Biases in multivariate neural population codes

Sander W. Keemink and Mark C.W. van Rossum

February 28, 2017

Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh,10 Crichton Street, Edinburgh EH8 9AB, UK
s.keemink@sms.ed.ac.uk, mvanross@inf.ed.ac.uk

Abstract

Throughout the nervous system information is typically coded in activity distributed over large population of neurons with broad tuning curves. In idealized situations where a single, continuous stimulus is encoded in a homogeneous population code, the value of an encoded stimulus can be read out without bias. Here we find that when multiple stimuli are simultaneously coded in the population, biases in the estimates of the stimuli and strong correlations between estimates can emerge. Although bias produced via this novel mechanism can be reduced by competitive coding and disappears in the complete absence of noise, the bias diminishes only slowly as a function of neural noise level. A Gaussian Process framework allows for accurate calculation of the bias and shows that a bimodal estimate distribution underlies the bias. The results have implications for neural coding and behavioral experiments.

In many brain areas information is distributed across neurons using population codes, in which many neurons respond collectively to a single stimulus. Given its ubiquity, understanding population coding is believed to be crucial to understand coding of information in the brain. By pooling across neurons, population codes allow for accurate estimation of a stimulus from the population response even when neural noise is present. Numerous studies have quantified, among other issues, the role of the tuning curves (Zhang and Sejnowski, 1999), noise-correlations (Sompolinsky et al., 2002; Moreno-Bote et al., 2014), and heterogeneity (Shamir and Sompolinsky, 2006; Ecker et al., 2011; Shamir, 2014) on the coding accuracy.

However, coding accuracy is not the only performance metric. When the same stimulus is repeatedly estimated from a population response and these estimates are averaged over many trials, a systematic difference between the mean estimated value and its true value might remain; this is called bias. In many idealized cases biases are absent from population coding estimation schemes. First, in the limit of low noise, estimators such as the maximum likelihood decoder can be shown to be unbiased (Kay, 1993). Secondly, the coding problem might have an intrinsic symmetry that abolishes bias, that is, over- and underestimation of the stimulus are equally likely - a typical example being the estimation of the orientation of a visual grating from a homogeneous population. Either condition by itself is sufficient to warrant unbiased estimation. For instance, while the maximum likelihood decoder is sub-optimal for high noise, it remains unbiased for one dimensional direction estimations (Xie, 2002).

Yet, in perception biases are common. To explain these, theoretical studies rely on mechanisms that modulate the neural response without adjusting the decoder to break the homogeneity, such as can occur with adaptation (e.g. Stocker and Simoncelli, 2006; Seriès et al., 2009; Cortes et al., 2012) or with contextual changes in the neural tuning (e.g. Schwartz et al., 2007). In contrast to those studies we show that even in homogeneous population codes biases can occur. We consider the case where multiple variables are simultaneously coded in a population, such as occurs in visual cortical area MT when two overlapping transparent random dot motion patterns are presented. We find that in these situations biases in estimation emerge from the decoder itself, even though the decoder has full knowledge of the coding process. Furthermore, when multiple overlapping stimuli are presented, the number of perceived stimuli can be fewer than the number presented, resembling psycho-physical findings (Treue et al., 2000; Edwards and Greenwood, 2005). We develop a mathematical framework based on Gaussian Processes to calculate and understand these effects and discuss their consequences for neural computation and perceptual biases.

Results

To examine the emergence of biases we consider a population of neurons described by their firing rates. The average response of each neuron is given by its tuning curve $f(\boldsymbol{s})$, where $\boldsymbol{s}$ is a vector of stimulus parameters encoded by the neuron. Gaussian white noise ν_i with mean zero and variance σ^2 is added to the response, so that on a given trial the firing rate r_i of neuron i is

$$r_i = f_i(\boldsymbol{s}) + \nu_i. \tag{1}$$

Commonly one studies the case where $\boldsymbol{s}$ is one-dimensional. Here we consider the coding of two stimuli $\boldsymbol{s} = (s_1{,}s_2)$ simultaneously. For concreteness we consider the coding of two overlapping random dot motion patterns in area MT; in this case s_1 and s_2 represent the two motion directions, Fig. 1A. The response of MT neurons to such a stimulus has been modeled by the linear average, or, equivalently for our purposes, the sum of the tuning curves to the individual stimuli (van Wezel et al., 1996; Treue et al., 2000),

$$f_i(\boldsymbol{s}) = g_i(s_1) + g_i(s_2) \tag{2}$$

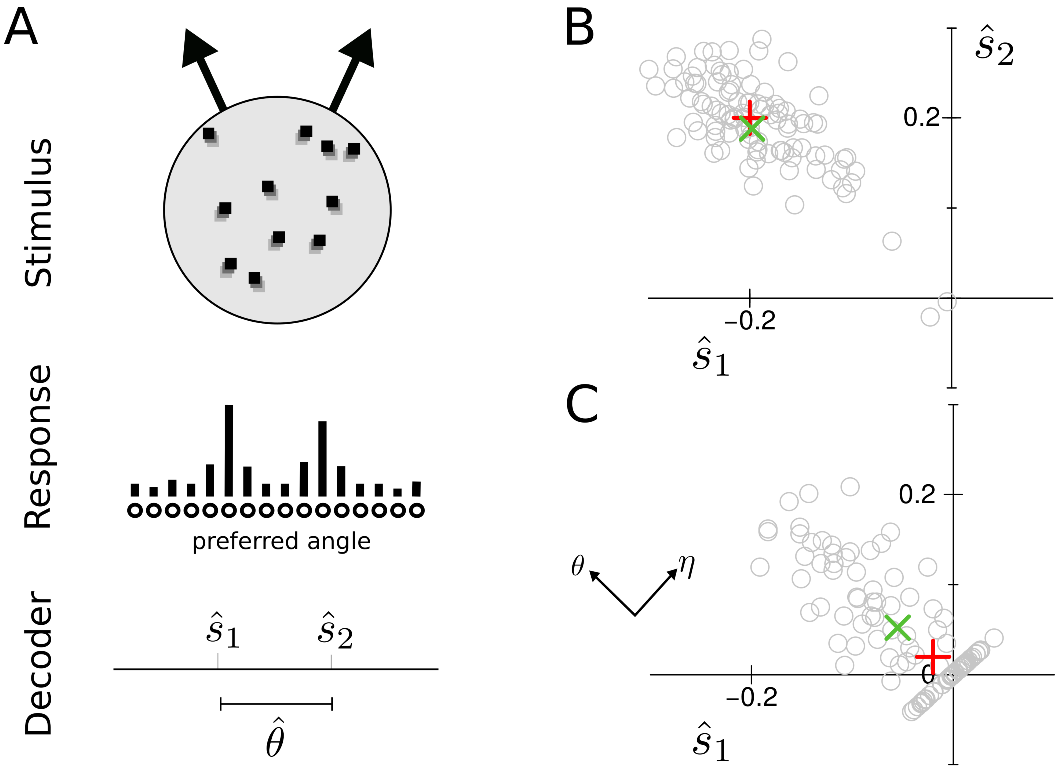

Figure 1: A) Basic encoding-decoding setup. The stimulus consists of two overlapping moving random dot patterns. A population of neurons codes for the two simultaneous stimuli. The task is to estimate the stimulus parameters, here the motion directions $\hat{s}_1$ and $\hat{s}_2$, from the noisy population response. B) Maximum likelihood estimates across a number of trials. For a wide opening angle $\boldsymbol{s} = (-0.2, 0.2)$, the distribution of estimates follows approximately a Gaussian distribution. True stimulus (red plus) and average estimate (green X) overlap. C) For narrow opening angles, $\boldsymbol{s} = (-0.02, 0.02)$, the distribution of estimates falls into two roughly equal parts, a Gaussian-shaped distribution and a distribution along the line $\hat{s}_1 = \hat{s}_2$. True stimulus and average estimate now diverge, i.e. the estimate is biased. The sum and difference angles are indicated by η and θ, respectively. (all angles in radians).

where $g_i(s)$ is the bell-shaped tuning of neuron i to a single stimulus (Methods). More competitive interactions between the responses have also been proposed; these are considered below.

Decoding of the neural response

We draw stochastic responses from the above model (see Methods for details) and then decode the stimulus parameters from the noisy population response using the maximum likelihood (ML) decoder. That is, estimates of the stimulus $\hat{\boldsymbol{s}}$ are obtained by finding the stimulus vector that was most likely given the noisy neural response vector $\boldsymbol{r}$,

$$\hat{\boldsymbol{s}} = \text{argmax}_{\boldsymbol{s}} \log P(\mathbf{r}|\boldsymbol{s}).$$

The hat indicates estimates throughout. Because the encoder loses the identity of the stimuli, we additionally impose that $s_2 \geq s_1$.

We first consider the case when the two peaks in the tuning curve are far apart ($|s_1 - s_2| \gg w$, where w is the tuning width). In this case the stimulus estimates follow a two-dimensional Gaussian distribution centered around the true stimulus value, Fig. 1B. The true stimulus value (cross) and the mean estimate (X) coincide.

However, when the motion directions are instead almost the same so that the peaks in the population response partly overlap, the distribution radically changes shape, Fig. 1C. Now the estimates fall essentially in two categories: Either the estimates are strongly positively correlated, and cluster on the diagonal where $\hat{s}_1 = \hat{s}_2$. In this case the most likely explanation for the neural response is that the two motion directions are the same. Alternatively, on other trials the estimates are negatively correlated, and the angular difference in the motion direction is over-estimated (repulsion). The mean of neither component of the distribution coincides individually with the true stimulus vector, nor does the mean of the full distribution; in other words, the estimate is biased.

To more easily understand these results we transform the coordinates and describe the system in the sum and difference of the angles. The sum of the angles, $\eta = s_1 + s_2$ follows a Gaussian distribution and is unbiased as dictated by the rotational invariance of the setup. More interesting, however, is the opening angle $\Theta = s_2 - s_1$. Estimator bias b is defined as the difference between mean estimate and true stimulus value, $b(\Theta) = \langle \hat{\theta} \rangle - \Theta$, where the angular brackets denote the average over trials and $\hat{\theta}$ are the estimates. The estimator bias is shown as a function of true value Θ in Fig. 2A. When the opening angle Θ is small, the bias is repulsive (the apparent angle is larger than the true value). As the opening angle increases, the bias changes sign and becomes attractive, before reducing to zero for even larger angles, Fig. 2A.

One can wonder whether the repulsive bias is simply caused by imposing $s_2 \geq s_1$. However, the distribution of estimates is unexpectedly bi-modal, with a gap between $\hat{\theta} = 0$ and the secondary peak, Fig. 2C. Furthermore, the change in the sign of the bias is unexpected from such an interpretation. If the ordering of s_1 and s_2 were randomly assigned, the estimate distribution would become tri-modal with some estimates lying on the diagonal, and others clustering in clouds on the anti-diagonal on either side of the origin.

The bias not unique to the use of maximum likelihood decoder and with a Bayesian decoder similar biases emerge. The Bayesian decoder calculates the full distribution of possible stimulus estimates given the response and the noise model, $P_B(\theta|\boldsymbol{r})$. For a flat prior for Θ, this is proportional to $P(\boldsymbol{r}|\theta)$. Whereas the maximum likelihood decoder takes the maximum of this distribution, using a square loss function the Bayesian estimate equals the mean of this distribution, $\hat{\theta}_B = \int \theta P_B(\theta|\boldsymbol{r})\, d\theta$ (Kay, 1993; Salinas and Abbott, 1994). The bias is slightly more pronounced with a Bayesian decoder, Fig S1. However, as the Bayesian decoder does not allow for theoretical treatment, we concentrate on the maximum likelihood decoder.

In summary, in this relatively simple coding problem biphasic biases emerge. Next, we attempt to understand why this occurs.

Emergence of bias

We now analyze the Maximum Likelihood estimator in detail. For independent Gaussian noise the maximum likelihood estimate is equivalent to minimizing the Mean Squared Error (MSE) E between observed and expected response

$$\hat{\boldsymbol{s}} = \mathrm{argmin}_{\boldsymbol{s}} E(\boldsymbol{s})$$

where $E(\boldsymbol{s}) = \sum_{i=1}^{N}[r_i - f_i(\boldsymbol{s})]^2$. The emergence of the bias and the underlying distribution of estimates can be understood from the mean square error that the estimator seeks to minimize. The MSE is a smooth function but its precise shape varies from trial to trial, Fig. 2B. To write the MSE as a Gaussian Process (Williams and Rasmussen, 2006) we first split it up as

$$E(\theta) = E_{\mathrm{mean}}(\theta) + E_{\mathrm{noise}}(\theta) + C,$$

where C a stimulus independent term, and θ denotes the candidate stimulus. The stimulus dependent part consists of two terms: the first term is the mean $E_{\mathrm{mean}}(\theta) = \sum_{i=1}^{N}[f_i(\theta) - f_i(\Theta)]^2$ that is identical across trials and attains its minimal value of zero at the true stimulus value, Θ. The second term is the noise term $E_{\mathrm{noise}}(\theta) = -2\sum_{i=1}^{N} \nu_i f_i(\theta)$.

Of particular interest is the limiting case of $\Theta = 0$. While somewhat contrived as the presented motion directions are identical in that case, exact results can be obtained in this limit that approximately hold for any small Θ. In this limit the term $E_{\mathrm{mean}}(\theta)$ is lowest at $\theta = 0$, as expected, Fig. 2A, black curve. Because of symmetry in the combined tuning curves, Eq. 6, not only all odd derivatives, but also the second derivative of E_{mean} is zero. Thus

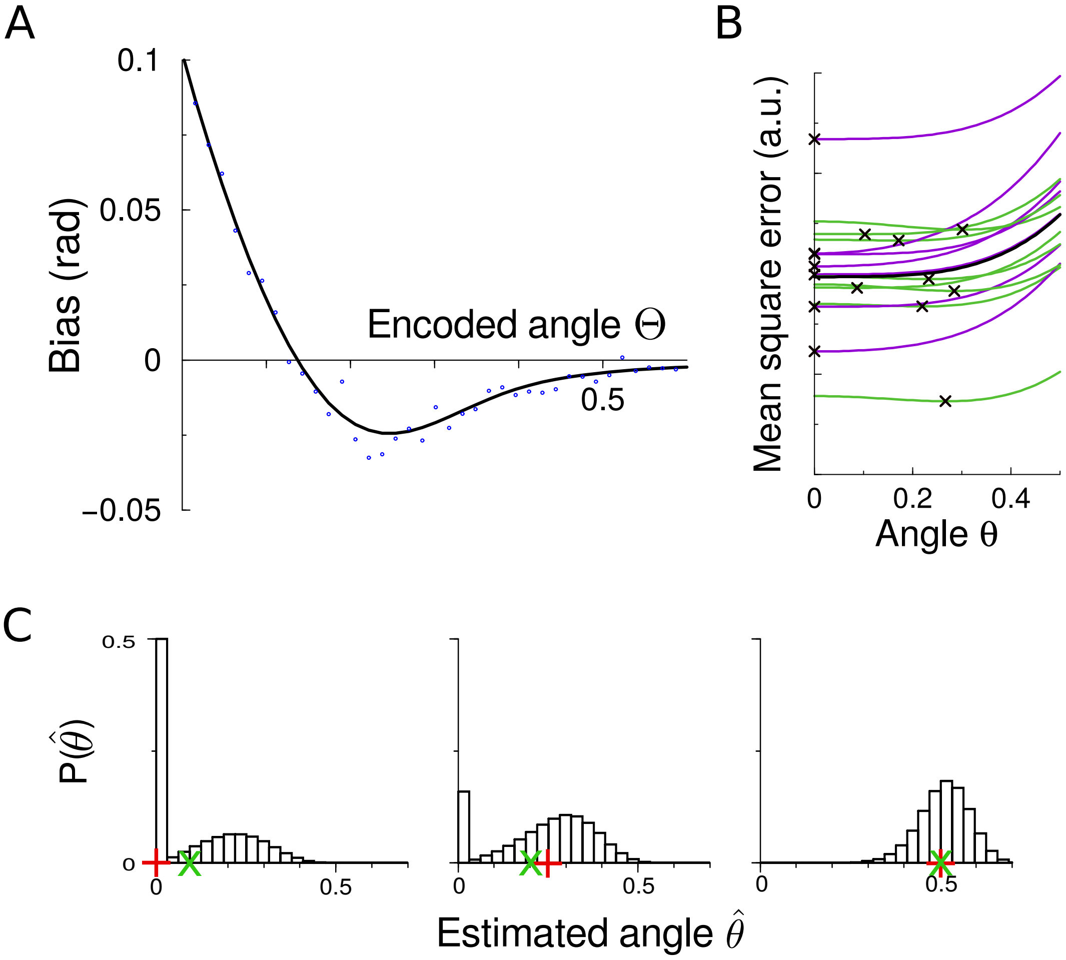

Figure 2: Decoding biases of the opening angle and the underlying decoding distribution.
a) Bias in estimation of the opening angle as a function of its true value, showing both a repulsive bias at small angles and a attractive one at larger angles. The curve was calculated using the algorithm given in the Methods. Also shown for comparison are simulations (dots) averaged over 1000 simulations per point.
b) Samples of the Mean Square Error in case the true opening angle is zero, the minima of the MSE correspond to the estimate of maximum likelihood estimator. While the average MSE has a minimum at the true value (black curve), on a given noisy trial the estimate can either be exactly $\theta = 0$ (shown in purple), or repulsed away from it (shown in green). The black crosses indicate the estimates, i.e. the angle that minimizes the error, on the individual trials.
c) Distribution of estimates that underlies the bias. When the true stimulus value is $\Theta = 0$ where the bias is repulsive (left), when the true stimulus value is $\Theta = 0.25$ where the bias is attractive (middle), and when $\Theta = 0.5$ where the bias is virtually absent (right). The true stimulus value is indicated with the red plus on the x-axis, the mean estimate is denoted with the green X. (all angles in radians).

$E_{\text{mean}}(\theta) \sim O(\theta^4)$.

The noise term E_{noise} is also symmetric and smooth in θ, however its second derivative is non-zero. In leading order it is, depending on the noise, either an upward or downward curved parabola centered around the origin. For small θ this parabola will dominate over E_{mean}. Therefore, if the parabola is U-shaped and thus with a minimum at $\theta = 0$, the total MSE also has a global minimum there, Fig. 2B, purple curves. If, on the other hand, the noise term has a maximum at $\theta = 0$, the global minimum will repulsed away from the true solution, Fig. 2B, green curves. As a result the distribution shows a sharp peak at 0, and a smeared peak further away, Fig. 2C (left). Furthermore, when the encoded angle $\Theta = 0$, exactly half of the estimates will be at $\theta = 0$ (i.e. fall on the diagonal in Fig. 1C) and the other half not. As Θ increases, the probability to find estimates $\hat{\theta} = 0$ will decrease and the second distribution will gain more mass until the mass at zero disappears, Fig. 2C (middle and right). The net effect is that this will first decrease the repulsive bias, then turn into an attractive bias, and finally the bias disappears.

In the Methods we describe how the Gaussian Process approach can be used to calculate the probability of estimates $P(\hat{\theta}|\Theta)$ in a numerically exact way without relying on simulations. This method was used to create Fig. 2A+C, and compares well to explicit simulations over many trials (dots in Fig. 2A).

Dependence of bias on noise

The bias curve depends on the neural noise level, Fig. 2A and other system parameters. In the limit of small angles the bias can be found by estimating the expected location of the minima of the Mean Square Error (markers in Fig. 2B). As shown in the Methods this gives for the tuning curves used,

$$b(0) = c\sqrt{\frac{\sigma}{A}}.\sqrt[4]{\frac{w^3}{N}}. \tag{3}$$

where σ is the std.dev. of the neural noise, w is the tuning width, A is the maximum neural response amplitude, N is the number of neurons and $c \approx 1.2$ is a numerical constant. Therefore to, say, half the bias, one needs 4 times less noise, or 16 times as many neurons. The second effect of the noise level is a shift in the angle at which repulsion becomes attraction, i.e. where the curve in crosses the x-axis (see Fig. 4A). Because the slope of the bias is exactly -1 at the origin ($b(\Theta) \approx b(0) - \Theta + O(\Theta^2)$, see below), the location of this transition point is well approximated by the bias at zero.

Interestingly, as the noise is reduced, the distribution of estimates remains bi-modal. While in the limit of zero noise the bias disappears as the theory of maximum likelihood estimation requires, the transition in the limit of small angles is not due to a collapse of $P(\hat{\theta})$ into a single Gaussian distribution, rather it is due to the two peaks in the distribution of estimates moving

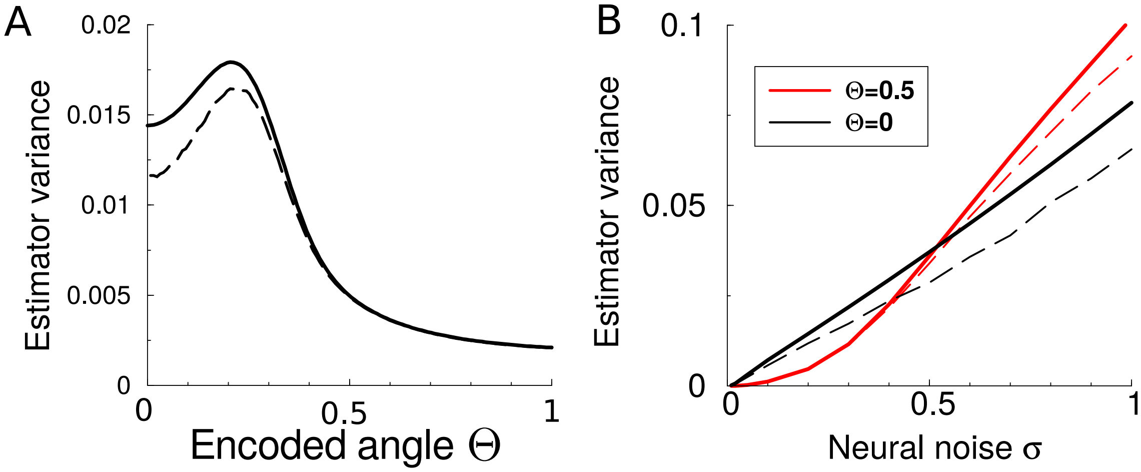

Figure 3: Variance and efficiency of the maximum likelihood decoder and its dependence on encoded angle and neural noise.
A. Variance in the estimates depends non-monotonically on the encoded angle. Dashed curve correspond to the Cramer-Rao bound; no estimator can achieve a lower variance.
B. Variance in the estimates as a function of the neural noise comparing large and small encoded angles. At small angles the strong bias alters the expected square dependence on noise into linear behaviour. Dashed curves correspond to the Cramer-Rao bound.

closer and closer together.

Effect of the bias on estimator efficiency

The quality of population code readout is not quantified by the bias alone, but also by the amount of trial-to-trial variations in the estimates, i.e. the variance in the distributions in Fig.1B+C. The variance in the ML decoder estimates follows directly from the distribution of estimates $P(\hat{\theta}|\Theta)$ that our Gaussian Process approach yields. The variance in the estimator at a particular noise level is plotted in Fig. 3A.

The minimal variance any estimator can achieve is limited by the Fisher Information through the Cramer-Rao bound which states that the variance of any estimator obeys (Methods)

$$\mathrm{var}(\hat{\theta}) \geq \frac{[1 + b'(\Theta)]^2}{\mathcal{I}(\Theta)}, \tag{4}$$

where b' is the derivative of the bias, and the Fisher Information $\mathcal{I}(\Theta)$ is given by Eq.8, Methods. The efficiency of an estimator expresses how close it comes to this bound. The resulting Cramer-Rao bound is indicated by the dashed curve in Fig. 3A. The estimator performs at the theoretical limit for large opening angles, and diverges (although it remains close) for smaller opening angles.

For large Θ, the estimate distributions are Gaussian with a width proportional to the neural noise; the bias plays a minor role. As expected from the Fisher Information, the variance of the estimator is proportional to the square of the neural noise, Fig. 3B, red curve. However, near $\Theta = 0$, the bias has a profound effect on the estimator. The estimator variance at $\Theta = 0$ can be approximately found by describing the estimate distribution $P(\hat{\theta}|\Theta)$ as a peak at zero and a Gaussian, Fig.2C. As for small angles the spread of the Gaussian is smaller than its mean, the variance is similar to the bias squared, and thus its parameter dependence can be calculated from squaring Eq. 3. Therefore, in contrast to square dependence at large angles, the variance in the estimates is only linear in the neural noise, as is confirmed in Fig.3B, black curve. Finally note that for low noise levels the variance at small angles is larger than the variance at larger angles, but that this switches at high noise.

Interestingly, Cramer-Rao bound follows the estimator performance and is linear in the neural noise, Fig.3B, dashed curves. In contrast, the Fisher Information is proportional to the neural noise squared σ^2, Eq. 8. The reason is that for small angles the Fisher Information goes to zero (Eq.8), but, because the estimator is a smooth, symmetric function, its derivative at the origin equals $b'(0) = -1$. Hence at small angles, both numerator and denominator of the Cramer-Rao bound, Eq. 4 go to zero and the net result is a linear dependence on the neural noise. For the parameters used, the MLE always achieves an efficiency $\geq 95\%$.

As an aside, in calculating the Cramer-Rao bound another advantage of the Gaussian Process approach shows. With simulations the bias and in particular its derivative are hard to calculate accurately, even using a large number of realizations, Fig. 2A dots. However, the numerically exact method to calculate the bias (Methods) allows for a precise calculation of the bias and its derivative. For instance, for a Bayesian decoder the precise bias is much harder to obtain, Fig. S1.

In summary, the bias does not simply lead to a small correction in estimator performance, but fundamentally alters it.

Competitive coding reduces bias

The estimation bias depends on the encoding model, that is, how the stimuli are coded in the neural response. Above it was assumed that the neural response to two simultaneous stimuli was the sum of the responses to the individual stimuli. While there is some experimental evidence for such a linear interaction, in other studies evidence for more competitive interaction has been found in area MT (Britten and Heuer, 1999), as well as other visual cortices (Gawne and Martin, 2002; Lampl et al., 2004; Oleksiak et al., 2011).

Such interactions have been modeled using a maximum-like interaction, so that instead of Eq. 2, the response of a single neuron to two simultaneous stimuli is

$$f_i(s_1, s_2) = \max[g_i(s_1), g_i(s_2)]. \tag{5}$$

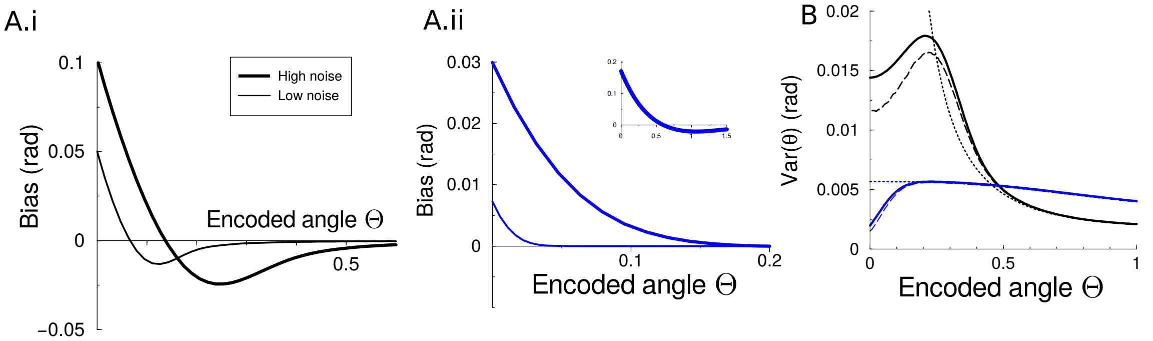

Figure 4: Bias and variance in a competitive coding model. A.i Bias in the estimates in the linear coding model for two different noise levels. The standard deviation of the neural noise was either high ($\sigma = 0.2$) or low ($\sigma = 0.05$).

A.ii Bias in the estimates in a competitive coding model where the response of any neuron to two stimuli equals the maximum response to the individual stimuli for the same noise levels as used in panel A.i. Only at very high noise levels ($\sigma = 1$), the attractive bias manifests itself (inset). Note the difference in scales.

B. Trial-to-trial variance in the estimates for the linear coding model (black) and the competitive coding model (blue). Dashed curve correspond to the Cramer-Rao bound. Dotted line corresponds to the Cramer-Rao bound uncorrected for the presence of bias (i.e. the inverse Fisher Information).

When the simulations are repeated for this encoding model, the bias is still present, but it is substantially smaller, Fig. 3A.ii. The repulsive bias is now approximately linear in the noise and the attractive component of the bias is smaller and becomes only apparent at even higher noise levels (see inset). Also the variance in the estimate is smaller, Fig. 3B, however, the bias still substantially alters the performance limit - compare the Cramer-Rao bound before (dotted) and after bias correction (dashed).

The mathematical reason for the reduced bias and variance, is that in this case the mean term in the MSE is not quartic but quadratic, reducing the bias. Thus we find that the precise coding model is an important determinant of the size of the bias, and these findings suggest a functional role for competitive interactions.

Discussion

Traditionally, theoretical studies of population codes have focused on estimator variance. Whenever biases have been studied, they have been explained from inhomogeneities in the neural encoding. Here we find that when multiple stimuli are encoded simultaneously in a relatively simple coding problem, substantial biases arise, which unlike previous mechanisms, are intrinsic to the decoder. That biases occur is in itself not surprising. Apart from cases where symmetry rules out biases, the absence of biases can only be proven in the limit of low noise, and in general an ML decoder will not be unbiased (Kay, 1993; Seriès et al., 2009; Pilarski and Pokora, 2015), nor efficient (Xie, 2002). Yet, the rich structure of the biases in these simple models, including its biphasic character and its relative persistence at low noise, is surprising.

The reason for the biases is the bimodal distribution of decoding estimates. By using a Gaussian Process approach we have developed an analytical theory of ML decoders which allows calculation of this distribution and the bias. Although the biases will disappear in the limit of zero noise, the bias diminishes only slowly as noise is reduced (proportional to the square root of the std. dev. of the neural noise). The persistence at low noise contrasts other studies of bias and efficiency where effects disappear abruptly when noise is lowered (Kay, 1993; Xie, 2002).

The results generalize in a number of directions. While we have only shown results for additive Gaussian noise, simulations show that our results extend to Poisson and multiplicative Gaussian noise, as well as correlated noise. Competitive encoding, such as the maximum coding, can reduce the bias, but will not abolish it, Fig. 4. Similarly, while our analysis relies on the maximum likelihood decoder, we find that the results are not unique to using the maximum likelihood estimator and occur with Bayesian decoders as well (Fig. S1). The nature of neural decoding mechanism is currently not clear, although it has been argued that it is straightforward to implement ML decoders neurally (Jazayeri and Movshon, 2006).

The results bear upon psycho-physical experiments where two overlapping random dot motion patterns with different directions are presented and subjects are asked to guess the angle between the two directions. In such experiments repulsive biases have commonly been observed (Marshak and Sekuler, 1979, but see Braddick et al., 2002). Several effects have been hypothesized to underlie these biases, including adaptation (the bias *increases* with presentation time, Rauber and Treue, 1999), cortical interactions (Carandini and Ringach, 1997) and repulsion from the cardinal directions (Rauber and Treue, 1998). The bias described here, is not at odds with those explanations, but presents a novel contribution to the total bias that is intrinsic to the decoder and which should be most prominent at small angles and for short presentation times.

The estimate distribution can be seen to reflect an ambiguity between the presence of one or two stimuli. Apart from predicting a bias, the theory predicts a bi-modal distribution of direction difference estimates and for small angles about half the time the two motions should be perceived as one. In experiments the number of stimuli that can simultaneously be perceived using overlapping motions is limited (e.g. Edwards and Greenwood, 2005) and when three or five overlapping motions are presented, they can sometimes be perceived as two (so called metamers, Treue et al., 2000); an effect which previously has been explained using the probabilistic population code framework (Zemel et al., 1998; Zemel and Dayan, 1999). The results here suggest that differences in the numerosity between presented and perceived stimuli already emerge with maximum likelihood decoders. Quantitative verification of this prediction of our study should be possible but might be challenging as participants' expectations and, similarly, natural priors for perceiving a single motion direction instead of two directions can influence results.

Acknowledgments

SK was supported by the EuroSpin Erasmus Mundus program and the EPSRC NeuroInformatics DTC. The authors would like to thank Udo Ernst, Richard van Wezel, Lawrence York, Peggy Series, and Chris Williams for discussions.

Methods

Neural population response

We use a population of $N = 100$ neurons. The tuning of neuron i to a single stimulus is given by $g_i(s) = A \exp\left[-\frac{(s-\phi_i)^2}{2w^2}\right]$. Here A is the response amplitude (arbitrarily set to 1), w is the width of the tuning curve (set to 1/2). The preferred directions ϕ_i of the neurons are equally spaced between 0 and 2π. As is common, we assume that the angles involved are relatively small, so that we don't have to worry their circularity, which would add complication through the need for circular statistics but does not change the results qualitatively.

When multiple stimuli are present, the neural response is modeled as the sum of the responses to the individual stimuli. After transforming the variables to the sum angle η and the difference angle Θ (see Main text) we can set η to zero, so that the tuning of neuron i becomes

$$f_i(\Theta) = g_i(\Theta/2) + g_i(-\Theta/2). \tag{6}$$

By replacing A by half its value one can obtain a model where the joint tuning curve equals the average (instead of the sum) of the tuning curves. The default value of the std. dev. of the noise in Eq. 1 was $\sigma = 0.2$.

Scaling of the bias

Here we calculate the bias for small angles analytically and estimate how the bias scales with the model parameters. We use that in case of small Θ and the limit of small candidate angles θ, the mean square error can be Taylor expanded as:

$$\begin{aligned} E_{\text{mean}}(\theta) &= \sum_i [f_i(\theta) - f_i(\Theta)]^2 \\ &\approx \frac{3\sqrt{\pi}}{64} \frac{\rho A^2}{w^3} \theta^4 \equiv \alpha \theta^4 \end{aligned}$$

where we replaced the sum by an integral and where ρ is the coding density (the number of neurons per unit angle, $\rho = N/2\pi$). Similarly, the noise term on a given trial can expanded as

$$E_{\text{noise}}(\theta) \approx [-2\sum_i \nu_i f_i''(\Theta)]\theta^2$$

The coefficient in the square brackets is a Gaussian random variable with zero mean and a variance $4\sigma^2 \sum_i [f_i''(0)]^2 \approx \frac{3\sqrt{\pi}}{4}\sigma^2\rho A^2/w^3$. We are interested in the cases where the coefficient will be negative as these are the repulsive trials, which happens in half of the trials. The mean value of a Gaussian truncated below zero is $-\sqrt{2/\pi}$ times the standard deviation, so that for these cases $\langle E_{\text{noise}}(\theta)\rangle \approx -\beta\theta^2$, with $\beta^2 = \frac{3}{2\sqrt{\pi}}\sigma^2\rho A^2/w^3$.

The approximate location of the repulsed minimum is given by $\frac{dE(\theta)}{d\theta}|_{\theta=\hat{\theta}} = 0$, or $\frac{d}{d\theta}(\alpha\theta^4 - \beta\theta^2)|_{\theta=\hat{\theta}} = 0$, and thus $\hat{\theta}^2 = \frac{\beta}{2\alpha}$. The bias in the other half of the trials is zero (purple traces in Fig. 2B), hence the total bias is $b(0) = \frac{1}{2}\sqrt{\frac{\beta}{2\alpha}}$. This yields the relation in the main text, Eq.3. The dependency of Eq.3 on all its parameters was confirmed numerically.

Calculation of maximum likelihood estimate

Here we demonstrate how to calculate the distribution of estimates $P(\hat{\theta}|\Theta)$ of the ML estimator in a numerically exact manner. Given a noisy response $\mathbf{r}$, we run over all candidate stimulus estimates and find the probability that it minimizes the Mean Square Error. Because the Mean Square Error is a smooth Gaussian process, and nearby E's are correlated, we can finely discretize θ. We define a set of M candidate estimates $(\theta_1, ..., \theta_M)$. To calculate that the probability that a certain estimate θ_m yields the lowest MSE, it is compared to the MSE that all other $M-1$ estimates yield. We define the $M-1$ dimensional set of MSE differences as $\mathbf{D}_m = E(\theta_m) - E(\mathbf{\Phi}_m)$, where $\mathbf{\Phi}_m = \{\theta_1, ..., \theta_M\}\backslash\theta_m$.

The distribution of differences $\mathbf{D}_m$ is a $(M-1)$-dimensional multivariate normal distribution

$$p(\mathbf{D}_m|\Theta) = \mathcal{N}(\boldsymbol{\mu}^m, \Sigma^m),$$

where $\boldsymbol{\mu}^m = E_{\text{mean}}(\theta_m) - E_{\text{mean}}(\mathbf{\Phi}_m)$ and the $(M-1)\times(M-1)$ covariance matrix has entries $\Sigma^m_{ab} = 4\sigma^2\sum_{i=1}^N [f_i(\theta_m) - f_i(\theta_a)][f_i(\theta_m) - f_i(\theta_b)]$. The probability that θ_m has a lower MSE than all other candidate estimates, is

$$p(\mathbf{D}_m < \mathbf{0}|\Theta) \quad = \quad \int_{-\infty}^{0} ... \int_{-\infty}^{0} p(\mathbf{D}_m|\Theta)\mathrm{d}\boldsymbol{D_m}, \tag{7}$$

which is a multi-variate cumulative normal distribution.

While this orthant integral is not analytically tractable, efficient algorithms exist that calculate it to a high precision for values of M up to in the hundreds. We used the quasi-Monte Carlo integration function `mvnun` from Scipy (Genz, 1992, 1998) with $M = 100$ and $\theta = 0\ldots\pi$

(using a larger M had negligible effects), and evaluated the integral for all values of m. This yields $P(\hat{\theta}|\Theta)$.

We note that while we applied it here to the estimation of Θ, this approach is general; it allows for arbitrary tuning curves and correlated Gaussian noise. It also extends to higher dimensional stimuli, but as one needs to discretize the stimulus space, a limitation is the efficient calculation of the integrals. Algorithms that calculate them for even higher dimensions exist (e.g. Azzimonti and Ginsbourger, 2016).

Calculation of the Cramer-Rao bound

Here we show how the Fisher Information is calculated which we use to compare to the variance in the estimator. The Fisher Information matrix for additive, uncorrelated Gaussian noise is given by $\mathcal{I}_{kl} = \frac{1}{\sigma^2}\sum_{i=1}^{N} \partial_{s_k} f_i(\boldsymbol{s}) \partial_{s_l} f_i(\boldsymbol{s})$. In the limit of dense tuning curves, the sum becomes an integral. While in the original $\boldsymbol{s}$-coordinates the Information matrix has off-diagonal elements (Orhan and Ma, 2015), in the coordinates (Θ, η) it becomes diagonal

$$\mathcal{I}(\Theta) = \frac{A^2\rho\sqrt{\pi}}{8w^3\sigma^2}\begin{pmatrix} 2w^2 + (\Theta^2 - 2w^2)\mathrm{e}^{-\Theta^2/4w^2} & 0 \\ 0 & 2w^2 + (2w^2 - \Theta^2)\mathrm{e}^{-\Theta^2/4w^2} \end{pmatrix}, \qquad (8)$$

The diagonal nature confirms the intuition that the opening and the sum angles can be estimated independently. Further note that both information components depend on the opening angle Θ, but neither depends on the sum angle η. This is due to the rotation invariance of the problem w.r.t. η. Finally, the Fisher Information for Θ, that is $\mathcal{I}_{11}$, is zero for small Θ (Amari and Nakahara, 2005).

An estimator is called *efficient* if its decoding covariance satisfies the Cramer-Rao bound (CRB) (Rao, 1945; Cramér, 1946; Rao, 2008). In the oft studied case of un-biased, one-dimensional estimators, the CRB is $\mathrm{var}(\hat{\theta}) \geq 1/\mathcal{I}(\Theta)$. In the case of biased vector parameters the CRB states that the matrix

$$C - B\mathcal{I}^{-1}B^T,$$

should be a positive definite matrix (Cover and Thomas, 1991; Kay, 1993). Here C is the covariance matrix of the stimuli that the estimator yields, B is the sum of the Jacobian matrix of $\mathbf{b}$ and the identity matrix. In our case this reduces to the bound in the main text.

Fisher Information in max-coding model

For max-coding the Fisher Information is identical for both sum and difference angles, $\mathcal{I}(\Theta) = \frac{A^2\rho}{8w^2}\{\sqrt{\pi}w[1 + \mathrm{erf}(\Theta/2w)] - \Theta\mathrm{e}^{-\Theta^2/4w^2}\}I$, where I is the 2×2 identity matrix. This is a monotonic function in Θ. When there are two separate peaks in the population response

$(\Theta \gg w)$, the information is twice that when $\Theta = 0$, where there is a single peak in the tuning.

References

S.-i. Amari and H. Nakahara. Difficulty of singularity in population coding. *Neural computation*, 17(4):839–858, 2005.

D. Azzimonti and D. Ginsbourger. Estimating orthant probabilities of high dimensional gaussian vectors with an application to set estimation. *arXiv preprint*, arXiv:1603.05031, 2016.

O. J. Braddick, K. A. Wishart, and W. Curran. Directional performance in motion transparency. *Vision Res*, 42(10):1237–1248, 2002.

K. H. Britten and H. W. Heuer. Spatial summation in the receptive fields of mt neurons. *Journal of Neuroscience*, 19(12):5074–5084, 1999.

M. Carandini and D. L. Ringach. Predictions of a recurrent model of orientation selectivity. *Vision research*, 37(21):3061–3071, 1997.

J. M. Cortes, D. Marinazzo, P. Series, M. W. Oram, T. J. Sejnowski, and M. C. W. van Rossum. The effect of neural adaptation on population coding accuracy. *J Comput Neurosci*, 32(3): 387–402, 2012.

T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley, New York, 1991.

H. Cramér. *Mathematical Methods of Statistics.* NJ: Princeton Univ. Press., 1946.

A. S. Ecker, P. Berens, A. S. Tolias, and M. Bethge. The effect of noise correlations in populations of diversely tuned neurons. *J Neurosci*, 31(40):14272–14283, 2011.

M. Edwards and J. A. Greenwood. The perception of motion transparency: A signal-to-noise limit. *Vision Research*, 45(14):1877–1884, 2005.

T. J. Gawne and J. M. Martin. Responses of Primate Visual Cortical V4 Neurons to Simultaneously Presented Stimuli. *J Neurophysiol*, 88:1128–1135, 2002.

A. Genz. Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Statist.*, 1(2):141–149, 1992.

A. Genz. MVNDST: Software for the numerical computation of multivariate normal probabilities, available from web page at http://www. sci. wsu. edu/math/faculty/genz/homepage. 1998.

M. Jazayeri and J. A. Movshon. Optimal representation of sensory information by neural populations. *Nature neuroscience*, 9(5):690–696, 2006.

S. Kay. *Fundamentals of statistical signal processing: Estimation theory.* Prentice-Hall, NJ, 1993.

I. Lampl, D. Ferster, T. Poggio, and M. Riesenhuber. Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *J Neurophysiol*, 92(5):2704–2713, 2004.

W. Marshak and R. Sekuler. Mutual repulsion between moving visual targets. *Science*, 205 (4413):1399–1401, 1979.

R. Moreno-Bote, J. Beck, I. Kanitscheider, X. Pitkow, P. Latham, and A. Pouget. Information-limiting correlations. *Nat Neurosci*, 17(10):1410–1417, 2014.

A. Oleksiak, P. C. Klink, A. Postma, I. J. M. van der Ham, M. J. Lankheet, and R. J. A. van Wezel. Spatial summation in macaque parietal area 7a follows a winner-take-all rule. *J Neurophysiol*, 105(3):1150–1158, 2011.

A. E. Orhan and W. J. Ma. Neural population coding of multiple stimuli. *The Journal of Neuroscience*, 35(9):3825–3841, 2015.

S. Pilarski and O. Pokora. On the cramér–Rao bound applicability and the role of fisher information in computational neuroscience. *Biosystems*, 136:11–22, 2015.

C. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(81-89), 1945.

C. Rao. Cramér-Rao bound. *Scholarpedia*, 3(8):6533, 2008. doi:10.4249/scholarpedia.6533.

H.-J. Rauber and S. Treue. Reference repulsion when judging the direction of visual motion. *Perception*, 27(4):393–402, 1998.

H.-J. Rauber and S. Treue. Revisiting motion repulsion: evidence for a general phenomenon? *Vision research*, 39(19):3187–3196, 1999.

E. Salinas and L. F. Abbott. Vector reconstruction from firing rates. *J. of Comput. Neurosc.*, 1:89–107, 1994.

O. Schwartz, A. Hsu, and P. Dayan. Space and time in visual context. *Nat Rev Neurosci*, 8(7): 522–535, 2007.

P. Seriès, A. Stocker, and E. Simoncelli. Is the homunculus "aware" of sensory adaptation? *Neural Comput*, 21:3271–3304, 2009.

M. Shamir and H. Sompolinsky. Implications of neuronal diversity on population coding. *Neural Comput*, 18(8):1951–1986, 2006.

M. Shamir. Emerging principles of population coding: in search for the neural code. *Curr Opin Neurobiol*, 25:140–148, 2014.

H. Sompolinsky, H. Yoon, K. Kang, and M. Shamir. Population coding in neuronal systems with correlated noise. *Phys. Rev E*, 64:51904, 2002.

A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nat Neurosci*, 9(4):578–585, 2006.

S. Treue, K. Hol, and H. J. Rauber. Seeing multiple directions of motion-physiology and psychophysics. *Nat Neurosci*, 3(3):270–276, 2000.

R. J. van Wezel, M. J. Lankheet, F. A. Verstraten, A. F. Marée, and W. A. van de Grind. Responses of complex cells in area 17 of the cat to bi-vectorial transparent motion. *Vision research*, 36(18):2805–2813, 1996.

C. K. Williams and C. E. Rasmussen. Gaussian processes for machine learning. *MIT Press*, 2 (3):4, 2006.

X. Xie. Threshold behaviour of the maximum likelihood method in population decoding. *Network: Computation in Neural Systems*, 13:447–456, 2002.

R. S. Zemel, P. Dayan, and A. Pouget. Probabilistic interpretation of population codes. *Neural Comput*, 10(2):403–430, 1998.

R. S. Zemel and P. Dayan. Distributional population codes and multiple motion models. *Advances in neural information processing systems*, pages 174–182, 1999.

K. Zhang and T. J. Sejnowski. Neuronal Tuning: to sharpen or to broaden? *Neural Comp.*, 11:75–84, 1999.

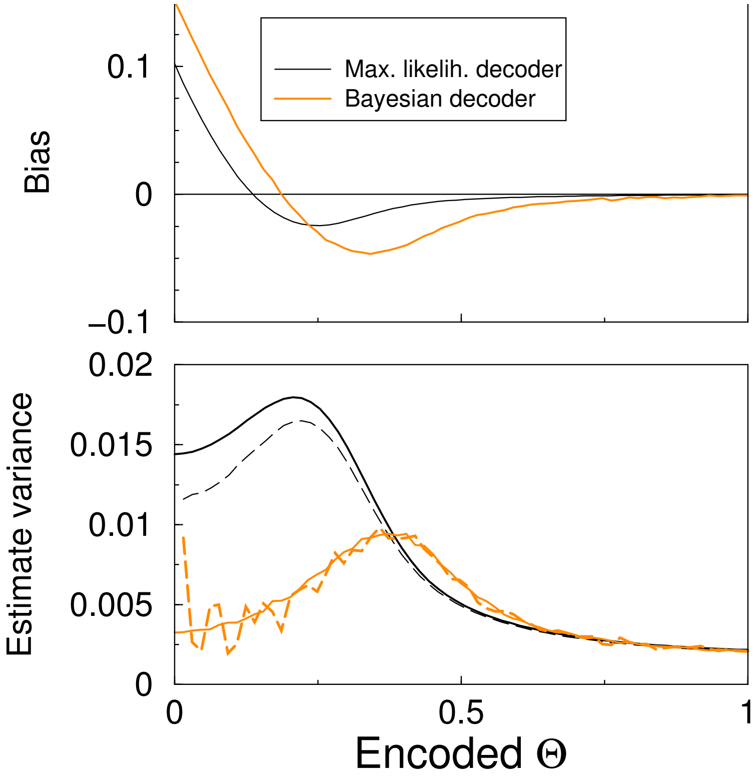

Figure S1: Bias and efficiency of a Bayesian decoder. Top: Bias in the estimate when using a Bayesian decoder (orange). Shown for comparison the ML decoder used in the main text (black). The biases are of comparable magnitude and share the biphasic character.
Bottom: Standard deviation in the estimator. Dashed line shows the Cramer-Rao bound. The ML decoder is shown in black. Note that due to its dependence on the bias, the bounds for the estimators are different. The Cramer-Rao bound for the Bayesian decoder is more variable because it relies on a bias that needs to be extracted from simulations.

Supplementary figure