ForumAcusticum
2 0 0 5

# Neuronal Predictions of Sparse Linear Representations

Barak A. Pearlmutter
National University of Ireland Maynooth, Co. Kildare, Ireland, e-mail: **barak@cs.nuim.ie**,

Hiroki Asari, Anthony M. Zador
Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, NY 11724, USA, e-mail: **{asari,zador}@cshl.edu**

A striking feature of many sensory processing problems is that there appear to be many more neurons engaged in the internal representations of the signal than in its transduction. For example, humans have about 30,000 cochlear neurons, but at least a thousand times as many neurons in the auditory cortex. Such apparently redundant internal representations have sometimes been proposed as necessary to overcome neuronal noise. We instead posit that they directly subserve computations of interest. We first review how sparse overcomplete linear representations can be used for source separation, using a particularly difficult case, the HRTF cue (the differential filtering imposed on a source by its path from its origin to the cochlea) as an example. We then explore some robust and generic predictions about neuronal representations that follow from taking sparse linear representations as a model of neuronal sensory processing.

## 1 Sparse Separation

For expository purposes, we will review sparse separation [1, 2] in the context of just one sort of cue—that provided by the differential filtering (HRTF) imposed on a source by its path from its origin in space to the cochlea [3]. For this example, all sounds from a given position are defined to belong to the same source, and any sounds from a different position are defined to belong to different sources. We will focus on the separation problem, and assume that source localisation occurs by other mechanisms.

Consider $N$ acoustic sources $x_i(t)$, for $i = 1, \ldots, N$, located at known distinct positions. Associated with each position is a known "HRTF" filter $h_i(t)$. The signal at the ear and consists of a superposition of the filtered sources

$$y(t) = \sum_{i=1}^{N} h_i(t) * x_i(t) = \sum_{i=1}^{N} \tilde{x}_i(t) \tag{1}$$

where $*$ indicates convolution and $\tilde{x}_i(t) \equiv h_i(t) * x_i(t)$ is the post-filter $i^{\text{th}}$ source in isolation. Our goal is to recover the $x_i(t)$ from $y(t)$, using knowledge of the $h_i(t)$.

We now invoke "sparseness", and assume that each source can be expressed as a linear combination of a potentially overcomplete, and not necessarily orthogonal, set of signal dictionary elements $q_j(t)$,

$$x_i(t) = \sum_j c_{ij}\, q_j(t). \tag{2}$$

Under our sparseness assumption, we can recover

$$\mathbf{c}_i = \arg\min_{\mathbf{c}_i} \|\mathbf{c}_i\|_1 \text{ subject to } \mathbf{Q}\,\mathbf{c}_i = \mathbf{x}_i \tag{3}$$

where the $q_j$ form the columns of signal dictionary $\mathbf{Q}$, and $\mathbf{c}_i$ and $\mathbf{x}_i$ are column vectors holding the elements of $c_{ij}$ and $x_i(t)$. This is a convex optimisation problem which is often solved using linear programming.

We now build a new signal dictionary $\mathbf{D}$ consisting of all the elements $q_j(t)$ filtered by each filter $h_i(t)$. Column $ij$ of $\mathbf{D}$ is constructed by convolution.

$$d_{ij}(t) = h_i(t) * q_j(t) \tag{4}$$

The signal received at the ear can be decomposed as

$$y(t) = \sum_i h_i(t) * x_i(t) = \sum_{ij} c_{ij}\, d_{ij}(t) \tag{5}$$

and all the coefficients $c_{ij}$ recovered jointly using

$$\mathbf{c} = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 \text{ subject to } \mathbf{D}\mathbf{c} = \mathbf{y} \tag{6}$$

where $\mathbf{c}$ is a single column vector consisting of all the coefficients $c_{ij}$, and $\mathbf{y}$ is a column vector holding $y(t)$.

The recovered coefficients can be seen both as representing the separated contributions of the individual sources in sensor space (as received at the ear),

$$\tilde{x}_i(t) = \sum_j c_{ij} d_{ij}(t) \tag{7}$$

and as representing the separated sources source space, reconstructed using Eq. 2. Thus the procedure not only performs unmixing (separation), but also unfiltering (deconvolution).

This framework can exploit other cues. Let us consider one: the binaural cues of differential attenuation and latency. To incorporate these, each $h_i(t)$ function is made single-input two-output, and the lengths of the column vectors $\mathbf{d}_{ij}$ the observation vector $\mathbf{y}$ are doubled.

Eq. 6 can be sensitive to noise. Fortunately a noise model can be added while keeping the problem convex. In particular, we can assume that the total amount of noise is

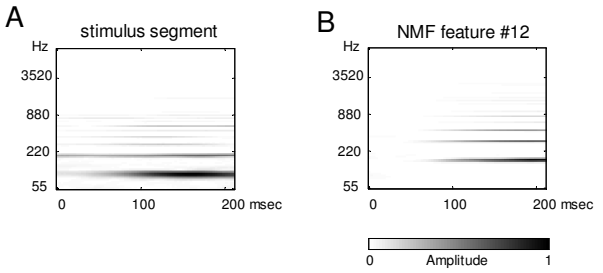


Figure 1: **Non-negative matrix factorisation (NMF) decomposition of musical sources. (A)** Spectrogram of a brief segment of a cello. Each musical piece was broken into an ensemble of such segments. **(B)** NMF was used to find a compact representation of the ensemble in (A). The spectrogram of a sample element is shown here. Note that power is concentrated in the fundamental frequency, reflecting a musical note, but that higher harmonics are clearly visible. When all elements were played in sequence, the elements sound like a crude musical scale. Similar ensembles were computed for other instruments (harp, violin, *etc.*) Note that this element, which reflects statistical correlations present in the sources, is an example of $q_j(t)$ defined in Eq. 2; it is the filtered versions $d_{ij}(t)$ that determine the $c_{ij}$ via Eq. 6.

bounded, resulting in the reformulation

$$\arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 \text{ subject to } \|\mathbf{Dc} - \mathbf{y}\|_p \leq \beta \qquad (8)$$

where $\beta$ is proportional to the noise level and with $p = 1$, 2, or $\infty$. The Gaussian noise case, $p = 2$, can be solved by semidefinite programming methods. Both $p = 1$ and $p = \infty$ can be solved using linear programming. All approaches yield qualitatively similar results. The solutions presented here all used $p = 1$.

# 2 Sample Application

We use the above procedure to separate acoustic sources consisting of mixtures of music, natural sounds and speech. This requires a signal dictionary $q_j(t)$, which is a problem outside the scope of this paper.

## 2.1 Finding a Signal Dictionary using NMF

Finding good overcomplete dictionaries from samples of a stimulus ensemble is a subject of ongoing research [4]. We used nonnegative matrix factorisation (NMF) to generate a set of basis features from spectrograms obtained from samples of solo instrumental music, natural sounds and speech (Fig. 1). NMF is an algorithm for factoring a data matrix—a matrix whose columns contain the snippets of solos—under non-negativity constraints [5]

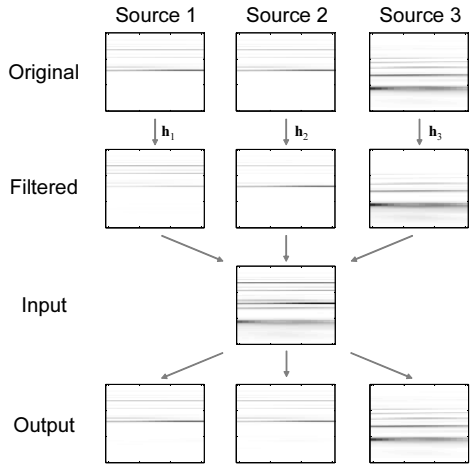


Figure 2: **Separation of three musical sources.** Three musical instruments at three distinct spatial locations were filtered (by $\mathbf{h}_1, \ldots, \mathbf{h}_3$, respectively) and summed to produce the *input* $\mathbf{y}$, and then separated using a sparse overcomplete representation to produce the *output*. Note that two of the sources (a flute playing the note "B", *left and center*) were chosen to be identical; this example is thus particularly challenging, since the only cue for separating the sources is the filtering imposed by the HRTF. Nevertheless, separation was good (compare *top* and *bottom* rows).

When applied to music, NMF typically yielded elements suggestive of musical notes, each with a strong fundamental frequency and weaker harmonics at higher frequencies. In many cases, listeners could easily use timbre to identify the instrument from which a particular element was derived. When applied to sounds from other ensembles (speech and natural sounds), NMF yielded elements that had rich harmonic structure, but it was not in general easy to "interpret" the elements (*e.g.* as vowels). Nonetheless, these elements still captured aspects of the statistical structure of the underlying ensemble of sounds.

## 2.2 Separation Performance

To test the model's ability to separate sources, we generated digital mixtures of three sources positioned at three distinct positions in space (Fig. 2). On the *top row* are the spectrograms of the sources at their origin. Note that two of the sources (a flute playing the note "B", *left and center*) were chosen to be identical; this example is thus particularly challenging, since the only cue for separating the sources is the filtering imposed by the HRTF.

Separation was nevertheless quite successful (compare *top* and *bottom* rows). These results were typical: whenever the underlying assumptions about the sparseness of the stimulus were satisfied, sources consisting of mixtures of music, natural sounds or speech were all sep-

arated well. Fig. 3 shows that separation without pre-filtering by the HRTF was unsuccessful, as was separation using a "dense" representation obtained via the pseudoinverse of the signal dictionary with $\mathbf{c} = \mathbf{D}^* \mathbf{y}$.

The representations underlying separation provide insight into these results. Fig. 4A shows the representations of each of the three sources (the same as in Fig. 2) presented in isolation. In each panel, the activity in a population of 3,600 neurons is indicated by the intensity of points on a $60 \times 60$ grid. Since the sources occupy three positions $j$, there are three copies of the basis $\mathbf{q}_i$ in each panel (corresponding to the three filters $\mathbf{h}_j$). The activity patterns are sparse; only a relatively small number of units are active in each representation. Note that because the left and the middle sources in this example were chosen to be identical, the left and middle neural representations differ only by a shift.

The procedure for recovering a source is straightforward: the estimate of the left source is simply the summed activity of the left third of the neurons—those representing features pre-filtered by the HTRF corresponding to the leftmost position in space; and likewise for the middle and right thirds. The HRTF can thus be seen as a kind of "tag" for grouping together elements from a single source. This suggests dividing source separation into two conceptually distinct steps (although in practice the steps occur simultaneously). In the first step, the stimuli are decomposed into the appropriate features. In the second step, the features are tagged and bundled together with other features from the same source. It is for this bundling step that the HRTF is essential.

The failure of the dense representation to separate sources (Fig. 3) results from a failure of the first step. The failure of even the sparse approach when the spectral cues induced by the HRTF are absent results from a failure at the second step. That is, the sparse approach finds a useful decomposition at the first step even without the HRTF, but without the HRTF cues the active features are not tagged, and so the features cannot be assigned appropriately to distinct sources.

# 3 Neural Model

We have reviewed how finding a maximally sparse linear representation in a suitably chosen overcomplete basis can directly solve a difficult signal processing problem. We now take this as a model of neuronal sensory processing, and posit that neuronal activities in primary auditory cortex (A1) correspond to the coefficients $c_{ij}$.

Although simply assumed in Eq. 2 to constitute a suitable sparse dictionary, the elements $q_j(t)$ reflect statistical correlations within sources; each source typically consists of several such features. These features can be thought

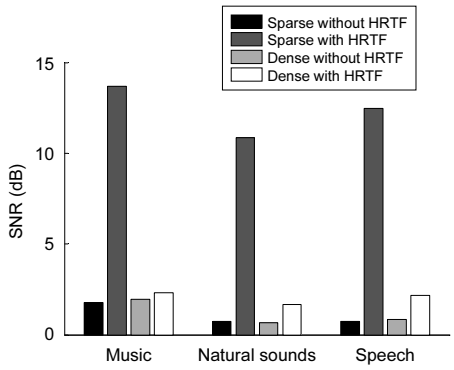


Figure 3: **Performance of different separation approaches on ensembles of music, natural sounds, and speech.** The SNR, across sources, is shown (**left**) for the musical ensemble, (**center**) the natural sounds ensemble, and (**right**) the speech ensemble. Excellent separation was achieved in all cases when the HRTF was known and a sparse prior was assumed.

of as an internal model of the components of acoustic sources, in the same way that edges might be thought of as components of visual sources (objects). However, because the neural representation involves pre-filtering with the HRTF (Eq. 4), each feature $d_{ij}(t)$ is better thought of as representing the hypothesis that an element $q_j(t)$ is present at position $i$. In the same way, neurons in the primary visual cortex can be thought of as representing the hypothesis ($d_{ij}$) that an oriented edge ($q_j$) is present at a particular position ($i$) in the visual field. In other words, the elements $q_j(t)$ reflect only the properties of the stimulus, whereas the features $d_{ij}(t)$ arise from interaction of these elements with the sense organs.

Our choice of NMF was merely one of convenience. We would not expect to find a precise correspondence between the features obtained by NMF and those observed in the auditory cortex. For this reason, our emphasis below will be not on the signal dictionary elements themselves, but rather on how they work together to form a representation that separates sources. We therefore turn our attention to making generic predictions: predictions which are not sensitive to the details of the signal dictionary, or for that matter to the specifics of the noise model or the particular measure of sparseness.

# 4 Experimental predictions

Our model makes at least three experimentally testable predictions about the nature of the neural representation underlying source separation.

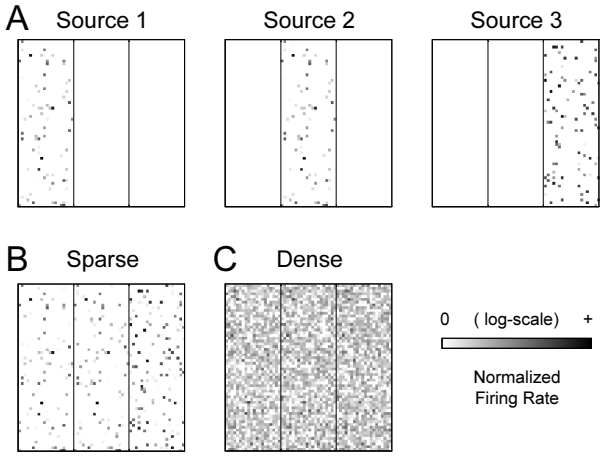


Figure 4: **Representations underlying source separation.** Each panel shows the activity of a population of 3,600 neurons, corresponding to the 3,600 features $\mathbf{d}_{ij} = \mathbf{h}_j * \mathbf{q}_i$. The intensity of each dot in the $60 \times 60$ grid is proportional to the log of the firing rate of each neuron. Since the sources occupy three positions $j$, there three copies of the basis $\mathbf{q}_i$ in each panel (corresponding to the three filters $\mathbf{h}_j$). The copies are arranged from left to right for convenience, and separated by vertical lines. However, the arrangement is for purposes of illustration only; we do not mean to imply any spatial organisation of sources within the cortex. The sources are the same as in the previous figure. **(A)** Sparse representations of the three sources (corresponding to the *original* spectrograms in Fig. 2) presented in isolation. Only a relatively small number of units are active in each panel. **(B)** Sparse representation of the mixed sources (*input* spectrogram in Fig. 2). Note that activity is approximately the sum of the activities of the isolated sources in (A). **(C)** Dense representation of the mixed sources. Note that most units are active.

## 4.1 Linear decoding & nonlinear encoding

Our model predicts that the encoding function is nonlinear but that the optimal decoding function is linear. Here *decoding* refers to the process of "reading out" a neural representation (*e.g.* by forming an estimate or reconstruction of the stimulus), whereas *encoding* refers to the process by which the nervous system constructs a pattern of neural activities from a stimulus.

It is sparseness that induces the nonlinear encoding; more precisely, the $L_1$ measure of sparseness induces a *piecewise linear* encoding function (Fig. 5). Sparseness implies that only at most $N_{\text{row}}$ out of the possible $N_{\text{col}}$ features $\mathbf{d}_{ij}$ are active in the representation of a particular stimulus; the precise subset of active neurons changes for different stimuli. Piecewise linearity arises because the encoding function is linear for all stimuli that activate the same subset of features, but changes for different subsets.

The prediction that there is an asymmetry between the linearity of the decoding function and the nonlinearity of the encoding function can be tested experimentally. Given an ensemble of stimulus-response pairs (*i.e.* the neural responses to an ensemble of sounds), our model predicts that a stimulus reconstruction approach (*i.e.* a decoding model) will outperform a "forward" (*i.e.* encoding) model.

The idea that a linear approximation is better suited for the neural decoding than encoding function was first exploited to estimate the information rate of fly visual neurons [6]. Our model provides a novel, principled explanation for this asymmetry in the context of overcomplete sparse representations. To our knowledge, this asymmetry has not been reported for high-level auditory representations. This prediction thus provides a strong test, since failure to observe the asymmetry will falsify our model.

## 4.2 Context-dependence of STRFs

A further prediction that follows from the piecewise linearity of the encoding function is that the linear component of receptive fields should depend on the acoustic context. Following conventional usage in auditory physiology, we will use the term spectrotemporal receptive field, or STRF, to refer only to the *linear* component of the encoding function, even though the encoding function itself may be highly nonlinear [7–9]. (In visual physiology, "STRF" is used to refer to the "*spatial* temporal receptive field," but the quantities are analogous). The STRF is the analog (in a high-dimensional input space) of the slope of a neuron's tuning curve in

In an experimental setting, piecewise linearity predicts that the STRF should depend on the acoustic context. We define the acoustic context of a feature $\mathbf{d}_{ij}$ with respect to a stimulus $\mathbf{y}$ as the collection of other features activated simultaneously by that stimulus. In music, for example, the features tend to resemble musical notes, and the acoustic context can be thought of as the set of notes (*e.g.* in a chord) that accompany a given note. Fig. 6 shows the STRF of the same neuron (a trumpet feature) in two different contexts (either clarinet or flute). The gross features of the STRF (*e.g.* the excitatory band around 880 Hz) are preserved in both contexts, but the secondary features (*e.g.* the addition of an inhibitory sideband) is context-sensitive. Changes in the STRF for different features and different contexts can be larger or smaller than in this example. Stimulus context thus changes the neural encoding function, suggestive of the non-classical receptive field modulation observed in visual cortex [10].

Context-dependence as defined here is stronger than simple nonlinearity. Specifically, the prediction is that there should exist extended subregions of stimulus space where the encoding function of a given target neuron is one lin-

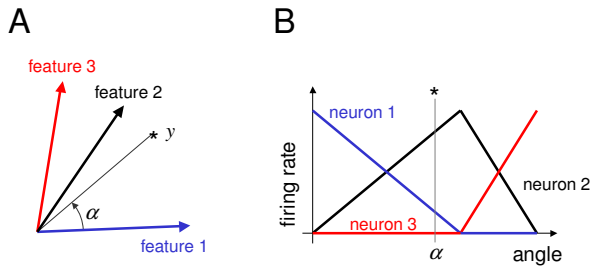


Figure 5: **Piecewise linear encoding.** **(A)** Three features in two dimensions constitute an overcomplete basis. A sample signal $\mathbf{y}$ is indicated with an '$*$'. **(B)** Tuning curves for the three features are piecewise linear. The firing rate of each of the three units in (A) is given as a function of angle for stimuli of unit length; the point $\mathbf{y}$ in (A) is at about 45°. Because the sample space is two-dimensional, any given point is represented by at most two active neurons. Decoding is linear: the point $\mathbf{y}$ is recovered by a weighted sum of the features, with the corresponding neural activities constituting the weights. Encoding, however, is nonlinear: the slope of all active neurons' activation functions can change at the boundaries, whenever any neuron becomes active or inactive. This principle generalises to the other examples in this paper, in which the dimensionality (given by the number of elements in the spectrogram) is much higher.

ear function, and across some boundary in stimulus space switch to a second linear function. These boundaries are demarcated by the activation of another (non-target) neuron in the population and the de-activation of a second (non-target) neuron (Fig. 5). This prediction could be tested using a multi-neuron recording technique.

The locally linear encoding induced by sparseness may help reconcile some of the apparent contradictions in the auditory literature. STRFs obtained using a "moving ripple" basis can predict responses to linear combinations of basis elements [7]. However, linear encoding (STRF) models fail to predict neural responses when the stimulus domain is extended to include a wide selection of complex sounds [11, 12], consistent with the idea that ripples represent a subspace within which encoding is linear. Context sensitivity may also provide an explanation for a proposed neural correlate of comodulation masking release in which the addition of a pure tone can suppress the response to temporally-modulated noise [13]; this form of contextual modulation cannot be explained by any purely linear encoding model.

## 4.3 Optimal feature estimation requires multi-neuron recording

In our model, the firing rate $c_{ij}$ of a neuron $\{i\,j\}$ is maximised when the stimulus matches that neuron's feature,

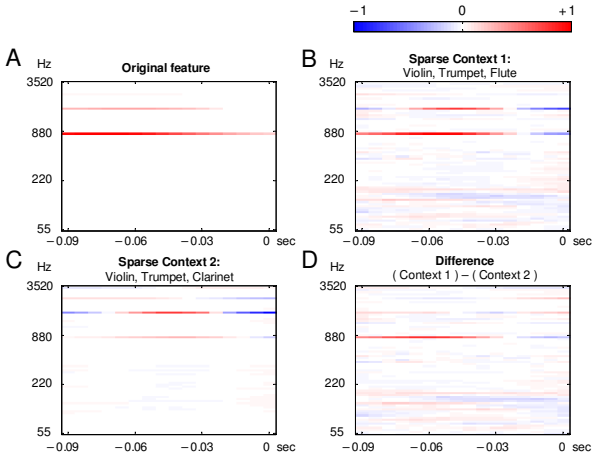


Figure 6: **Dependence of STRF on context.** **(A)** Spectrogram of trumpet feature, showing a strong fundamental around 880 Hz and several higher harmonics. **(B,C)** The STRFs corresponding to the feature in (A) when that feature is played in two different contexts (clarinet or flute played simultaneously), derived under the assumption of a sparse neural representation. The STRF provides the *encoding* from the stimulus to neural activity. The colour at any point of the STRF indicates the value (in spikes/second) of the kernel which is convolved with the spectrogram of the stimulus to generate a neural response. Under the sparse assumption, the encoding is piecewise linear, and the STRFs shown are two out of the many possible pieces. The STRF is obtained from the appropriate row of the matrix $\mathbf{D}_k^{\diamond}$ (see *Methods*). **(D)** The difference between the two spectrograms. Note that they show the same basic harmonic structure, but differ in details such as the relative contributions of the excitatory and inhibitory sidebands. The differences can be as large as the STRFs themselves.

*i.e.* when $\mathbf{y} = \mathbf{d}_{ij}$. Since the feature $\mathbf{d}_{ij}$ is used in the linear reconstruction of the stimulus from the neural activities (Eq. 5), one might imagine that the optimal stimulus (*i.e.* the stimulus that maximises the firing rate) can be obtained by estimating the optimal linear decoder. Experiments based on this idea have shown that the optimal linear decoder can sometimes drive neurons in the auditory cortex to fire vigorously [14].

Surprisingly, our model predicts that this linear estimate of the decoder is *not* the optimal stimulus, even though the optimal decoder is linear. Instead, finding the optimal stimulus requires recording from *all* neurons involved in the representation. This is because the $\mathbf{d}_{ij}$ are not orthogonal. Note that in our model, optimal decoding (Eq. 5) need not take neural correlations into account, even when they are present. The phenomenon is illustrated in Fig. 7.

When the optimal linear decoder is estimated from only the target neuron, the firing rate is sub-maximal. As the

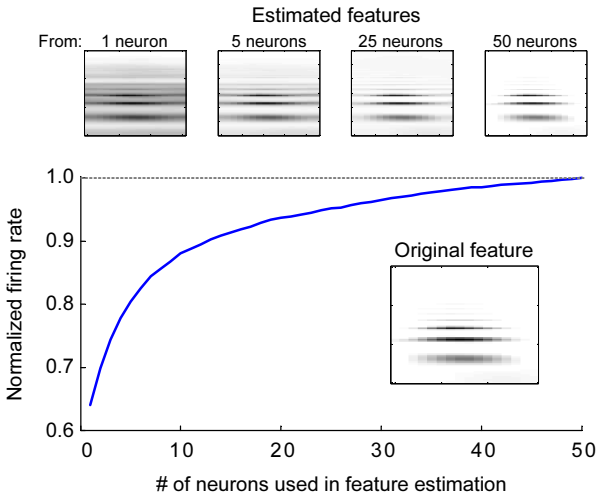


Figure 7: **Stimulus optimisation requires multi-neuron recording.** The $y$-axis shows the simulated firing rate of a target neuron (normalized to its maximum firing rate) in response to the presentation of the optimal linear decoder constructed by recording the activity of a target neuron and a variable number of other neurons. When the optimal linear decoder is estimated from only the target neuron, the firing rate is sub-maximal. As the number of neurons used in this simulation to estimate the optimal linear decoder is increased ($x$-axis), the response of the target neuron converges to unity, indicating that optimal decoder has converged to the target neuron's feature.

number of neurons used to estimate the optimal linear decoder is increased ($x$-axis), the response of the target neuron converges to unity, indicating that optimal decoder has converged to the target neuron's feature. This represents a novel and testable prediction of the model. Note that although in principle the activity of all neurons involved in the representation must be recorded, in practice the activity of even a few can be useful.

## Acknowledgements

Supported by Higher Education Authority of Ireland (An tÚdarás Um Ard-Oideachas) and Science Foundation Ireland grant 00/PI.1/C067 (BAP), a Farrish-Gerry Fellowship (HA) and the Sloan Foundation, Mathers Foundation, NIH, Packard Foundation and the Redwood Neuroscience Institute (AMZ).

## References

[1] P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal Processing*, 81(11):2353–2362, 2001.

[2] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural. Comput.*, 13(4):863–882, Apr. 2001.

[3] B. A. Pearlmutter and A. M. Zador. Monaural source separation using spectral cues. In *Fifth International Conference on Independent Component Analysis*, LNCS 3195, pages 478–485, Granada, Spain, Sep. 22–24 2004. Springer-Verlag.

[4] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural. Comput.*, 15(2):349–396, 2003.

[5] D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matric factorization. *Nature*, 401:788–791, 1999.

[6] W. Bialek, F. Rieke, R. R. de Ruyter van Stevenick, and D. Warland. Reading a neural code. *Science*, 252:1854–1857, 1991.

[7] N. Kowalski, D. A. Depireux, and S. A. Shamma. Analysis of dynamic spectra in ferret primary auditory cortex II: Prediction of unit responses to arbitrary dynamic spectra. *J. Neurophysiol.*, 76(5): 3524–3534, 1996.

[8] F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network*, 12(3):289–316, 2001.

[9] F. E. Theunissen, K. Sen, and A. J. Doupe. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neuroscience*, 20(6):2315–2331, 2000.

[10] W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

[11] J. F. Linden, R. C. Liu, M. Sahani, C. E. Schreiner, and M. M. Merzenich. Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. *J. Neurophysiol.*, 90(4):2660–2675, 2003.

[12] C. K. Machens, M. S. Wehr, and A. M. Zador. Linearity of cortical receptive fields measured with natural sounds. *J. Neurosci.*, 24(5):1089–1100, 2004.

[13] I. Nelken, Y. Rotman, and O. B. Yosef. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, 397(6715):154–157, 1999.

[14] R. C. deCharms, D. T. Blake, and M. M. Merzenich. Optimizing sound features for cortical neurons. *Science*, 280(5368):1439–1443, 1998.