Supplementary Discussion.

Optimal filtering in a nonlinear system. The simple argument leading to Eq. (1) may appear reminiscent of the redundancy reduction principle[40, 41, 45-47]. However we do not assume that the response is linear, impose a particular constraint, or specify the optimality measure. We simply assume that the optimality measure, whatever it may be, is preserved under a change in ensemble. Due to the generality of this argument, we cannot make predictions for the optimal shape of frequency tuning, only for the relative changes in tuning upon a change in the input power spectra. For a linear system, the redundancy reduction arguments predict[40, 41, 47] that neural filters should completely remove second-order correlations present in the input ensembles, i.e. the product L(k)P(k) should be constant across frequencies for sufficiently small frequencies for any ensemble. Although this argument may reasonably describe subcortical visual processing[41, 43, 47], it does not appear to describe visual cortex either in response to natural stimuli or to noise (Fig. 2c,f), where L(k)P(k) depends on k. Therefore nonlinearities of simple cells and/or alternative optimization principles appear essential in describing optimal filter properties in the primary visual cortex.

In the Discussion of the main text, we point out that the adaptation observed here may share some underlying mechanisms with previous observations of cortical pattern-specific adaptation. Indeed, it has been proposed[40, 48, 49] that such pattern-specific adaptation arises from anti-Hebbian or decorrelating mechanisms that would more generally lead to adaptation to the stimulus power spectrum like that observed here. These models of adaptation[40, 48, 49] are closely related to the redundancy reduction arguments just discussed and, more generally, to principles of optimal encoding[37, 41, 45-47]

that have been proposed to govern the design and operation of the nervous system. Despite the specific disagreements just discussed, our results support these general ideas in two respects. First, we have found that adaptation acts to reduce relative responsiveness to patterns that have relatively greater stimulus power, as these theories predict. Second, we have found that neural filters adapt to changes in stimulus ensemble in a manner that increases the information transmitted, relative to the information that would be transmitted if filters did not adapt (as seen by the decreased information when filter and ensemble are swapped, Fig. 3).

The optimality argument (1) for a nonlinear system analyzing Gaussian inputs predicts that the nonlinearity does not change its functional form. This is supported in our data by the fact that the average information values are roughly equal under natural and noise stimulation. The information I can be rewritten in terms of the nonlinear function $f(x)$ of the filter output x and the probability $P(x)$ that the filter output has value x: $I = \int dx\, P(x) f(x) \log_2 f(x)$. One way to preserve this sum is to use the strategy of our optimality argument: to leave $P(x)$ unchanged, which for a Gaussian ensemble is accomplished by changing the filter according to Eq. 1, and to leave the nonlinearity $f(x)$ unchanged. The extent to which this strategy is followed by our two example cells can be seen in Figure 1 and Supplementary Figure 3: the pink curves illustrate $P(x)$, while the blue curves illustrate $f(x)$, which in Figure 1 is also scaled by the firing rate. As can be seen by comparing the curves for the noise MID to that for the natural MID, the curves are at least roughly preserved.

Sensitivity of simple cells to multiple stimulus dimensions. We find that even simple cells in primary visual cortex are sensitive to more than one stimulus dimension, in

agreement with other recent work[26, 29]. A single filter corresponds to a single stimulus dimension; the filter output tells the strength of the stimulus along that dimension. The ratio I/I_{spike} (Fig. 3, bottom) tells the proportion of the information encoded about the stimulus in the neuron's spikes that can be accounted for by the output of the most informative filter[30]. For simple cells, the dominant filter accounts for only about 35% of the overall information. Thus, other stimulus dimensions must significantly influence the neuron's firing[10, 26, 29]. Presumably all of these relevant dimensions also shift with changes in stimulus ensemble, of which we analyzed here only the dominant one. It is also possible that adaptive changes in the structure of each of the relevant dimensions will change their relative importance for eliciting a spike. In particular, the dominant filter for one input ensemble might become secondary in encoding the other input ensemble. The fact that we did not see qualitative changes in the structure of the dominant filter between natural and noise stimulation suggests that such shifts in the relative role of dimensions are not common. Future studies will extend the adaptation analysis to include other relevant dimensions beyond the dominant filter.

Optimal spatial frequency and orientation under natural and noise stimulation.

Filters derived from noise and natural stimuli had similar optimal orientation and spatial frequency. The optimal values were obtained as the position of the maximum of the 2D Fourier transform in space at the temporal frequency of the grating (2Hz). We found a small but statistically significant shift in the optimal spatial frequency, with filters derived from noise inputs having a 21% (±3% s.d.) higher value of the optimal spatial frequency than filters derived from natural inputs ($p<10^{-4}$). This shift in optimal spatial frequency was small enough that neither the noise ensemble estimate nor the natural ensemble estimate was significantly different from direct measurements of the preferred spatial frequencies of these cells with gratings. We note that the measurements with gratings were done separately, before exposure to the noise or natural ensembles, and do

not represent tests of grating spatial frequency sensitivity in the states of adaptation to white noise or natural stimuli. We note also that our conclusions about optimal coding depend on the sensitivity throughout the entire range of spatial frequencies and not on the position of the maximum of the spatial frequency tuning curve (the "optimal" spatial frequency) for a particular cell.

In agreement with previous findings[25, 26], we did not see statistically significant changes in optimal stimulus orientation between grating, natural ensemble, or noise ensemble estimates. Natural stimuli have anisotropic power spectra with increased power at horizontal and vertical orientations[50], and therefore one might have expected some shifts in optimal stimulus orientation away from horizontal or vertical for the natural filter relative to the noise filter. Adaptation to orientation is strongest when the difference between the preferred orientation of the neuron and the adapting orientation is between 20-60 degrees, and acts to shift the preferred orientation away from the adapting stimulus[11]. Thus, shifts due to over-representation of vertical and horizontal orientations would both tend to occur on neurons preferring oblique orientations, and would be in opposite directions. We speculate that the two effects tend to cancel.

Dynamics of Adaptation to Natural Stimuli. Here we argue against certain artifactual explanations of Figure 4a. It could be argued that the increase of information with time seen in Figure 4a may occur because of correlations between the stimuli used for the information calculation (the "test set") and those used in calculating the filters themselves (the "training set"). Natural movies tend to have correlations that diminish in time as a power law rather than an exponential[31, 33, 35, 36], and in that sense are long-lasting. The training set was the last half of the movies, so it might be argued that, as time progresses from the beginning of the movies, the correlation of the test set with the training set would increase and this might explain the increase in information. One argument against

this explanation is that information saturates after the first quarter of stimulus presentations, whereas the correlation with the training set would continue to increase throughout the first half. We tested this explanation more directly by using an alternative training set. We calculated the filters from the middle half of the movies (136 to 410 sec) and then calculated information on the first quarter and the last quarter. Now the first quarter and the last quarter are equally distant in time from the training set, and so if this explanation were correct we would expect them to be mirror images of each other: information would go up during the first quarter and go down by an equal amount during the last quarter. On the contrary, and in support of the adaptation argument, we see the same rise in information during the first quarter as before, even though the first quarter is now much closer in time to the training set, and we see no fall in information during the last quarter, cf. Supplementary Figure 6a. An exponential fit gave a time constant of 55±9 seconds, which agrees with the time constant of 42±9 seconds derived from information during the first half of the data, cf. Figure 4. Also, against the more general argument that the rise or fall in information in Figure 4 might be due to some non-stationarity in the stimulus movies, we show that relevant stimulus components, such as the mean and the standard deviation of the outputs of the neural filters applied to these movies, are stable, cf. Supplementary Figure 6 (b-e).

Supplementary Methods.

Dataset Selection.

The present dataset is obtained from 4 animals and included 133 single units which were clustered using a manual spike sorter. For 85 of the 133 neurons, a reliable non-zero filter was obtained from natural inputs, as judged by visual inspection. We found that this

subjective criterion correlated well with an objective criterion of having a significantly positive information value for the filter applied to its own ensemble (after finite-size corrections[51] are applied). The information was positive for all 85 cells, and exceeded its standard deviation in 81/85 cells. We used the latter criterion to select the dataset of 71 cells with reliable filter estimates to both noise and natural stimuli, of which 40 were classified as simple based on their responses to moving sinusoidal gratings of optimal orientation and spatial frequency. Specifically, simple cells were those with ratio of F1/F0>1, where F1 is the response modulation (Fourier component at the frequency of the stimulus grating) and F0 is the mean response to the optimal grating. Because results of Figure 4 are based only on natural stimuli filters, we have included 5 additional simple cells for which the natural stimulus filter was reliable and noise stimulus filter was not.

Response Reconstruction: Neural Filters and Corresponding Nonlinearities

In the framework of the LN model, the probability of response to a particular input **S** is given by an arbitrary nonlinear function f which only depends on the product of the input signal **S** and the neural filter **L**:

$$f=f(\mathbf{L}*\mathbf{S}). \qquad (2)$$

More generally, reconstruction might require description in terms of a nonlinear function of the outputs of several filters, or curved subspaces instead of a strictly linear projection between signals and filters. However, in this paper we focus on the analysis of properties of the dominant filter **L** of the LN model obtained with noise or natural inputs. We note that the assumption of a single linear filter is more general than the assumption that the cell is linear overall, because the input/output function can be strongly nonlinear and is usually well described by a threshold or threshold-linear function.

In the case of white noise inputs, the linear filter can be found using the reverse correlation method, also known as the spike-triggered average (STA):

$$\hat{e}_{\mathrm{STA}} = \left\langle \mathbf{S} P(\mathrm{spike} \mid \mathbf{S}) \right\rangle - P(\mathrm{spike}) \left\langle \mathbf{S} \right\rangle , \qquad (3)$$

where the expectations are taken over the stimulus ensemble probability distribution P(**S**). In other words, the STA vector is computed by taking the average stimulus weighted by the number of spikes it elicits and subtracting the average stimulus multiplied by the overall number of spikes. The magnitude of the filter is irrelevant, because its change can be accommodated by an appropriate rescaling of the input/output function (2), which converts stimulus components along the relevant filter into spike probability. Therefore, we normalize all of the derived filters to unit length or measure them with respect to the noise level.

If inputs are taken from a Gaussian distribution with correlations (colored noise), then the linear filter can be estimated by computing the STA according to Eq. (3) with a subsequent correction for input correlations. The decorrelated STA (dSTA) is obtained by multiplying the STA with the inverse of the stimulus covariance matrix C_{ij}:

$$\hat{e}_{\mathrm{dSTA}} = C^{-1} \hat{e}_{\mathrm{STA}} \qquad (4)$$

In the case of correlated Gaussian inputs, the dSTA filter Eq. (4) represents the solution of both the purely linear model and the LN model. This is no longer true for natural inputs, which are not Gaussian[30]. Therefore we calculate and treat the dSTA for the natural ensemble as the prediction of the purely linear model. It is known that higher signal-to-noise ratios and smoother filters can be achieved by various forms of regularization of the decorrelation process, including low-pass filtering the STA or imposing a high-frequency cutoff on the covariance matrix[23, 25, 26]. The increase in

predictive power upon such regularization happens for three reasons. First, due to finite data or simply the nature of the stimulus ensemble, the covariance matrix might be singular or nearly so, so that its inversion would result in uncontrollably large eigenvalues for high frequencies where power in the stimulus ensemble is small. We have found that this is not the case for our covariance matrix: calculation of the dSTA according to Eq. (4) without any regularization, in numerical simulations for model linear cells, led to excellent agreement between the dSTA and the filter of the model cell with correlation coefficients >0.99 [30] (and unpublished data). Second, due to finite amounts of data, there is noise in the estimation of the STA. If this noise has a relatively flat spectrum, then at high frequencies where signal in the true STA is low, decorrelation may preferentially amplify noise rather than signal. Again, our results with the linear model with a finite number of spikes (e.g. 1000 spikes) suggest that this is not a problem, although we cannot be certain that the noise problem is not worse for real nonlinear neurons. Third, because the dSTA is a biased estimate of the filter of an LN neuron probed with natural scenes, the estimate might be improved by deviating from the linear model. This can be done by adding a parameter (a low-pass cutoff) and tuning this parameter on a cell-by-cell basis to maximize predictive power of the resulting filter[23, 25]. However, it is not clear to what degree a change in just one parameter could account for all deviations between filters of the fully linear model and those of the LN framework. For all of these reasons, we refrained from regularization in our calculations of the dSTA except in the illustrations of example cells in Figure 1; we otherwise treated the dSTA calculated by Eq. (4) as the prediction of the fully linear model. It should also be noted that the inclusion of an ad-hoc low-pass filter parameter would make it impossible to

reliably estimate the higher-frequency parts of the filter; this, along with the bias of the unregularized dSTA, is why the MID method was necessary for us to assay changes in the spatial frequency tuning across ensembles. In Figure 1, for comparison purposes, we illustrate both regularized and unregularized forms of the dSTA. Regularization was based on selecting a cutoff on the eigenvalues of the covariance matrix *C* below which none of the eigenvalues with the corresponding eigenvectors contributed to the inverse C^{-1} in Eq. (7), making it a pseudo-inverse[23, 25]. For each possible value of the cutoff parameter, the dSTA vector was calculated according to Eq. (7) based on a trial set using 7/8 of the data. The optimal cutoff value was selected as that for which the corresponding dSTA provided maximal information on the remaining 1/8 of the data designated as a test set.

In addition to the above methods, we also derived neural filters using the method of most informative dimensions[30], see next section. For all of the above methods, jackknife analysis of neural filters was performed: 8 filters were computed, each with 1/8 of the data left out. When information was computed for a filter on its own ensemble, it was calculated only on this 1/8 of the data that was not used for computing the filter, except in Figure 4 where a single filter was calculated from ½ of the data and information was calculated on segments of the other half. In all other cases, information values reported are an average over the 8 values found with the 8 jackknife estimates. To establish statistical significance of the difference between filters derived with any two different methods and/or two stimulus ensembles, all 16 of the corresponding jackknife estimates (8 for each combination of method and ensemble) were projected on the direction of the difference between the mean filters describing the two groups, and an

unpaired Students t-test was used on these projections. To calculate the signal-to-noise level of receptive fields shown in Figure 1, we compute the average standard deviation across all components of the receptive field across all the jackknife estimates (normalized to unit length) and display receptive field values relative to that noise level.

Once the filter **L** has been obtained as either the STA (3), dSTA (4), or the MID[30], we can calculate the nonlinear input/output function (2) directly from the data. According to its definition it is given by the normalized spike probability given the stimulus **S**:

$$f(\mathbf{S}*\mathbf{L}) = \frac{P(\text{spike} \mid \mathbf{S})}{P(\text{spike})}.$$

When working in the framework of the linear-nonlinear model we assume that the spike probability only depends on stimulus components along the filter **L** of interest: P(spike| **S**)=P(spike| **S***L**). Therefore the nonlinear input/output function can also be written as:

$$f(\mathbf{S}*\mathbf{L}) = \frac{P(\text{spike} \mid \mathbf{S}*\mathbf{L})}{P(\text{spike})}.$$

The last expression can be transformed using Bayes' rule:

$$f(\mathbf{S}*\mathbf{L}) = \frac{P(\mathbf{S}*\mathbf{L} \mid \text{spike})}{P(\mathbf{S}*\mathbf{L})}. \qquad (5)$$

That is, the nonlinear input/output function f is evaluated as a ratio of probability distributions of stimulus components along the filter **L,** P(**S***L), and of the probability distribution of stimulus components P(**S***L|spike) conditional on a spike. Both of the probability distributions are readily available from the experimental data.

Reconstruction of Receptive Fields as Most Informative Dimensions. The justification for the method of most informative dimensions as a way to calculate neural receptive fields is described elsewhere[30], where performance of the method is illustrated on model

visual and auditory neurons. For the convenience of the reader we describe here the methodology of maximizing information to find the receptive fields. It was shown that the information between the output of a particular vector **L** in the input space and the neuron's response, regarded as a spike or no spike in each time bin, can be computed, to lowest order in the probability P(spike) of a spike in the time bin, as the Kullback-Leibler distance between the probability distributions P(x) and P(x|spike):

$$I(\mathbf{L}) = \int dx P_{\mathbf{L}}(x|\text{spike}) \log_2 \left[\frac{P_{\mathbf{L}}(x|\text{spike})}{P_{\mathbf{L}}(x)} \right], \qquad (6)$$

where $P_{\mathbf{L}}(x)$ is the probability distribution of stimulus projections x onto the vector **L** in the input ensemble, and $P_{\mathbf{L}}(x|\text{spike})$ is the probability distribution of stimulus projections x onto the vector **L** among inputs that led to a spike. We compute these two probability distributions as histograms in 21 bins covering the range of projection values (the same number of bins was used in finding MIDs from neural responses to noise and natural ensemble). For each trial vector, we also compute the gradient of information as:

$$\nabla_{\mathbf{L}} I = \int dx P_{\mathbf{L}}(x) \left[\langle \mathbf{S}|x,\text{spike} \rangle - \langle \mathbf{S} \mid x \rangle \right] \frac{d}{dx} \left[\frac{P_{\mathbf{L}}(x|x,\text{spike})}{P_{\mathbf{L}}(x)} \right], \qquad (7)$$

where <**S**|x> is the average of the stimuli having projection value of x onto the vector **L** (using the same binning of x as for the probability distributions $P_{\mathbf{L}}(x)$ and $P_{\mathbf{L}}(x|\text{spike})$). Similarly, <**S**|$x$,spike>is the average of the stimuli that led to a spike that had projection value of x onto the vector **L**. We evaluate the derivative at a particular value of x using Savitsky-Golay coefficients (W.H. Press et al., *Numerical Recipes*, Cambridge University Press 1998) based on two adjacent bins on either side of the bin with the value x; if projections values from any one of these bins were not encountered in the stimulus ensemble, the corresponding average did not contribute to the derivative. We find that the

use of Savitsky-Golay smoothing coefficients is not required, but helps improve convergence of the algorithm [note that in the search algorithm, described below, the trial vectors are accepted based on information values, which are evaluated without smoothing.] This analysis requires that stimuli and spike trains are binned at the same time resolution (33 ms for natural stimuli and 16 ms for noise stimuli). Therefore occasional stimuli correspond to multiple spikes in a bin. If that happened, projections values of such stimuli were counted as many times as there were spikes for all the probability distributions and averages in Eqs. (6) and (7).

The search for the most informative dimension (MID) is initialized by setting the starting vector equal to the STA. To generate a new trial vector, we perform a line maximization (W.H. Press et al., *Numerical Recipes*, Cambridge University Press 1998) along the line defined by the gradient (7), and choose, on average, the one with the largest information. Because information (6) as a function of components of the vector **L** has local maxima, smaller information values are accepted with Boltzmann probability, $\exp(-\Delta I/T)$, where ΔI is the decrease in information between the new and old trial vector measured in units of the information I_{spike} carried by the arrival of a single spike, and the parameter T is called the effective temperature of the simulated annealing cooling scheme. Information values in these units are typically less than one (unless there is overfitting). Therefore, we start the simulated annealing scheme with T=1, and decrease it by a factor of 0.95 after each line maximization. If the search appears to have converged with a fraction precision of 5×10^{-5} and the effective temperature $T\leq10^{-5}$, then the effective temperature is increased by a factor of 5, but not to exceed the starting temperature value. This results in repeated "cooling" and "remelting", and is equivalent

to restarting the algorithm multiple times. We limit the total number of line maximizations to 3000. The best vector found in terms of information during the overall maximization procedure is taken as the most informative dimension **L**. Cross-validation is performed by leaving out 1/8 of data and treating that 1/8 as a test set. We compute information on the test set after every 100 line maximizations, and if the information value has dropped on the test set by 25% of its maximum value, the optimization procedure is stopped and the current filter taken as the MID. Such early stopping seldom occurs when we compute receptive fields from responses to natural scenes, but is common when receptive fields are computed from noise ensembles. This is due to the fact that the starting point, the STA, is very close to the optimal value when neural responses to the noise ensemble are analyzed.

Because the MID method is based on a search in a high-dimensional space for an information maximum, there is of course a concern that our search might become stuck in a local maximum. We believe this is not a concern for the following reasons. First, as just noted, our search procedure is equivalent to restarting the search algorithm multiple times from multiple starting points, only the first of which is the STA, and we take the maximum of information over the entire search. Second, in studies of model cells[30] (and unpublished data), we have found that the error (measured as 1 minus the projection between the true model filter and the MID found by the search) decreases as 1/N where N is the number of spikes used to estimate the filter. This is the dependence predicted theoretically[30], and would not be expected to hold if the true maximum were not being found. Third, we have previously verified on model cells that beginning with a random starting point rather than the STA does not produce better solutions. The STA represents

a natural choice of a starting point in that it is clearly a stimulus direction that carries nonzero information about the neuron's response.

The MID method produces an unbiased estimate.

In this section we provide a detailed derivation for the fact, first published in Ref.[30], that the MID method produces unbiased estimates of neural filters within a single-filter LN model. We will first consider the case of infinite data, and then go through details of the argument with finite data.

While the MID filter can be calculated with respect to any particular pattern of spikes[30], in this paper we have concentrated on finding filters associated with single spikes. Therefore we will do so in this section as well. Information carried by individual spikes about the incoming stimuli is given by[37]:

$$I_{\text{spike}} = \int d^D\mathbf{S} P(\mathbf{S}) \frac{P(\text{spike}\,|\,\mathbf{S})}{P(\text{spike})} \log_2 \frac{P(\text{spike}\,|\,\mathbf{S})}{P(\text{spike})} \qquad (8)$$

Because this is the information between single spikes and full, unfiltered, stimuli, information between spikes and stimuli filtered along any dimension may not exceed (8). To verify that the only filter that leads to an equal amount of information between spikes and stimuli filtered with it is the neural receptive field **L**, we invoke the main assumption of the single-filter LN model: P(spike|**S**)=P(spike|**S*L**), so that:

$$I_{\text{spike}} = \int d^D\mathbf{S} P(\mathbf{S}) \frac{P(\text{spike}\,|\,\mathbf{S}*\mathbf{L})}{P(\text{spike})} \log_2 \frac{P(\text{spike}\,|\,\mathbf{S}*\mathbf{L})}{P(\text{spike})}$$

The integration $d^D\mathbf{S}$ with along all stimulus dimensions can be carried out separately along the relevant stimulus dimension, **S*L,** and along the rest of stimulus dimensions, which we denote as $\mathbf{S}^{\perp}$:

$$I_{\text{spike}} = \int d(\mathbf{S}*\mathbf{L}) \frac{P(\text{spike}\,|\,\mathbf{S}*\mathbf{L})}{P(\text{spike})} \log_2 \frac{P(\text{spike}\,|\,\mathbf{S}*\mathbf{L})}{P(\text{spike})} \int d^{D-1}\mathbf{S}_\perp P(\mathbf{S}_\perp, \mathbf{S}*\mathbf{L})$$

Integration with respect to all of the irrelevant stimulus dimensions $\mathbf{S}_\perp$ results in:

$$I_{\text{spike}} = \int d(\mathbf{S}*\mathbf{L}) \frac{P(\text{spike}\,|\,\mathbf{S}*\mathbf{L})}{P(\text{spike})} \log_2 \frac{P(\text{spike}\,|\,\mathbf{S}*\mathbf{L})}{P(\text{spike})} P(\mathbf{S}*\mathbf{L})$$

which is precisely the information along the filter **L**, cf. Eq. (6). We have thus shown that information along the filter that represents the neural receptive field achieves the maximal information possible, I_{spike} and describes the encoding $\mathbf{S} \rightarrow \mathbf{S}*\mathbf{L} \rightarrow$ spikes. Filtering along any other dimension **V** will correspond to encoding $\mathbf{S} \rightarrow \mathbf{S}*\mathbf{V} \rightarrow \mathbf{S}*\mathbf{L} \rightarrow$ spikes or $\mathbf{S} \rightarrow \mathbf{S}*\mathbf{L} \rightarrow \mathbf{S}*\mathbf{V} \rightarrow$ spikes and, by the data processing inequality (Cover and Thomas, John Wiley Inc. 1991), leads to a lower information processing value. The data processing inequality applies to stochastic inputs but presumes that we know exact probabilities such as P(**S*L**| spike) and P(**S*V**| spike). This shows that the MID method is unbiased in the limit of infinite data and stochastic neurons.

With finite data, we have only a limited number of samples to measure the probability distributions P(**S*L**| spike) and P(**S**| spike). With N spikes, our empirical estimates of these probability distributions P_N(**S*L**| spike) and P_N(**S**| spike) will differ from experiment to experiment in such a way that the average across trials produces the true distribution and the variance across trials acquires a term of the order of 1/N:

$$\left\langle P_N(\mathbf{S}\,|\,\text{spike}) \right\rangle = P(\mathbf{S}\,|\,\text{spike}) \qquad (9)$$

$$\left\langle P_N^2(\mathbf{S}\,|\,\text{spike}) \right\rangle = P^2(\mathbf{S}\,|\,\text{spike}) + \frac{1}{N} P(\mathbf{S}\,|\,\text{spike})\big(1 - P(\mathbf{S}\,|\,\text{spike})\big), \qquad (10)$$

where we have used the properties of the binomial distribution; each particular stimulus **S** can occur with a spike anywhere between 0 and N times, if N is the total number of spikes. Similar relations can be used with other probability distributions involved.

The deviation between the true filter and the MID filter obtained with a particular data set, $\delta\mathbf{V}$, is proportional to the gradient of information (evaluated with finite data) at the position of the true filter: $\delta\mathbf{V} \sim \nabla I(\mathbf{L})$. Here we show that, as was stated in Ref.[30], the gradient of information is zero, after averaging across trials, for the true filter. To verify this we represent information $I_N(\mathbf{L})=I(\mathbf{L})+\delta I_N(\mathbf{L})$, as the information obtained with infinite data and the deviation from it due to finite sampling. The gradient of the information is zero at the true filter **L.** The deviation

$$\delta I_N(\mathbf{L}) = \int dx \delta P_N(x|\text{spike}) \log_2\left[\frac{P(x|\text{spike})}{P(x)}\right] + \int dx \delta P_N(x|\text{spike}),$$

where $x=S*L$, $\delta P_N(x|\text{spike})=P_N(x|\text{spike})-P(x|\text{spike})$ is the difference between the empirical and true distributions, and there is no need to consider noise in the stimulus distribution P(x) because it might be taken as the one actually used in the experiment. Next we take into account that the empirical distribution obeys a normalization constraint, such that $\int dx P_N(x|\text{spike}) = 1$, and therefore $\int dx\, \delta P_N(x|\text{spike}) = 0$, so that:

$$\delta I_N(\mathbf{L}) = \int dx \delta P_N(x|\text{spike}) \log_2\left[\frac{P(x|\text{spike})}{P(x)}\right], \qquad (11)$$

But the average of the empirical distributions is the true distribution (9), so $\delta I_N(\mathbf{L}) = 0$ in the first-order approximation in the deviations between empirical and true distributions. The second-order approximation results, using the property Eq. (10), in a

uniform correction: $\delta I_N(\mathbf{L}) \sim \frac{N_{\text{bins}} - 1}{N_{\text{spike}}}$, where N_{spike} is the number of spikes and N_{bins} is the number of bins used in estimating the probability distribution P(x|spike). Because this correction is independent of the direction in the stimulus space, it provides a zero contribution to the gradient at the position of the true filter. The second-order terms determine the variance of the MID filters on a trial-by-trial basis, because while the deviations themselves $\delta\mathbf{V} \sim \nabla I(\mathbf{L})$ are proportional to the gradient of information, their variance $\left\langle \delta\mathbf{V}_i \delta\mathbf{V}_j \right\rangle \sim \left\langle \nabla_i I(\mathbf{L}) \nabla_j I(\mathbf{L}) \right\rangle$ is proportional to pairwise gradient correlations. Using Eqs. (11) and (10), one can show that the leading term determining this variance behaves as $\sim 1/N_{spike}$. The exact coefficient can be found in Ref. [30]. This means that while different MID filters obtained based on different empirical distributions deviate from each other and from the true filter, these deviations have zero mean and finite variance that decreases as $\sim 1/N_{spike}$ with increasing number of spikes. While there may be terms $\sim N_{spike}^{-2}$ describing a shift in the mean, these will be masked by a much larger effect of variance between estimates decreasing as $\sim N_{spike}^{-1}$. This is what we mean by saying that the MID method is unbiased. Note that the gradient of information evaluated at the filters of the linear model (STA or decorrelated STA) will be non-zero, with terms of order O(1), which do not depend on the number of spikes and remain finite even in the limit of infinite data.

However there are ways in which the stimulus ensemble can influence the single MID even in a neuron that does not adapt, if the relevant subspace (RS) has two or more dimensions. In this case, as shown in Ref. [30], Appendix B, the single MID for that ensemble may include a component outside of the RS if the ensemble is such that the

average stimulus given the projections along the relevant dimensions is not a linear function of each projection (as can occur for non-Gaussian ensembles). Any such effects, however, would be instantaneous and would not yield a time-dependence to the calculation of information as in Fig. 4.

Details of stimulus presentation and filter analysis. The visual input signals were presented as two-dimensional spatiotemporal patterns of light intensities on a video monitor with a refresh rate of 120 Hz. The frame update rate was 60 Hz in the case of the white noise stimulus ensemble and 30 Hz in the case of the natural stimulus ensemble (our commercial cameras did not provide higher temporal resolution than that of television, which is 30Hz). No corrections were made for the camera nonlinear amplitude to intensity transformation function.

The optimal orientation was determined from responses to a set of evenly spaced orientations at 10° intervals, with a spatial frequency of 0.5 cycles/degree and a temporal frequency of 2 Hz. The optimal spatial frequency was derived from responses to a set of moving gratings of optimal orientation and variable spatial frequencies (approximately logarithmically spaced between 0.1 and 4 cycles/degree).

Spatial frequency profiles were obtained by taking the Fourier transform in time and, with zero-padding to 32x32, in space. Linear interpolation between pixels of the 2D transform was used to derive one-dimensional profiles along the preferred orientation of each cell. Before averaging across cells, the spatial frequency profiles of individual cells were normalized to unit length across all spatial and temporal frequencies. Identical procedures were used for receptive fields and stimuli comprising the input ensembles (averaging over all three frame subsequences, e.g. 1-2-3, 2-3-4, etc.).

In Figure 3, the information I was calculated from jackknife estimates of the filters. For each cell, for either the natural or noise ensemble, eight jackknife estimates were derived, each from 7/8 of the data with the remaining 1/8 of the data serving as a test set on which the information was calculated. The mean of these 8 estimates was assigned as information I that cell and ensemble. I_{spike} is calculated from responses to 50-150 repetitions of an 11s-long segment of the natural or noise ensemble. Finite-size corrections[37] were applied to both I and I_{spike}. As a control for the information calculation, we calculated natural MID filters for a series of model simple cells with a static filter where the number of spikes emitted over the course of the test set varied from 80-13,000. The calculated information, of course, decreased substantially at low numbers of spikes, but it did so similarly whether the filter was applied to the natural or the noise ensemble. There was no significant difference between the information about the natural ensemble and about the noise ensemble for any choice of nonlinearity, that is, for any signal-to-noise ratio.

Additional References.

45. Barlow, H. B. in Sensory Communication (ed. Rosenblith, W. A.) 217-234 (MIT Press, Cambridge, MA, 1961).
46. Barlow, H. Redundancy reduction revisited. Network 12, 241-53 (2001).
47. Atick, J. J. & Redlich, A. N. Towards a theory of early visual processing. Neural Comput. 2, 308-320 (1990).
48. Barlow, H. in Vision: coding and efficiency (ed. Blakemore, C.) (Cambridge University Press, Cambridge, UK, 1990).
49. Barlow, H. & Foldiak, P. in The computing neuron (eds. Durbin, R., Miall, C. & Mitchinson, G.) (Addison-Wesley, New York., 1989).
50. Coppola, D. M., Purves, H. R., McCoy, A. N. & Purves, D. The distribution of oriented contours in the real world. Proc Natl Acad Sci U S A 95, 4002-6 (1998).
51. Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. & Bialek, W. Entropy and information in neural spike trains. Phys. Rev. Let. 80, 197-200 (1998).

Supplementary Figure 1

This figure shows the spatial frequency profiles of receptive fields from the two example cells of Figure 1. Spatial frequency sensitivity at zero temporal frequency (a,c) and at 10 Hz (b,d). Red indicates filter derived from responses to noise ensemble, blue indicates filter derived from responses to natural ensemble. The second of the two example cells is typical in all respects. The first of the two cells is atypical in that it did not change its sensitivity at low spatial frequency between natural and noise stimulation at 0 Hz, but exhibited an appropriate change in its tuning at 10 Hz, see Supplementary Figure 2.

Supplementary Figure 2

Spatial frequency sensitivity on a cell-by-cell basis for the first 9 spatial frequencies from Figure 2 (here called k1 to k8 from lowest to highest) for temporal frequencies of 0 and 10 Hz respectively. P-values on top of each graph show significance in sensitivity differences of filters derived from noise vs. natural stimulation. Color for each cell codes sensitivity to noise filter at lowest frequency (k1) and is retained in the plots of higher frequencies. The two example cells of Supplementary Figure 1 are marked as a '+' and an 'X' respectively. Note that the cell marked by a '+' is atypical in its behavior at 0Hz.

Supplementary Figure 3.

Panels (a,b) show that the nonlinear input/output function f(x)=P(spike|x)/P(spike) associated with the MID filters for two exemplary cells of Figure 1 overlap under natural (solid) and noise (dashed) stimulation when stimulus projection *x* along the corresponding receptive fields is measured in units of its standard deviation (x-axis). For comparison, in Figure 1 we plot the input/output function f(x) scaled by the firing rate, P(spike|x) – the probability of a spike in 33ms window given a stimulus projection value *x* along the receptive field. Therefore the difference in scale for the nonlinearities observed between natural and noise conditions in Figure 1, as for example cell 856 2, reflects only a change in the mean firing rate under the two conditions. Panels (c,d) show the probability distributions of projections *x* for natural (solid) and noise (dashed) stimulation.

Supplementary Figure 4.

The increase in information on a cell-by-cell basis when the noise filter is applied to the noise vs. natural ensemble (a) or when the natural filter is applied to the natural vs. noise ensemble (b). Panels (c,d) show this effect in units of I_{spike}. Notations are as in Figure 3.

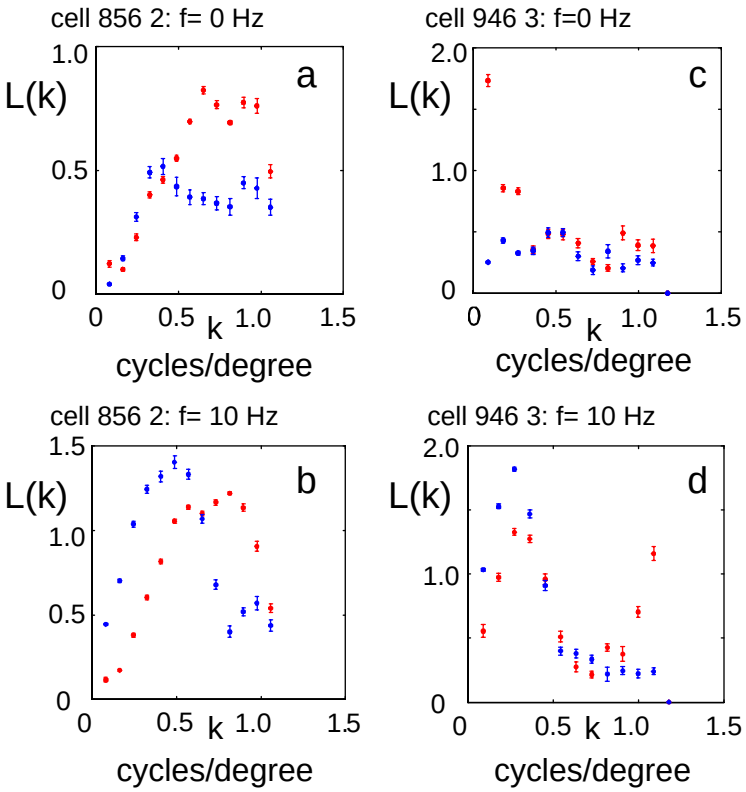
Supplementary Figure 5. Coarse evolution of adaptive neural filters. (a,d) Comparison of neural filters derived from the first half (a,d), middle half (b,e) or last half (c,f) of stimulation with noise and natural inputs. Notations are as in Figure 2(a,d). In panels (g,h) we plot only natural filters to show that they overlap. In panels (i,j) we compare three of the noise filters derived from the first half of the data (magenta), middle half of the data (yellow), and last of the data (red) to the natural filters of the last half of the data. With time, noise neural filters diverge from natural filters.

Supplementary Figure 6. (a) The neural filter derived from the middle half of natural stimulation is applied to the first and last quarter of the natural input ensemble. Notations are as in Figure 4. The solid line is an exponential fit, dashed lines show one standard deviation based on the Jacobian of the fit, p=0.007. The remaining panels show that the relevant statistical properties of the input ensemble are stable and cannot account for the time dependence seen in Figure 4. Here we show the mean and standard deviations (in arbitrary units) for natural and noise input stimuli filtered differently: (b) natural stimuli (first half of the data) filtered with natural neural filters computed from second half of the data; (c) natural stimuli (all duration) filtered with noise neural filters; (d) noise stimuli (all duration) filtered with natural neural filters; (e) noise stimuli (first half of the data) filtered with noise neural filters obtained from the second half.

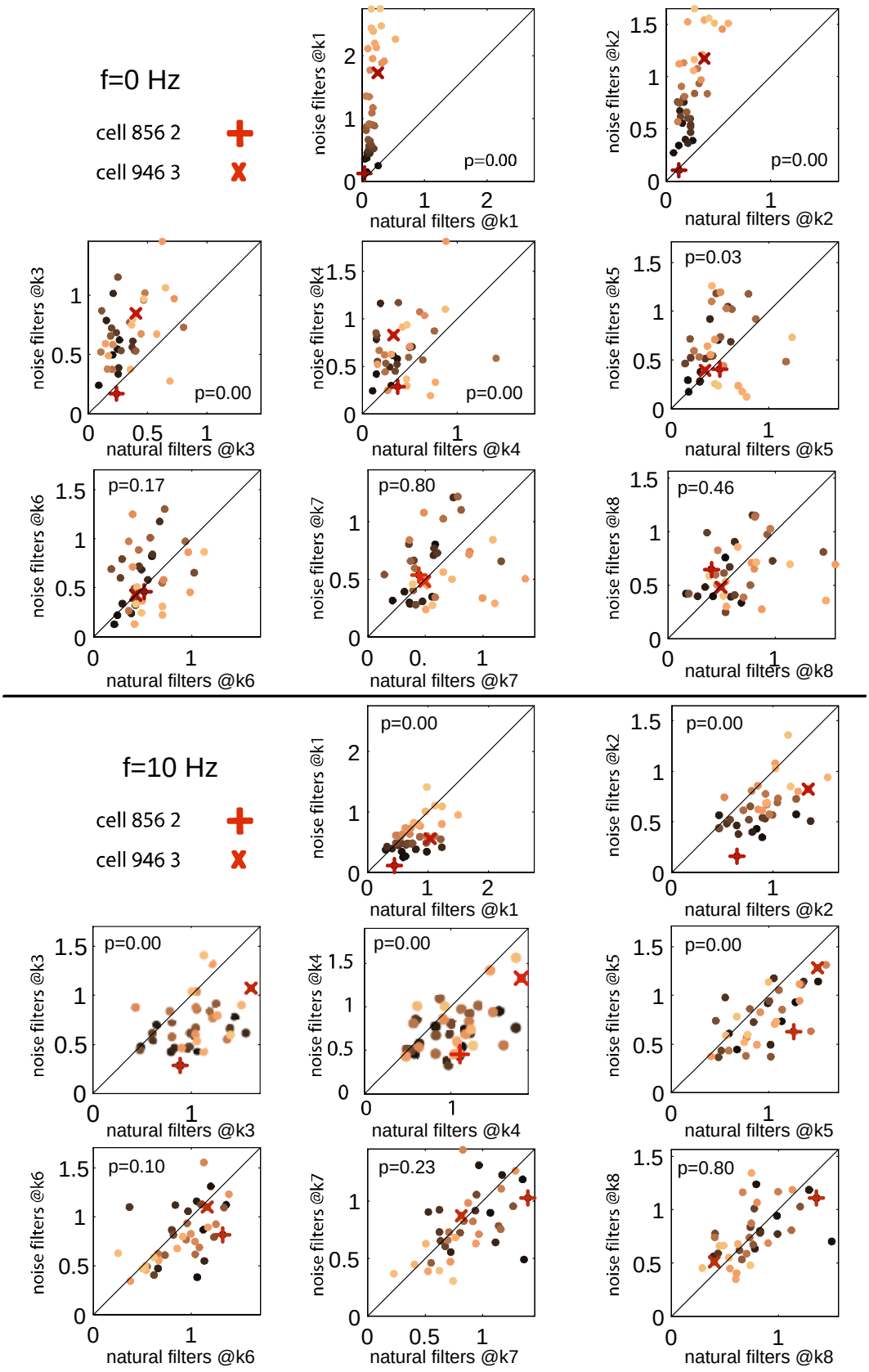
Supplementary Figure 7

Information carried by the noise filter about the neuron's response, as a function of time after exposure to the noise ensemble (a) or natural stimulus ensemble (b). Information values were evaluated along the noise filter derived from the second half (a) and from full recording (b) of noise stimulation. No significant time dependence could be established. Notations are as in Figure 4. Left and right blue bars show average information carried by noise filter about responses to noise ensemble (taller bar) or natural ensemble (shorter bar). Note that the average information values computed for the short time segments for the noise filter applied to the noise ensemble (a) are all smaller than the average information computed over the whole noise ensemble (right bar in a). This suggests that these short-time estimates are too noisy to be reliable in the case of the noise filter, which

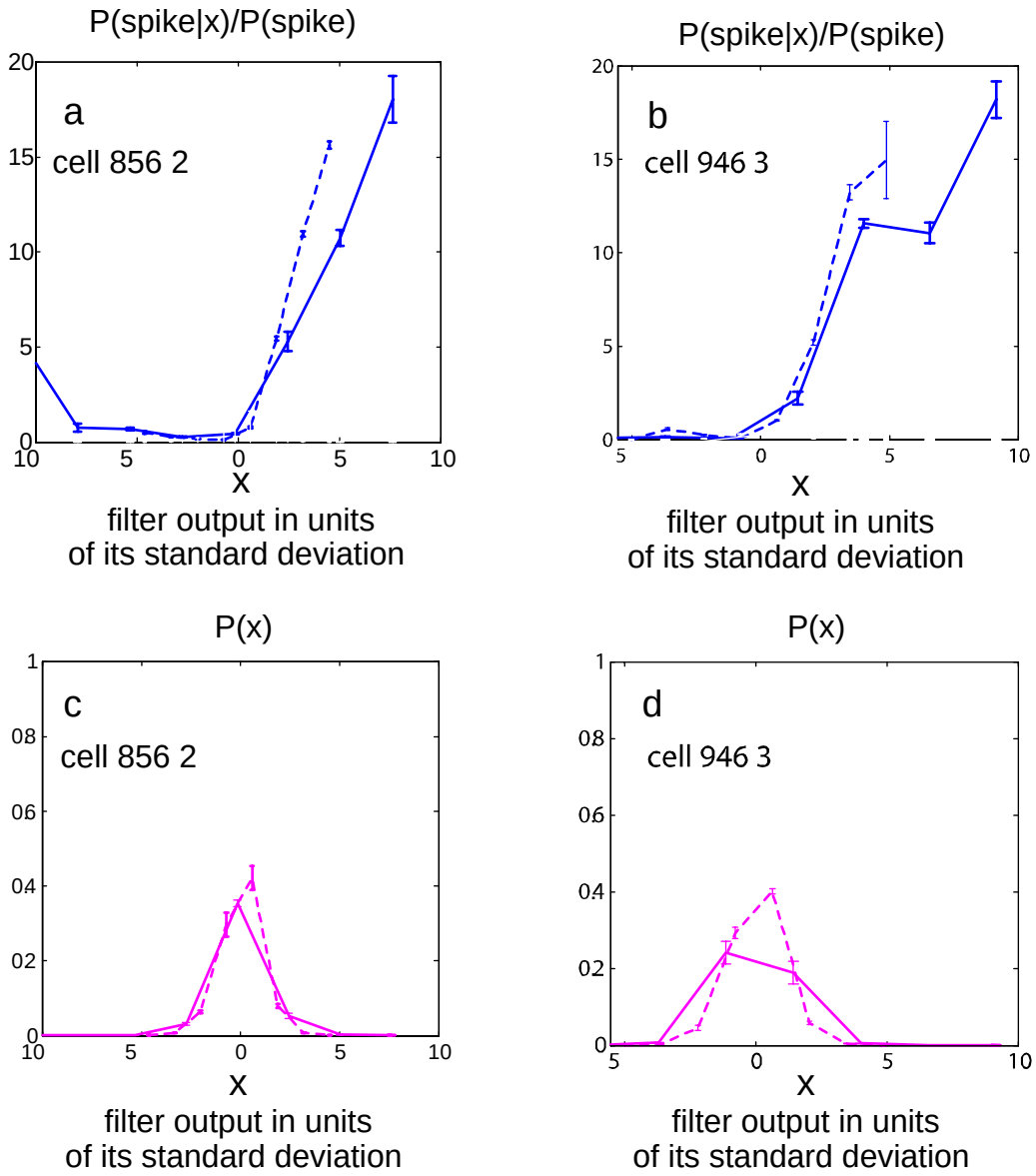
may provide another reason that we could observe no trend for the noise filter. A similar problem can be seen in (b). Note that a similar problem did not arise for the natural filter (main text, figure 4): short-time estimates were equal in size to the estimate over the whole ensemble after adaptation. We used the filter from the full recording in (b) (unlike in main text, figure 4, where the same filter was used in (a) and (b) for consistency) because the short-time estimates for the filter from the second half of the recording showed an even stronger tendency to have low information values; using the full recording helps fight noise and so improves the situation, but not sufficiently.



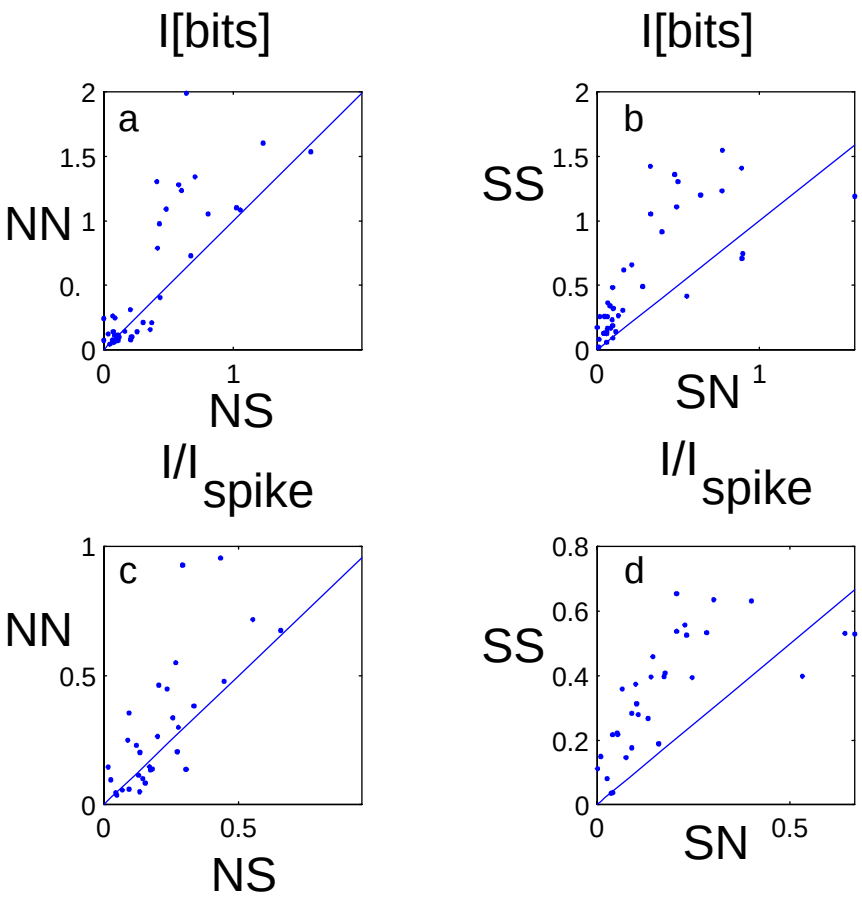
supplementary figure 1



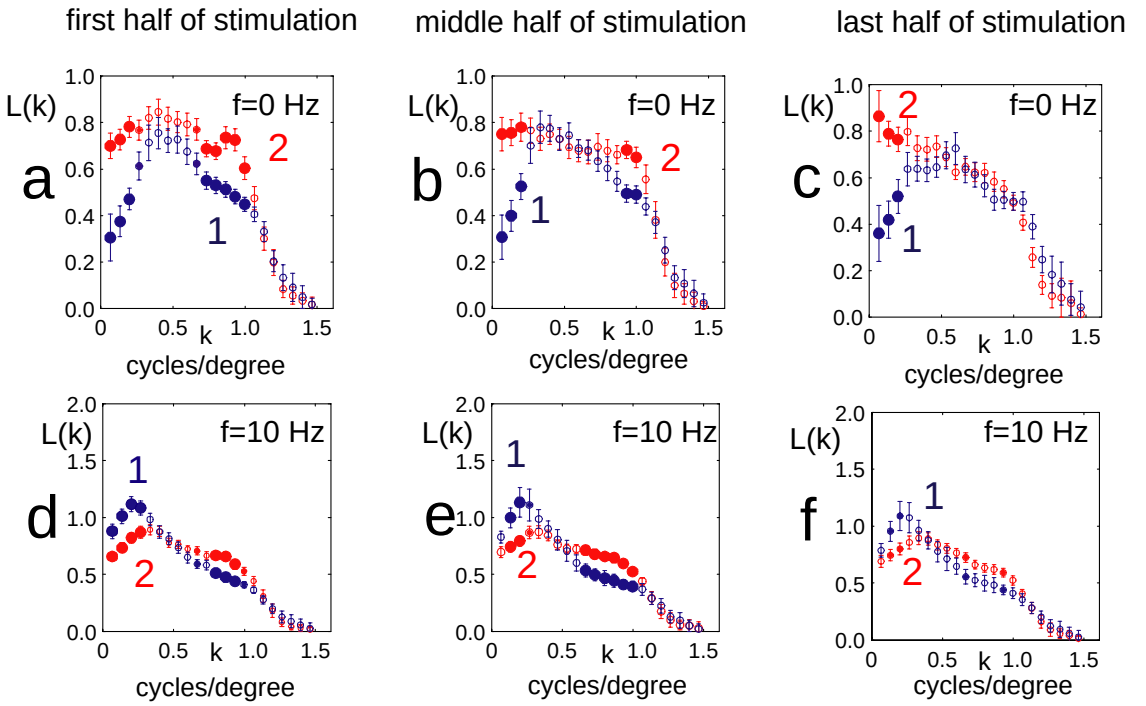
supplementary figure 2

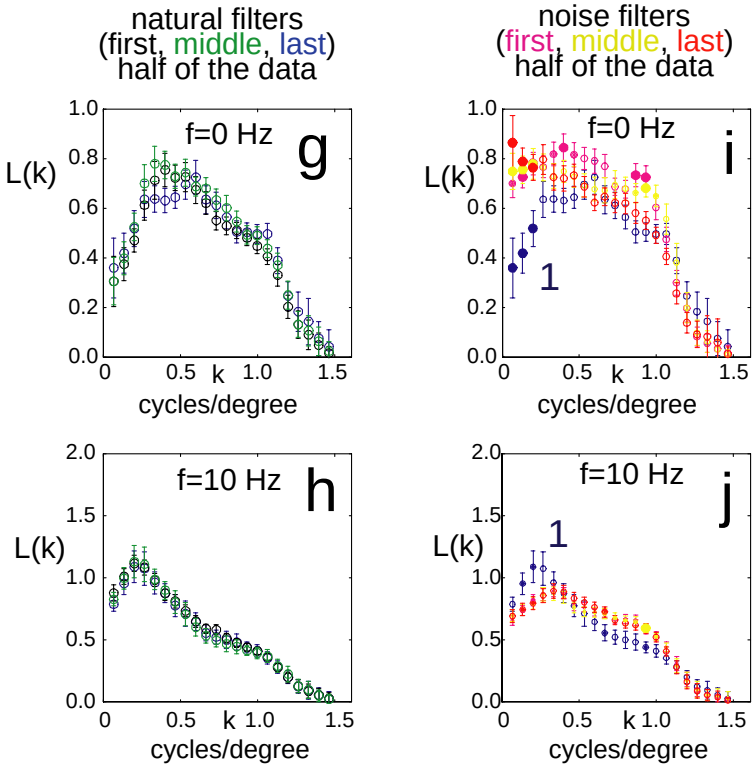


supplementary figure 3

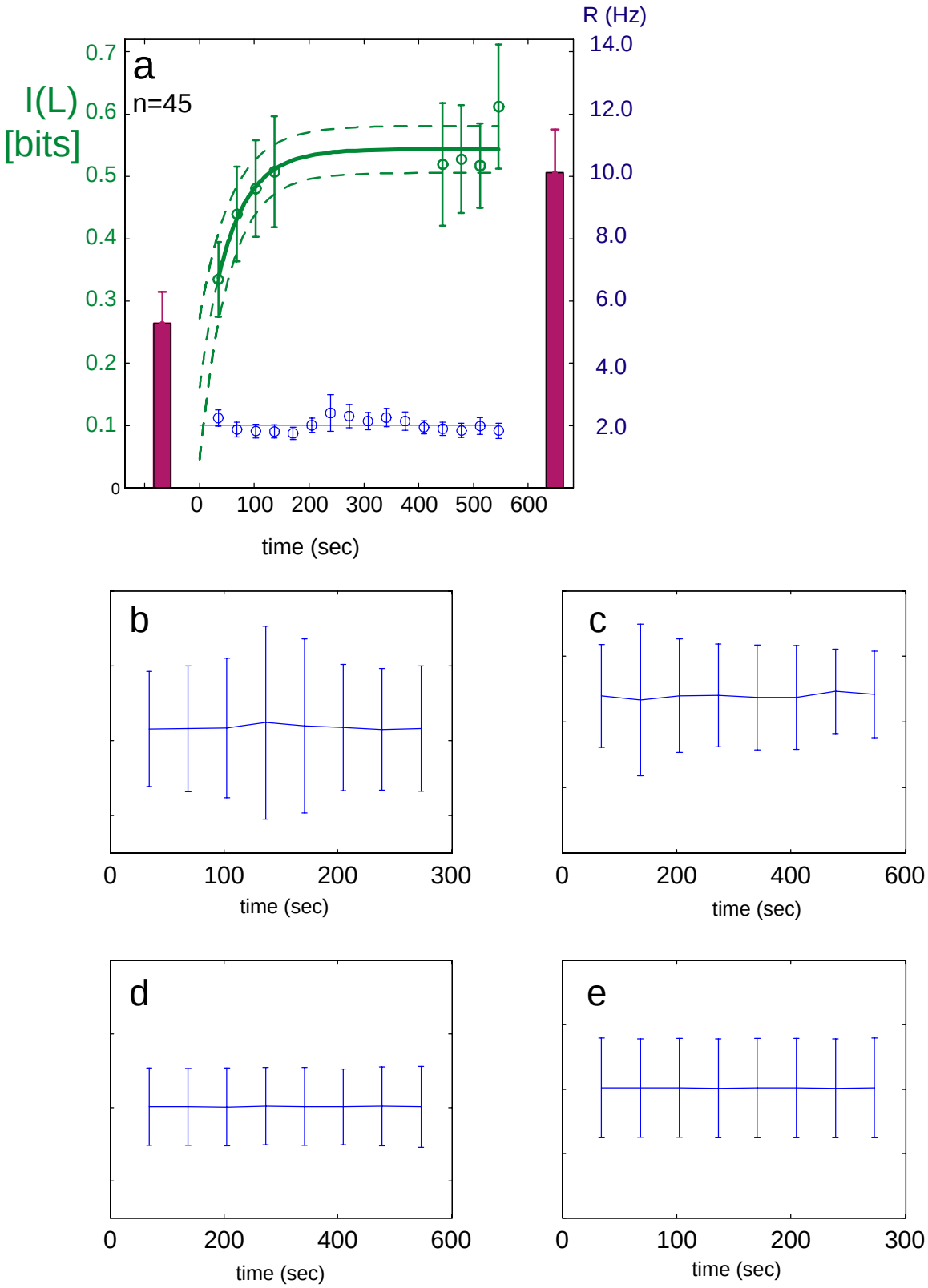


supplementary figure 4

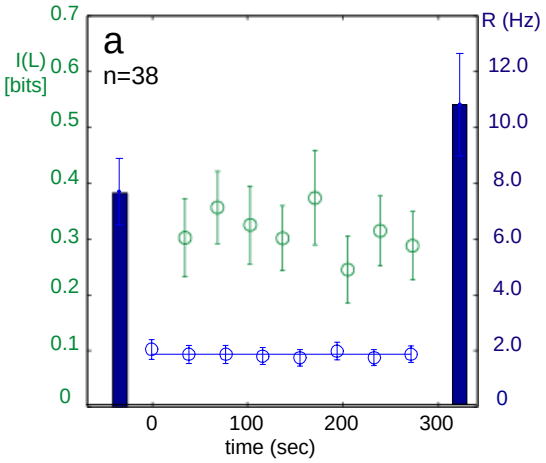


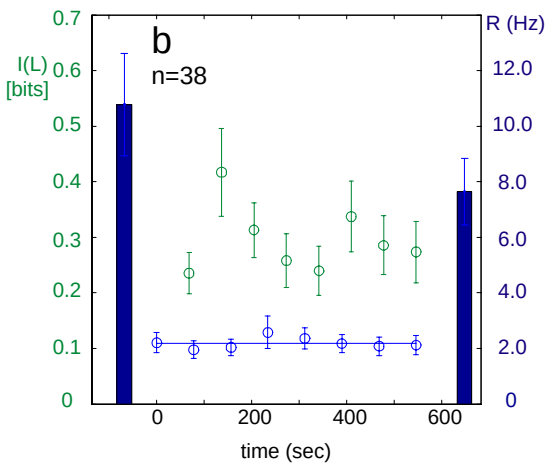


supplementary figure 5



supplementary figure 6





supplementary figure 7