OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Receptive Field Inference with Localized Priors

Mijung Park[1]*, Jonathan W. Pillow[2]*

1 Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas, United States of America, **2** Center for Perceptual Systems, Department of Psychology and Section of Neurobiology, The University of Texas at Austin, Austin, Texas, United States of America

Abstract

The linear receptive field describes a mapping from sensory stimuli to a one-dimensional variable governing a neuron's spike response. However, traditional receptive field estimators such as the spike-triggered average converge slowly and often require large amounts of data. Bayesian methods seek to overcome this problem by biasing estimates towards solutions that are more likely *a priori*, typically those with small, smooth, or sparse coefficients. Here we introduce a novel Bayesian receptive field estimator designed to incorporate *locality*, a powerful form of prior information about receptive field structure. The key to our approach is a hierarchical receptive field model that flexibly adapts to localized structure in both spacetime and spatiotemporal frequency, using an inference method known as empirical Bayes. We refer to our method as *automatic locality determination* (ALD), and show that it can accurately recover various types of smooth, sparse, and localized receptive fields. We apply ALD to neural data from retinal ganglion cells and V1 simple cells, and find it achieves error rates several times lower than standard estimators. Thus, estimates of comparable accuracy can be achieved with substantially less data. Finally, we introduce a computationally efficient Markov Chain Monte Carlo (MCMC) algorithm for fully Bayesian inference under the ALD prior, yielding accurate Bayesian confidence intervals for small or noisy datasets.

Citation: Park M, Pillow JW (2011) Receptive Field Inference with Localized Priors. PLoS Comput Biol 7(10): e1002219. doi:10.1371/journal.pcbi.1002219

Editor: Olaf Sporns, Indiana University, United States of America

Received April 4, 2011; **Accepted** August 19, 2011; **Published** October 27, 2011

Funding: This work was supported by the UT Austin Center for Perceptual Systems. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mjpark@mail.utexas.edu (MP); pillow@mail.utexas.edu (JWP)

Introduction

A fundamental problem in systems neuroscience is to determine how sensory stimuli are functionally related to a neuron's response. A popular mathematical description of this encoding relationship is the "cascade" model, which consists of a linear filter followed by a noisy nonlinear spiking process. The linear stage in this model is commonly identified as the neuron's *spatiotemporal receptive field*, which we will refer to simply as the receptive field (RF) or "filter". The RF describes how a neuron sums up its inputs across space and time. It can also be conceived as the spatiotemporal stimulus pattern that optimally drives the neuron to spike. A large body of literature in sensory neuroscience has addressed the problem of estimating a neuron's RF from its responses to a rapidly fluctuating stimulus, a problem known generally as "neural characterization" [1–17].

Here we focus on a highly simplified encoding model that describes neural responses in terms of a linear filter and additive Gaussian noise [5,11,18]. Although this model gives an imperfect description of real neural responses, the RF estimators that arise from it (such as the spike-triggered average) are consistent under a much larger class of models [7,19,20]. The maximum likelihood filter estimate under the linear-Gaussian model is the whitened spike-triggered average (STA), also known as *linear regression*, *reverse correlation*, or the *first-order Weiner kernel* [1–3]. The STA has an extensive history in neuroscience and has been used to characterize RFs in a wide variety of areas, including retina [4,7,13,21,22], lateral geniculate nucleus [23,24], primary visual cortex [5,25], and peripheral as well as central auditory brain areas [8,9,11,26–28].

The STA is often high-dimensional (containing tens to hundreds of parameters) and generally requires large amounts of data to converge. With naturalistic stimuli, the whitened STA is often corrupted by high-frequency noise because natural scenes contain little power at high frequencies. A common solution is to regularize the filter estimate by penalizing unlikely parameter settings, generally by biasing parameters towards zero (also known as "shrinkage"). Statisticians have long known that biased estimators can achieve substantially lower error rates in high-dimensional inference problems [29,30], and Bayesian methods formalize such biases in terms of a prior distribution over the parameter space. In neuroscience applications, priors for sparse (having many zeros) or smooth (having small pairwise differences) filter coefficients have been used to obtain substantially more accurate RF estimates [9,11,12,15,31].

However, neural receptive fields are more than simply sparse or smooth. They are *localized* in both spacetime and spatiotemporal frequency. This is a structured form of sparsity: RFs contain many zeros, but these zeros are not uniformly distributed across the filter. Rather, the zeros tend to occur outside some region of spacetime and, in the Fourier domain, outside some region of spatiotemporal frequency. Although this property of receptive fields is well-known [32,33], it has not to our knowledge been previously exploited for receptive field inference. Here we introduce a family of priors that can flexibly encode locality. Our approach is to first estimate a localized prior from the data, and then find the maximum a posteriori (MAP) filter estimate under this prior. This general approach is known in statistics as parametric empirical Bayes [34,35]. Our method is directly inspired by previous empirical Bayes estimators designed to

Author Summary

A central problem in systems neuroscience is to understand how sensory neurons convert environmental stimuli into spike trains. The receptive field (RF) provides a simple model for the first stage in this encoding process: it is a linear filter that describes how the neuron integrates the stimulus over time and space. A neuron's RF can be estimated using responses to white noise or naturalistic stimuli, but traditional estimators such as the spike-triggered average tend to be noisy and require large amounts of data to converge. Here, we introduce a novel estimator that can accurately determine RFs with far less data. The key insight is that RFs tend to be localized in spacetime and spatiotemporal frequency. We introduce a family of prior distributions that flexibly incorporate these tendencies, using an approach known as empirical Bayes. These methods will allow experimentalists to characterize RFs more accurately and more rapidly, freeing more time for other experiments. We argue that locality, which is a structured form of sparsity, may play an important role in a wide variety of biological inference problems.

incorporate sparsity [36] and smoothness [11]. We show that locality can be an even more powerful source of prior information about neural receptive fields, and introduce a method for simultaneously inferring locality in two different bases, yielding filter estimates that are both sparse (local in a spacetime basis) and smooth (local in a Fourier basis).

Results

The results section is organized as follows. First, we will describe the linear-Gaussian encoding model and the empirical Bayes framework for receptive field estimation. Second, we will review several previous empirical Bayes RF estimators, to which we will compare our method. Third, we will derive three new receptive field estimators that we collectively refer to as *automatic locality determination* (ALD). We will apply ALD to simulated data and to neural data recorded in primate V1 and primate retina. Finally, we will describe an extension from empirical Bayes to "fully Bayesian" inference under the ALD prior.

Model-based receptive field estimation

A typical neural characterization experiment involves rapidly presenting stimuli from some statistical ensemble and recording the neuron's response in discrete time bins. Let $\mathbf{x}_i$ denote the (vector) stimulus and y_i the neuron's (scalar) spike response at time bin i. Here, $\mathbf{x}_i$ is a vector of spacetime stimulus intensities over some preceding time window that affects the spike response at time bin i.

We will model the neuron's response as a linear function of the stimulus plus Gaussian noise:

$$y_i = \mathbf{k}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0,\sigma^2), \tag{1}$$

where $\mathbf{k}$ denotes the neuron's receptive field and ε_i is a sample of zero-mean, independent Gaussian noise with variance σ^2. This model is the simplest type of cascade encoding model (depicted in Fig. 1 A), and plays an important role in the theory of neural encoding and decoding [5,11,17,28,37,38]. For a complete dataset with n stimulus-response pairs, likelihood is given by

$$P(Y|X,\mathbf{k}) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}\exp\left(-\frac{1}{2\sigma^2}(Y-X\mathbf{k})^{\mathrm{T}}(Y-X\mathbf{k})\right), \tag{2}$$

where $Y = [y_1, y_2, \ldots, y_n]^{\mathrm{T}}$ is a column vector of neural responses and $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^{\mathrm{T}}$ is the stimulus design matrix, with i'th row equal to $\mathbf{x}_i^{\mathrm{T}}$. The maximum likelihood (ML) receptive field estimate is:

$$\hat{\mathbf{k}}_{ml} = \arg\max_{\mathbf{k}} P(Y|X,\mathbf{k}) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y. \tag{3}$$

This estimate, also known as the whitened spike-triggered average, and is proportional to the ordinary spike-triggered average if the stimulus ensemble is uncorrelated, meaning $X^{\mathrm{T}}X \propto I$.

A major drawback of the maximum likelihood estimator is that it typically requires large amounts of data to converge, especially when $\mathbf{k}$ is high-dimensional. This problem is exacerbated for correlated or naturalistic stimulus ensembles, because the high-frequency components of $\mathbf{k}$ are not well constrained by the data. In the Bayesian framework, regularization is formalized in terms of a prior distribution $P(\mathbf{k})$, which tells us that we should bias our estimate of $\mathbf{k}$ toward regions of parameter space that are more probable *a priori*. The posterior distribution, which captures the combination of likelihood and prior information, is given by Bayes' rule:

$$P(\mathbf{k}|X,Y) = \frac{P(Y|X,\mathbf{k})P(\mathbf{k})}{P(Y|X)}. \tag{4}$$

The most probable filter given the data and prior is known as the *maximum a posteriori* (MAP) estimator:

$$\begin{aligned}\hat{\mathbf{k}}_{map} &= \arg\max_{\mathbf{k}} P(Y|X,\mathbf{k})P(\mathbf{k}) \\ &= \arg\min_{\mathbf{k}} \left[\frac{1}{2\sigma^2}(Y-X\mathbf{k})^{\mathrm{T}}(Y-X\mathbf{k}) - \log P(\mathbf{k})\right].\end{aligned} \tag{5}$$

The log prior behaves as a "penalty" on the solution to an ordinary least-squares problem, forcing a tradeoff between minimizing the sum of squared prediction errors and maximizing $\log P(\mathbf{k})$.

Biased estimators can achieve substantial improvements over the maximum likelihood, particularly for high-dimensional problems, without giving up desirable features such as consistency (i.e., converging to the correct value in the limit of infinite data). However, the important question arises: how should one select a prior distribution? (Choosing the *wrong* prior can certainly lead to a worse estimate!)

One common method is to set the prior (or "penalty") by cross-validation. This involves dividing the data into a "training" and "test" set, and selecting the prior for which $\hat{\mathbf{k}}_{map}$ (estimated on the training set) achieves maximal performance on the test set. However, this approach is computationally expensive and may be intractable for a prior with multiple hyperparameters. Empirical Bayes is an alternative method for prior selection that does not require separate training and test data.

Empirical Bayes

Empirical Bayes can be viewed as a maximum-likelihood procedure for estimating the prior distribution from data. It is also known in the literature as *evidence optimization*, *Type II maximum*

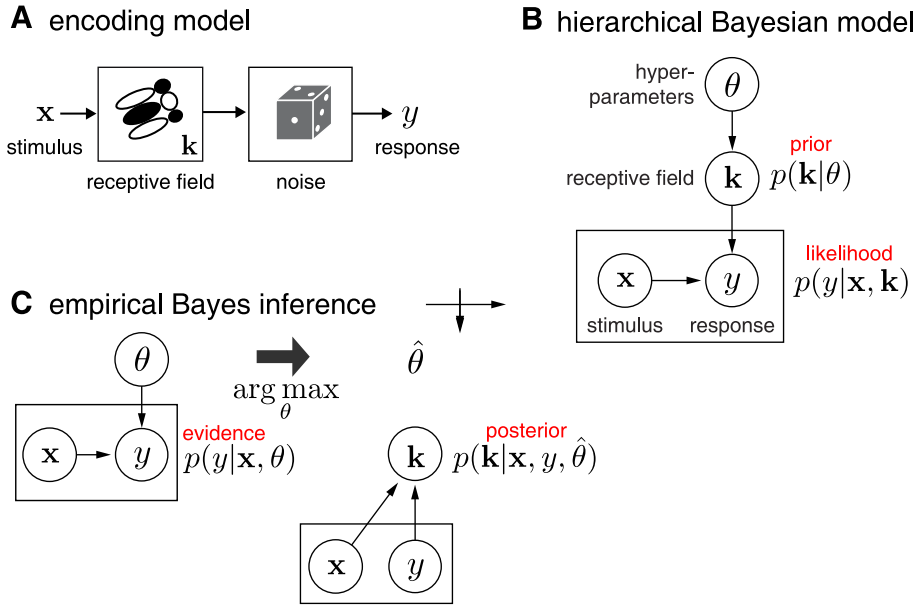


Figure 1. Neural encoding model and empirical Bayes receptive field inference. (**A**) Linear Gaussian encoding model: the stimulus $\mathbf{x}$ is projected on the receptive field $\mathbf{k}$ and Gaussian noise is added to produce the neural response y. (**B**) Graphical model for a hierarchical Bayesian receptive field model. The hyperparameters θ specify a prior over the receptive field $\mathbf{k}$, which together with stimulus $\mathbf{x}$ determines the conditional probability of neural response y. Circles indicate variables, arrows indicate conditional dependence, and the square denotes a pair of variables (stimulus $\mathbf{x}$ and response y) that are observed many times. (**C**) Empirical Bayes involves a two-stage inference procedure: first, maximize the evidence $p(y|\mathbf{x},\theta)$ for θ (*left*), which can be computed by integrating out $\mathbf{k}$ from the generative model in (B); second, maximize the posterior over $\mathbf{k}$ given the data and estimated hyperparameters $\hat{\theta}$ (*right*). See text for details.
doi:10.1371/journal.pcbi.1002219.g001

likelihood, and *maximum marginal likelihood* [11,34,39–41]. The basic idea is that we can compute the probability of the data given a set of hyperparameters governing the prior by "integrating out" the model parameters. This probability is really just a likelihood function for the hyperparameters, so maximizing it results in a maximum-likelihood estimate for the hyperparameters. (Technically, this is *parametric empirical Bayes*, since we will assume a particular parametric form for the prior; see [34,35,42] for a more general discussion).

Let θ denote a set of hyperparameters controlling the prior distribution over $\mathbf{k}$, which we will henceforth denote $P(\mathbf{k}|\theta)$. The posterior distribution over the RF (eq.4) can now be written:

$$P(\mathbf{k}|X,Y,\theta)=\frac{P(Y|X,\mathbf{k})P(\mathbf{k}|\theta)}{P(Y|X,\theta)}. \tag{6}$$

The denominator in this expression is known as the *evidence* or *marginal likelihood*. (Note that we ignored this denominator when finding the MAP estimate (eq.5), since it does not involve $\mathbf{k}$). The evidence is the probability of the responses Y given the stimuli X and the hyperparameters θ, which we can compute by integrating the numerator (eq.6) with respect to $\mathbf{k}$:

$$P(Y|X,\theta)=\int_{\Omega} P(Y|X,\mathbf{k})P(\mathbf{k}|\theta)d\mathbf{k}, \tag{7}$$

where Ω is the parameter space for $\mathbf{k}$. Maximizing the evidence for θ therefore amounts to a maximum likelihood estimate of the hyperparameters. The MAP estimate for $\mathbf{k}$ under this prior is an empirical Bayes estimate, since the prior is learned "empirically" from the data.

Empirical Bayes can therefore be described as a two-stage procedure: **(1)** Maximize the evidence to obtain $\hat{\theta}_{ml}=\mathrm{argmax}_{\theta}P(Y|X,\theta)$; **(2)** Find the MAP estimate for $\mathbf{k}$ under the prior $P(\mathbf{k}|\hat{\theta}_{ml})$. Fig. 1 shows a diagram for this hierarchical receptive field model the steps for empirical Bayesian inference.

Zero-mean Gaussian priors

Following earlier work [11,36,43,44], we will take the prior distribution to be a Gaussian centered at zero:

$$P(\mathbf{k}|\theta)=\mathcal{N}(0,C(\theta)), \tag{8}$$

where $C(\theta)$ is a covariance matrix that depends on hyperparameters θ in some yet-to-be-specified manner. This Gaussian prior together with a Gaussian likelihood (eq.2) ensures the posterior is also Gaussian:

$$P(\mathbf{k}|X,Y,\theta)=\mathcal{N}(\mu,\Lambda),\ \Lambda=(\frac{1}{\sigma^2}X^{\mathrm{T}}X+C^{-1})^{-1},\ \mu=\frac{1}{\sigma^2}\Lambda X^{\mathrm{T}}Y, \tag{9}$$

where μ and Λ are the posterior mean and covariance. The MAP filter estimate $\hat{\mathbf{k}}_{map}$ is simply the posterior mean μ, since the mean and maximum of a Gaussian are the same. Moreover, the evidence (eq.7) can be computed in closed form, since it is the integral of a product of two Gaussians. This allows for rapid optimization of θ. We will in practice maximize the log-evidence, given by:

$$\begin{aligned}\mathcal{E}(\theta)&=\log P(Y|X,\theta)\\&=-\frac{n}{2}\log|2\pi\sigma^2|-\frac{1}{2}\log|C\Lambda^{-1}|+\frac{1}{2}\mu^{T}\Lambda\mu-\frac{1}{2\sigma^2}Y^{\mathrm{T}}Y,\end{aligned} \tag{10}$$

where n is the number of samples (rows) in X and Y. All that remains is to specify the prior covariance $C(\theta)$, which we will explore in detail below.

Before continuing, we wish to distinguish two distinct notions of "dimensionality" for a receptive field. First, dimensionality may refer to the number of parameters or coefficients in $\mathbf{k}$. We will refer to this as the *parameter dimensionality* of the filter, denoted d. Second, dimensionality may refer to the dimensionality of the coordinate space in which the filter is defined. In this sense, a filter with $d = 100$ elements arranged as a 100×1 vector is 1-dimensional (e.g., a temporal filter), while a filter with the same number of elements arranged in a 10×10 matrix is 2-dimensional (e.g., an image filter). We will refer to this as the *coordinate dimensionality* of the filter, denoted D.

Previous methods

We will examine three empirical Bayes RF estimators from the literature: ridge regression [45], Automatic Relevance Determination (ARD) [36,43,44], and Automatic Smoothness Determination (ASD) [11]. Fig. 2 provides an illustrative comparison of these methods, using a simulated example consisting with a 100-element vector filter ($d = 100, D = 1$), stimulated with correlated ("1/F") Gaussian noise stimuli. The true filter was a difference of two Gaussians, and the maximum likelihood estimate (middle left) is badly corrupted by high frequency noise.

First, *ridge regression* assumes a prior with covariance matrix proportional to the identity matrix: $C = \theta^{-1} I$. This treats the filter coefficients as drawn *i.i.d.* from a zero-mean Gaussian prior with precision ("inverse variance") θ. Ridge regression is penalized least-squares estimate with a penalty (eq.5) on the squared L^2 norm of the filter, given by $\sigma^2 \theta \mathbf{k}^{\mathrm{T}} \mathbf{k}$. This penalty shrinks the coefficients of $\mathbf{k}$ towards zero. Larger θ yields smaller filter coefficients, and in the limit of infinite θ, the MAP estimate shrinks to all-zeros. Set correctly, the ridge prior can provide substantial improvement over maximum likelihood, especially when the stimulus autocovariance is ill-conditioned, as it is for naturalistic stimuli (see Fig. 2). Ridge regression is perhaps the most popular and well-known regularization method. Although it is not usually employed in an empirical Bayes framework, it is straightforward (and fast) to maximize the evidence for the ridge parameter θ using a fixed-point rule [36,45]. (See Methods).

Second, *Automatic Relevance Determination* (ARD) [36] assumes a diagonal prior covariance matrix with a distinct hyperparameter θ_i for each element of the diagonal. This resembles the ridge prior covariance except that the prior variance of each filter coefficient is set independently. The prior covariance matrix can be written $C_{ii} = \theta_i^{-1}$, where i ranges over the number of elements in $\mathbf{k}$. It would be intractable to use cross-validation to estimate all the elements in θ (a 100-element vector in Fig. 2), so empirical Bayes plays a critical role for inference. In practice, evidence maximization drives many of the prior variances to zero, making the posterior a delta function at zero for those coefficients. The MAP estimate for these coefficients is therefore zero, making the ARD estimate sparse. The ARD estimate can be computed rapidly using fixed-point methods, expectation-maximization, or variational methods [43,44,46–49]. Fig. 2 (middle column) shows the ARD and the *lasso* estimate [50], the latter of which is the MAP estimate under an exponential (or L^1) prior. We set the lasso parameter here by cross-validation. Both estimates are sparse. The ARD

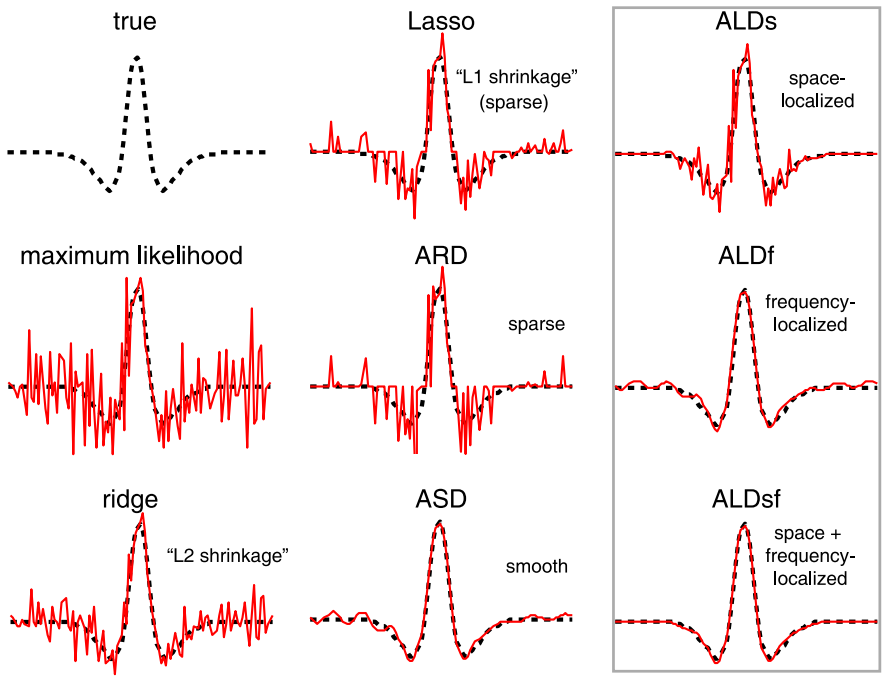


Figure 2. Comparison of estimators for 1D simulated example. A 1D difference-of-Gaussians receptive field with 100 elements was stimulated with 2000 samples of correlated (1/F) Gaussian noise. **Left column:** True filter (top), maximum likelihood (linear regression) estimate (middle), and empirical Bayes ridge regression (L2-penalized) estimate (bottom). **Middle:** Lasso (L1-penalized) estimate (top) and ARD (middle) produce sparse estimates but fail to capture smoothness. The ASD estimate (bottom) captures smoothness, but exhibits spurious oscillations in the tails. **Right column**: Three variants of automatic locality determination (ALD): Spacetime localization (ALDs, top), which identifies a spatial region in which the filter coefficients are large; frequency localization (ALDf, middle), which identifies a local region of the frequency domain in which Fourier coefficients are large, leading to a smooth estimate that closely resembles ASD; and joint localization in spacetime and frequency (ALDsf, bottom), which simultaneously identifies a local region in spacetime and frequency, yielding an estimate that is both smooth and sparse.
doi:10.1371/journal.pcbi.1002219.g002

estimate is actually sparser and less biased towards zero for large coefficients, but both fail to provide a close match to the smooth filter used in this example.

Third, *Automatic Smoothness Determination* (ASD) [11] assumes a non-diagonal prior covariance, given by a Gaussian kernel [51], which is parametrized so that the correlation between filter coefficients falls off as a function of their separation distance. The rationale here is that RFs are smooth in both space and time, so nearby coefficients should be highly correlated, while more distant ones should be more nearly independent. For a 1D filter, the ASD prior covariance takes the form of a "fuzzy ridge", with Gaussian decay on either side of the diagonal. The (i,j)'th element is given by $C_{ij}=\exp(-\rho-\Delta_{ij}/2\delta^2)$, where Δ_{ij} is the squared distance between the filter coefficients k_i and k_j in pixel space, and the hyperparameters $\theta=[\rho,\delta]$ control the scale (analogous to the ridge parameter) and smoothness (the width of the fuzzy ridge), respectively. For filters with higher coordinate dimension (e.g., a 2D spatial filter), the hyperparameters include additional hyperparameters to control smoothness in each direction. Optimization of $\theta=[\rho,\delta_x,\delta_y,\ldots]$ can be achieved by gradient ascent of the log-evidence (see Methods). For our simulated example (Fig. 2, bottom middle), the ASD estimate is indeed smooth due to the correlations in the inferred prior.

Note that for smooth RFs, the ASD prior covariance matrix becomes ill-conditioned, as some of its eigenvalues are very close to zero. This implies that the ASD estimate is sparse, but (unlike ARD) it is not sparse in the pixel basis. Rather, the ASD estimate is sparse in a basis that depends on the hyperparameters (since the eigenvectors of the ASD prior covariance vary with the hyperparameters). The small-eigenvalue eigenvectors tend to have high-frequency oscillations, meaning that the ASD estimate is sparse in a Fourier-like basis, with the prior variance of high-frequency modes set near to zero. In our view, ASD is the current state-of-the-art method for linear filter estimation and indeed (as shown in Fig. 2) it performs far better than previous methods for realistic neural RFs.

Automatic Locality Determination (ALD)

The motivation for our approach is the observation that neural receptive fields tend to be localized in space, time, and spatiotemporal frequency (i.e., Fourier space). Neurons in the visual pathway, for example, tend to integrate light only within some restricted region of visual space and some finite window of time, and respond only to some finite range of spatiotemporal frequencies [25,32,52,53]. This is tantamount to a structured form of sparsity: large groups of coefficients (e.g., those outside some spacetime region) that fall to zero in a dependent manner. Here we describe three prior distributions for exploiting this structure. We refer to these methods collectively as *automatic locality determination (ALD)*.

Locality in spacetime (ALDs). First we formulate a prior covariance matrix $C(\theta)$ that can capture the tendency for RFs to have a limited extent in space and time. We can achieve this with a diagonal covariance matrix, but instead of using a constant diagonal (as in ridge regression) or a vector of hyperparameters along the diagonal (as in ARD), we use a functional form for the diagonal that allows the prior variance to be large for coefficients within some region, and small (decaying to zero) for coefficients outside that region.

We parametrize the local region with a Gaussian form, so that prior variance of each filter coefficient is determined by its Mahalanobis distance (in coordinate space) from some mean location v under a symmetric positive semi-definite matrix Ψ. The diagonal prior covariance matrix is given by:

$$C_{ii}=\exp\left(-\frac{1}{2}(\chi_i-v)^{\mathrm{T}}\Psi^{-1}(\chi_i-v)-\rho\right), \qquad (11)$$

where χ_i is the spacetime location (i.e., filter coordinates) of the i'th filter coefficient $\mathbf{k}_i$, Ψ is a covariance matrix determining the shape and extent of the local region, and ρ sets the overall scale of the prior variance (as in ASD). We refer to this method as ALDs, for automatic locality determination in *spacetime coordinates*. The hyperparameters governing the ALDs prior are $\theta=\{\rho,v,\Psi\}$, which can specify an arbitrary elliptical region of coordinate space where prior variance is large.

Fig. 2 shows the ALDs estimate for the 1D example discussed above. As expected, the RF coefficients are large within a central region, and decay to zero outside it. Fig. 3 (top row) shows the prior variance underlying this estimate (i.e., the diagonal of the prior covariance matrix C) at the maximum-evidence θ. The method can be extended to filters of higher coordinate dimensionality D. In this case, with v is a $D\times 1$ vector and Ψ is a $D\times D$ symmetric, positive definite matrix specified by $D(D+1)/2$ parameters.

Computationally, ALDs is faster than ASD because, although its parametrization is similar, the prior covariance matrix is diagonal. As the localized region described by the hypearparemeters becomes smaller, the prior variance of outer filter pixels falls arbitrarily close to zero, and we can prune these coefficients (as in ARD) because the prior effectively pins them to zero. This reduces the dimensionality of $\mathbf{k}$, making it sparse in pixel space, and making evaluation of the log-evidence (eq.10) faster. The key difference from ARD, however, is that pruning does not take place independently for each coefficient, but occurs systematically as a function of distance from v, the center of some spatiotemporal region.

Note that the ALDs estimator does not assume any functional form for the filter itself. Rather, it seeks to determine (via evidence optimization) only whether there is some elliptical region beyond which the filter coefficients fall to zero. If an RF is *not* localized, the evidence will be maximal when the width of the region specified by Ψ becomes much larger than the area covered by the RF coefficients. In this limit, the diagonal of the prior covariance will be nearly constant, where the ALDs prior is equivalent to the ridge regression prior.

Although ALDs correctly identifies spacetime locality in simulated examples, the estimates it provides are not smooth. The use of a diagonal prior covariance C means that the filter coefficients are independent *a priori* given θ. We can address this shortcoming by considering a different basis for the RF coefficients.

Locality in frequency (ALDf). Neural receptive fields are localized in spatiotemporal frequency as well as in spacetime, which is apparent from their Fourier power spectra [53]. That is, a neuron typically responds to sine waves over some limited range of spatiotemporal frequencies, and is insensitive beyond this range. We can design a prior covariance matrix to capture this structure by employing the ALDs prior in the Fourier domain. We refer to this as the ALDf, for automatic locality determination in *frequency coordinates*.

We can define a Gaussian prior over the Fourier-transformed RF coefficients $\tilde{\mathbf{k}}$ using a diagonal covariance matrix C with diagonal:

$$C_{ii}=\exp\left(-\frac{1}{2}(|M\boldsymbol{\omega}_i|-v)^{\mathrm{T}}(|M\boldsymbol{\omega}_i|-v)-\rho\right), \qquad (12)$$

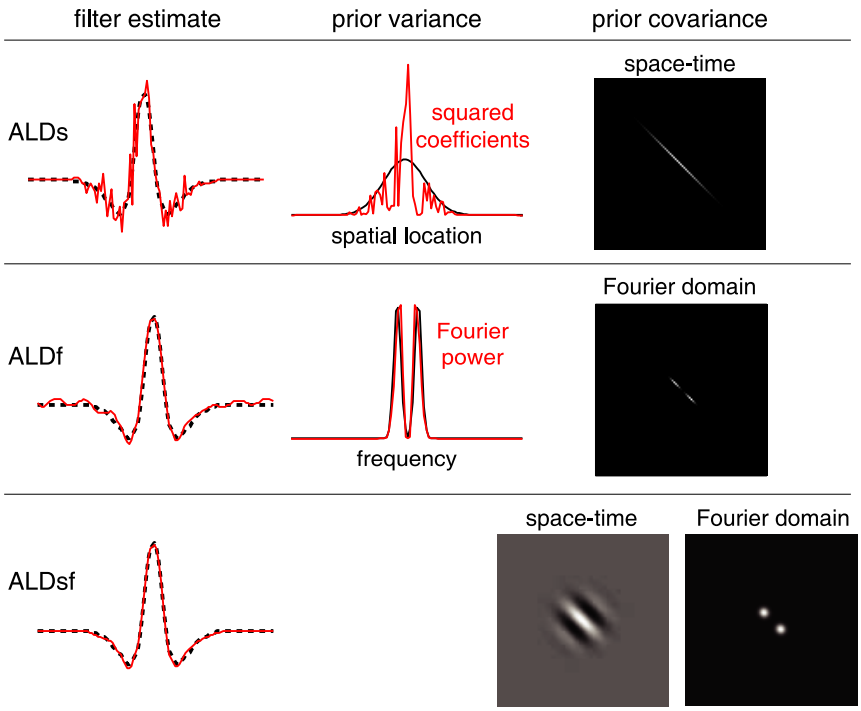


Figure 3. Estimated filters and prior covariances for ALD methods. (Same example filter as shown in Fig. 2). Left column shows the true filter (dotted black) and ALD estimates (red) replotted from the right-most column of Fig. 2. **Top:** Space-localized estimate. The estimated prior variance (black trace, middle) is a Gaussian form that controls the falloff in amplitude of filter coefficients (red) as a function of position. The prior covariance (right) is a diagonal matrix with this Gaussian along the diagonal. The prior is thus independent with location-dependent variance. **Middle:** Frequency-localized estimate. A Gaussian form (reflected around the origin due to symmetries of the Fourier transform) specifies the prior variance as a function of frequency (black trace, middle). The Fourier power of the filter estimate (red) drops quickly to zero outside the estimated region. The prior covariance matrix (right) is diagonal in the Fourier domain, meaning the Fourier coefficients are independent with frequency-dependent variance. **Bottom:** Space and frequency localized estimate. The estimated prior covariance matrix is not diagonal in spacetime or frequency, but takes the form of a "sandwich matrix" that combines the prior covariances from ALDs and ALDf (see text). The resulting prior covariance matrix can be visualized in either the spacetime domain (left) or the Fourier domain (right). It is localized (has a local region of large prior variance) in both coordinate frames, but has strong dependencies (off-diagonal elements), particularly across space.
doi:10.1371/journal.pcbi.1002219.g003

where $\boldsymbol{\omega}_i$ denotes the frequency coordinates for the i'th coefficient of $\tilde{\mathbf{k}}_i$, M is a symmetric matrix, and v describes the mean of (symmetric) elliptical regions in Fourier space. The absolute value ensures reflection symmetry through the origin, a property of the Fourier transform of any real signal, while allowing localized Fourier energy to exhibit orientations in spacetime and to extend over different frequency ranges for different coordinate dimensions. The hyperparameters are $\theta=\{\rho,v,M\}$. (See Methods for details).

The ALDf estimate can be computed efficiently by taking the discrete Fourier transform of stimuli $\{\tilde{\mathbf{x}}_i\}$, maximizing evidence for θ under the diagonal ALDf prior (eq.10), computing $\tilde{\mathbf{k}}_{map}$ in the Fourier domain (eq.9), and then taking the inverse Fourier transform to obtain the spacetime filter $\hat{\mathbf{k}}_{map}$. (See Methods). Note that a filter in D coordinate dimensions requires the D-dimensional Fourier transform. Fig. 2 shows the ALDf estimate for the simulated 1D example, and Fig. 3 (second row) shows the diagonal of the estimated prior covariance and Fourier spectrum of the ALDf estimate, which exhibits modes at ± 4 Hz. Note that this filter is more sparse in the Fourier domain than the space domain, and that the ALDf estimate exhibits correspondingly smaller error than the ALDs. Thus, locality in frequency is more useful than locality in spacetime for smooth RFs.

Although the ASD and ALDf estimates look similar for this 1D example, the latter achieves slightly lower error due to the fact that it also suppresses low frequencies (e.g., the DC component), which are also small for this filter. The ASD prior, in contrast, always assigns highest prior variance to the lowest frequency Fourier components. (This can be seen by inspecting the ASD prior covariance matrix in the Fourier basis). ALDf can be expected to outperform ASD whenever the Fourier spectrum is not a monotonically decreasing function of frequency; however, for realistic examples we considered, the two perform very similarly. The main limitation of both methods is a failure to account for locality in spacetime, which is evident in the ripples present in the tails of both estimates (Fig. 2).

Locality in spacetime and frequency (ALDsf). The two methods described above exploit locality by estimating a diagonal prior covariance matrix in either a spacetime basis (ALDs) or a Fourier basis (ALDf). However, neural receptive fields generally exhibit both kinds of locality at once. One would therefore like to design a prior that simultaneously captures both forms of locality. We can accomplish this by forming a "sandwich" matrix out of the two prior covariance matrices defined above. We define the prior covariance to be:

$$C = C_s^{\frac{1}{2}}(B^{\mathrm{T}} C_f B)C_s^{\frac{1}{2}}, \tag{13}$$

where $C_s^{\frac{1}{2}}$ is the square root of the diagonal ALDs prior covariance (eq.11), C_f is the diagonal ALDf prior covariance matrix (eq.12),

and B is an orthogonal basis matrix for the D-dimensional discrete Fourier transform. (That is, $\tilde{\mathbf{k}} = B\mathbf{k}$, $\mathbf{k} = B^{\mathrm{T}}\tilde{\mathbf{k}}$, and $B^{\mathrm{T}}B = BB^{\mathrm{T}} = I$.) This formulation effectively imposes the two forms of locality in series: first, the spacetime prior covariance (outer matrix); then Fourier transform and the frequency domain covariance (inner matrix). Although there are other combination schemes are possible (see Discussion), we found this one to give the best performance on simulated data. We call the resulting estimate ALDsf, for automatic locality determination in *spacetime and frequency*.

The hyperparameters for ALDsf are union of the ALDs and ALDf hyperparameters, $\theta = [v_s, v_f, \Psi, M, \rho]$, where subscripts s and f indicate parameters for the spatial and frequency domain matrices C_s and C_f, respectively. We perform evidence optimization over this full set of hyperparameters, although it is helpful to initialize with the values estimated for each of the two above methods individually to avoid sub-optimal local maxima. Fig. 2 (bottom right) shows the ALDsf estimate for our 1D example, which is nearly indistinguishable from the true filter. Fig. 3 shows the estimated prior covariance matrix C, represented in both pixel and Fourier bases. As expected, the prior covariance exhibits locality in both coordinates (bases), but is no longer diagonal in either. This indicates that the resulting prior covariance imposes dependencies between neighboring coefficients in both $\mathbf{k}$ and its Fourier transform $\tilde{\mathbf{k}}$.

One useful feature of ALDsf is that it defaults to ALDf if the filter is not localized in space, to ALDs if not localized in frequency, or to ridge regression if not localized in either basis. When the filter is not localized, the evidence will favor regions that are sufficiently broad (i.e., sufficiently large Ψ and M^{-1}) that the matrices C_s or C_f (or both) will approximate the identity matrix, eliminating the prior preference for locality in the corresponding basis. When both C_s and C_f are the identity matrix, the resulting covariance matrix C corresponds to the ridge regression prior.

Application to simulated data

To compare performance with previous receptive field estimators, we began with simulated data. We generated six different 2D spatial receptive fields with varying degrees of locality in space and frequency. Each filter consisted of a 2D array of 20×20 pixels, making for a parameter space of $d = 400$ dimensions. Noisy responses were simulated using 1600 samples of 1/F correlated Gaussian noise according to (eq.1). Results are shown in Fig. 4.

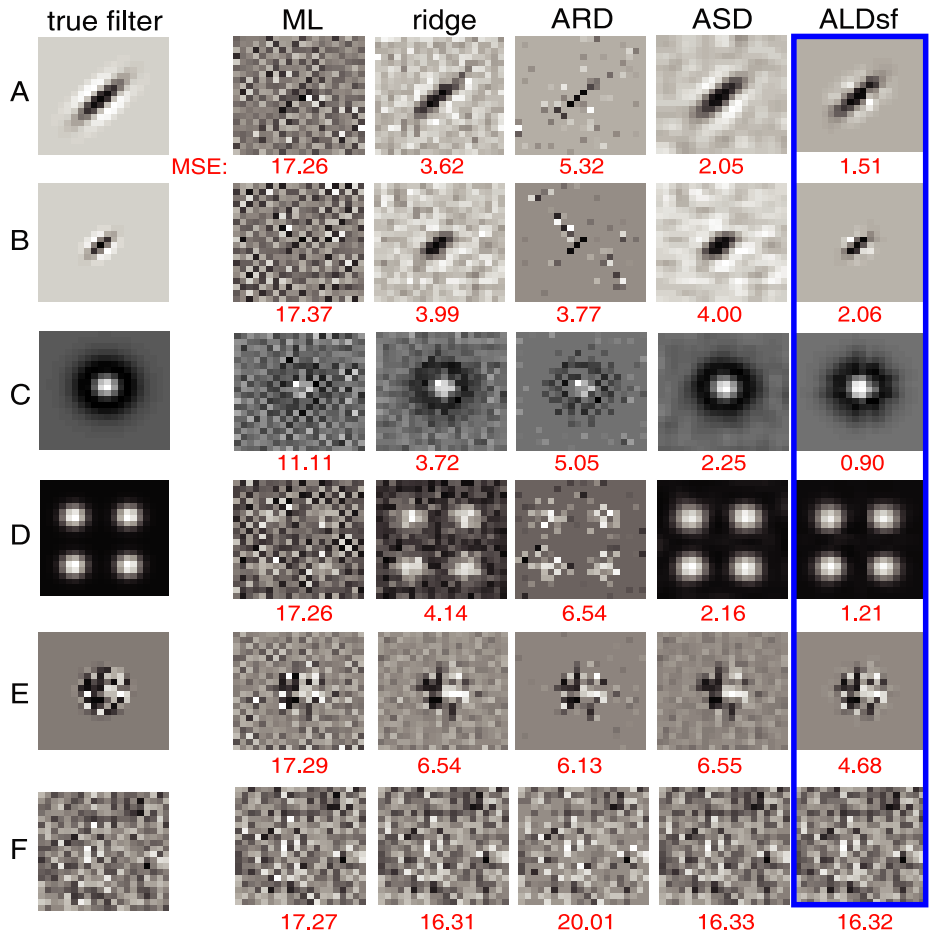


Figure 4. Menagerie of simulated examples. Noisy responses to 1600 random 1/F Gaussian stimuli were simulated and used for training. The leftmost column shows the true filter (a 20×20 pixel image), while subsequent columns show various estimates. The mean squared error of each estimate is indicated below in red. Filters shown include: (**A**) Oriented Gabor filter, typical of a V1 simple cell; (**B**) Smaller Gabor filter; (**C**) center-surround "difference-of-Gaussians" filter, typical of retinal ganglion cells; **D**) grid cell with multiple non-zero regions (localized in the Fourier domain but not in space); (**E**) circularly windowed Gaussian white noise (localized in space but not in frequency); (**F**) full field Gaussian noise (not localized in space or frequency). ALDsf performs at or near the optimum for all examples we examined.
doi:10.1371/journal.pcbi.1002219.g004

Each row of Fig. 4 shows one of the six filters, and the estimates provided by maximum likelihood (ML), ridge regression, ARD, ASD, and (highlighted in blue) ALDsf. The numbers in red below each estimate indicate the mean squared error between the true filter and the estimate. (We did not show ALDs or ALDf because ALDsf always performed best of the three new methods). The simulated examples included: (A) a large Gabor filter; (B) a small Gabor filter; (C) a retina-like center-surround RF; (D) a grid cell RF with several non-zero regions; (E) circularly windowed Gaussian white noise; and (F) a pure Gaussian white noise filter. The grid cell filter did not exhibit strong locality in space, while the windowed white noise did not exhibit locality in frequency, and the pure white noise filter did not exhibit locality in either space nor frequency. Nevertheless, the ALDsf estimate had the smallest error by a substantial margin for all examples except the white noise filter. For the white noise filter, the ridge prior (i.i.d. zero-mean Gaussian) was in fact the "correct" prior. For this example, the ASD and ALDsf estimates were not distinguishable from the ridge regression estimate, consistent with the expectation that both should default to the ridge prior when the evidence did not favor smoothness (ASD) nor locality (ALDsf).

We examined the convergence properties of the various estimators as a function of the amount of data collected. We simulated responses from the first filter in Fig. 4A according to (eq.1), using two kinds of stimuli: Gaussian white noise, and 1/F correlated Gaussian noise, which more closely resembles natural stimuli. The results (Fig. 5) show that the ALDsf estimate achieved the smallest error for both kinds of stimuli, regardless of the number of training samples. The upper plots in Fig. 5 show that for white noise stimuli, traditional estimators (ML and ridge regression) needed more than four times more data than ALDsf to achieve the same error rate. For naturalistic stimuli, traditional estimators needed twenty to thirty times more data. The bottom row of plots shows the ratio of the average mean-squared error (MSE) for each estimate to the average MSE for the ALDsf estimate, showing that the next best method (ASD) exhibits errors nearly 1.8 times larger than ALDsf.

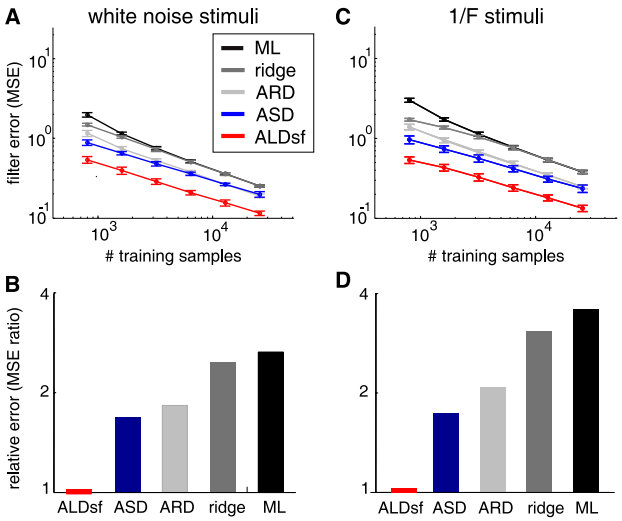


Figure 5. Comparison of error rates on simulated data. Responses of a 20×20 pixel Gabor filter (shown in Fig. 4 A) were simulated using white noise stimuli (left) or "naturalistic" 1/F Gaussian stimuli (right). (**A**): Filter error using white noise stimuli, for varying amounts of training data (See Methods). (**B**) Average filter error under each method. (**C–D**) Analogous to A–B, but for 1/F stimuli. For both kinds of stimuli, ALDsf achieved error rates almost 2 times smaller than ASD, the next best method. By examining horizontal slices through panels (A) and (C), it is apparent that traditional methods (ML and ridge regression) required four times more data on white noise stimuli, and twenty to thirty times more data on 1/F stimuli, to achieve the same error rate as ALDsf.
doi:10.1371/journal.pcbi.1002219.g005

Application to neural data

Next, we compared the various estimators using neural data recorded from simple cells in primate V1 [53]. The stimuli consisted of 16 "flickering bars" aligned with each cell's preferred orientation. We took the receptive field to have a length of 16 time bins, resulting in a 16×16 filter with two coordinate dimensions (space $\times$ time), resulting in a 256-dimensional parameter space. Because the "true" filter was not known, we quantified performance using relative cross-validation error, defined as the prediction error on an 8-minute test set (See Methods). We varied the amount of data used for training, and performed 100 repetitions with randomly selected subsets of the full training data to obtain accurate estimates for each size training set.

Fig. 6 (left) shows ML, ridge regression and ALDsf estimates for an example cell with a 1, 2 or 4 minutes of training data. Numbers in red indicate the average cross-validation error of each estimate. Note that with only 1 minute of data, ALDsf performed nearly as well as ML and ridge regression with 4 minutes of data. The middle panel shows a summary of cross-validation error for each of the five empirical Bayes estimators discussed previously, as a function of the amount of training data. ALDsf once again achieved substantially lower error than other methods. The right panel shows how many times more data were required to achieve the same level of cross-validation error as ALDsf. On average, ALDsf required 1.7 times less data than the next best method (ASD) and five times less data than maximum likelihood.

Fig. 7 shows the ML and ALDsf estimates for all 16 V1 simple cells in the population obtained with 1 minute of training data, as well as the ML estimate obtained using all the data available for each cell (40 minutes of data, on average). Note that for ALDsf recovers the qualitative structure of these RFs even when the underlying RF structure is barely discernible in the 1-minute ML estimate. Also note that the population exhibits substantial variability in RF shape, with many neurons whose RFs would not be well described by a fixed parametric form such as a Gabor filter.

We examined a second dataset of retinal ganglion cells (RGCs) in primate retina, which stimulated with 2D spatiotemporal white noise ("binary flicker") [54,55]. The RFs considered had 3 coordinate dimensions (space $\times$ space $\times$ time), and a 2500-dimensional parameter space (10×10 pixels in space $\times$ 25 8.33 ms-bins in time). Fig. 8 shows the spatial (2D) and the temporal (1D) slices through the estimated 3D RFs (schematized at left). Even with only 1 minute of training data, the ALDsf estimate recovered the qualitative structure of the RF at all time points, including the filters' departure from spacetime separability (i.e., the center pixel has different timecourse than surround). By contrast, the ML estimate is indistinguishable from noise in many places, indicating that ALDsf can reveal qualitative structure that is not visible in the ML estimate. We examined 3 ON and 3 OFF RGCs, and found that error was 18 times higher in ML estimates and 6 times higher in ridge regression estimates than in ALDsf (where error was computed with respect to the ML estimate using a full 20 minutes of data).

Quantifying uncertainty: Bayesian confidence intervals

How can we quantify uncertainty in a receptive field estimate? The error bars shown in Figs. 5 and 6 represent variability in $\hat{\mathbf{k}}$

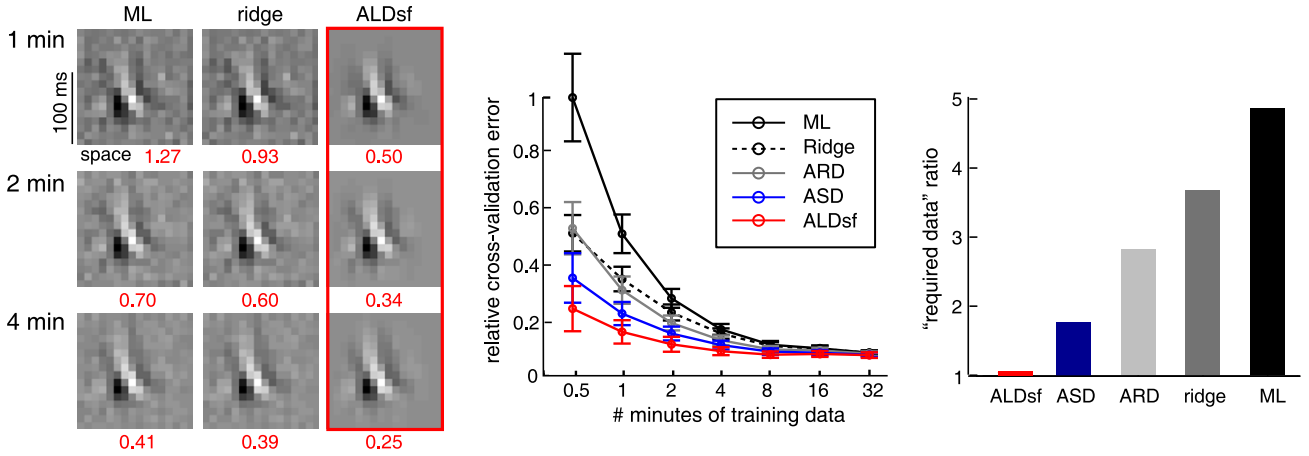


Figure 6. Receptive field estimates for V1 simple cells. (Data from [53]). **Left**: Filter estimates obtained by ML, ridge regression, and ALDsf, for three different amounts of training data (1, 2, and 4 min). Numbers in red beneath each filter indicate relative cross-validation error. **Middle**: Relative cross validation error for each method, averaged across 16 neurons. ALDsf achieved the lowest average error, for all amounts of training data. **Right**: Number of times more training data required by each method to obtain the same error level of as ALDsf with 30s of training data. On average, the ML estimator required 5 times more training data, while ASD required 1.7 times more training data to match the performance of ALDsf.
doi:10.1371/journal.pcbi.1002219.g006

across resampled or permuted datasets. However, we would like to be able to measure the uncertainty in a single estimate given a single set of training data. Given the hyperparameters $\hat{\theta}_{ml}$, the model specifies a Gaussian posterior (eq.9) with mean $\hat{\mathbf{k}}_{map}$ and covariance Λ. The diagonal of Λ specifies the posterior variance for each element of $\mathbf{k}$, giving us 95% credible intervals (Bayesian confidence intervals) of the form

$$c_i = \hat{\mathbf{k}}_i \pm \left[1.96\sqrt{\Lambda_{ii}}\right]. \quad (14)$$

The interpretation of these credible intervals is that, given the data and $\hat{\theta}_{ml}$, $P(\mathbf{k}_i \in c_i) = .95$. More generally, for any unit vector $\mathbf{u}$, the credible interval of size $(1-\alpha)$ for the projection $\mathbf{u}^{\mathrm{T}}\mathbf{k}$ is $c_\mu = \mathbf{u}^{\mathrm{T}}\hat{\mathbf{k}} \pm \left[\Phi^{-1}(-\alpha/2)\sqrt{\mathbf{u}^{\mathrm{T}}\Lambda\mathbf{u}}\right]$, where $\Phi^{-1}(\cdot)$ is the inverse normal cumulative density function.

However, these credible intervals, and the associated Gaussian posterior for $\mathbf{k}$, are conditioned on maximum-evidence estimate of the hyper-parameters $\hat{\theta}_{ml}$. These intervals fail to take into account uncertainty in θ, which may be substantial if the evidence $P(Y|X,\theta)$ is not tightly concentrated around its maximum. The true uncertainty in $\mathbf{k}$ will therefore generally be greater than that captured by the posterior covariance Λ.

Fully Bayesian inference

To accurately quantify uncertainty, we may wish to perform fully Bayesian inference under the priors introduced above. Empirical Bayes (EB) inference can be interpreted as an approximate form of fully Bayesian (FB) inference in a hierarchical model [35,45]. If we incorporate a prior $P(\theta)$ over the hyperparameters at the top level of the graphical model shown in Fig. 1 B, also known as a *hyperprior*, we will have a complete

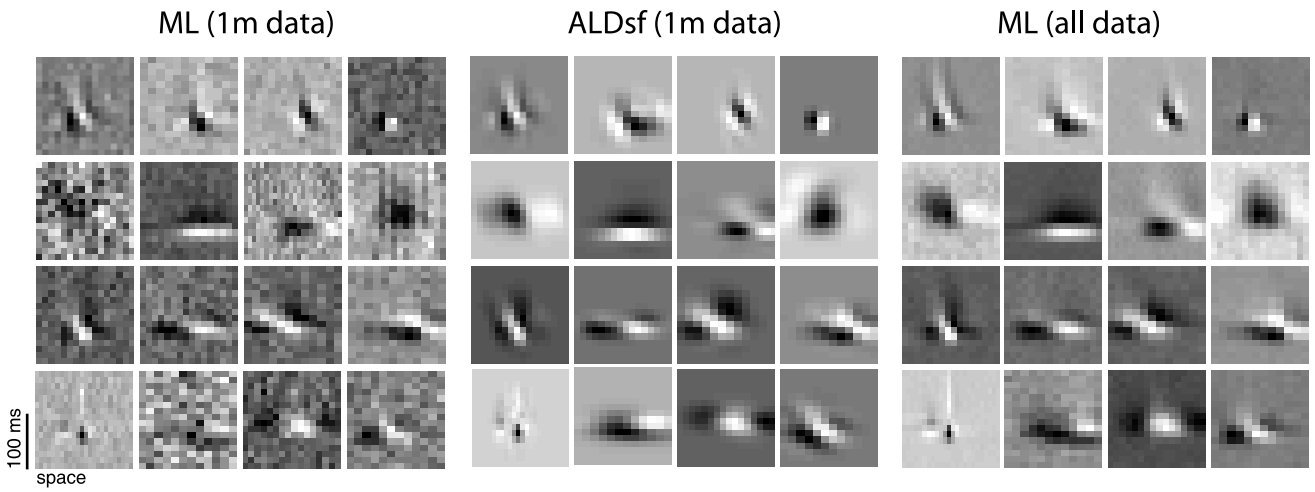


Figure 7. Receptive field estimates for the full set of sixteen V1 simple cells analyzed. (Data from [53]). **Left**: ML filter estimates from 1 minute of training data. **Middle**: ALDsf estimates from 1 minute of training data. **Right**: ML estimates from all data (an average of approximately 40 minutes of data per cell). Note the heterogeneity across cells, and that ALDsf captures the qualitative RF structure even when the 1-minute ML estimate is nearly indistinguishable from noise.
doi:10.1371/journal.pcbi.1002219.g007

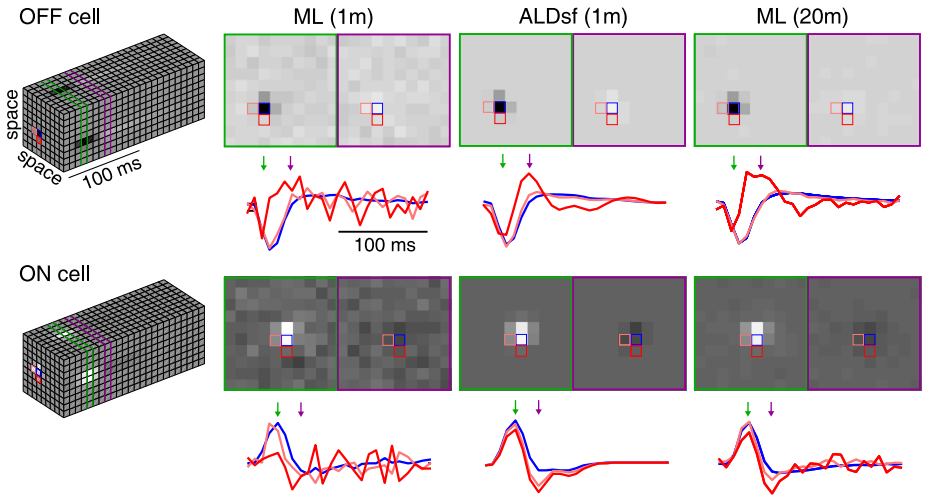


Figure 8. Comparison of 3D receptive field estimates for retinal data. (Data from Chichilnisky lab, [55]). **Top row**: Maximum likelihood and ALDsf estimates for an OFF retinal ganglion cell (RGC) receptive field, stimulated using 1 minute of binary spatiotemporal white noise. Left column shows a schematic of the 10×10 pixel $\times25$ time bin receptive field, containing 2500 total coefficients. Each time bin was 8.33 ms, corresponding to a frame rate of 120 Hz. Colored lines indicate specific pixels whose timecourses shown at right, and spatial time-slices, depicted as images at right (taken at the 4th and 8th time bins, indicated by green and purple arrows, respectively). The ML and ALDsf estimates with 1 minute of training data are shown alongside the ML estimate computed from 20 minutes of data. Pixel time-courses were rescaled to be unit vectors, so that differences in temporal profiles (i.e., spacetime non-separability of filter) can be observed. **Bottom row**: Similar plots for an ON RGC, with spatial profiles shown for the 5th and 8th time bins. In both cases, the ALDsf accurately recovered the shape and timecourse of the RF, while the ML estimate was often indistinguishable from noise. We examined RF estimates from 3 ON and 3 OFF cells, and found that, with 1 minute of training data, the average mean-squared-error between each estimate and a reference estimate (the ML estimate computed with 20 minutes of data) was 18 times larger for ML and 6.6 times larger for ridge regression than for ALDsf.
doi:10.1371/journal.pcbi.1002219.g008

hierarchical model of the neural response. The difference between EB and FB inference for $\mathbf{k}$ comes down to the fact that the FB prior involves marginalizing over θ:

$$P(\mathbf{k})=\int P(\mathbf{k}|\theta)P(\theta)d\theta, \tag{15}$$

while the EB prior is just the conditional distribution $P(\mathbf{k}|\hat{\theta}_{ml})$. When are these priors equivalent or, more importantly, when do the EB and FB estimates agree?

The relationship between EB and FB inference can be understood by examining the posterior distribution over $\mathbf{k}$. The full posterior is

$$P(\mathbf{k}|X,Y)=\int P(\mathbf{k},\theta|X,Y)d\theta=\int P(\mathbf{k}|\theta,X,Y)P(\theta|X,Y)d\theta, \tag{16}$$

where $P(\mathbf{k}|\theta,X,Y)$ is the posterior over k given θ, and $P(\theta|X,Y)$ is proportional to the evidence (i.e. the exponential of (eq.10)) times the hyperprior:

$$P(\theta|X,Y)=\frac{1}{Z}P(Y|X,\theta)P(\theta), \tag{17}$$

where Z is a normalizing constant. Note that if the evidence is proportional to a delta function at its maximum, then the posterior over θ is itself a delta function, $P(\theta|X,Y)=\delta(\theta-\hat{\theta}_{ml})$. The full posterior then reduces to

$$P(\mathbf{k}|X,Y)=\frac{1}{Z}\int P(\mathbf{k}|\theta,X,Y)\delta(\theta-\hat{\theta}_{ml})d\theta=P(\mathbf{k}|X,Y,\hat{\theta}_{ml}), \tag{18}$$

which is the EB posterior (i.e., the posterior over $\mathbf{k}$ conditioned on $\theta=\hat{\theta}_{ml}$). Thus, EB and FB inference are identical when the evidence is proportional to a delta function, and the two methods will in general give similar results whenever the evidence is highly concentrated around its maximum [45]. In general, EB and FB estimates will always agree given enough data, since by central limit theorem, the evidence will concentrate around its maximum with variance that falls as $1/n$. However, for finite datasets, the two may differ.

To examine the proximity of EB and FB estimates and credible intervals, we developed a sampling-based algorithm to perform FB inference under the ALD prior. The factorization shown in (eq.16) suggests an efficient method for sampling from $P(\mathbf{k}|X,Y)$ via Markov Chain Monte Carlo (MCMC), using a Markov chain over the space of the hyperparameters whose stationary distribution is proportional to the evidence. The summary of the algorithm for sampling $P(\mathbf{k}|X,Y)$ is as follows:

$$\begin{aligned}&(\mathbf{1})\text{Sample } \theta_t\sim P(\theta|X,Y) \text{ via MCMC}\\ &(e.g.,\text{Metropolis}-\text{Hastings}[56]),\\ &(\mathbf{2})\text{for each } \theta_t,\text{sample } \mathbf{k}_t\sim P(\mathbf{k}|X,Y,\theta_t),\\ &\text{which is Gaussian(eq.9).}\end{aligned} \tag{19}$$

A nice feature of this approach is that the hyperparameters live in relatively low-dimension (e.g., 5 for a 1D filter and 11 for a 2D filter under ALDsf). The Markov Chain therefore only has to explore this low-dimensional space, instead of the high-dimensional space of $\mathbf{k}$, which contains tens to thousands of parameters in typical cases [57]. Samples $\mathbf{k}_t$ are obtained by drawing from the Gaussian conditioned on each MCMC sample θ_t. These samples may be averaged to the posterior mean $E[\mathbf{k}|X,Y]$, also known as

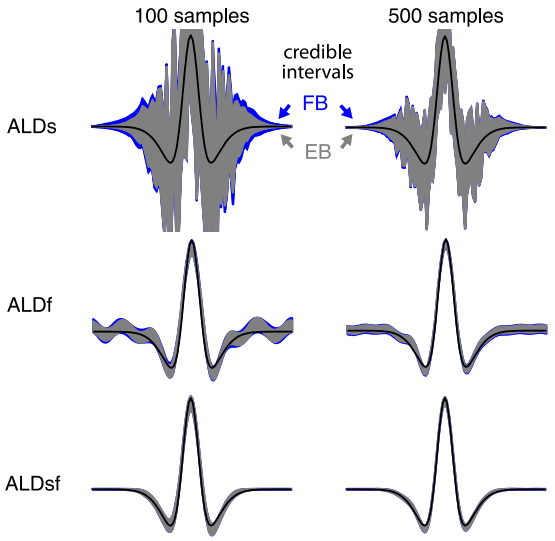


Figure 9. Empirical Bayes (EB) and fully Bayes (FB) credible intervals on simulated data. Left: FB and EB 95% credible intervals, computed from 100 samples of training data, for ALDs (above), ALDf (middle), and ALDsf (bottom). The true filter is shown in black. FB intervals are larger than EB intervals, due to the incorporation of uncertainty in the hyperparameters under fully Bayesian inference. **Right**: Credible intervals computed from 500 samples of training data. As the amount of training data increases, the FB and EB credible regions became indistinguishable, indicating that the evidence is tightly constrained around its maximum. For both amounts of training data, the posterior mean under FB and EB were virtually identical.
doi:10.1371/journal.pcbi.1002219.g009

the *Bayes Least-Squares* estimate, and their quantiles provide credible intervals. (See Method).

Fig. 9 shows a comparison of EB and FB estimates and credible intervals for the 1D simulated example shown previously. The hyperprior $P(\theta)$ was taken to be uniform over a large region (See Methods). For a small dataset, the FB credible intervals were noticeably larger than the EB credible intervals, as expected, owing to the effects of uncertainty in θ [35]. For larger datasets, this discrepancy was much smaller, and was smaller in general for ALDsf than ALDs or ALDf intervals. The EB and FB (Bayes least-squares) filter estimates, however, did not differ noticeably even for small amounts of data. Fig. 10 shows a comparison of EB and FB inference for the V1 neural data presented in Fig. 6. For small datasets, the FB credible intervals were larger than EB intervals, but cross-validation error did not differ noticeably across dataset sizes. This suggests that the higher computational cost of FB inference may not be justified unless one is interested in obtaining accurate quantification of uncertainty from a small or noisy dataset.

Discussion

We have described a new family of priors for Bayesian receptive field estimation that seek to simultaneously exploit locality in spacetime and spatiotemporal frequency. We have shown that empirical Bayes estimates under a localized prior are more accurate than those obtained under alternative priors designed to incorporate sparsity and smoothness. Although the ALD prior does not explicitly impose sparseness or smoothness, the estimates obtained with realistic neural data were both sparse and smooth. Sparsity arises from the fact that pixels outside a central region fall to zero, while smoothness arises from the fact that Fourier coefficients outside some low-frequency region fall to zero. However, for a receptive field dominated by high frequency components, ALD should outperform ASD and other smoothed estimates (e.g., smooth RVM [47], fused lasso [58]), since it can also select regions centered on high frequencies.

We have also derived an algorithm for performing fully Bayesian inference under ALD, ASD, and ridge regression priors. The algorithm exploits the low-dimensionality of the hyperparameter space and the tractability of the evidence to perform MCMC sampling of the posterior over hyperparameters. The full prior takes the form of a *Gaussian scale mixture* [59,60], a mixture of zero-mean Gaussians with covariances $C(\theta)$ and mixing weights $P(\theta)$, resulting in a Gaussian posterior over $\mathbf{k}$ given θ that is trivial to sample. MCMC sampling allows for the calculation of fully Bayesian credible intervals over RF coefficients, which we found to be systematically larger than empirical Bayesian intervals. Nevertheless, we found no differences in the quantitative

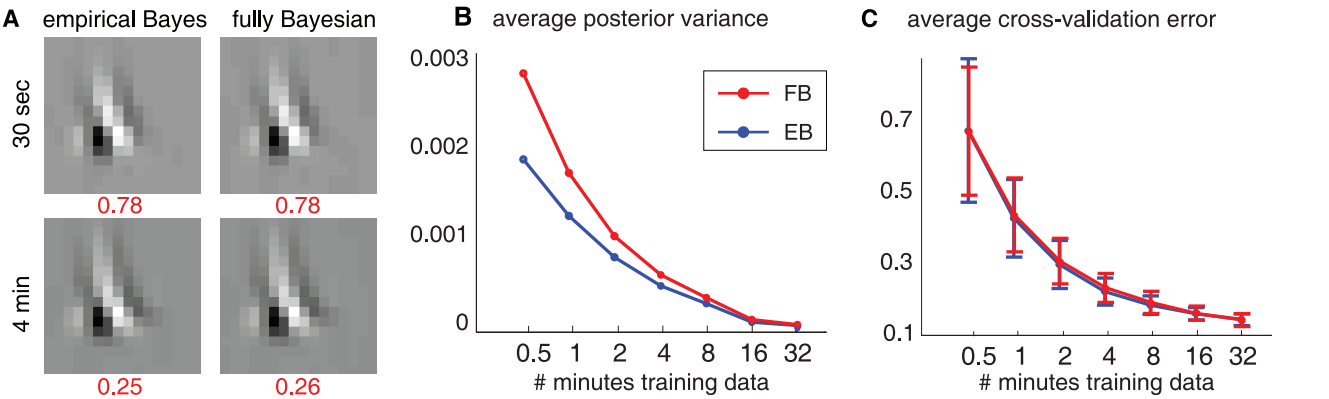


Figure 10. Empirical Bayes (EB) and fully Bayesian (FB) estimates on V1 data. (**A**) ALDsf estimates for a single V1 simple cell under EB and FB inference, from 30 seconds (above) and 4 minutes (below) of training data. There was no significant difference in cross-validation error (numbers below in red, averaged over 100 resampled training sets). (**B**) Marginal posterior variance of RF coefficients, averaged across pixels and across all 16 cells, under EB and FB inference. As expected, FB estimates of the posterior variance were higher, especially for small datasets, reflecting the effects of posterior uncertainty in the hyperparameters. (**C**) Average cross-validation error across 16 cells for FB and EB estimates. For all amounts of training data, error rates were nearly identical, indicating that the FB posterior mean (computed via MCMC) is not superior to the more computationally inexpensive EB estimate.
doi:10.1371/journal.pcbi.1002219.g010

performance of EB and FB receptive field estimates with either simulated or real neural data (Figs. 9 and 10). Of course, both intervals rely on the linear-Gaussian model of the neural response, which may be inaccurate in cases where the neural response noise is highly non-Gaussian (e.g., heavy-tailed).

More generally, this work highlights the advantages of locality as an additional source of prior information in biological inference problems. Shrinkage and sparsity have attracted considerable attention in statistics, and they have advantageous properties for a variety of high-dimensional inference problems [29,50,61,62]. ALD exploits a stronger form of prior information, assuming that large groups of coefficients go to zero in a correlated manner. This may not hold for generic regression problems; for a sparse filter with randomly distributed non-zero coefficients, the ARD estimate substantially outperforms ALD (not shown), but such filters are unlikely to arise in neural systems.

Two general ideas that arise from ALD may be useful for thinking about statistical inference in other biological and non-biological systems. The first is the idea of exploiting an underlying coordinate system or topography. Whenever the regression coefficients can be arranged topographically (e.g., temporally, spatially, spectrally), it may be possible to design a prior that exploits dependencies within this topography using a small number of hyperparameters. This idea is central to ALD as well as to ASD, which uses the distances between RF pixels to set their prior correlation. But other coordinates and prior parameterizations are possible. For example, although ALD performs reasonably well for a simulated grid cell (Fig. 4 D), locality in space does not hold for grid cells, and a prior that exploits the "natural" parameters of grid cell responses (e.g., grid spacing, size, orientation, phase) might perform even better. Optimizing the hyperparameters governing such a prior is tractable with empirical Bayes. The second idea that arises from ALD is that of simultaneously constraining a set of regression coefficients in two (or more) different bases. The ALDsf method combines a local prior in a spacetime basis and a local prior in Fourier basis via a "sandwich matrix" (eq.13), which effectively applies prior constraints in series: first in spacetime and then in frequency. Another solution would be to combine the two priors symmetrically, e.g., using prior covariance $C=(C_s{}^{-1}+C_f^{-1})^{-1}$. (This is the covariance that results from taking the product of the ALDs and ALDf Gaussian priors). We found this formulation to perform slightly worse on test data, but results were similar. Note that the sum of prior covariances $C=(C_s+C_f)$ would *not* achieve the desired goal of imposing the prior constraints simultaneously, since it would prune only those coefficients in the (effective) null space of both C_s and C_f. A large literature has examined regularization and feature selection in overcomplete dictionaries (e.g., "basis pursuit") [62–65], but combining structured prior information defined in different bases poses an intriguing open problem.

One potential criticism of ALD is that the linear-Gaussian encoding model (eq.1) is overly simplistic. Despite its simplicity, this model has a long history in the neural characterization literature [5,11,18], and the estimators considered here are consistent (i.e., converge asymptotically) for responses generated by any linear-nonlinear response model, so long as the stimuli are elliptically symmetric and the expected STA is non-zero [20]. We addressed whether the linear-Gaussian modeling assumption undermines our results by re-analyzing the V1 simple cell data with maximally informative dimensions (MID) [66], an information-theoretic estimator that incorporates neural nonlinearities and Poisson spiking. The results (shown in Supporting Information (Text S1), Fig. S1), indicate that MID errors were large, comparable in size to those of the maximum likelihood (linear regression) estimate. Even when comparing to the MID filter computed from test data, ALDsf outperformed MID by a substantial margin. This shows that the limitations of the linear-Gaussian model do not substantially undermine its performance on simple cells. However, we have applied ALD only to neurons whose responses exhibit a quasi-linear relationship to the stimulus. ALD would indeed fail for a neuron with a symmetric nonlinearity (e.g., squaring) and cannot recover multiple filters (e.g., those driving a complex cell). A variety of techniques exist estimating multi-dimensional feature spaces (e.g., spike-triggered covariance (STC) [67–69], MID [20,66], iSTAC [70], spike-triggered ICA [71]). However, the "kernel trick" [17,41], which involves using linear methods on nonlinearly transformed stimuli, provides the simplest method for extending ALD to nonlinear response models. Many nonlinear transformations (e.g., transforming the stimulus to its Fourier power [72]) preserve the topography of the underlying stimulus, making this approach directly applicable to ALD.

One advantage of the linear-Gaussian model is its computational tractability. ALD is fast because the evidence can be calculated and optimized entirely from the sufficient statistics $X^{\mathrm{T}}X$, $X^{\mathrm{T}}Y$, and $Y^{\mathrm{T}}Y$ (the raw stimulus covariance, the STA, and sum of squared responses, respectively). This means that the computational cost does not scale with the amount of data (unlike MID and maximum-likelihood point process methods). Evidence optimization is also much faster than cross-validation, particularly with the $5\sim15$ hyperparameters employed by ALDsf. The computational cost of ALD is still at least $O(d^3)$ in the number of filter coefficients, since evidence evaluation requires left-division by matrices of size $d\times d$. However, the number of approximately zero coefficients often falls considerably during optimization, and eliminating these coefficients by thresholding small eigenvalues of C can speed convergence considerably.

Given the hyperparameters, the log-posterior over $\mathbf{k}$ is concave, with a single maximum that can be computed in closed form (eq.5). Although the log-evidence (eq.10) is *not* concave in the hyperparameters θ, there are far fewer hyperparameters than parameters, making ALD far easier than non-convex optimization in the full space of $\mathbf{k}$ (e.g., as in MID). We can maximize the evidence more rapidly by using its first and second derivatives, which we can compute analytically (see Methods). We also exploit a heuristic strategy for initializing the ALDsf hyperparameters using the estimates from ridge regression (to identify the scale), ALDs (to identify a spatiotemporal region) and ALDf (to identify a Fourier region). Although it is substantially more computationally expensive, the fully Bayesian estimate based on MCMC avoids the issue of local maxima because it explores the entire evidence surface, not just its modes.

However, we do not ultimately view ALD and other model-based or information-based methods as in conflict. Rather, we regard ALD as providing a prior distribution over RFs that can be combined with any likelihood. Computing and optimizing the evidence under nonlinear models with non-Gaussian noise represents an important direction for future work. We suggest that locality is a general feature of neural information processing and anticipate that it will be useful for neural characterization in a wide variety of brain areas, including those where response properties are not yet well understood [73]. We expect hierarchical models and empirical and fully Bayesian inference methods to find application to a wide range of problems where structured prior information can be usefully defined.

Methods

Implementation of RF estimators

Ridge regression. For simulated and real datasets, we computed the empirical Bayes ridge regression estimate as follows. First, we initialized the noise variance to $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{k}_0^{\mathrm{T}}\mathbf{x}_i)^2$ where $\mathbf{k}_0 = (X^{\mathrm{T}}X)^{-1}X^T Y$ was the ML estimate, and the inverse prior variance to $\theta = 10^{-6}$. Then, we ran an iterative fixed-point algorithm [36,45] to optimize the evidence for θ and σ^2,

$$\theta^{new} = \frac{d - \theta \mathrm{Tr}(\Lambda)}{\sum_j \mu_j^2}, \quad (\sigma^2)^{new} = \frac{(Y - \mu^{\mathrm{T}}X)(Y - \mu^{\mathrm{T}}X)^{\mathrm{T}}}{n - \sum_j (1 - \theta\Lambda_{jj})}, \quad (20)$$

where d is the parameter dimensionality (number of RF coefficients), n is the number of samples, μ is the posterior mean (eq.9), and Λ_{jj} are the diagonal elements of the posterior covariance. The posterior mean and covariance are recomputed after each update to σ^2 and θ. Note that (following [11,36]) we treat the noise variance σ^2 as a hyperparameter, maximizing instead of integrating it out, which is computationally more tractable, although technically it appears in the likelihood rather than the prior.

Automatic Relevance Determination (ARD). We initialized the noise variance and ARD hyperparameters using the maximum-evidence values obtained from the ridge regression prior: $\sigma^2 = \sigma^2_{ridge}$ and inverse prior variance $\theta_i = \theta_{ridge}$. Then, we updated θ_i and σ^2 using the fixed-point rule given in [36]:

$$\theta_i^{new} = \frac{1 - \theta_i \Lambda_{ii}}{\mu_i^2}, \quad (\sigma^2)^{new} = \frac{(Y - \mu^{\mathrm{T}}X)(Y - \mu^{\mathrm{T}}X)^{\mathrm{T}}}{n - \sum_i (1 - \Lambda_{ii}\theta_i)}. \quad (21)$$

Lasso. We computed the Lasso estimate using the algorithm introduced in [74,75], using software available at http://www-stat.stanford.edu/~tibs/glmnet-matlab. This implementation performs cyclical coordinate descent in a pathwise fashion. We used a test dataset with 2000 samples to find the optimal value of the lasso parameter (Fig. 2).

Automatic Smoothness Determination (ASD). We computed the ASD estimate by gradient ascent of the log-evidence function, following the methods in [11]. Briefly, we initialized using the hyperparameter estimates from ridge regression: $\sigma^2 = \sigma^2_{ridge}$, and $\rho = -\log(\theta_{ridge})$, initialized the smoothness parameter δ to 1, then minimized the negative log-evidence for (σ^2, ρ, δ) using analytically computed gradients (provided in [11]) and Hessians, which we derive below. We performed minimization using *fmincon* in MATLAB, with boundary conditions for the hyperparameters and the noise variance set to $-20 \le \rho \le 20$, $10^{-6} \le \delta \le 10^6$, and $10^{-6} \le \sigma^2 \le 10^6$, which we selected to be far larger than the range of probable values.

Automatic Locality Determination (ALD). We computed ALD estimates by numerical optimization of the log-evidence using the analytically computed gradient and Hessian (second derivative matrix). For notational convenience, we will denote $\frac{\partial}{\partial\theta}A$, the first derivative of a quantity A with respect to a parameter θ, as $A_{(\theta)}$, and denote the second derivative $\frac{\partial^2}{\partial\theta_2\partial\theta_1}A$ as $A_{(\theta_1\theta_2)}$.

The first derivatives of the log-evidence $\mathcal{E}$ with respect to the hyperparameters θ and the observation noise σ^2 are given by [11]:

$$\begin{aligned} \mathcal{E}_{(\theta)} &= \frac{1}{2} Tr\left[(C - \Lambda - \mu\mu^T)C^{-1}_{(\theta)}\right], \\ \mathcal{E}_{(\sigma^2)} &= \frac{1}{2\sigma^2}\left(-n + d - Tr[\Lambda C^{-1}]\right) + \frac{1}{2\sigma^4}R^2, \end{aligned} \quad (22)$$

where $C^{-1}_{(\theta)}$ is the derivative of C^{-1} with respect to θ, n is the number of training samples, d is dimensionality of $\mathbf{x}$, $R^2 = (Y - X\mu)^T(Y - X\mu)$ is the squared residual error, and Λ and μ are the posterior covariance and mean, respectively (eq.9). The corresponding second derivatives are given by:

$$\begin{aligned} \mathcal{E}_{(\theta_i\theta_j)} &= \mathrm{Tr}\left[\left(C_{(\theta_j)} - \Lambda_{(\theta_j)} - 2\mu\mu^T_{(\theta_j)}\right)C^{-1}_{(\theta_i)}\right] \\ &\quad + \mathrm{Tr}\left[(C - \Lambda - \mu\mu^T)C^{-1}_{(\theta_i\theta_j)}\right], \\ \mathcal{E}_{((\sigma^2)^2)} &= -\frac{1}{2\sigma^4}\left(-n + d - \mathrm{Tr}[\Lambda C^{-1}]\right) - \frac{1}{2\sigma^2}\mathrm{Tr}[\Lambda_{(\sigma^2)}C^{-1}] \\ &\quad - \frac{1}{\sigma^6}R^2 + \frac{1}{\sigma^4}\left(-Y^T X\mu_{(\sigma^2)} + \mu X^T X\mu_{(\sigma^2)}\right), \\ \mathcal{E}_{(\theta\sigma^2)} &= -\frac{1}{2}\mathrm{Tr}\left[\left(\Lambda_{(\sigma^2)} + 2\mu\mu^T_{(\sigma^2)}\right)C^{-1}_{(\theta)}\right]. \end{aligned} \quad (23)$$

These expressions involve the derivatives of Λ and μ and with respect to θ and σ^2, which are matrices and vectors of the same size as Λ and μ, given by:

$$\begin{aligned} \Lambda_{(\theta)} &= -\Lambda C^{-1}_{(\theta)}\Lambda, \qquad \mu_{(\theta)} = -\Lambda C^{-1}_{(\theta)}\mu, \\ \Lambda_{(\sigma^2)} &= \frac{1}{\sigma^4}\Lambda(X^T X)\Lambda, \qquad \mu_{(\sigma^2)} = -\frac{1}{\sigma^2}\Lambda C^{-1}\mu. \end{aligned} \quad (24)$$

Here, $C^{-1}_{(\theta)} = -C^{-1}C_{(\theta)}C^{-1}$. Note that C^{-1} is numerically unstable when C becomes ill-conditioned. Thus we never compute the inverse C explicitly. Instead, we exploit the Woodbury matrix identity to compute the evidence and other quantities using matrices that are well-conditioned. The resulting expressions involve matrix left division instead of inversion (computed via the backslash operator in Matlab; see Supplementary Information for details). Code is available from the last authors website (http://pillowlab.cps.utexas.edu/code.html). Below, we provide the partial derivatives $C_{(\theta)}$ for the various ALD hyperparameters, which are all that is required for computing the gradient and Hessian of $\mathcal{E}$.

ALD in spacetime (ALDs). The hyperparameters governing the ALDs prior covariance matrix C are (v, Ψ, ρ), where v defines the mean of the localized RF, Ψ defines its elliptical extent, and ρ defines the scale of the prior variance (as in ASD). For filters with coordinate dimension $D > 2$, v is a vector and Ψ is a $D \times D$ matrix that we parametrized (e.g., for $D = 2$) as

$$\Psi = \begin{pmatrix} \psi_1^2 & \phi\psi_1\psi_2 \\ \phi\psi_1\psi_2 & \psi_2^2 \end{pmatrix}. \quad (25)$$

For a one-dimensional filter, $\Psi = \psi^2$, while for $D = 3$, we have Ψ defined in terms of six hyperparameters $[\psi_1, \psi_2, \psi_3, \phi_1, \phi_2, \phi_3]$. The first and second derivatives of C with respect to these hyperparameters (for $D = 1$) are given by:

$$
\begin{aligned}
&C_{(\rho)} = -C, \qquad C_{(\rho^2)} = C, \qquad C_{(\rho v)} = -C_{(v)},\\
&C_{(v)} = \frac{1}{\psi^2}\Delta C, \qquad C_{(v^2)} = \left(-\frac{1}{\psi^2} + \frac{1}{\psi^4}\Delta^2\right)C,\\
&\qquad C_{(\rho\psi)} = -C_{(\psi)},\\
&C_{(\psi)} = \frac{1}{\psi^3}\Delta^2 C, \qquad C_{(\psi^2)} = \left(-\frac{3}{\psi^4}\Delta^2 + \frac{1}{\psi^6}\Delta^4\right)C,\\
&\qquad C_{(v\psi)} = \left(-\frac{2}{\psi^3}\Delta + \frac{1}{\psi^5}\Delta^3\right)C,
\end{aligned}
\tag{26}
$$

where $\Delta = (\chi - v)$ is a matrix of differences between pixel coordinates in spacetime and v.

During optimization, local maxima can be a problem if one initializes with a prior region that does not cover the location where the receptive field is largest. To avoid local maxima, we initialized σ^2 and v to the noise variance from ridge regression and to the center of mass of the ridge regression estimate, respectively. We used a coarse grid search for initializing Ψ, with off-diagonal terms $\phi = 0$. From the best initial point, we then descended the negative log-evidence using *fmincon* in MATLAB, with analytically computed gradients and Hessians given above. Hyperparameters were constrained to fall within the ranges $10^{-6} \le \sigma^2 \le 10^6$, $-1 \le v_i \le d_i + 1$, $0.1 \le \psi_i \le 2d_i$, $-1 \le \phi \le 1$, $-20 \le \rho \le 20$, where d_i is the number of filter elements along the i'th coordinate dimension.

ALD in spatiotemporal frequency (ALDf). We implemented ALDf using the method similar to that described above for ALDs, after performing an orthogonal fast Fourier transform (FFT) on the stimuli X, which amounts to a change-of-basis (i.e., multiplication by a unitary matrix).

As described in Results, for filters with coordinate dimension D, v is a $D \times 1$ vector. M is a $D \times D$ symmetric matrix, parametrized as

$$
M = \begin{pmatrix} \theta_{11} & \cdots & \theta_{1D} \\ \vdots & \ddots & \vdots \\ \theta_{1D} & \cdots & \theta_{DD}. \end{pmatrix}
\tag{27}
$$

The first and second derivatives of C in $1D$ with respect to these hyperparameters are given by:

$$
\begin{aligned}
&C_{(\rho)} = -C, \qquad C_{(\rho^2)} = C, \qquad C_{(\rho v)} = -C_{(v)},\\
&C_{(v)} = \Delta C, \qquad C_{(v^2)} = -C + \Delta^2 C, \qquad C_{(\rho M)} = -C_{(M)},\\
&C_{(M)} = -\xi\Delta C, \quad C_{(M^2)} = -\xi^2 C + (\xi\Delta)^2 C, \quad C_{(vM)} = \xi C - \xi\Delta^2 C,
\end{aligned}
\tag{28}
$$

where $\Delta = (|M\omega| - v)$ is a matrix of differences between pixel coordinates in frequency and v, and $\xi = \mathrm{sign}(M\omega)$.

Analogously to ALDs, to avoid local maxima, we initialized σ^2 and v to the noise variance from ridge regression and to the centroid of the region of maximal power in the Fourier transform of the ridge-regression estimate, respectively. We used a coarse grid search for initializing M. From the best initial point, we then performed optimization as described above, with boundary conditions for the noise variance and hyperparameters $10^{-6} \le \sigma^2 \le 10^6$, $-1 \le v_i \le \frac{1}{2}d_i + 1$, $10^{-6} \le M_i \le 10^6$, $-20 \le \rho \le 20$, where d_i is the number of filter elements along the i'th coordinate dimension. Once we found the filter estimate in Fourier domain, we projected it back to the spacetime domain via the inverse FFT.

ALD in spacetime and frequency (ALDsf). For the jointly localized prior, we first obtained the maximum-evidence estimates for the ALDs and ALDf covariance matrices C_s and C_f (eqs. 11 and 12). We then performed the optimization of the log-evidence for the full set of ALDsf hyperparameters using *fmincon* in MATLAB, using analytic gradient and Hessian (introduced above), with the boundary conditions for the noise variance and hyperparameters set to the same values as above.

Application to simulated and real neural data

For the simulated data shown in Fig. 5 , we used a 2-dimensional Gabor filter (shown in Fig. 4 A) and two types of stimuli: Gaussian white noise and "naturalistic spectrum" noise–Gaussian noise with a $1/F$ power spectrum. Simulations were carried out with various numbers of stimulus samples $n \in \{800, 1600, 3200, 6400, 12800, 25600\}$, noise variance $\sigma^2 = 2$, signal variance of 1, and a 20×20 pixel filter (coordinate dimension $D = 2$, filter dimension $d = 400$). To quantify performance, we defined the filter error e_f as $e_f = (\sum_{i=1}^{d} (\mathbf{k}_i - \hat{\mathbf{k}}_i)^2)^{\frac{1}{2}}$, where $\mathbf{k}$ is the true filter and $\hat{\mathbf{k}}$ is an estimate. To obtain reliable estimates of mean error, we ran 100 simulations at each sample size. To calculate the relative error (Fig. 5 B and D), we computed the error $e_f(\hat{\mathbf{k}})$ for each method, and then computed the geometric mean of the error ratio $e_f(\hat{\mathbf{k}})/e_f(\hat{\mathbf{k}}_{ALDsf})$ across datasets.

For V1 data shown in Fig. 6 , the data and experimental methods are described in [53]. Briefly, cells were stimulated with 1D spatiotemporal binary white noise stimuli ("flickering bars") aligned with each neuron's preferred orientation. Stimuli were presented at a frame rate of 100 Hz. The number of bars d_x varied for different neurons, $d_x \in \{12, 16, 24\}$. The linear receptive field was assumed to extend over a time window of $d_t = 16$ frames before a spike (a 160 ms time interval). The full dimensionality of the filter was thus $d_t \times d_x$, ranging from 192 to 384 parameters.

For retinal ganglion cell data shown in Fig. 8 , the data and experimental methods are described in [54,55]. Briefly, cells were stimulated with the spatiotemporal binary white noise stimuli presented at a frame rate of 120 Hz, contained in 10×10 pixels in space. We assumed the size of the linear receptive field to be 10×10 pixel $\times 25$ time bin, making for 2500 total coefficients in the RF.

We used cross-validation to quantify the performance of the various estimators (Fig. 6), and resampled the training data to examine performance as a function of training sample size. To quantify error reliably, we performed 100 repetitions for each sample size, drawing the training data randomly without replacement in blocks of size 2s, which helped to minimize the effects of non-stationarities in the data. To quantify cross-validation performance, we used relative cross-validation e_{xv}, defined as $e_{xv} = \frac{1}{n}\sum_{j=1}^{n} (y_{test,j} - X_{test,j}\hat{\mathbf{k}})^2 - \frac{1}{n}\sum_{j=1}^{n} (y_{test,j} - X_{test,j}\hat{\mathbf{k}}_{test})^2$, where n is the number of samples of test data, $y_{test,j}$ is a spike count in the j'th time bin in the test set, $X_{test,j}$ is the jth row of the design matrix X_{test}, $\hat{\mathbf{k}}$ is the RF estimate obtained by each method (from training data), and $\hat{\mathbf{k}}_{test}$ is the ML estimate obtained on the test data. Essentially, this is the ordinary test error minus the error of the ML estimator trained on test data (which provides an absolute lower bound on the performance of any linear model). We computed the relative cross-validation errors from five methods (ML, Ridge, ARD, ASD, and ALDsf) using 8 minutes of test data. In Fig. 6, we normalized the errors by dividing them by maximum average error across methods (the ML estimate using 30 seconds of data yielded

the maximum cross-validation error). We computed the standard deviation of the normalized cross-validation error across 100 different training sets for each dataset size.

Fully Bayesian inference (MCMC)

To perform fully Bayesian inference, we used Metropolis-Hastings (MH) sampling to sample from the distribution over hyperparameters θ given the data X, Y. We used an isotropic Gaussian proposal distribution with variance given by the largest eigenvalue of inverse Hessian of the log-evidence around $\hat{\theta}_{map}$. (More advanced proposal distributions and sampling methods are found in [76,77], but this simple proposal sufficed for our purposes and mixed reasonably quickly). Thus, we first optimized the evidence to obtain the mode $\hat{\theta}_{map}$ of $\log P(\theta|X,Y)$, which is the mode of $\log P(X,Y|\theta)+\log P(\theta)$. We assumed a non-informative hyperprior $P(\theta)$, taken to be uniform over the range of values permitted during constrained optimization of the log-evidence (see above).

To carry out MH sampling, we sampled from the Gaussian proposal distribution centered on the current state θ_t of the Markov chain, $\theta^* \sim \mathcal{N}(\theta_t, \sigma^2 I)$, then computed $\alpha = \frac{p(\theta*)}{p(\theta_t)}$, with the $p(\theta) = P(\theta|X,Y)$. We accepted the proposal randomly with probability $\min(1,\alpha)$, setting $\theta_{t+1} = \theta^*$, and otherwise rejected it, setting $\theta_{t+1} = \theta_t$. Given each sample θ_t, we drew a sample of the receptive field $\mathbf{k}_t \sim P(\mathbf{k}|X,Y,\theta_t)$. These samples were averaged to compute the posterior mean (or Bayes Least Squares estimator). Their quantiles were used to compute credible intervals for each filter coefficient.

In Fig. 10 , we compared fully Bayesian (FB) and empirical Bayes (EB) filter estimates obtained from V1 simple cell data [53]. For each set of training data, we drew 5000 samples using MH to compute the posterior mean and credible intervals. The average acceptance rate of the MH sampler was 0.12. For Fig. 10 A, we computed the average of the EB and FB error from 100 repetitions with independently drawn sets of training data. We computed the average cross-validation error of both estimates of the example cell (in red). For Fig. 10 B, we computed the average posterior variance by averaging the posterior variances in the estimates from the 100 iterations in each cell, which we then averaged across all 16 cells. For Fig. 10 C, we computed the average cross-validation error by averaging the errors from the 100 iterations in each cell, and we averaged these across 16 cells. The same 8 minutes of held out test data was used for cross-validation, for all training iterations.

Supporting Information

Figure S1 Figure S1 shows the comparison of ALD and MID estimates.
(EPS)

Text S1 Supporting information to 1) compare the performance of the ALDsf estimator under the linear Gaussian model to the MID estimator which is equivalent to the maximum likelihood estimator under the linear-nonlinear Poisson cascade model; 2) provide expressions for the quantities for computing the log-evidence.
(PDF)

Acknowledgments

We would like to thank NC Rust and JA Movshon for V1 data, and thank J Shlens, AM Litke, A Sher, and EJ Chichilnisky for retinal data. We are grateful to L Paninski, M Sahani, and EP Simoncelli for helpful discussions.

Author Contributions

Analyzed the data: MP JWP. Wrote the paper: MP JWP. Designed the software used in analysis: MP JWP.

References

1. Lee Y, Schetzen M (1965) Measurement of the wiener kernels of a non-linear system by crosscorrelation (Wiener kernels of nonlinear system based on cross-correlation techniques). Int J Control 2: 237–254.
2. deBoer E, Kuyper P (1968) Triggered correlation. IEEE T Bio-Med Eng 15: 169–179.
3. Marmarelis PZ, Naka K (1972) White-noise analysis of a neuron chain: an application of the Wiener theory. Science 175: 1276–1278.
4. Victor JD, Shapley RM (1979) Receptive field mechansims of cat x and y retinal ganglion cells. J Gen Physiol 74: 275–298.
5. Jones JP, Palmer LA (1987) The two-dimensional spatial structure of simple receptive fields in the cat striate cortex. J Neurophysiol 58: 1187–11211.
6. Ringach D, Shapiro G, Shapley R (1997) A subspace reverse correlation technique for the study of visual neurons. Vision Res 37: 2455–2464.
7. Chichilnisky EJ (2001) A simple white noise analysis of neuronal light responses. Network 12: 199–213.
8. Depireux D, Simon J, Klein D, Shamma S (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85: 1220.
9. Theunissen F, David S, Singh N, Hsu A, Vinje W, et al. (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network 12: 289–316.
10. Ringach D, Hawken M, Shapley R (2002) Receptive field structure of neurons in monkey primary visual cortex revealed by stimulation with natural image sequences. J Vision 2: 12–24.
11. Sahani M, Linden J (2003) Evidence optimization techniques for estimating stimulus-response functions. Adv Neural Inf Process Syst 15: 301–308.
12. 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. J Neurosci 23: 4746–4759.
13. Simoncelli EP, Pillow JW, Paninski L, Schwartz O (2004) The Cognitive Neurosciences MIT Press, 3 edition. pp 327–338.
14. Paninski L (2004) Maximum likelihood estimation of cascade point-process neural encoding models. Network 15: 243–262.
15. David S, Gallant J (2005) Predicting neuronal responses during natural vision. Network 16: 239–260.
16. Pillow JW, Simoncelli EP (2003) Biases in white noise analysis due to non-Poisson spike generation. Neurocomputing 52: 109–115.
17. Wu M, David S, Gallant J (2006) Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci 29: 477–505.
18. Korenberg M, Hunter I (1996) The identification of nonlinear biological systems: Volterra kernel approaches. Ann Biomed Eng 24: 250–268.
19. Bussgang J (1952) Crosscorrelation functions of amplitude-distorted gaussian signals. RLE Technical Reports 216: 14–30.
20. Paninski L (2003) Convergence properties of some spike-triggered analysis techniques. Network 14: 437–464.
21. Victor JD (1987) The dynamics of the cat retinal x cell centre. J Physiol 386: 219–246.
22. Meister M, Pine J, Baylor DA (1994) Multi-neuronal signals from the retina: acquisition and analysis. J Neurosci Methods 51: 95–106.
23. Reid R, Alonso J (1995) Specificity of monosynaptic connections from thalamus to visual cortex. Nature 378: 281–283.
24. Reid R, Shapley R (2002) Space and time maps of cone photoreceptor signals in macaque lateral geniculate nucleus. J Neurosci 22: 6158–6175.
25. DeAngelis G, Ohzawa I, Freeman R (1993) Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. ii. linearity of temporal and spatial summation. J Neurophysiol 69: 1118.
26. Eggermont JJ, Johannesma PIM, Aertsen AMHJ (1983) Reverse-correlation methods in auditory research. Q Rev Biophys 16: 341–414.
27. Klein D, Depireux D, Simon J, Shamma S (2000) Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. J Comput Neurosci 9: 85–111.
28. Sahani M, Linden J (2003) How linear are auditory cortical responses? Adv Neural Inf Process Syst 15: 125–132.
29. James W, Stein C (1960) Estimation with quadratic loss. Proc Fourth Berkeley Symp Math Statist Probab 1: 361–379.
30. Efron B, Morris C (1973) Stein's estimation rule and its competitors–an empirical Bayes approach. J Am Stat Assoc 68: 117–130.
31. Stevenson I, Rebesco J, Hatsopoulos N, Haga Z, Miller L, et al. (2009) Bayesian inference of functional connectivity and network structure from spikes. IEEE T Neural Syst Rehabil Eng 17: 203–213.

32. DeAngelis G, Ohzawa I, Freeman R (1995) Receptive-field dynamics in the central visual pathways. Trends Neurosci 18: 451–458.
33. deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science 280: 1439–1443.
34. Casella G (1985) An introduction to empirical Bayes data analysis. Am Stat. pp 83–87.
35. Kass R, Steffey D (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J Am Stat Assoc 84: 717–726.
36. Tipping M (2001) Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 1: 211–244.
37. Warland D, Reinagel P, Meister M (1997) Decoding visual information from a population of retinal ganglion cells. J Neurophysiol 78: 2336–2350.
38. Pillow JW, Ahmadian Y, Paninski L (2011) Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains. Neural Comput 23: 1–45.
39. Robbins H (1956) An empirical Bayes approach to statistics. Proc Third Berkeley Symp Math Statist Probab 1: 157–163.
40. Morris C (1983) Parametric empirical Bayes inference: theory and applications. J Am Stat Assoc 78: 47–55.
41. Bishop CM (2006) Pattern recognition and machine learning. New York: Springer.
42. Raphan M, Simoncelli EP (2011) Least squares estimation without priors or supervision. Neural Comput 23: 374–420.
43. Faul A, Tipping M (2002) Analysis of sparse Bayesian learning. Adv Neural Inf Process Syst 14: 383–389.
44. Wipf D, Nagarajan S (2008) A new view of automatic relevance determination. Adv Neural Inf Process Syst 22: 1625–1632.
45. MacKay D (1991) Bayesian interpolation. Neural Comput 4: 415–447.
46. Tipping M, Faul AC (2003) Fast marginal likelihood maximisation for sparse Bayesian models. Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics 9: 1–13.
47. Schmolck A, Everson R (2007) Smooth relevance vector machine: a smoothness prior extension of the RVM. Mach learn 68: 107–135.
48. Wipf D, Nagarajan S (2009) Sparse estimation using general likelihoods and non-factorial priors. Adv Neural Inf Process Syst 23: 2071–2079.
49. Wipf D, Nagarajan S (2010) Latent variable Bayesian models for promoting sparsity. Technical report, UCSF.
50. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 58: 267–288.
51. Rasmussen C, Williams C (2006) Gaussian Processes for Machine Learning MIT Press.
52. Ohzawa I, DeAngelis G, Freeman R (1997) The neural coding of stereoscopic depth. NeuroReport 8: 3–12.
53. Rust NC, Schwartz O, Movshon JA, Simoncelli EP (2005) Spatiotemporal elements of macaque V1 receptive fields. Neuron 46: 945–956.
54. Shlens J, Field G, Gauthier J, Grivich M, Petrusca D, et al. (2006) The structure of multi-neuron firing patterns in primate retina. J Neurosci 26: 8254–8266.
55. Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, et al. (2008) Spatio-temporal correlations and visual signaling in a complete neuronal population. Nature 454: 995–999.
56. Robert C, Casella G (2005) Monte Carlo Statistical Methods Springer.
57. Schummers J, Cronin B, Wimmer K, Stimberg M, Martin R, et al. (2007) Dynamics of orientation tuning in cat V1 neurons depend on location within layers and orientation maps. Front Neurosci 1: 145–159.
58. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused Lasso. J Roy Stat Soc B Met 67: 91–108.
59. Wainwright M, Simoncelli E (2000) Scale mixtures of gaussians and the statistics of natural images. Adv Neural Inf Process Syst 12: 855–861.
60. Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103: 681–686.
61. Donoho D, Johnstone I (1995) Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90: 1200–1224.
62. Donoho D, Elad M (2003) Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. Proc Natl Acad Sci U S A 100: 2197–2202.
63. Chen S, Donoho D, Saunders M (1999) Atomic decomposition by basis pursuit. SIAM J Sci Comp 20: 33–61.
64. Wipf D, Rao B (2004) Sparse Bayesian learning for basis selection. IEEE T Signal Proces 52: 2153–2164.
65. Mineault PJ, Barthelme S, Pack CC (2009) Improved classification images with sparse priors in a smooth basis. J Vision 9: 1–24.
66. Sharpee T, Rust NC, Bialek W (2004) Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput 16: 223–250.
67. de Ruyter van Steveninck R, Bialek W (1988) Real-time performance of a movement-senstivive neuron in the blowy visual system: coding and information transmission in short spike sequences. Proc R Soc Lond B 234: 379–414.
68. Bialek W, de Ruyter van Steveninck R (2005) Features and dimensions: Motion estimation in y vision. arXiv: q–bio/0505003.
69. Schwartz O, Pillow JW, Rust NC, Simoncelli EP (2006) Spike-triggered neural characterization. J Vis 6: 484–507.
70. Pillow JW, Simoncelli EP (2006) Dimensionality reduction in neural models: An informationtheoretic generalization of spike-triggered average and covariance analysis. J Vision 6: 414–428.
71. Saleem A, Krapp H, Schultz S (2008) Receptive field characterization by spike-triggered independent component analysis. J Vis 8: 1–16.
72. David S, Vinje W, Gallant J (2004) Natural stimulus statistics alter the receptive field structure of V1 neurons. J Neurosci 24: 6991.
73. David SV, Hayden BY, Gallant JL (2006) Spectral receptive field properties explain shape selectivity in area V4. J Neurophysiol 96: 3492–3505.
74. Friedman J, Hastie T, Tibshirani R (2009) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33: 1–22.
75. Friedman J, Hastie T, Höing H, Tibshirani R (2007) Pathwise coordinate optimization. Ann Appl Stat 1: 302–332.
76. Neal RM (2003) Slice sampling. Ann Stat 31: 705–741.
77. Ahmadian Y, Pillow JW, Paninski L (2011) Efficient Markov chain Monte Carlo methods for decoding neural spike trains. Neural Comput 23: 46–96.

Supporting Information

1 Comparison to MID estimator

One reasonable concern is that our results might reflect limitations of the linear Gaussian encoding model rather than the virtues of the ALD prior. That is, the evidence optimization framework relies on a linear Gaussian model of the neural response (Fig. 1), which fails to take into account neural response nonlinearities or the discrete noise distribution underlying spike counts; perhaps the maximum likelihood estimator under a more realistic encoding model would perform better than any of the estimators considered here, making ALD unnecessary.

To address this possibility concretely, we compared the performance of ALDsf to the MID (maximally informative dimensions) estimator [1], which is equivalent to the maximum likelihood estimator under the linear-nonlinear Poisson cascade model [2, 3]. This estimator takes the neural nonlinearity into account, and models the response noise as Poisson, which is clearly more accurate than Gaussian. We fitted the MID estimate using a spline with 6 knots to parametrize the distributions $P(x)$ and $P(x|spike)$, which are the necessary ingredients for computing the "single-spike information" (or equivalently, log-likelihood).

We fit MID, ALD, and Gaussian ML estimates to the data of the same V1 cell shown in (Fig. 6 left) for 100 different resamplings of the original data, for each size dataset. We used the MID estimate computed on 25 minutes of independent data as our "test" filter estimate, and computed the mean squared error between each "training estimate" and this test filter (If anything, this comparison should favor MID, since the comparison filter was computed using the same model).

Figure S1 shows that MID error rate was comparable to the Gaussian-ML estimate (i.e., linear regression), and that ALDsf achieved significantly lower error. This demonstrates that benefits conferred by the ALD prior are not compromised by the assumption of a linear Gaussian response model. In fact, the MID estimate performed slightly worse than the linear regression estimate, perhaps due to the fact that it effectively has more free parameters (including those governing the nonlinearity) and therefore has even greater need of regularization.

2 Essential quantities

Here we provide expressions for many of the quantities required for computing the log-evidence $\mathcal{E}$, which are useful for numerical optimization of the ALD model parameters. Although these expressions are all available in the published literature [4–6], it is useful to have them compiled in one place, using the same notation.

More importantly, we provide expressions for evaluating the evidence and posterior mean that avoid inverting the prior covariance C or the posterior covariance Λ. This is important for cases where the prior becomes ill-conditioned due to pixels or frequencies are effectively pruned from the model. We use the $\backslash$ ("backslash" operator in matlab) to indicate left-division, which is a faster and more numerically stable way to left-multiply by the inverse of a matrix. (Note that the matrices to which we apply the backslash operator here are always well-conditioned).

Likelihood. from linear-Gaussian encoding model:

$$\begin{aligned} P(Y|X,\mathbf{k},\sigma^2) &= \frac{1}{|2\pi\sigma^2 I|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{k}-m)^\top L^{-1}(\mathbf{k}-m)\right] \\ &= |L|^{\frac{1}{2}}\sigma^{-n}\mathcal{N}(m,L) \end{aligned} \tag{1}$$

where

$$L = \sigma^2(X^TX)^{-1} \quad \text{and} \quad m = \frac{1}{\sigma^2}LX^\top Y = (X^TX)^{-1}X^\top Y. \tag{2}$$

Note $L^{-1} = \frac{X^TX}{\sigma^2}$.

Prior. Zero mean Gaussian with covariance C:

$$P(\mathbf{k}|\theta) = \mathcal{N}(0, C(\theta)). \tag{3}$$

Posterior. Gaussian, proportional to product of likelihood and prior:

$$P(\mathbf{k}|X,Y,\theta,\sigma^2) = \mathcal{N}(\mu,\Lambda) \tag{4}$$

where

$$\begin{aligned} \Lambda &= (L^{-1}+C^{-1})^{-1} && (5)\\ &= (\tfrac{1}{\sigma^2}CX^TX + I_d)^{-1}C && (6)\\ &= \left(\tfrac{1}{\sigma^2}CX^TX + I_d\right)\setminus C, && (7) \end{aligned}$$

and

$$\begin{aligned} \mu &= \Lambda L^{-1}m = \tfrac{1}{\sigma^2}\Lambda X^\top Y && (8)\\ &= (X^TX + \sigma^2C^{-1})^{-1}X^\top Y && (9)\\ &= \left[(CX^TX + \sigma^2 I_d)\setminus C\right]X^\top Y, && (10) \end{aligned}$$

where I_d is a $(d \times d)$ identity matrix and d is the parameter dimensionality of $\mathbf{k}$.

Evidence:

$$\begin{aligned} P(Y|X,\theta,\sigma^2) &= \int P(Y|X,\mathbf{k},\sigma^2)P(\mathbf{k}|\theta)d\mathbf{k} \\ &= \frac{|2\pi\Lambda|^{\frac{1}{2}}}{|2\pi\sigma^2 I_n|^{\frac{1}{2}}|2\pi C|^{\frac{1}{2}}} \exp\left[\frac{1}{2}\left(\mu^\top\Lambda^{-1}\mu - m^\top L^{-1}m\right)\right] \end{aligned} \tag{11}$$

where I_n is a $(n \times n)$ identity matrix and n is the number of data points.

Log-Evidence:

Define $\mathcal{E} = \log P(Y|X,\theta,\sigma^2)$ and we have

$$\mathcal{E} = -\tfrac{n}{2}\log|2\pi\sigma^2| - \tfrac{1}{2}\log|C\Lambda^{-1}| + \tfrac{1}{2}\mu^T\Lambda^{-1}\mu - \tfrac{1}{2\sigma^2}Y^\top Y. \tag{12}$$

Special case: what happens when C goes to all-zeros (or equivalently, all coefficients are pruned)? If $C = 0$, then $C^{-1} = \infty$. So $\Lambda = 0$. In this case, the log-evidence reduces to:

$$\mathcal{E} = -\tfrac{n}{2}\log|2\pi\sigma^2| - \tfrac{1}{2\sigma^2}Y^\top Y. \tag{13}$$

References

1. Sharpee T, Rust NC, Bialek W (2004) Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput 16: 223–250.

2. Kouh M, Sharpee TO (2009) Estimating linear-nonlinear models using Renyi divergences. Network 20: 49–68.

3. Williamson RS, Sahani M, Pillow JW (2011) On information-theoretic and likelihood-based methods for spike-triggered neural characterization. In: Computational and Systems Neuroscience (CoSyNe) Abstracts.

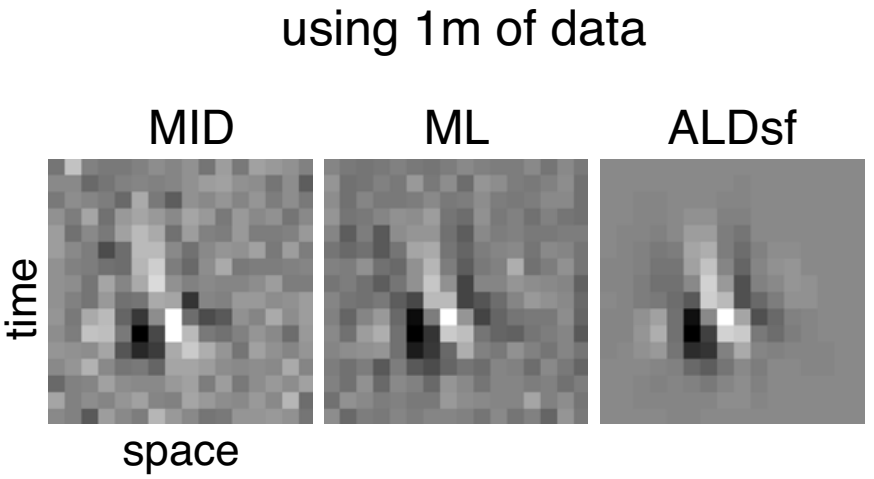
4. MacKay D (1991) Bayesian interpolation. Neural Comput 4: 415–447.

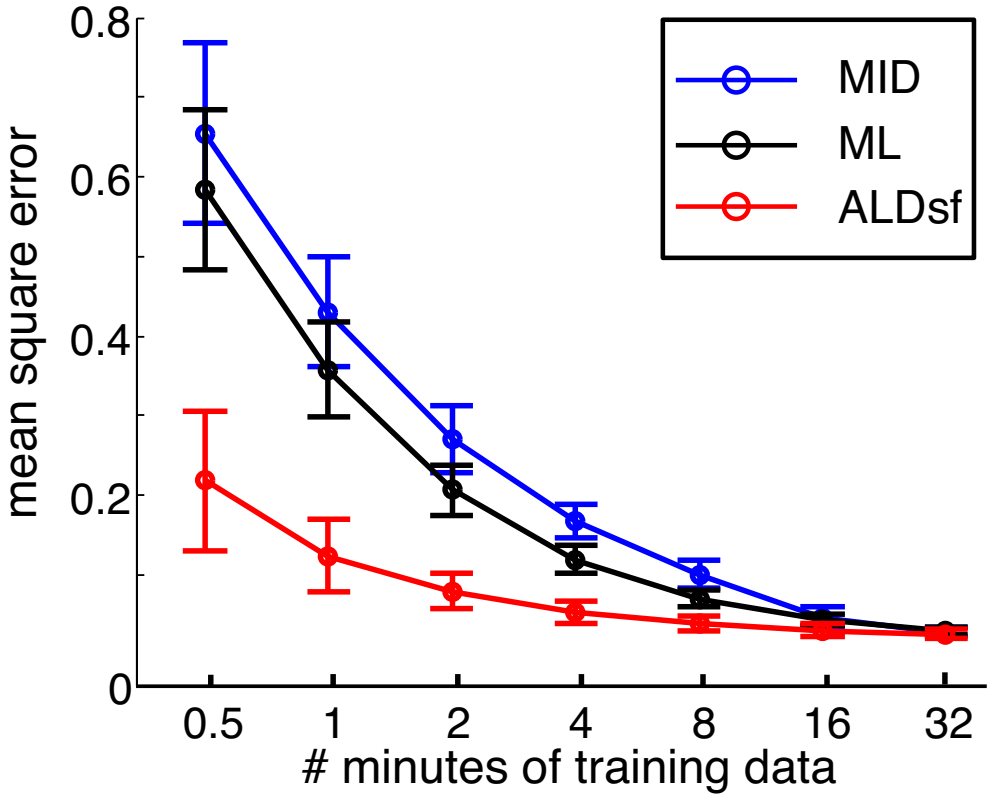
5. Tipping M (2001) Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 1: 211–244.

6. Sahani M, Linden J (2003) Evidence optimization techniques for estimating stimulus-response functions. Advances in Neural Information Processing Systems 15: 301–308.

7. Rust NC, Schwartz O, Movshon JA, Simoncelli EP (2005) Spatiotemporal elements of macaque V1 receptive fields. Neuron 46: 945–956.

Figure S1. Comparison of MID and ALDsf estimates.
Left: RF estimates from MID, ML (linear Gaussian model), and ALDsf, using one minute of data from a V1 simple cell (data from [7]). **Right**: Mean squared error between MID, ML and ALDsf estimates and an MID estimate computed on an independent 25m test set, as a function of the amount of training data. Each point represents an average of 100 training datasets randomly sub-sampled from the original data.


using 1m of data
MID
ML
ALDsf
time
space


0.8
0.6
0.4
0.2
0
mean square error
MID
ML
ALDsf
0.5
1
2
4
8
16
32
minutes of training data