel determined by $\boldsymbol{\theta}$. After optimizing $Q_{\boldsymbol{\phi}}(\boldsymbol{z})$, the value of the evidence lower bound (16) also lower-bounds the likelihood of the data for the chosen hyperparameters. This allows one to compare the quality of the fit for different choices of $\boldsymbol{\theta}$ (e.g. Fig. 3b).

We optimized $\boldsymbol{\theta}$ using a hill-climbing grid search. We start with a heuristic guess (*Methods:* Heuristic initialization) and recursively search nearby values of $\boldsymbol{\theta}$ until we find a global maximum. The solutions for the parameters of the variational posterior from previous choices of $\boldsymbol{\theta}$ can be

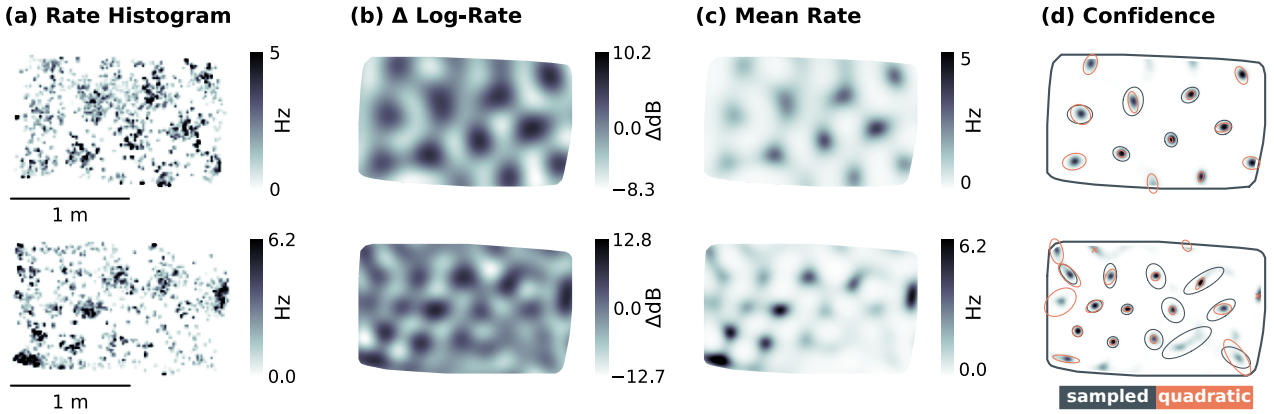

Figure 4: **Flexible smoothing of grid-cell firing-rate maps using LGCP regression. (a)** Rate histograms from two different cells, one with activity throughout the arena (top) and another with uneven activity (bottom), from Krupic et al. (2018). **(b)** Posterior log-rate map from LGCP inference using the optimized grid-cell scale and prior variance. (Background variations in log firing-rate not included). **(c)** Expected firing rate calculated using (20) from the variational posterior. **(d)** Probability density of grid-field peaks (shaded), as well as 95% confidence intervals for field location calculated either using a locally-quadratic approximation (ochre) (30) or sampling (dashed purple). For the quadratic case, peaks whose 95% confidence ellipse exceeded the grid period have been removed (red "x"). The LGCP estimator estimator smooths data out to nearest-neighbor interactions without merging adjacent grid fields, and adapts to the level of information available in each portion of the arena.

used as the starting point for optimizing new $\boldsymbol{\theta}$, reducing computational cost.

Sampling spatial statistics

Once obtained, the Gaussian-process posterior can be used to sample the distribution of likely firing-rate maps. For example, one may wish to obtain the probability distribution of the peaks of individual grid fields (Fig. 4d).

Given a Gaussian posterior $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, one can draw samples as $\boldsymbol{z} \leftarrow \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{\eta}_M$ where $\boldsymbol{\eta}_M \sim \mathcal{N}(0, \mathrm{I}_M)$ is a vector of M Gaussian random numbers with unit-variance and zero-mean. However, obtaining $\Sigma^{1/2}$ becomes impractical when M is large. Sampling in the low-rank ($D<M$) space $\tilde{\boldsymbol{z}} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$ is more efficient (Methods: *Working in a low-rank subspace*). Samples can be drawn as

$$\boldsymbol{z} \leftarrow \tilde{\mathrm{R}}\left(\tilde{\boldsymbol{\mu}} + \tilde{\Sigma}^{1/2}\,\boldsymbol{\eta}_D\right), \tag{26}$$

where $\tilde{\mathrm{R}}$ maps samples from the low-rank subspace into the full (spatial) representation, and is described in (44) and (45). The factor $\tilde{\Sigma}^{1/2}$ can be calculated as $\tilde{\Sigma}^{1/2} = \tilde{\mathrm{R}}\ \mathrm{chol}[\mathrm{diag}[\tilde{\boldsymbol{\xi}}^{-1}] + \boldsymbol{X}\boldsymbol{X}^\top]^{-1}$, where $\boldsymbol{X} = \tilde{\mathrm{R}}^\top\ \mathrm{diag}[\bar{\boldsymbol{\lambda}}^{1/2}]$. We describe this in more detail in Methods: *Calculating the expected firing rate*; (46).

Figures 4d uses sampling to visualize uncertainty in grid-field locations. We generated a "peak density" map by plotting the fraction of samples that contain a local maximum within a radius of $P/2$, where P is the grid cell's spatial period. We segmented the arena into Voronoi cells associated with each grid field, and calculated 95% confidence ellipses by fitting a 2D Gaussian to each segmented grid field's peak distribution.

Peak-location confidence intervals

For well-identified grid fields, one can calculate confidence intervals from the posterior distribution using a locally quadratic approximation. Consider a local maximum in the posterior mean $\boldsymbol{\mu}(\boldsymbol{x})$ at location $\boldsymbol{x}_0$. How much does $\boldsymbol{x}_0$ change if a perturbation $\boldsymbol{\varepsilon}(\boldsymbol{x}) \sim \mathcal{N}[0, \Sigma(\boldsymbol{x}, \boldsymbol{x}')]$, sampled from the posterior covariance, is added to $\boldsymbol{\mu}(\boldsymbol{x})$?

This can be calculated via a Taylor expansion in $\Delta_{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{x}_0$ of $\boldsymbol{\mu}(\boldsymbol{x})$ at $\boldsymbol{x}_0$. The slope at $\boldsymbol{x}_0$ is zero, since it is a local maximum, so a Taylor expansion out to second order has only 0th- and 2nd-order (curvature) terms. The curvature in $\boldsymbol{x}$ is defined by the Hessian matrix $\mathrm{H}_{\boldsymbol{z}} := \nabla_{\boldsymbol{x}}\boldsymbol{\mu}(\boldsymbol{x}_0)\nabla_{\boldsymbol{x}}^\top$. Out to second order, then, our grid-field log-firing-rate is:

$$\boldsymbol{\mu}(\boldsymbol{x}) \approx \boldsymbol{\mu}(\boldsymbol{x}_0) + \tfrac{1}{2}\Delta_{\boldsymbol{x}}^\top \mathrm{H}_{\boldsymbol{z}} \Delta_{\boldsymbol{x}}. \tag{27}$$

Now, add a first-order approximation $\boldsymbol{\varepsilon}(\boldsymbol{x}) \approx \boldsymbol{\varepsilon}(\boldsymbol{x}_0) + \mathrm{J}_{\boldsymbol{\varepsilon}}^\top \Delta_{\boldsymbol{x}}$ of the noise (posterior uncertainty) to (27), where $\mathrm{J}_{\boldsymbol{\varepsilon}} := \nabla_{\boldsymbol{x}}\boldsymbol{\varepsilon}(\boldsymbol{x}_0)$ is the gradient of $\boldsymbol{\varepsilon}$ at $\boldsymbol{x}_0$:

$$\boldsymbol{z}(\boldsymbol{x}) \approx \boldsymbol{z}(\boldsymbol{x}_0) + \boldsymbol{\varepsilon}(\boldsymbol{x}_0) + \mathrm{J}_{\boldsymbol{\varepsilon}}^\top \Delta_{\boldsymbol{x}} + \tfrac{1}{2}\Delta_{\boldsymbol{x}}^\top \mathrm{H}_{\boldsymbol{z}} \Delta_{\boldsymbol{x}}. \tag{28}$$

Setting the derivative of (28) in $\Delta_{\boldsymbol{x}}$ to zero and solving for $\Delta_{\boldsymbol{x}}$, we find that:

$$\Delta_{\boldsymbol{x}} = -\mathrm{H}_{\boldsymbol{z}}^{-1}\left[\nabla_{\boldsymbol{x}}\boldsymbol{\varepsilon}(\boldsymbol{x}_0)\right]. \tag{29}$$

We can construct a covariance matrix "$\Sigma_{\Delta_{\boldsymbol{x}}}$" for the location of the peak using (29).

$$\begin{aligned} \Sigma_{\Delta_{\boldsymbol{x}}} &= \langle \Delta_{\boldsymbol{x}} (\Delta_{\boldsymbol{x}})^\top \rangle \\ &= \mathrm{H}_{\boldsymbol{z}}^{-1} \left\langle \mathrm{J}_{\boldsymbol{\varepsilon}} \mathrm{J}_{\boldsymbol{\varepsilon}}^\top \right\rangle \mathrm{H}_{\boldsymbol{z}}^{-1} \\ &= \mathrm{H}_{\boldsymbol{z}}^{-1} \nabla_{\boldsymbol{x}} \Sigma(\boldsymbol{x}_0, \boldsymbol{x}_0) \nabla_{\boldsymbol{x}}^\top \mathrm{H}_{\boldsymbol{z}}^{-1}. \end{aligned} \tag{30}$$

Equation (30) depends on the posterior covariance, which is impractical to construct for large environments. Instead, we use the low-rank approximation $\Sigma \approx \tilde{\mathrm{R}}\tilde{\Sigma}\tilde{\mathrm{R}}^\top$ as

in (26), where $\tilde{\Sigma} \in \mathbb{R}^{D \times D}$ is the low-rank covariance and $\tilde{\mathrm{R}} \in \mathbb{R}^{L^2 \times D}$ is a semi-orthogonal operator defining our low-rank basis. We can use the Cholesky decomposition to obtain $\mathrm{Q} \in \mathbb{R}^{D \times X}$ such that $\tilde{\Sigma} = \mathrm{Q}\mathrm{Q}^\top$ and calculate (30) as

$$\Sigma_{\Delta_{\boldsymbol{x}}} = \mathrm{B}\mathrm{B}^\top, \text{ where } \mathrm{B} = \mathrm{H}_{\boldsymbol{z}}^{-1}[\nabla_{\boldsymbol{x}}\tilde{\mathrm{R}}\mathrm{Q}](\boldsymbol{x}_0). \tag{31}$$

Figure 4d compares field-location confidence intervals obtained either by sampling, or quadratic approximation. These methods agree for well-localized peaks. For poorly-localized peaks, the quadratic approximation yields more conservative (i.e. looser) confidence intervals.

Discussion

We have introduced a variational Bayesian approach to analyzing data from grid cells. We focused on challenges associated with working with grid cells in larger environments, and prioritized computational efficiency to facilitate exploratory analysis of large datasets.

Out method incorporates prior assumptions of grid-cell periodicity, is computationally efficient, yields approximate confidence intervals, and provides a formal way to compare different prior assumptions (i.e. optimize the kernel hyperparameters).

Technical remarks Our approach connects pseudo-point methods for log-Gaussian Cox processes and Poisson generalized linear models. We used a simple yet expressive parameterization of the variational posterior covariance in terms of a diagonal precision update. The resulting inference procedure is tantamount to repeatedly applying Newton-Raphson for fitting GLMs (sometimes called iteratively re-weighted least-squares), incorporating a covariance correction to the expected firing rate. For a fixed posterior covariance, this optimization is convex and exhibits quadratic convergence to a global optimum.

We used the fact that translation-invariant prior covariance kernels amount to convolutions, which can be calculated as point-wise multiplication in Fourier space. We optimized the variational posterior in a low-rank spatial frequency domain, retaining only $D < L^2$ non-negligible components.

Caveats The posterior covariance of a GP regression is only interpretable if the kernel is a true model for the data's correlation structure, and only after optimizing the kernel's hyperparameters. Our design of a periodic prior mitigates this issue, but does not eliminate it. If the grid structure breaks down in a portion of the arena, then a nearest-neighbour periodic kernel becomes mis-specified for this region. Optimizing the kernel hyperparameters by maximizing (16) can accommodate some local mismatch by increasing the prior marginal variance σ_0^2. This relaxes the influence of the prior, allowing the locally non-grid-like observations to more strongly influence the GP regression. Research into faster methods to allow non-stationary (or in our case, spatially inhomogeneous) kernels is ongoing (e.g. Paun et al. 2023), but we are aware of no advances suitable for the specific problem considered here. In principle, one could merge maps from different priors in different regions of the arena *post hoc*, but we will leave such explorations for the future.

Instead, we recommend applying robust controls for hypothesis testing, such as shuffle tests that remove any purported effect form the underlying data time series. Such tests can be used to measure the variance in the GP regression as an estimator of peak location, and complements the GP posterior covariance (which measures the uncertainty assuming that the prior assumptions are true). These are computationally intensive but feasible, requiring resources similar to shuffle controls for a large GLM regressions.

Generality and applicability The GP estimators described here provide a useful extension of traditional methods. Our method combines aspects of traditional kernel-density smothers and Poisson generalized linear models with a principled (periodic) prior. The advantages of GP regression over the KDE are that GP regression (1) adapts to smooth more where data are limited (2) can be used to infer parameters of the spatial correlation structure (e.g. identify grid period), and (3) can employ priors with nearest-neighbor interactions, allowing the smoothing radius to exceed the size of a single grid field.

Our numerical approach is broadly applicable to problems that are large enough for the cubic complexity of naïve GP estimators to become burdensome, but which are unsuitable for existing sparse or low-rank solutions for large problems. Beyond grid cells, it can be applied to intermediate-size spatiotemporal inference problems with prior kernels with many negligible frequency components.

Although we focused on the Poisson observation model, our approach generalizes to other observation models in the natural exponential family. The main caveat is other observation models may lack a convenient closed-form expression for the expected firing rate, and these terms may need to be approximated via sampling or another more intensive procedure.

Future work The methods we have describe are suitable as a drop-in replacement for the smoothing and kernel density estimation approaches currently used to analyze grid cell activity. In addition to providing a principled smoother that is aware of a grid cell's spatial scale, this approach provides approximate measures of the confidence of the smoothed map. These algorithms have considerable potential, and could be extended e.g. to incorporate tuning to additional behavioral covariates, or additional latent rate fluctuations that are correlated in time.

We plan to apply these methods in our own work. We hope that others will as well, in addition to building upon these approaches to design new algorithms for spatiotemporal statistics in the study of spatial navigation.

Code and data availability

We have provided a reference implementation of these methods, along with tutorial notebooks, online at Github (http://github.com/michaelerule/lgcpspatial).

Acknowledgments

M.E.R. is supported by a Leverhulme and Isaac Newton Trust fellowship ECF-2020-352. P.C.V is supported by MRC, Frank Elmore Fund, and University of Cambridge School of Clinical Medicine. J.K. is a Wellcome Trust/Royal Society Sir Henry Dale Fellow (206682/Z/17/Z) and is supported by Dementia Research Institute (DRICAMKRUPIC18/19), Isaac Newton Trust/Wellcome Trust ISSF/University of Cambridge Joint Research Grant, Kavli Foundation Dream Team project (RG93383), Isaac Newton Trust [17.37(t)], and NVIDIA Corporation. T.O'L. is supported by ERC grant 716643 FLEXNEURO and HFSP grant RGY0069/2017.

Methods

Finding the posterior mode

One approach to estimating the firing rate map is to find the $\boldsymbol{z}$ that maximizes the log-posterior (7). This is the *Maximum a Posteriori* (MAP) estimator, and can be solved via gradient ascent. The gradient of (7) in $\boldsymbol{z}$ is:

$$\nabla_{\boldsymbol{z}} \ln \Pr(\boldsymbol{z}|\boldsymbol{k}) = -\Sigma_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z) + \boldsymbol{k} - \tilde{\boldsymbol{\lambda}}, \tag{32}$$

where we have defined $\tilde{\boldsymbol{\lambda}} := \boldsymbol{n} \circ \boldsymbol{\lambda}$ as the vector of estimated firing-rates weighted by the number of visits to each location $\boldsymbol{n}$ ($\circ$ denotes element-wise multiplication).

The (negative) log-posterior (7) is convex, and well approximated as locally-quadratic. The Newton-Raphson method is applicable, and much faster than gradient descent. This uses the curvature (Hessian, "$\nabla_{\boldsymbol{z}}\nabla_{\boldsymbol{z}}^\top$") of (7) in $\boldsymbol{z}$:

$$\nabla_{\boldsymbol{z}}\nabla_{\boldsymbol{z}}^\top \ln \Pr(\boldsymbol{z}|\boldsymbol{k}) = -\Sigma_z^{-1} - \operatorname{diag}[\tilde{\boldsymbol{\lambda}}]. \tag{33}$$

On each iteration, Newton-Raphson must solve the update

$$\begin{aligned} \hat{\boldsymbol{z}}_{t+1} &\leftarrow \hat{\boldsymbol{z}}_t - \mathrm{H}_{\hat{\boldsymbol{z}}}^{-1}\mathrm{J}_{\hat{\boldsymbol{z}}}, \text{ where} \\ \mathrm{J}_{\hat{\boldsymbol{z}}} &:= \nabla_{\hat{\boldsymbol{z}}} \ln \Pr(\hat{\boldsymbol{z}}|\boldsymbol{k}) \text{ and} \\ \mathrm{H}_{\hat{\boldsymbol{z}}} &:= \nabla_{\hat{\boldsymbol{z}}}\nabla_{\hat{\boldsymbol{z}}}^\top \ln \Pr(\hat{\boldsymbol{z}}|\boldsymbol{k}). \end{aligned} \tag{34}$$

The Hessian and Jacobian for the MAP estimator are the same as those for the variational posterior mean (21), but with the expected rate $\bar{\boldsymbol{\lambda}}$ replaced by the point estimate $\tilde{\boldsymbol{\lambda}}$.

To compute $\mathrm{H}_{\hat{\boldsymbol{z}}}^{-1}\mathrm{J}_{\hat{\boldsymbol{z}}}$ quickly in high dimensions, we used an inexact Newton-Raphson method (Dembo et al., 1982) that approximates $\mathrm{H}_{\hat{\boldsymbol{z}}}^{-1}\mathrm{J}_{\hat{\boldsymbol{z}}}$ each iteration of (34) via a preconditioned Krylov method (Methods: *Newton-Krylov methods*)

Newton-Krylov methods

Naïve algorithms for multiplying or inverting dense matrices have cubic complexity, making expressions such as (34) computationally prohibitive in higher dimensions (c.f. Liu et al., 2020). Thankfully, modern Krylov-subspace algorithms for solving large linear systems $\boldsymbol{A}^{-1}\boldsymbol{v}$ only require a function that can compute the matrix-vector-product $\boldsymbol{u} \mapsto \boldsymbol{A}\boldsymbol{u}$. This can be computed quickly if our matrix $\boldsymbol{A}$ has special structure (Chan and Jackson, 1984; Brown and Saad, 1990; see Knoll and Keyes, 2004 for review). Fast solutions exist if the covariance kernel (or its inverse) is sparse (e.g. Luttinen and Ilin, 2009; Kiiveri and De Hoog, 2012; Gal et al., 2014; Cseke et al., 2016) or Toeplitz/circulant (eg. Jensen et al., 2021).

Algorithms combining Newton-Raphson iteration with Krylov methods were first developed for very large, sparse, Gaussian process models (Kiiveri and De Hoog, 2012; Cseke et al., 2016), but apply to any problem where $\boldsymbol{A}\boldsymbol{u}$ can be computed quickly but $\boldsymbol{A}^{-1}\boldsymbol{u}$ is impractical. In our case, our prior covariance Σ_z is a convolution, and we calculate $\Sigma_z^{-1}\boldsymbol{u}$ quickly as point-wise multiplication in in the spatial-frequency domain (Methods: *Working in a low-rank subspace*). We can therefore calculate the Hessian-vector product $\mathrm{H}_{\hat{\boldsymbol{z}}}\boldsymbol{u}$ quickly, which allows us to expediently calculate $\mathrm{H}_{\hat{\boldsymbol{z}}}^{-1}\mathrm{J}_{\hat{\boldsymbol{z}}}$ using a Krylov-subspace solver.

In our tests, we found that Scipy's (Virtanen et al., 2020) implementation of the minimum residual Krylov-subspace algorithm (MINRES; Paige and Saunders, 1975) provided the best balance of speed and stability. To make complex-valued spatial-frequency components compatible with Krylov solvers designed for real-valued matrices, we used the Hartley transform rather than the Fourier transform (Methods: *Working in a low-rank subspace*).

Krylov-subspace algorithms benefit from a preconditioner "M" that approximates the inverse ($\mathrm{H}_{\hat{\boldsymbol{z}}}^{-1}$ in our case). We used the prior covariance kernel for this, $M = \Sigma_z$, also computed as a convolution via pointwise multiplication in a low-rank spatial-frequency domain (see Algorithm 1).

Connection to the Laplace Approximation

It is instructive to compare the Laplace approximation for Poisson generalized linear models to the variational methods described here. The Laplace approximation (Fig. 2b) models the posterior uncertainty in the MAP-estimated log-rate $\hat{z}$ as a Gaussian centered at $\boldsymbol{\mu} = \hat{z}$, and with the covariance equal to the negative-inverse of the Hessian (33) evaluated at $\hat{z}$:

$$\begin{aligned} \Pr(\hat{z}) &\approx \mathcal{N}(\boldsymbol{\mu} = \hat{z}, \hat{\Sigma}) \\ \hat{\Sigma}^{-1} &= \Sigma_z^{-1} + \operatorname{diag}[\tilde{\boldsymbol{\lambda}}(\hat{z})]. \end{aligned} \tag{35}$$

Intuitively, (35) says that the Laplace approximation models the posterior precision as a sum of the prior precision, plus a diagonal matrix representing new information from the spikes at each location.

Note the similarity between the derivatives of the variational mean (21) and those of the MAP estimator (32) and (33), and the similarity between the Laplace-approximated posterior variance (35) and the covariance update for variational Bayes (23) and (48). Optimizing the variational mean is tantamount to calculating the MAP estimator using the expected rate $\langle \boldsymbol{\lambda} \rangle_z$ rather than a point estimate $\boldsymbol{\lambda}$. Likewise, updating the posterior covariance is tantamount to applying the Laplace approximation that the variational mean, again using the expected rate rather than a point estimate.

Derivatives

In this section, we derive the gradients of the evidence lower bound (19) with respect to Σ and $\boldsymbol{q}$ (Equations (22) and (24) in the main text, respectively).

First, we obtain the derivative of the evidence lower bound (19) with respect to the posterior covariance matrix Σ. Consider the derivative of the term term $\boldsymbol{n}^\top \langle \boldsymbol{\lambda} \rangle$ with respect to individual elements Σ_{ij}. We use Einstein summation notation, wherein sums over repeated indecies is implied:

$$\begin{aligned}
\partial_{\Sigma_{ij}} [\boldsymbol{n}^\top \langle \boldsymbol{\lambda} \rangle] &= \partial_{\Sigma_{ij}} [n_k \langle \lambda_k \rangle] \\
&= n_k \partial_{\Sigma_{ij}} \langle \lambda_k \rangle \\
&= n_k \partial_{\Sigma_{ij}} \exp(\mu_k + \tfrac{1}{2}\Sigma_{kk}) \\
&= n_k \exp(\mu_k + \tfrac{1}{2}\Sigma_{kk}) \partial_{\Sigma_{ij}} (\mu_k + \tfrac{1}{2}\Sigma_{kk}) \\
&= \tfrac{1}{2} n_k \langle \lambda_k \rangle \partial_{\Sigma_{ij}} \Sigma_{kk} \\
&= \tfrac{1}{2} n_k \langle \lambda_k \rangle \delta_{ik} \delta_{jk} \\
&= \tfrac{1}{2} n_i \langle \lambda_i \rangle \delta_{ij} \\
\Rightarrow \quad \partial_\Sigma [\boldsymbol{n}^\top \langle \boldsymbol{\lambda} \rangle] &= \tfrac{1}{2} \operatorname{diag}[\boldsymbol{n} \circ \boldsymbol{\lambda}],
\end{aligned} \tag{36}$$

where δ_{ab} is the Kronecker delta (1 if $a = b$ and 0 otherwise).

The derivative of the term $\operatorname{tr}(\Sigma_z^{-1}\Sigma)$ in (19) is given by identity (100) in The Matrix Cookbook (Petersen et al., 2008), and is Σ_z^{-1}. The derivative of the term $\ln|\Sigma|$ is given by identity (57) in The Matrix Cookbook (Petersen et al., 2008), (assuming $|\Sigma| \neq 0$), and is Σ^{-1}. Overall, then, the derivative of (19) in Σ is:

$$\begin{aligned}
\partial_\Sigma \mathcal{L} &= \partial_\Sigma \left\{ \tfrac{1}{2} \left[\ln|\Sigma| - \operatorname{tr}(\Sigma_z^{-1}\Sigma) \right] - \boldsymbol{n}^\top \langle \boldsymbol{\lambda} \rangle \right\} \\
&= \tfrac{1}{2} \left\{ \Sigma^{-1} - \Sigma_z^{-1} - \operatorname{diag}[\boldsymbol{n} \circ \boldsymbol{\lambda}] \right\}.
\end{aligned} \tag{37}$$

Let $\bar{\boldsymbol{\lambda}} = \boldsymbol{n} \circ \boldsymbol{\lambda}$. For the parameterization $\Sigma^{-1} = \Sigma_z^{-1} + \operatorname{diag}[\boldsymbol{q}]$, (37) simplifies to:

$$\begin{aligned}
\partial_\Sigma \mathcal{L} &= \tfrac{1}{2} \left\{ \Sigma^{-1} - \Sigma_z^{-1} - \operatorname{diag}[\bar{\boldsymbol{\lambda}}] \right\} \\
&= \tfrac{1}{2} \left\{ (\Sigma_z^{-1} + \operatorname{diag}[\boldsymbol{q}]) - \Sigma_z^{-1} - \operatorname{diag}[\bar{\boldsymbol{\lambda}}] \right\} \\
&= \tfrac{1}{2} \left\{ \operatorname{diag}[\boldsymbol{q}] - \operatorname{diag}[\bar{\boldsymbol{\lambda}}] \right\} \\
&= \tfrac{1}{2} \operatorname{diag}[\boldsymbol{q} - \bar{\boldsymbol{\lambda}}]
\end{aligned} \tag{38}$$

We can now obtain the derivative of the evidence lower bound (19) in $\boldsymbol{q}$ via the chain rule, $\partial_{q_k} \mathcal{L} = [\partial_{\Sigma_{ij}} \mathcal{L}][\partial_{q_k} \Sigma_{ij}]$. To obtain $\partial_{q_k} \Sigma_{ij}$, we will need identity (59) in The Matrix Cookbook (Petersen et al., 2008), which provides the chain rule for the derivative of the inverse of a matrix, $\partial_x [\boldsymbol{A}(x)^{-1}] = -\boldsymbol{A}^{-1} [\partial_x \boldsymbol{A}] \boldsymbol{A}^{-1}$. We will let $\boldsymbol{A} = \Sigma^{-1}$:

$$\begin{aligned}
\partial_{q_k} \Sigma_{ij} &= \partial_{q_k} [\boldsymbol{A}^{-1}]_{ij} \\
&= [-\boldsymbol{A}^{-1} (\partial_{q_k} \boldsymbol{A}) \boldsymbol{A}^{-1}]_{ij} \\
&= -[\Sigma (\partial_{q_k} \Sigma^{-1}) \Sigma]_{ij} \\
&= -[\Sigma (\partial_{q_k} [\Sigma_z^{-1} + \operatorname{diag}[\boldsymbol{q}]]) \Sigma]_{ij} \\
&= -\Sigma_{ia} (\partial_{q_k} \operatorname{diag}[\boldsymbol{q}])_{ab} \Sigma_{bj} \\
&= -\Sigma_{ia} (\delta_{ka} \delta_{kb}) \Sigma_{bj} \\
&= -\Sigma_{ik} \Sigma_{kj}.
\end{aligned} \tag{39}$$

Combining (38) and (39) gives:

$$\begin{aligned}
\partial_{q_k} \mathcal{L} &= -\tfrac{1}{2} \operatorname{diag}[\boldsymbol{q} - \bar{\boldsymbol{\lambda}}]_{ij} \Sigma_{ik} \Sigma_{kj} \\
&= -\tfrac{1}{2} \Sigma_{ki} \operatorname{diag}[\boldsymbol{q} - \bar{\boldsymbol{\lambda}}]_{ij} \Sigma_{jk} \\
&= -\tfrac{1}{2} [\Sigma \operatorname{diag}[\boldsymbol{q} - \bar{\boldsymbol{\lambda}}] \Sigma]_{kk} \\
\Rightarrow \partial_{\boldsymbol{q}} \mathcal{L} &= -\tfrac{1}{2} \operatorname{diag}\{\Sigma \operatorname{diag}[\boldsymbol{q} - \bar{\boldsymbol{\lambda}}] \Sigma\}
\end{aligned} \tag{40}$$

Working in a low-rank subspace

Working in a low-rank subspace can make large problems tractable. We first find a low-rank approximation of the prior covariance Σ_z, and then perform inference within this subspace.

The prior covariance Σ_z is defined by a convolution kernel. The components of this kernel "$\boldsymbol{\xi}$" in spatial-frequency (Fourier) space are the eigenvalues of Σ_z:

$$\Sigma_z = \mathrm{F} \operatorname{diag}[\boldsymbol{\xi}] \, \mathrm{F}^\dagger, \tag{41}$$

where F is the unitary Fourier transform (which is its own inverse) and $\dagger$ denotes the conjugate (Hermitian) transpose.

In practice, many of these spatial frequency components will be close to zero. These are frequencies where the prior assigns very little probability. We work in a low-rank space consisting only of those directions in Σ_z where the prior has assigned non-negligible variance. We retain the $D \leq L^2$ components "$\tilde{\boldsymbol{\xi}}$" whose magnitude in the prior covariance kernel is at least 10% of the eigenvalue of the prior covariance with the largest magnitude "$\xi_{\max}$":

$$\tilde{\boldsymbol{\xi}} = \{\xi_m \text{ such that } |\xi_m| > 0.1 |\xi_{\max}|\}. \tag{42}$$

The low-rank approximation to the posterior covariance can then be calculated as

$$\begin{aligned}\Sigma &= (\Sigma_z^{-1} + \mathrm{diag}[\bar{\boldsymbol{\lambda}}])^{-1} \\ &= (\mathrm{F}\,\mathrm{diag}[\boldsymbol{\xi}^{-1}]\,\mathrm{F}^\dagger + \mathrm{diag}[\bar{\boldsymbol{\lambda}}])^{-1} \\ &= \mathrm{F}(\mathrm{diag}[\boldsymbol{\xi}^{-1}] + \mathrm{F}^\dagger\,\mathrm{diag}[\bar{\boldsymbol{\lambda}}]\,\mathrm{F})^{-1}\mathrm{F}^\dagger \\ &\approx \tilde{\mathrm{F}}(\mathrm{diag}[\tilde{\boldsymbol{\xi}}^{-1}] + \tilde{\mathrm{F}}^\dagger\,\mathrm{diag}[\bar{\boldsymbol{\lambda}}]\,\tilde{\mathrm{F}})^{-1}\tilde{\mathrm{F}}^\dagger,\end{aligned} \tag{43}$$

where $\tilde{\mathrm{F}}$ is the (unitary) Fourier transform retaining only the non-negligible components $\tilde{\boldsymbol{\xi}}$. $\tilde{\mathrm{F}}$ is not invertable, but since it is semi-orthogonal, $\tilde{\mathrm{F}}^\dagger$ is its pseudoinverse.

Note that the Kullback-Leibler divergence contribution to the evidence lower bound (17) contains a constant factor $-\frac{1}{2}M$ that depends on the number of dimensions $M = L^2$ in the our multivariate-Gaussian prior. When working in a low-rank $D < M$ subspace, this term should be replaced by $-\frac{1}{2}D$ to ensure that the evidence lower-bound can be compared between models with low-rank subspaces of different ranks.

Fourier coefficients can take on complex values. This creates compatibility and performance issues with standard numerical linear algebra software. To address this, we use a real-valued relative of the Fourier transform called the Hartley transform (Hartley, 1942).

We denote the Hartley transform as R, and the transform with negligible frequencies discarded as $\tilde{\mathrm{R}}$. The Hartley transform is calculated by summing the real and imaginary components of the Fourier transform

$$\mathrm{R} = \Re(\mathrm{F}) + \Im(\mathrm{F}). \tag{44}$$

If F is the the unitary Fourier transform, then R is also unitary.

Equations in (41)-(46) work similarly with the Hartley transform, replacing F with R, and replacing Hermitian transposes with ordinary transposes. Since the (circulant) prior Σ_z is symmetric, its Fourier coefficients $\boldsymbol{\xi}$ are real-valued, and the Hartley-transform coefficients for the prior are identical to the Fourier coefficients. Overall, (43) calculated using the Hartley transform can be written as

$$\Sigma \approx \tilde{\mathrm{R}}(\mathrm{diag}[\tilde{\boldsymbol{\xi}}^{-1}] + \tilde{\mathrm{R}}^\top\,\mathrm{diag}[\bar{\boldsymbol{\lambda}}]\,\tilde{\mathrm{R}})^{-1}\tilde{\mathrm{R}}^\top. \tag{45}$$

We denote this approximation as $\tilde{\Sigma}$.

The frequency-subspace representation simplifies some of the matrix calculations. Let L be the size of the environment, and let L^2 be the total number of spatial bins. A length-M vector can be converted into frequency space using the 2D fast Fourier transform, which costs $\mathcal{O}(L^2\log(L))$. If we retain only D components, the relevant transform has dimensions L^2D and the cost is $\mathcal{O}(L^2D)$. Ordinary matrix multiplication can outperform the FFT when $D \sim \mathcal{O}(\log(L))$. Since grid cells display only a narrow range of spatial scales, D can be small, and the complexity of each optimization iteration is competitive with simpler estimators.

We perform most calculations in this low-rank space, and never explicitly construct the full $L^2 \times L^2$ posterior covariance. The only calculation which cannot be performed in the low-rank space is the calculation of the expected firing-rates at each location, $\bar{\boldsymbol{\lambda}}$, which we address in *Methods: Calculating the expected firing rate*. This has complexity $\mathcal{O}(D^2(L^2 + D))$ owing to the cubic complexity of some of the matrix operations.

When operating in a low-rank subspace, it is important that the mean of the variational posterior also be expressed in this subspace. Leaving $\boldsymbol{\mu}$ in the full-rank space creates a poorly-conditioned problem, since several directions will be ignored when calculating gradients using low-rank approximations.

Calculating the expected firing rate

For Gaussian $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and exponential firing-rate nonlinearity, the firing rate $\boldsymbol{\lambda} = \exp(\boldsymbol{z})$ is log-normally distributed. The mean of this distribution, $\langle\boldsymbol{\lambda}\rangle$, has the closed-form expression $\exp(\boldsymbol{\mu} + \frac{1}{2}\,\mathrm{diag}[\Sigma])$ (c.f. Rule and Sanguinetti, 2018). Evaluating this expression requires the diagonal of the posterior covariance matrix.

In the low-rank subspace (45), these diagonal elements Σ_{ii} can be calculated with the following procedure:

$$\begin{aligned}\boldsymbol{X} &\leftarrow \tilde{\mathrm{R}}^\top\,\mathrm{diag}[\bar{\boldsymbol{\lambda}}^{1/2}] = \tilde{\mathrm{R}}^\top \circ \bar{\boldsymbol{\lambda}}^{1/2} \\ \boldsymbol{\Lambda} &\leftarrow \mathrm{diag}[\tilde{\boldsymbol{\xi}}^{-1}] + \boldsymbol{X}\boldsymbol{X}^\top \\ \tilde{\Sigma}^{1/2} &\leftarrow \tilde{\mathrm{R}}\,\mathrm{chol}[\boldsymbol{\Lambda}]^{-1} \\ \tilde{\Sigma}_{ii} &\leftarrow \textstyle\sum_j [\tilde{\Sigma}^{1/2}]^2_{ij}\end{aligned} \tag{46}$$

In (46), we first project the (square root of the) precision update $\mathrm{diag}[\bar{\boldsymbol{\lambda}}^{1/2}]$ into the low-rank subspace. We then obtain the inverse posterior covariance in the low-rank space, "$\boldsymbol{\Lambda}$". Rather than invert this directly, we compute its Cholesky factorization and use a triangular inverse solver. We expand this from the low-rank subspace using the inverse FFT. This provides a factor $\tilde{\Sigma}^{1/2}$ of the low-rank approximation to the posterior covariance such that $\tilde{\Sigma} = \tilde{\Sigma}^{1/2}(\tilde{\Sigma}^{1/2})^\top$, from which we extract the diagonal variances. This factor is also useful for sampling from the variational posterior (26).

Iteratively estimating $\boldsymbol{q}$

From (24), we see that $\boldsymbol{q}$ must equal $\bar{\boldsymbol{\lambda}}$ to maximize the evidence lower bound (19) (for fixed $\boldsymbol{\mu}$). However, since $\bar{\boldsymbol{\lambda}}$ depends on $\boldsymbol{q}$, this must be solved self consistently. This can be solved by ascending the simpler gradient $\Delta\boldsymbol{q} \propto \bar{\boldsymbol{\lambda}} - \boldsymbol{q}$. Taking discrete steps yields the following fixed-point iteration:

$$\begin{aligned}&\boldsymbol{q}_{t+1} \leftarrow \bar{\boldsymbol{\lambda}}(\boldsymbol{q}_t), \quad \text{where} \\ &\bar{\boldsymbol{\lambda}}(\boldsymbol{q}_t) = \boldsymbol{n} \circ \exp(\boldsymbol{\mu} + \tfrac{1}{2}\,\mathrm{diag}\left[\Sigma(\boldsymbol{q}_t)\right]) \quad \text{and} \\ &\Sigma(\boldsymbol{q}_t) = \left[\Sigma_z^{-1} + \mathrm{diag}[\boldsymbol{q}]\right]^{-1}.\end{aligned} \tag{47}$$

In practice, we implement this by iterating the marginal posterior variances $\boldsymbol{v} = \mathrm{diag}[\Sigma]$. This is amounts to a different parameterization of (47):

$$\begin{aligned} \mathrm{v}_{t+1} &\leftarrow \mathrm{diag}[\Sigma(\boldsymbol{v}_t)], \quad \text{where} \\ \Sigma(\boldsymbol{v}_t) &= \left[\Sigma_z^{-1} + \mathrm{diag}[\bar{\boldsymbol{\lambda}}(\boldsymbol{v}_t)]\right]^{-1} \\ \bar{\boldsymbol{\lambda}}(\boldsymbol{v}_t) &= \boldsymbol{n} \circ \exp(\boldsymbol{\mu} + \tfrac{1}{2}\boldsymbol{v}_t). \end{aligned} \tag{48}$$

Challis and Barber (2013) note that the iteration in (47) may diverge. In practice, we have found that the reparameterized iteration in (48) always converges when starting from $\boldsymbol{v} = 0$, provided one re-optimizes the posterior mean before each step according to (21), and provided the prior is sufficiently well-conditioned. Note that $\Sigma(\boldsymbol{v}_t)$ remains bounded in the parameterization in (48) as $0 \preccurlyeq \Sigma(\boldsymbol{v}_t) \preccurlyeq \Sigma_z$. This implies that the iteration cannot diverge to infinity if the prior precision Σ_z^{-1} is full rank ($\preccurlyeq$ denotes the Loewner order of positive semidefinite matrices). Since these iterations are simply gradient descent with a step size of 1, instability can be remedied by choosing a smaller step-size, if encountered.

Binning of spatial data

For clarity, we presented the derivations in this manuscript in terms of piecewise-constant spatial basis functions. In practice, linearly interpolated binning provides slightly better resolution for a given grid size, and this is what we used in the provided reference implementation.

For each visit and/or spike at location location $\boldsymbol{x}$, we distributed the point mass at $\boldsymbol{x}$ over a 2×2 neighborhood of adjacent bins via linear interpolation. This amounts to using square-pyramidal basis functions to provide a piecewise-linear model the inferred firing-rate map (compare to figure 1 in Cseke et al., 2016).

Since most calculations are performed directly on the spatial-frequency components of the grid map, choices for spatial binning only affect the numeric integration of the data likelihood (5) over the spatial domain. Locally-constant binning amounts to using the Riemann sum to compute this integral, and linearly interpolated binning amounts to using the trapezoid rule. Linear interpolation improves resolution compared to a piecewise-constant model, but conceptually there are no substantive differences.

Heuristic initialization

Kernel period P: We estimated the grid period P using the radial autocorrelogram of a heuristically-estimate firing-rate $\hat{\boldsymbol{\lambda}}$. The radial autocorrelogram "R_ρ" for an ideal hexagonal grid with period P is given by Equation (11):

$$R_\rho(\|\Delta\boldsymbol{x}\|) \propto J_0\left(\tfrac{2\pi}{P}\|\delta\boldsymbol{x}\|\right) + \text{constant}. \tag{49}$$

The location "Δ_p" of the first nonzero peak of $R_\rho(\|\Delta\boldsymbol{x}\|)$ depends on P, and we can solve for P given Δ_p as

$$P = \tfrac{2\pi}{k_{1,2}}\Delta_p \tag{50}$$

where $k_{1,2}$ is the second zero of the first-order Bessel function of the first kind.

To estimate this autocorrelogram, we first computed a heuristic rate as $\hat{\boldsymbol{\lambda}}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})/\boldsymbol{n}(\boldsymbol{v})$, where $\boldsymbol{k}(\boldsymbol{x})$ are spike-counts and $\boldsymbol{n}(\boldsymbol{v})$ are the number of visits (Locations with no visits are set to zero). We obtained a radial autocorrelogram by multiplying this firing-rate histogram by a Hanning window, computing the 2D autocorrelation, and averaging these values radially.

Heuristic $\boldsymbol{\mu}$ and prior mean $\boldsymbol{\mu_z}$: We used a Gaussian kernel density smoother to estimate foreground $\boldsymbol{\lambda}_f = \mathcal{K}_{\sigma_f} * \hat{\boldsymbol{\lambda}}$ and background $\boldsymbol{\lambda}_b = \mathcal{K}_{\sigma_b} * \hat{\boldsymbol{\lambda}}$ rate maps ($\sigma_f = P/\pi; \sigma_b = 5\sigma_f$). The foreground map contains information about grid fields, and the background rate map contains gradual variations in firing rate across the environment. We use this background log-rate map for the prior mean $\boldsymbol{\mu_z}$ in variational inference. We used the foreground as an initial guess when optimizing the posterior mean.

Kernel height σ_0^2 and constant offset c: We calculated an initial estimate of the log-rate as the difference between the log-foreground and log-background maps. The variance of this map was then used to initialize the kernel height, σ_0^2.

The kernel's constant offset c controls how confident we are in our prior assumptions about the average log-firing-rate across the environment. The average log-rate is the average ("DC") component of the prior mean $\boldsymbol{\mu_z}$. We set c to a large, arbitrary value ($c = 10^3$) to leave this weakly constrained. This leaves the variational inference procedure free to select the appropriate average (log) firing rate.

Algorithm 1: *Iterative procedure for variational-Bayesian log-Gaussian Cox process regression;* $\circ$ denotes element-wise vector and matrix products, with $\boldsymbol{u} \circ \mathrm{A} := \mathrm{diag}[\boldsymbol{u}]\mathrm{A}$ and $\mathrm{A} \circ \boldsymbol{u} := \mathrm{A}\,\mathrm{diag}[\boldsymbol{u}]$, and $(\cdot)^{\circ\cdot}$ denotes element-wise power.

variable definitions:

$\boldsymbol{n} \in \mathbb{R}^{L^2}$:	Vector of visit counts to each location
$\boldsymbol{y} \in \mathbb{R}^{L^2}$:	Vector of spike counts at each location
$\tilde{\boldsymbol{\xi}} \in \mathbb{R}^{D}$:	Low-rank Fourier coefficients of the prior covariance kernel
$\boldsymbol{\mu} \in \mathbb{R}^{L^2}$:	Posterior mean
$\boldsymbol{v} \in \mathbb{R}^{L^2}$:	Diagonal (marginal) posterior variances $\boldsymbol{v} := \mathrm{diag}[\Sigma]$
$\tilde{\mathrm{R}} \in \mathbb{R}^{D \times L^2}$:	Unitary Hartley transform truncated to $D \leq L^2$ components (a semi-orthogonal matrix)
$\tilde{\boldsymbol{\mu}} \in \mathbb{R}^{D}$:	$\tilde{\boldsymbol{\mu}} := \tilde{\mathrm{R}}\,\boldsymbol{\mu}$ Posterior mean in low-rank Frequency space
$\tilde{\boldsymbol{\mu}}_o \in \mathbb{R}^{D}$:	Prior mean in low-rank Frequency space

/* *The low-rank posterior mean* $\tilde{\boldsymbol{\mu}}$ *and marginal posterior variances* $\boldsymbol{v}$ *suffice to keep track of parameterization* (23). */

Starting from some initial $(\tilde{\boldsymbol{\mu}}, \boldsymbol{v})$, **repeat**

  for $i \leftarrow 0$ **to** MEAN_ITERATIONS **do**

    $\tilde{\boldsymbol{\mu}} \leftarrow$ `MeanUpdate(`$\tilde{\boldsymbol{\mu}}_i, \boldsymbol{v}$`)`

  for $i \leftarrow 0$ **to** VARIANCE_ITERATIONS **do**

    $\boldsymbol{v} \leftarrow$ `MarginalVariances(`$\tilde{\boldsymbol{\mu}}, \boldsymbol{v}_i$`)`

until $\tilde{\boldsymbol{\mu}}$ and $\boldsymbol{v}$ have reached the desired level of convergence.

return $\tilde{\boldsymbol{\mu}}, \boldsymbol{v}$

/* *Mean updates resemble fitting a log-linear Poisson GLM with a Gaussian prior* (34), *replacing the posterior mode with the posterior mean* $\langle \boldsymbol{\lambda} \rangle$ (20). *All operations are diagonal in either 2D space or Fourier space. We use the Hartley transform* $\tilde{\mathrm{R}}$ (44) *to switch between these spaces while keeping coefficients real-valued.* */

subroutine `MeanUpdate(`$\tilde{\boldsymbol{\mu}}, \boldsymbol{v}$`)`**:**

  $\bar{\boldsymbol{\lambda}} \leftarrow \boldsymbol{n} \circ \exp(\boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{v} + \tilde{\mathrm{R}}^\top \tilde{\boldsymbol{\mu}})$ // Posterior mean-rate

  $\mathrm{J} \leftarrow \tilde{\boldsymbol{\xi}} \circ \tilde{\boldsymbol{\mu}} + \tilde{\mathrm{R}}\,(\bar{\boldsymbol{\lambda}} - \boldsymbol{n} \circ \boldsymbol{y})$ // Jacobian

  let $\mathrm{H} : \boldsymbol{u} \mapsto \tilde{\boldsymbol{\xi}}^{-1} \circ \boldsymbol{u} + \tilde{\mathrm{R}}\,(\bar{\boldsymbol{\lambda}} \circ \tilde{\mathrm{R}}^\top \boldsymbol{u})$ // Hessian-vector product (function)

  let $\mathrm{M} : \boldsymbol{u} \mapsto \tilde{\boldsymbol{\xi}} \circ \boldsymbol{u}$ // Preconditioner (function)

  return $\tilde{\boldsymbol{\mu}} -$ `Minres(`$\mathrm{H}, \mathrm{J}, \mathrm{M}$`)` // Newton-Raphson update

/* *The variance update resembles the Laplace approximation* (25) *evaluated at* $\langle \boldsymbol{\lambda} \rangle$. *Matrix inverses are performed in the low-rank subspace* (45)–(48). *Combining Cholesky factorization with triangular linear system solvers improves speed by constant factors.* */

subroutine `MarginalVariances(`$\tilde{\boldsymbol{\mu}}, \boldsymbol{v}$`)`**:**

  $\bar{\boldsymbol{\lambda}} \leftarrow \boldsymbol{n} \circ \exp(\boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{v} + \tilde{\mathrm{R}}^\top \tilde{\boldsymbol{\mu}})$ // Posterior mean-rate

  $\boldsymbol{X} \leftarrow \bar{\boldsymbol{\lambda}}^{1/2} \circ \tilde{\mathrm{R}}$ // Precision update square-root

  $\mathrm{C} \leftarrow \tilde{\mathrm{R}}\ \mathrm{chol}(\mathrm{diag}[\tilde{\boldsymbol{\xi}}^{-1}] + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}$ // (Use a triangular solver)

  return $[\mathrm{C}^{\circ 2}]1,$ // "1" is a column vector of ones

References

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., & Wanderman-Milne, S. (2018). JAX: composable transformations of Python+NumPy programs http://github.com/google/jax.

Brandman, D. M., Burkhart, M. C., Kelemen, J., Franco, B., Harrison, M. T., & Hochberg, L. R. (2018). Robust closed-loop control of a cursor in a person with tetraplegia using Gaussian process regression. *Neural computation*, *30*(11), 2986–3008, https://doi.org/10.1162/neco_a_01129.

Brandon, M. P., Bogaard, A. R., Libby, C. P., Connerney, M. A., Gupta, K., & Hasselmo, M. E. (2011). Reduction of theta rhythm dissociates grid cell spatial periodicity from directional tuning. *Science*, *332*(6029), 595–599, https://doi.org/10.1126/science.1201652.

Brown, P. N. & Saad, Y. (1990). Hybrid Krylov methods for nonlinear systems of equations. *SIAM Journal on Scientific and Statistical Computing*, *11*(3), 450–481, https://doi.org/10.1137/0911026.

Challis, E. & Barber, D. (2013). Gaussian Kullback-Leibler approximate inference. *Journal of Machine Learning Research*, *14*(8) http://www.jmlr.org/papers/v14/challis13a.html.

Chan, T. F. & Jackson, K. R. (1984). Nonlinearly preconditioned Krylov subspace methods for discrete Newton algorithms. *SIAM Journal on scientific and statistical computing*, *5*(3), 533–542, https://doi.org/10.1137/0905039.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*(2), 129–157, https://doi.org/10.1111/j.2517-6161.1955.tb00188.x.

Cseke, B., Zammit-Mangion, A., Heskes, T., & Sanguinetti, G. (2016). Sparse approximate inference for spatio-temporal point process models. *Journal of the American Statistical Association*, *111*(516), 1746–1763, https://doi.org/10.1080/01621459.2015.1115357.

Dembo, R. S., Eisenstat, S. C., & Steihaug, T. (1982). Inexact Newton methods. *SIAM Journal on Numerical analysis*, *19*(2), 400–408, https://doi.org/10.1137/0719025.

Duncker, L. & Sahani, M. (2018). Temporal alignment and latent Gaussian process factor inference in population spike trains. *bioRxiv*, (pp. 331751)., https://doi.org/http://proceedings.neurips.cc/paper/2018/hash/d1ff1ec86b62cd5f3903ff19c3a326b2-Abstract.html.

Frigola, R., Chen, Y., & Rasmussen, C. E. (2014). Variational Gaussian process state-space models. In *Advances in neural information processing systems* (pp. 3680–3688). http://proceedings.neurips.cc/paper/2014/hash/139f0874f2ded2e41b0393c4ac5644f7-Abstract.html.

Gal, Y., van der Wilk, M., & Rasmussen, C. E. (2014). Distributed variational inference in sparse gaussian process regression and latent variable models. https://doi.org/10.48550/arXiv.1402.1389.

Gerlei, K., Passlack, J., Hawes, I., Vandrey, B., Stevens, H., Papastathopoulos, I., & Nolan, M. F. (2020). Grid cells are modulated by local head direction. *Nature communications*, *11*(1), 1–14, https://doi.org/10.1038/s41467-020-17500-1.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806, https://doi.org/10.1038/nature0372.

Hartley, R. V. (1942). A more symmetrical fourier analysis applied to transmission problems. *Proceedings of the IRE*, *30*(3), 144–150, https://doi.org/10.1109/JRPROC.1942.234333.

Jensen, K., Kao, T.-C., Tripodi, M., & Hennequin, G. (2020). Manifold GPLVMs for discovering non-euclidean latent structure in neural data. *Advances in Neural Information Processing Systems*, *33*, 22580–22592 http://proceedings.neurips.cc/paper/2020/hash/fedc604da8b0f9af74b6cfc0fab2163c-Abstract.html.

Jensen, K. T., Kao, T.-C., Stone, J. T., & Hennequin, G. (2021). Scalable Bayesian GPFA with automatic relevance determination and discrete noise models. *bioRxiv* http://proceedings.neurips.cc/paper/2021/hash/58238e9ae2dd305d79c2ebc8c1883422-Abstract.html.

Keeley, S. & Pillow, J. (2018). Introduction to Gaussian processes http://pillowlab.princeton.edu/teaching/statneuro2018/slides/notes12_GPs.pdf.

Keeley, S., Zoltowski, D., Yu, Y., Smith, S., & Pillow, J. (2020). Efficient non-conjugate Gaussian process factor models for spike count data using polynomial approximations. In *International Conference on Machine Learning* (pp. 5177–5186).: PMLR http://proceedings.mlr.press/v119/keeley20a.html.

Keinath, A. T., Epstein, R. A., & Balasubramanian, V. (2018). Environmental deformations dynamically shift the grid cell spatial metric. *Elife*, *7*, e38169, https://doi.org/10.7554/eLife.38169.

Kiiveri, H. & De Hoog, F. (2012). Fitting very large sparse Gaussian graphical models. *Computational Statistics & Data Analysis*, *56*(9), 2626–2636, https://doi.org/10.1016/j.csda.2012.02.007.

Killian, N. J., Jutras, M. J., & Buffalo, E. A. (2012). A map of visual space in the primate entorhinal cortex. *Nature*, *491*(7426), 761–764, https://doi.org/10.1038/nature11587.

Knoll, D. A. & Keyes, D. E. (2004). Jacobian-free Newton–Krylov methods: a survey of approaches and applications. *Journal of Computational Physics*, *193*(2), 357–397, https://doi.org/10.1016/j.jcp.2003.08.010.

Krupic, J., Bauza, M., Burton, S., & O'Keefe, J. (2018). Local transformations of the hippocampal cognitive map. *Science*, *359*(6380), 1143–1146, https://doi.org/10.1126/science.aao496.

Langston, R. F., et al. (2010). Development of the spatial representation system in the rat. *Science*, *328*(5985), 1576–1580, https://doi.org/10.1126/science.1188210.

Liu, H., Ong, Y.-S., Shen, X., & Cai, J. (2020). When Gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems*, *31*(11), 4405–4423, https://doi.org/10.1109/TNNLS.2019.2957109.

Luttinen, J. & Ilin, A. (2009). Variational Gaussian-process factor analysis for modeling spatio-temporal data. *Advances in neural information processing systems*, *22*, 1177–1185 http://proceedings.neurips.cc/paper/2009/hash/4a47d2983c8bd392b120b627e0e1cab4-Abstract.html.

MacKay, D. J. (1998). Introduction to Gaussian processes. *NATO ASI series F computer and systems sciences*, *168*, 133–166.

Paige, C. C. & Saunders, M. A. (1975). Solution of sparse indefinite systems of linear equations. *SIAM journal on numerical analysis*, *12*(4), 617–629, https://doi.org/10.1137/0712047.

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, *15*(4), 243, https://doi.org/10.1088/0954-898X_15_4_002.

Park, M., Weller, J. P., Horwitz, G. D., & Pillow, J. W. (2014). Bayesian active learning of neural firing rate maps with transformed Gaussian process priors. *Neural computation*, *26*(8), 1519–1541, https://doi.org/10.1162/NECO_a_00615.

Paun, I., Husmeier, D., & Torney, C. J. (2023). Stochastic variational inference for scalable non-stationary gaussian process regression. *Statistics and Computing*, https://doi.org/10.1007/s11222-023-10210-w.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, *7*(15), 510.

Quinonero-Candela, J. & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, *6*, 1939–1959 http://www.jmlr.org/papers/v6/quinonero-candela05a.html.

Rad, K. R. & Paninski, L. (2010). Efficient, adaptive estimation of two-dimensional firing rate surfaces via Gaussian process methods. *Network: Computation in Neural Systems*, *21*(3-4), 142–168, https://doi.org/10.3109/0954898X.2010.532288.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning* (pp. 63–71).: Springer.

Rowland, D. C., Roudi, Y., Moser, M.-B., & Moser, E. I. (2016). Ten years of grid cells. *Annual review of neuroscience*, *39*, 19–40, https://doi.org/10.1146/annurev-neuro-070815-013824.

Rule, M. & Sanguinetti, G. (2018). Autoregressive point processes as latent state-space models: A moment-closure approach to fluctuations and autocorrelations. *Neural computation*, *30*(10), 2757–2780, https://doi.org/10.1162/neco_a_01121.

Rule, M. E., Schnoerr, D., Hennig, M. H., & Sanguinetti, G. (2019). Neural field models for latent state inference: Application to large-scale neuronal recordings. *PLoS computational biology*, *15*(11), e1007442, https://doi.org/10.1371/journal.pcbi.1007442.

Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B. L., Witter, M. P., Moser, M.-B., & Moser, E. I. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, *312*(5774), 758–762, https://doi.org/10.1126/science.1125572.

Seeger, M. (1999). Bayesian methods for support vector machines and Gaussian processes. Master's thesis, École Polytechnique Fédérale de Lausanne https://infoscience.epfl.ch/record/175479.

Truccolo, W. (2016). From point process observations to collective neural dynamics: Nonlinear Hawkes process GLMs, low-dimensional dynamics and coarse graining. *Journal of Physiology-Paris*, *110*(4), 336–347, https://doi.org/10.1016/j.jphysparis.2017.02.004.

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., & Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, *93*(2), 1074–1089, https://doi.org/10.1152/jn.00697.2004.

Virtanen, P., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272, https://doi.org/10.1038/s41592-019-0686-2.

Wu, A., Roy, N. A., Keeley, S., & Pillow, J. W. (2017). Gaussian process based nonlinear latent structure discovery in multivariate spike train data. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3499–3508). http://proceedings.neurips.cc/paper/2017/hash/b3b4d2dbedc99fe843fd3dedb02f086f-Abstract.html.

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., & Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of neurophysiology*, *102*(1), 614–635

http://proceedings.neurips.cc/paper/2008/hash/ad972f10e0800b49d76fed33a21f6698-Abstract.html.

Zhao, Y. & Park, I. M. (2017). Variational latent Gaussian process for recovering single-trial dynamics from population spike trains. *Neural computation*, *29*(5), 1293–1316, https://doi.org/10.1162/NECO_a_00953.