Scalable Variational Inference for Low-Rank Spatiotemporal Receptive Fields

Lea Duncker
lduncker@stanford.edu
Wu Tsai Neurosciences Institute and Howard Hughes Medical Institute,
Stanford University, Stanford, CA 94302, U.S.A.

Kiersten M. Ruda
kruda@bidmc.harvard.edu
Beth Israel Deaconess Medical Center, Harvard University,
Boston, MA 02115, U.S.A.

Greg D. Field
field@neuro.duke.edu
Department of Neurobiology, Duke University, Durham, NC 27708, U.S.A.

Jonathan W. Pillow
pillow@princeton.edu
Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, U.S.A.

An important problem in systems neuroscience is to characterize how a neuron integrates sensory inputs across space and time. The linear receptive field provides a mathematical characterization of this weighting function and is commonly used to quantify neural response properties and classify cell types. However, estimating receptive fields is difficult in settings with limited data and correlated or high-dimensional stimuli. To overcome these difficulties, we propose a hierarchical model designed to flexibly parameterize low-rank receptive fields. The model includes gaussian process priors over spatial and temporal components of the receptive field, encouraging smoothness in space and time. We also propose a new temporal prior, temporal relevance determination, which imposes a variable degree of smoothness as a function of time lag. We derive a scalable algorithm for variational Bayesian inference for both spatial and temporal receptive field components and hyperparameters. The resulting estimator scales to high-dimensional settings in which full-rank maximum likelihood or a posteriori estimates are intractable. We evaluate our approach on neural data from rat retina and primate cortex and show that it substantially outperforms a variety of existing estimators. Our modeling approach will have useful extensions to a variety of other high-dimensional inference problems with smooth or low-rank structure.

1 Introduction

A key problem in computational and systems neuroscience is to understand the information carried by neurons in the sensory pathways (Jones & Palmer, 1987; Ringach et al., 1997). A common approach to this problem is to estimate the linear receptive field (RF), which provides a simple characterization of a neuron's stimulus-response properties. The RF consists of a set of linear weights that describe how a neuron integrates a sensory stimulus over space and time (Chichilnisky, 2001; deBoer & Kuyper, 1968; Jones & Palmer, 1987; Lee & Schetzen, 1965; Meyer et al., 2017; Ringach, 2004; Schwartz et al., 2006; Sharpee et al., 2004; Simoncelli et al., 2004; Theunissen et al., 2001). Estimating the RF from imaging or electrophysiological recordings can thus be seen as a straightforward regression problem. However, characterizing RFs in realistic settings poses a number of important challenges.

One major challenge for RF estimation is the high dimensionality of the inference problem. The number of RF coefficients is equal to the product of the number of spatial stimulus elements (e.g., pixels in an image) and number of time lags in the temporal filter. In realistic settings, this can easily surpass thousands of coefficients. Classic RF estimators such as the least-squares regression suffer from high computational cost and low statistical power in high-dimensional settings. The computational cost of the regression estimate scales cubically with the number of RF coefficients, while memory cost scales quadratically. Moreover, the standard regression estimate does not exist unless there are more samples than dimensions and typically requires large amounts of data to achieve a high level of accuracy. High dimensionality is thus a limiting factor in terms of both computational resources and statistical power.

Past efforts to improve statistical power, and thereby reduce data requirements, have relied on various forms of regularization. Regularization reduces the number of degrees of freedom in the RF by biasing the estimator toward solutions that are more likely a priori and can thus provide for better estimates from less data. Common forms of regularization involve smoothness or sparsity assumptions and have been shown to outperform maximum likelihood (ML) estimation in settings of limited data or correlated stimuli (Aoi & Pillow, 2017; Calabrese et al., 2011; David & Gallant, 2005; Gerwinn et al., 2010; Linden et al., 2003; Park & Pillow, 2011; Smyth et al., 2003). However, the computational cost of such estimators is generally no better than that of classical estimators, and may be worse due to the need to optimize hyperparameters governing the strength of regularization. These poor scaling properties make it difficult to apply such estimators to settings involving high-dimensional stimulus ensembles.

In this article, we propose to overcome these difficulties using a model that parameterizes the RF as smooth and low rank. A low-rank receptive field can be described as a sum of a small number of space-time

separable filters (Linden et al., 2003; Park & Pillow, 2013; Pillow et al., 2008; Pillow & Simoncelli, 2006; Qiu et al., 2003). This choice of parameterization significantly reduces the number of parameters in the model and is the key to the scalability of our method. To achieve smoothness, we use gaussian process (GP) priors over both the spatial and temporal filters composing the low-rank RF. For the temporal filters, we introduce the temporal relevance determination (TRD) prior, which uses a novel covariance function that allows for increasing smoothness as a function of time lag. To fit the model, we develop a scalable algorithm for variational Bayesian inference for both receptive field components and hyperparameters governing shrinkage and smoothness. The resulting estimator achieves excellent performance in settings with correlated stimuli and scales to high-dimensional settings in which full-rank estimators are intractable.

The article is organized as follows. Sections 2 and 3 review relevant background and the literature: section 2 introduces the linear encoding model and low-rank linear receptive fields; section 3 reviews previously proposed priors for receptive fields. In section 4, we introduce TRD, a new prior for temporal receptive fields, while in section 5 we introduce our variational low-rank (VLR) receptive field estimator. Section 6 shows applications to simulated as well as real neural data sets, and section 7 provides discussion of our results and suggests directions for future work.

2 The Linear-Gaussian Encoding Model

A classic approach to neural encoding is to formulate a parameteric statistical model that describes the mapping from stimulus input to neural response (Efron & Morris, 1973; James & Stein, 1961; Jones & Palmer, 1987; Park & Pillow, 2011; Ringach et al., 1997). Here we focus on linear models, where the neural response is described as a linear or affine function of the stimulus plus gaussian noise (Park & Pillow, 2011; Sahani & Linden, 2003) (see Figure 1). Formally, the model can be written as

$$y_t = \mathbf{k}^\top \mathbf{x}_t + b + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{2.1}$$

where y_t is the neuron's scalar response at the tth time bin, $\mathbf{k}$ is the vector receptive field, $\mathbf{x}_t$ is the vectorized stimulus at the tth time bin, b is an additive constant or bias, and ϵ_t denotes zero-mean gaussian noise with variance σ^2.

2.1 The Receptive Field Tensor. This model description neglects the fact that the stimulus and receptive field are often more naturally described as multidimensional tensors. In vision experiments, for example, the stimulus and receptive field are typically third-order tensors, with two dimensions of space and one dimension of time. In this case, the full stimulus movie shown during an entire experiment can be represented by a tensor
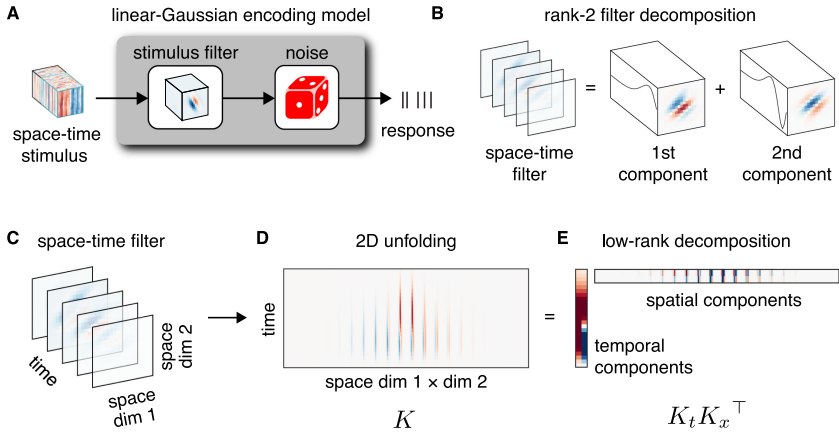
Figure 1: (A) The linear-gaussian encoding model is a two-stage model consisting of a linear stage followed by a noise stage. First, the high-dimensional spatiotemporal stimulus is filtered with the linear receptive field, which describes how the neuron integrates the stimulus values over space and time. The output of this filter is a scalar that gives the expected response at a single time bin. We then add noise to obtain the neural response for the given time bin. (B) A low-rank spatiotemporal receptive field can be described as a sum of two space-time-separable (rank 1) filters. (C, D) Illustration of different STRF representations as a third order tensor (C), matrix (D), or in terms of low-rank factors (E).

of size $n_{x_1} \times n_{x_2} \times T$, consisting of $n_{x_1} \times n_{x_2}$ pixel images over T total time bins or frames. In this case, a given element $\mathbf{X}_{i,j,t}$ in the stimulus tensor could represent the light intensity in a gray-scale image at time t at spatial location (i, j) in the image.

In such settings, the spatiotemporal receptive field is also naturally defined as a tensor, $\mathbf{K} \in \mathbb{R}^{n_{x_1} \times n_{x_2} \times n_t}$, with weights that determine how the neuron integrates light over the $n_{x_1} \times n_{x_2}$ spatial pixels and the n_t preceding time bins (see Figure 1C). Thus, the dot product between the receptive field $\mathbf{k}$ and vector stimulus $\mathbf{x}_t$ in equation 2.1 is equal to the following linear function defined by summing over the product of all corresponding elements of the RF tensor $\mathbf{K}$ and corresponding portion of the stimulus tensor $\mathbf{X}$ at time t,

$$\mathbf{k}^\top \mathbf{x}_t = \sum_{i=1}^{n_{x_1}} \sum_{j=1}^{n_{x_2}} \sum_{\tau=0}^{n_t-1} \mathbf{K}_{i,j,\tau} \mathbf{X}_{i,j,t-\tau}, \tag{2.2}$$

where $\mathbf{K}$ and $\mathbf{X}_t$ are both tensors of size $n_{x_1} \times n_{x_2} \times n_t$. The vectorized RF and stimulus are thus given by $\mathbf{k} = \text{vec}(\mathbf{K})$, and $\mathbf{x}_t = \text{vec}(\mathbf{X}_{i,j,t-\tau})$ for $i \in \{1, \dots, n_{x_1}\}$, $j \in \{1, \dots, n_{x_2}\}$ and $\tau \in \{1, \dots, n_t\}$.

2.2 Low-Rank Receptive Fields. A key feature of neural receptive fields is that they can typically be described by a relatively small number of spatial and temporal components (Adelson & Bergen, 1985; Linden et al., 2003; Park & Pillow, 2013; Pillow et al., 2008; Pillow & Simoncelli, 2006; Qiu et al., 2003). This means that we do not need to use a full set of $n_{x_1} \times n_{x_2} \times n_t$ independent parameters to represent the coefficients in **K**. Instead, we can accurately describe **K** with a small number of spatial components, each corresponding to an image with $n_{x_1} \times n_{x_2}$ coefficients, and a corresponding number of temporal components, each with n_t coefficients. The number of paired spatial and temporal components needed to represent **K** is known as the rank of the tensor, which we denote r.

Figure 1B illustrates a scenario with $r = 2$, in which the tensor **K** is the sum of two rank 1 components, each defined as a single spatial and temporal weighting function. These rank 1 components are commonly referred to as "space-time separable" filters. Note that a rank r tensor has only $r(n_{x_1} n_{x_2} + n_t)$ parameters, which generally represents a significant savings over the $(n_{x_1} n_{x_2} n_t)$ parameters needed to parameterize the full-rank tensor. Furthermore, having an explicit description of the temporal and spatial filters composing **K** increases the interpretability of the RF.

For modeling low-rank filters of this form, it is convenient to unfold the third-order receptive field tensor **K** into a matrix. Let $K \in \mathbb{R}^{n_t \times (n_{x_1} \cdot n_{x_2})}$ denote the matrix unfolding of the receptive field, where the two spatial dimensions have been concatenated (see Figures 1C and 1D). This representation makes it possible to represent low-rank receptive fields with a product of matrices,

$$K = K_t K_x, \tag{2.3}$$

where $K_t \in \mathbb{R}^{n_t \times r}$ is a matrix whose columns are temporal filters and $K_x \in \mathbb{R}^{(n_{x_1} n_{x_2}) \times r}$ is a matrix whose columns are (reshaped) spatial filters (see Figure 1E). In section 5, we will develop a Bayesian hierarchical model for efficient estimation of low-rank receptive fields using this parameterization.

3 Existing Receptive Field Priors

In high-dimensional settings, or settings with highly correlated stimuli, receptive field estimation can be substantially improved by regularization. Here we review previously proposed prior distributions for regularizing receptive field estimates. The general family of priors that we consider takes the form of a zero-mean multivariate gaussian distribution,

$$\mathbf{k} \sim \mathcal{N}(0, C_\theta), \tag{3.1}$$

where C_θ denotes a covariance matrix that depends on a set of hyperparameters θ. Different choices of prior arise by selecting different functional
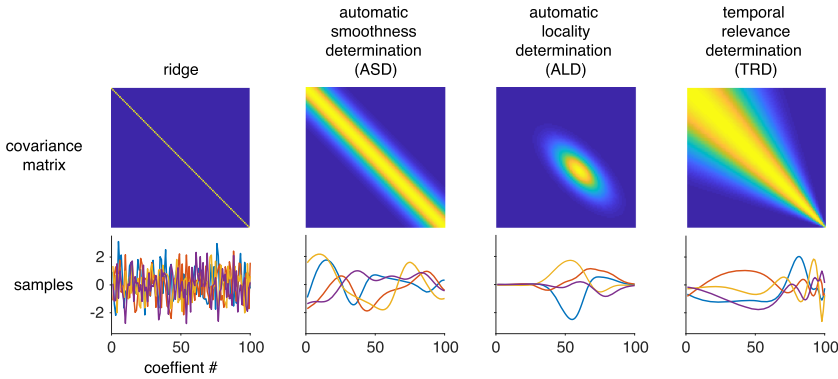
Figure 2: Illustration of different priors for use in receptive field estimation. (Top) Prior covariance matrices C_θ under different choices of covariance function. Under the ridge prior, all off-diagonal elements of the covariance matrix are zero, and receptive field coefficients are independent. In the ASD prior covariance, neighboring coefficients are highly correlated, and the correlation decreases with increasing distance between coefficient locations in the RF. The ALD prior covariance contains high correlations for neighboring coefficients within a local region and sets the prior covariance of the RF coefficient to zero outside of this local region, a form of structured sparsity. Finally, in the TRD covariance matrix, the correlation of neighboring RF coefficients increases as a function of the coefficient location. (Bottom) Four samples from a multivariate gaussian distribution with zero mean and covariance matrix shown above, illustrating the kinds of receptive fields that are typical under each choice of prior.

forms for the covariance matrix C_θ. We will review several popular choices of covariance (see Figure 2) before introducing a novel prior covariance in section 4.

3.1 Ridge Prior. Ridge regression (Hoerl & Kennard, 1970) is classically viewed as an added L_2 norm penalty on the receptive field weights in the context of least-squares regression. However, it also has a Bayesian interpretation as resulting from a zero-mean gaussian prior with covariance given by,

$$C_\theta^{\text{ridge}} = \frac{1}{\lambda}I, \tag{3.2}$$

where the hyperparameter $\theta = \{\lambda\}$ is known as the ridge parameter and I denotes the identity matrix. This prior has the effect of biasing the estimate toward zero, a phenomenon also known as L_2 shrinkage.

3.2 Automatic Smoothness Determination. The automatic smoothness determination (ASD) prior (Sahani & Linden, 2003) goes beyond shrinkage to incorporate the assumption that the receptive field changes smoothly in time and/or space. The ASD prior covariance matrix relies on the radial basis function (RBF) or "gaussian" covariance function that is well known in the gaussian process literature (Rasmussen & Williams, 2006). The (i, j)th element of this covariance matrix is given by

$$[C_\theta^{\text{ASD}}]_{ij} = \rho \exp\left(-\frac{1}{2}(\chi_i - \chi_j)^\top \Gamma^{-1}(\chi_i - \chi_j)\right), \tag{3.3}$$

where $\{\chi_i\}$ is a 3D vector containing the locations of RF pixels in space-time, thus indicating both the 2D spatial locations of the RF coefficients and the 1D temporal locations (or lags). And $\Gamma = \text{diag}(\ell_1^s, \ell_2^s, \ell^t)$ is a diagonal matrix containing length-scale parameters. The covariance matrix is thus controlled by four hyperparameters: $\theta = \{\rho, \ell_1^s, \ell_2^s, \ell^t\}$. ρ is the marginal variance (and equivalent to $1/\lambda$ in the ridge prior above), and the length-scale parameters $\ell_{1,2}^s$ and ℓ^t determine the degree of smoothness in space and time, respectively. Recent work has exploited the Kronecker and Toeplitz structure of the ASD covariance matrix to show that it has an exact diagonal representation in the Fourier domain, which allows for dramatic improvements in computational efficiency (Aoi & Pillow, 2017).

3.3 Automatic Locality Determination. The automatic locality determination (ALD) prior (Park & Pillow, 2011) goes beyond the smoothness of the ASD prior by encouraging RF coefficients to be localized in space-time and spatiotemporal frequency.

It relies on a covariance function that encourages both the space-time coefficients and the spatiotemporal frequency components of the RF to be zero outside of some localized region, resulting in a form of structured sparsity (Wu et al., 2014). This prior includes the ASD smoothness prior as a special case, namely, when the region of nonzero spatiotemporal frequency components is centered at zero, and there is no locality in space or time. However, the ALD prior also allows for bandpass filters in which the RF is composed primarily of intermediate frequencies.

The ALD covariance matrix can be written in terms of a pair of diagonal matrices and a discrete Fourier transform matrix,

$$C_\theta^{\text{ALD}} = \rho \, C_\theta^{s\,\frac{1}{2}} B^\top C_\theta^f B C_\theta^{s\,\frac{1}{2}}, \tag{3.4}$$

where C_θ^s and C_θ^f are diagonal matrices that specify a region of nonzero coefficients in space-time and the Fourier domain, respectively, and B is the discrete-time Fourier transform matrix.

The space-time locality matrix is a diagonal matrix with diagonal elements given by

$$[C_\theta^s]_{ii} = \exp\left(-\frac{1}{2}(\chi_i - \mathbf{m})^\top \Psi^{-1}(\chi_i - \mathbf{m})\right), \tag{3.5}$$

where $\{\chi_i\}$ are the locations of RF pixels in space-time, and $\mathbf{m}$ and Ψ denote the mean and covariance of the region where the RF coefficients are nonzero. The Fourier-domain locality matrix takes a similar form,

$$[C_\theta^f]_{ii} = \exp\left(-\frac{1}{2}(\tilde{\chi}_i - \tilde{\mathbf{m}})^\top \tilde{\Psi}^{-1}(\tilde{\chi}_i - \mathbf{m})\right), \tag{3.6}$$

where $\{\tilde{\chi}_i\}$ are the spatiotemporal frequencies for each Fourier coefficient, and $\tilde{\mathbf{m}}$ and $\tilde{\Psi}$ denote the mean and covariance of the region in Fourier space where RF Fourier coefficients are nonzero. The hyperparameters governing the ALD prior are thus $\theta = \{\rho, \mathbf{m}, \tilde{\mathbf{m}}, \Psi, \tilde{\Psi}\}$. As before, ρ governs the marginal variance of the RF coefficients, analogous to the ridge parameter. For a 3D tensor receptive field, $\mathbf{m}$ and $\tilde{\mathbf{m}}$ are both 3-vectors, and Ψ and $\tilde{\Psi}$ are both 3×3 positive semidefinite covariance matrices.

The full ALD covariance matrix (see equation 3.4), which multiplies together the diagonal matrices C_θ^s and C_θ^f with a discrete Fourier domain operator in between, has the effect of simultaneously imposing locality in both space-time and frequency. It is identical to ASD under the setting that $C_\theta^s = I$, $\tilde{\mathbf{m}} = 0$, and $\tilde{\Psi}$ is diagonal (Aoi & Pillow, 2017). Empirically, ALD has been shown to outperform both the ridge and ASD priors, as well as other sparsity-inducing priors, in applications to neural data in the early visual system (Park & Pillow, 2011).

4 A New Prior for Temporal Receptive Fields

Our first contribution is to propose a new prior for temporal receptive fields. The priors in ASD and ALD both assume a constant degree of smoothness across the receptive field. However, an assumption of stationary smoothness is less appropriate for temporal receptive fields, which are typically sharper at short time lags and smoother at longer time lags. In order to incorporate this variable form of smoothness, we introduce the temporal recency determination (TRD) prior. This prior uses a smoothing covariance function that is stationary in log-scaled time and is therefore nonstationary in linear time.

The TRD covariance function can be described as a time-warped version of the ASD covariance. Specifically, the (i, j)th element of the prior covariance is given by

$$[C_\theta^{\mathrm{TRD}}]_{ij} = \rho \exp\left(-\frac{1}{2\ell^2}\|\tau_\alpha(t_i) - \tau_\alpha(t_j)\|^2\right). \tag{4.1}$$

As was the case for the ASD covariance function, ρ describes the marginal variance of the RF prior and ℓ is a length-scale parameter determining the smoothness of the RF. Additionally, $\tau_\alpha(t)$ is a nonlinear warping function given by

$$\tau_\alpha(t) = \frac{T}{\log(1 + \exp(\alpha) * T)} \log(1 + \exp(\alpha)t). \tag{4.2}$$

Here, T is the length of the temporal RF (in seconds), t_i is the time lag for temporal RF coefficient i, and α is a parameter determining how quickly the RF smoothness increases with time. In summary, the TDR covariance function is determined through a set of three hyperparameters, $\theta = \{\rho, \ell, \alpha\}$ and can incorporate the prior structure that temporal receptive fields become smoother for increasing time lags. Figure 2 shows an illustration of the TRD prior, alongside the other RF priors discussed in section 3.

5 A Probabilistic Model for Low-Rank Receptive Fields

Our second contribution is a model and inference method for smooth, low-rank receptive field fields. As noted in section 2.2, a low-rank parameterization for spatiotemporal receptive fields can offer a dramatic reduction in dimensionality without loss of accuracy. In particular, a rank r spatiotemporal filter requires only $r \cdot (n_{x_1} \cdot n_{x_2} + n_t)$ coefficients versus $N = n_{x_1} \cdot n_{x_2} \cdot n_t$ for a full-rank filter.

To place our model on a solid probabilistic footing, we place independent zero-mean gaussian priors over the r temporal (1D) and r spatial (2D) components of the receptive field (see Figure 3A). If we use a TRD prior for the temporal components and an ALD prior for the spatial components, then the prior over the ith spatial and temporal receptive field components can be written as

$$\mathbf{k}_{ti} \sim \mathcal{N}(0, C_{\theta_t}^{\text{TRD}}), \tag{5.1}$$

$$\mathbf{k}_{xi} \sim \mathcal{N}(0, C_{\theta_x}^{\text{ALD}}), \tag{5.2}$$

where $C_{\theta_t}^{\text{TRD}}$ denotes the TRD covariance (see equation 4.1), $C_{\theta_x}^{\text{ALD}}$ is the ALD covariance (see equation 3.4), which we apply here to a 2D spatial receptive field. Although we selected these covariance functions because of their suitability for the structure of neural receptive fields, our modeling framework is general and could easily accommodate other choices.

Under this modeling choice, the full receptive field $K = K_t K_x^\top = \mathbf{k}_{t1}\mathbf{k}_{x1}^\top + \cdots + \mathbf{k}_{tr}\mathbf{k}_{xr}^\top$ represents the product of two gaussian random variables. This effective prior over K is thus non gaussian. This entails that posterior inference is not analytically tractable due to the fact that the prior is not conjugate to the gaussian likelihood of the encoding model (see
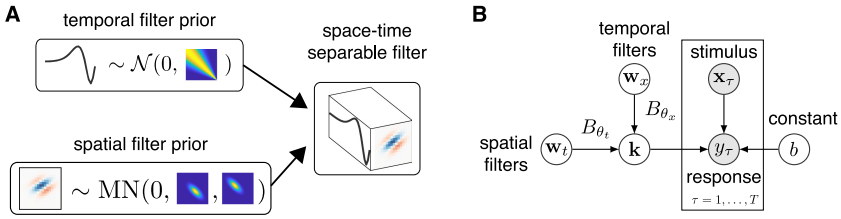
Figure 3: Graphical model representation of the low-rank STRF model. (A) We combine a multivariate normal temporal relevance determination (TRD) prior over temporal components (above) and a matrix normal prior over spatial components (below), where row and column covariances have an automatic locality determination (ALD) parameterization, to obtain a prior over each rank-1 component or space-time separable filter making up a low-rank filter. (B) The graphical model for the full model can be represented in terms of linear transformations $B_{\theta_t}^t$ and $B_{\theta_x}^x$, which parameterize the transformation from whitened temporal and spatial receptive fields $\mathbf{w}_t$ and $\mathbf{w}_x$, which are combined to form the STRF $\mathbf{k}$. The linear transformations depend on hyperparameters θ_t and θ_x. The spatial and temporal components are combined to form a low-rank STRF $\mathbf{k}$, which acts to integrate a stimulus $\mathbf{x}_\tau$ over space and time to generate a neural response y_τ, with constant offset b.

equation 2.1). We therefore develop a scalable variational inference algorithm for setting the hyperparameters θ governing the prior covariance matrices and inferring the receptive field coefficients K_x and K_t, which we describe in sections 5.1 and 5.2. We introduce additional modeling choices that facilitate scalable estimation through a further reduction in the total number of receptive field coefficients in section 5.1 and describe the resulting algorithm in detail in section 5.2.

5.1 Scalability via Sparse Spectral Representations of Covariance Matrices. Performing inference in the model we have already introduced still requires building and inverting large covariance matrices. To reduce the dimensionality of the spatial and temporal receptive field inference further and to achieve scalability to high-dimensional settings, we make use of sparse spectral representations of the prior covariance matrices.

This sparse spectral representation involves expressing the prior covariance of $\mathbf{k}$ as $C_\theta \approx B_\theta B_\theta^\top$. When B_θ contains the eigenvectors of C_θ, scaled by their square-rooted associated eigenvalues, then this approximation is exact. However, we might be able to incur only a small approximation error by retaining only the leading eigenvectors of C_θ in B_θ. For example, for larger length-scale parameters (high degree of smoothness), the covariance matrix associated with the ASD covariance function has eigenvalues that quickly decay toward zero (Rasmussen & Williams, 2006). By truncating its

eigenvalues near zero, it is possible to find a low-rank approximation to the covariance matrix C_θ that is exact to machine precision. For all choices of prior covariance matrices we discussed in section 3, it is possible to obtain an analytic expression for B_θ (Aoi & Pillow, 2017).

Given B_θ, we can express the prior distribution over the receptive field in terms of a linear transformation of a multivariate standard normal, or "whitened" random variable: $\mathbf{k} = B_\theta \mathbf{w}$, with $\mathbf{w} \sim \mathcal{N}(0, I)$. The resulting prior covariance of $\mathbf{k}$ is equal to $C_\theta = B_\theta B_\theta^\top$ and thus unchanged from before. However, performing inference over $\mathbf{w}$ instead of $\mathbf{k}$ allows us to circumvent costly inversions of C_θ. Furthermore, if C_θ is represented to sufficient accuracy using fewer dimensions in B_θ, then $\mathbf{w}$ will contain fewer dimensions than $\mathbf{k}$, leading to additional improvements in scalability.

To incorporate low-rank receptive field structure into this approach, we define our model in the transformed, "whitened" space. The receptive field priors are described by multivariate standard normal distributions over the whitened temporal receptive field $\mathbf{w}_t$ and whitened spatial receptive field $\mathbf{w}_x$:

$$\mathbf{w}_t \sim \mathcal{N}(0, I_{p_t \cdot r}), \tag{5.3}$$

$$\mathbf{w}_x \sim \mathcal{N}(0, I_{p_x \cdot r}), \tag{5.4}$$

where $\mathbf{w}_t$ and $\mathbf{w}_x$ are vectors containing the concatenated temporal and spatial components, respectively, and are of dimensions $p_t \cdot r$ and $p_x \cdot r$, where $p_{x,t}$ depend on the number of dimensions needed to approximate the respective covariance matrices to sufficient accuracy and r corresponds to the STRF rank. We define the matrix reshaping of these vectors as

$$\mathbf{w}_t = \text{vec}(W_t), \quad \mathbf{w}_x = \text{vec}(W_x^\top), \tag{5.5}$$

where $W_t \in \mathbb{R}^{p_t \times r}$ and $W_x \in \mathbb{R}^{p_x \times r}$. The full STRF can then be represented as

$$\mathbf{k} = \text{vec}(B_{\theta_t}^t W_t W_x^\top B_{\theta_x}^x), \tag{5.6}$$

where the analytic decompositions of the prior covariances are denoted as $C_\theta^t = B_{\theta_t}^t {B_{\theta_t}^t}^\top$ and $C_\theta^x = B_{\theta_x}^x {B_{\theta_x}^x}^\top$ for the temporal and the spatial receptive field prior covariances, respectively.

Given this representation of the STRF, the rest of the model remains unchanged. This model description is illustrated as a graphical model in Figure 3B.

5.2 Variational Inference and Hyperparameter Learning for Low-Rank STRFs (VLR). Fitting our low-rank model to data involves performing posterior inference over the receptive field components $\mathbf{w}_x$ and $\mathbf{w}_t$ and learning the hyperparameters θ that determine the shape of the basis

functions in B_θ. To do this, we derive a variational optimization framework, similar to the classic empirical Bayes approach reviewed in appendix A.2. We rely on a variational approach here, since the effective non gaussian prior over **k** makes obtaining an exact expression for the marginal log-likelihood intractable.

A lower bound to the marginal log-likelihood can be obtained by applying Jensen's inequality,

$$\log P(Y|\Phi, \theta) = \log \int P(Y, \mathbf{w}_t, \mathbf{w}_x|\Phi, \theta) d\mathbf{w}_t d\mathbf{w}_x \tag{5.7}$$

$$\geq \int q(\mathbf{w}_t, \mathbf{w}_x) \log \frac{P(Y, \mathbf{w}_t, \mathbf{w}_x|\Phi, \theta)}{q(\mathbf{w}_t, \mathbf{w}_x)} d\mathbf{w}_t d\mathbf{w}_x \stackrel{\text{def}}{=} \mathcal{F}(q, \theta), \tag{5.8}$$

where $q(\mathbf{w}_t, \mathbf{w}_x)$ is a distribution over the whitened temporal and spatial receptive field parameters $\mathbf{w}_t$ and $\mathbf{w}_x$. $\mathcal{F}(q, \theta)$ is often called the variational "free energy" or "evidence lower bound" (ELBO). Y denotes the recorded neural responses, while Φ is the stimulus design matrix (see appendix A.1 for details).

Instead of computing and optimizing the model evidence directly, we can instead perform coordinate ascent and alternatingly maximize $\mathcal{F}$ with respect to the distribution $q(\mathbf{w}_t, \mathbf{w}_x)$ and with respect to the parameters θ as part of the expectation maximization (EM) (Dempster et al., 1977) algorithm. The optimization over $q(\mathbf{w}_t, \mathbf{w}_x)$ can either be performed exactly (which amounts to setting q equal to the posterior joint distribution over $\mathbf{w}_t$ and $\mathbf{w}_x$) or under additional constraints on q, such as restricting it to a particular family of variational distribution.

The free energy can be written in two equivalent ways:

$$\mathcal{F}(q(\mathbf{w}_t, \mathbf{w}_x), \theta) = \log P(Y|\Phi, \theta) - \mathrm{KL}[q(\mathbf{w}_t, \mathbf{w}_x)\|p(\mathbf{w}_t, \mathbf{w}_x|Y, \Phi, \theta)], \tag{5.9}$$

$$\mathcal{F}(q(\mathbf{w}_t, \mathbf{w}_x), \theta) = \langle \log P(Y, \mathbf{w}_t, \mathbf{w}_x|\Phi, \theta)\rangle_{q(\mathbf{w}_t, \mathbf{w}_x)} + \mathbf{H}[q(\mathbf{w}_t, \mathbf{w}_x)], \tag{5.10}$$

where $\mathbf{H}[q]$ is the entropy of q. The EM algorithm involves iteratively maximizing this lower bound with respect to a distribution over $\mathbf{w}_t$ and $\mathbf{w}_x$ (E-step) and with respect to the hyperparameters θ (M-step). From equation 5.9, it is apparent that the free energy is maximized when $q(\mathbf{w}_t, \mathbf{w}_x)$ is equal to the posterior distribution,

$$P(\mathbf{w}_t, \mathbf{w}_x|Y, \Phi, \theta) = \frac{P(Y, \mathbf{w}_t, \mathbf{w}_x|\Phi, \theta)}{P(Y|\Phi, \theta)}, \tag{5.11}$$

at which the Kullback-Leibler divergence in equation 5.9 vanishes and the free energy provides a tight lower bound to the marginal log-likelihood.

Performing this computation exactly is intractable in our model. We therefore choose a variational approximation of the form

$$q(\mathbf{w}_t, \mathbf{w}_x) = q(\mathbf{w}_t)q(\mathbf{w}_x), \tag{5.12}$$

where we assume that the approximate posterior distribution factorizes over the spatial and temporal receptive field. We now seek to find the distribution $q(\mathbf{w}_t, \mathbf{w}_x)$ that lies within the family of distributions that factorize over $\mathbf{w}_t$ and $\mathbf{w}_x$ and maximizes the free energy. Taking variational derivatives of the free energy, the variational updates for our approximating distributions are found to be of the general form

$$q(\mathbf{w}_t) \propto \exp\langle \log P(Y, \mathbf{w}_t, \mathbf{w}_x | \Phi, \theta) \rangle_{q(\mathbf{w}_x)}, \tag{5.13}$$

$$q(\mathbf{w}_x) \propto \exp\langle \log P(Y, \mathbf{w}_t, \mathbf{w}_x | \Phi, \theta) \rangle_{q(\mathbf{w}_t)}, \tag{5.14}$$

where angled brackets denote expectations. Evaluating the above in our model, we obtain multivariate gaussian distributions of the form $q(\mathbf{w}_t) = \mathcal{N}(\mathbf{w}_t | \mu_t, \Sigma_t)$ and $q(\mathbf{w}_x) = \mathcal{N}(\mathbf{w}_x | \mu_x, \Sigma_x)$. The variational updates for the posterior means μ_t, μ_x, and covariances Σ_t, Σ_x are available in closed form. Detailed derivations and the exact update equations are given in appendix B.

The M-step in our variational EM algorithm involves maximizing the free energy with respect to the model parameters b, σ and hyperparameters $\theta = \{\theta_t, \theta_x\}$ and requires solving

$$\theta^*, b^*, \sigma^* \leftarrow \arg\max \langle \log P(Y, \mathbf{w}_t, \mathbf{w}_x | \Phi, \theta) \rangle_{q(\mathbf{w}_t, \mathbf{w}_x)} \tag{5.15}$$

using gradient-based optimization. We update the hyperparameters in $B_{\theta_t}^t$ and $B_{\theta_x}^x$, and the model parameters b and σ in separate conditional M-steps. Performing conditional M-steps allows one to project the high-dimensional stimulus onto either the temporal or spatial basis and never requires building or inverting the full stimulus design matrix Φ. Thus, this strategy further exploits the lower-dimensional decomposition of the full receptive field and provides an efficient algorithm, even for high-dimensional data. Further details are provided in appendix C.

6 Results

6.1 Data Efficient Estimation under Correlated Stimulus Distributions. We first tested the VLR estimator using synthetic data. We simulated data from a linear-gaussian encoding model with a rank-2 receptive field that was stimulated with correlated stimuli.
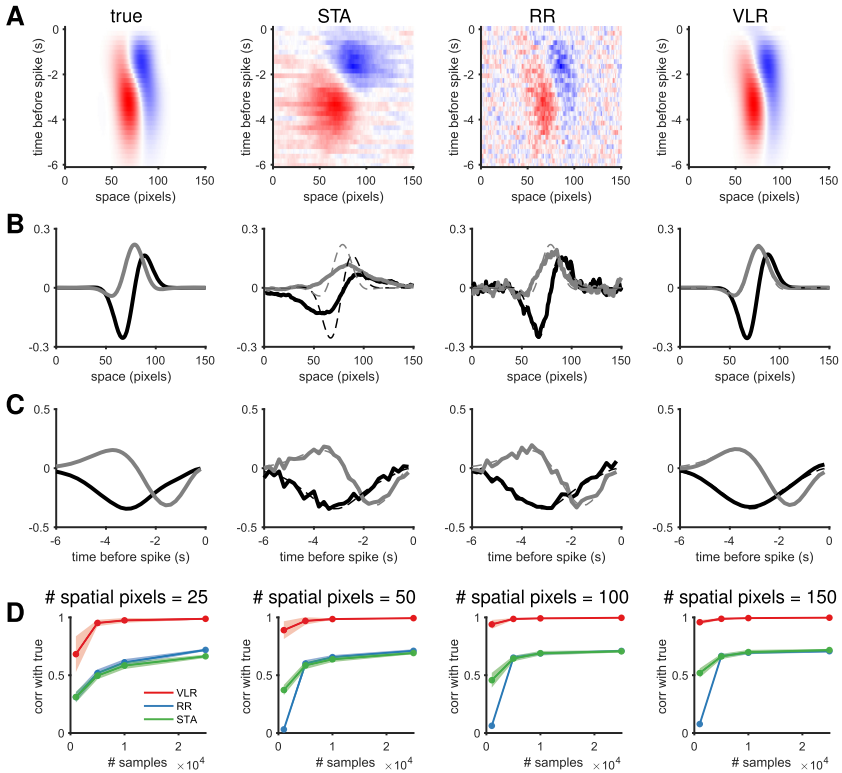
Figure 4: Synthetic data example. (A) True STRF used for generating synthetic responses and recovered estimates using STA, RR, and VLR (rank 2), with ALD priors for the 2D spatial and the TRD prior for the temporal components. Shown estimates are obtained using $1e^4$ training samples. (B, C) The top two spatial (B) and temporal (C) components of the true and recovered STRF estimates. Dashed lines show components of the true STRF for reference. For STA and RR, spatial and temporal STRF components were obtained by taking the leading left and right singular vectors of the STRF in matrix form and rescaling by the size of the true STRF. (D) Correlation between the true and estimated STRF for different numbers of STRF coefficients and amounts of training data. Lines represent the average correlation across 20 repeated simulations with different random seeds; shaded regions represent ± 1 standard error.

Figure 4 shows results under exponentially filtered gaussian stimuli. To assess performance, we computed the STRF estimates using different numbers of receptive field coefficients and different amounts of simulated training data. We compared VLR to the classic spike-triggered average (STA) and Bayesian ridge regression (RR). Figure 4A and 4B show that the STA
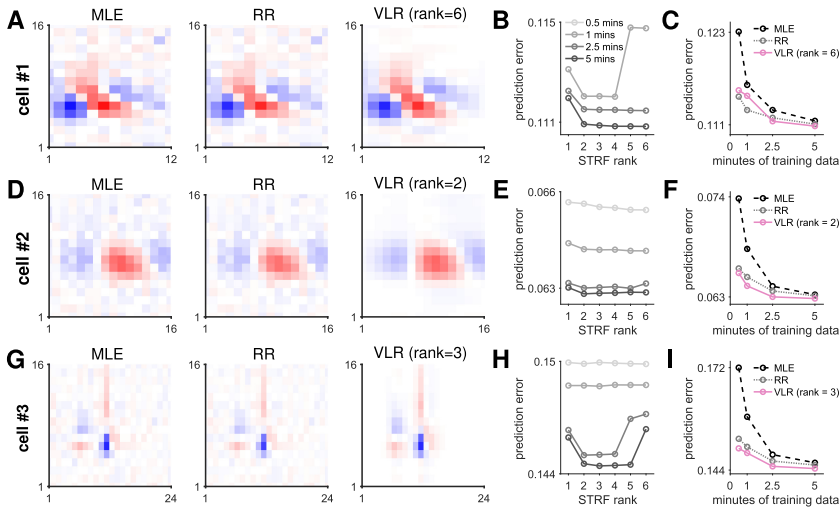
Figure 5: Application to V1 simple cells (A, D, G) Example STRF estimates using MLE, RR, and our VLR approach. Shown estimates are obtained using 5 minutes of training data. (B, E, H) Mean squared error on 10 minutes of held-out test data using different ranks and amounts of training data for the VLR approach. (C, F, I) Mean squared error on 10 minutes of held-out test data when using different amounts of training data for MLE, RR, and an example rank for VLR.

estimate suffers from severe bias due to the correlated stimulus distribution. The RR estimate provides an improvement in terms of the bias but exhibits a high level of variability. Our proposed VLR estimator, in contrast, yields accurate estimates under correlated stimulus distributions, even when using only small amounts of training data. We obtained similar results using stimuli extracted from natural movies (see Figure D.1).

6.2 Application to Macaque V1 Simple Cells. Next, we examined recordings from V1 simple cells in response to random binary "bar" stimuli, previously published in Rust et al. (2005). The stimulus contained 12 to 24 1D spatial bars on each frame that were aligned with each cell's preferred orientation. We used 16 time bins (160 ms) of stimulus history to define the temporal receptive field, so the entire STRF had between 12×16 and 24×16 total coefficients. We examined the performance of the estimator for different ranks and different amounts of training data, and evaluated the mean squared error on held-out test data. For comparison, we also computed the RR estimate and the maximum likelihood estimate (MLE), which in this setting corresponds to the ordinary least squares regression estimate (see Figure 5). We found that VLR outperformed the MLE and RR

estimators, achieving lower prediction errors on held-out test data under varying amounts of training data (see Figures 5C, 5F, and 5I). These results illustrate the heterogeneity of STRF ranks in the V1 simple cell population, as the three example STRFs shown in Figures 5A, 5D, and 5G achieved minimal prediction error for different choices of STRF rank (see Figures 5B, 5E, and 5H).

6.3 Application to Rat Retinal Ganglion Cells. Previous work has identified multiple cell types of retinal ganglion cells (RGCs; Baden et al., 2016). Functional cell-type distinctions are usually based on receptive field properties such as ON/OFF regions, temporal filtering properties, or other structures that can be extracted from STRFs (Chichilnisky & Kalmar, 2002; Field et al., 2007; Frechette et al., 2005). Other differences in response properties, such as STRF rank, may prove valuable for further distinguishing functional cell types.

To further illustrate the power of the VLR estimator, we examined a data set consisting of electrophysiological recordings from rat RGCs in response to binary white noise stimuli. We selected a region of 25×25 pixels in space based on the center of the RF (as estimated by the STA) from the 80×40 overall stimulus, and we considered 30 time lags (0.5 seconds) of stimulus history. This yielded a total of $N = 18{,}750$ STRF coefficients. Standard regression estimators like the MLE require large amounts of training data for STRFs of this size. Bayesian estimators like ASD and ALD are computationally intractable in high-dimensional settings due to the need to store and invert matrices of size N^2. Thus, we compare the VLR estimator with the STA, which can readily be computed even for large STRF sizes such as those considered here.

Figure 6 shows the performance of STA and VLR for different receptive field ranks. VLR achieved lower prediction errors on held-out test data, even when using only small amounts of training data. Figure 6A shows that VLR outperformed the STA estimate computed on 83 minutes of training data using as little as 4.2 minutes of training data and under all assumed STRF ranks (see Figure 6B). Comparing the STA and VLR estimates in the top panels of Figures 6D and 6E, VLR managed to more successfully extract signal from the data and reduce speckle noise in the estimate. The temporal and spatial components of the STRF can be extracted as the leading left and right singular vectors of the matrix unfolding of the STRF tensor, each weighted by the square root of the associated singular value. The STA estimate was dominated by noise and provided much poorer estimates of the spatial and temporal components of the STRF, as demonstrated in the bottom panels of Figures 6D and 6E. The second spatial component of the STA estimate is dominated almost exclusively by speckle noise (see Figure 6D, lower left), while it is structured in the VLR estimate (see Figure 6E, lower left). The rank of the receptive field indicates how many rank-1 space-time components are required to reconstruct the STRF. The singular values of the
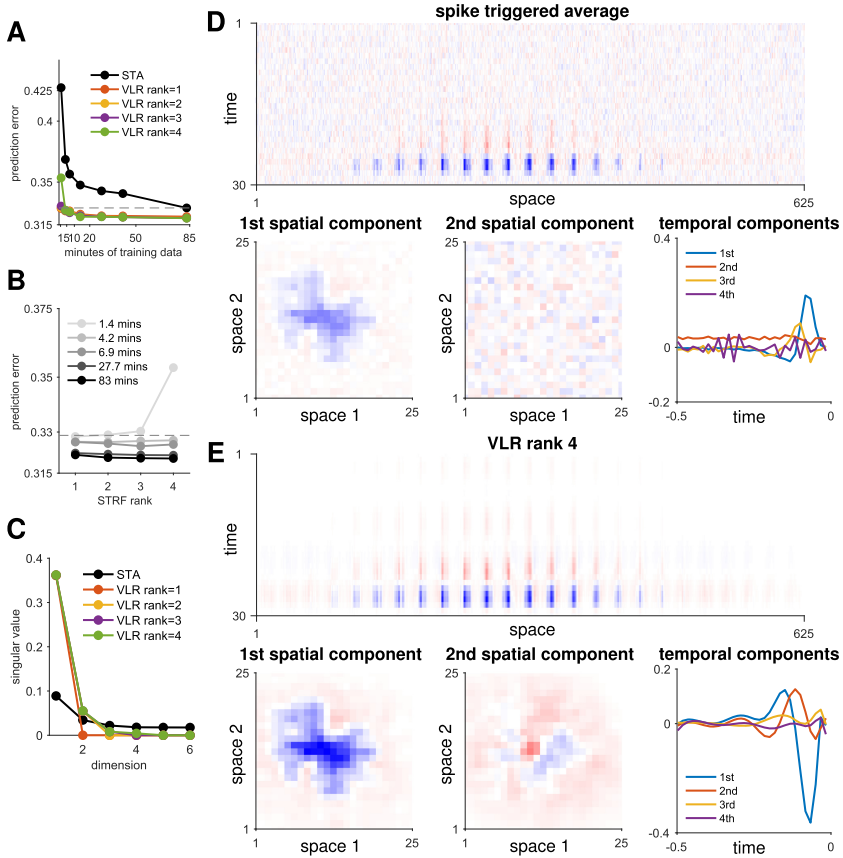
Figure 6: Application to rat retinal ganglion cell (RGC) responses. (A) Mean squared error on 27.66 minutes of held-out test data using different ranks and amounts of training data. Dashed gray line indicates prediction error of the STA using the maximum amount of training data. (B) Same as panel A but plotting the prediction error against ranks for the VLR approach. (C) The top singular values computed on the STRF estimate using STA or the VLR approach under different assumed ranks and using the maximum amount (83 minutes) of training data. (D) STRF estimate computed using STA on 13.83 minutes of training data together with the top two spatial and top four temporal components, determined as the leading left and right singular vectors of the STRF, scaled by the square root of the corresponding singular value. (E) Same as panel D but using the VLR approach with an assumed rank of 4.

STRF indicate the associated weight of each rank-1 space-time component in this reconstruction. As the assumed rank of the VLR estimate grows, our model has the capacity to fit more complex structures in the STRF.

Figure 6C shows the singular values of the STA and and VLR STRF esti-
mates under different assumed ranks. As the assumed rank of the VLR
estimate grows, the associated singular values decay to zero. This demon-
strates that the VLR estimator is able to successfully prune out space-time
components that do not reflect signal in the training data and thus prevent
overfitting to noise. This feature can be attributed to the non gaussian
prior distribution over the effective STRF weights of our model and is a
favorable regularization property. As a result, the prediction error plateaus
after a sufficient rank has been reached, demonstrating that even estimates
obtained under a higher assumed rank generalize well to unseen data (see
Figure 6B).

Finally, in Figure 7 we show summary statistics across a population
of 15 RGCs, which can be grouped into three cell types depending on
STRF properties such as ON/OFF regions or STRF shape. Figure 7A
shows that VLR robustly outperforms the STA, while Figure 7B shows that
STRF ranks are diverse across the assigned cell types, adding to previous
reports on diverse response properties within ON/OFF pathways in mam-
malian RGCs (Ravi et al., 2018). Figure 7C shows the leading spatial com-
ponent of the inferred STRFs across different cells and cell types. Further
examples along with the temporal receptive field components are shown in
Figure D.2.

7 Discussion

In this article, we have introduced a novel method for inferring low-rank
STRFs from neural recordings and have shown a substantial improvement
over previous methods for STRF estimation. Previous approaches like the
STA have low computational complexity and are hence applicable to STRFs
with large sizes but provide biased estimates under correlated stimulus dis-
tributions and classically require large amounts of training data to arrive at
accurate estimates. Previous Bayesian approaches have been shown to be
more data efficient and can provide more accurate STRF estimates, even un-
der correlated stimulus distributions. However, they require matrix inver-
sions that are computationally costly. Thus, applying Bayesian approaches
to large stimulus ensembles remains challenging. Furthermore, all of these
approaches parameterize the full STRF and do not explicitly make use of
potential low-rank structure in the STRF. These methods are therefore gen-
eral but require estimation of large numbers of parameters, which prohibits
their application to large-scale problems.

The VLR estimator we have introduced in this article addresses lim-
itations of previous approaches to STRF estimation in linear encoding
models. We have shown that our method provides accurate estimates in
settings when stimuli are correlated, the receptive field dimensionality
is high, or training data are limited. We developed a new prior covari-
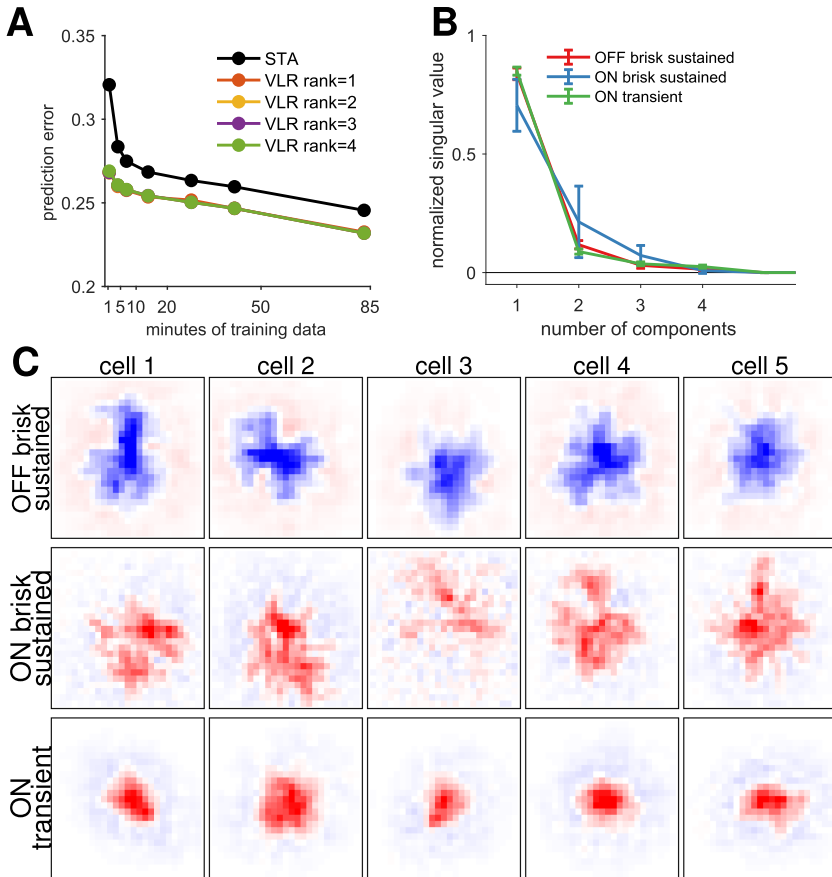ance function that captures the nonstationary smoothness that is typical of

Figure 7: Application to a population of rat retinal ganglion cells. (A) Mean squared error on 27.66 minutes of held-out test data using different ranks and amounts of training data, averaged across 15 cells. (B) Mean and ± 1 standard deviation of the singular values for the rank 4 VLR estimate, separated by cell type. Singular values are normalized to sum to one. Cell types have been classified by hand by the experimenters based on STRF properties. (C) Leading spatial STRF component for five example cells of each of three different cell types.

temporal receptive fields. In combination with previous priors capturing localized smooth structure for spatial receptive fields, our probabilistic model provided a powerful framework for low-rank STRF estimation. While we have focused on these modeling choices in this article, our modeling framework is general and can easily accommodate other choices of prior covariance functions.

The ability to accurately estimate STRFs from limited amounts of data will be important for quantifying changes of receptive field properties over time or during learning. Furthermore, it will open up possibilities for closed-loop experiments, where stimuli are selected actively on each frame to reduce uncertainty about the STRF. Furthermore, being able to use high-dimensional stimuli and correlated stimulus distributions will be important for studying population responses. In this setting, VLR will be useful for estimating the receptive fields of many cells in parallel. Finally, VLR allows for cross-validated estimates of the STRF rank, which may prove important for quantifying single-cell response properties and categorizing cells into specific functional classes.

While we have presented our method in the context of gaussian linear encoding models, it is also possible to extend the variational inference algorithm we presented here to non gaussian observation models in a generalized linear encoding model (GLM). Here, the ability to perform closed-form updates for the posterior means and covariances will be lost and will instead require direct optimization of the variational free energy. Previous approaches have made use of basis functions to improve the scalability of RF estimation in Poisson GLMs (Pillow et al., 2008). However, this typically requires setting parameters like the number, location, and shape of basis functions by hand. Recent work has started to consider circumventing this cumbersome by-hand selection of basis functions by using sparse variational gaussian process approximations instead (Dowling et al., 2020). This approach relies on inducing point methods that have recently gained popularity in improving the scalability of gaussian process methods more generally (Hensman et al., 2013; Titsias, 2009). The framework we presented here is closely related to the basis function approach for GLMs and can be used to determine parameters such as the number, location, and shape of basis functions automatically. The improvements in scalability we achieve due to whitened representations of the receptive field also apply in the context of GLMs. Ultimately, algorithms that can automatically determine such hyperparameters and do not require user-dependent input will contribute to the robustness and reproducability of modeling results.

Appendix A: Bayesian Receptive Field Estimation

Here we review the standard approaches for estimating receptive fields from data under the linear-gaussian encoding model (section 2) using Gaussian priors (section 3).

A.1 Maximum A Posteriori (MAP) Estimation. Given a fixed setting of the hyperparameters θ, it is straightforward to compute the maximum a posteriori (MAP) estimate, which is simply the maximum of the posterior distribution over **k** given the data. The posterior can be computed using Bayes' rule:

$$p(\mathbf{k} \mid \mathbf{y}, \mathbf{X}, \sigma^2, \theta) = \frac{p(\mathbf{y} \mid \mathbf{X}, \mathbf{k}, \sigma^2)p(\mathbf{k} \mid \theta)}{p(\mathbf{y} \mid \mathbf{X}, \theta, \sigma^2)}. \tag{A.1}$$

The numerator consists of the likelihood times the prior, where the likelihood term comes from the gaussian encoding model,

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{k}, \sigma^2) = \prod_{t=1}^{T} \mathcal{N}(y_t \mid \mathbf{k}^\top \mathbf{x}_t + b, \sigma^2) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t - \mathbf{k}^\top \mathbf{x}_t)^2}{2\sigma^2}}, \tag{A.2}$$

and the prior is

$$p(\mathbf{k}|\theta) = \mathcal{N}(\mathbf{k} \mid 0, C_\theta) = \frac{1}{|2\pi C_\theta|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{k}^\top C_\theta^{-1} \mathbf{k}}. \tag{A.3}$$

The denominator term, $p(\mathbf{y} \mid \mathbf{X}, \theta)$, is known as the evidence or marginal likelihood, and represents a normalizing constant that we can ignore when computing the MAP estimate.

In this setting, where the likelihood and prior are both gaussian in $\mathbf{k}$, the posterior is also gaussian. The posterior mean, which is also the MAP estimate, has a closed-form solution,

$$\hat{\mathbf{k}}_{map} = (\Phi^\top \Phi + \sigma^2 C_\theta^{-1})^{-1} \Phi^\top \mathbf{y}, \tag{A.4}$$

where $\mathbf{y}$ denotes the vector of responses and Φ is the design matrix, which contains the corresponding stimulus vectors along its rows:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \qquad \Phi = \begin{bmatrix} - \mathbf{x}_1^\top - \\ \vdots \\ - \mathbf{x}_T^\top - \end{bmatrix}. \tag{A.5}$$

If needed, the posterior covariance is equal to

$$\Lambda = \left(\frac{1}{\sigma^2} \Phi^\top \Phi + C_\theta^{-1} \right)^{-1}. \tag{A.6}$$

A.2 Evidence Optimization and Empirical Bayes. A critical question we have not yet discussed is how to set the noise variance σ^2 and the hyperparameters θ governing the prior. A common approach for selecting hyperparameter values is cross-validation over a grid of candidate values. However, this approach is unwieldy for models with more than one or two hyperparameters.

A popular alternative in Bayesian settings is evidence optimization, also known as type II maximum likelihood estimation (Bishop & Nasrabadi, 2006; MacKay, 1992; Park & Pillow, 2011; Sahani & Linden, 2003). The idea is to compute maximum likelihood estimates for σ^2 and θ using the marginal likelihood, or evidence, which is obtained by marginalizing over the model parameters $\mathbf{k}$. Incidentally, the marginal likelihood is also the denominator in Bayes' rule (see equation A.1). In the current setting, the marginal likelihood is given by

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}, \sigma^2, \theta) &= \int p(\mathbf{y}|\mathbf{X}, \mathbf{k}, \sigma^2) p(\mathbf{k}|\theta) \, d\mathbf{k} \\
&= \frac{|\Lambda|^{\frac{1}{2}}}{(2\pi)^{\frac{T}{2}} |C_\theta|^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2} \mathbf{y}^\top \mathbf{y} + \frac{1}{2} \hat{\mathbf{k}}_{map}^\top \Lambda^{-1} \hat{\mathbf{k}}_{map}\right), \quad (A.7)
\end{aligned}
$$

where $\hat{\mathbf{k}}_{map}$ and Λ are the posterior mean and covariance defined above.

In practice, one performs inference for σ^2 and θ by numerically optimizing the log of the marginal likelihood:

$$
\hat{\sigma^2}, \hat{\theta} = \underset{\sigma^2, \theta}{\operatorname{argmax}} \log p(\mathbf{y}|\mathbf{X}, \sigma^2, \theta), \tag{A.8}
$$

although there are also fixed-point methods available for the ridge regression case (MacKay, 1992; Park & Pillow, 2011).

Once this numerical optimization is complete, we can compute the MAP estimate for the receptive field $\mathbf{k}$ conditioned on these point estimates. This two-step estimation procedure (evidence optimization followed by MAP estimation) is known as *empirical Bayes*:

$$
\hat{\mathbf{k}}_{EB} = \underset{\mathbf{k}}{\operatorname{argmax}} \, p(\mathbf{k}|\mathbf{y}, \mathbf{X}, \hat{\sigma^2}, \hat{\theta}) = (\Phi^\top \Phi + \hat{\sigma^2} C_{\hat{\theta}}^{-1})^{-1} \Phi^\top \mathbf{y}. \tag{A.9}
$$

This approach has been shown to obtain substantially improved receptive field estimates in settings with limited data and correlated stimulus distributions, particularly using ASD or ALD priors (see Park & Pillow, 2011; Sahani & Linden, 2003). However, computing the evidence (see equation A.7) or even the simple MAP estimate (see equation A.4) requires storing and inverting a matrix of size $N \times N$, where $N = n_{x_1} \cdot n_{x_2} \cdot n_t$ is the total number of coefficients in $\mathbf{k}$. The storage costs thus scale as $O(N^2)$, while the computational costs scale as $O(N^3)$, cubically in the number of receptive field coefficients. This severely limits the feasibility of MAP and empirical Bayesian estimators in high-dimensional settings.

Appendix B: Variational Updates for Low-Rank STRF Inference _____

The variational updates require computing

$$q(\mathbf{w}_t) \propto \exp\langle \log P(Y, \mathbf{w}_t, \mathbf{w}_x | \Phi, \theta) \rangle_{q(\mathbf{w}_x)}, \text{ and}$$

$$q(\mathbf{w}_x) \propto \exp\langle \log P(Y, \mathbf{w}_t, \mathbf{w}_x | \Phi, \theta) \rangle_{q(\mathbf{w}_t)}.$$

Both of these distribution will turn out to be gaussians, such that the variational update, or E-step, reduces to computing means and covariances that fully specify the distributions $q(\mathbf{w}_t) = \mathcal{N}(\mathbf{w}_t | \mu_t, \Sigma_t)$ and $q(\mathbf{w}_x) = \mathcal{N}(\mathbf{w}_x | \mu_x, \Sigma_x)$.

To find the variational update for $q(\mathbf{w}_t)$, we evaluate the expected log-joint distribution of the response and receptive field with respect to $q(\mathbf{w}_x)$. The log-joint distribution can be expressed as

$$\log P(Y, \mathbf{w}_t, \mathbf{w}_x | \Phi, \theta) = \log P(Y | \mathbf{w}_t, \mathbf{w}_x, \Phi, \theta) + \log P(\mathbf{w}_t) + c, \quad \text{(B.1)}$$

where c absorbs all terms that are constant with respect to $\mathbf{w}_t$. The log-likelihood of the response can be expressed as

$$\log P(Y | \mathbf{w}_t, \mathbf{w}_x, \Phi, \theta) = -\frac{1}{2\sigma^2}(Y - b - \Phi\mathbf{k})^\top (Y - b - \Phi\mathbf{k}) + c, \quad \text{(B.2)}$$

$$= -\frac{1}{2\sigma^2}(Y - b - \Phi B_\theta \mathbf{w})^\top (Y - b - \Phi B_\theta \mathbf{w}). \quad \text{(B.3)}$$

Using properties of the Kronecker product, we can note that $\mathbf{w}$ can be rewritten in two ways:

$$\mathbf{w} = \text{vec}(W_t W_x^\top) = (I_{p_x} \otimes W_t)\mathbf{w}_x = (W_x \otimes I_{p_t})\mathbf{w}_t,$$

with $\mathbf{w}_t = \text{vec}(W_t)$ and $\mathbf{w}_x = \text{vec}(W_x^\top)$. Substituting this into the expression for the log-joint distribution and taking expectations (denoted by angled brackets), we have

$$\langle \log P(Y, \mathbf{w}_t, \mathbf{w}_x | \theta, X) \rangle_{q(\mathbf{w}_x)} \quad \text{(B.4)}$$

$$= -\frac{1}{2\sigma^2} \langle (Y - b - \Phi B_\theta (W_x \otimes I_{p_t})\mathbf{w}_t)^\top (Y - b - \Phi B_\theta (W_x \otimes I_{p_t})\mathbf{w}_t) \rangle_{q(\mathbf{w}_x)}$$

$$- \frac{1}{2}\mathbf{w}_t^\top \mathbf{w}_t + c \quad \text{(B.5)}$$

$$= -\frac{1}{2\sigma^2} (\text{Tr}[\langle (W_x \otimes I_{p_t})^\top B_\theta^\top \Phi^\top \Phi B_\theta (W_x \otimes I_{p_t}) \rangle_{q(\mathbf{w}_x)} \mathbf{w}_t \mathbf{w}_t^\top] \quad \text{(B.6)}$$

$$- 2\mathbf{w}_t^\top (M^x \otimes I_{p_t})^\top B_\theta^\top \Phi^\top (Y - b)) - \frac{1}{2}\mathbf{w}_t^\top \mathbf{w}_t + c \quad \text{(B.7)}$$

$$= -\frac{1}{2}[\mathbf{w}_t^{\top}(I_{p_tr} + \frac{1}{\sigma^2}\left(\sum_{n=1}^{N}(I_r \otimes S_n)P(\Sigma_x + \mu_x\mu_x^{\top})P^{\top}(I_r \otimes S_n)^{\top}\right)\mathbf{w}_t \quad \text{(B.8)}$$

$$- \frac{2}{\sigma^2}\mathbf{w}_t^{\top}(M_x \otimes I_{p_t})^{\top}B_\theta^{\top}\Phi^{\top}(Y-b)] + c, \quad \text{(B.9)}$$

where $\mu_x = \text{vec}(M^{x\top}) = \text{vec}(\langle W_x^{\top}\rangle_{q(\mathbf{w}_x)})$, P is a commutation matrix (Magnus & Neudecker, 1979) such that $\text{vec}(M_x) = P\text{vec}(M^{x\top}) = P\mu_x$, and $S_n = \text{reshape}(\tilde{\phi}_n, [p_t, p_x])$ and $\tilde{\phi}_n$ is the nth row of ΦB_θ. The above can be brought into a multivariate gaussian form by completing the square, yielding the variational updates:

$$\Sigma_t \leftarrow \left(I_{p_tr} + \frac{1}{\sigma^2}\left(\sum_{n=1}^{N}(I_r \otimes S_n)P(\Sigma_x + \mu_x\mu_x^{\top})P^{\top}(I_r \otimes S_n)^{\top}\right)\right)^{-1} \quad \text{(B.10)}$$

$$\mu_t \leftarrow \frac{1}{\sigma^2}\Sigma_t(M^x \otimes I_{p_t})^{\top}B_\theta^{\top}\Phi^{\top}(Y-b). \quad \text{(B.11)}$$

Similarly, to evaluate the expectation with respect to $q(\mathbf{w}_t)$,

$$\langle \log P(Y, \mathbf{w}_t, \mathbf{w}_x | \theta, X)\rangle_{q(\mathbf{w}_t)}$$

$$= -\frac{1}{2\sigma^2}\langle(Y - b - \Phi B_\theta(W_x \otimes I_{p_t})\mathbf{w}_t)^{\top}(Y - b - \Phi B_\theta(W_x \otimes I_{p_t})\mathbf{w}_t)\rangle_{q(\mathbf{w}_t)}$$

$$- \frac{1}{2}\mathbf{w}_x^{\top}\mathbf{w}_x + c$$

$$= -\frac{1}{2\sigma^2}(\text{Tr}[(W_x \otimes I_{p_t})^{\top}B_\theta^{\top}\Phi^{\top}\Phi B_\theta(W_x \otimes I_{p_t})\langle\mathbf{w}_t\mathbf{w}_t^{\top}\rangle_{q(\mathbf{w}_t)}]$$

$$- 2\mu_t^{\top}(W_x \otimes I_{p_t})^{\top}B_\theta^{\top}\Phi^{\top}(Y-b)) - \frac{1}{2}\mathbf{w}_x^{\top}\mathbf{w}_x + c$$

$$= -\frac{1}{2\sigma^2}(\text{Tr}[(W_x \otimes I_{p_t})^{\top}B_\theta^{\top}\Phi^{\top}\Phi B_\theta(W_x \otimes I_{p_t})(\Sigma_t + \mu_t\mu_t^{\top})]$$

$$- 2\mathbf{w}_x^{\top}(I_{p_x} \otimes M^t)^{\top}B_\theta^{\top}\Phi^{\top}(Y-b)) - \frac{1}{2}\mathbf{w}_x^{\top}\mathbf{w}_x + c$$

$$= -\frac{1}{2\sigma^2}\left(\mathbf{w}_x^{\top}P^{\top}\left(\sum_{n=1}^{N}(I_r \otimes S_n^{\top})(\Sigma_t + \mu_t\mu_t^{\top})(I_r \otimes S_n)\right)P\mathbf{w}_x\right.$$

$$\left. - \mathbf{w}_x^{\top}(I_{p_x} \otimes M_t)^{\top}B_\theta^{\top}\Phi^{\top}(Y-b)\right) - \frac{1}{2}\mathbf{w}_x^{\top}\mathbf{w}_x + c,$$

where $\mu_t = \text{vec}(M^t) = \text{vec}(\langle W_t\rangle_{q(\mathbf{w}_t)})$, and S_n and P are the same as before.

We can again complete the square to obtain a gaussian form in $\mathbf{w}_x$, leading to the variational updates:

$$\Sigma_x \leftarrow \left(P^\top \left(I_{p_x r} + \frac{1}{\sigma^2} \sum_{n=1}^{N} ((I_r \otimes S_n^\top)(\Sigma_t + \mu_t \mu_t^{\ \top})(I_r \otimes S_n) \right) P \right)^{-1}$$

$$\mu_x \leftarrow \frac{1}{\sigma^2} \Sigma_x (I_{p_x} \otimes M^t)^\top B_\theta^\top \Phi^\top (Y - b).$$

Appendix C: M-Step Updates for Low-Rank STRF Inference

The M-step updates for the hyperparameters $\theta = \{\theta_x, \theta_t\}$ can be carried out most efficiently by performing conditional M-steps for either the spatial or temporal parameters while holding the other set of parameters fixed.

Dropping all terms that do not depend on the hyperparamaters θ from the free energy,

$$\mathcal{F} = -\frac{1}{2\sigma^2} \left\langle \sum_{n=1}^{N} (y_n - \text{Tr}(W_t^\top B_{\theta_t}^{t\ \top} \Phi_n B_{\theta_x}^x W_x))^2 \right\rangle$$

$$= -\frac{1}{2\sigma^2} \left\langle \sum_{n=1}^{N} \text{Tr}(W_t^\top B_{\theta_t}^{t\ \top} \Phi_n B_{\theta_x}^x W_x)^2 - 2y_n \text{Tr}(W_t^\top B_{\theta_t}^{t\ \top} \Phi_n B_{\theta_x}^x W_x) \right\rangle$$

$$= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} \langle \text{Tr}(W_t^\top B_{\theta_t}^{t\ \top} \Phi_n B_{\theta_x}^x W_x)^2 \rangle + \frac{1}{\sigma^2} \text{Tr}\left(M_t^\top B_{\theta_t}^{t\ \top} \left(\sum_{n=1}^{N} y_n \Phi_n \right) B_{\theta_x}^x P M_x \right).$$

Here, Φ_n is the nth row of Φ (which is equal to $\mathbf{x}_n$), reshaped to have the same size as the STRF.

To evaluate the first term efficiently, it can be rewritten as follows:

$$\left\langle \text{Tr}(W_t^\top B_{\theta_t}^{t\ \top} \Phi_n B_{\theta_x}^x W_x)^2 \right\rangle$$

$$= \left\langle \left(\sum_{i=1}^{r} W_{t:,i}^\top B_{\theta_t}^{t\ \top} \Phi_n B_{\theta_x}^x W_{x:,i} \right)^2 \right\rangle$$

$$= \left\langle \sum_{i=1}^{r} \text{Tr}\left(B_{\theta_t}^{t\ \top} \Phi_n B_{\theta_x}^x W_{x:,i} W_{x:,i}^\top B_{\theta_x}^{x\ \top} \Phi_n^\top B_{\theta_t}^t W_{t:,i} W_{t:,i}^\top \right) \right\rangle$$

$$+ 2 \left\langle \sum_{i=1}^{r} \sum_{i<j} \text{Tr}\left(B_{\theta_t}^{t\ \top} \Phi_n B_{\theta_x}^x W_{x:,i} W_{x:,j}^\top B_{\theta_x}^{x\ \top} \Phi_n^\top B_{\theta_t}^t W_{t:,j} W_{t:,i}^\top \right) \right\rangle$$

$$= \sum_{i=1}^{r} \mathrm{Tr}\left({B_{\theta_t}^t}^{\top} \Phi_n B_{\theta_x}^x \Lambda_{ii}^x {B_{\theta_x}^x}^{\top} \Phi_n^{\top} B_{\theta_t}^t \Lambda_{ii}^t \right)$$

$$+ 2 \sum_{i=1}^{r} \sum_{i<j} \mathrm{Tr}\left({B_{\theta_t}^t}^{\top} \Phi_n B_{\theta_x}^x \Lambda_{ij}^x {B_{\theta_x}^x}^{\top} \Phi_n^{\top} B_{\theta_t}^t \Lambda_{ji}^t \right).$$

Here, the blocks of the second moment of the vectorized STRF are expressed as

$$\Lambda^x = P \langle \mathbf{w}_x \mathbf{w}_x^{\top} \rangle P^{\top} = P(\Sigma_x + \mu_x \mu_x^{\top}) P^{\top} = \begin{bmatrix} \Lambda_{11}^x & \cdots & \Lambda_{1r}^x \\ \vdots & \ddots & \vdots \\ \Lambda_{r1}^x & \cdots & \Lambda_{rr}^x \end{bmatrix}$$

$$\Lambda^t = \langle \mathbf{w}_t \mathbf{w}_t^{\top} \rangle = \Sigma_t + \mu_t \mu_t^{\top} = \begin{bmatrix} \Lambda_{11}^t & \cdots & \Lambda_{1r}^t \\ \vdots & \ddots & \vdots \\ \Lambda_{r1}^t & \cdots & \Lambda_{rr}^t \end{bmatrix}.$$

It is apparent from the above expression that the conditional M-steps will rely on statistics of the stimulus projected onto the basis we are not optimizing over. In order to evaluate each M-step update as efficiently as possible, these statistics can be precomputed and stored throughout the M-step update. Below, we provide further details for each conditional M-step.

C.1 Spatial Parameter Updates.

$$\mathcal{F}_x = -\frac{1}{2\sigma^2} \sum_{i=1}^{r} \mathrm{Tr}\left({B_{\theta_x}^x}^{\top} \left(\sum_{n=1}^{N} \Phi_n^{\top} B_{\theta_t}^t \Lambda_{ii}^t {B_{\theta_t}^t}^{\top} \Phi_n \right) B_{\theta_x}^x \Lambda_{ii}^x \right)$$

$$- \frac{1}{\sigma^2} \sum_{i=1}^{r} \sum_{i<j} \mathrm{Tr}\left(\left(\sum_{n=1}^{N} \Phi_n^{\top} B_{\theta_t}^t \Lambda_{ji}^t {B_{\theta_t}^t}^{\top} \Phi_n \right) B_{\theta_x}^x \Lambda_{ij}^x {B_{\theta_x}^x}^{\top} \right)$$

$$+ \frac{1}{\sigma^2} \mathrm{Tr}\left(M^{t\top} \left(\sum_{n=1}^{N} y_n {B_{\theta_t}^t}^{\top} \Phi_n \right) B_{\theta_x}^x M^x \right).$$

Now, before computing the parameter updates for the spatial parameters, we can precompute:

$$S_{1_{ji}}^t = \left(\sum_{n=1}^{N} \Phi_n^{\top} B_{\theta_t}^t \Lambda_{ji}^t {B_{\theta_t}^t}^{\top} \Phi_n \right) \qquad S_2^t = \left(\sum_{n=1}^{N} y_n {B_{\theta_t}^t}^{\top} \Phi_n \right).$$

This involves $\binom{r+1}{r-1}$ matrices of size $p_x \times p_x$. While building this matrix is expensive, the size of p_x is pruned using ALD. This cost is incurred only once before performing repeated updates to the spatial hyperparameters.

C.2 Temporal Parameter Updates.

$$
\mathcal{F}_t = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} \sum_{i=1}^{r} \mathrm{Tr}\left({B_{\theta_t}^t}^\top \Phi_n B_{\theta_x}^x \Lambda_{ii}^x {B_{\theta_x}^x}^\top \Phi_n^\top B_{\theta_t}^t \Lambda_{ii}^t \right)
$$

$$
- \frac{1}{\sigma^2} \sum_{n=1}^{N} \sum_{i=1}^{r} \sum_{i<j} \mathrm{Tr}\left({B_{\theta_t}^t}^\top \Phi_n B_{\theta_x}^x \Lambda_{ij}^x {B_{\theta_x}^x}^\top \Phi_n^\top B_{\theta_t}^t \Lambda_{ji}^t \right)
$$

$$
+ \frac{1}{\sigma^2} \mathrm{Tr}\left(M^{t\top} {B_{\theta_t}^t}^\top \left(\sum_{n=1}^{N} y_n \Phi_n \right) B_{\theta_x}^x M^x \right).
$$

Now, before computing the parameter updates for the spatial parameters, we can precompute:

$$
S_{1_{ij}}^x = \left(\sum_{n=1}^{N} \Phi_n B_{\theta_x}^x \Lambda_{ij}^x {B_{\theta_x}^x}^\top \Phi_n^\top \right) \qquad S_2^x = \left(\sum_{n=1}^{N} y_n \Phi_n B_{\theta_x}^x \right).
$$

Analogous to the spatial parameter updates, this involves precomputing and storing $\binom{r+1}{r-1}$ matrices of size $p_t \times p_t$.
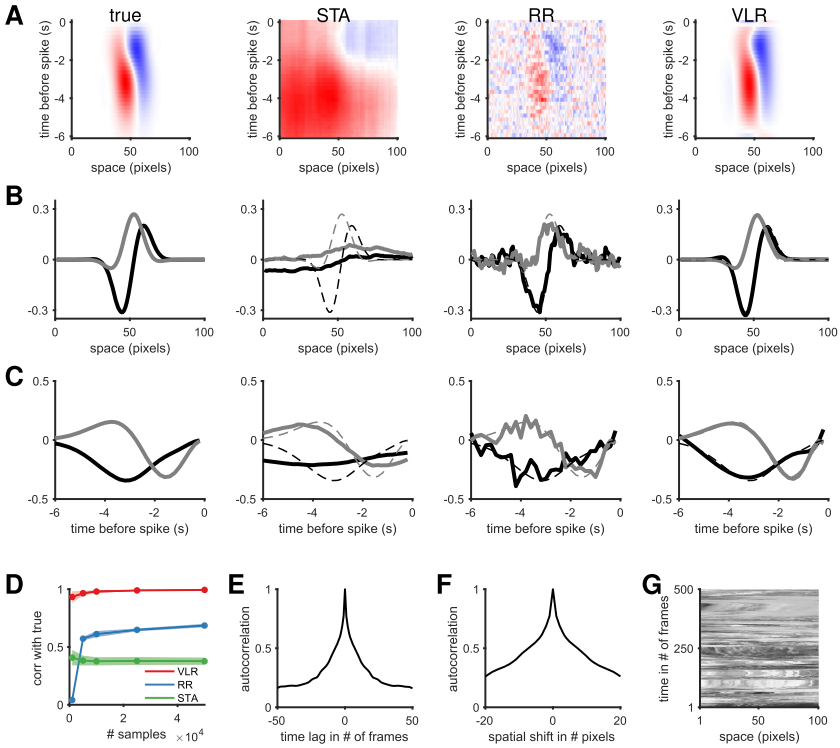
Appendix D

Figure D.1: Simulated responses using natural movie stimulus. Stimuli were extracted as a single row of pixels from a movie in the CatCam data set in Betsch et al. (2004) and used to simulate responses using the same low-rank STRF as in Figure 4. (A) The true STRF that was used for generating synthetic responses and recovered estimates using STA, RR, and our VLR approach with a rank of 2, with ALD for the 2D spatial and TRD for the temporal component. Shown estimates are obtained using 1e4 training samples. (B, C) Same as panel A but plotting the top two spatial (B) and temporal (C) components of the true and recovered STRF estimates. Dashed lines show the spatial and temporal components of the true STRF for reference. Spatial and temporal components of the STA and RR were not accessible by estimation; they were obtained by taking the leading left and right singular vectors of the STRF in matrix form and rescaling by the size of the true STRF. (D) Correlation of the STRF estimate and the true STRF used to generate data for different amounts of training data. Lines represent the average correlation across 15 repeated simulations with different natural movie stimuli; shaded regions represent $\pm$ 1 standard error. (E) Temporal autocorrelation of the natural stimulus. (F) Spatial autocorrelation of the natural stimulus. (G) Five hundred example frames of the natural stimulus.
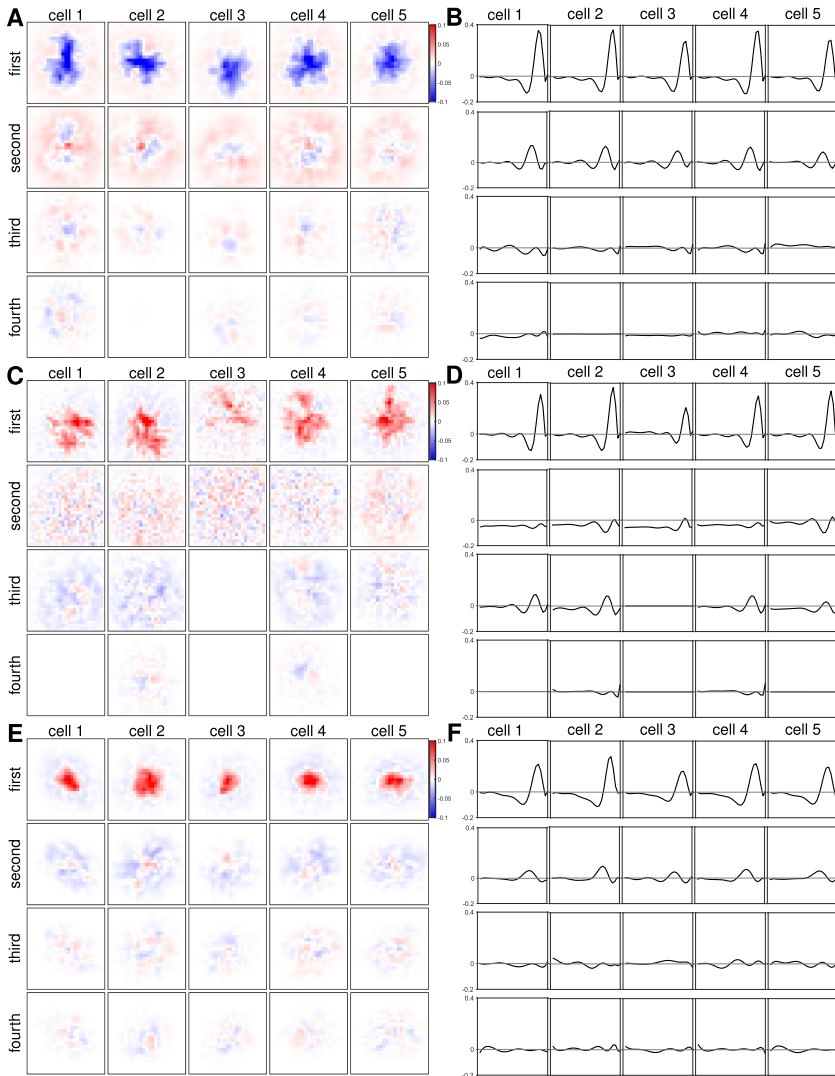
Figure D.2: Inferred spatial and temporal components using VLR for different cell types and example cells. Each panel shows the four spatial (A, C, E) and temporal (B, D, F) low-rank receptive field components for five example cells that have been classified as *OFF brisk sustained* (A, B), *ON brisk sustained* (C, D), or *ON transient* (E, F).

Code Availability

We provide Matlab code for fitting low-rank STRFs under flexible choices of prior covariance functions at https://github.com/pillowlab/VLR-STRF.

Author Contributions

Acknowledgments

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2, 284–299. 10.1364/JOSAA.2.000284

Aoi, M. C., & Pillow, J. W. (2017). *Scalable Bayesian inference for high-dimensional neural receptive fields*. bioRxiv:212217.

Baden, T., Berens, P., Franke, K., Román Rosón, M., Bethge, M., & Euler, T. (2016). The functional diversity of retinal ganglion cells in the mouse. *Nature*, *529*(7586), 345–350. 10.1038/nature16468

Betsch, B. Y., Einhäuser, W., Körding, K. P., & König, P. (2004). The world from a cat's perspective: Statistics of natural videos. *Biological Cybernetics*, *90*(1), 41–50. 10.1007/s00422-003-0434-6

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer.

Calabrese, A., Schumacher, J. W., Schneider, D. M., Paninski, L., & Woolley, S. M. N. (2011). A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. *PLOS One*, *6*(1), e16104.

Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, *12*(2), 199–213. 10.1080/713663221

Chichilnisky, E., & Kalmar, R. S. (2002). Functional asymmetries in on and off ganglion cells of primate retina. *Journal of Neuroscience*, *22*(7), 2737–2747. 10.1523/JNEUROSCI.22-07-02737.2002

David, S. V., & Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems*, *16*(2), 239–260.

deBoer, E., & Kuyper, P. (1968). Triggered correlation. *IEEE Transact. Biomed. Eng.*, *15*, 169–179. 10.1109/TBME.1968.4502561

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, *39*(1), 1–38.

Dowling, M., Zhao, Y., & Park, I. M. (2020). *Non-parameteric generalized linear model*. arXiv:2009.01362.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, *68*(341), 117–130.

Field, G. D., Sher, A., Gauthier, J. L., Greschner, M., Shlens, J., Litke, A. M., & Chichilnisky, E. (2007). Spatial properties and functional organization of small bistratified ganglion cells in primate retina. *Journal of Neuroscience*, *27*(48), 13261–13272. 10.1523/JNEUROSCI.3437-07.2007

Frechette, E. S., Sher, A., Grivich, M. I., Petrusca, D., Litke, A. M., & Chichilnisky, E. (2005). Fidelity of the ensemble code for visual motion in primate retina. *Journal of Neurophysiology*, *94*(1), 119–135. 10.1152/jn.01175.2004

Gerwinn, S., Macke, J. H., & Bethge, M. (2010). Bayesian inference for generalized linear models for spiking neurons. *Frontiers in Computational Neuroscience*, *4*(January). 10.3389/fncom.2010.00012

Hensman, J., Fusi, N., & Lawrence, N. D. (2013). Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence* (pp. 282–290).

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. 10.1080/00401706.1970.10488634

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (vol. 1, pp. 361–379).

Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1187–1211. 10.1152/jn.1987.58.6.1187

Lee, Y., & Schetzen, M. (1965). Measurement of the Wiener kernels of a non-linear system by cross-correlation. *International Journal of Control*, *2*(3), 237–254. 10.1080/00207176508905543

Linden, J. F., Liu, R. C., Sahani, M., Schreiner, C. E., & Merzenich, M. M. (2003). Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. *Journal of Neurophysiology*, *90*(4), 2660–2675. 10.1152/jn.00751.2002

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, *4*(3), 415–447. 10.1162/neco.1992.4.3.415

Magnus, J. R., & Neudecker, H. (1979). The commutation matrix: Some properties and applications. *Annals of Statistics*, *7*(2), 381–394. 10.1214/aos/1176344621

Meyer, A. F., Williamson, R. S., Linden, J. F., & Sahani, M. (2017). Models of neuronal stimulus-response functions: Elaboration, estimation, and evaluation. *Frontiers in Systems Neuroscience*, *10*, 109. 10.3389/fnsys.2016.00109

Park, M., & Pillow, J. W. (2011). Receptive field inference with localized priors. *PLOS Comput Biol.*, *7*(10), e1002219. 10.1371/journal.pcbi.1002219

Park, M., & Pillow, J. W. (2013). Bayesian inference for low rank spatiotemporal neural receptive fields. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, 26 (pp. 2688–2696). Curran.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999. 10.1038/nature07140

Pillow, J. W., & Simoncelli, E. P. (2006). Dimensionality reduction in neural models: An information-theoretic generalization of spike-triggered average and covariance analysis. *Journal of Vision*, *6*(4), 9–9. 10.1167/6.4.9

Qiu, A., Schreiner, C. E., & Escabí, M. A. (2003). Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition. *Journal of Neurophysiology*, *90*(1), 456–476. 10.1152/jn.00851.2002

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT Press.

Ravi, S., Ahn, D., Greschner, M., Chichilnisky, E., & Field, G. D. (2018). Pathway-specific asymmetries between on and off visual signals. *Journal of Neuroscience*, *38*(45), 9728–9740. 10.1523/JNEUROSCI.2008-18.2018

Ringach, D. L. (2004). Mapping receptive fields in primary visual cortex. *J. Physiol.*, *558*(Pt. 3), 717–728. 10.1113/jphysiol.2004.065771

Ringach, D. L., Sapiro, G., & Shapley, R. (1997). A subspace reverse-correlation technique for the study of visual neurons. *Vision Research*, *37*(17), 2455–2464. 10.1016/S0042-6989(96)00247-7

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron*, *46*(6), 945–956. 10.1016/j.neuron.2005.05.021

Sahani, M., & Linden, J. F. (2003). Evidence optimization techniques for estimating stimulus- response functions. In S. Becker, S. Thrun, & K. Overmayer (Eds.), *Advances in neural information processing systems*, 15 (p. 317). MIT Press.

Schwartz, O., Pillow, J. W., Rust, N. C., & Simoncelli, E. P. (2006). Spike-triggered neural characterization. *Journal of Vision*, *6*(4), 484–507. 10.1167/6.4.13

Sharpee, T., Rust, N. C., & Bialek, W. (2004). Analyzing neural responses to natural signals: Maximally informative dimensions. *Neural Computation*, *16*(2), 223–250. 10.1162/089976604322742010

Simoncelli, E. P., Pillow, J. W., Paninski, L., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In M. Gazzaniga (Ed.), *The cognitive neurosciences*, *III* (pp. 327–338). MIT Press.

Smyth, D., Willmore, B., Baker, G., Thompson, I., & Tolhurst, D. (2003). The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *Journal of Neuroscience*, *23*, 4746–4759. 10.1523/JNEUROSCI.23-11-04746.2003

Theunissen, F., David, S., Singh, N., Hsu, A., Vinje, W., & Gallant, J. (2001). Estimating spatiotemporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, *12*, 289–316. 10.1080/net.12.3.289.316

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (pp. 567–574).

Wu, A., Park, M., Koyejo, O. O., & Pillow, J. W. (2014). Sparse Bayesian structure learning with "dependent relevance determination" priors. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *27* (pp. 1628–1636). Curran.