www.sciencemag.org/cgi/content/full/331/6013/83/DC1

Supporting Online Material for

Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment

Pietro Berkes,* Gergő Orbán, Máté Lengyel, József Fiser

*To whom correspondence should be addressed. E-mail: berkes@brandeis.edu

Published 7 January 2011, *Science* **331** 83 (2010)
DOI: 10.1126/science.1195870

This PDF file includes

Methods

Animal preparation and data acquisition

Neural activity was recorded from 16 ferrets at different stages of visual development (Table S1). The details of the surgical procedures have been described in detail previously (*S1*, *S2*), and were approved by the University of Rochester Committee on Animal Research. Briefly, a linear array of 16 electrodes, spaced by 200 micrometers, was implanted in layer 2-3 of the primary visual cortex (V1) under isoflurane anesthesia. The electrodes typically provided clear multi-unit (and occasionally single-unit) signal on each channel. The signal was pre-processed by band-pass filtering ($600 - 6000$ Hz) and digitized at 10 kHz. Spike discrimination was performed offline by manually setting a separate voltage threshold for each electrode. Stable recordings were maintained for $8 - 12$ hours.

Visual stimulation

Shortly after recovery from surgery, neural activity in response to different stimulus ensembles was recorded in awake animals. Animals rested on a padded platform with their head fixed to a rigid metal post and were free to make natural eye movements. Stimuli were displayed on a 4×3 feet back-projection screen at a distance of 30 cm from the head, covering 130×100 degrees of visual angle. The screen resolution was 1024×768 pixels, with a refresh rate of 75 Hz.

Animals were presented with four stimulus conditions:

- *Movie evoked activity*: Stimuli consisted of a movie (the trailer for the film *The Matrix*), presented at a resolution of 720×480 pixels and a frame rate of 24 Hz. This stimulus ensemble is meant to capture the distribution of the statistics of natural stimuli at the level of the simple visual elements encoded by V1 neurons.

- *Noise evoked activity*: Random noise was generated as a grid of black and white squares, each occupying 5×5 degrees of visual angle. A new pattern was generated at random at each screen refresh, with white squares appearing independently with probability $1/4$.

- *Grating evoked activity*: Stimuli consisted of a sequence of full-field, drifting sinusoidal gratings with random frequency and orientation (2 sec per grating, 5 frequencies at $0.5 - 8$ deg/cycle, 9 orientations at intervals of 20 deg).

- *Spontaneous activity*: Neural activity was recorded in complete darkness eliminating all light sources. Control tests found no significant differences in spontaneous neural signals recorded in the experimental setup vs. in a closed box padded with black clothes (*S1*).

Recordings with different stimulus ensembles were performed in interleaved trials of 100 sec, 15 trials for each ensemble, for a total of 25 min of recording in each condition. Table S1 reports the number of animals for each age group recorded in each condition.

Data analysis

Spike timing data over the 16 electrode channels was discretized in 2 ms bins and binarized for each stimulus condition, a, thus yielding a 16-bit binary word, $\mathbf{r}_t^{(a)}$, in each time bin t (Fig. 2A). Neural activity in each condition was thus represented as a sequence of these words from time bin 1 through T, $\mathbf{r}_{1:T}^{(a)}$. We then constructed the empirical distribution of patterns in each condition:

$$\hat{p}^{(a)}_{\text{static}}(\mathbf{r}) = \frac{1}{T}\sum_{t=1}^{T} \delta_{\mathbf{r},\mathbf{r}_t^{(a)}} \qquad \text{(Eq. S1)}$$

simply representing the frequency with which activity pattern $\mathbf{r}$ was observed in $\mathbf{r}_{1:T}^{(a)}$.

To assess the contribution of the correlational structure in our results, we also constructed surrogate distributions by assuming all channels were independent, and computing the product of the marginal empirical distributions

$$\hat{p}^{(a)}_{\text{factorized static}}(\mathbf{r}) = \prod_i \hat{p}^{(a)}_{\text{static}}(r_i) \ , \qquad \text{(Eq. S2)}$$

where the marginal distribution $\hat{p}^{(a)}_{\text{static}}(r_i) = \sum_{\mathbf{r}'} \delta_{r_i,r_i'} \hat{p}^{(a)}_{\text{static}}(\mathbf{r}')$ represented the frequency with which channel i took on the value r_i (irrespective of the other channels). This manipulation left the marginal distributions (and thus the firing rate over individual channels) intact, but removed all statistical dependencies between channels.

The empirical distributions of transitions between neural activity patterns, $\hat{p}^{(a)}_{\text{transition}}(\mathbf{r}_{t+\tau}, \mathbf{r}_t)$, were similarly collected by measuring the frequency with which each pattern $\mathbf{r}_t$ was followed by pattern $\mathbf{r}_{t+\tau}$ after τ msec in condition a. The transition distributions over all the 16 channels would have required filling 2^{32} histogram bins in the empirical distributions, which would have resulted in severe undersampling of the histogram even with the large amount of data at our disposal. Analyses on transition probabilities were thus limited to every second channel in the array, thus keeping the total number of histogram bins equal to that in the static case. The contribution of temporal dependencies was assessed by factorizing the distribution over time, assuming that successive time bins were independent, which was obtained by replacing the full transition distribution by

$$\hat{p}^{(a)}_{\text{factorized transition}}(\mathbf{r}_{t+\tau}, \mathbf{r}_t) = \hat{p}^{(a)}_{\text{static}}(\mathbf{r}_{t+\tau}) \cdot \hat{p}^{(a)}_{\text{static}}(\mathbf{r}_t) \ . \qquad \text{(Eq. S3)}$$

This manipulation kept all spatial (across-channel) correlations of all orders intact, but removed statistical dependencies in time.

Dissimilarity of neural activity distributions

We compared the distributions of neural activity patterns under different stimulus conditions using a standard information-theoretical measure of dissimilarity known as the Kullback-Leibler (KL) divergence (*S3*):

$$\mathrm{KL}\Big[p^{(1)} \parallel p^{(2)}\Big] = \sum_{\mathbf{r}} p^{(1)}(\mathbf{r}) \log \frac{p^{(1)}(\mathbf{r})}{p^{(2)}(\mathbf{r})} \ , \qquad \text{(Eq. S4)}$$

for two distributions, $p^{(1)}$ and $p^{(2)}$, over activity patterns, $\mathbf{r}$. The KL divergence is zero if and only if $p^{(1)} = p^{(2)}$ and increases with increasing dissimilarity. An intuitive interpretation of this measure is the cost in bits of encoding the activity patterns $\mathbf{r}$, distributed as $p^{(1)}$, using distribution $p^{(2)}$ instead of the optimal encoding distribution, i.e., $p^{(1)}$ (*S4*). An advantage of this measure is that it takes into account statistical dependencies of all orders between all channels, as opposed for example to comparing the distribution of correlations between pairs of channels across stimulus conditions.

In the case of transition probability distributions, we compared conditional distributions, $p^{(1)}(\mathbf{r}_{t+\tau}|\mathbf{r}_t)$ and $p^{(2)}(\mathbf{r}_{t+\tau}|\mathbf{r}_t)$, instead of the joint distributions, in order to eliminate differences that are due to dissimilarities in the spatial distributions, $p^{(1)}(\mathbf{r}_t)$ and $p^{(2)}(\mathbf{r}_t)$, since these have been considered already in the analysis for the static case. This was done by averaging the KL divergence over $p^{(1)}(\mathbf{r}_t)$:

$$\left\langle \mathrm{KL}\Big[p^{(1)}(\mathbf{r}_{t+\tau}|\mathbf{r}_t) \parallel p^{(2)}(\mathbf{r}_{t+\tau}|\mathbf{r}_t)\Big] \right\rangle_{p^{(1)}(\mathbf{r}_t)} \tag{Eq. S5}$$

$$= \mathrm{KL}\Big[p^{(1)}(\mathbf{r}_{t+\tau}, \mathbf{r}_t) \parallel p^{(2)}(\mathbf{r}_{t+\tau}, \mathbf{r}_t)\Big] - \mathrm{KL}\Big[p^{(1)}(\mathbf{r}_t) \parallel p^{(2)}(\mathbf{r}_t)\Big] \ , \tag{Eq. S6}$$

i.e., the average KL divergence between conditional transition probabilities could be computed as the difference between the KL divergence of the joint transition probabilities and the KL divergence of the static probabilities.

A successful estimation of KL divergence between neural activity distribution must take into account the uncertainty about the real underlying distributions that is the result of observing only a limited amount of samples, and a potential bias in the estimation of information theoretical quantities (*S5*). We addressed these issues using a Bayesian estimator for the KL, followed by the extrapolation of the number of samples to infinity, as described in the next section.

Estimation of the Kullback-Leibler divergence between neural activity distributions

The Kullback-Leibler (KL) divergence is a measure of dissimilarity between two distributions that is also appropriate to measure the efficiency of an encoding system. However, estimating KL divergence reliably when the two distributions whose divergence needs to be measured, $p^{(1)}$ and $p^{(2)}$, are only known through samples requires two issues to be resolved.

First, under experimental conditions we do not observe $p^{(1)}$ and $p^{(2)}$ directly, but only a limited amount of samples from them. Because the empirical distributions of samples are slightly different from the underlying distributions we cannot determine $p^{(1)}$ and $p^{(2)}$ with absolute certainty. We addressed this issue by taking into account the resultant uncertainty about $p^{(1)}$ and $p^{(2)}$ explicitly through Bayesian inference: we computed posterior distributions over these distributions (represented as multinomial vectors*, $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$) given the experimental data recorded

$$\mathrm{P}\Big(\boldsymbol{\pi}^{(a)}|\mathbf{r}^{(a)}_{1:T}\Big) \quad \propto \quad \mathrm{P}\Big(\mathbf{r}^{(a)}_{1:T}|\boldsymbol{\pi}^{(a)}\Big)\,\mathrm{P}\Big(\boldsymbol{\pi}^{(a)}\Big) \qquad a \in \{1,2\}\ . \tag{Eq. S7}$$

*For the purposes of this section we represent probability distributions over the 2^{16} possible 16-bit binary patterns with vectors of 2^{16} elements, $\boldsymbol{\pi}^{(a)}$, containing the parameters of the corresponding multinomial distributions, such that the probability of a 16-bit binary pattern $\mathbf{r}$, $p^{(a)}(\mathbf{r})$, is given by element $\pi^{(a)}_j$, with scalar index $j = \left(2^0\, 2^1\, 2^2 \ldots 2^{16}\right) \cdot \mathbf{r}$. The KL divergence between two such vectors is computed equivalently to Eq. S4 as $\mathrm{KL}\big[\boldsymbol{\pi}^{(1)} \parallel \boldsymbol{\pi}^{(2)}\big] = \sum_j \pi^{(1)}_j \log \frac{\pi^{(1)}_j}{\pi^{(2)}_j}$.

Activity patterns were assumed to be sampled i.i.d. from the respective multinomial distributions (but see the previous section for the analysis of temporal dependencies), $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$:

$$\mathrm{P}\Big(\mathbf{r}_{1:T}^{(a)}|\boldsymbol{\pi}^{(a)}\Big) \;=\; \prod_{t=1}^{T} p^{(a)}\Big(\mathbf{r}_t^{(a)}\Big) = \prod_j \Big[\pi_j^{(a)}\Big]^{T\,\hat{\pi}_j^{(a)}} \qquad a \in \{1,2\}\;, \tag{Eq. S8}$$

where $\hat{\boldsymbol{\pi}}^{(a)}$ was the multinomial vector corresponding to the empirical distribution (as defined in Eq. S1), and so $T\,\hat{\pi}_j^{(a)}$ was the number of times the activity pattern with scalar index j was observed in $\mathbf{r}_{1:T}^{(a)}$. The prior over the underlying distributions was

$$\boldsymbol{\pi}^{(a)} \;\sim\; \mathrm{Dirichlet}\Big(\boldsymbol{\alpha}^{(a)}\Big)\,,\ \text{where } \boldsymbol{\alpha}^{(a)} = \mathbf{1} \qquad a \in \{1,2\}\;, \tag{Eq. S9}$$

which allowed us to re-write the posterior in Eq. S7 in a more compact form:

$$\boldsymbol{\pi}^{(a)}|\mathbf{r}_{1:T}^{(a)} \;\sim\; \mathrm{Dirichlet}\Big(\boldsymbol{\alpha}'^{(a)}\Big)\,,\ \text{where } \boldsymbol{\alpha}'^{(a)} = \boldsymbol{\alpha}^{(a)} + T\,\hat{\boldsymbol{\pi}}^{(a)} \qquad a \in \{1,2\}\;. \tag{Eq. S10}$$

We then estimated the KL divergence between $p^{(1)}$ and $p^{(2)}$ as the average over these posterior distributions:

$$\overline{\mathrm{KL}}_T \;=\; \int \mathrm{KL}\Big[\boldsymbol{\pi}^{(1)} \parallel \boldsymbol{\pi}^{(2)}\Big]\ \mathrm{P}\Big(\boldsymbol{\pi}^{(1)}|\mathbf{r}_{1:T}^{(1)}\Big)\ \mathrm{P}\Big(\boldsymbol{\pi}^{(2)}|\mathbf{r}_{1:T}^{(2)}\Big)\ \mathrm{d}\boldsymbol{\pi}^{(1)}\,\mathrm{d}\boldsymbol{\pi}^{(2)}\;, \tag{Eq. S11}$$

where $\overline{\mathrm{KL}}_T$ was the estimate of the KL divergence of the two neural activity distributions from which the observed data, $\mathbf{r}^{(1)}$ and $\mathbf{r}^{(2)}$, were sampled. The subscript T indicates that we took into consideration data points from time 1 to T (the last data point). It was possible to derive a closed-form equation for Eq. S11 by solving the integral:

$$\overline{\mathrm{KL}}_T \;=\; \sum_j \frac{\alpha_j'^{(1)}}{\alpha_0'^{(1)}}\Big[\psi\Big(\alpha_j'^{(1)}+1\Big) - \psi\Big(\alpha_0'^{(1)}+1\Big)\Big] - \sum_j \frac{\alpha_j'^{(1)}}{\alpha_0'^{(1)}}\Big[\psi\Big(\alpha_j'^{(2)}\Big) - \psi\Big(\alpha_0'^{(2)}\Big)\Big] \tag{Eq. S12}$$

$$\text{where} \qquad \alpha_0'^{(a)} = \sum_j \alpha_j'^{(a)} \qquad a \in \{1,2\}\;, \tag{Eq. S13}$$

and ψ is the digamma function.

A second, well-known issue that needed to be addressed was that the estimation of information theoretical quantities from a limited number of samples is biased (*S5*). In order to compensate for this bias, we followed the method proposed in (*S6*) for the estimation of the entropy of neural distributions: we computed the value of the estimator for fractions of the data and extrapolated for the number of samples tending to infinity. This was done by fitting $\overline{\mathrm{KL}}_{T'}$ (Eq. S12) as a polynomial function of $1/T'$ with parameters β_0, β_1, and β_2 at $T' = T$, $T/2$, and $T/4$:

$$\overline{\mathrm{KL}}_{T'} = \beta_0 + \frac{\beta_1}{T'} + \frac{\beta_2}{T'^2} \quad . \tag{Eq. S14}$$

When using fractions of the data, we computed the average estimate for multiple non-overlapping blocks. For example, for $T' = T/2$ we computed $\overline{\mathrm{KL}}_{T/2}$ for the first and second half of the data

separately and then averaged these two estimates. The final value for the KL estimation was obtained by taking the limit $T' \to \infty$ in Eq. S14, which gave

$$\overline{\mathrm{KL}}_\infty = \beta_0 \quad . \tag{Eq. S15}$$

Validation experiments show that the method described above was appropriate for accurately estimating the KL divergence of distributions of the same size as the neural activity distributions analyzed in the main paper (2^{16} possible patterns), given the same number of samples ($750,000$ samples). For each run, we drew two distributions over 2^{16} states, $p^{(1)}$ and $p^{(2)}$, from a Dirichlet prior with parameters $\boldsymbol{\alpha}^{(1)} = \boldsymbol{\alpha}^{(2)} = \mathbf{1}$, and measured their true KL divergence. We then drew 750,000 samples from each of the two distributions, $\mathbf{r}^{(1)}_{1:T}$ and $\mathbf{r}^{(2)}_{1:T}$, from which we estimated the KL divergence according to the method described above. Figure S4a shows the histogram of percent error over multiple runs. The results show that the method is unbiased and the error is very small (average percent error is -0.003 ± 0.3).

Model selection test

In the paper the divergences of neural activity distributions were often compared under different conditions (e.g., comparing the divergence between movie-EA and SA with the divergence between movie-EA and surrogate-SA, Fig. 3, or comparing the divergence between movie-EA and SA with the divergence between noise-EA and SA, Fig. 4). Establishing whether two samples are significantly different from one another is known as the two-sample problem, and is most often addressed by applying Student's t-test. However, our data have two characteristics which are known to cause the t-test to have low statistical power: a small number of samples from each population (between 2 and 6 samples, i.e., animals), and widely different variances. To mitigate this problem, we used a more flexible and powerful statistical test based on model selection. Even though the construction of the test follows the standard hypothesis testing paradigm, we report all the steps here in detail, as (perhaps surprisingly) it is uncommon to design a statistical test adapted to the task at hand. At the end of the section we extensively validate the test by showing that it typically outperforms a two-samples t-test in cases similar to the ones analyzed in the paper.

The model selection test proceeds by computing the probability of observations (here: estimated KL divergences) coming from the two populations, $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}$ to be compared under three different hypotheses.

- M_0 is the null-hypothesis and assumes that all the data was drawn from a single Gaussian distribution with unknown mean, μ_0, and variance, σ_0^2:

$$\mathbf{y}^{(1)}, \mathbf{y}^{(2)} \quad \sim \quad \mathrm{Normal}\big(\mu_0, \sigma_0^2\big) \ . \tag{Eq. S16}$$

- M_1 assumes that the data was drawn from two Gaussian distributions with unknown but different means, μ_1 and μ_2, and equal variances, σ_{12}^2:

$$\mathbf{y}^{(1)} \quad \sim \quad \mathrm{Normal}\big(\mu_1, \sigma_{12}^2\big) \tag{Eq. S17}$$
$$\mathbf{y}^{(2)} \quad \sim \quad \mathrm{Normal}\big(\mu_2, \sigma_{12}^2\big) \ . \tag{Eq. S18}$$

- M_2 assumes that the data was drawn from two Gaussian distributions with unknown but different means, μ_1 and μ_2, and different variances, σ_1^2 and σ_2^2:

$$\mathbf{y}^{(1)} \sim \text{Normal}\big(\mu_1, \sigma_1^2\big) \quad \text{(Eq. S19)}$$
$$\mathbf{y}^{(2)} \sim \text{Normal}\big(\mu_2, \sigma_2^2\big) \ . \quad \text{(Eq. S20)}$$

Using these two alternative hypotheses ensures that the test has good statistical power independently of whether the underlying distributions have the same or different variance.

The test statistic for hypothesis testing expresses the relative strength of belief that at least one of the alternative hypotheses, M_1 or M_2, is more probable than the null hypothesis, M_0. For this, we defined the test statistic as the maximum of the two marginal likelihood ratios given the data, $\mathbf{y} = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}\}$:

$$m = \max\left\{\frac{\mathrm{P}(\mathbf{y}|M_1)}{\mathrm{P}(\mathbf{y}|M_0)}, \frac{\mathrm{P}(\mathbf{y}|M_2)}{\mathrm{P}(\mathbf{y}|M_0)}\right\} \quad \text{(Eq. S21)}$$

High values of m indicated that one of the alternative hypotheses had higher probability than the null hypothesis given the observed data, and thus that the two populations were likely to be distinct.

Hypothesis testing proceeded as standard for statistical tests by computing the distribution of m under M_0, and reporting the P-value for the test statistics computed on observed data:

$$P = \mathrm{P}(m > m_{\text{observed}}|M_0) \quad \text{(Eq. S22)}$$

The distribution of m depends on the number of samples, N_1 and N_2, in the two populations. Thus, we computed tables for these distributions based on Monte Carlo simulations by drawing different numbers, $N_{1/2}$, of random samples from M_0 and computing m for each $N_{1/2}$ according to Eq. S21.

In order to compute the individual marginal likelihoods, $\mathrm{P}(\mathbf{y}|M_i)$, the unknown parameters, θ (here: means and variances of the normal distributions) were marginalized out by Monte Carlo integration:

$$\mathrm{P}\Big(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|M_i\Big) = \int \mathrm{P}\Big(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\theta, M_i\Big)\ \mathrm{P}(\theta|M_i)\ \mathrm{d}\theta \quad \text{(Eq. S23)}$$
$$\simeq \frac{1}{N_{\text{samples}}} \sum_{\theta_j \sim \mathrm{P}(\theta|M_i)} \mathrm{P}\Big(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\theta_j, M_i\Big)\ , \quad \text{(Eq. S24)}$$

with $i \in \{0, 1, 2\}$. Integrating over the parameters ensured that the alternative hypotheses, M_1 and M_2, did not have higher marginal likelihoods simply for having a higher number of parameters, due to the "automatic Occam's razor" property of Bayesian model selection (*S4*, Ch. 28): models with multiple parameters are automatically penalized because the probability mass over parameters, $\mathrm{P}(\theta|M_i)$, is spread over a larger volume that does not necessarily overlap with the parts of parameter space where the likelihood, $\mathrm{P}\big(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\theta, M_i\big)$, has high values. As a consequence, the three alternative hypotheses are directly comparable, which makes it possible to combine them in a single test statistic (Eq. S21). This is in contrast with likelihood ratio tests, that use as a test statistic the ratio of the likelihood of the data given maximum likelihood parameters. In such a case, models with a higher number of parameters typically have a higher likelihood, which may lead to overfitting (particularly with a small number of samples), and thus to a high probability of Type I errors. Note that being careful about these issues is only important for increasing the

power of our test, and it does not influence Type I error rates as this will be computed using the distribution of our statistics as in classical statistical tests, and will thus be guaranteed to be below threshold.

For computing the marginal likelihood of each hypothesis (Eq. S24) we defined prior distributions over the unknown parameters of the corresponding distribution(s), i.e. the mean(s) and variance(s), as described below. Our choice for the prior distributions was based on two basic ideas: First, the uncertainty over the mean of the distributions should be inversely related to the number of data points. We thus defined the prior over the means to be a normal distribution with a standard deviation that scaled as the standard error of the mean (s.e.m.), i.e., as $1/\sqrt{N}$. This implied that the uncertainty over the mean under M_0 would be smaller than that of the means under M_1 and M_2, since M_0 considered the data points from both populations. Second, we chose a broad, uniform distribution as the prior over the standard deviation of the data. In order to simplify the notation, and without loss of generality, we assume in the following that $\mathbf{y}$ has a mean of 0 and an empirical variance of 1, which can always be achieved by shifting and rescaling the data.

- Under the null hypothesis, M_0 (Eq. S16), the prior over μ_0 was a normal distribution centered at the empirical mean, $\overline{\mathbf{y}}$, with a standard deviation equal to the standard error of the mean, $1/\sqrt{N}$. For the standard deviation, σ_0, we chose an interval bounded below by zero (since σ_0 is a positive quantity), and above by 3 times the empirical standard deviation of the data (which is 1 due to normalization). This interval is quite conservative, as the mean would need to be far outside the observed range for the real standard deviation to reach that value. We thus defined the prior as a uniform distribution between these two bounds. In summary,

$$\mu_0|M_0 \sim \text{Normal}\left(\overline{\mathbf{y}}, \frac{1}{N}\right) \quad \text{(Eq. S25)}$$

$$\sigma_0|M_0 \sim \text{Uniform}(\epsilon, 3) \quad \text{(Eq. S26)}$$

  where $\epsilon = 10^{-3}$ was a value close to zero to avoid numerical errors, and $N = N_1 + N_2$ was the total number of samples from the two populations.

- For the first alternative hypothesis, M_1 (Eq. S17-Eq. S18), the priors over the parameters were defined as for M_0, with the exception of the range for σ_{12}, which was decreased to 1 in order to reflect the fact that the the data from the two individual populations must have a smaller standard deviation than the combined data:

$$\mu_1|M_1 \sim \text{Normal}\left(\overline{\mathbf{y}^{(1)}}, \frac{1}{N_1}\right) \quad \text{(Eq. S27)}$$

$$\mu_2|M_1 \sim \text{Normal}\left(\overline{\mathbf{y}^{(2)}}, \frac{1}{N_2}\right) \quad \text{(Eq. S28)}$$

$$\sigma_{12}|M_1 \sim \text{Uniform}(\epsilon, 1)\ , \quad \text{(Eq. S29)}$$

  where N_1 and N_2 were the number of samples from the two populations.

- Finally, priors for the second alternative hypothesis, M_2, were defined similarly as for M_1:

$$\mu_1|M_2 \sim \text{Normal}\left(\overline{\mathbf{y}^{(1)}}, \frac{1}{N_1}\right) \quad \text{(Eq. S30)}$$

$$\mu_2|M_2 \sim \text{Normal}\left(\overline{\mathbf{y}^{(2)}}, \frac{1}{N_2}\right) \quad \text{(Eq. S31)}$$

$$\sigma_1|M_2 \sim \text{Uniform}(\epsilon, 1) \quad \text{(Eq. S32)}$$

$$\sigma_2|M_2 \sim \text{Uniform}(\epsilon, 1) \quad \text{(Eq. S33)}$$

The P-value corresponds to the Type I error of rejecting the null hypothesis even though the data was drawn from M_0, and is thus the quantity that we were interested in reporting for our data. In order to compare the test with others, we also needed to consider the Type II error of failing to reject the null hypothesis even though the data came from two different distributions. We computed the Type II error by sampling N samples from two Gaussian distributions, one with zero mean and unit variance, and the second with mean 1 and standard deviation between 0.25 and 2. Type I error was kept fixed at 5% by rejecting the null-hypothesis whenever the P-value of the test was smaller than 0.05. Figure S4b–d shows the Type II error over $50,000$ runs for the t-test (Fig. S4b) and the model selection test (Fig. S4c). For both tests, the Type II error decreased with the number of samples and with decreasing standard deviation of the second population, as $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ became more separated. The Type II error of the model selection test showed a remarkable improvement of up to 25% over the t-test for the vast majority of cases (Fig. S4d). The single exception was when the assumptions of the t-test were exactly matched, with the two population distributions having the same variance, in which case there was a small increase in Type II error of at most 1% (Fig. S4d, green line).

Sparseness measures

We measured both population sparseness (i.e., the sparseness of the population response at any given time) and lifetime sparseness (i.e., the sparseness of the response of individual neurons over time) of the firing rates $r_{i,t}$ on channel i at time t, collected in 10 msec bins.[†]

The definitions of the sparseness measures were as follows. For lifetime sparseness we used a standard measure of sparseness (*S7*):

$$\text{TR}^i_{\text{lifetime}} = \frac{1}{1-1/T}\left[1 - \frac{\left(\sum_{t=1}^T r_{i,t}/T\right)^2}{\sum_{t=1}^T r_{i,t}^2/T}\right] \quad \text{(Eq. S34)}$$

where T is the total number of data points. TR is defined between zero (less sparse) and one (more sparse), and depends on the shape of the firing rate distribution. For monotonic, non-negative distributions, an exponential decay corresponds to TR = 0.5, and values smaller and larger than 0.5 indicate distributions with lighter and heavier tails, respectively (*S8*). For every animal, we report the value of $\text{TR}^i_{\text{lifetime}}$, averaged over all channels, i (Fig. S2).

[†]Since these firing rates were measured in time bins of longer duration, they were not binarized as in previous sections.

The first measure of population sparseness is a direct adaptation of Eq. S34, where sparseness is measured across neurons instead of time. To eliminate differences in sparseness due to changes in global neural activity across age, we rescaled firing rates as

$$\hat{r}_{i,t} = \frac{r_{i,t}}{\sqrt{\sum_{t'=1}^{T} r_{i,t'}^2 / T}} \tag{Eq. S35}$$

and defined

$$\mathrm{TR}^{t}_{\mathrm{population}} = \frac{1}{1 - 1/N} \left[1 - \frac{\left(\sum_{i=1}^{N} \hat{r}_{i,t}/N \right)^2}{\sum_{i=1}^{N} \hat{r}_{i,t}^2/N} \right] \tag{Eq. S36}$$

where N is the total number of channels. Note that sums are now taken over channels (index i) rather than time (index t). We discarded bins with no neural activity, as population sparseness is undefined in such cases. In Fig. S2 we report the distribution of $\mathrm{TR}^{t}_{\mathrm{population}}$, averaged over time, t, for each animal.

The second measure of population sparseness is known as *activity sparseness*, and it directly quantifies the number of neurons active at any time (*S9*):

$$\mathrm{AS} = \frac{1}{T} \sum_{t=1}^{T} \left(1 - \frac{n_t}{N} \right) \tag{Eq. S37}$$

where n_t is the number of active channels at time t. Channel i is defined to be active at time t if its firing rate, $r_{i,t}$, is above the upper 68th percentile of the distribution of firing rates over time (i.e., the equivalent of one standard deviation for non-negative distributions). AS is thus invariant to multiplicative and additive changes in firing rate. However, since it discards most of the information about the shape of the distribution, it is a less sensitive measure than $\mathrm{TR}_{\mathrm{population}}$. AS is defined between 0 and 1, with 1 meaning that no neuron was active above the threshold, and 0 meaning that all neurons were active at all times.

Since our recordings consisted of multi-unit activity, the reported sparseness is a lower bound on the true sparseness. Nevertheless, this did not affect our conclusions because they were based on the relative comparison of sparseness over age. Changes in average firing rate with age also cannot account for our results, as they are factored out in the population sparseness measures, $\mathrm{TR}_{\mathrm{population}}$ and AS. $\mathrm{TR}_{\mathrm{lifetime}}$ is unaffected by multiplicative changes in firing rate, but not by additive changes. This is because lifetime sparseness is motivated by metabolic costs, and should thus depend on the overall level of activity. These results are also discussed in *S10*.

Testing trends over development

Trends over development (Figs. 2B, 3A, 3B, 3D, 4C, S2, S3) were tested using Spearman's correlation. These tests were performed using the exact age (as opposed to the age group) of animals as the independent variable. In other words, we measured trends across the individual animals rather than their age groups.

Supporting Text

Defining statistical optimality of an internal model

An internal model of visual stimuli incorporates assumptions about a set of visual features (anything from low level oriented edges, through visual chunks of medium complexity, to high level objects) that are relevant to describe images. To formalize such an internal model, two components need to be defined: the likelihood of features, $\mathrm{P}(\text{input}|\text{features}, \text{parameters}, \text{model})$ describing the probability with which any given input image can be expected to arise from a particular combination of features (thereby implicitly defining the "meaning" of these features), and the prior distribution of features, $\mathrm{P}(\text{features}|\text{parameters}, \text{model})$, describing the probability with which any particular combination of features can be expected to occur. Based on these, the posterior distribution describing the probability that any given combination of features may have given rise to a particular input can be computed by Bayes' rule:

$$\begin{aligned}&\mathrm{P}(\text{features}|\text{input}, \text{parameters}, \text{model}) \propto \\ &\quad \propto \mathrm{P}(\text{input}|\text{features}, \text{parameters}, \text{model})\ \mathrm{P}(\text{features}|\text{parameters}, \text{model})\end{aligned} \tag{Eq. S38}$$

For such a statistical model to be optimal, two conditions must be fulfilled. (1) Inferences about features in an image must be expressed as posterior distributions for optimal decision making and learning (*S11*). (2) The posterior distribution has to be consistent with the true features that generated the input. In order to achieve this latter goal, it is important that the parameters of the model, that the above formulation made explicit[‡], are adapted to the statistical properties of the input data according to a statistically well-founded criterion. A standard criterion is that the parameters should be such that their likelihood, $\mathrm{P}(\text{input}|\text{parameters}, \text{model})$ (not to be confused with the likelihood of features described above), is maximised:

$$\text{parameters}_{\text{ML}} = \underset{\text{parameters}}{\operatorname{argmax}}\, \mathrm{P}(\text{input}|\text{parameters}, \text{model})\ . \tag{Eq. S39}$$

The likelihood of the parameters can be computed as

$$\begin{aligned}&\mathrm{P}(\text{input}|\text{parameters}, \text{model}) = \\ &\quad = \int \mathrm{P}(\text{input}|\text{features}, \text{parameters}, \text{model})\, \mathrm{P}(\text{features}|\text{parameters}, \text{model})\, \mathrm{d}\,\text{features}\,.\end{aligned} \tag{Eq. S40}$$

As we explain in the next section, testing the maximum likelihood criterion of optimality in neural data is challenging. However, it is easy to see that maximizing the likelihood of parameters is equivalent to minimizing the KL divergence between the true distribution of inputs, $\mathrm{P}^*(\text{input})$, and the distribution of inputs predicted by the model with its parameters (*S4*):

$$\text{parameters}_{\text{ML}} = \underset{\text{parameters}}{\operatorname{argmin}}\, \mathrm{KL}[\mathrm{P}^*(\text{input}) \parallel \mathrm{P}(\text{input}|\text{parameters}, \text{model})] \tag{Eq. S41}$$

[‡]For simplicity, we dropped the parameters from the notation in the main text.

The minimal possible value of the KL divergence is zero, and for a model that attains this minimum the KL divergence between the average posterior and the prior over features will also be zero:

$$\begin{aligned}&\mathrm{KL}[\mathrm{P}^*(\text{input}) \parallel \mathrm{P}(\text{input}|\text{parameters}, \text{model})] = 0 \Rightarrow \\ &\quad \Rightarrow \mathrm{KL}\left[\int \mathrm{P}(\text{features}|\text{input}, \text{parameters}, \text{model})\, \mathrm{P}^*(\text{input})\, \mathrm{d}\,\text{input} \parallel \mathrm{P}(\text{features}|\text{parameters}, \text{model})\right] = 0\end{aligned} \tag{Eq. S42}$$

This motivated the use of the latter KL divergence as a benchmark of the optimality of statistical models (*S12*, *S13*), and this is the measure we also adopted in our study (Eq. 1 of the main text) because we could relate it directly to neural data.[§]

Identifying the hallmarks of statistical optimality in neural activity

Identifying neural correlates of statistical optimality in cortical activity has proven to be challenging mainly because basic aspects of internal representations in the cortex are unknown. Therefore, the problem has been addressed by indirect approaches that start from specific assumptions about the precise nature (parametric form) of the internal model and search for neural correlates based on these assumptions.

For example, one dominant indirect approach widely employed in low-level vision is based on a class of internal models working on the assumption that retinal images are linear combinations of underlying visual features (*S14*, *S15*). In this case, the parameters of the underlying statistical model determine the linear features that the model uses to describe images, and so the optimality of the model can be measured by the likelihood of its parameters (as in Eq. S40): the probability with which the linear combinations of features result in natural images. Consequently, the statistical optimality of the internal model encoded in the primary visual cortex (V1) is typically assessed in this approach by comparing receptive fields in V1 to the set of visual features with maximal likelihood, as required by Eq. S39 (*S7*, *S13*, *S14*, *S16*, *S17*).

There are two shortcomings of the indirect approach demonstrated in the example above. First, asserting statistical optimality remains conditional on the specific assumptions about the parametric form of the internal model, which are typically hard to validate. In the above case, the assumption about the linear combination of visual features contradicts the strong non-linear interactions between visual objects due to occlusion and other effects. Even in models where the specific assumption about the linear combination of features was relaxed (*S18*), other assumptions (parametric forms for the model) needed to be made, e.g. that features were sparse (*S14–S17*) or slowly changing (*S19*, *S20*). Confirming the validity of these assumptions proved to be fraught with theoretical and experimental difficulties (*S10*, *S21*, *S22*). Moreover, many of these models did not represent full posterior distributions over features, and so it is unclear how well the parameters that their optimization algorithms found were truly maximising the likelihood as required by Eq. S39 (*S23*). The second shortcoming of these indirect approaches (regardless of whether or not

[§]We note that the direction of causation in Eq. S42 cannot be reversed, so the latter KL divergence (our benchmark measure) being zero does not necessarily imply the former being zero (i.e., that the maximum likelihood values of parameters have been found). A degenerate case in which our benchmark yields zero divergence is when the posterior over features does not depend on the image (i.e., the model is decoupled from its inputs). Importantly, this was not the case in our study, as demonstrated by Figs. 4 and S3.

they assumed linear superposition of features) is that receptive fields alone provide an incomplete description of neural responses (*S24*, *S25*) and therefore analyzing other aspects of neural activity may be necessary for testing the statistical optimality of a model (*S12*, *S26*, *S27*).

The direct approach employed in our study circumvented both of these shortcomings. First, it required making only minimal assumptions about the parametric form of the internal model (see main text, and next section). Second, it used a benchmark of statistical optimality that also took into account aspects of neural variability and, in general, higher order moments of response distributions that had been ignored by previous approaches which only addressed the mean responses of cells. Specifically, we compared the distribution of evoked neural activity patterns averaged over a stimulus ensemble (aEA) to the distribution of spontaneous activity patterns, collected in the dark (Fig. 1). This novel combination of analyses also made our approach distinct from other approaches to neural data analysis (Table S2).

The relation between spontaneous activity and the prior distribution

In the main text we argued that neural activity is the result of the interaction between an internal model of the environment, embedded in the underlying neural circuit, and the sensory input. This interaction corresponds to the way the posterior distribution is computed from the likelihood and the prior according to Bayes' rule (Eq. S38). For high levels of brightness or contrast, this posterior will be dominated by the likelihood of possible interpretations of the input, but as brightness or contrast decreases, the internal model will need to rely more heavily on the prior of the internal model, and so spontaneous activity in darkness will be dominated by this prior distribution (Fig. 1A). In the following, we provide a formal basis for this intuitive argument by showing that in efficient natural image models the posterior distribution in darkness reduces to the prior distribution, in a way consistent with experimental observations in behavioral and electrophysiological studies (*S11*).

The key insight is that an efficient model of images should be able to represent a uniformly dark stimulus with a very simple description. Since such a description would require fixing only a small number of variables in the representation, the rest of the variables would be constrained only by their prior expectations, and would thus be free to match the prior. This effect is a special case of explaining away (*S28*).

This general idea can be illustrated with a toy model of natural images (*S11*). In the model, each input image, $\mathbf{y}$, is represented as a linear superposition of a set of basic, oriented features, $\mathbf{w}_i$ (for simplicity, we will consider only two features in the following), with variables x_i representing the (continuous) local contrast or luminance level at which each feature is present. In addition, the model has a global contrast (or luminance) variable, b. The model is defined through prior distributions over these feature values, $\mathrm{P}(x_1, x_2)$ and $\mathrm{P}(b)$, and the likelihood of features:

$$\mathbf{y}|x_1, x_2, b \quad \sim \quad \mathrm{Normal}\big(b \cdot (\mathbf{w}_1 x_1 + \mathbf{w}_2 x_2), \sigma_y^2 \mathbf{I}\big) \qquad \text{(Eq. S43)}$$

where $\mathbf{I}$ is the identity matrix, thus formalizing the notion that the image is a linear combination of local features with contrast levels x_1 and x_2 scaled by the global contrast level b, with additive (zero-mean) Gaussian pixel noise.

When presented with an image, the goal of inference in this model is to simultaneously infer the local contrast levels of the linear features and the global contrast level, i.e., the posterior distribution

over all the variables:

$$\mathrm{P}(x_1, x_2, b|\mathbf{y}) \qquad \text{(Eq. S44)}$$

A change in global contrast will affect only b, while the posterior distribution over feature-encoding variables will remain largely unaffected (Fig 1A, central panel). This is consistent with the luminance-invariant behavior that has been observed in perceptual (*S29*) and electrophysiological studies (*S30*, *S31*). When the input is a uniform dark stimulus, the global contrast feature is inferred to be tightly distributed around 0. The zero contrast level explains the whole content of the image, and so the posterior distribution over the linear features is unconstrained by the likelihood and free to match the prior (Fig. 1A, right panel).

More formally, the posterior distribution of the global contrast level can be written as

$$\begin{aligned} \mathrm{P}(b|\mathbf{y}=\mathbf{0}) &\propto \mathrm{P}(b) \int \mathrm{P}(\mathbf{y}=\mathbf{0}|x_1, x_2, b)\ \mathrm{P}(x_1, x_2)\ \mathrm{d}x_1\, \mathrm{d}x_2 \\ &= \mathrm{P}(b) \int \exp\left[-\frac{1}{2\pi\sigma_y^2}\|b \cdot (\mathbf{w}_1 x_1 + \mathbf{w}_2 x_2)\|^2\right] \mathrm{P}(x_1, x_2)\ \mathrm{d}x_1\, \mathrm{d}x_2 \end{aligned} \qquad \text{(Eq. S45)}$$

It is easy to see that for $b = 0$ the first term in the integral will be 1 for any setting of x_1 and x_2, while for any other value of b it will quickly decay towards zero for most settings of x_1 and x_2. As a consequence, the integral (the likelihood of b) will be much larger for $b = 0$ than for any other value of b, resulting in a posterior that is heavily skewed towards zero and which thus can be well approximated as $\mathrm{P}(b|\mathbf{y}=\mathbf{0}) \approx \delta(b)$. Using this approximation to express the posterior over x_1 and x_2 we obtain:

$$\begin{aligned} \mathrm{P}(x_1, x_2|\mathbf{y}=\mathbf{0}) &= \int \mathrm{P}(x_1, x_2|b, \mathbf{y}=\mathbf{0})\ \mathrm{P}(b|\mathbf{y}=\mathbf{0})\ \mathrm{d}b = \int \mathrm{P}(x_1, x_2|b, \mathbf{y}=\mathbf{0})\ \delta(b)\, \mathrm{d}b \\ &\propto \mathrm{P}(x_1, x_2)\ \underbrace{\mathrm{P}(\mathbf{y}=\mathbf{0}|x_1, x_2, b=0)}_{=1} = \mathrm{P}(x_1, x_2) \end{aligned} \qquad \text{(Eq. S46)}$$

and so the posterior over x_1 and x_2 reduces to the corresponding prior distribution. Note that the global contrast level, b, influences only indirectly the uncertainty over x_1 and x_2: while in the model global and local feature contrasts can be varied independently, for a given image an inferred low global contrast level will influence the belief in the contrast of the individual local features through the multiplicative interaction in Eq. S43.

The explaining away effect illustrated above is typical of a large class of natural image models. The toy model proposed above is representative of models that take into account the co-variation of visual features (*S12*, *S17*). Considering these common modulations leads to a more efficient representation of images, since it removes redundant information (e.g., luminance) from the representation of individual features. The same effect would also be obtained in models that represent the presence and appearance of visual elements separately (*S32*), in which case the dark stimulus would be explained away by the absence of all elements. Finally, image models that include occlusion (*S18*) would explain dark images as one dark object occluding the rest of the scene. Even though our example is a directed, causal model of images, this is not necessary for our argument. Similar effects would be observed in an undirected model, as long as a uniform stimulus could be explained away by a small subset of the variables.

Supporting Figures

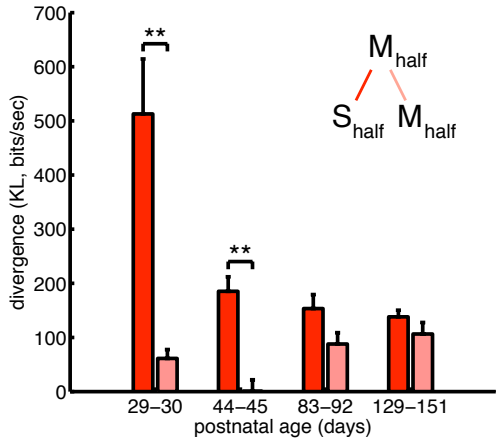

Fig. S1: Baseline KL divergence between evoked and spontaneous activity. Comparing the divergence between movie-evoked activity and spontaneous activity to the baseline divergence between movie-evoked activity and itself. Divergences are computed by splitting the data into two halves and averaging over all comparisons. For instance, for the baseline divergence between movie-evoked activity with itself we split the neural responses into two halves, $\mathbf{r}^{(\mathrm{M})}_{1:\frac{1}{2}}$ and $\mathbf{r}^{(\mathrm{M})}_{\frac{1}{2}:T}$, where $\mathbf{r}^{(\mathrm{M})}_{t_1:t_2}$ is movie-evoked neural activity between time bins t_1 and t_2, and computed $\frac{1}{2}\left[\overline{\mathrm{KL}}_\infty\left(\mathbf{r}^{(\mathrm{M})}_{1:\frac{T}{2}}, \mathbf{r}^{(\mathrm{M})}_{\frac{T}{2}:T}\right) + \overline{\mathrm{KL}}_\infty\left(\mathbf{r}^{(\mathrm{M})}_{\frac{T}{2}:T}, \mathbf{r}^{(\mathrm{M})}_{1:\frac{T}{2}}\right)\right]$. Error bars represent s.e.m in all figures, significance levels are $^{**}p < 0.01$, otherwise $p > 0.05$ (m-test); see Methods for further details.

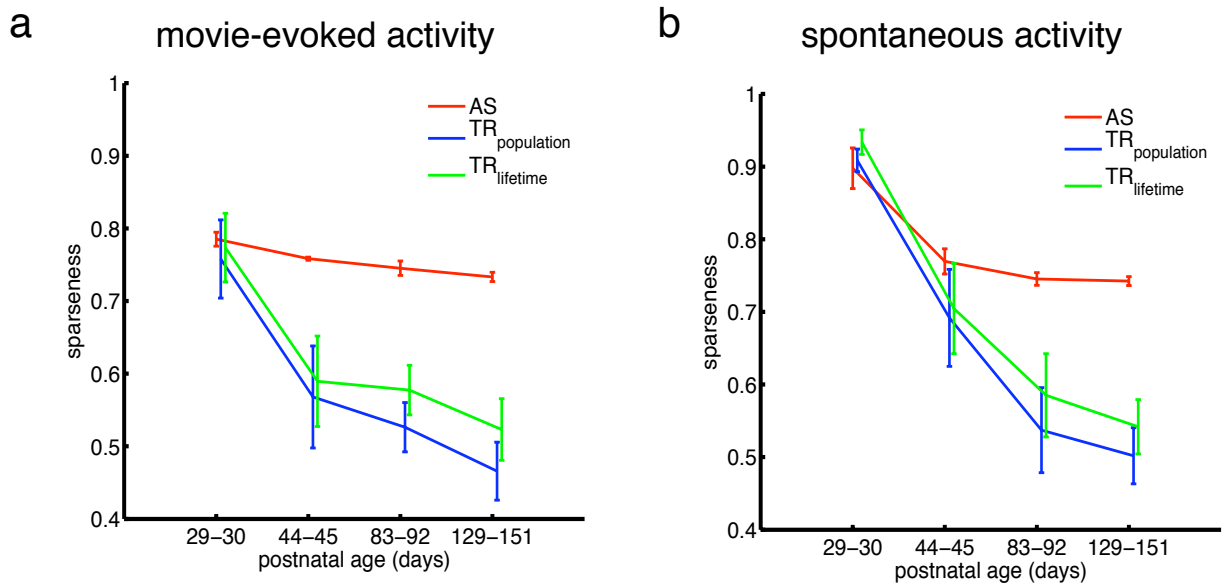

Fig. S2: Sparseness of neural representations over development. Changes over development of three measures of sparseness: activity sparseness (AS, red), population sparseness ($\mathrm{TR}_{\mathrm{population}}$, blue), and lifetime sparseness ($\mathrm{TR}_{\mathrm{lifetime}}$, green), as defined in Methods. All three measures decreased significantly over development, both for movie-evoked activity (activity sparseness: Spearman's $\rho = -0.79$, $P = 0.0004$; population sparseness: $\rho = -0.75$, $P = 0.001$; lifetime sparseness: $\rho = -0.65$, $P = 0.009$), and spontaneous activity (activity sparseness: Spearman's $\rho = -0.61$, $P = 0.01$; population sparseness:$\rho = -0.78$, $P = 0.0006$; lifetime sparseness: $\rho = -0.75$, $P = 0.001$). Thus, neural activity distributions became less sparse as the internal model adapted to natural image statistics (Fig. 2B in main text), in contrast to the trend predicted by sparse coding models (*S10*). See Methods for details.

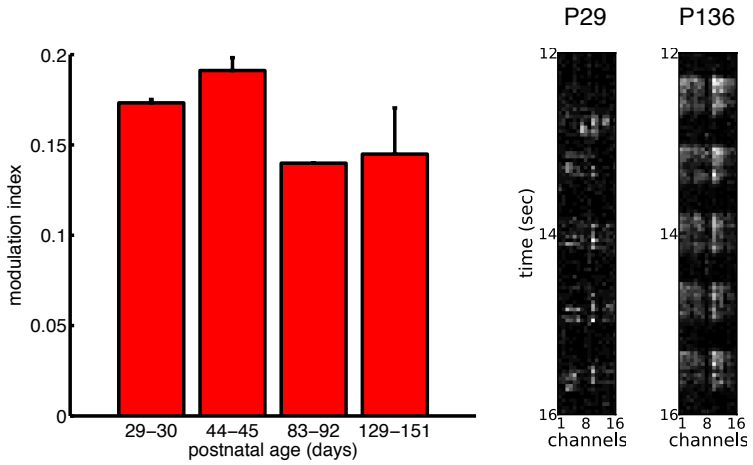

Fig. S3: Modulation of neural activity by grating stimuli over development. For each animal and channel we computed the average Fourier spectrum of neural response in 2-second windows. We defined the *modulation index* as the ratio of the power at the temporal frequency of the oscillation of the grating to the power at zero frequency, in order to normalize for changes in average firing rate. We then averaged the modulation index across channels in each animal. The histogram shows the average ($\pm$ s.e.m.) modulation index in the four age groups. The modulation index did not change significantly over age (Sperman's $\rho = -0.27$, p=0.49), suggesting that neurons were equally responsive to external stimuli at different stages of development. The two panels on the right show two examples of neural activity on the 16 electrodes for a young animal at age P29 (left) and an adult animal at age P136 (right) in response to a grating stimulus.

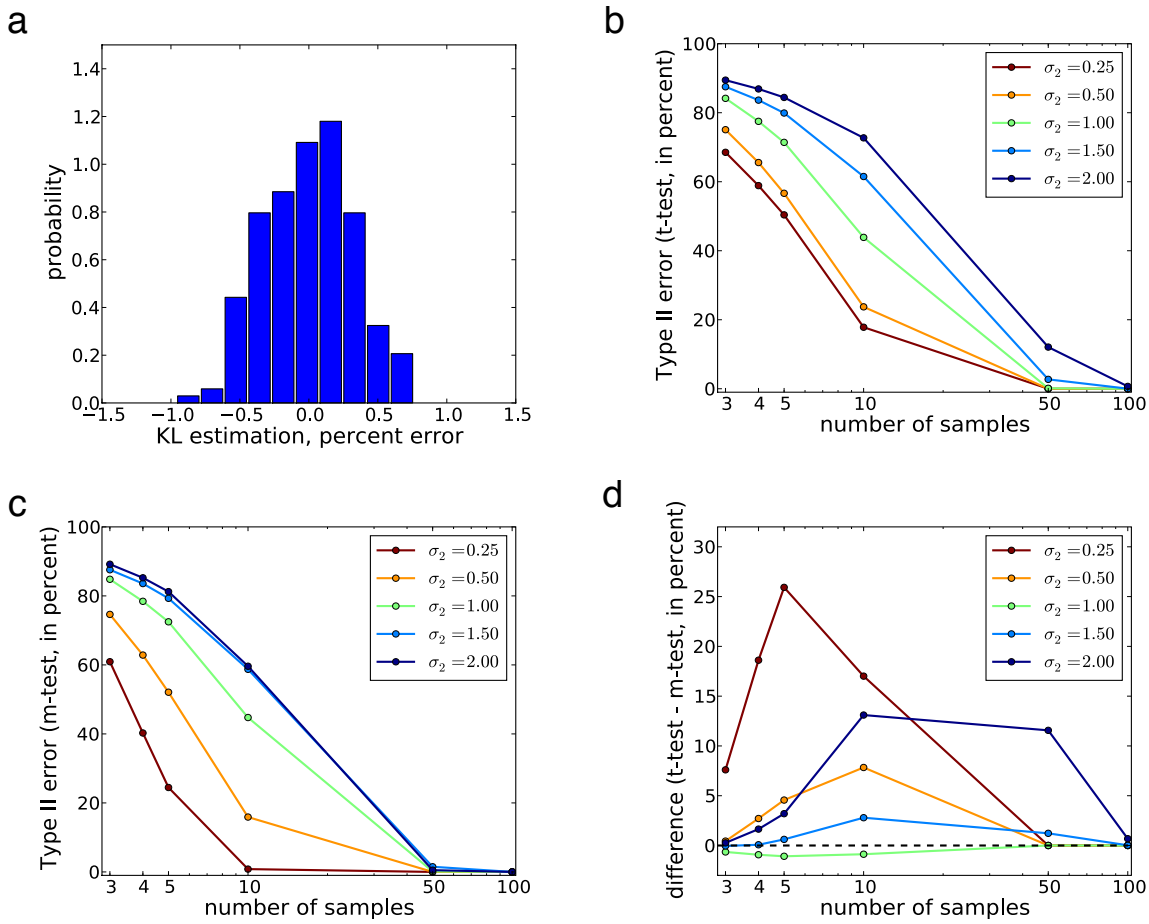

Fig. S4: Kullback-Leibler divergence estimation and m-test validation. **a**, Percent error for the estimation of the KL divergence between two high-dimensional, random distributions. For each run, we drew two distributions over 2^{16} states from a uniform Dirichlet prior, and measured their true KL divergence. From each of the two distributions we then drew $750,000$ samples, from which we estimated the KL divergence (see Methods). The data shown in the figure is taken over 197 runs. **b–d** Type II error was measured for t-test and model selection test (m-test) when the probability of Type I error was fixed at 0.05. The error was computed over $50,000$ Monte Carlo simulations where a given number of samples (x-axis) was drawn from two normal distributions, the first with mean 0 and standard deviation 1, and the second with mean 1 and standard deviation varying from 0.25 to 2.0 (colors from red to blue). **b**, Type II error in percent for the t-test. **c**, Type II error in percent for the model selection test. **d**, Difference in Type II error between t-test and model selection test. Positive numbers indicate lower error (greater statistical power) for the m-test.

Supporting Tables

age groups	N
P29-30	3
P44-45	3
P83-92	4
P129-151	6
TOTAL	16

Movie	Noise	Grating	Dark
3	3	2	3
3	3	2	3
3	4	1	4
3	5	4	6
15	15	9	16

Table S1: Number of animals used in the study. Number of animals per age group (left), and number of animals recorded in each condition (right).

spontaneous activity	evoked activity: stimulus-dependent $P(\mathbf{r}\|\mathbf{y})$	evoked activity: stimulus-averaged $\int P(\mathbf{r}\|\mathbf{y})P^*(\mathbf{y})\,d\mathbf{y}$
ignored	*"classical" neural coding* (e.g., *Rieke et al., 1997*) information transmission	*Schneidman et al., 2006* error correction
analyzed $P(\mathbf{r})$	*Luczak et al., 2009* coding robustness	*this paper* probabilistic inference

Table S2: Approaches to neural activity analysis. Our approach assesses the statistical optimality of an internal model for probabilistic inference and therefore it is conceptually different from standard approaches in neural coding that focus on the optimality of information transmission in a neural circuit by quantifying how easily a stimulus, $\mathbf{y}$ can be recovered from the evoked neural responses, $\mathbf{r}$ (*S33*, *S34*). To provide a statistical description of the mapping between stimuli and EA responses, standard neural coding analyses focus on responses to individual stimuli (left column). These approaches either compute the average stimulus eliciting a particular EA pattern for constructing receptive fields by reverse correlation methods (*S35*), or characterize the distribution of EA patterns in response to each individual stimuli, e.g. by computing its average for constructing tuning curves (*S36*) or by computing higher order moments for characterizing the roles of "noise" correlations in multi-neural responses (*S37*). In contrast, for assessing the statistical optimality of an internal model, the average distribution of EA patterns needs to be computed, where the average is taken over many individual stimuli sampled from a particular stimulus ensemble, $P^*(\mathbf{y})$ (Fig. 1 in the main text). In addition, this distribution then needs to be compared to the distribution of SA patterns (lower right). While averaged EA distributions have been recently analysed to study the error correcting properties of EA patterns (*S38*, upper right), SA has typically been excluded from such analyses. Recent analyses of SA have been limited to comparing patterns of neural activity recorded during spontaneous activity with those evoked by particular stimuli, finding similar structure and repeating motives (*S39*, *S40*, lower left), but see also (*S1*).

Supporting references

S1. Fiser, J., Chiu, C. & Weliky, M. Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature* **431**, 573–578 (2004).

S2. Chiu, C. & Weliky, M. Spontaneous activity in developing ferret visual cortex in vivo. *Journal of Neuroscience* **21**, 8906–8914 (2001).

S3. Cover, T. & Thomas, J. *Elements of information theory* (Wiley-Interscience, Hoboken, NJ, 2006).

S4. MacKay, D.J.C. *Information theory, inference, and learning algorithms* (Cambridge University Press, 2003).

S5. Panzeri, S., Senatore, R., Montemurro, M.A. & Petersen, R.S. Correcting for the sampling bias problem in spike train information measures. *Journal of Neurophysiology* **98**, 1064–72 (2007).

S6. Treves, A. & Panzeri, S. The upward bias in measures of information derived from limited data samples. *Neural Computation* **7**, 399–507 (1995).

S7. Vinje, W. & Gallant, J. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).

S8. Földiák, P. & Endres, D. Sparse coding. *Scholarpedia* **3**, 2984 (2008).

S9. Willmore, B. & Tolhurst, D. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems* **12**, 255–270 (2001).

S10. Berkes, P., White, B.L. & Fiser, J. No evidence for active sparsification in the visual cortex. in *Advances in Neural Information Processing Systems, 22* (eds. Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I. & Culotta, A.) 108–116 (MIT Press, Cambridge, MA, 2009).

S11. Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Science* **14**, 119–130 (2010).

S12. Schwartz, O. & Simoncelli, E. Natural signal statistics and sensory gain control. *Nature Neuroscience* **4**, 819–25 (2001).

S13. Teh, Y.W., Welling, M., Osindero, S. & Hinton, G.E. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research* **4**, 1235–60 (2003).

S14. Olshausen, B.A. & Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).

S15. Bell, A.J. & Sejnowski, T.J. The "independent components" of natural scenes are edge filters. *Vision Research* **37**, 3327–38 (1997).

S16. Rehn, M. & Sommer, F.T. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *Journal of Computational Neuroscience* **22**, 135–146 (2007).

S17. Karklin, Y. & Lewicki, M.S. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* **457**, 83–87 (2009).

S18. Lücke, J., Turner, R., Sahani, M. & Henniges, M. Occlusive components analysis. in *Advances in Neural Information Processing Systems, 22* , 1069–1077, (2009).

S19. Körding, K.P., Kayser, C., Einhäuser, W. & König, P. How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology* **91**, 206–12 (2004).

S20. Berkes, P. & Wiskott, L. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision* **5**, 579–602 (2005).

S21. Lehky, S.R., Sejnowski, T.J. & Desimone, R. Selectivity and sparseness in the responses of striate complex cells. *Vision Research* **45**, 57–73 (2005).

S22. Tolhurst, D.J., Smyth, D. & Thompson, I.D. The sparseness of neuronal responses in ferret primary visual cortex. *Journal of Neuroscience* **29**, 2355–70 (2009).

S23. Turner, R., Berkes, P. & Sahani, M. Two problems with variational Expectation Maximisation for time-series models. in *Proceedings of the Inference and Estimation in Probabilistic Time-Series Models Workshop, Cambridge* , (2008).

S24. Allman, J.M., F, M. & E, M. Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review of Neuroscience* **8**, 407–430 (1985).

S25. Schwartz, O., Pillow, J.W., Rust, C.N. & Simoncelli, E.P. Spike-triggered neural characterization. *Journal of Vision* **6**, 484–507 (2006).

S26. Olshausen, B.A. & Fields, D.J. How close are we to understanding V1? *Neural Computation* **17**, 1665–1699 (2005).

S27. Rao, R.P. & Ballard, D.H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2**, 79–87 (1999).

S28. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference* (Morgan Kaufmann Publishers, 1988).

S29. Adelson, E. Perceptual organization and the judgment of brightness. *Science* **262**, 2042–2044 (1993).

S30. Rossi, A., Rittenhouse, C. & Paradiso, M. The representation of brightness in primary visual cortex. *Science* **273**, 1104–1107 (1996).

S31. Churchland, *et al.*. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience* **13**, 369–78 (2010).

S32. Berkes, P., Turner, R. & Sahani, M. A structured model of video reproduces primary visual cortical organisation. *PLoS Computational Biology* **5**, e1000495 (2009).

S33. Rieke, F., Warland, D., Steveninck, R.D.R.V. & Bialek, W. *Spikes: exploring the neural code* (MIT Press, Cambridge, MA, 1997).

S34. Dayan, P. & Abbott, L.F. *Theoretical Neuroscience* (MIT Press, Cambridge, MA, 1999).

S35. Marmarelis, P. & Marmarelis, V. *Analysis of physiological systems: The white-noise approach* (Plenum Press, 1978).

S36. Hubel, D. & Wiesel, T. Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology* **148**, 574–591 (1959).

S37. Montani, F., Kohn, A., Smith, M. & Schultz, S. The role of correlations in direction and contrast coding in the primary visual cortex. *Journal of Neuroscience* **27**, 2338–2348 (2007).

S38. Schneidman, E., Berry, M.J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007–12 (2006).

S39. Tsodyks, M., Kenet, T., Grinvald, A. & Arieli, A. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science* **286**, 1943–1946 (1999).

S40. Luczak, A., Barthó, P. & Harris, K.D. Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron* **62**, 413–25 (2009).