J Comput Neurosci () :
DOI 10.1007/s10827-006-6357-1

Effects of stimulus transformations on estimates of sensory neuron selectivity

Alexander G. Dimitrov · Tomáš Gedeon

Received: 1 March 2005 / Revised: 24 October 2005 / Accepted: 28 November 2005 / Published online: xx xx

Abstract Stimulus selectivity of sensory systems is often characterized by analyzing response-conditioned stimulus ensembles. However, in many cases these response-triggered stimulus sets have structure that is more complex than assumed. If not taken into account, when present it will bias the estimates of many simple statistics, and distort the estimated stimulus selectivity of a neural sensory system. We present an approach that mitigates these problems by modeling some of the response-conditioned stimulus structure as being generated by a set of transformations acting on a simple stimulus distribution. This approach corrects the estimates of key statistics and counters biases introduced by the transformations. In cases involving temporal spike jitter or spatial jitter of images, the main observed effects of transformations are blurring of the conditional mean and introduction of artefacts in the spectral decomposition of the conditional covariance matrix. We illustrate this approach by analyzing and correcting a set of model stimuli perturbed by temporal and spatial jitter. We apply the approach to neurophysiological data from the cricket cercal sensory system to correct the effects of temporal jitter.

Keywords Stimulus transformation · Sensory systems · Neural coding · Dejitter

Action Editor: Matthew Weiner

A. G. Dimitrov (✉) · T. Gedeon
Center for Computational Biology, Montana State University, Bozeman, Montana, USA

T. Gedeon
Department of Mathematical Sciences, Montana State University, Bozeman, Montana, USA

1. Introduction

The mean and covariance of spike-conditioned stimulus sets are frequently used to characterize stimulus selectivity of neural sensory cells. The spike-conditioned mean is often interpreted as the stimulus "feature" to which a cell responds (Jones and Palmar, 1987; Meister et al., 1994; Poon and Yu, 2000; Reid and Alonso, 1995; Simoncelli et al., 2004). It has been proposed recently that the spike-conditioned covariance (STC) and its spectral decomposition can provide additional information about stimulus structures to which a cell responds as well (Agüera y Arcas and Fairhall, 2003; de Ruyter van Steveninck and Bialek, 1988; Rust et al., 2004; Schwartz et al., 2002; Theunissen et al., 2004). Many widely used characteristics of stimulus selectivity in neural sensory systems, like reverse Wiener kernels (Rieke et al., 1997), spatio-temporal receptive fields (STRF, DeAngelis et al. (1993) and Theunissen et al. (2004)) or spectro-temporal receptive fields (Eggermont et al., 1983; Poon and Yu, 2000), rely on similar simple statistics of response-conditioned stimuli. However, these response-conditioned statistics may be distorted by the action of several confounding processes, associated with uncertainty and non-uniqueness of neural system responses. The distortion can be substantial and lead to significant misrepresentation of the cells'functional characteristics. In this paper we present an approach that analyzes and corrects these distortions by explicitly modeling some of the response-conditioned noise sources.

As an example of the effects to which we refer, consider temporal uncertainty in the generation of single action potentials. In a classic experiment (Bryant and Segundo, 1976; Mainen and Sejnowski, 1995), a stimulus waveform generated by a band-limited white noise process is presented to a cell multiple times (frozen noise). On repeated presentation of the same sensory stimulus, the cell does not respond at exactly the same times. It exhibits a certain temporal jitter, typically captured in the stimulus-conditioned firing rate (PSTH). Imagine now that a spike-triggered statistic is estimated from the same dataset, as a proxy for the cell's functional properties. In this case typically the stimuli are aligned on the time of occurrence of individual spikes. Thus, the temporal jitter of spikes is translated into uncertainty in the time of occurrence of the spike-triggered stimuli. This will affect the estimates of statistical quantities, including mean (as illustrated recently in Aldworth et al. (2005)) and covariance. If these spike-conditioned quantities are used to represent stimulus-related function of this cell, they will lead to a distorted description of the cell's stimulus selectivity.

A similar effect also manifests itself when considering eye jitter and microsaccades in the visual system (Forte et al., 2002; Martinez-Conde et al., 2002). While the visual system may receive proprioceptor input with information about such events, this input is currently not available to researchers. So images in the response-conditioned stimulus ensemble will be contaminated by random spatial jitter. This will again distort the estimates of various statistical quantities. Stimulus selectivity estimated without taking this jitter into account will differ from the actual stimulus selectivity of a cell in the visual system.

These two examples can be seen as special cases of a more general phenomenon, which involves the action of some class of transformations on the stimulus, that leave the response unchanged. The two cases above are examples of 1-dimensional translation in time (temporal jitter) and 2-dimensional translation in space (spatial jitter). They leave the response invariant: in all cases a hypothetical single spike occurs at relative time zero with respect to the spike-triggered stimulus. Although these two examples deal exclusively with physiological noise, the invariance may also be due to the lossy nature of neural processing, where many different stimuli lead to identical responses. Many other transformations may conceivably modify the stimulus and not affect the response, including spatio-temporal translations, rotations, spatial or temporal stretching, and scaling to name just a few.

In this paper we present a framework in which to model, analyze and correct the effects of such transformations. The approach explicitly represents the effect of transformations on the stimulus and isolates them in a separate probability model. After the transformations are removed, the stimulus residual is processed in the conventional way. Statistics computed with the corrected stimulus will not contain artefacts introduced when these transformations are present.

In Section 2 we present the basic modeling framework. Using this framework, we describe the effects of transformations on the spike-triggered mean and covariance in the general case, and specialize to the case of temporal jitter. Section 3 develops tools with which to correct the biases in the mean and covariance introduced by transformations, and reverse their action on the stimulus by inferring the most likely set of transformations that could have produced the observed response-conditioned stimulus set. In Section 4 the tools developed in this framework are validated in two cases: (1) a model of temporal jitter of spike trains; (2) a model of spatial jitter in two dimensions, with model receptive filed similar to a a simple primary visual neuron (V1 simple cell model). In the same Section we also apply the methodology to the study of temporal jitter in an identified interneuron of the cricket cercal sensory system. The main effects that our theory predicts and we observe for these cases are:

- The mean, estimated in the presence of jitter (raw mean) is a blurred version of the true mean.
- The conditional covariance matrix, estimated in the presence of jitter (raw covariance), has artefactual eigenvectors. They resemble the derivatives (temporal or spatial) of the true mean when the jitter is small.

In Section 5 we discuss the implications of this work in the context of general neural sensory processing, and its relations to other research. Mathematical details of this investigation are relegated to the Appendix.

2. Sources of uncertainty in response-conditioned stimuli

We shall model the space of inputs preceding a distinct neural response as a probability space X with elements $x \in X$. We denote by $p(x \mid r)$ the conditional probability of x given that a response r occurs. This is a stimulus reconstruction, or "reverse" type of model. In principle, a model of neural response generated by the stimulus ("forward" model) can be obtained from the reverse model through Bayes' theorem by $p(r \mid x) = p(x \mid r)p(r)/p(x)$. However here we take the animal-centric stimulus reconstruction point of view and study $p(x \mid r)$. To simplify the notation, we shall denote the conditional stimulus probability simply as $p(x)$, implicitly assuming a fixed response type. We further restrict our attention to response sequences consisting of isolated single spikes, in order to avoid confounding effects arising from interaction between spikes. However, this approach can be applied to stimuli conditioned on any sequence of spikes, groups of spike patterns (Dimitrov and Miller, Victor and Purpura), or discriminable instances of other measures of neural activity (e.g., rates).

We shall model some of the sources of uncertainty in response-conditioned stimuli as being generated by random transformations that act on the stimulus and leave the response invariant (Grenander, 1996). As an example, the uncertainty in the timing of a spike given a stimulus can be interpreted as an invariance of the cell's response to small temporal shifts of the stimulus. In other words, if we slightly shift in time a given stimulus, the timing of the response spike will not change. The probability that a transformation leaves the response invariant will be modeled as a distribution on the set of transformations (Grenander, 1963). That is, the invariance of the response to stimuli is probabilistic: some transformations are less likely to leave the response unchanged compared to others.

We model the effects of transformations by following closely the transformation-invariant clustering formalism developed by Frey and Jojic (2003). There will be three spaces involved in this discussion: the space of observable (*raw*) stimuli Z, the set of *true* stimuli to which the cell is assumed to respond, X, and the space of transformations $\mathcal{T}$ that act on the true stimuli to produce the raw stimuli in Z. We parameterize the set $\mathcal{T}$ by $t \in T$ with probability $p(t)$ and denote the corresponding transformation by $g_t \in \mathcal{T}$. Thus the complete description of the system is given by the triple $(z, x, t) \in Z \times X \times T$, and the probability $p(z, x, t)$ in this product space. In this paper the only transformations considered are those for which the true space X coincides with the raw space Z ($X \equiv Z$), that is, $\mathcal{T}$ is a set of automorphisms.

The assumption that a raw stimulus z is obtained by the action of a transformation > upon a true stimulus means that

$$p(z \mid x, t) = p(z \mid g_t x)$$

where $g_t x$ is the action of a transformation $\mathrm{g_t}$ on a stimulus x. For practical purposes, we will always assume, as in Frey and Jojic (2003), that $p(z \mid g_t x) = \mathcal{N}(z; g_t x, \Psi)$ is a multivariate normal distribution with mean $g_t x$ and instrument noise given by the covariance matrix Ψ. We assume that Ψ has simple structure (diagonal or spherical) and is much smaller than other sources of noise in the problem (e.g. the maximal eigenvalue of Ψ is much smaller than the maximal eigenvalue of any other covariance matrix present in the problem). As such, it is unlikely to randomly generate transformations on the same scale as the effects we are looking for. A further simplification we will make when convenient is that $\Psi = 0$, in which case $z = g_t x$. The instrument noise model is a useful technical abstraction, that makes all the quantities of interest random variables, and allows for a completely probabilistic treatment of the problem.

With these assumptions,

$$p(z, x, t) = \mathcal{N}(z; g_t x, \Psi) P(x, t)$$

We also assume that the joint probability factorizes:

$$P(x, t) = p(x) p(t),$$

that is, transformations are independently applied to stimuli. This brings us to the final probability model,

$$p(z, x, t) = \mathcal{N}(z; g_t x, \Psi) p(x) p(t) \tag{1}$$

From here onward we shall set the instrumental noise Ψ to 0, except when explicitly stated otherwise. In this case, $z = g_t x$.

In addition to the terms *true* and *raw*, describing the stimuli in spaces X and Z correspondingly, we shall use the term *dejittered* to denote our estimate of the true stimulus.

2.1. Effects of transformations on the conditional mean and covariance: general case

Typically, when analyzing a relation between stimuli and neural responses, we are interested in statistics of the true stimulus distribution $p(x)$. However, in the presence of transformations we can obtain immediate statistics only for the raw distribution $p(z) = E_{P(x,t)} p(z, x, t)$, as the other two variables are latent (unobservable). Equation (1) implies that the action of transformations modifies the raw response-conditioned stimulus distribution. We first describe the effects of transformations on the estimate of the conditional mean

$$\bar{x} = E_{p(x)} x \tag{2}$$

taken as a representative of the cell's stimulus preference. When we compute the average of the raw collection (1), we are actually estimating the parameter

$$\bar{z} = E_{p(z)} z = E_{p(z,x,t)} z.$$

As shown in Lemma 2 of Appendix A, if g_t are linear transformations, the relation between the true mean $\bar{x}$ and the mean in the presence of transformation (raw mean), $\bar{z}$, is

$$\bar{z} = E_{p(t)} \bar{x}_t, \tag{3}$$

where $\bar{x}_t := g_t \bar{x}$. That is, the raw mean $\bar{z}$ is the average over all transformations of the transformed true mean $\bar{x}_t$.

The transformations also affect the estimate of the covariance when this estimate is based on the raw set (1). There are differences between the true covariance matrix

$$C_x = E_{p(x)} (x - \bar{x})(x - \bar{x})^T \tag{4}$$

and the covariance matrix computed in the presence of transformations (raw covariance)

$$C_z = E_{p(z)} (z - \bar{z})(z - \bar{z})^T.$$

Using techniques similar to the ones applied to the analysis of the mean (3), in Lemma 3 of Appendix A we show that

$$C_z = \bar{C}_x + C_t, \tag{5}$$

when $\Psi = 0$. Here $\bar{C}_x = E_{p(t)}\, g_t C_x\, g_t^T$ is the expected transformed covariance and $C_t = E_{p(t)}\,(\bar{x}_t - \bar{z})\,(\bar{x}_t - \bar{z})^T$ is a covariance term induced by the difference between the transformed true mean $\bar{x}_t$ and the raw mean $\bar{z}$.

2.2. Model of the temporal uncertainty in neural cell responses

We now specialize our model of uncertainty to temporal uncertainty of spikes. In this case $\mathcal{T}$ is a set of time shifts acting on stimulus waveforms and the action of $g_t \in \mathcal{T}$ on the stimulus is

$$g_t x(\tau) := x(\tau - t). \tag{6}$$

We assume that the probability of a spike elicited at time t given a stimulus at time τ is distributed in time around the mean spike time, represented by the probability of spike at time τ given stimulus at the same time τ. The natural delay in response is build into the stimulus at time τ. In other words we have

$$p(\text{spike}(t)\,|\,\text{input}(\tau)) = p(t-\tau)p(\text{spike}(\tau)\,|\,\text{input}(\tau)).$$

For the analysis developed here, we need $p(\text{input}(\tau)\,|\,\text{spike}(t))$, which we obtain by Bayes' theorem:

$$\begin{aligned} p(\text{input}(\tau)\,|\,\text{spike}(t)) &= p(\text{spike}(t)\,|\,\text{input}(\tau))p(\text{input}(\tau)) \\ &\quad /p(\text{spike}(t)) \\ &= p(t-\tau)p(\text{spike}(\tau)\,|\,\text{input}(\tau)) \\ &\quad p(\text{input}(\tau))/p(\text{spike}(\tau)) \\ &= p(t-\tau)p(\text{input}(\tau)\,|\,\text{spike}(\tau)), \end{aligned}$$

as $p(\text{spike}(t)) = p(\text{spike}(\tau))$ is a constant, inversely proportional to the mean spike rate. In this case (3) specializes to

$$\bar{z}(\tau) = E_{p(t)}\bar{x}(\tau - t) = \int p(t)\bar{x}(\tau - t)dt =: p * \bar{x}, \tag{7}$$

where $*$ denotes the convolution operation. That is, for temporal jitter the raw mean is obtained by convolving the true mean with the jitter distribution. Correspondingly, (5) specializes to

$$C_z(\tau) = \int p(t)C_{x(t-\tau)}dt + \int p(t)(\bar{x}_t - \bar{z})(\bar{x}_t - \bar{z})^T dt$$

3. Analyzing and correcting the effects of transformations

Expression (7) points to a way to undo the effects of temporal jitter on the estimates of the spike-triggered average. The convolution with the distribution of jitters acts in exactly the same way as blurring (point spread function) in optical systems. Standard algorithms from image processing (Wiener deconvolution, regularized deconvolution, Gonzalez and Woods (1992) can be used to perform the deconvolution. All rely on some assumptions about the form of the convolution kernel $p(t)$, and about the level of noise, on which to base the regularization. We discuss some natural choices of those parameters in Appendix B.

It is harder to analyze the effects of jitter on the covariance matrix, since it depends non-trivially on the transformations. Here we approach this problem by assuming that the density $p(t)$ is sharply peaked around zero with small standard deviation σ_t and thus the distortions caused by transformations can be treated as perturbations. As we show in Lemma 4 of Appendix A in this case, the expression (5) becomes

$$C_z \approx C_x + \sigma_t^2\left(C_{Ax} + C^S_{A^2x} + C_A\right) \tag{8}$$

where A is the generator of the set of transformations, $C_{Ax} = E_{p(x)}A(x - \bar{x})\,(A(x - \bar{x}))^T$ is the expectation of the transformed residual, $C^S_{A^2x} = \frac{1}{2}(C_{A^2x} + C^T_{A^2x})$ is the symmetrized second order analog of C_{Ax}, and $C_A = (A\bar{x})(A\bar{x})^T$ depends only on the transformed mean $\bar{x}$.

Since expression (8) links C_z and C_x directly, it allows us to predict the effect of the transformations on the form and structure of eigenvectors of the raw covariance matrix C_z. We will apply this approximation to the case of temporal jitter (6). The approximation for temporal uncertainty is (see (A.17) in Appendix A)

$$x(\tau - t) \approx x(\tau) - \frac{dx}{dt}(\tau)t + \frac{d^2x}{dt^2}(\tau)\frac{t^2}{2}.$$

Then (8) becomes

$$\begin{aligned} C_z \approx C_x + \sigma_t^2 \int \left(\frac{d}{dt}(x - \bar{x})\frac{d}{dt}(x - \bar{x})\right)^T p(x)dx \\ + \frac{\sigma_t^2}{2}\int \left(\left(\frac{d^2}{dt^2}(x - \bar{x})\right)(x - \bar{x})^T \right.\\ \left. +(x - \bar{x})\left(\frac{d^2}{dt^2}(x - \bar{x})\right)^T\right) p(x)dx + \sigma_t^2\left(\frac{d\bar{x}}{dt}\right)\left(\frac{d\bar{x}}{dt}\right)^T. \end{aligned} \tag{9}$$

The spectral decomposition of the covariance matrix has gained a lot of popularity of recently as a way to uncover additional stimulus dimensions which can modulate neural responses independently of the mean (de Ruyter van Stevenick and Bialek, 1988; Rust et al., 2004; Schwartz et al., 2002). In particular, the space spanned by the leading or lagging eigenvectors is considered one such set of relevant stimulus dimensions. It is thus imperative to address the question of which of those eigenvectors are real and which are artefactually induced by the transformations. Expression (9) allows us to estimate how the leading eigenvectors of the raw covariance C_x and C_z are related. While we leave the details of the argument to the Appendix A, we remark that if the last term in (9) dominates the other terms then the leading eigenvector of C_z will be approximately equal to $\frac{d\bar{x}}{dt}$, the sole eigenvector of the last term. This perturbation technique is only able to explain some effects in the special case of peaked distribution of transformations and relatively small noise around the mean. Without these simplifying assumptions the situation is even more problematic, since the spectral decomposition of the covariance matrix will be transformed in less predictable ways, and more of its components will be affected. When applying this theory (Section 4), we empirically observe that several of the top eigenvalues and eigenvectors seem to be either pure artefacts of the transformations, or are heavily modified from the true distribution.

In the following section we discuss tools that allow for the general correction of such artefacts, without the assumption of small perturbation stated above. While these tools do not provide an explicit form of the artefacts, they do remove them to a great degree, and allow further analysis of the conditional mean and covariance structure. Similar tools have been developed by researchers in machine vision and automated object recognition (Amit et al., 1991; Frey and Jolic, 2003; Miller et al., 2000; Rao and Ruderman, 1999).

3.1. Estimating transformation parameters for individual samples: the dejittering procedure

Here we attempt to reverse the transformation on a sample-by-sample basis. The approach we take is similar to the transformation-invariant clustering developed in Frey and jojic (1999, 2003). According to our assumptions (1), $p(z, x, t) = \mathcal{N}(z; g_t x, \Psi) p(x) p(t)$. Using this distribution we can infer the pair (x, t) that is associated with an observed raw z. Assuming we know $p(z, x, t)$, this can be done by considering

$$\begin{aligned} p(x, t \mid z) &= p(z, x, t)/p(z) \\ &= \mathcal{N}(z; g_t x, \Psi) p(x) p(t)/p(z). \end{aligned} \tag{10}$$

This expression gives us a distribution over possible pairs (x, t). We shall select the pair (x^*, t^*) that maximizes (10). Since $p(z)$ is a constant for a fixed z, this is equivalent to maximizing the joint probability $\mathcal{N}(z; g_t x, \Psi) p(x) p(t)$. To simplify our computations further we again set $\Psi = 0$. Therefore $z = g_t x$ and hence $x = g_t^{-1} z$ is a deterministic function of z. Thus the only variable that remains to be optimized is t, and the problem to be solved is (M-step in an EM algorithm) (Dempster et al., 1977)

$$t^* = \arg\max_t p\left(g_t^{-1} z\right) p(t). \tag{11}$$

After finding t^*, set $x^* := g_t^{*-1} z$, obtaining the pair (x^*, t^*) which is most likely to have produced the observed z.

In reality, the distributions $p(x)$ and $p(t)$ are unknown and are initialized to arbitrary initial models $p_0(x)$ and $p_0(t)$. Once the pairs (x_i^*, t_i^*) are inferred for each sample z_i, the models for $p(x)$ and $p(t)$ are updated (E-step in an EM algorithm). As the two models are independent, the expectations for their parameters are run independently over the x_i^* and t_i^* sets inferred from the observations. The parameters that are estimated through the expectations depend on the types of models that are used for $p(x)$ and $p(t)$. The whole cycle is then iterated.

We now discuss one particular choice of models for $p(x)$ and $p(t)$. Consider $x \propto \mathcal{N}(x; \bar{x}, C_x)$, $t \propto \mathcal{N}(t; 0, \sigma_t)$ and $z(\tau) = g_t x(\tau) := x(\tau - t)$. The probability for a raw observation $z(\tau)$ to have come from this model is given by

$$p(x)p(t) = \mathcal{N}(z(\tau + t); \bar{x}, C_x) \mathcal{N}(t; 0, \sigma_t). \tag{12}$$

The optimal pair $(g_t^{*-1} z, t^*)$ is obtained as the solution to

$$t^* = \arg\max_t \mathcal{N}(z(\tau + t); \bar{x}, C_x) \mathcal{N}(t; 0, \sigma_t). \tag{13}$$

Note that here we are assuming (and enforcing) the mean of the t distribution to be $\bar{t} = 0$. For the first step, we initialize $p(x)$ with the estimates of the raw mean and covariance, $\bar{z}, C_z$, and $p(t)$ with a physiologically relevant σ_t. Given that the parameters of $p(t)$ are guessed anyway, a better starting point would be to assign $\bar{x}$ to the deconvolved $\bar{z}$ (7), and approximate C_x with $C_x = C_z - \sigma_t^2 (C_{Ax} + C^S_{A^2 x} + C_A)$ (see Eq. (8)).

For computational purposes it is better to write expression (13) in terms of the negative log likelihood of the transformed observation. This monotonic transformation does not change the position of any extremum, but dramatically increases the numerical precision. The non-constant portion of the log likelihood is a quadratic form of the variables, and hence a distance,

$$d((z,t),(\bar{x},0)) = \left(g_t^{-1}z - \bar{x}\right) C_x^{-1} \left(g_t^{-1}z - \bar{x}\right)^T + t^2/\sigma_t^2. \tag{14}$$

A minimal distance here implies maximal likelihood in (11).

In the case where t are temporal shifts, we have performed the procedure outlined in (14) under several simplifying assumptions about the structure of the covariance matrix C_x of the stimulus model, similar to the ones made by Dimitrov et al. (2003) One simplification to (14) constrains C_x to a diagonal matrix that can have different values (variances) on the diagonal. In this case the distance (14) is expressed by

$$d((z,t),(\bar{x},0)) = \sum_i \left(\left(g_t^{-1}z\right)_i - \bar{x}_i\right)^2/\sigma_{x_i}^2 + t^2/\sigma_t^2, \tag{15}$$

where z_i and $\bar{x}_i$ are the i-th coordinate of the raw stimulus sample and true mean, correspondingly. This distance will tend to accentuate (weigh more) coordinates with low variance, and disregard coordinates with high variance. Of course this is also automatically done by the full covariance C_x^{-1} in (14), but one typically needs many more samples for a reliable estimate of C_x from observations.

This distance, without the penalty term and in a probability form (exponentiated), was used by Chang et al. (2005) as a similarity index with which to correct the spectro-temporal receptive fields of rat auditory neurons, with results similar to the ones reported below and by Aldworth et al. (2005). In the context of the formalism presented here, this translates to assuming a uniform jitter distribution. This assumption is problem-dependent and may lead to the introduction of additional artefacts when not fulfilled, as random features far in time may be pulled towards and aligned to the template.

The simplest case in this series is when $C_x = \sigma_x I$, that is, the stimulus distribution is modeled as a spherical Gaussian. In this case the distance (15) further simplifies to

$$d((z,t),(\bar{x},0)) = \left|g_t^{-1}z - \bar{x}\right|^2/\sigma_x^2 + t^2/\sigma_t^2, \tag{16}$$

which is essentially an Euclidean distance between the inversely transformed stimulus and the true mean, penalized by the squared temporal shift needed to reverse the transformation.

4. Application

In this section we apply the tools developed in the previous section to two models of sensory processing, where we explicitly introduce transformations of a known kind. We also use the tools to analyze the stimulus selectivity of a sensory interneuron in the cricket cercal sensory system.

4.1. Analysis of temporal processing and temporal jitter: model studies

A simple model of the conditional stimulus illustrates the application of this analysis to neural signal processing. The model is a multivariate Gaussian, the mean of which is the putative target to which a cell responds in its assigned function of a signal discriminator. The model mean waveform was obtained by slightly modifying a spike-triggered average of a cricket sensory interneuron. We use two different models for the noise covariance: one with a spherical noise model around the mean (model 1), and another with an autoregressive noise model (model 2), the correlation function of which is similar to the one observed in physiological recordings in the cricket cercal sensory system. The mean and correlation functions for both models can be seen on Fig. 1. The covariance matrix of each model was obtained as a Töplitz matrix of the autocorrelation function. For model 1, this resulted in a multiple of the identity matrix (spherical noise model). For model 2, a more complex covariance matrix resulted, more similar to signal covariances estimated from physiological recordings. Both models are in 25 dimensional space at 1 ms temporal resolution; waveforms were interpolated to 0.1 ms for visualization purposes. Additionally, model 1 has a single variance parameter to describe the spherical noise structure around the mean. For model 2, the first 15 principle components (PC-s) account for $>95\%$ of the total variance in the model.

For both models we applied the transformation procedure outlined in Section 2: sample a stimulus from the multivariate normal model, and shift it by a time t. The shift times in both cases were sampled from a normal distribution $p(t) = \mathcal{N}(0 \text{ ms}, 1.5 \text{ ms})$. The results of the analysis for the more physiologically relevant autoregressive model 2 are presented in Fig. 2. The results for the spherical model 1 are very similar, and are not presented here in detail. The steps of sampling, jittering to obtain a simulated raw dataset and dejittering with the diagonal distance function (15) are presented in panels A, B and C correspondingly. The transformations acts on the mean as expected, by blurring it (green trace on panel D). Reversing the effects of jitter was successful: the true mean (blue) and reconstructions through dejittering (red) and deconvolution (magenta) essentially overlap. Panel E explicitly shows the top eigenvector of the raw covariance C_z, which will be shown to be an artefact from the transformation; it bears no resemblance to the top 3 eigenvectors computed from the true covariance matrix C_x (blue), or the top 3 eigenvectors of the dejittered covariance matrix (red). This was further confirmed by the angle between subspaces spanned by those eigenvectors. The

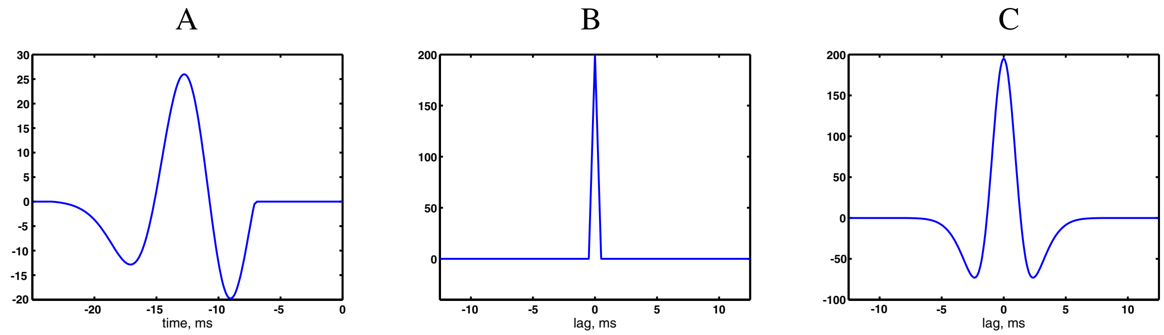

Fig. 1 Model parameters. A. Conditional mean of both models. B. Autocorrelation function of the residual for model 1. The autocorrelation peak is at temporal lag 0. C. Autocorrelation function for the residual for model 2. The autocorrelation peak is at temporal lag 0.

angle between the true (model) and dejittered subspaces was approximately 15°. The angle between the true and jittered subspaces was 77°, meaning that those 2 subspaces were almost orthogonal, a distortion caused by the transformations. The dejittering procedure cannot guarantee an exact recovery of the eigenvectors, as small perturbations in the top few eigenvectors may lead to relatively large changes of the whole eigensystem, due to the orthogonality imposed by the properties of the covariance matrix. The top 10 eigenvalues of the true covariance, the raw covariance, and the covariance estimated after dejittering (dejittered covariance) can be seen in Panel F. For eigenvalues obtained from estimated covariance matrices (jittered, dejittered), we obtained error margins by bootstrapping the eigenvalue estimates and computing the standard deviation of the bootstrap samples (Efron and Tibshirani, 1993). Estimates were based on 2000 sample drawn from model 2. The model covariance matrix defines model parameters, and hence model eigenvalues computed from it do not contain sampling uncertainty. The two largest eigenvalues of the raw covariance differ significantly (more than 95% level) from the corresponding values of the true covariance, implying that the spectral decomposition was significantly changed in at least 2 dimensions. Dejittering restores the original spectrum: red and blue values don't differ significantly. We discuss these effects in more detail in Fig. 4.

To establish if the dejittering procedure helps in explaining the observations better, we applied the model selection criteria described in C. Briefly, we fitted two different multivariate normal models to the observations. One was fitted to the set of samples (x_i, t_i) of stimuli and transformations. The second was fitted to the set of raw samples $z_i = g_{t_i} x_i$. After the models were estimated, we computed the log likelihood ratio between the two models with the same set of observations, and the corresponding difference of AIC values (Akaike's Information Criterion, see C). We report the average value of both criteria (per sample), so it can be compared for cases with different number of samples. Positive values in both cases favor the true process model; negative values favor the raw model. For the synthetic case discussed so far, the average log likelihood ratio was 0.6075 per sample. Since this is a logarithmic measure, it means that on the average, each sample was about 2 times more likely to be explained by the true model than by the raw model. The corresponding average difference of AIC criteria, which takes into account the small difference in model complexity, was 1.214, again favoring the true model. To obtain the corresponding values for the whole set of 2000 observations, the average values have to be multiplied by 2000, stressing the enormous advantage that the true process model has above the model directly estimated on observables.

4.2. Analysis of temporal processing and temporal jitter: physiological studies in the cricket cercal sensory system

The same procedures were applied to stimulus/response data from the cricket cercal sensory system. This mechanosensory system mediates the detection and analysis of low velocity air currents, and is considered a low-frequency, near-field extension of the animal's auditory system (Bacon and Murphey, 1984; Jacobs et al., 1986; Kämpar and Kleindienst, 1990; Kanou and Shimozawa, 1984; Miller et al., 1991; Roddey and Jacobs;1996; Theunissen et al., 1996). The data analyzed here consists of sensory stimuli and intracellular record of stimulus-evoked spike trains from the axon of the primary sensory interneuron IN10-3, kindly provided by Zane Aldworth. The sensory stimulus used to drive IN10-3 was a dynamic air current moving across the animal's body with Gaussian white noise (GWN) velocity profile band-passed at 5–150 Hz, which brackets the range of frequencies to which this cell is known to respond. The physiological protocols used here are detailed in Aldworth et al. (2005). The analysis reported below is based on 13,600 samples of isolated single spikes. The stimulus samples conditioned on isolated single spikes were represented as vectors in 20 dimensional space at 1 ms temporal resolution; waveforms were interpolated to 0.1 ms for visualization purposes. Additionally, the first

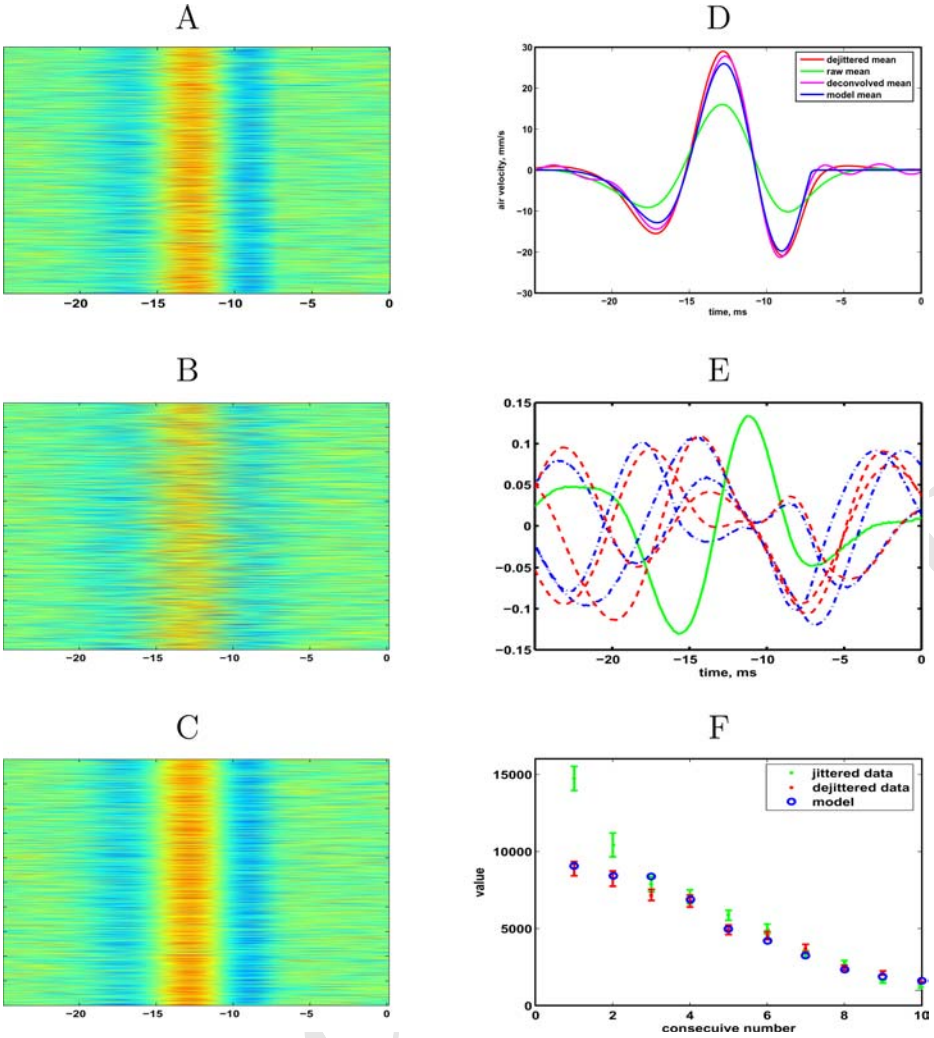

Fig. 2 Effects of temporal jitter on spike-triggered statistics: model studies. (A) Rasters of waveforms sampled from the autoregressive conditional stimulus model of interneuron function. (B) The samples from (A) are shifted randomly in time, with a distribution of shifts $p(t) = \mathcal{N}(0 \text{ ms}, 1.5 \text{ ms})$ to obtain a raw dataset that models a spike-triggered stimulus ensemble. (C) The effects of temporal jitter are removed from the raw dataset by dejittering with the cost function in Eq. (16). (D) Comparison between the true model mean (blue), raw mean (green), dejittered mean (red) and deconvolved mean (magenta). As expected, the raw mean is a blurred version of the true mean. The corrections to the mean, obtained either by dejittering or deconvolution, closely match the true mean. (E) Evidence that eigenvectors of the raw covariance may be artefacts of the transformations. In particular, the top eigenvector of the raw covariance C_z (solid green line) bears no resemblance to any of the top 3 eigenvectors of the true covariance matrix C_x (dot-dash blue lines), or of the dejittered covariance matrix (dashed red lines). The eigenvectors of the true and dejittered covariance matrices are similar. (F) Top 10 eigenvalues of the true covariance (blue), the raw covariance (green) and the dejittered covariance (red). Eigenvalues obtained from estimates of the covariance matrix (red, green) are shown with 95% confidence intervals. The two largest eigenvalues of the raw covariance differ significantly from the corresponding values of the true covariance. Dejittering restores the original spectrum: red and blue values do not differ significantly

8 PC-s account for > 95% of the total variance around the sample mean.

The results from the analysis of this dataset using the diagonal distance function (15) are reported in Fig. 3 in the same format as the results reported for the synthetic data. The obvious exception in the case of an actual sensory system is that the set of true stimuli, mean and covariance are not available, hence the top right panel and some traces in other panels are missing. The panels are labeled consecutively, thus the labels do not correspond to the labels in Fig. 2. As with the model studies in Fig. 2, the raw dataset on Panel A was dejittered to obtain the raster on Panel B. The standard deviation of the jitter was estimated to be $\sigma_t = 1.27$ ms. Unlike the model case, now there is not a true model mean and covariance to which to compare the results of dejittering. However, the waveforms on Panel C follow the general pattern established in the corresponding Panel D of Fig. 2: the raw mean (green) is a blurred version of the dejittered mean; dejittering (red) and deconvolution sharpen its features and in general increase in size. Comparing the top raw

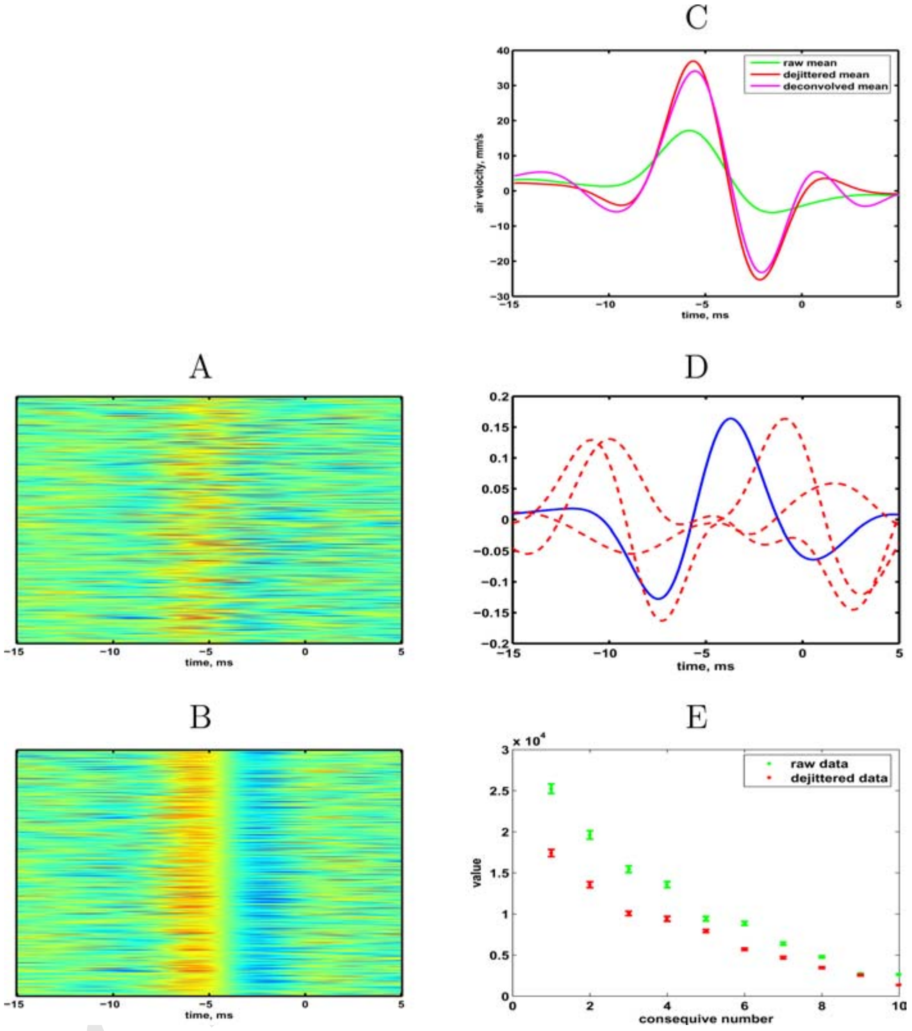

Fig. 3 Effects of temporal jitter on spike-triggered statistics: physiological studies. (A) Rasters of stimulus waveforms preceding isolated single spikes of IN10-3 in the cricket cercal sensory system. The spikes occur at relative time 0 on this plot. (B) The effects of temporal jitter are removed from the raw dataset by using the cost function in Eq. (15). (C) Comparison between the raw mean (green), dejittered mean (red) and deconvolved mean (magenta). The corrected means differ significantly from the raw mean, and agree with one another. (D) Evidence that eigenvectors of the raw covariance can be artefacts of the transformations. In particular, the top eigenvector of the raw covariance C_z (solid blue line) bears no resemblance to any of the top 3 eigenvectors of the dejittered covariance matrix C_x (dashed red lines), which is the most likely estimate of the true covariance. (E) Top 10 eigenvalues of the raw covariance (green) and the dejittered covariance (red). The top eigenvalue of the raw covariance differs significantly from the corresponding value of the dejittered covariance

eigenvector (solid blue) on Panel D to the top three eigenvectors of the dejittered covariance again demonstrates that some of the spectral components of the spike-triggered covariance may be artefacts of temporal jitter. The top 10 eigenvalues of the raw (green) and dejittered (red) covariances in Panel E suggest that here there are a number of eigenvalues that differ significantly (more than 95% level).

There are similarities and differences in the application of the dejittering methods to models and sensory data. Most of the results are quite similar to the ones obtained from our study of synthetic data. This distinctions are manifested in panels C and E of Fig. 3. In Panel C one can notice somewhat larger differences between the mean corrected by deconvolution, and the one recovered by the dejittering procedure. There were essentially no noticeable differences in the corresponding panel of Fig. 2. One possibility is that in the real system there may be more transformations acting on the stimulus, and undoing the effects of one still leaves nontrivial noise sources to affect the mean waveform. Panel E shows multiple eigenvalues differing between the raw and dejittered spectra, compared to two on the corresponding panel of Fig. 2. This highlights the observation that even small levels of jitter ($\sigma_t \approx 1.5$ ms in this case) can lead to large distortions of the conditional covariance spectrum. It still leaves open the possibility that there are more artefacts generated by other transformations.

We again apply the model selection criteria described in C. Positive values in both cases favor the true process model; negative values favor the model of observables. For the physiological observations, the average log likelihood ratio was 1.06 per sample. Since this is a logarithmic measure, it means that on the average, each sample was about 3 times more likely to be explained by the true model than by the raw model. The corresponding average difference of AIC criteria was 2.12, again favoring the true model. To obtain the corresponding values for the whole set of 13,600 observations, the average values have to be multiplied by 13,600, stressing the overwhelming advantage that the true process model has above the model estimated directly on raw observables.

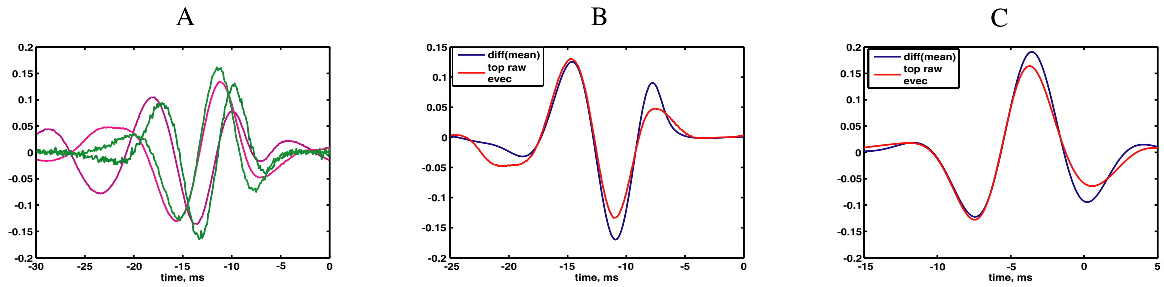

Fig. 4 Evidence for the artefactual origin of the top eigenvectors of the spike-triggered covariance. (A) Similarity between the top two eigenvectors of the raw covariance for data sampled from the diagonal normal model (green lines) and data sampled from the autoregressive normal model (magenta lines). Pairs of eigenvectors are very similar to one another (bright green and bright magenta; dark green and dark magenta), even though the true stimuli have very different covariance matrices and corresponding spectral decomposition. (B) Similarity between the top eigenvector of the raw covariance matrix in Fig. 2 (top raw evec) and the normalized derivative of the true mean (diff(mean)). (C) Similarity between the top eigenvector of the raw covariance matrix in Fig. 3 and the normalized derivative of the dejittered mean

The spectral decomposition of the raw covariance matrix in both figures deserves more attention. As can be seen in Eq. (5), transformations induce artefactual structures in the raw covariance C_z, which are otherwise not present in C_x. In Fig. 4 we present evidence that the top eigenvectors of the raw covariance may be artefactual. In Panel A we compare the top two eigenvectors of the raw covariances obtained from model 1 and model 2. To remind the reader, we sample a set of stimuli from each true model, and shift them by random time $t \propto p(t)$ to obtain raw stimuli. The raw covariances are then estimated from those raw stimuli. Recall that both models have the same true mean, but very different true covariance structures. Model 1 has a spherical covariance structure—the covariance matrix is $C_x = \sigma^2 I$. Thus any vector is an eigenvector of C_x. Model 2 on the other hand has autoregressive covariance, the top three eigenvectors of which were shown in Panel E of Fig. 2. In Panel A we show the top two eigenvectors of the raw covariance for both model 1 and model 2. Despite the big differences in the true covariances, the spectral decomposition of the *raw* covariances derived from those models are strikingly similar. This is a strong indication that these eigenvectors are artefacts of the temporal shifts.

As we discussed in Section 3, when σ_t is relatively small and when C_A in (9) dominates the other terms, the analysis in Appendix A predicts that in the case of temporal jitter the leading eigenvector of C_z is approximately the derivative of the true mean, $\frac{d\bar{x}}{dt}$. We hasten to state that, even though currently the results of the perturbation analysis (9) can explain just the top raw eigenvector, it by no means implies that just a single artefactual eigenvector is generated. Evidence for that is shown in Panel A of Fig. 4, where we see two artefactual eigenvectors, and in panel F of Fig. 2, where two eigenvalues were found to be significantly different from the expected spectrum.

We tested the perturbation assumptions for both model 2 and the cricket data. In the case of the model, the largest eigenvalue of C_x is (approximately) 9 $10^3 \sigma_t$ is set to 15 (= 1.5 ms at 10 kHz sampling rate), the largest eigenvalue of C_{Ax} is 36, the largest eigenvalue of $C^S_{A^2x}$ is 0.11 and the only nonzero eigenvalues of C_A is 115. Therefore, since $\sigma_t^2\|C_A\| \approx 2.6 \times 10^4$, the last term dominates the rest in (9). Currently we cannot estimate analytically for what range of σ_t the approximation (9) is valid. Instead we present the eigenvectors with corresponding normalized derivatives of the mean in Panel B of Fig. 4. For the cricket data the mean and covariance were estimated by deconvolution and dejittering, as outlined above. The largest eigenvalue of C_x was 1.7×10^4, the value of σ_t was 21.5 (2.15 ms at 10 kHz sampling rate), the largest eigenvalue of C_{Ax} was 11.2, the largest eigenvalue of $C^S_{A^2x}$ was 8.2 and the only nonzero eigenvalue of C_A was 234. In this case, as before, the largest eigenvalue of C_x is much smaller than the size of $\sigma_t^2\|C_A\| \approx 1.1 \times 10^5$, and visual inspection of the leading eigenvector of C_z on Panel C reveals that it also strongly resembles the eigenvector of C_A, that is, $\frac{d\bar{x}}{dt}$.

4.3. Analysis of visual processing and spatial jitter: model studies

Extension of this framework and algorithms to two dimensional shifts is straightforward. For related work from the perspective of computer vision the reader should consult (Frey and Jolic, 1999, 2003; Miller and Chef'dhotel, 2003; Miller et al., 2000; Rao and Ruderman, 1999). Here we study the effects of spatial jitter on a model of a simple V1 cell. We use a classic model of simple V1 cells: the Gabor function (Jones and Palmer, 1987; Marcelja, 1980). The model cell has the receptive field (true mean), shown in Fig. 5A, that is a 32 × 32 pixels Gabor wavelet with Gaussian $\sigma = 3.5$ pixels and sine wavelength $k = 2\sqrt{2}\sigma$. We use arbitrary non-dimensional units instead of spatial angle to keep the model general. The noise for the model cell was an independent Gaussian noise with standard deviation σ for each pixel, approximately of the order of the maximum RF value. The data on which the algorithms operated was generated by sampling from this model. Once frames were sampled, they were shifted in the plane by shifts consisting of a horizontal and vertical component, both drawn independently from a normal distribution with mean zero and $\sigma_{x,y} = 2.5$ pixels (spatial jitter).

We report results from the analysis of a model simple visual cell in Fig. 5 and 6. Panel B of Fig. 5 shows the blurring caused by the action of spatial shifts. Panel C demonstrates that this effect can be corrected, in this case by deconvolving the estimate in panel B with the 2-d distribution of spatial shifts.

The spectral analysis of the conditional covariance can also be extended to higher dimensions, with equally important consequences. As mentioned above, the noise model for this model cell was independent for each pixel. Thus the true covariance matrix here is proportional to the unit matrix, and any specific eigen-basis of the estimated covariance would be induced at random by the finite number of samples. However, as can be seen on Fig. 6, the covariance matrix estimated from the raw data has some very specific structures (panels B, D and F there). We can show that some of those structures (the 3 shown here) are generated solely by the action of the transformations on the stimulus. In these cases, the first derivatives of the receptive field in x (A) and y (C), and the second derivative in x (E) matched almost exactly eigenvectors 1, 4 and 3, respectively. The above derivatives emerge from perturbation analysis similar to the one performed for the 1-d case, which is not discussed in detail here.

The first order perturbation analysis result in (9) can provide an approximation to the top eigenvector of the raw covariance. In reality, more eigenvectors and eigenvalues will be affected. For example, in Panel F on Fig. 2, at least **two** eigenvalues are significantly affected, as judged by the eigenvalue spectrum. As we just discussed, in Fig. 6 at least three are affected. The first order expansion presented in the Appendix cannot explain more than one such artefactual eigenvector. However a second- and higher-order expansions can provide further insight in this process when necessary. It bears repeating that the dejittering procedure discussed above, not relying on perturbation analysis, can in principle remove all effects of transformations. The drawbacks there are the increased computational cost of the current implementation of this procedure, and the use of specific models, the choice of which may affect the final results. The practice that we have adopted was to first search for signatures of the transformations in the raw covariance matrix, which is a relatively quick process. If such signatures were found, we applied the dejittering procedure to remove the effects of transformations not just for the top eigenvector, but from the whole ensemble of spike-triggered stimuli.

5. Discussion

Biological sensory systems, and more so individual neurons, do not represent external stimuli exactly. This obvious statement is a consequence of the almost infinite richness of the sensory world compared to the relative paucity of neural resources that are used to represent it. Even if the intrinsic uncertainty present in all biological systems is disregarded, there will always be a many-to-one representation of whole regions of sensory space by indistinguishable neural responses. One direction of research in sensory neuroscience, espoused by us and others, is to identify and model such regions, with the goal of eventually completely describing neural sensory function as the partitioning of sensory space into distinguishable regions, associated to different response states of a sensory system.

In pursuing this agenda, the vastness of sensory space imposes a certain style of analysis that explicitly addresses the problem ensuing from the availability of relatively small datasets with which to provide description of relatively large sensory regions. Typically, response-conditioned stimuli are represented by parametric models with few free parameters. Multivariate Gaussians, characterized by center (mean) and

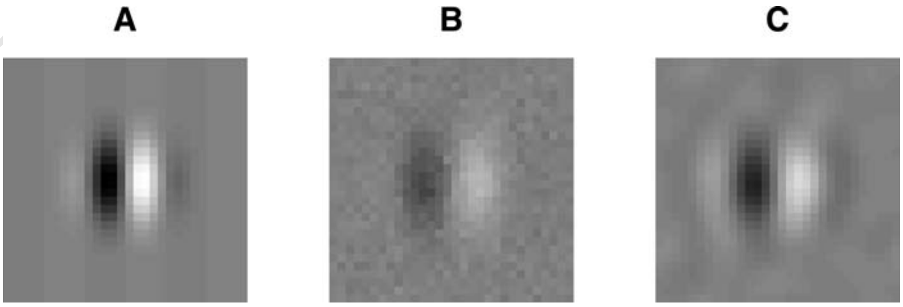

Fig. 5 Effects of spatial jitter on receptive field estimates of a model V1 simple cell. All images are plotted on a common grayscale map. (A) Receptive field of the model V1 simple cell: a Gabor patch with Gaussian spread $\sigma = 3.5$ pixels and sine wavelength $k = 2\sqrt{2}\sigma$. (B) Estimate of the receptive field in the presence of random spatial shifts with $\sigma_{x,y} = 2.5$ pixels. (C) The mean in (B) after deconvolution with a rotationally symmetric Gaussian kernel with $\sigma = 2.5$ pixels is a much better estimate of the true mean in (A)

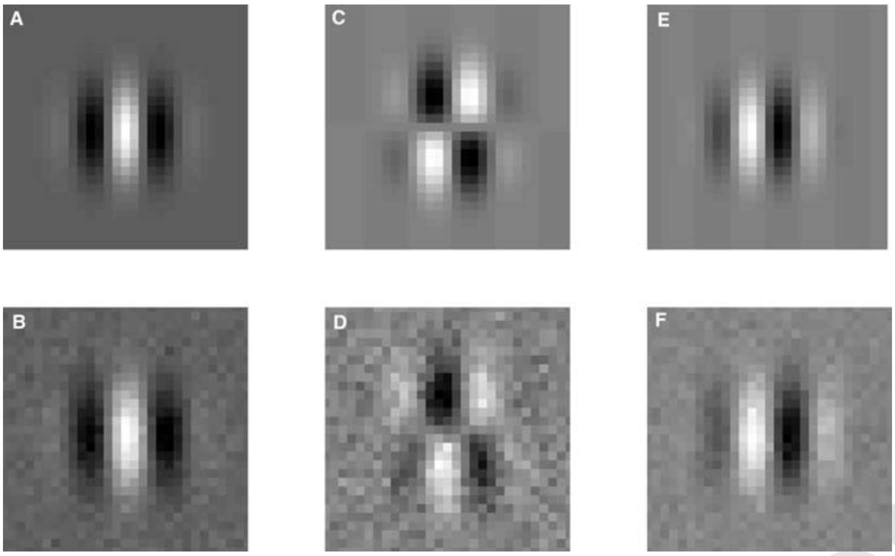

Fig. 6 Evidence that eigenvectors of the raw spatial covariance of the model V1 simple cell can be artefacts due to the presence of random spatial translations. The panels show relations between eigenvectors of the raw stimulus covariance matrix and functions of the receptive field for the model V1 simple cell. All images are plotted on a common grayscale map. On the top row are shown several of the spatial derivatives of the receptive field from Fig. 5A (A) The first horizontal derivative ($\partial/\partial x$); (C) the first vertical derivative ($\partial/\partial y$); (E) the second horizontal derivative ($\partial^2/\partial x^2$). All derivatives were estimated numerically. On the bottom row are shown several of the eigenvectors of the raw covariance matrix. (B) The eigenvector corresponding to the largest eigenvalue; (D) The eigenvector corresponding to the 4th largest eigenvalue. (F) The eigenvector corresponding to the 3rd largest eigenvalue. Eigenvectors and corresponding derivatives are strikingly similar

covariance structure around it, are one such set of models. Once such models are obtained, their parameters are interpreted as neural functions in the context of sensory processing: stimulus features to which the system is selective, or filters and discriminant functions used to represent neural stimulus selectivity.

The analysis presented here provides tools with which to obtain more precise “reverse” models of the sensory regions associated with distinct neural responses. It achieves this by explicitly identifying sources of non-uniqueness and uncertainty in the stimulus, and providing specific models for those sources. This leaves a stimulus residual with smaller variance, which is more likely to be explained by the general parametric models discussed above. Furthermore, parameters of the stimulus models will not be contaminated anymore by the presence of those noise sources. Any interpretation of these parameters in the context of stimulus selectivity will be free of distortions formerly induced by the unaccounted noise sources. So, at the cost of at most a minor increase of model complexity, and possibly a decrease (due to the simplification of the set that needs to be explained), the analytical tools discussed here achieve a much better description response-conditioned stimulus space. Quantitatively, “more precise” refers to the evidence presented here that models which explicitly represent transformations consistently outperform by a sizable margin in both log likelihood ratio and AIC tests equivalent models with implicit representation.

In this work we model some of the effects of uncertainty and non-uniqueness of neural responses as a set of transformations that act on the stimulus and leave the response invariant. We demonstrate how stimulus transformations, when not taken into account explicitly, can bias the estimates of response-conditioned statistics. In particular, we show that the conditional mean is “blurred” with a point-spread function given by the distribution of transformations. The conditional covariance is affected in a more complex manner (5). However, in some special cases we can associate the top eigenvectors of the raw covariance matrix with transformation-induced functions of the conditional mean (temporal, spatial or spatio-temporal derivatives in the case of corresponding shifts). Thus, according to this line of research, such eigenvectors have *no* relation to stimulus selectivity, but are artefacts of the transformations acting on the stimulus. Both of these effects have been confirmed in models and their presence verified with observations in the cricket’s cercal sensory system.

The results we report are also relevant to spike-triggered covariance analysis (Agüera y Arcas and Fairhall, 2003; de Ruyter van Stveninck and Bialek, 1988; Rust et al., 2004; Schwartz et al., 2002; Theunissen et al., 2004), in which special meaning is assigned to eigenvectors of the conditional covariance matrix, whose eigenvalues differ significantly from those of the unconditional stimulus covariance. Here, without referring to the unconditional spectrum, we demonstrated that some of the top conditional eigenvectors may be artefacts of transformations. Moreover, these results seems consistent with eigenvector structures observed for temporal stimuli (Agüera y Arcas and Fairhall, 2003; Agüera y Arcas et al., 2003; Schwartz et al., 2002) and 1-space, 1-time stimuli (Pillow et al., 2003; Rust et al., 2004; Schwartz et al., 2002 Simoncelli et al., 2004), although we have not re-analyzed data from the above publications to confirm this statement. Certainly not all of the structures reported in these articles are due solely to the uncertainty (spatial or temporal) of neural responses. However, when functional significance is attributed to eigenvectors of the covariance, any close similarity between derivatives (spatial or temporal) of the true response-triggered average and eigenvectors of the raw covariance matrix should be studied carefully to avoid possible artefacts due to the processes described above, irrespective of the properties of the unconditional covariance matrix. The work reported by Agüera y Arcas and Fairhall (2003) and Agüera y Arcas et al. (2003) is especially interesting, as such structures appear there despite the fact that the authors used deterministic models in their work, so no biophysical noise sources are present. As we discuss below, the other major source of uncertainty is the major compression performed by the early sensory system, which will generate effective temporal uncertainties that can be modeled as temporal jitter.

Fortunately, in many cases it is not too difficult to remove the action of the transformations and obtain a data set and response-conditioned model that are free of this confounding influence. We propose an iterative algorithm for a set of 1-parametric shifts that selects inverse shifts that are maximally likely under a joint model of stimulus and shifts, $P(x, t)$, and then re-estimates the model to obtain better parameters. In particular, we assume that stimulus and transformations are independent. In cases where the conditional stimulus distribution is not as simple as assumed here (e.g., is bimodal or multi-modal), the method can easily be extended by modeling the stimulus distribution $P(x)$ with a mixture model. Nothing else changes in the formalism of Section 2 except the form of $P(x)$ with which we model the stimulus. Similarly, the distribution of transformations can be modeled with parametric models other than Gaussian when the problem demands it.

The analysis shown here was performed predominantly with the assumption that the action of the transformations on the stimulus is parametrized by a single scalar parameter, t. It can be extended easily to higher dimensional transformations, with essentially identical results. Similar ideas for the more general case of arbitrary affine transformation has been proposed by Frey and Jojic (1999, 2003), for problems in Computer Vision. Both of these cases can also be treated in the common framework of Pattern Theory (Grenander, 1996). Results of the 2-d case shown here are relevant for the analysis of visual systems, especially regarding the concepts of spatial receptive fields (Schwartz et al., 2002), 1-space, 1-time receptive fields (Rust et al., 2004; Theunissen et al., 2004), and spike-triggered covariance analysis (Agüera y Arcase and Fairhall, 2003; Agüera y Arcase et al., 2003; Pillow et al., 2003; Rust et al., 2004, Schwartz et al., 2002; Simoncelli et al., 2004).

Interpretations of the parameters of the transformation noise models depend on the specific problems and sensory systems being analyzed. For example, here we attribute the transformation noise predominantly to biophysical sources, while Aldworth et al. (2005) interpreted the standard deviation of temporal jitter as a mixture of intrinsic biophysical noise and external stimuli leading to variable precision. In the visual system model discussed here, the noise was considered due solely to invariance of the response to such transformations, that is, its source was assumed to have a signal-processing origin. Any of those cases, or a mixture, may be present in a biological sensory system, which makes the parameter interpretation more difficult and problem specific than the actual analytical tools developed here. Furthermore, there are interesting limiting cases—jitter approaching zero, and jitter dominating the variability, that can further complicate the interpretation of these processes. We view temporal jitter as fundamentally different from other transformation-induced noise. In threshold biological systems, many distinct noise sources will manifest themselves at least partially as temporal jitter: any variability in the membrane potential will cause either a delay or speed-up of a spike. Thus, when several types of transformations are considered, temporal jitter may be correlated with other transformation-induced noise. To unravel these effects will require a more detailed noise models. Purely biophysical noise sources can be addressed with the stochastic neuronal models recently developed by Paninski (2004) and Paninski et al. (2005). However, invariance-based and mixed noise sources are beyond the current reach of those types of models. Additional techniques may have to be developed to address such issues as they arise.

Appendix A: Mathematical details

We follow the notation established in the main body of the paper.

Appendix A.1. Effects of transformations on the conditional mean and covariance

We first describe the effects of transformation on the estimate of the conditional mean

$$\bar{x} = E_{p(x)}x \qquad (A.1)$$

as a representative of the cell's stimulus preference. When we compute the raw mean of the observed collection (1), we are estimating

$$\bar{z} = E_{p(z,x,t)} g_t x.$$

Our analysis is based on the following straightforward observation regarding the linearity of expectation:

Lemma 1. *If the action of the transformations g_t is linear, then the transformation commutes with the expectation in x*

$$E_{p(x)} g_t x = g_t E_{p(x)} x. \tag{A.2}$$

The relation between $\bar{z}$ and $\bar{x}$ is addressed in the following

Lemma 2. *Assume that the joint probability factorizes $P(x,t) = p(x)p(t)$ and that the action of transformations g_t is linear. Then*

$$\bar{z} = E_{p(t)} g_t \bar{x}. \tag{A.3}$$

Proof: Since $P(x,t) = p(x)p(t)$, the raw conditional mean $\bar{z}$ can be written as

$$\bar{z} = E_{p(z,x,t)} z = E_{P(x,t)} E_{\mathcal{N}(z;g_t x,\Psi)} z = E_{p(t)} E_{p(x)} g_t x.$$

By (A.2) the last expression is $E_{p(t)} g_t E_{p(x)} x = E_{p(t)} g_t \bar{x}$.

Next we discuss the differences between the true covariance matrix

$$C_x = E_{p(x)}(x - \bar{x})(x - \bar{x})^T \tag{A.4}$$

and the covariance matrix computed from the collection of observations (raw covariance) (1)

$$C_z = E_{p(z,x,t)}(z - \bar{z})(z - \bar{z})^T. \tag{A.5}$$

□

Lemma 3. *Assume that $P(x,t) = p(x)p(t)$ and that transformations g_t act linearly. Then*

$$C_z = \bar{C}_x + C_t + \Psi \tag{A.6}$$

where $\bar{C}_x = E_{p(t)} g_t C_x g_t^T$ *and* $C_t = E_{p(t)}(g_t \bar{x} - \bar{z})(g_t \bar{x} - \bar{z})^T$.

Proof: First write

$$z - \bar{z} = (z - g_t x) + (g_t x - \bar{z})$$

and compute the conditional covariance

$$\begin{aligned}
C_{z|x,t} &= E_{p(z|x,t)}(z - \bar{z})(z - \bar{z})^T \\
&= E_{\mathcal{N}(z;g_t x,\Psi)}((z - g_t x) \\
&\quad + (g_t x - \bar{z}))((z - g_t x) + (g_t x - \bar{z}))^T \\
&= E_{\mathcal{N}(z;g_t x,\Psi)}(z - g_t x)(z - g_t x) \\
&\quad + E_{\mathcal{N}(z;g_t x,\Psi)}(z - g_t x)(g_t x - \bar{z})^T \\
&\quad + E_{\mathcal{N}(z;g_t x,\Psi)}(g_t x - \bar{z})(z - g_t x)^T \\
&\quad + E_{\mathcal{N}(z;g_t x,\Psi)}(g_t x - \bar{z})(g_t x - \bar{z})^T
\end{aligned}$$

The first term here is the instrument noise covariance Ψ. In the second and third terms, $(g_t x - \bar{z})$ is independent of z, and $E_{\mathcal{N}(z;g_t x,\Psi)}(z - g_t x) = 0$ as the expected residual around the mean. Nothing depends on z in the last term, hence

$$\begin{aligned}
C_{z|x,t} &= E_{\mathcal{N}(z;g_t x,\Psi)}(g_t x - \bar{z})(g_t x - \bar{z})^T \\
&= (g_t x - \bar{z})(g_t x - \bar{z})^T + \Psi.
\end{aligned} \tag{A.7}$$

Now consider the expression for C_z (A.5). By (A.7) we can write

$$\begin{aligned}
C_z &= E_{p(z,x,t)}(z - \bar{z})(z - \bar{z})^T \\
&= E_{p(x,t)} E_{p(z|x,t)}(z - \bar{z})(z - \bar{z})^T \\
&= E_{p(x,t)}(C_{z|x,t} + \Psi) \\
&= \Psi + E_{p(x,t)}(g_t x - \bar{z})(g_t x - \bar{z})^T
\end{aligned} \tag{A.8}$$

since Ψ does not depend on $p(x,t)$. Define $\bar{x}_t := g_t \bar{x}$ to be the transformed mean $\bar{x}$. We write

$$g_t x - \bar{z} = (g_t x - \bar{x}_t) + (\bar{x}_t - \bar{z})$$

and compute the last term of (A.8)

$$\begin{aligned}
& E_{p(t)} E_{p(x)}(g_t x - \bar{z})(g_t x - \bar{z})^T \\
&= E_{p(t)} E_{p(x)}(g_t x - \bar{x}_t)(g_t x - \bar{x}_t)^T \\
&\quad + E_{p(t)} E_{p(x)}(\bar{x}_t - \bar{z})(\bar{x}_t - \bar{z})^T \\
&\quad + E_{p(t)} E_{p(x)}(g_t x - \bar{x}_t)(\bar{x}_t - \bar{z})^T \\
&\quad + E_{p(t)} E_{p(x)}(\bar{x}_t - \bar{z})(g_t x - \bar{x}_t)^T
\end{aligned} \tag{A.9}$$

We analyze successively all the terms in this expression. The first term

$$\begin{aligned}&E_{p(t)}E_{p(x)}(g_t x - \bar{x}_t)(g_t x - \bar{x}_t)^T\\ &= E_{p(t)}E_{p(x)}g_t(x-\bar{x})(x-\bar{x})g_t^T\\ &= E_{p(t)}g_t(E_{p(x)}\\ &(x-\bar{x})(x-\bar{x}))g_t^T = \bar{C}_x \qquad \text{(A.10)}\end{aligned}$$

by (A.2). The second term in the expression (A.9) does not depend on x and we can write

$$E_{p(t)}E_{p(x)}(\bar{x}_t-\bar{z})(\bar{x}_t-\bar{z})^T = E_{p(t)}(\bar{x}_t-\bar{z})(\bar{x}_t-\bar{z})^T = C_t. \qquad \text{(A.11)}$$

Finally, we look at the third expression in (A.9)

$$\begin{aligned}&E_{p(t)}E_{p(x)}(g_t x-\bar{x}_t)(\bar{x}_t-\bar{z})^T\\ &\quad = E_{p(t)}E_{p(x)}g_t(x-\bar{x})(\bar{x}_t-\bar{z})^T\\ &\quad = E_{p(t)}\left(\underbrace{E_{p(x)}g_t(x-\bar{x})}_{=0}\right)(\bar{x}_t-\bar{z})^T = 0\end{aligned}$$

where we again used (A.2). An analogous argument applies to the last expression in (A.9).

Combining (A.9), (A.10) and (A.2) the expression (A.8) takes the form

$$C_z = \bar{C}_x + C_t + \Psi.$$

Appendix A.2. Effects of small perturbations on the mean and the covariance

The expression (A.6) that we obtained for the raw covariance matrix is not entirely satisfactory, since it does not allow conclusions about the relationship between eigenvectors and eigenvalues of C_z and C_x. Furthermore, the matrices C_z, $\bar{C}_x$ and C_t in Lemma 3 all depend in a complicated way on the distribution $p(t)$ and the set of transformations $\{g_t\}_{t\in T}$. We wish to simplify the expression for C_z in such a way that this dependence will be on certain characteristics of the set $\{g_t\}_{t\in T}$ and distribution $p(t)$, namely the infinitesimal generator of the set of transformations and the variance σ_t^2 of $p(t)$. In order to this we specialize here to the case which is most often found in applications, where the effect of the transformations $\mathrm{g_t}$ is small, that is, the value of σ_t is small. In other words we assume that $p(t)$ is sharply peaked around its mean, zero. In such case we would like to perform something akin to Taylor expansion of he expressions for $\bar{C}_x$ and C_t on the right-hand side of (A.6), similar to the expansion discussed by Rao and Ruderman (1999) for the purpose of invariant learning. In order to do that we need additional assumptions on the transformations $\mathrm{g_t}$, namely, that the collection $\{g_t\}_{t\in T}$ is a one-dimensional Lie group (Hamermesh, 1962).

Lemma 4. *Assume all assumptions of Lemma 3. In addition assume that the distribution of t is symmetric around zero and that the second moment of this distribution dominates the fourth moment ($\sigma_t^2 \gg E_{p(t)}t^4$). Furthermore, assume that the set of transformations g_t forms a one-dimensional Lie group. Then*

$$C_z \approx C_x + \sigma_t^2\left(C_{Ax} + C_A + \frac{1}{2}(C_{A^2x} + C_{A^2x}^T)\right) + \Psi \qquad \text{(A.12)}$$

where

$$\begin{aligned}C_{Ax} &:= E_{p(x)}A(x-\bar{x})(A(x-\bar{x}))^T\\ C_A &:= (A\bar{x})(A\bar{x})^T,\\ C_{A^2x} &:= E_{p(x)}A^2(x-\bar{x})(x-\bar{x})^T.\end{aligned}$$

This implies that the perturbation to the true covariance matrix is of the order σ_t^2.

Remark 5. The symmetry assumption on the transformation distribution is natural in the context of the problem and it implies that the first and third moments are zero $E_{p(t)}t = E_{p(t)}t^3 = 0$. The assumption that the second moment dominates the fourth moment implies that the t distribution does not have heavy tails. In particular, if the t distribution is normal with zero mean, then $E_{p(t)}t^4 = 3\sigma_t^4$ which satisfies the assumption, since σ_t is small.

Proof: Since the collection $\{g_t\}_{t\in T}$ forms a one-dimensional Lie group, we can write $g_t = e^{At}$, where A is the infinitesimal generator of $\{g_t\}_{t\in T}$. For small t we can approximate

$$g_t \approx I + At + \frac{A^2t^2}{2}, \qquad \text{(A.13)}$$

where I represents the identity transformation. With the approximation (A.13) we write (A.10) as

$$\begin{aligned}\bar{C}_x &= E_{p(t)}E_{p(x)}g_t(x-\bar{x})(x-\bar{x})^T g_t^T\\ &\approx E_{p(x)}E_{p(t)}\\ &\left(I + At + \frac{A^2t^2}{2}\right)(x-\bar{x})(x-\bar{x})^T\left(I + At + \frac{A^2t^2}{2}\right)^T\end{aligned}$$

$$
\begin{aligned}
&= E_{p(x)}E_{p(t)}(x-\bar{x})(x-\bar{x})^T\\
&+E_{p(x)}E_{p(t)}tA(x-\bar{x})(x-\bar{x})^T\\
&+E_{p(x)}E_{p(t)}t(x-\bar{x})(x-\bar{x})^TA^T\\
&+E_{p(x)}E_{p(t)}t^2A(x-\bar{x})(x-\bar{x})^TA^T\\
&+\frac{1}{2}E_{p(x)}E_{p(t)}t^2A^2(x-\bar{x})(x-\bar{x})^T\\
&+\frac{1}{2}E_{p(x)}E_{p(t)}t^2(x-\bar{x})(x-\bar{x})^T(A^2)^T\\
&+\frac{1}{2}E_{p(x)}E_{p(t)}t^3A^2(x-\bar{x})(x-\bar{x})^TA^T\\
&+\frac{1}{2}E_{p(x)}E_{p(t)}t^3A(x-\bar{x})(x-\bar{x})^T(A^2)^T\\
&+\frac{1}{4}E_{p(x)}E_{p(t)}t^4A^2(x-\bar{x})(x-\bar{x})^T(A^2)^T
\end{aligned}
$$

□

We now analyze these expressions one at a time. The first expression is C_x since $E_{p(t)}1 = 1$. The second expression can be rewritten as $(E_{p(t)}t)(E_{p(x)}A(x-\bar{x})(x-\bar{x})^T)$ and the first part is zero by assumption. The same argument applies to the third expression. The fourth expression can be written as

$$\left(E_{p(t)}t^2\right)\left(E_{p(x)}A(x-\bar{x})(A(x-\bar{x}))^T\right) = \sigma_t^2 C_{Ax}.$$

and the fifth is

$$\frac{1}{2}\left(E_{p(t)}t^2\right)\left(E_{p(x)}A^2(x-\bar{x})(x-\bar{x})^T\right) = \frac{1}{2}\sigma_t^2 C_{A^2x}.$$

The sixth term is the transpose of the fifth. By assumption, the cubic terms in t are zero since $E_{p(t)}t^3 = 0$ and the fourth order term is negligible. Therefore

$$\bar{C}_x \approx C_x + \sigma_t^2\left(C_{Ax} + \frac{1}{2}\left(C_{A^2x} + C^T_{A^2x}\right)\right).$$

Now we compute the approximation of the matrix C_t, when we use the approximation (A.13). First observation is that by Lemma 2

$$\bar{z} = E_{p(t)}g_t\bar{x} \approx E_{p(t)}\left(I + At + \frac{A^2t^2}{2}\right)\bar{x} = \bar{x} + \frac{\sigma_t^2}{2}A^2\bar{x}$$

since $E_{p(t)}1 = 1$ and $E_{p(t)}t = 0$. Then, using the fact that the first and third moment vanish and the second moment dominates the fourth, we get

$$
\begin{aligned}
C_t \quad &= E_{p(t)}(g_t\bar{x} - \bar{z})(g_t\bar{x} - \bar{z})^T\\
&\approx E_{p(t)}\left(\left(I + At + \frac{(At)^2}{2}\right)\bar{x} - \bar{x} - \frac{\sigma_t^2}{2}A^2\bar{x}\right)\\
&\times\left(\left(I + At + \frac{(At)^2}{2}\right)\bar{x} - \bar{x} - \frac{\sigma_t^2}{2}A^2\bar{x}\right)^T\\
&= \quad \sigma_t^2(A\bar{x})(A\bar{x})^T + \left(\frac{\sigma_t^4}{4} - \frac{\sigma_t^2}{2}E_{p(t)}t^2 + \frac{1}{4}E_{p(t)}t^4\right)\\
&\times(A^2\bar{x})(A^2\bar{x})^T\\
&\approx \sigma_t^2 C_A.
\end{aligned}
$$

We collect the results $C_z \approx C_x + \sigma_t^2(C_{Ax} + C_A + \frac{1}{2}(C_{A^2x} + C^T_{A^2x})) + \Psi$.

We return to eigenvalue problem with matrix (A.19)

$$(C_x + \epsilon C_A)\zeta = \lambda\zeta \tag{A.14}$$

where we seek a regular expansion of λ and ζ in ϵ

$$\lambda = \lambda_0 + \epsilon\lambda_1 + \cdots, \quad \zeta = \zeta_0 + \epsilon\zeta_1 + \cdots.$$

Plugging these expressions to (A.14) we get the order $O(1)$ equation

$$C_x\zeta_0 = \lambda_0\zeta_0$$

and the order $o(\epsilon)$ equation

$$(C_x - \lambda_0 I)\zeta_1 = -C_A\zeta_0 + \lambda_1\zeta_0. \tag{A.15}$$

Assume that λ_0 is a simple eigenvalue of C_x. The necessary condition for solvability of (A.15) is that the right hand side is in the range of $C_x - \lambda_0 I$. Since the matrix $C_x - \lambda_0 I$ is symmetric, and its kernel is spanned by ζ_0, by the Fredholm alternative the condition of solvability for (A.15) is

$$\langle\zeta_0, -C_A\zeta_0 + \lambda_1\zeta_0\rangle = 0.$$

This yields

$$\lambda_1 = \frac{\langle\zeta_0, C_A\zeta_0\rangle}{\|\zeta_0\|^2} = \frac{\|\zeta_0 \cdot v\|^2}{\|\zeta_0\|^2} \tag{A.16}$$

where we used the fact that $C_A = vv^T$. The Eq. (A.15) for ζ_1 then becomes

$$(C_x - \lambda_0 I)\zeta_1 = -C_A\zeta_0 + \frac{\|\zeta_0 \cdot v\|^2}{\|\zeta_0\|^2}\zeta_0.$$

We first observe that the separation of scales we have used assumes that the eigenvalue λ_0 of matrix C_x is order 1. Assume now that the matrix C_x has a few dominant eigenvalues or order 1 and the rest of the eigenvalues are of order ϵ. This is realistic assumption for C_x a covariance matrix of a spike triggered ensemble in the presence of noise.

Assume further that the projection of v onto the dominant eigenvectors of C_x is of order ϵ. It follows from (A.16) that this assumption implies $\lambda_1 = O(\epsilon^2)$ for a dominant eigenpair (λ_0, ζ_0). Hence these eigenpairs will be perturbed very little by the matrix ϵC_A. On the other hand with this assumption we have that $C_x v = O(\epsilon)$ and thus

$$(C_x + \epsilon C_A)v = C_x v + \epsilon C_A v = O(\epsilon) + \epsilon \|v\|^2 v.$$

Both terms on the right hand side are of the order ϵ. In order for the second term to be of order 1 we must have that $\|v\|^2 = O(\frac{1}{\epsilon})$. Since we assume that the dominant eigenvalues of C_x are of order 1, this means that v must be an order of magnitude larger then the largest eigenvalues of C_x.

In our example from the data this condition is satisfied: the largest eigenvalue of C_x was 1.7×10^4, the value of σ_t was 21.5 (2.15 ms at 10 kHz sampling rate), the largest eigenvalue of C_{Ax} was 11.2, the largest eigenvalue of $C^S_{A^2x}$ was 8.2 and the only nonzero eigenvalue of C_A was 234. Thus the largest eigenvalue of C_x is much smaller than the size of $\sigma_t^2 \|C_A\| \approx 1.1 \times 10^5$. Visual inspection of the leading eigenvector of C_z on Panel C of Fig. 4 reveals that it also strongly resembles the eigenvector of C_A.

Appendix A.3. Analysis of temporal shifts

The expression (A.12) allows us to predict the effect the transformations have on the form and structure of eigenvectors of C_z in certain cases. It follows from (A.12) that the distortion depends on the relative size of the eigenvalues of C_x, the variance σ_t and the eigennvalues of C_A, C_{Ax} and C_{A^2x}. Rather then analyze the general case, we show that in the case when $\{g_t\}$ act as time shifts, and under some additional conditions, one of the leading eigenvectors of C_z resembles a time derivative of the true mean $\bar{x}$.

The approximation (A.13) for temporal uncertainty takes the form

$$\begin{aligned} x(\tau - t) &\approx x(\tau) - \frac{dx}{d\tau}(\tau)t + \frac{d^2x}{(d\tau)^2}(\tau)\frac{t^2}{2} \\ &= \left(I + At + \frac{(At)^2}{2}\right)x(\tau). \end{aligned} \tag{A.17}$$

It follows that the action of the linear operator A is defined by $Au(\theta) := -\frac{du}{d\theta}(\theta)$ and $A^2u(\theta) := \frac{d^2u}{(d\theta)^2}(\theta)$. Since expectations here are computed by integrals for time shifts we will use integrals instead of general expectation notation used in this section so far. In this case (A.12) becomes

$$\begin{aligned} C_z \approx C_x \; &+ \sigma_t^2 \int \left(\frac{d}{dt}(x - \bar{x})\right)\left(\frac{d}{dt}(x - \bar{x})\right)^T p(x)dx \\ &+ \frac{\sigma_t^2}{2} \int \left(\left(\frac{d^2}{(dt)^2}(x - \bar{x})\right)(x - \bar{x})^T \right. \\ &\left. + (x - \bar{x})\left(\frac{d^2}{(dt)^2}(x - \bar{x})\right)^T\right) p(x)dx \\ &+ \sigma_t^2 \left(\frac{d\bar{x}}{dt}\right)\left(\frac{d\bar{x}}{dt}\right)^T + \Psi. \end{aligned} \tag{A.18}$$

Observe that the expressions for C_{Ax} and C_{A^2x} depend on the distribution $p(x)$ and hence will change depending on the problem at hand. Therefore it is very difficult to make general conclusions that would be valid for all such problems. However, in our analysis of the cricket cercal system the norm of these matrices have been an order of magnitude smaller then that of matrix C_A. Therefore we concentrate on a question how the matrix C_A affects the eigenvalue of the matrix perturbation problem

$$C_x + \sigma_t^2 C_A. \tag{A.19}$$

In this analysis we set $\epsilon := \sigma_t^2$ to indicate that σ_t^2 is assumed small. We first analyze the term

$$C_A = \sigma_t^2 \left(\frac{d\bar{x}}{dt}\right)\left(\frac{d\bar{x}}{dt}\right)^T.$$

Notice that this is a matrix of the size $N \times N$ where N is the size of vector $\frac{d\bar{x}}{dt}$, with $N - 1$ dimensional null space and one dimensional range. Let $v := \frac{d\bar{x}}{dt}$. Then

$$\frac{d\bar{x}}{dt}\left(\frac{d\bar{x}}{dt}\right)^T v = \left|\frac{d\bar{x}}{dt}\right|^2 v$$

and so v is the unique eigenvector of $\frac{d\bar{x}}{dt}(\frac{d\bar{x}}{dt})^T$ with eigenvalue $|\frac{d\bar{x}}{dt}|^2$.

Appendix B: Deconvolution parameters

Appendix B.1. Form of the deconvolution kernel

As a starting point we assume that time shifts are distributed with a normal distribution with standard deviation σ_t around a mean spike arrival time: $p(t) = N(t; 0, \sigma_t)$. However, if better models of the shift distribution are available, they can be used instead. For use with the dejittering algorithm, the assumed distribution of time shifts can be modified to match the empirically recovered distribution after dejittering.

Appendix B.2. Regularization parameters

For deconvolution we use standard deconvolution routines from Matlab®'s Image Processing toolbox (deconvwnr, deconvreg). In both cases, a regularization parameters is estimated based on information about signal and noise power in the target to be corrected. In our case, the target is an average of multiple samples, so we have a direct way to estimate signal and noise power. The noise power is estimated as the average (per coordinate) squared standard error of stimulus far from a registered response. This can be estimated directly as $\langle Var(x)\rangle/n$, or computed from know statistical properties of the stimulus (e.g., if a GWN stimulus is generated, the variance of the stimulus can be used). The signal power is estimated as the average (per coordinate) sum of squares in a region where a feature was evident. A single trial will tend to under-estimate the signal power, since it is based on the blurred raw mean. However, this can be amended by performing several re-estimates of the signal power based on results from prior deconvolutions, until a stable estimate of both signal power and deconvolved target is reached.

Appendix C: Model selection

To test whether dejittering improves our understanding of the data, we compare two models on different representation of the observations. The first model is the true process model $p(x)p(t) = \mathcal{N}(x; \bar{x}, C_x)\mathcal{N}(t; 0, \sigma_t)$ (Eq. (12)) in the joint space $X \times T$. This model explicitly takes the transformations into account. The second model is the model of the observables $g_t x = z \in X$ with $p(z) = \mathcal{N}(z; \bar{z}, C_z)$. This operates in a smaller space (X vs $X \times T$) and accounts for the transformations only implicitly, through the covariance matrix in the smaller space. The stimulus portions of the two models ($p(x)$, $p(z)$) have the same dimensionality and number of parameters; the true process model has a single additional parameter: the variance σ_t of the distribution of transformations $\mathcal{N}(t; 0, \sigma_t)$.

To evaluate which of the models explains the observations better, we fit the two models to the equivalent representations $y_i = (x_i, t_i)$ and $y_i = g_{t_i} x_i$ correspondingly. We evaluate the likelihood function $L = \prod_i P(y_i)$ (Krzanowski and Marriott (1995), p. 100) on the same set of observations and then evaluate the log of the likelihood ratio $\log L_{xt}(\{y_i\}) - \log L_z(\{y_i\})$ between the two models. Here $\{y_i\}$ denotes the set of observations. A positive value here implies that the true process model explains the observations better than the model of observables. A negative value implies the reverse. To compare between cases with different number of samples, we report the average (per sample) log likelihood ratio. The actual value can be obtained by multiplying the average ratio by the reported number of samples.

As the first model has one extra parameter, it could be argued that it would be *a priori* favored by the log likelihood ratio test. To address this, we apply Akaike's Information Criterion (AIC) to each model and subtract the observables AIC from the true model AIC. As smaller value of the AIC is indicative of a better model, a positive difference will select the true model, and vice verse. Since AIC criterion for a model with m parameters is defined as (Krzanowski and Marriott (1995), p. 101)

$$AIC(\{y_i\}) = -2 \log L(\{y_i\}) + 2m,$$

for our case the difference $AIC_z - AIC_{xt} = 2(\log L_{xt}(\{y_i\}) - \log L_z(\{y_i\})) - 2$, that is, twice the log likelihood ration minus two. Hence the two criteria yield almost identical results when the number of observations is large. Again, we report the average AIC difference (AIC per sample).

Acknowledgments We thank Zane Aldworth and Melissa A. Sheiko for fruitful discussions and preliminary design for some of the figures, John P. Miller, Liam Paninski and the three anonymous reviewers for critical comments on early versions of this manuscript. The data from the cercal sensory system were graciously provided by Zane Aldworth. This work was partially supported by a grant from US National Science Foundation (BITS-0129895) and NIH-NCRR INBRE grant PR16455 to TG.

References

Agüera y Arcas B, Fairhall AL (2003) What causes a neuron to spike? Neur. Comp. 15: 1789–1807.

Agüera y Arcas B, Fairhall AL, Bialek W (2003) Computation in a single neuron: Hodgkin and Huxley revisited. Neur. Comp. 15: 1715–1749.

Aldworth ZN, Miller JP, Gedeon T, Cummins GI, Dimitrov AG (2005) Dejittered spike-conditioned stimulus waveforms yield improved estimates of neuronal feature sensitivity. J. Neurosci. 25(22): 5323–5332.

Amit Y, Grenander U, Piccioni M (1991) Structural image restoration through deformable templates. JASA 86(414): 376–387.

Bacon JP, Murphey RK (1984) Receptive fields of cricket (acheta domesticus) are determined by their dendritic structure. J. Physiol. (Lond.) 352: 601–613.

Bryant HL, Segundo JP (1976) Spike initiation by transmembrane current: a white-noise analysis. J. Physiol. 260: 279–314.

Chang T-R, Chung P-C, Chiu T-W, Poon PW-F (2005) A new method for adjusting neural response jitter in the STRF obtained by spike-trigger averaging. BioSystems 79: 213–222.

de Ruyter van Steveninck RR, Bialek W (1988) Coding and information transfer in short spike sequences. Proc. Roy. Soc. Lond. B 234: 379–414.

DeAngelis GC, Ohzawa I, Freeman RD (1993) Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. J. Neurophys. 69(14): 1091–1117.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc., B 39(1): 1–38.

Dimitrov AG, Miller JP (2001) Neural coding and decoding: communication channels and quantization. Network: Computation in Neural Systems 12(4): 441–472.

Dimitrov AG, Miller JP, Gedeon T, Aldworth Z, Parker AE (2003) Analysis of neural coding through quantization with an information-based distortion measure. Network: Computation in Neural Systems 14: 151–176.

Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Monographs on Statistics & Applied Probability. Chapman & Hall CRC, New York.

Eggermont JJ, Sersten AM, Johannesma PI (1983) Prediction of the responses of auditory neurons in the midbrain of grass frog based on the spectro-temporal receptive field. Hear. Res. 10: 191–202.

Forte J, Peirce J, Kraft JM, Krauskopf J, Lennie P (2002) Residual eye-movements in macaque and their efects on visual responses of neurons. Vis. Neurosci. 19(1): 31–38.

Frey BJ, Jojic N (1999) Estimating mixture models of images and inferring spatial transformations using the em algorithm. In IEEE Computer Vision and Pattern Recognition, pp. 416–422.

Frey BJ, Jojic N (2003) Transformation-invariant clustering using the em algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(1): 1–17.

Gonzalez RC, Woods RE (1992) Digital Image Processing, Addison-Wesley Publishing Company, Inc.

Grenander U (1963) Probabilities on Algebraic Structures, John Wiley and Sons, qA273.G69.

Grenander U (1996) Elements of Pattern Theory. Johns Hopkins University Press.

Hamermesh M (1962) Group theory and its applications to physical problems. Dover Books on Physics. Dover Publications, Inc., New York.

Jacobs GA, Miller JP, Murphy RK (1986) Cellular mechanisms underlying directional sensitivity of an identi.ed sensory interneuron. J. Neuroscience 6: 2298–2311.

Jones JP, Palmer LA (1987) An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. J. Neurophys. 58: 1233–1258.

Kämper G, Kleindienst H-U (1990) Oscillation of cricket sensory hairs in a low frequency sound field. J. Comp. Physiol. A. 167: 193–200.

Kanou M, Shimozawa TA (1984) Threshold analysis of cricket cercal interneurons by an alternating air-current stimulus. J. Comp. Physiol. A 154: 357–365.

Krzanowski WJ, Marriott FHC (1995) Multivariate Analysis Part 2 Classification, Covariance Structures and Repeated Measurements. Kendall's Library of Statistics 2. Edward Arnold, London.

Mainen ZG, Sejnowski TJ (1995) Reliability of spike timing in neocortical neurons. Science 268(5216): 1503–1506.

Marcelja S (1980) Mathematical description of the responses of simple cortical cells. J. Opt. Soc. Am. A 70: 1297–1300.

Martinez-Conde SL, Macknik SH, Hubel D (2002) The function of bursts of spikes during visual .xation in the awake primate lateral geniculate nucleus and primary visual cortex. Proc Natl Acad Sci USA 99(21): 13920–13925.

Meister M, Pine J, Baylor DA (1994) Multi-neuronal signals from the retina: acquisition and analysis. J. Neurosci. Methods. 51(1): 95–106.

Miller EG, Chef'dhotel C (2003) Practical non-parametric density estimation on a transformation group for vision. In: IEEE Conference on Computer Vision and Pattern Recognition.

Miller EG, Matsakis N, Viola P (2000) Learning from one example through shared densities on transforms. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 464–471.

Miller JP, Jacobs GA, Theunissen FE (1991) Representation of sensory information in the cricket cercal sensory system. I. Response properties of the primary interneurons. J. Neurophys 66: 1680–1689.

Paninski L (2004) Maximum likelihood estimation of cascade point-process neural encoding models. Network 15: 243–262.

Paninski L, Pillow J, Simoncelli E (2005) Maximum likelihood estimation of a stochastic integrate-and-fire neural model. Neur. Comp. 17: 1480–1507.

Pillow JW, Simoncelli EP, Chichilnisky EJ (2003) Characterization of nonlinear spatiotemporal properties of macaque retinal ganglion cells using spike-triggered covariance. In: The Society for Neuroscience Annual Meeting.

Poon PW-F, Yu PP (2000) Spectro-temporal receptive fields of midbrain auditory neurons in the rat obtained with frequency modulated stimulation. Neurosci. Lett. 289: 9–12.

Rao R, Ruderman D (1999) Learning Lie groups for invariant visual preception. In: Kearns, MS, Solla, SA, Cohn, DA eds., Advances in NIPS, Vol. 11, The MIT Press, pp. 810–816.

Reid RC, Alonso, JM (1995) Specifcity of monosynaptic connections from thalamus to visual cortex. Nature 378(6554): 281–284.

Rieke F, Warland D, de Ruyter van Steveninck RR, Bialek W (1997) Spikes: Exploring the neural code, The MIT Press.

Roddey JC, Jacobs GA (1996) Information theoretic analysis of dynamical encoding by filiform mechanoreceptors in the cricket cercal system. J. Neurophysiol. 75: 1365–1376.

Rust NC, Schwartz O, Movshon JA, Simoncelli E (2004) Spiketriggered characterization of excitatory and suppressive stimulus dimensions in monkey V1. Neurocomputing 58–60: 793–799.

Schwartz O, Chichilniksy EJ, Simoncelli EP (2002) Characterizing neural gain control using spike-triggered covariance. In: Dietterich, TG, Becker, S, Ghahramani, Z. eds., Advances in Neural Information Processing Systems, Vol. 14, MIT Press, pp. 269–276.

Simoncelli EP, Paninski L, Pillow J, Schwartz O (2004) Characterization of neural responses with stochastic stimuli. In: Gazzaniga, M Ed., The New Cognitive Neurosciences, 3rd edn., MIT Press.

Theunissen F, Roddey JC, Stu. ebeam S, Clague H, Miller JP (1996) Information theoretic analysis of dynamical encoding by four primary interneurons in the cricket cercal system. J. Neurophysiol. 75: 1345–1364.

Theunissen FE, Woolley SM, Hsu A, Fremouw T (2004) Methods for the analysis of auditory processing in the brain. Ann NY Acad Sci 1016: 187–207.

Victor JD, Purpura K (1997) Metric-space analysis of spike trains: theory, algorithms, and application. Network: Computation in Neural Systems 8: 127–164.