Journal of Mathematical Psychology 82 (2018) 26–37

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Classification images as descriptive statistics

Peter Neri

Laboratoire des Systèmes Perceptifs (CNRS UMR 8248), Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France

HIGHLIGHTS

- Classification images are psychophysical estimates of perceptual mechanisms that resemble 'filters'.
- They are almost invariably connected with human discrimination via a template-matching operation, however the connection is far more opaque than envisaged by this operation.
- Extension to higher-order statistical properties of the classified noise is necessary for adequately constraining potentially underlying circuit models.
- Classification images are best thought of as rich descriptors of data structure, rather than intuitively interpretable snapshots of system components.

ARTICLE INFO

Article history:
Received 20 February 2017
Received in revised form 22 October 2017
Available online 24 November 2017

Keywords:
System identification
Reverse correlation
Perceptual weight
Psychophysics
Sensory processing

ABSTRACT

Classification images have become popular tools in psychophysics, yet difficulties associated with their interpretation have often hindered their application. Alternative methods for characterizing perceptual filters have been proposed, and the discussion has often focussed on the degree to which classification images are optimal statistical estimators of system components (e.g. kernels). This technical note argues that those difficulties become irrelevant once the tool is situated within a data-driven interpretational framework. Within this framework, classification images and their nonlinear derivatives are understood *not* as transparent estimates of system components, but instead as transparent descriptors of data structure. The many pitfalls associated with the former approach, and the power of the latter, are demonstrated via combination of counter-intuitive computer simulations with empirical examples from published literature. A change in perspective over the manner in which this tool is understood and utilized may lead to a more productive engagement with this methodology.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

1.1. What is a classification image, and why is it useful?

We constantly 'filter' the world around us. Our sensors (eyes, ears, nose, tongue, skin) are bombarded by signals of various kinds, and our ability to discriminate between two such signals (e.g. blue versus red colours) relies on perceptual filters that retain one signal and throw out the other. Our brain then exploits the activity of many such filters to perform specific actions for the purpose of successfully interacting with the environment. In its simplest account, this filtering process can be summarized by a trace (bell-shaped curve in Fig. 1A) that records the response of the perceptual filter (plotted on the y axis) to different values of an environmental characteristic, such as the position of an object along the horizon (plotted on the x axis). From Fig. 1A we infer that this specific filter is selective for objects sufficiently close to the middle position along the x axis (stimulus in Fig. 1C), but stops responding when the object is moved further away from the midpoint (Fig. 1D).

The filtering stage outlined above returns a continuous value. Our behavioural decisions, however, do not come in this format: we either decide to run away from a predator, or stay put; we either eat a potentially poisonous food item, or we drop it. In other words, most decisions we take on how we use our sensory representation to interact with the world are discrete (typically binary): we either choose to take an action, or we choose not to. How do we go from our perceptual representation, which comes in the form 'it is 2× more likely that a predator is hiding behind that bush than not', to the decision 'run away!'?

The simplest model of how this conversion may happen involves a threshold (Green & Swets, 1966): if the ratio between the likelihood of 'predator' versus 'non-predator' is greater than some value, e.g. 1, we run away; if it is smaller than that value, we stay put. Because we tend to produce this kind of response somewhat erratically, i.e. our estimate of the likelihood is not always identical under the same environmental conditions due to noise in our sensors and our decisional process (Green, 1964;

E-mail addresses: neri.peter@gmail.com, peter.neri@ens.fr.
URL: https://sites.google.com/site/neripeter/.

https://doi.org/10.1016/j.jmp.2017.10.004
0022-2496/© 2017 Elsevier Inc. All rights reserved.

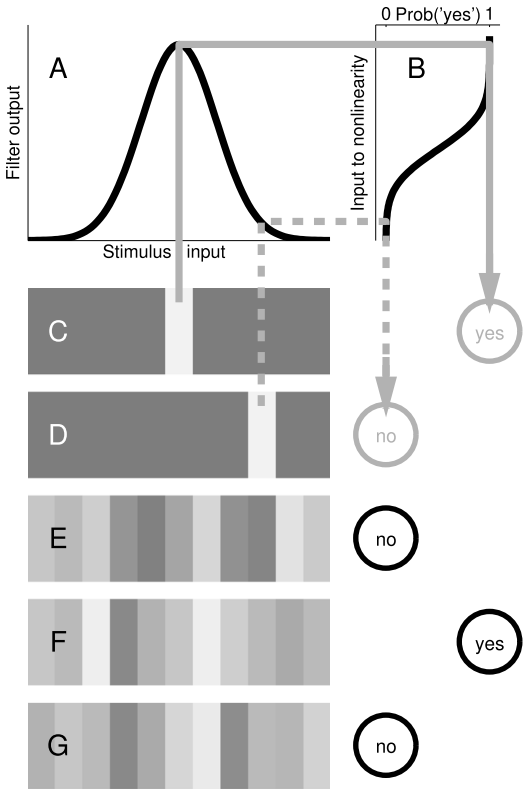

Fig. 1. The linear–nonlinear model consists of a linear filtering stage (A) followed by a nonlinear static nonlinearity (B). The linear filter specifies weights associated with different portions of the incoming stimulus (e.g. different bar positions along the *x* axis) and converts each stimulus into a single decision variable. The static nonlinearity takes the decision variable as its own input, and converts it into a psychophysical choice ('yes' versus 'no' in this example) according to a specified probabilistic rule. Solid/dashed grey lines indicate processing paths through this model for well-matched and mismatched example stimuli (C/D). When stimuli contain multiple bars (E–G), the filter in A acts as a weighting function that sums across bars.

Neri, 2010a), any model of what decision we take must be itself probabilistic: it can only predict that we will run with probability x. To this end, the model in Fig. 1A converts the output from the filter (on the y axis) onto the probability that it will lead to one of two binary choices (e.g. 'yes' versus 'no'). The 'link' function is called a static nonlinearity (an example is shown in Fig. 1B). This function is necessary if one is to re-format the output of the filtering stage into the currency of real-world actions.

The combination of the filtering stage and the static nonlinearity in Fig. 1A–B is termed a 'linear–nonlinear' (LN) model (Murray, 2011; Ostojic & Brunel, 2011): the 'linear' part is the filtering stage (A), the 'nonlinear' part is the static nonlinearity (B). This is a minimal model: anything simpler will not provide a description of sensory processing that is even passable (Neri, 2015). It is therefore understandable that this model is used as a reference point in computational accounts of sensory processing by neurons (Ostojic & Brunel, 2011) as well as observers (Murray, 2011): it is a sensible building block to start with; more complex models can be constructed using blocks of this kind (Carandini et al., 2005) if called for by the phenomenon under study (e.g. Fournier, Monier, Pananceau, & Fregnac, 2011). In particular with relation to the topic discussed in this article, the LN model is often regarded as the theoretical foundation for computing classification images in sensory psychophysics (Ahumada, 2002; Murray, 2011), which brings us to the question: what are classification images?

If we accept the LN model as an adequate representation of the sensory process at hand, say human vision, the classification image is an 'image' of the filtering stage in Fig. 1A: when the underlying sensory filter takes on the shape in Fig. 1A, so will the classification image (Ahumada, 2002). In other words, *if our viewpoint is informed by the LN model*, the classification image technique is a tool for deriving a picture of the filtering stage (Murray, 2011). Why should we want to obtain such a picture?

Classical approaches to sensory processing in animals, e.g. Fechnerian psychophysics, have traditionally emphasized performance: the experimenter focuses on measuring *how well* the animal can detect/discriminate among different signals (Green & Swets, 1966). From these measurements, inferences are sometimes made about the possible shape of the filtering stage, but this is typically achieved via indirect routes (e.g. poorly constrained models with several free parameters) or not at all: the transduction from stimulus to filter output is modelled as a static nonlinear function, effectively incorporating it into the N portion of the LN model and shifting the focus of the investigation onto this component alone (Solomon, 2009). With classification images, the opposite approach is taken: the focus is shifted onto the filtering stage, while the decisional nonlinearity that maps filter output onto choice is bypassed (Neri, 2010b). In this sense, the two approaches are complementary and should be used synergistically whenever possible (Neri, 2011b, 2014b).

There are two critical ingredients that enable this technique to take a snapshot of the filtering stage in a way that is not accessible to e.g. Fechnerian psychophysics. First, the injection of a controlled small perturbation into the stimulus: external noise. To provide a simple example, if observers are asked to discriminate between a bar in the middle (Fig. 1C) and a bar to the side (Fig. 1D), the luminance of each image bar along the x axis may be independently jittered by a random source (see toy examples in Fig. 1E–G). In this way, the output of the filtering stage in the observer's brain will not always be the same in response to the central bar, due to small fluctuations introduced by the added pixel noise; further, the decision taken by the observer on the basis of the filter output will also vary from trial to trial (see 'yes'/'no' responses corresponding to Fig. 1E–G), and those variations will depend on the fluctuations introduced by the noise. Sometimes, the added pixel noise will make a bar-in-the-middle stimulus look very much like a bar-to-the-side stimulus; on those trials, the observer will likely classify the bar-in-the-middle stimulus as bar-to-the-side. On other trials, the added noise will emphasize those features of the bar-in-the-middle stimulus that set it apart from bar-to-the-side; on those trials, the observer will likely classify the bar-in-the-middle stimulus as containing a bar in the middle. The term 'classification' in 'classification images' comes from the classification carried out by the observer as just described.

The addition of noise *per se*, however, is not in itself new: there is a long tradition of using stimulus noise to study its impact on performance (Ahumada, 1987; Pelli & Farell, 1999). The additional ingredient is that, when adding noise, the experimenter keeps track of the *specific* noise sample that was added on every separate perturbation that led to a classification by the observer (Ahumada, 1967; Ahumada & Marken, 1971). This is different than classical approaches where, say, 1000 trials are run at some noise intensity x_1 to measure observer performance p_1 (whatever metric is used to assess it), this process is repeated for different noise intensity values x_2, x_3 and so on, and finally the relationship between x and p becomes the main subject of investigation. In those approaches, the specific noise samples presented during the 1000 trials at intensity x_1 are all lumped into one class without regard for the fact that, on some of those 1000 trials, the specific noise sample that was added to the target signal may have made it easier to detect, while the opposite may have been true for other noise samples in the 1000 trial sequence. In the classification image technique, different noise samples (even if generated by a noise source of

fixed intensity x_1) are treated differently, and are separated into different classes depending on the individual choices generated by the observer in response to those specific samples (no–yes–no sequence in Fig. 1E–G). For example, if the observer responded 'bar-in-the-middle' on the 13th, 17th, and 23rd trials, we average the noise samples presented on those specific trials to obtain a 'classification image' of the typical (i.e. average) noise sample that would lead observers to respond 'bar-in-the-middle'. In this way, we obtain a picture of the filter engaged by the observer to filter out a bar in the middle: we throw noise at it, study how specific noise fluctuations drive specific responses from the observer, and make the connection between the two. In the LN model, the connection is instantiated by the linear filter; therefore, by working out the connection, we are also characterizing the filter itself.

1.2. Should we take the linear–nonlinear model out of the picture?

As outlined in the previous section, classification images are used to characterize the way in which sensory stimuli are processed by humans engaged in specific detection/discrimination tasks (Murray, 2011). They are easy to compute: a human participant is asked to classify stimuli into categories (e.g. 'target present' versus 'target absent'), and simple statistical properties of the stimulus samples associated with the different categories are summarized for inspection (Ahumada, 2002). In a typical example, we may separately average all noise samples associated with false-alarm trials (those on which the observer was presented with a stimulus containing only noise but reported seeing the target), misses, hits and correct rejections; the resulting four averages can be combined via simple rules (adding false alarms and hits, subtracting misses and correct rejections) to obtain a picture of the underlying perceptual mechanism (the filtering stage introduced in the previous section). The accessibility of this procedure has been a benefit and a curse. On the one hand, it has led to its popularity in contemporary psychophysics (Murray, 2011). On the other hand, it has obscured complex issues regarding the interpretability of the outcome, a correct evaluation of which requires far more sophistication than suggested by the simplicity of the rules by which it is obtained (Victor, 2005).

It is not surprising that these difficulties should arise. The initial theoretical drive behind classification images was the linear–nonlinear model introduced in the previous section, also termed 'template matcher' (Brunelli & Poggio, 1997) in the perceptual literature. This is still widely accepted as the appropriate conceptual tool for thinking about classification images (Ahumada, 2002; Murray, 2011; Neri, 2015). There are two fundamental problems with the linear–nonlinear framework and its association with classification images: (1) this framework rarely applies to human sensory processing (Neri, 2010c; Tjan & Nandy, 2006), in fact it may never fully apply (Neri, 2015); (2) when it does not apply (e.g. Abbey & Eckstein, 2006; Neri, 2009, 2011a; Tjan & Nandy, 2006), any inference drawn from classification images that relies on this framework is tentative at best, grossly misleading at worst. In this article it is not only emphasized that classification images *can* be computed without reference to the linear–nonlinear framework, but more importantly that they *should* be computed in this manner: the linear–nonlinear framework should play no role because it represents more of a burden than an aid, and classification images become more powerful tools when relieved of this burden.

The view expressed here is illustrated by the following simple example drawing on familiar concepts. We measure one dimension of a given phenomenon and obtain 1000 measurements. We need to inspect and evaluate these measurements, and we need to communicate them to fellow scientists. Communication to others must be transparent: it must allow the recipient to retain proximity with the raw data within the limits of feasibility afforded by the medium of communication (e.g. article, conference, etc.). For the hypothetical dataset just considered, it is typical to offer surrogate descriptors such as the average value across the 1000 data points. When we report the mean, we do not assume an underlying distribution for the data points: we simply report the mean. Whoever is on the receiving end of our communication will have no difficulty relating to this object, thanks to its simplicity and transparent relationship with the data.

A different approach involves setting up a model of the phenomenon under study, fitting the output of this model to the 1000 measurements, estimating the parameter(s) of the model, and reporting the best-fit parameter(s) in place of the mean. This approach is valuable and often more informative than reporting the mean. If we accept the underlying model and carry out optimal inference of the parameters, it is the approach we should favour. However, it is completely different than reporting a transparent descriptive statistic like the mean. In general, data inspection via the model parameter affords less proximity with the raw data, so that the term 'descriptive statistic' is not entirely appropriate: the best-fit parameter is not a transparent summary of the data; it is meant to convey information about the data in the form of inference, not merely description, and cannot be arrived at without committing to a model. A descriptive statistic like the mean, on the contrary, can be computed regardless, and its relationship with the raw data can be transparently gauged by anyone familiar with the basic notion of averaging.

Classification images should be regarded as descriptive statistics like the mean or the median, rather than estimates of underlying model components. If the goal is to obtain a simple statistical relationship between noise input and response output, there is no simpler way of computing this description than the classification image: the current procedure for computing a classification image is justified on the mere basis of attempting a minimal statistical description, without any need for a model. When viewed from this perspective, several difficulties disappear. In particular, the thorny issue of what relationship a classification image may have with the underlying 'perceptual filter' (Murray, 2011; Neri & Levi, 2006) (whatever that may be) becomes immaterial: the classification image is *not* computed with relation to the (vague) concept of perceptual filter or any similar constructs, it is simply computed using a straightforward combination of basic operations (Ahumada, 2002; Neri, 2010b) (averaging, summing, subtracting) on the raw data for the purpose of providing a transparent description of said data. Specific conclusions about the underlying perceptual process may be inferred from classification images in the same way that one may infer some aspects of a measurement distribution from descriptive statistics like the mean or the median, but this secondary process bears no implication for the validity of the classification image itself.

If we adopt this perspective, the classification image is then a valid description of the data that can be computed regardless of any underlying model, and that retains close proximity with the raw structure of the data. No more, no less. This state of affairs differs from those associated with approaches that, on the surface and in the literature, have been discussed in connection with classification images, like statistical inference of system kernels via generalized linear models or related techniques (Gilkey & Robinson, 1986; Knoblauch & Maloney, 2008; Neri, 2004; Oberfeld, 2008): those approaches necessitate a model, do not retain proximity with the raw data, and do not fall into the category of descriptive statistics. These issues are clarified in the remaining part of the article with the goal of enhancing classification images as effective and reliable tools for understanding human sensory processing.

We conclude this introductory section with a brief reference to the distributed aperture technique (Haig, 1985) (DAT), termed 'Bubbles' in subsequent developments (Gosselin & Schyns, 2001).

This methodology is often discussed alongside classification images (Gosselin & Schyns, 2004; Murray & Gold, 2004), however it is not included in the present article because we believe that it is largely distinct from classification images, and that the reasons for discussing it in conjunction with classification images are more superficial than commonly believed in mainstream literature. In essence, the point of contact between the two techniques is that they both introduce a random perturbation into the stimulus and capitalize on trial-by-trial fluctuations of said perturbation, an approach more generally termed 'molecular psychophysics' (Green, 1964). It should be noted that other techniques rely on these same features (e.g. the double-pass protocol for estimating internal noise, Burgess & Colborne, 1988; Neri, 2010a) yet are rarely discussed alongside classification images (with some exceptions Ahumada, 2002), meaning that this commonality between DAT/Bubbles and image classification is not essential. Beyond this common feature, the differences outdo the commonalities (see below).

In DAT/Bubbles, noise comes in the form of an envelope modulation of the existing signal-to-be-detected (or signal-to-be-discriminated), rather than an independent additive perturbation (Murray & Gold, 2004). For this reason, DAT/Bubbles is better characterized as a form of masking, rather than a variant of classification images (it is telling in this respect that the early DAT literature (Haig, 1985) does not refer to pre-existing published work on classification images (Ahumada, 1967; Ahumada & Marken, 1971; Ahumada, Marken, & Sandusky, 1975); this debatable connection was introduced by later literature (Gosselin & Schyns, 2004; Murray & Gold, 2004)). Similarly to masking, DAT/Bubbles does not retrieve the perceptual template, but rather the intersection between the template and the signal-to-be-detected (Gosselin & Schyns, 2001, 2002). This is different than image classification, where the noisy perturbation can generate features that are not present in the signal-to-be-detected and is therefore in a position to map template characteristics besides those directly targeted by the signal-to-be-detected (Murray, 2011). A further point of departure is represented by the different domains of enquiry that are best investigated by the two methodologies: DAT/Bubbles presents several advantages over noise image classification when studying higher-level perceptual phenomena such as faces (Gosselin & Schyns, 2001; Haig, 1985), while noise image classification may be preferable when studying lower-level phenomena and in particular when attempting computational characterizations of those phenomena (Murray, 2011; Neri, 2015). In conclusion, DAT/Bubbles and noise image classification are distinct methodologies to an extent that the former falls beyond the scope of the present article.

2. Results and discussion

2.1. Defining the scope of the LN characterization adopted here

As mentioned in the Introduction, it is (unfortunately) impossible to discuss classification images (henceforth CI's for brevity) without reference to the linear–nonlinear (LN) model (Murray, 2011; Neri, 2015). To avoid ambiguity in discussing this simple model for application to the examples considered in this article, we briefly define its implementation here in relation to the human observer: if the observer is asked to choose between stimulus $\mathbf{s}^{[1]}$ and stimulus $\mathbf{s}^{[2]}$ (both defined by vectors), the LN model applies filter $\mathbf{f}$ to $\mathbf{s}^{[1]} - \mathbf{s}^{[2]}$ via inner product $\langle \mathbf{s}^{[1]} - \mathbf{s}^{[2]}, \mathbf{f} \rangle$, the output o of this operation (a scalar) is corrupted by the addition of an additive internal noise source (typically Gaussian), followed by conversion to a binary response (e.g. 'stimulus #1' if $o > 0$, 'stimulus #2' if $o < 0$). This formulation is equivalent to that presented in Fig. 1 (see further below for clarification of this point).

The above definition of LN transduction may seem restrictive: it only applies to two-alternative-forced-choice (2AFC) protocols, and its only nonlinearity is the binary conversion at the output. As a matter of fact, most considerations in this article apply to a wider range of protocols and LN-like model architectures; we restrict the definition so that we can discuss specific example implementations of the relevant topics. We choose the 2AFC protocol because it is by far the most common and most reliable protocol for psychophysical characterization: in the class of binary response protocols, the other common alternative is the yes–no single-interval design, however it is well-known that this paradigm suffers from potential confounds associated with bias in the response criterion (Green & Swets, 1966). There are laboratory situations when the experiment of interest cannot be satisfactorily formulated according to a 2AFC design, in which case yes–no paradigms must be chosen; in all other instances, however, the 2AFC design should be given priority to avoid criterion confounds (Neri, 2010b), and for this reason it is given priority here.

Our choice to restrict the nonlinear stage to the output binary conversion is consistent with established literature on this topic (Pritchett & Murray, 2015). It coincides with the linear amplifier model (LAM), the standard tool for discussing template-matching in psychophysics and its connection to CI's (Ahumada, 2002; Murray, 2011; Murray, Bennett, & Sekuler, 2005). It is also consistent with analogous definitions for single neurons, where the nonlinearity is represented by Poisson conversion to spike output (Priebe & Ferster, 2008): if we treat human observers and single neurons as input–output devices within a common framework, spike conversion corresponds to psychophysical choice (Neri, 2010b) (both implemented by a static nonlinearity). Finally, under some conditions the LAM provides an accurate description for the operations of human vision (Neri, 2015): it is not merely an abstract tool, it is also a useful tool for practical applications.

It may appear that the above formulation is distinct from those where the nonlinear function is made explicit and potentially parameterized in a number of different ways: in those formulations (Neri, 2010b; Nykamp & Ringach, 2002), the output is expressed as $\Psi(\langle \mathbf{s}, \mathbf{f} \rangle)$ where Ψ is a static nonlinear function typically sigmoidal in shape. In fact, this formulation largely overlaps with the one we adopt here because Ψ specifies the probability of producing response #1 as opposed to response #2 in a binary response protocol (Neri, 2010b; Nykamp & Ringach, 2002). For the binary conversion model adopted in this article, Ψ is a step function in the absence of internal noise, and a cumulative density function in the presence of additive internal noise. Internal noise was adopted for the simulations in Figs. 3 and 6, not in Fig. 4 (in Fig. 3, it enabled targeting of specific d′ values to match the distributions in Fig. 3C; in Fig. 6, a realistic level of internal noise (1.3 ratio between internal and external noise standard deviation Neri, 2010a) was included to ensure that the simulations could replicate the human data under plausible conditions; in Fig. 4, internal noise had no impact on the simulations (except for reducing reliability) and was therefore omitted). The presence of internal noise affects the shape of Ψ, but besides this change of parameterization for Ψ we do not explore a wide range of specifications because that is unnecessary for the argument put forward in this study: here we wish to demonstrate some interpretational issues associated with classification images under plausible scenarios, and binary conversion is entirely plausible (Pritchett & Murray, 2015). Furthermore, our formulation of the LN model belongs to the same family as those where the nonlinear function is made explicit in the form of a probabilistic transducer between the output of the linear stage and the end binary response, as clarified above (Neri, 2010b; Nykamp & Ringach, 2002).

2.2. The connection (or lack thereof) between sensitivity and the classification image

The LN model is a highly tractable one which we understand in great detail, and for which we can make clear-cut quantitative predictions (Ahumada, 2002). An important result relating to this model and its connection with CI's is that, if the model applies, we can use the CI to predict absolute efficiency (Murray et al., 2005). This connection remains important even when we ignore its quantitative nature (which is what we normally do when thinking about LN models): we can still reason qualitatively about CI's without exact predictions of the associated efficiency. For example, if we know that the signal-to-be-detected is an impulse at position 0 on the screen, and we find a Gaussian-shaped CI centred on 0 with a spread of 0.1° under condition #1, versus a spread of 1° under condition #2, we expect that d′ will be higher under condition #1 as opposed to condition #2 (we are assuming that stimulus SNR does not change between the two conditions). This expectation derives from the way we think about 'filters' or 'templates', and we have developed a manner of reasoning about these objects that is intimately connected with the LN model (Neri & Levi, 2006; Spillmann, 2006).

It is often taken for granted that, even if the LN model does not apply exactly but only approximately, our broad conclusions about the connection between CI's and sensitivity remain valid: we accept that the human observer does not behave like an LN model (Neri, 2015); nevertheless, we maintain that by reasoning about filters and kernels as we normally do, we will still get it right 'more or less' (Neri & Levi, 2006; Spillmann, 2006). We may under or overestimate the amount by which filter #1 should under/outperform filter #2, but the direction of the difference should stay the same: we can still guess which one will do better. As discussed below, this approach can in fact produce spectacularly wrong answers. It is comparable to devising an explanation for the Hermann grid illusion based on centre-surround filters: sometimes it works (Spillmann, 1971; Spillmann, Ransom-Hogg, & Oehler, 1987), other times it is entirely misleading (Geier, Bernath, Hudak, & Sera, 2008).

The connection between the CI and sensitivity (d′) is broken in both directions: from sensitivity to the CI, and from the CI to sensitivity. Here 'broken' means that it is not a mathematical function in that it is not injective even when formulated in its most general sense: if we take the relation in the direction sensitivity→CI, given the set of all possible d′ values and the set of all possible CI's for a given experiment (i.e. specified stimulus and task), we cannot always assign a unique CI to a given d′ value. In other words, there may be two different CI's that produce the same d′ value or, said the other way around, if we are given a certain d′ value and asked to predict what the associated CI will look like, we cannot say, not even in rough terms.

Clearly, such state of affairs greatly reduces the utility of the relation for quantitative purposes. But that in itself would not be hugely problematic, were it not for the fact that the breakdown is so extreme as to render it useless for qualitative reasoning too. The latter statement will be demonstrated by the lack of injective mapping with respect not just to the magnitude of the objects involved, but to their sign: given a positive d′ value, we cannot say whether the CI will look like function $\mathbf{f}$ or its negative image $-\mathbf{f}$. Conversely when taking the relation in the direction CI→sensitivity, given a CI with say positive weight on the target signal, we cannot say that the associated d′ value will be positive — it may be negative, i.e. the observer may perform *below* chance, effectively reporting the non-target stimulus as being the target stimulus. These extreme scenarios may appear puzzling to some readers, but they can be implemented via relatively simple computational schemes.

2.3. Sensitivity→CI is not injective

We start with the intuitively more accessible demonstration: radically different CI's may correspond to identical d′ values. To a limited extent, this result can be demonstrated with LN models alone and without the aid of computer simulations. Consider an input stimulus vector $\mathbf{s}$ with n entries (we chose $n = 5$ in Fig. 2A), each defined by a normally distributed variable. In the non-target stimulus, all entries have mean 0. In the target stimulus, all entries have mean 0 except the kth entry having mean > 0 (this fact is indicated by the upward-pointing arrow above panel A for which $k = 1$). If we choose $\mathbf{f}$ to take value $a > 0$ at the kth entry and value b at all other entries, the corresponding CI will be $\propto \mathbf{f}$ and $d' > 0$ under the LN model. The same d′ value, however, will be returned by replacing b with $-b$ (black versus grey traces in Fig. 2A). It is easy to see why this happens: all entries except k consist of Gaussian noise with 0 mean; whether these entries are weighted positively or negatively is immaterial to the final output r, because noise fluctuates randomly in both directions. This means that signal detectability must remain unchanged (same d′), except the CI will look radically different: consider that in Fig. 2A the black CI is a lowpass filter, while the grey CI is a bandpass filter. And yet, expected d′ is the same.

It is possible to devise infinitely many other examples with similar characteristics. Fig. 2B shows an interesting variant, and one for which we have encountered an approximate empirical instantiation (Neri, 2014b). In this example, the target stimulus contains a signal at the location indicated by the solid arrow, while the non-target stimulus contains a signal of equal magnitude at the location indicated by the dashed arrow. If we choose $\mathbf{f}$ so that it is narrowly tuned around the target signal and broadly tuned around the non-target signal (black trace in Fig. 2B), the associated LN model will return a d′ value identical to that returned by swapping tuning parameters between target and non-target, i.e. by choosing $\mathbf{f}$ so that it is broadly tuned around the target signal and narrowly tuned around the non-target signal (grey trace in Fig. 2B). We encountered a similar scenario in experiments where target and non-target signals consist of orthogonally oriented wavelets, so that they correspond to signals occupying different locations along the orientation axis of a stimulus representation projected across orientation (Fig. 2C). An experimental manipulation that did not involve SNR changes (upside-down image inversion) produced broader tuning around the target signal and sharper tuning around the non-target signal at the level of the retrieved CI (compare black with grey traces in Fig. 2C), without any concomitant change in sensitivity (Neri, 2011b, 2014b) (see inset).

The above examples are interesting, but not as extreme as promised in the previous section: they do not show that a CI shaped like $\mathbf{f}$ and one shaped like $-\mathbf{f}$ are associated with the same d′ value. This is not a minor point, because for example the scenario in Fig. 2A only works if the kth entry into $\mathbf{f}$ (corresponding to the target signal) retains the same sign when we switch between conditions: we can invert sign for entries *away* from the signal-to-be-detected without affecting d′, but we cannot invert sign for weights assigned to the signal. If we change sign, d′ will also change sign. If we do not change sign, under the LN model this means the CI will retain same sign for the kth entry. So we cannot have it both ways (change sign for the kth entry into the CI without changing sign for d′).

To rephrase this concept using simple expressions, the d′ associated with the LN model is $\langle \mathbf{f}, \mathbf{t}^{[1]} - \mathbf{t}^{[0]} \rangle / \sigma_N$ for unit-energy (normalized) filter $\mathbf{f}$ and noise standard deviation σ_N (this term incorporates both external and internal noise sources; more specifically, it is the Pythagorean sum of the standard deviations associated with these two sources, Green & Swets, 1966). For the example we considered, the non-target signal $\mathbf{t}^{[0]}$ is 0 everywhere,

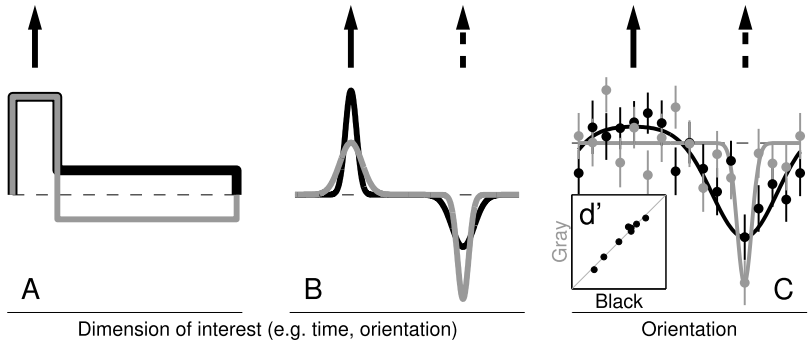

Fig. 2. The black and grey templates in A–B return equal sensitivity for discriminating a target signal injected at the location indicated by the solid arrow as opposed to the non-target signal indicated by the dashed arrow (in A the non-target signal is 0). C shows empirically measured aggregate CI's approximating the scenario in B; inset plots d′ values associated with black/grey CI's (on x/y axes respectively) for individual observers (diagonal line indicates equality). Observers were asked to discriminate between a target signal containing an oriented wavelet with orientation indicated by the solid arrow, and a non-target signal with orientation indicated by the dashed arrow. Orientation noise was added to the stimulus in the form of oriented wavelets spanning the orientation axis and taking on random contrast values. The target orientation was *not* fixed (e.g. vertical); it was instead defined by the local orientation content of a natural scene surrounding the wavelet. The black CI refers to trials on which the natural scene was in upright configuration, while the grey CI refers to trials on which it was inverted upside-down. Readers are referred to the original publication (Neri, 2014b) for further details.

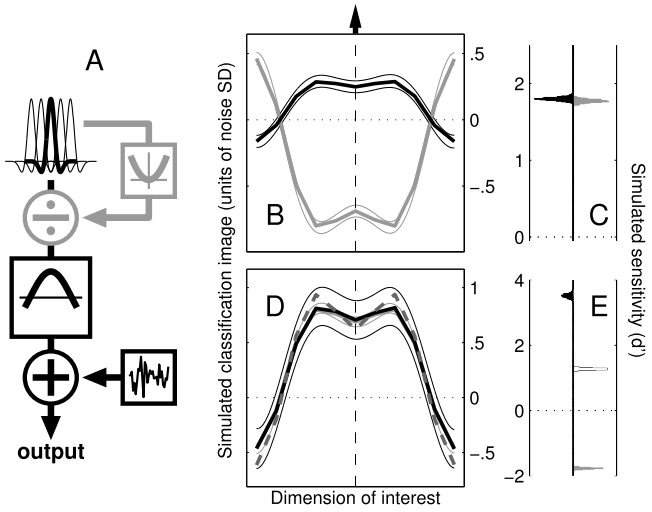

Fig. 3. The grey simulated CI's in B, D were generated by the circuit in A. The circuit generates a decisional variable (output) via front-end convolution (top Mexican-hat shaped functions), linear weighting by function inside large square, addition of late additive internal noise (small square). Gain control was implemented by squaring (grey square) and division (÷ symbol; please see supplied Matlab code for details). Black CI's were generated by a template matcher using the grey CI as template, followed by additive internal noise. The output of this process is similar (though not identical) to the output generated by the black portion of the circuit in A. C, E plot corresponding sensitivity (d′) distributions (across multiple simulations) for detecting the target signal injected at the location indicated by top arrow and vertical dashed line. The main difference between B,C and D,E is that in the latter case the weighting function in A (large square) was sign-inverted for the grey simulation. Dashed trace in D was generated by the same circuit used to generate the grey trace when challenged with stimulus SNR reduced to 1/4; the corresponding d′ distribution is shown by open histogram in E. Thinner lines in B, D show ±1 SD across 1k iterations of 10k trials each.

and the target signal $\mathbf{t}^{[1]}$ is 0 everywhere except its kth entry. The above expression for d′ reduces to $\mathbf{f}(k)\mathbf{t}^{[1]}(k)/\sigma_N$. For a given experiment, the sign of the signal pulse is fixed (say $\mathbf{t}^{[1]}(k) > 0$), so that the sign of the expression for d′ is univocally determined by the sign of $\mathbf{f}(k)$: if we invert the sign of the kth entry into $\mathbf{f}$, the sign of d′ must also change, i.e. we cannot demonstrate the stronger result that $\mathbf{f}$ and $-\mathbf{f}$ are associated with the same d′ value using an LN model.

To demonstrate the stronger result we resort to models outside the LN family. In Fig. 3B, the black CI was generated by an LN model, while the grey CI was generated by a model involving a form of divisive normalization (see diagram in Fig. 3A). More specifically, the input stimulus was first convolved with a Gabor filter $\mathbf{f}$: $\mathbf{r} = \mathbf{f} * \mathbf{s}$ ($\mathbf{r}$ is a vector because transduction of $\mathbf{s}$ through $\mathbf{f}$ is not effected via inner product as in previous sections, but by convolution; $\mathbf{f}$ was specified by a cosine carrier frequency of 2 cycles per whole range of dimension of interest (doi) with 180° phase (negative peak) and a standard deviation of Gaussian envelope equal to 11% of whole range of doi). Each value in the convolution layer was then self-normalized: $\mathbf{r}/(k+\mathbf{r}^2)$, where $k \sim 11\times$ the collective standard deviation of the values in $\mathbf{r}$ across all entries and iterations (i.e. k was fixed, not variable from trial to trial). Finally, the resulting vector was weighted by a sinusoidal weighting function $\mathbf{w}$ to generate the decision variable $o = \langle \mathbf{w}, \mathbf{r} \rangle$ ($\mathbf{w}$ was specified by a cosine function of frequency 0.625 cycles per whole range of the doi peaking at the centre value of the doi). This process was repeated for both target and non-target stimuli to obtain $o^{[1]}$ and $o^{[0]}$ (o drives the psychophysical response according to the usual rule based on the sign of $o^{[1]} - o^{[0]}$: if > 0 the system responds correctly, incorrectly otherwise). The resulting CI (grey trace in Fig. 3B), which we denote $\mathbf{f}^*$, was then sign-inverted and used as linear filter for the LN model that generated the black trace in Fig. 3B: the decision variable was $o = \langle \mathbf{f}^*, \mathbf{s} \rangle + \epsilon$ where ϵ is additive Gaussian internal noise (please refer to Matlab code for specific values used in the simulations).

The two models detailed above generate sign-inverted CI's (Fig. 3B) *and* deliver matched d′ distributions (Fig. 3C). Fig. 3B–C do not merely show that different CI's may be associated with equal sensitivity: the two CI's are not just 'different', rather they are *opposite* across their entire domain. Other similar examples can be constructed; we chose parameters and model structures (Fig. 3A) that are reminiscent of typical components used in the computational literature (Carandini & Heeger, 2011; Heeger, Simoncelli, & Movshon, 1996). These results may seem counter-intuitive to some readers; in the next section we offer an intuitive explanation of why the models behave in this way (see below).

2.4. CI→sensitivity is not injective

Using the model that generated the result in Fig. 3B–C, we demonstrate the equally counter-intuitive result that two identical CI's may be associated with d′ values of opposite sign (this is achieved by simply inverting the sign of the weighting function $\mathbf{w}$ for the grey CI, leading to a pattern of results that is mirror symmetric to that in Fig. 3B). The result is shown in Fig. 3D–E. Again, it is hard to reconcile the grey CI with negative d′, because the CI shows positive weight at the location occupied by the target

signal (indicated by vertical dashed line): intuitively, this should correspond to *positive* d′. And yet the simulations in Fig. 3D–E show that it may correspond to positive or negative d′, undermining typical qualitative thinking about CI's and their relationship to sensitivity (see below). The structure of this simulation is very similar to the one adopted for Fig. 3B–C (see caption to Fig. 3; code is included with this article).

To make this simulated scenario even more counter-intuitive and opaque, the model that generates negative d′ values above is not *structurally* incapable of positively detecting the target: it can also generate positive d′ values, without any change of its parameters. This can be simply achieved by reducing stimulus SNR to a quarter of its previous value, which produces the d′ distribution shown by the black open histogram in Fig. 3E; however, the corresponding CI does not change at all (dashed trace in Fig. 3D). Therefore, by merely changing a stimulus property (SNR) that is routinely manipulated in the laboratory, the same mechanism may produce positive or negative sensitivity while returning identical CI's. The reason for this apparently herratic behaviour is simple: divisive normalization occurs unit-by-unit in the model (see code), essentially reducing to a term of the form $x/(1+x^2)$. This function is non-monotonic, giving rise to the switch from positive to negative sensitivity. It is not at all unreasonable that a computation of this kind may be operating in the human observer (Neri, 2015). The non-monotonic nature of the normalized characteristic also underlies the puzzling result in Fig. 3B: if the input is large enough around the target region, it will inhibit rather than excite filter outputs due to self-normalization; the best stimulus profile for producing positive d′ (grey distribution in Fig. 3C) will then be one where the input is smaller, rather than greater, near the target region. Intuitively, this is the reason for the inverted grey CI shape in Fig. 3B.

In the auditory literature, CI equivalents are usually called 'perceptual weights' (Kortekaas, Buus, & Florentine, 2003; Oberfeld, 2008), and this term is also often used in the vision literature to translate the significance of CI's into more intuitive language (Murray, 2011; Neri & Levi, 2006). It comes with the implied notion that observers place variable weights on different portions of the incoming stimulation, so that some stimulus parts are weighted positively and others are weighted negatively (or ignored when the weight is 0). Fig. 3D–E demonstrate that this notion is problematic: how can observers be placing positive weight on the stimulus region containing the target signal (grey trace in Fig. 3D), and yet consistently report the non-target signal as being the target (grey distribution in Fig. 3E)? There is no sense in which the notion of 'perceptual weight', however loosely one may define it, can be applicable to the scenarios simulated in Fig. 3. This does not necessarily invalidate previous literature that relied on this notion: it is very likely that, for the large majority (if not all) of the phenomena studied by prior work, the notion of perceptual weight was at least approximately applicable. However the simulations in Fig. 3 show that this need not be the case and that, when it is not the case, the notion of perceptual weight may be entirely inadequate.

We emphasize that, under the LN model, the above results cannot be obtained. If we fix $\mathbf{f}$ for a LN model, the resulting CI will be $\propto \mathbf{f}$. There is of course a range of d′ values that are compatible with a given $\mathbf{f}$ due to the presence of internal noise: the larger the internal noise source, the lower the associated d′. Varying internal noise, however, can only modify the *magnitude* of d′, not its sign: if $d' > 0$ in the absence of internal noise, the introduction of internal noise will make it smaller, but not <0. Again, this is a direct consequence of the simple expression detailed above for d′: when the intensity of the internal noise source is modified, σ_N also changes its magnitude but remains positive by definition (it is a standard deviation term), leaving the sign of the expression unaltered. Needless to say, a sign change for d′ carries much greater conceptual significance than a mere amplitude change (Green & Swets, 1966), as it indicates a qualitative transition from a target detector (a system that selects the target signal over the non-target signal) to a non-target detector (a system that selects the non-target signal over the target signal). For this reason, the inability of LN models to account for d′ values of opposite sign is particularly relevant to our present discussion.

2.5. From linear to nonlinear descriptors

If we accept the notion that the CI is first and foremost a descriptive statistic, it becomes natural to enhance it with other related descriptors, just as we may choose to communicate the structure of a dataset by reporting not only its mean but also its spread (in the form of, say, variance). Prior work has extensively demonstrated that second-order CI's (obtained by computing covariance of classified noise samples in place of computing their average Neri, 2004) can be informative about specific perceptual mechanisms beyond the characterization provided by first-order CI's (Neri, 2010b). Rather than replicating relevant literature here, we consider unpublished simulations that offer compelling examples of how inadequate first-order CI's may be for recovering system structure, in line with the above demonstrations of how inadequate they are for predicting sensitivity. These examples also demonstrate the power of full-scale (i.e. not limited to first-order) CI's as descriptive statistics.

For this demonstration we consider two separate noise probes, labelled α and β in Fig. 4 (we offer a concrete example of how this scenario may translate to an experimental setting in Fig. 5). As in previous simulations, on every trial the system is challenged with two stimuli, the target stimulus $\mathbf{s}^{[1]}$ and the non-target stimulus $\mathbf{s}^{[0]}$, but each stimulus now consists of two components $\mathbf{s}_\alpha^{[q]}$ and $\mathbf{s}_\beta^{[q]}$ ($q = 1$ for target stimulus, $q = 0$ for non-target stimulus). For these simulations, a signal-to-be-detected is only added to the α probe for the target stimulus: $\mathbf{s}_\alpha^{[q]} = \mathbf{t}^{[q]} + \mathbf{n}_\alpha^{[q]}$ where $\mathbf{t}^{[1]}$ is zero everywhere except the middle entry set at $1.5\times$ the noise standard deviation and $\mathbf{t}^{[0]} = 0$ everywhere (as before, $\mathbf{n}$ is Gaussian noise). For the β probe $\mathbf{s}_\beta^{[q]} = \mathbf{n}_\beta^{[q]}$ (only noise) on both target ($q = 1$) and non-target ($q = 0$) intervals.

The two probes target two different visual operators (e.g. processing different regions of the visual field), each with a front-end linear filter shaped like a Mexican-hat function defined across some dimension of interest (e.g. orientation). We refer to the two filter functions as $\mathbf{f}_\alpha$ and $\mathbf{f}_\beta$ (e.g. they could be tuning functions characterizing orientation selectivity within two different regions of the visual field), and to their outputs following template matching with the stimulus as $r_\alpha^{[q]}$ and $r_\beta^{[q]}$: $r_\alpha^{[q]} = \langle \mathbf{s}_\alpha^{[q]}, \mathbf{f}_\alpha \rangle$ (if $\mathbf{f}$ is an orientation tuning function, $\mathbf{s}$ would be a vector specifying oriented energy within the stimulus at different orientations). The expression for the β operator is identical except α is replaced by β. Fig. 4A shows CI's (first-order) for the case where the β operator is left out of the decision variable, i.e. $o^{[q]} = r_\alpha^{[q]}$. In this case, there is obviously no measurable tuning for the β probe (grey trace), while the α probe (black trace) returns the front-end filter $\mathbf{f}_\alpha$ as expected for a template matcher (Ahumada, 2002; Murray, 2011).

Fig. 4E shows the tuning functions returned by the circuit immediately to the left of that panel: they are virtually identical to those in Fig. 4A meaning that, on the basis of first-order statistics alone, there is no way of distinguishing between a system solely driven by α and one conforming to the circuit architecture on the second row of Fig. 4 where β is also contributing to the final response. In this circuit arrangement, output from the β operator is added to the output from the α operator, but only after squaring: the decision variable is $o = r_\alpha + r_\beta^2$. As expected, the β operator is

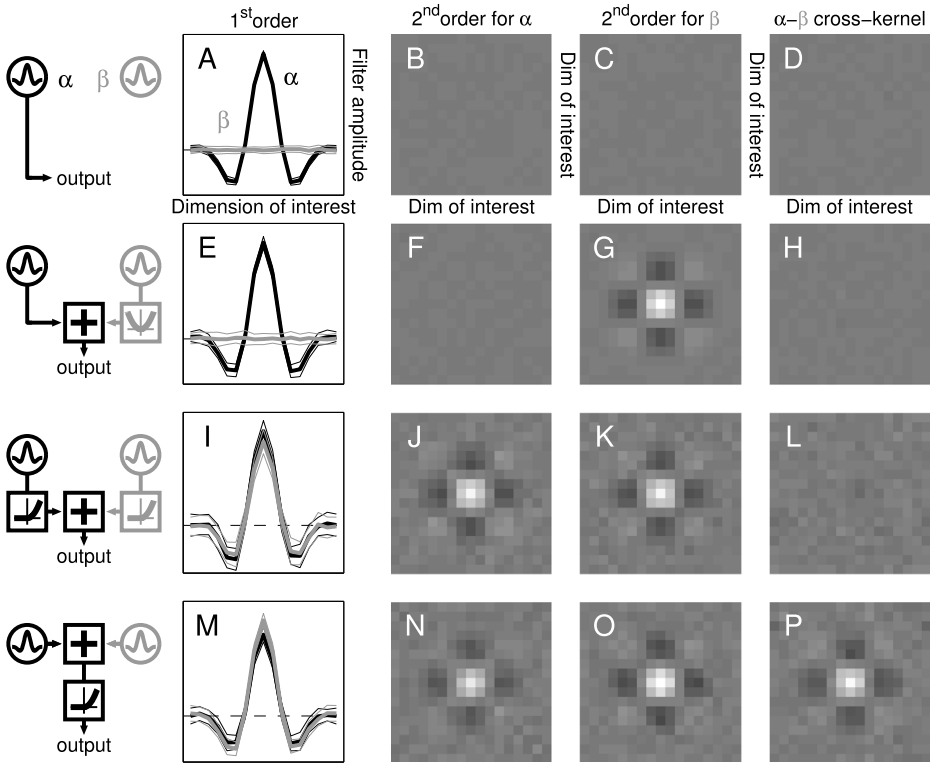

Fig. 4. CI analysis is applied separately to noise samples independently delivered to two front-end Mexican-hat-shaped filters α and β, enclosed within circles in circuit diagrams on the left. A shows first-order CI's (black/grey for α/β) for a model where only output of α unit is used to generate decisional variable. B–C show second-order CI's for the two units separately, while D shows the cross-unit second-order CI. E–H show same for circuit on second row: squaring of β (parabola within square) then sum of α and β; I–L for circuit on third row: static nonlinearity then sum; M–P for circuit on bottom row: sum then static nonlinearity. Thinner lines in A, E, I, M show ± 1 SD across simulations.

transparent to first-order statistics, in the same way that complex-cell-like receptive field structure is not exposed by the spike-triggered average (Schwartz, Pillow, Rust, & Simoncelli, 2006); in this sense, the outcome in Fig. 4E may seem trivial.

The above simulation becomes interesting when we carefully consider some peculiarities of the two-probe design in a psychophysical context. As a starting point, we consider that the analogy with complex-like operators and spike-triggered averaging only applies for a noisy input with no added signal: if a target signal is added to the β probe, it is no longer the case that the resulting CI is flat as in Fig. 4E (Neri, 2010c; Tjan & Nandy, 2006). In a realistic psychophysical context, there are several reasons why adding a signal-to-be-detected is a near necessity (see Neri, 2010b for relevant discussions of this issue). Therefore, if the stimulus only consisted of one probe as in previous simulations, the result in Fig. 4E could not be obtained because the probe would contain a signal-to-be-detected. In the presence of two distinct probes, however, it is entirely conceivable (and empirically observed as in Fig. 5) that observers are asked to detect a signal in one probe only, while their decision is also affected by stimulus modulations in the other probe, despite the latter being devoid of any task-relevant information (i.e. containing pure noise with no added signal). If observers behave like the circuit on the second row of Fig. 4, first-order analysis of the associated CI's would incorrectly lead to the conclusion that the β probe is entirely ignored by their perceptual system.

It is however possible to expose the contribution from the β operator by extending the analysis to second-order CI's (Neri, 2004, 2010b); two such objects, one for the α unit and one for the β unit, are shown in Fig. 4B–C for the scenario where β is irrelevant (top row), and in Fig. 4F–G for the corresponding interaction circuit considered so far. Because under the latter scenario the β unit interacts nonlinearly with the stimulus, the nonlinear descriptor for β in Fig. 4G presents clear modulations that are absent from the corresponding descriptor for the irrelevant probe (Fig. 4C). By adopting nonlinear descriptors we can therefore discriminate between the two scenarios outlined above.

Fig. 4I–P demonstrate that the two circuits on third and fourth rows of Fig. 4 cannot be effectively discriminated using either first-order (Fig. 4I, M) or unit-specific second-order descriptors (Fig. 4J–K, N–O). In this case it becomes necessary to compute an additional nonlinear descriptor that captures the *interaction* between the two units, shown in Fig. 4L, P. Like the second-order kernel, the cross-kernel comes from a simple covariance calculation, the only difference being that in the case of the second-order kernel it is the covariance of input noise samples from one probe (covariance between α and itself, or β and itself), while in the case of the cross-kernel it is the covariance of noise samples from the α probe with noise samples from the β probe (this also explains why second-order kernels are symmetric, while cross-kernels are not). In the circuit on the third row, the two units interact only *after* the nonlinearity: the decisional variable is $o = \exp(r^{[\alpha]}) + \exp(r^{[\beta]})$. Because the nonlinearity ($\exp(.)$) is applied separately to each unit before their outputs are combined, the corresponding interaction kernel (Fig. 4L) shows no sign of nonlinear processing. In the circuit on the fourth row, the two units interact (via sum) *before* the nonlinearity (the decisional variable being $o = \exp(r^{[\alpha]} + r^{[\beta]})$); the interaction kernel (Fig. 4P) therefore retains the corresponding nonlinear signature and allows discrimination between the two models.

2.6. A realistic example of the double probe scenario

The two-probe scenario envisaged by the simulations in Fig. 4 has practical implications for realistic experimental settings. To

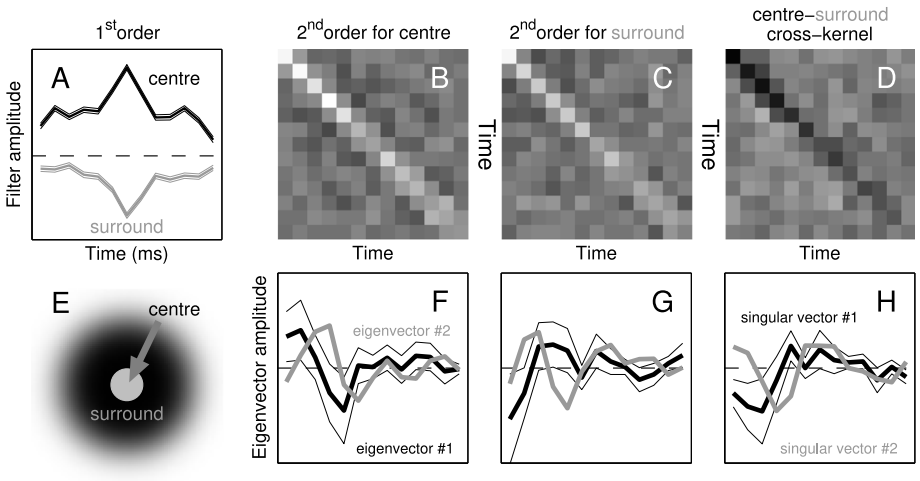

Fig. 5. A–D plot the same descriptors as Fig. 4A–D with α = centre and β = surround for real experiments involving discrimination of a centre-surround stimulus (E). Observers were asked to detect a brief luminance increase in the centre (see main text for details). F–G plot the two eigenvectors of B–C associated with the largest positive eigenvalues; H plots singular vectors of D associated with the largest singular values. Shaded regions show ±1 SD across simulations. Please refer to the original report (Neri, 2009) for further details on stimulus specification.

illustrate this point, we draw from our previous work (Neri, 2009) with centre-surround displays where the luminance values of 'centre' and 'surround' regions (Fig. 5E) were independently modulated by Gaussian noise every 20 ms over a temporal window of 260 ms (13 time points). An increment-to-be-detected was added to the centre region within the middle 20 ms (between 120 and 140 ms), but not to the surround region. Centre and surround therefore correspond to α and β in the above simulations. Fig. 5A–D plot first- and second-order CI's as in Fig. 4A–D, demonstrating that significant modulations can be measured at the level of all CI's. It is possible to construct a relatively simple model that captures all features of these measurements; for details, readers are referred to the original publication (Neri, 2009). We do not consider those specific models here because they are not interesting for the present discussion. However, we will consider one feature of these measurements that is directly relevant to the additional information that may be gained by considering second-order CI's, and that is not transparently available at the level of first-order CI's (see below).

There is a relevant issue pertaining to the visualization of second-order CI's. As we have discussed in previous sections, visualization is central to the argument put forward in this article, because the emphasis here is on retaining the connection with raw data structure while adopting relatively transparent summary descriptors (such as CI's). The issue with second-order CI's is that their dimensionality is twice that of the input stimulus; when stimulus probes are defined across two dimensions (as is often the case Neri & Heeger, 2002; Neri, 2014a), the associated second-order CI's are four-dimensional and cannot be readily visualized. We can address this issue using dimensionality reduction (Fournier et al., 2011; Sandler & Marmarelis, 2015), an approach we demonstrate for the centre-surround second-order CI's. Their dimensionality (2D) is halved via eigenvector decomposition; the resulting eigenvectors (Fig. 5F–H) are therefore directly comparable to the first-order CI's (Fig. 5A). For illustrative purposes we have deliberately chosen a 2D→1D example so that readers can inspect both full and reduced second-order kernels directly and evaluate the connection, however as noted above this approach is more productive for the case of 4D second-order CI's (Fournier et al., 2011).

An obvious difference between first-order CI's (Fig. 5A) and second-order eigenvectors/singular-vectors (Fig. 5F–H) is that the former are primarily lowpass, while the latter are primarily bandpass. As discussed in the original publication (Neri, 2009), it is reasonable to interpret this discrepancy as reflecting an inability of first-order CI's to resolve the known bandpass nature of the underlying front-end filter (Stromeyer & Martini, 2003). This lack of resolving power is likely caused by temporal blurring (see Watson, 1982 for an early example of how the filtering characteristics of average descriptors from perturbation techniques may be misinterpreted (Roufs & Blommaert, 1981) due to temporal convolution). A bandpass temporal impulse response is not only consistent with the structure of second-order CI's, but can also be successfully incorporated into a model that accounts for the empirically measured characteristics of both first-order and second-order CI's (Neri, 2009). In this respect, dimensionality reduction is therefore useful in exposing interesting features of second-order CI's (bandpass characteristics in this example) that are not available from first-order CI's.

It is relevant in this context that eigenvalue decomposition maps to an intuitively graspable model of how eigenvectors combine with the first-order CI to drive system output (Fournier et al., 2014). If $\mathbf{f}$ is the first-order CI, $\mathbf{f}_n^+$ is the nth eigenvector associated with a positive eigenvalue, and $\mathbf{f}_n^-$ is the nth eigenvector associated with a negative eigenvalue, these descriptors can be viewed as components of a model where the final decision variable is $\langle \mathbf{f}, \mathbf{s} \rangle + \sum_{n=1}^{N^+} \langle \mathbf{f}_n^+, \mathbf{s} \rangle^2 - \sum_{n=1}^{N^-} \langle \mathbf{f}_n^-, \mathbf{s} \rangle^2$ where N^+ is the number of statistically significant positive eigenvalues and N^- is the number of significant negative eigenvalues. In words, the eigenvectors (also termed principal dynamic modes in the engineering literature Sandler & Marmarelis, 2015) derived from second-order kernels of an identified system act as linear filters on the input stimulus $\mathbf{s}$ just like the first-order kernel $\mathbf{f}$, except their output is squared before being added/subtracted to the final system output. This formulation naturally incorporates common computational tools such as the energy model (Adelson & Bergen, 1985), which maps to a quadrature pair of eigenvectors. We highlight this connection here because it provides a useful starting point for developing CI-driven computational models, however our favoured approach remains one in which the CI's (whether first- or second-order) are treated as descriptive statistics, not least because second-order psychophysical CI's are biased estimates of second-order system kernels in the Volterra framework (see Neri, 2010b for detailed discussion of the associated distortions), which is the framework that underlies the eigenvector interpretation outlined above (Fournier et al., 2014; Sandler & Marmarelis, 2015).

There is a sense in which it is not surprising that more detailed (higher-order) characterization of relevant statistical structure

may achieve better discrimination among candidate models. The simulations in Fig. 4 and the experimental results in Fig. 5 demonstrate that this notion is applicable to plausible scenarios (Heeger et al., 1996; Neri, 2009, 2015) (e.g. circuits in Fig. 4). We propose that CI's should be treated as descriptive statistics encompassing as much statistical structure as can be reliably characterized with the available data mass (Neri, 2010b).

2.7. Nonlinear can be tricky

We have already shown above (with simulations) that small departures from the LN model make interpretation of first-order CI's uninformative at best (Fig. 4A, E), misleading at worst (Fig. 3). From a theoretical standpoint, the most insidious consequence of introducing computational elements beyond the scope of the LN cascade is their potential interaction with the target signal (Neri, 2010b, c; Tjan & Nandy, 2006). When dealing with neurons, experimenters get away with presenting only noise (neurons respond anyhow), so that an unbiased stochastic perturbation is delivered as input to the system (Ringach & Shapley, 2004). With human participants, this approach is in general not applicable, because humans must be asked to detect some signal if they are to press buttons other than randomly. The experimenter may then choose to present only noise anyhow under the assumption that participants will look out for the specified signal (Gosselin & Schyns, 2003), however in the absence of an objective measure of performance there is no certainty that they are actually performing the assigned task: they may be relying on stimulus cues that are largely disconnected from the specified signal, in fact they may be adopting a processing strategy that would result in negative sensitivity for the specified signal, yet produce CI's that appear to indicate otherwise (as in Fig. 3D–E).

To illustrate with real-world data some of the interpretational issues that may arise in connection with signal-induced distortions of CI estimates under highly nonlinear models, Fig. 6 applies one such model to published measurements of orientation tuning and their dependence on target orientation bandwidth (Taylor, Bennett, & Sekuler, 2014). These experiments have reported that some features of first-order CI's vary as a function of specific characteristics of the signal-to-be-detected, a result that has been interpreted to reflect adjustable bandwidth of the underlying perceptual mechanism (Taylor et al., 2014): when the target signal spans a narrow orientation range, the corresponding CI presents a sharp Mexican-hat profile centred on the target (black traces in Fig. 6A–B); when the signal spans a much broader range, CI's become slightly broader and display signatures of orthogonal inhibition that are not present for the narrow range (grey traces in Fig. 6A–B).

Our simulations can account for these effects (Fig. 6D) using one model architecture (a variant of the uncertainty model, see Fig. 6C) where no parameters are adjusted to accommodate signal manipulations, suggesting that the observed change in CI structure may not in fact reflect any change in the underlying perceptual mechanism. More specifically, the model initially extracts orientation content within the stimulus via an energy detector consisting of two Gabor wavelets in quadrature pair oriented at angle θ: $\langle \mathbf{f}_{even}^{[\theta]}, \mathbf{s}\rangle^2 + \langle \mathbf{f}_{odd}^{[\theta]}, \mathbf{s}\rangle^2$ (wavelets are specified by Gaussian envelope of SD$\sim$6% of image-width and carrier frequency = 6.4 cycles/image-width). This operation is repeated for θ sampling the orientation axis at 8 distinct values, and it is performed in image space ($\mathbf{s}$ is the 2D pixel image of the stimulus that was projected on the monitor, not its power representation). The resulting 8-element vector was then circularly convolved with 3-element vector $(-0.545, 1, -0.545)$, thus implementing inhibition from nearby ($\pm 22.5°$) orientations (represented by top Mexican-hat shaped functions in Fig. 6C). The resulting vector was further

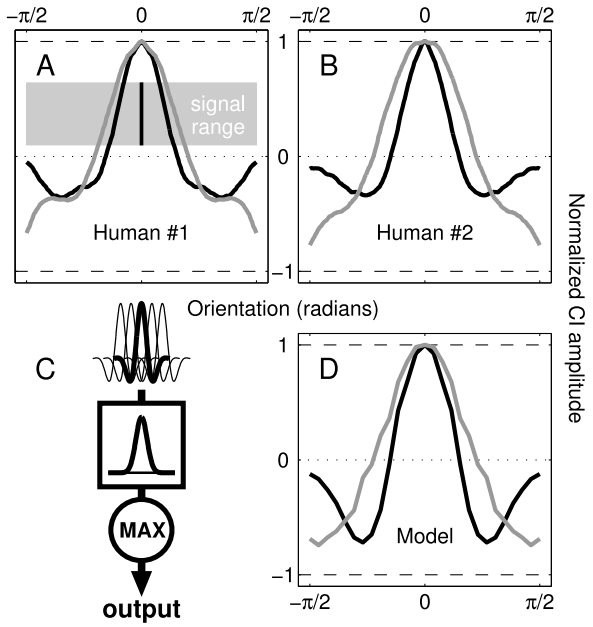

Fig. 6. A–B show empirically derived CI's for orientation tuning from two observers (extracted from Fig. 6 in Taylor et al., 2014); black trace for detecting a target spanning a 2-deg orientation energy centred on 0 (indicated by black rectangular area in A), grey trace for target spanning the entire orientation axis (grey shaded area). D show simulated CI's from the MAX model in C (average across 100 iterations of 2.5k trials each). Although the front-end filters (Mexican-hat-shaped functions at the top) are displayed here with reference to the dimension of orientation, the model operates on 2D stimulus images like those used with the human observers. Following orientation-selective processing by the front-end layer, outputs from orientation channels are weighted by the Gaussian function within the square outline (greater weight around 0) before taking the maximum output (MAX) as decision variable. Please refer to Taylor et al. (2014) for more details on the experiments, and to the main text (as well as the supplied Matlab code) for more details on the simulations.

weighted (multiplied element-by-element) by a narrow (SD = 15°) Gaussian envelope centred on the target orientation (0 on x axis in Fig. 6A–B, D; this stage is represented by the square symbol in Fig. 6C). Finally (but most importantly), the resulting 8-element vector was subjected to a MAX operation (retain maximum value as decisional variable; this stage is represented by the circle in Fig. 6C). The associated CI's are computed by first converting individual noise samples from image space to orientation energy (by extracting power from their spectrum within a wedge region sampling the orientation axis every 7° with wedge width of 22.5°), and then applying image classification to the resulting 1D traces (this procedure is similar to the one adopted for computing human CI's in Fig. 6A–B, Taylor et al., 2014).

How can this model capture the retuning observed experimentally (Fig. 6A–B), as demonstrated in Fig. 6D? The CI for a narrow signal range (indicated by the black rectangular region in Fig. 6A) is expected from previous analyses of the MAX model (Neri, 2010c; Tjan & Nandy, 2006): the MAX operation selects the most active channels, i.e. those in the vicinity of the signal (0 on x axis). These channels produce greater outputs for two reasons: they are driven by the signal (which is confined to a narrow region around 0), but additionally they are favoured by the Gaussian weighting function. The resulting CI (black trace in Fig. 6D) largely reflects the shape of the inhibitory orientation kernel (Mexican-hat shaped function in Fig. 6C). When the signal range is extended to the entire orientation axis (indicated by the grey rectangular region in Fig. 6A), additional filters are recruited by the MAX operation because their output level is increased by the driving signal. This additional contribution from nearby filters, combined with the Gaussian weighting

function and the MAX operation, results in a modified CI (grey trace in Fig. 6D) that resembles the tuning characteristic measured experimentally (Fig. 6A–B).

This interpretation does not call into question the broader conclusions of the original study (Taylor et al., 2014), however it does expose some unexpected difficulties with interpreting CI's as transparently connected with the properties of the underlying sensory system. The connection is often opaque, and is best avoided by viewing these objects as alternative descriptions of data structure. For the specific example considered in Fig. 6, it is not incorrect to *describe* the measured objects (the CI's) as displaying broader tuning in one condition as opposed to another (as was done by the original authors, Taylor et al., 2014): this is a consistent feature of the data, and should be presented as such. The difficulty arises when attempts are made at establishing a *transparent* connection between this experimental feature and the underlying perceptual mechanism: the connection is not transparent, as we demonstrate in Fig. 6C–D by replicating the dynamic experimental feature using a static perceptual mechanism. Related examples of similar scenarios exist in apparently distant areas of neuroscience research, such as specific adaptive properties of fly spiking neurons that may be mistaken as indicative of similarly adaptive characteristics within the mechanism itself (Fairhall, Lewen, Bialek, & de Ruyter Van Steveninck, 2001), when in fact those properties are exhibited by a non-adaptive Reichardt detector (Borst, Flanagin, & Sompolinsky, 2005).

2.8. Final remarks

During the past two decades since the introduction of classification images into vision (Ahumada, 1996), very substantial progress has been made in understanding their relationship to different theoretical and computational constructs (Murray, 2011; Neri, 2010b). This valuable body of knowledge is there to stay, largely unchallenged by the considerations made here. At the same time, an important lesson we have learnt from classification images is that they are most useful when approached as rich descriptions of data structure, rather than transparent images of underlying perceptual mechanisms. We have offered a specific example of how this approach may avoid potentially incorrect interpretations of measured changes in CI structure (Taylor et al., 2014) (Fig. 6). Below we discuss two relevant examples from the broader CI literature where this approach has produced useful developments.

An example of interest is the refined understanding of the often-measured differences between target-present and target-absent CI's (Abbey & Eckstein, 2006; Ahumada et al., 1975; Neri & Heeger, 2002; Solomon, 2002). It was realized in the early stages of psychophysical research on classified noise that those results present significant challenges for the interpretation of CI's, and that the observed differences most likely reflect the operation of a MAX-like operator connected with perceptual uncertainty (Ahumada et al., 1975). However, the underlying theoretical framework was not formalized until more recently, and in particular with two developments in the 2000's: an explicit analysis of the connection between uncertainty models and CI's on the one hand (Tjan & Nandy, 2006), and a general formulation of the connection between nonlinear operators and CI's on the other hand (Neri, 2004). When combined within a unified framework, these two developments lead to a formal theory of the expected differences between target-present and target-absent CI's (Neri, 2010c).

Another pertinent example comes from research on energy operators of the complex-cell-like family. It is well understood that these operators may not display clear modulations at the level of traditional CI measurements (we provide an example in Fig. 4E, grey trace), and early studies of these phenomena attempted to deal with the associated interpretational difficulties via the introduction of novel analytical methods (Neri & Heeger, 2002; Solomon, 2002). The development of such methods has been consistently pursued using a variety of different tools (e.g. Abbey & Eckstein, 2006; Morgenstern & Elder, 2012; Neri, 2011a), equipping us with more principled ways of interpreting the rich datasets returned by noise image classification.

Needless to say, dropping the LN model is not cost-free. If we can assume the LN model, classification images will find the linear template for us with minimal additional assumptions (Ahumada, 2002). Furthermore, we avoid the endless process of searching for the 'perfect' model whereby whenever we device a suitable model, we need to establish how 'good' it is in the sense of fitting the data, a procedure that often involves noisy techniques laden with assumptions of varying complexity (Claeskens & Hjort, 2008). We can establish lower and upper bounds on the maximum predictability of trial-by-trial human responses achievable by the theoretically optimal model (Neri & Levi, 2006), however those bounds are relatively loose and noisiness in the human measurements makes it difficult to be certain that the bound has been saturated. Clearly, the domain of LN models is within our 'comfort' zone in the sense that we understand their properties sufficiently well that we can interpret CI's without much difficulty (Murray, 2011). Unfortunately, existing evidence indicates that LN models are just not appropriate for modelling any visual operation, not even the most elementary one (Neri, 2015). The view advocated in this article is that we can turn the intricacies that arise from renouncing the LN model to our advantage, for example in the identification of specific components underlying nonlinear operators (Abbey & Eckstein, 2006; Beard & Ahumada, 1998; Neri, 2010b, c, 2011a, 2015; Neri & Heeger, 2002; Solomon, 2002).

Classification images are here intended as objects computed using standard statistical descriptors, such as mean (Ahumada, 2002) or covariance (Neri, 2004, 2010b): they should be approached with the same attitude with which we approach those descriptors, no more no less. From such perspective, it becomes immediately clear that there is really no more straightforward and transparent way of summarizing the statistics of a noise-based experiment that are relevant to the observer: if we are to connect noise samples with psychophysical responses, what simpler rule can be devised than the standard combination rule (Ahumada, 2002)? There are of course differences between CI's and the statistical operators used to compute them (e.g. the 'variance' component of a second-order CI can be negative while variance cannot) and those must be handled appropriately, but they do not undermine a transparent connection with the raw data. It is this connection that is vital to the communication and interpretation of empirical results. Further elaboration using computational/theoretical tools is also important (Murray, 2011; Neri, 2010b), but it must not come at the cost of obscuring/misinterpreting data structure.

Acknowledgements

Supported by Centre national de la recherche scientifique (France) and the Agence Nationale de la Recherche (grants ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL and ANR-16-CE28-0016).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmp.2017.10.004.

References

Abbey, C. K., & Eckstein, M. P. (2006). Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer. *Journal of Vision*, *6*, 335–355.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*, 284–299.

Ahumada, A. J. (1987). Putting the visual system noise back in the picture. *Journal of the Optical Society of America A*, *4*(12), 2372–2378.

Ahumada, A. J. (1996). Perceptual classification images from Vernier acuity masked by noise. *Perception*, *26*, 18.

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, *2*, 121–131.

Ahumada, A. J. (1967). Detection of tones masked by noise: A comparison of human observers with digital-computer-simulated energy detectors of varying bandwidths, Technical Report (Human Communications Laboratory, Department of Psychology, UCLA) 29.

Ahumada, A., & Marken, R. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America*, *49*, 1751–1756.

Ahumada, A. J., Marken, R., & Sandusky, A. (1975). Time and frequency analyses of auditory signal detection. *Journal of the Optical Society of America A (Optics, Image Science and Vision)*, *57*, 385–390.

Beard, B. L., & Ahumada, A. J. (1998). A technique to extract relevant image features for visual tasks. *Proceedings of SPIE*, *3299*, 79–85.

Borst, A., Flanagin, V. L., & Sompolinsky, H. (2005). Adaptation without parameter change: Dynamic gain control in motion detection. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 6172–6176.

Brunelli, R., & Poggio, T. (1997). Template matching: matched spatial filters and beyond. *Pattern Recognition*, *30*, 751–768.

Burgess, A. E., & Colborne, B. (1988). Visual signal detection. IV. Observer inconsistency. *Journal of the Optical Society of America A*, *5*, 617–627.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., et al. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, *25*, 10577–10597.

Carandini, M., & Heeger, D. J. (2011). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*, 51–62.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge, UK: Cambridge University Press.

Fairhall, A. L., Lewen, G. D., Bialek, W., & de Ruyter Van Steveninck, R. R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, *412*(6849), 787–792.

Fournier, J., Monier, C., Levy, M., Marre, O., Sari, K., Kisvarday, Z. F., et al. (2014). Hidden complexity of synaptic receptive fields in cat V1. *Journal of Neuroscience*, *34*(16), 5515–5528.

Fournier, J., Monier, C., Pananceau, M., & Fregnac, Y. (2011). Adaptation of the simple or complex nature of V1 receptive fields to visual statistics. *Nature Neuroscience*, *14*, 1053–1060.

Geier, J., Bernath, L., Hudak, M., & Sera, L. (2008). Straightness as the main factor of the Hermann grid illusion. *Perception*, *37*(5), 651–665.

Gilkey, R. H., & Robinson, D. E. (1986). Models of auditory masking: a molecular psychophysical approach. *Journal of the Acoustical Society of America*, *79*(5), 1499–1510.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261–2271.

Gosselin, F., & Schyns, P. G. (2002). RAP: a new framework for visual categorization. *Trends in Cognitive Sciences*, *6*(2), 70–77.

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science*, *14*, 505–509.

Gosselin, F., & Schyns, P. G. (2004). No troubles with bubbles: a reply to Murray and Gold. *Vision Research*, *44*(5), 471–477.

Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological Review*, *71*, 392–407.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Haig, N. D. (1985). How faces differ–a new comparative technique. *Perception*, *14*(5), 601–615.

Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (1996). Computational models of cortical visual processing. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 623–627.

Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision*, *8*(16), 1–19.

Kortekaas, R., Buus, S., & Florentine, M. (2003). Perceptual weights in auditory level discrimination. *Journal of the Acoustical Society of America*, *113*(6), 3306–3322.

Morgenstern, Y., & Elder, J. H. (2012). Local visual energy mechanisms revealed by detection of global patterns. *Journal of Neuroscience*, *32*(11), 3679–3696.

Murray, R. F. (2011). Classification images: A review. *Journal of Vision*, *11*(5), 1–25. http://dx.doi.org/10.1167/11.5.2.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2005). Classification images predict absolute efficiency. *Journal of Vision*, *5*, 139–149.

Murray, R. F., & Gold, J. M. (2004). Troubles with bubbles. *Vision Research*, *44*(5), 461–470.

Neri, P. (2004). Estimation of nonlinear psychophysical kernels. *Journal of Vision*, *4*, 82–91.

Neri, P. (2009). Nonlinear characterization of a simple process in human vision. *Journal of Vision*, *9*, 1–29.

Neri, P. (2010a). How inherently noisy is human sensory processing? *Psychonomic Bulletin & Review*, *17*, 802–808.

Neri, P. (2010b). Stochastic characterization of small-scale algorithms for human sensory processing. *Chaos*, *20*, 045118.

Neri, P. (2010c). Visual detection under uncertainty operates via an early static, not late dynamic, non-linearity. *Frontiers in Computational Neuroscience*, *4*, 151.

Neri, P. (2011a). Coarse to fine dynamics of monocular and binocular processing in human pattern vision. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 10726–10731.

Neri, P. (2011b). Global properties of natural scenes shape local properties of human edge detectors. *Frontiers in Psychology*, *2*, 172.

Neri, P. (2014a). Dynamic engagement of human motion detectors across space–time coordinates. *Journal of Neuroscience*, *34*(25), 8449–8461.

Neri, P. (2014b). Semantic control of feature extraction from natural scenes. *Journal of Neuroscience*, *34*(6), 2374–2388.

Neri, P. (2015). The elementary operations of human vision are not reducible to template matching. *PLoS Computational Biology*, *11*(11), e1004499.

Neri, P., & Heeger, D. J. (2002). Spatiotemporal mechanisms for detecting and identifying image features in human vision. *Nature Neuroscience*, *5*, 812–816.

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research*, *46*, 2465–2474.

Nykamp, D. Q., & Ringach, D. L. (2002). Full identification of a linear-nonlinear system via cross-correlation analysis. *Journal of Vision*, *2*, 1–11.

Oberfeld, D. (2008). Does a rhythmic context have an effect on perceptual weights in auditory intensity processing? *Canadian Journal of Experimental Psychology*, *62*(1), 24–32.

Ostojic, S., & Brunel, N. (2011). From spiking neuron models to linear-nonlinear models. *PLoS Computational Biology*, *7*(1), e1001056.

Pelli, D. G., & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America A (Optics, Image Science and Vision)*, *16*(3), 647–653.

Priebe, N. J., & Ferster, D. (2008). Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron*, *57*, 482–497.

Pritchett, L. M., & Murray, R. F. (2015). Classification images reveal decision variables and strategies in forced choice tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(23), 7321–7326.

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*, 147–166.

Roufs, J. A., & Blommaert, F. J. (1981). Temporal impulse and step responses of the human eye obtained psychophysically by means of a drift-correcting perturbation technique. *Vision Research*, *21*(8), 1203–1221.

Sandler, R. A., & Marmarelis, V. Z. (2015). Understanding spike-triggered covariance using Wiener theory for receptive field identification. *Journal of Vision*, *15*(9), 16.

Schwartz, O., Pillow, J. W., Rust, N. C., & Simoncelli, E. P. (2006). Spike-triggered neural characterization. *Journal of Vision*, *6*(4), 484–507.

Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision*, *2*, 105–120.

Solomon, J. A. (2009). The history of dipper functions. *Attention, Perception, and Psychophysics*, *71*, 435–443.

Spillmann, L. (1971). Foveal perceptive fields in the human visual system measured with simultaneous contrast in grids and bars. *Pflugers Archiv*, *326*, 281–299.

Spillmann, L. (2006). From perceptive fields to Gestalt. *Progress in Brain Research*, *155*, 67–92.

Spillmann, L., Ransom-Hogg, A., & Oehler, R. (1987). A comparison of perceptive and receptive fields in man and monkey. *Human Neurobiology*, *6*(1), 51–62.

Stromeyer, C. F., & Martini, P. (2003). Human temporal impulse response speeds up with increased stimulus contrast. *Vision Research*, *43*(3), 285–298.

Taylor, C. P., Bennett, P. J., & Sekuler, A. B. (2014). Evidence for adjustable bandwidth orientation channels. *Frontiers in Psychology*, *5*, 578.

Tjan, B. S., & Nandy, A. S. (2006). Classification images with uncertainty. *Journal of Vision*, *6*, 387–413.

Victor, J. D. (2005). Analyzing receptive fields, classification images and functional images: challenges with opportunities for synergy. *Nature Neuroscience*, *8*, 1651–1656.

Watson, A. B. (1982). Derivation of the impulse response: comments on the method of Roufs and Blommaert. *Vision Research*, *22*(10), 1335–1337.