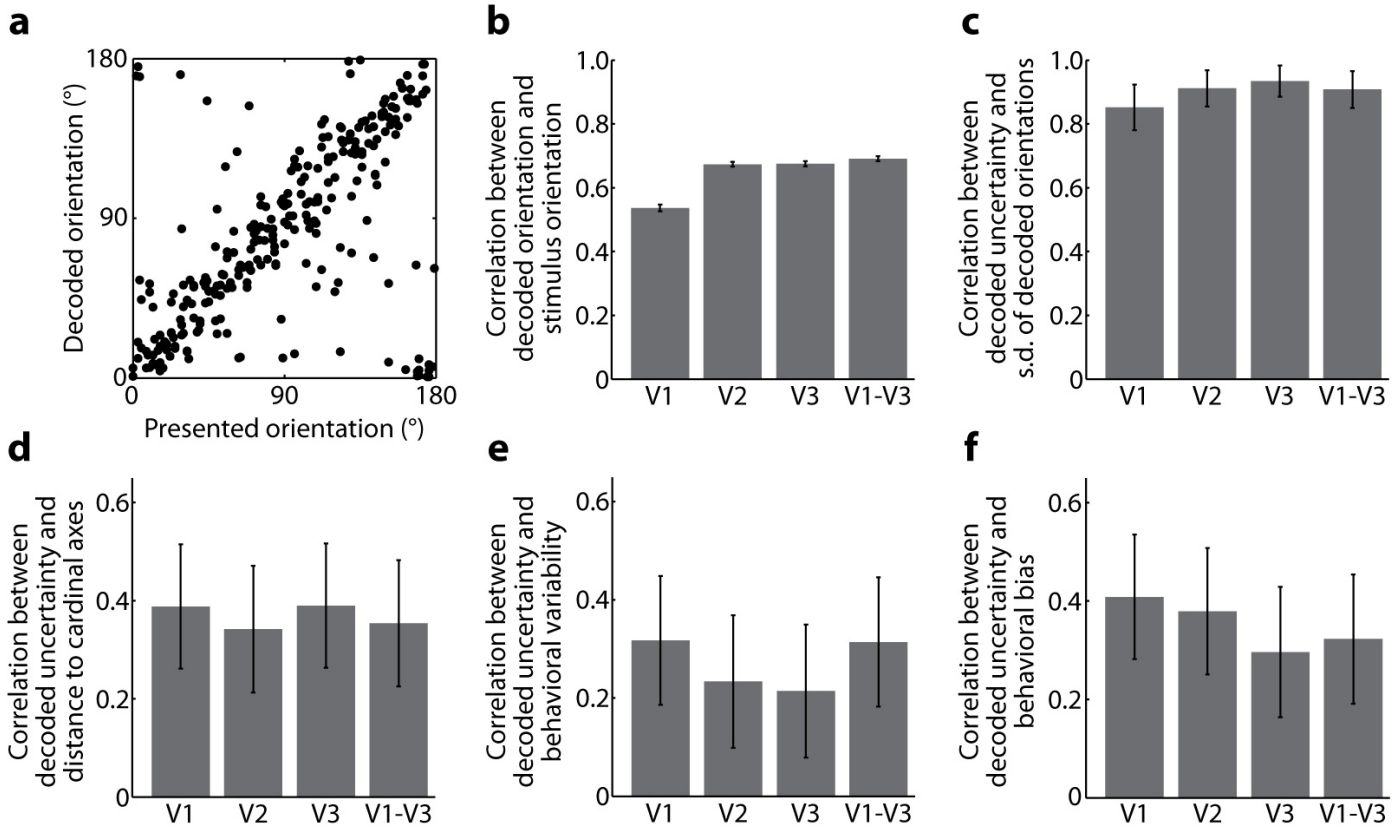
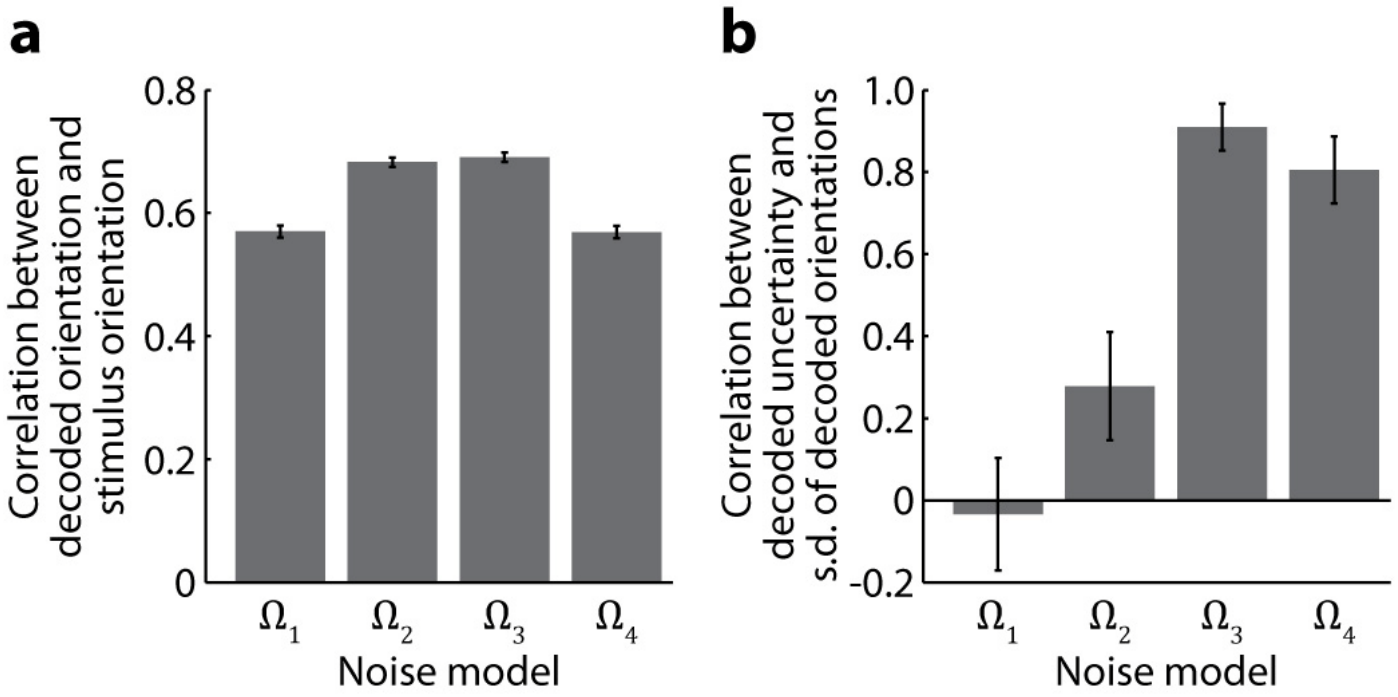


Supplementary Figure 1

Decoding results broken down for different ROIs

Decoding results for areas V1, V2, V3, and V1–V3 combined. (**a**) Decoded and presented orientations are strongly correlated in areas V1–V3 from a representative observer ($r = 0.74$, $p \approx 0$). (**b**) Similar results were found across observers. That is, decoding performance, plotted as the across-subject mean of the correlation between actual and decoded orientations, was highly significant in areas V1, V2, V3, and V1–V3 combined (all $p \approx 0$). (**c**) Posterior width reliably predicted the variability in decoded orientations in all visual ROIs (V1: $p = 0.2\times10^{-15}$; for each of V2, V3 and V1–V3 combined: $p \approx 0$). (**d**) The posterior distribution is reliably broader for more oblique orientations in areas V1 ($p = 0.003$), V2 ($p = 0.011$), V3 ($p = 0.003$) and V1–V3 combined ($p = 0.008$, see also Fig. 1b, main text). (**e**) Posterior width reliably predicts the trial-by-trial variability in behavioral orientation estimates in visual areas V1 ($p = 0.003$) and V1–V3 combined ($p = 0.021$; see also Fig. 1c, main text), with a trend towards significance in area V2 ($p = 0.09$). (**f**) The strength of the behavioral bias away from the cardinal axes is significantly correlated with posterior width in areas V1 ($p = 0.002$), V2 ($p = 0.005$), V3 ($p = 0.030$) and V1–V3 combined ($p = 0.017$, see also Fig. 1d, main text), with smaller behavioral biases for decreasing decoded uncertainty. In all plots, error bars represent +/– 1 SE.



Supplementary Figure 2

Comparison of different noise models

Comparison across different noise models. We considered four different covariance structures for the model described by equation 6. The first covariance structure was defined as:

$$\mathbf{\Omega}_1 = \mathbf{I} \circ \boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}}$$

where $\boldsymbol{\tau}$ is a vector that models the standard deviation of each voxel's Gaussian variability. The second covariance structure was specified by:

$$\mathbf{\Omega}_2 = \rho\boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} + (1-\rho)\mathbf{I} \circ \boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}}$$

where ρ models variability shared globally across voxels, irrespective of their tuning preference. The third noise structure that we considered was defined as:

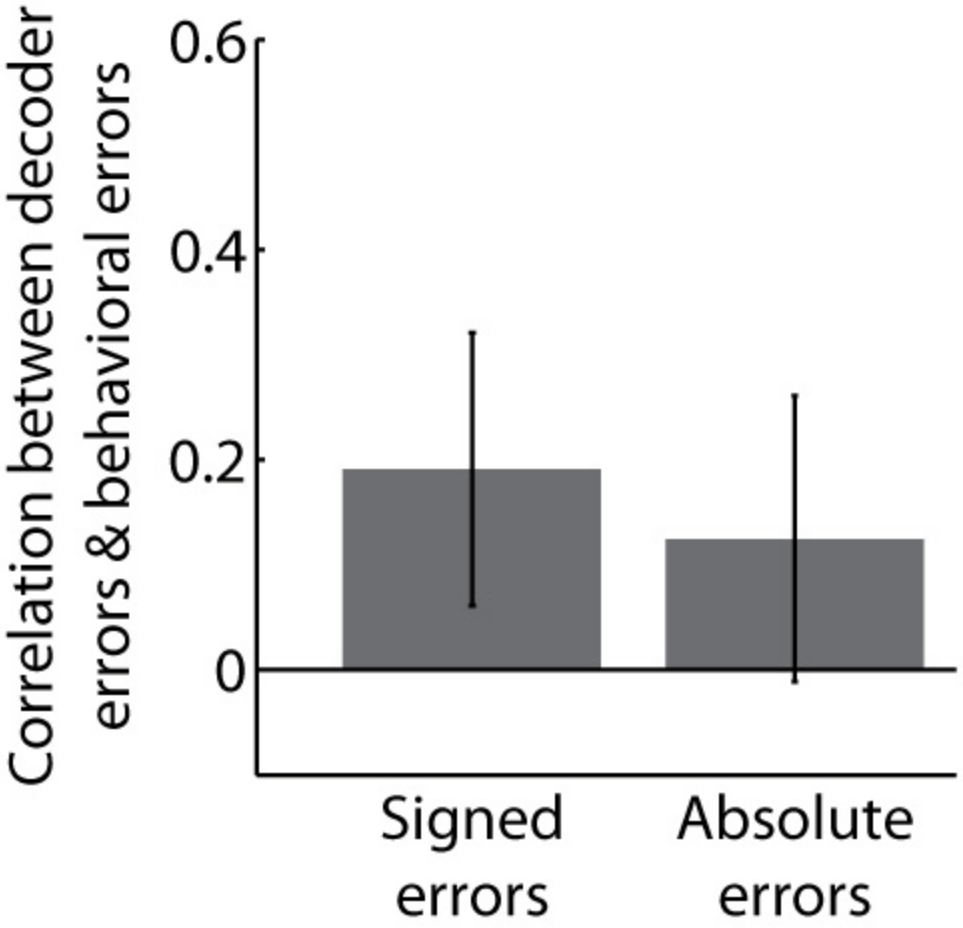
$$\mathbf{\Omega}_3 = \rho\boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} + (1-\rho)\mathbf{I} \circ \boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} + \sigma^2\mathbf{W}\mathbf{W}^{\mathrm{T}}$$

where σ^2 specifies the variance of independent and identically normally distributed noise shared across neural populations of similar orientation preference. The final noise model additionally described local correlations between voxels due to the BOLD point spread function (PSF; see e.g. Parkes et al., *Magn. Reson. Med.*, 2005):

$$\mathbf{\Omega}_4 = \alpha \exp(-\beta\mathbf{D}) \circ \boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} + \rho\boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} + (1-\rho-a)\mathbf{I} \circ \boldsymbol{\tau}\boldsymbol{\tau}^{\mathrm{T}} + \sigma^2\mathbf{W}\mathbf{W}^{\mathrm{T}}$$

This noise model assumed that the degree of shared variability due to the PSF decays exponentially with distance, with initial amplitude α and decay rate controlled by β, and where matrix **D** describes, for each pair of voxels, the absolute distance in millimeters between their center coordinates.

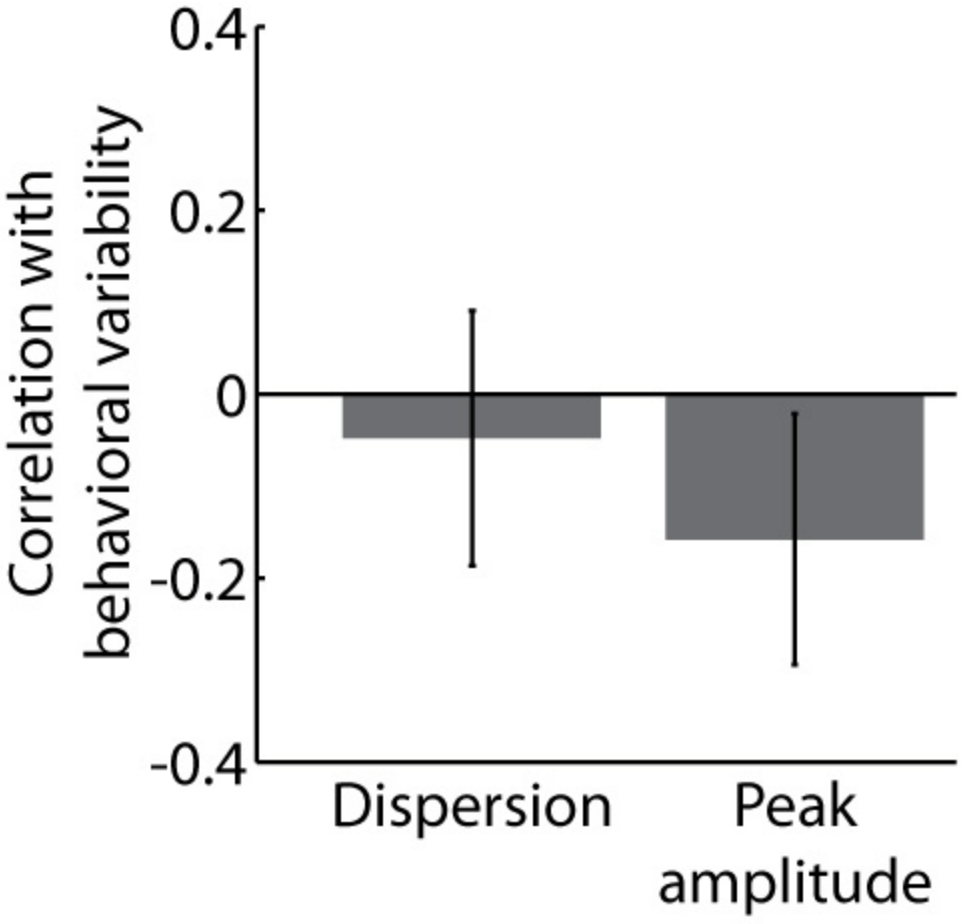
Model parameters were estimated using the fMRI data in a leave-one-run-out cross-validation procedure, using the two-step training procedure as described in Methods. After fitting each model to the training data set, we used the held-out testing data set to evaluate its performance on two relevant benchmark tests. (**a**) We first focused on each model's ability to identify the presented orientation from BOLD activity in areas V1–V3. Orientation decoding performance, quantified as the circular correlation coefficient between the decoded and presented stimulus orientation, was significantly better for models 2 and 3 than for models 1 and 4 (pairwise Z-tests; 2 vs. 1: $Z = 9.11$, $p \approx 0$; 2 vs. 4: $Z = 9.18$, $p = 1.2\times10^{-10}$; 3 vs. 1: $Z = 9.85$, $p \approx 0$; 3 vs. 4: $Z = 9.92$, $p = 3.4\times10^{-12}$). We found no reliable difference in orientation decoding performance between models 2 and 3 ($Z = 0.74$, $p = 0.46$). (**b**) We then evaluated each model's ability to characterize the degree of uncertainty about orientation. To the extent that the decoded posterior distribution appropriately models the covariance in fMRI data, broader distributions should be linked to increased variability in the decoder's estimates of the presented orientation. Accordingly, we divided each participant's data into four bins of increasing posterior width, calculated the across-trial variability in the decoder's orientation estimates for each of the bins, and computed the partial correlation coefficient between mean posterior width and variability in decoded orientations (while regressing out between-subject variability). Interestingly, only models 2–4 successfully characterized a sufficient degree of uncertainty in the fMRI data (model 1: $t(53) = -0.25$, $p = 0.81$, model 2: $t(53) = 2.11$, $p = 0.04$; model 3: $t(53) = 15.91$, $p \approx 0$; model 4: $t(53) = 9.85$, $p = 1.4\times10^{-13}$). Furthermore, model 3 reliably outperformed models 2 and 4 on this test (pairwise Z-tests on correlation coefficients; $Z = 6.31$, $p = 2.8\times10^{-10}$ and $Z = 2.10$, $p = 0.035$, respectively). Taking both benchmark tests together, these results indicate that the third model best captured the noise covariance in BOLD activity relevant to orientation decoding.



Supplementary Figure 3

Correlation between decoder errors and behavioral errors

Correlation between errors in decoded orientation estimates and errors in behavioral orientation reports, with error bars corresponding to +/– 1 SE. For each participant, trials were sorted into four bins of increasing signed or absolute decoder error. Within each bin, we calculated both the mean error (signed or absolute) in decoded orientation and the mean behavioral error (signed or absolute). We then used a multiple linear regression analysis to compute partial correlations between decoder and behavioral errors, controlling for mean differences between observers. Signed decoder errors were not significantly correlated with signed behavioral errors ($t(53) = 1.42$, $p = 0.16$), nor were larger decoder errors reliably associated with larger behavioral errors ($t(53) = 0.91$, $p = 0.12$) in areas V1–V3. Why do we nevertheless find that the variance of the posterior distribution is linked to behavioral biases and across-trial variability in behavioral errors? To see why, consider that the errors themselves are two independent random variables. As such, the correlation between the errors must necessarily be relatively weak, even when the mean and variance of their underlying distributions are linked. This observation exemplifies the utility of our uncertainty metric, which directly reflects the variance of the underlying distributions.



Supplementary Figure 4

Correlation between the estimated neural population response and behavioral variability

Estimating the neural population response. Does a channel-based approach (cf. [13,14]) similarly reflect the degree of uncertainty about orientation? It is important to realize that the posterior probability distribution characterizes the amount of information contained in the pattern of BOLD responses, rather than providing a direct estimate of the neural population response. That said, our model does allow for the estimation of neural population responses at a single trial level. Specifically, the population response **c** is described as the (idealized) tuning curves of the population plus noise (cf. equation 1):

$$\mathbf{c} = \boldsymbol{f}(s) + \boldsymbol{\eta}$$

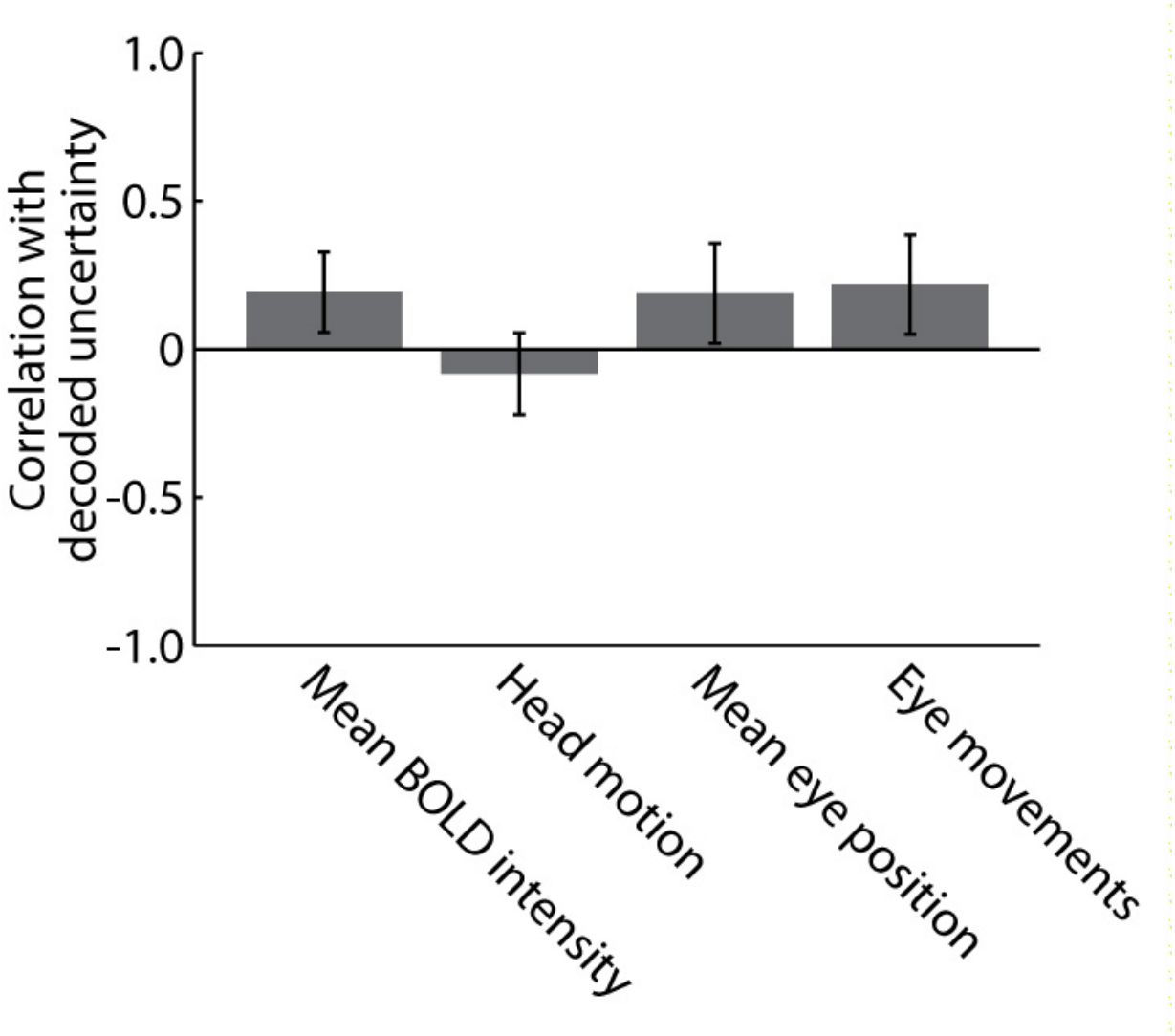
Thus, estimating **c** involves finding the most likely value for $\boldsymbol{\eta}$ by maximizing the joint likelihood $p(\mathbf{b}|s, \boldsymbol{\eta}; \widehat{\mathbf{W}}, \hat{\boldsymbol{\tau}}, \hat{\rho})p(\boldsymbol{\eta}|\hat{\sigma})$. Differentiating this likelihood with respect to $\boldsymbol{\eta}$ gives the following expression for the maximum likelihood estimate (MLE):

$$\hat{\boldsymbol{\eta}} = \hat{\sigma}^2 \widehat{\mathbf{W}}^{\mathrm{T}} \widehat{\boldsymbol{\Omega}}^{-1} \left(\mathbf{b} - \widehat{\mathbf{W}} \boldsymbol{f}(s) \right)$$

With these equations in hand, we first computed for each independent test trial the most likely neural

population response $\hat{\mathbf{c}}$. We then calculated the dispersion (circular standard deviation across channels) of the estimated neural population response, as well as the amplitude of the channel most strongly tuned to the presented stimulus orientation. Trials were subsequently sorted into four bins of increasing dispersion or peak response amplitude value. Summary statistics (mean dispersion, mean peak amplitude, and behavioral variability) were computed across all trials in each bin. Multiple regression analysis was used to compute partial correlation coefficients between dispersion and behavioral variability, as well as between peak amplitude and behavioral variability (regressing out distance to cardinal axes and between-subject variability). Interestingly, neither the dispersion, nor the peak amplitude, of the estimated population response reliably predicted behavioral variability ($r = -0.05$, $p = 0.73$ and $r = -0.16$, $p = 0.25$, respectively; in the figure, error bars indicate +/– 1 SE).

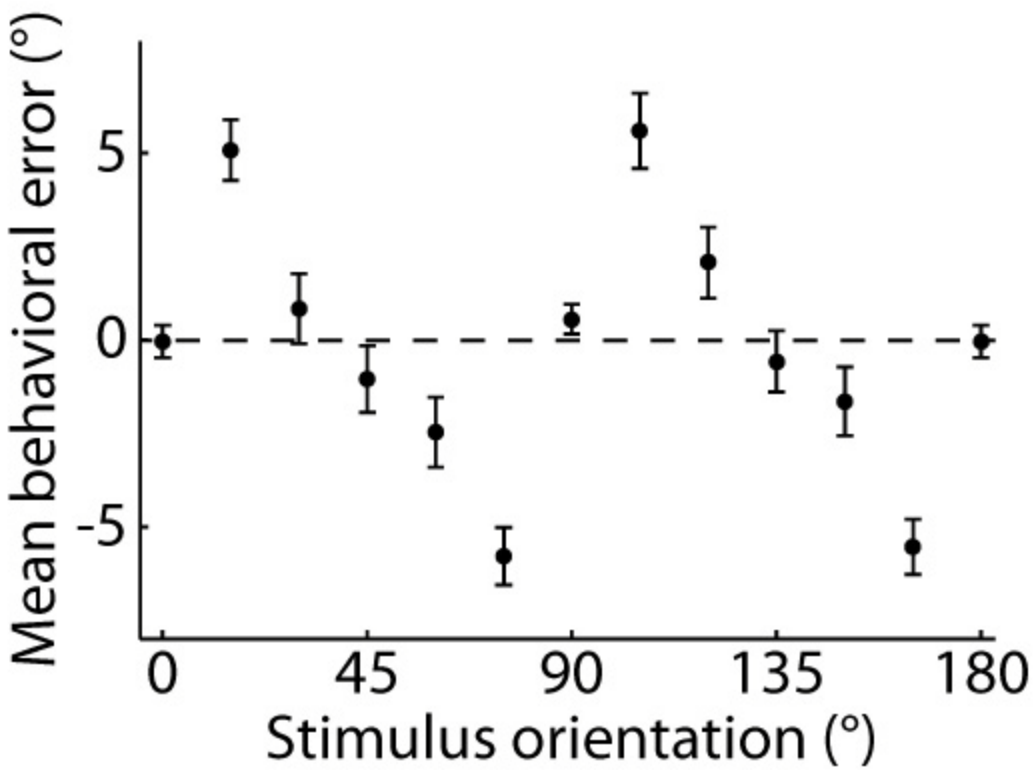
Why is the estimated neural population response a less reliable predictor of behavior than the posterior probability distribution? One reason may be that the posterior distribution combines many aspects of the population response in a single information metric. As such, the posterior distribution is more sensitive to changes in orientation information than any one property of the neural population response alone. It is also important to realize that $\hat{\mathbf{c}}$ reflects the *most likely* neural population response (i.e., a single estimate). BOLD activity, however, is rather noisy and typically consistent with a *whole range* of neural population responses. The posterior distribution explicitly reflects this full range of possibilities, thereby providing greater sensitivity to subtle changes in BOLD activity.



Supplementary Figure 5

Control analyses for mean BOLD, head motion and eye movements

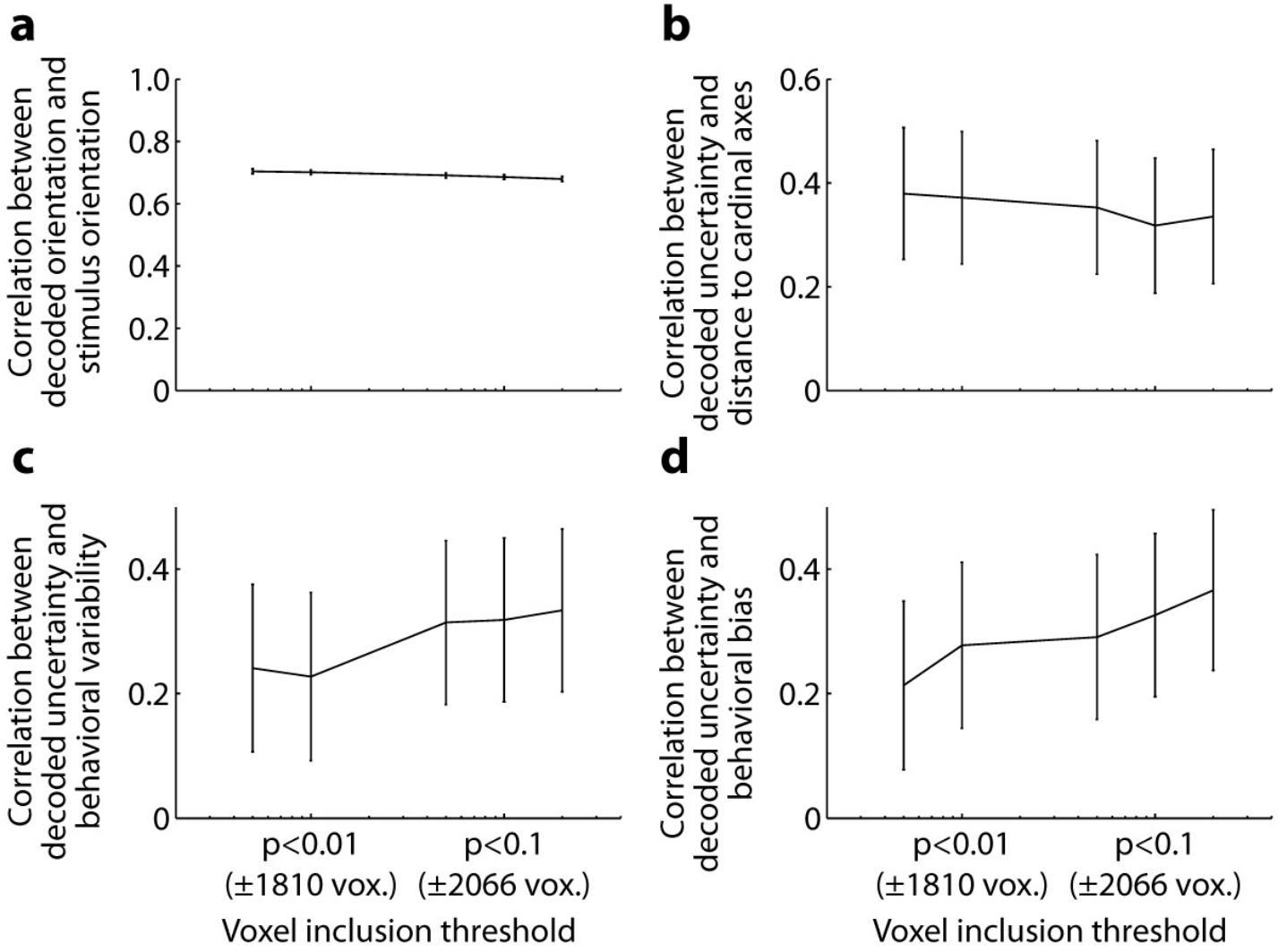
Control analyses. Partial correlation coefficients between decoded uncertainty and mean BOLD signal intensity, head motion, eye position and eye movements. We found no significant correlation between decoded uncertainty and any of these variables in areas V1–V3 ($p = 0.16$, $p = 0.55$, $p = 0.27$ and $p = 0.20$, respectively), indicating that gross BOLD amplitude, mean eye position (amount of) eye movements, and (amount of) subject head motion cannot account for the trial-by-trial uncertainty in cortical stimulus representations. This furthermore rules out simple explanations in terms of the amount of attentional effort put into the task, as overall BOLD amplitude tends to increase with effort (Ress, Backus & Heeger, *Nat Neurosci.*, 2000). Error bars represent +/– 1 SE.



Supplementary Figure 6

Bias in orientation reports as a function of stimulus orientation

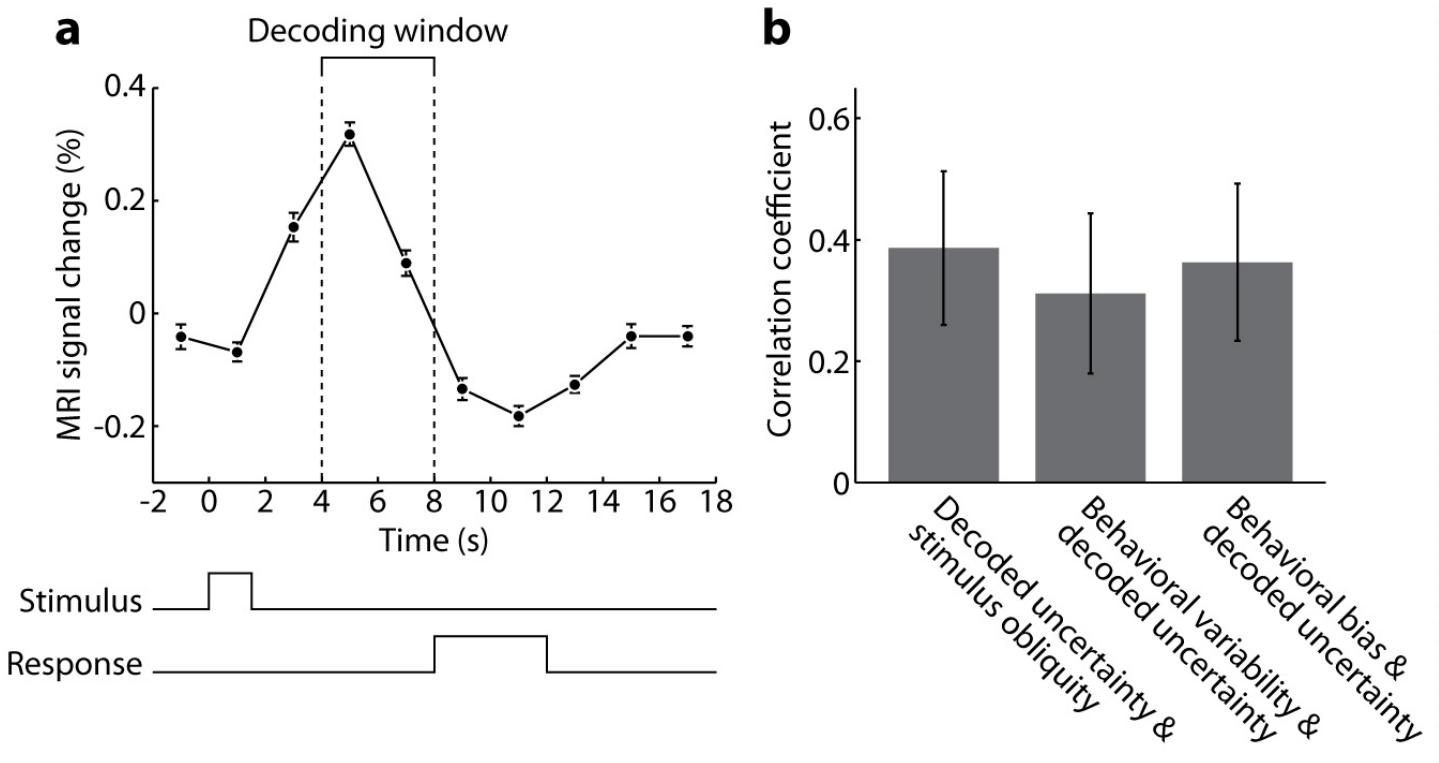
Behavioral orientation reports are biased away from the nearest cardinal axis. Plots show the average (signed) error across observers in the behavioral orientation judgments, as a function of stimulus orientation. Positive errors indicate clockwise deviations from the veridical stimulus orientation. For each observer, trials were binned based on stimulus orientation, and the average behavioral error was calculated within each bin. Error bars indicate +/– 1 SEM.



Supplementary Figure 7

Effect of number of voxels on main results

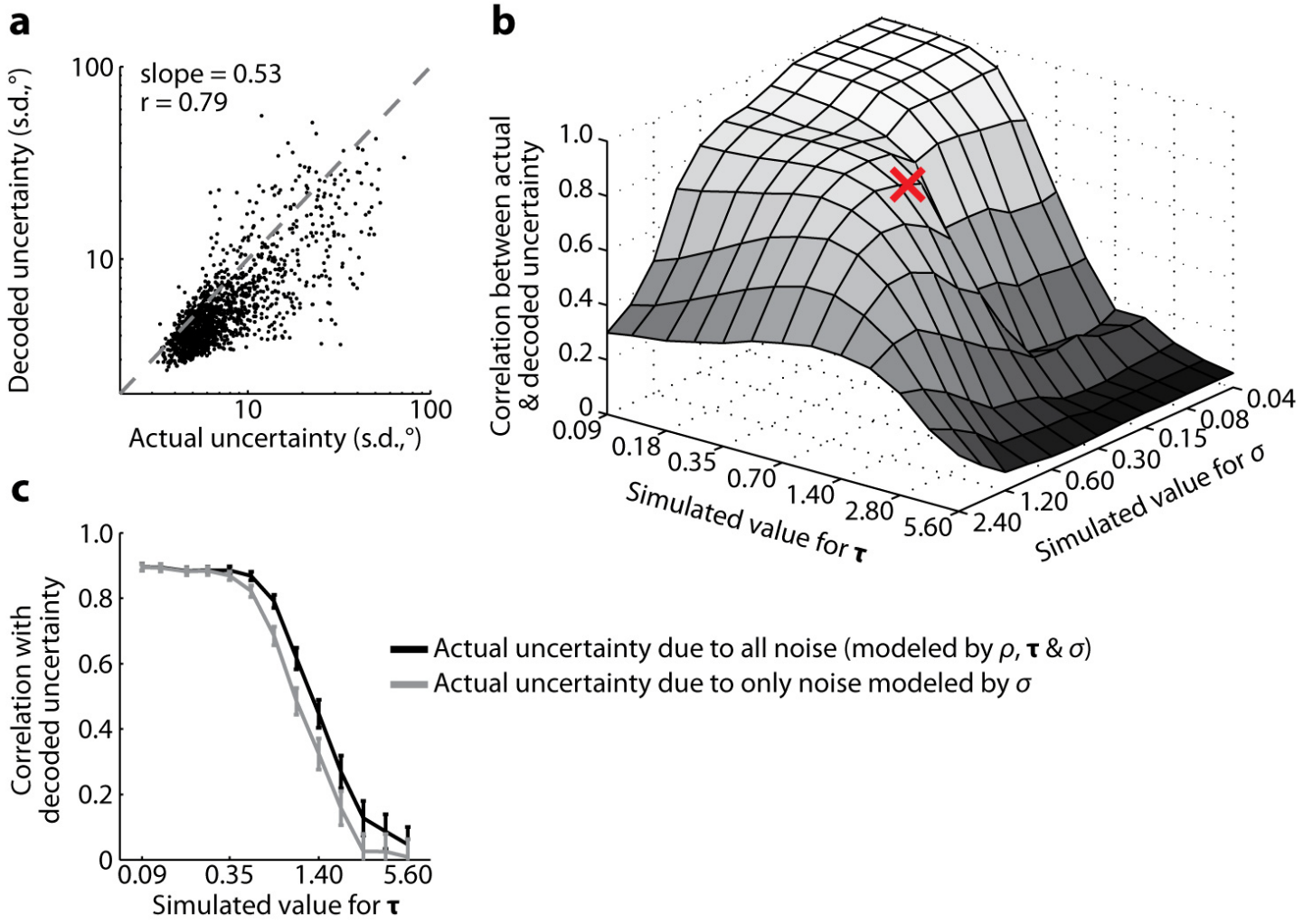
Decoding results as a function of number of voxels included within the ROI. Within areas V1–V3, all voxels whose response to the localizer stimulus met a chosen threshold (x-axis, uncorrected p-values) were selected for subsequent analysis. Shown is the correlation between (**a**) decoded and actual stimulus orientation (decoding performance), (**b**) decoded uncertainty and angle between stimulus orientation and nearest cardinal axis (oblique effect in decoded uncertainty), (**c**) decoded uncertainty and behavioral variability and (**d**) decoded uncertainty and behavioral bias. Error bars represent +/– 1 SE. All of these results proved robust to reasonable variations in the number of voxels included in the analysis.



Supplementary Figure 8

Hemodynamic response function and decoding window

Hemodynamic response function and decoding window. (**a**) Time course of mean BOLD activity in areas V1–V3 over the course of a trial. Time points between 4–8 s were averaged for subsequent decoding analysis. This relatively short time window (4 s) was chosen in order to ensure that activity from the response window was excluded from analysis. (**b**) Temporally expanding the time window to 2–8s did not greatly affect any of our results ($p = 0.004$, $p = 0.022$ and $p = 0.007$, respectively). Error bars correspond to +/– 1 SE.

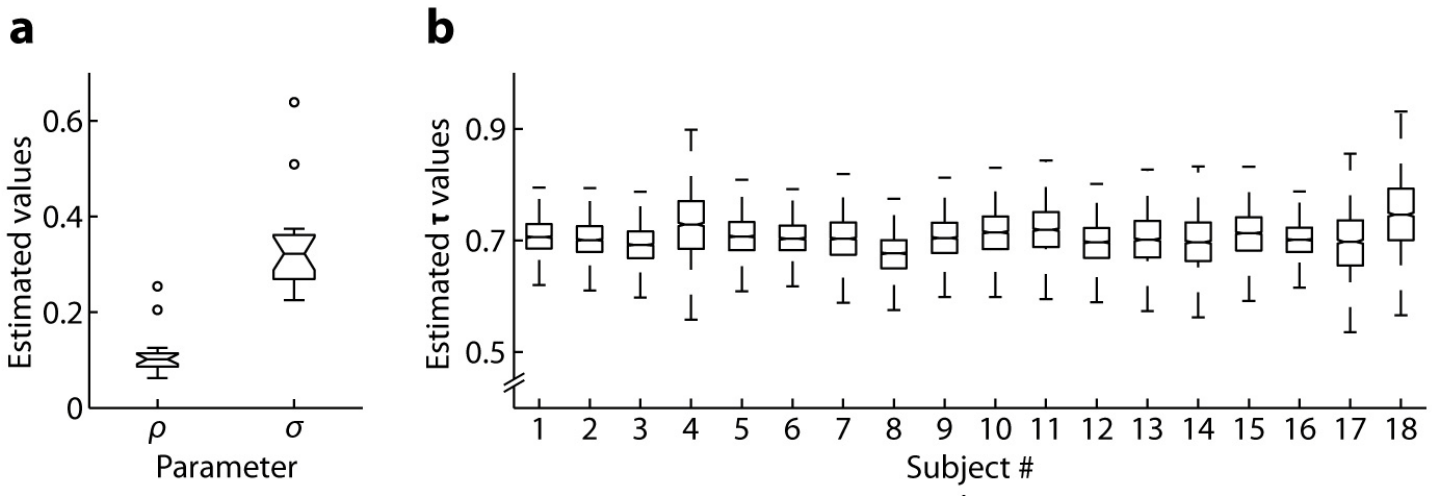


Supplementary Figure 9

Simulation results supporting the two-step model parameter estimation procedure

Estimation of **W** conditioned on the assumption that $\sigma = 0$. We simulated data according to the generative model described by equation 6, with $\rho = 0.05$, $\sigma = 0.3$, and $\boldsymbol{\tau} \sim \mathcal{N}(0.7, 0.035^2)$ for five different (hypothetical) observers. These parameter values were based on actual values estimated from 18 real observers. Similarly, **W** was simulated (independently for each observer) by sampling at random from a weight matrix that was estimated from the data of one of the participants to ensure a realistic distribution of orientation tuning across simulated voxels. Each simulated pattern of BOLD activity contained information about stimulus orientation, with varying degrees of uncertainty due to noise. To assess trial-by-trial uncertainty, we computed, for each trial of simulated data, the probability distribution over stimulus orientation given the parameters of the generative model (i.e. $p(s|\mathbf{b}; \mathbf{W}, \rho, \boldsymbol{\tau}, \sigma)$). We took the circular standard deviation of this distribution as the actual degree of uncertainty in the data. We then asked how well our decoder would recover this uncertainty, having estimated the model parameters from the simulated BOLD patterns using the two-step procedure described in Methods. Panel (**a**) plots decoded uncertainty against the actual degree of uncertainty in the data. While the decoder is slightly biased towards smaller values of uncertainty (slope = 0.53), the correlation between actual and decoded uncertainty is reasonable ($r = 0.79$, $p = 6.0 \times 10^{-15}$). (**b**) The simulations were repeated for a range of parameter values (σ and $\boldsymbol{\tau}$), in semi-octave steps around the empirical values used in **a** (red cross).

Decoded uncertainty becomes less accurate with larger degrees of noise (larger σ and $\boldsymbol{\tau}$), as more noisy data generally results in poorer parameter estimates. Overall, for realistic levels of noise, decoded uncertainty correlates well with the actual degree of uncertainty in the data. (**c**) To assess whether our approach is sensitive to fluctuations in neural response in particular, we additionally computed the uncertainty about orientation in the simulated data given only neural variability modeled by σ. Specifically, we computed for each trial of simulated data the posterior probability distribution $p(s|\mathbf{c};\sigma)$, where $\mathbf{c} = \boldsymbol{f}(s) + \boldsymbol{\eta}$ (cf. equation 1) and $\sigma = 0.3$. The circular standard deviation of this distribution served as the actual degree of uncertainty given the neural response alone. Interestingly, the correlation between decoded uncertainty and actual uncertainty reduced only slightly, suggesting that $\boldsymbol{\tau}$ and ρ contribute little to overall uncertainty. To see why $\boldsymbol{\tau}$ and ρ have a relatively small influence on decoded uncertainty, consider that the noise modeled by sigma is positively correlated with changes in signal (tuning curves) – such signal-dependent noise has the greatest impact on information (Smith & Kohn, *J. Neurosci.*, 2008; Abbot & Dayan, *Neural Comput.*, 1999; Averbeck, Latham & Pouget, *Nat. Rev. Neurosci.*, 2006). In contrast, ρ is unrelated to voxel tuning, while the sources of noise modeled by $\boldsymbol{\tau}$ can, to large extent, be averaged out across voxels. Altogether, these simulations confirm that our two-step parameter estimation approach captures a sufficient degree of uncertainty in the data.



Supplementary Figure 10

Noise model parameter estimates

Noise model parameter estimates obtained from the fMRI data. (**a**) Distributions across subjects of the mean estimated values for ρ and σ (averaged across all training partitions of the data), shown as box plots. Boxes extend from the first to the third quartiles, with notches at the medians. Whiskers span the full range of the data except for outliers, which were defined as values deviating from the median by more than 1.5 times the inter-quartile range, and are shown separately as open circles. (**b**) Since $\boldsymbol{\tau}$ contained a value for each voxel, we plot its estimates separately, showing the distribution of values across voxels within each participant. For clarity of exposition, no outliers are shown here (due to the large number of data). Otherwise, this panel follows the same conventions as in **a**.

Supplementary Table 1: Clarification of variables

M	number of voxels
K	number of hypothetical neural populations
i, k	indexes of voxels and neural populations, respectively
$\mathbf{b}$	an M x 1 vector $\{b_i\}$ of voxel responses
$\mathbf{W}$	an M x K matrix $\{W_{ik}\}$, containing for each voxel i the contribution of each neural population k to that voxel's orientation tuning function
s	stimulus orientation in degrees
$f_k(s)$	the orientation tuning function of the k^{th} neural population
η_k	noise in the response of the k^{th} neural population
$\mathbf{c}$	the neural population response; a K x 1 vector such that $\mathbf{c} = \boldsymbol{f}(s) + \boldsymbol{\eta}$
ν_i	noise in the response of the i^{th} voxel
$\boldsymbol{\Sigma}$	an M x M covariance matrix for the multivariate normal distribution of $\boldsymbol{\nu}$
$\boldsymbol{\Omega}$	an M x M covariance matrix for the multivariate normal distribution of $(\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\eta})$
$\boldsymbol{\tau}$	an M x 1 vector, where τ_i^2 is the marginal variance of ν_i
ρ	global noise correlation between all voxels
σ^2	variance of independent noise in neural populations tuned to the same orientation
α, β	amplitude and space constant (respectively) of an exponential decay function specifying correlations between neighboring voxels as a function of spatial separation (see Supplementary Fig. 2)

Nature Neuroscience: doi:10.1038/nn.4150