An Absolute Scale for Evaluating Neuronal Response Predictions

M. Sahani[1], J. F. Linden[2] and M. M. Merzenich[2]
[1]Gatsby Computational Neuroscience Unit, University College, London
[2]Keck Center for Integrative Neuroscience, UC San Francsico

Abstract

A prerequisite for quantitative understanding of perceptual systems is accurate modelling of sensory neuron responses. "Reverse correlation" has advanced this goal considerably, employing both linear models and nonlinear extensions. But how complex a model is needed? It is straightforward to rank the performance of two types of model, but not to judge a single type on a meaningful absolute scale. We propose such a scale based on a nonparametric estimate of the predictable power in a neural response. Using it, we evaluate spectrotemporal receptive field (STRF) models for rodent auditory cortex cells. While showing that the predictive abilities of the STRFs can be improved beyond that achieved by the Wiener kernel using Bayesian regression techniques, we conclude that the best STRF models (even when they incorporate static output non-linearities) explain less than half the predictable power in our recordings.

When repeatedly presented with identical stimulus sequences, the responses of neurons at intermediate stages of perceptual systems can vary dramatically, even in anaesthetised animals. While this variability may reflect meaningful changes in the internal state of the animal or else may be completely random, from the point of view of modelling the relationship between stimulus and neural response it must be treated as noise, and should therefore be excluded when setting an absolute standard against which a predictive model for a response — hereafter, a response function (RF) model — can be judged. The mutual information rate between stimulus and response is one example of a measure which excludes stimulus-independent variability in the neural response. However, while quantities derived from it can be used to determine whether an RF model retains relevant information about the stimulus, such measures are not sensitive to distortions in the model predictions (e.g., a sign inversion would not affect mutual information rates). In contrast, classic measurements of model performance, such as the R^2 statistic[5], make no attempt to discount noise; therefore, while they can be used to rank different models or to compare a model to a purely random null hypothesis, they do not provide an absolute scale for determining when a model is adequate in itself. A similar limitation applies to an approach in which the variability between two successive responses to the same stimulus is compared to that between the model prediction and a novel response. This method amounts to a model comparison scheme in which one half of the comparison is the "identity" model (i.e., a model which predicts that the output will be identical on different trials). Since trials are actually variable, the identity model is incorrect, and so this method will in fact tend to overestimate the validity of the RF model. To overcome the shortcomings of these approaches, we propose a power statistic which directly estimates just that part of the response which is repeated from trial to trial. This measure provides a reliable absolute benchmark for evaluating RF model performance.

The data to be used in this analysis consist of spike trains continuously recorded during presentation of a long, complex, rapidly varying stimulus. This stimulus is treated as a discrete-time process. In the auditory experiment considered here, the discretization was set by the duration of regularly clocked sound pulses of fixed length; in a visual experiment, the discretization might be the frame rate of a movie. The neural response can then be measured with the same level of precision, counting action potentials to estimate a response rate for each time bin, to obtain a response vector $\mathbf{r} = (r_t)_{t=1...T}$. We propose to measure model performance in terms of the fraction of *response power* predicted successfully, where "power" is used in the sense of average squared deviation from the mean: $P(\mathbf{r}) = \left\langle (r_t - \langle r_t \rangle)^2 \right\rangle$ ($\langle \cdot \rangle$ denoting averages over time). As argued above, only some part of the total response power is predictable, even in principle; fortunately, this *signal power* can be estimated by combining repeated responses to the same stimulus sequence. We present a method of moments[5] derivation of the relevant estimator.

Suppose we have N responses $\mathbf{r}^{(n)} = \boldsymbol{\mu} + \boldsymbol{\eta}^{(n)}$, where $\boldsymbol{\mu}$ is the common, stimulus-dependent component of the response and $\boldsymbol{\eta}^{(n)}$ is the (zero-mean) noise component of the nth trial. The expected power in each response is given by $P(\mathbf{r}^{(n)}) \overset{\mathcal{E}}{=} P(\boldsymbol{\mu}) + \left\langle (\eta_t^{(n)})^2 \right\rangle$ (where the symbol $\overset{\mathcal{E}}{=}$ means "equal in expectation"). We can now construct two trial-averaged quantities, similar to the sum-of-squares terms used in the analysis of variance[5]: the average power per response, and the power of the average response. Using $\bar{\cdot}$ to indicate trial averages:

$$P(\overline{\mathbf{r}^{(n)}}) \overset{\mathcal{E}}{=} P(\boldsymbol{\mu}) + P(\overline{\boldsymbol{\eta}^{(n)}}) \qquad \text{and} \qquad \overline{P(\mathbf{r}^{(n)})} \overset{\mathcal{E}}{=} P(\boldsymbol{\mu}) + \overline{P(\boldsymbol{\eta}^{(n)})}.$$

Assuming the noise in each trial is independent (although the noise in different time bins within a trial need not be), we have: $P(\overline{\boldsymbol{\eta}^{(n)}}) \overset{\mathcal{E}}{=} \overline{P(\boldsymbol{\eta}^{(n)})}/N$. Thus solving for $P(\boldsymbol{\mu})$ suggests the following estimator for the signal power

$$\hat{P}(\boldsymbol{\mu}) = \frac{1}{N-1}\left(NP(\overline{\mathbf{r}^{(n)}}) - \overline{P(\mathbf{r}^{(n)})}\right). \tag{1}$$

(A similar estimator for the noise power is obtained by subtracting this expression from $\overline{P(\mathbf{r}^{(n)})}$.) This estimator is unbiased (provided only that the noise distribution has defined first and second moments, and is independent between trials), as can be verified by explicitly calculating its expected value. Since each of the power terms in equation 1 is the mean of T numbers, the central limit theorem suggests that $\hat{P}$ will be approximately normally distributed for long experiments (for the experiment described below, $T = 3000$). Its variance is given by:

$$\mathcal{V}ar\left[\hat{P}\right] = \frac{4}{N}\left(\frac{1}{T^2}\boldsymbol{\mu}'\Sigma\boldsymbol{\mu} - \frac{2}{T}\mu\boldsymbol{\sigma}'\boldsymbol{\mu} + \mu\sigma\mu\right) + \frac{2}{N(N-1)}\left(\frac{1}{T^2}\mathrm{Tr}\left[\Sigma\Sigma\right] - \frac{2}{T}\boldsymbol{\sigma}'\boldsymbol{\sigma} + \sigma^2\right), \tag{2}$$

where Σ is the covariance matrix of the noise, $\boldsymbol{\sigma}$ is a vector formed by averaging each column of Σ, σ is the average of all the elements of Σ and μ is the time-average of the mean $\boldsymbol{\mu}$. Thus, $\mathcal{V}ar\left[\hat{P}\right]$ depends only on the first and second moments of the response distribution, and we can substitute the data-derived estimates of these into equation 2 to yield a standard error bar for the estimator. We have thereby obtained an estimate, with corresponding uncertainty, of the maximal possible performance of any predictive model.

To compare the performance of an estimated RF model to this maximal value, we must determine the power successfully predicted by the model. As the prediction may be inaccurate, this is not just the power of the predicted response. Instead, we first measure the residual power in the signal formed by subtracting the predicted response from the measured response. The difference between this *error power* and the total power of the response describes the extent to which the prediction was successful. We define this difference in powers to be the *predictive power* of the model; it is this value which can be compared to the estimated signal power.

To be able to describe more than one neuron, an RF model class must contain parameters that can be adapted to each case. Ideally, the appropriateness of the model class for a population of neurons would be judged using parameters that produced models closest to the true RFs (the *ideal models*), but we do not have *a priori* knowledge of those parameters. Instead, the parameters must be tuned in each case using measured neuronal responses or *training data*.

One way to choose the RF model parameters is to minimize the mean squared error (MSE) between the predicted response and the neural response in the training data; for example, the Wiener kernel minimizes the MSE for a model based on a finite impulse response (FIR) filter of fixed length, assuming that the input and system are stationary. This MSE is identical to the error power that would be obtained for a prediction made on the training data stimulus. Thus, by minimising the MSE (or error power), we maximize the predictive power on the training data. The resulting maximum value, hereafter the *training predictive power*, is likely to be an overestimate of the predictive ability of the ideal model, since the minimum-MSE parameters will be overfit to the training data. Overfitting is inevitable, because model estimates based on finite data will always capture some stimulus-independent response variability. More precisely, the expected value of the training predictive power is an upper bound on the true predictive power of the model class; we therefore refer to the training predictive power itself as an *estimated upper bound* on the model performance.

We can also obtain an *estimated lower bound*, defined similarly, by empirically measuring the generalisation performance of the model by cross-validation[4]. A subset of the data is used to choose model parameters, and then the performance of the model with those parameters is evaluated on the remaining data to obtain a *test predictive power*. This procedure is repeated a number of times, holding out a different subset of the available data in each case. The resulting test predictive power estimates are then averaged to obtain the *cross-validation predictive power*. Since overfitting to the training data will reduce the ability of a model to generalize to novel data, the cross-validation predictive power will tend to underestimate the predictive power of the ideal model, and thus provides the desired estimated lower bound. In general, the minimal-MSE parameter settings (such as the Wiener kernel) overfit severely, and will produce an overly loose lower bound. We have successfully used Bayesian techniques with appropriately optimized priors to improve the robustness of the parameter estimates in the cross-validation sense, and thus tighten the lower bound on predictive power.

For any one recording of finite length, the predictive power of the ideal model of a particular type (e.g. STRF models) can only be bracketed between these estimated upper and lower bounds (that is, between the training predictive power and the cross-validation predictive power). As the noise in the recording grows, the model parameters will overfit more and more to the noise, and hence both bounds will grow looser. Indeed, in high-noise conditions, the model may primarily describe the stimulus-independent part of the training data, and so the training predictive power might exceed the signal power, while the cross-validation predictive power may fall below zero. However, assuming that the predictive power of a single model class is similar for a population of similar neurons, it is possible to exploit this noise dependence to tighten the bounds by extrapolating within the population to the zero noise point. This extrapolation allows us to answer the sort of question posed at the outset: how well, in an absolute sense, can a particular type of model account for the responses of a population of neurons?

We used the techniques described above to quantify the extent to which the stimulus-dependent signal in a population of auditory cortical neurons could be predicted by a linear function of the spectrographic representation of an auditory stimulus; that is, how accurate a description of their responses could be provided by the STRF[1]. Neurons in the auditory cortex are thought to be insensitive to the phase of a sound, and so studies of response properties have tended to focus on the spectral content of the stimulus; in particular, a number of reverse-correlation studies have characterized neurons with reference to the stimulus spectrogram.

To this end, we analyzed 205 recordings collected from 68 recording sites within the thalamo-recipient layers of rodent primary auditory cortex. Animals (6 CBA/CaJ mice and 4 Long-Evans rats) were anaesthetised with either ketamine/medetomidine or sodium pentobarbital, and neural responses recorded using extracellular tungsten microelectrodes while a complex random chord stimulus was presented (see Fig. 1**a**). Recordings often reflected the activity of a number of neurons. We identified single neurons by Bayesian spike-sorting techniques[2, 6] whenever possible; at these sites, the properties of the isolated cell responses were similar to those of the aggregate activity, allowing us to treat all responses, whether single-cell or aggregate, together.

The recordings were binned to match the discretization rate of the stimulus (Fig. 1**c**) and the signal power estimated as above. A total of 141 recordings in the data set (75 from mouse, 66 from rat) showed signal power more than 2 standard errors above zero (Fig. 1**d–e**). The subsequent analysis was confined to these stimulus-responsive recordings.

For each such recording, we sought to build a model that predicted the number of spikes in each time bin using a linear function of the pulse amplitudes at the same and preceding 14 time-steps (in total, a 300ms window of the sound). In the discrete formulation this is a problem in linear regression, the standard minimum-MSE solution to which corresponds to Wiener kernel estimation

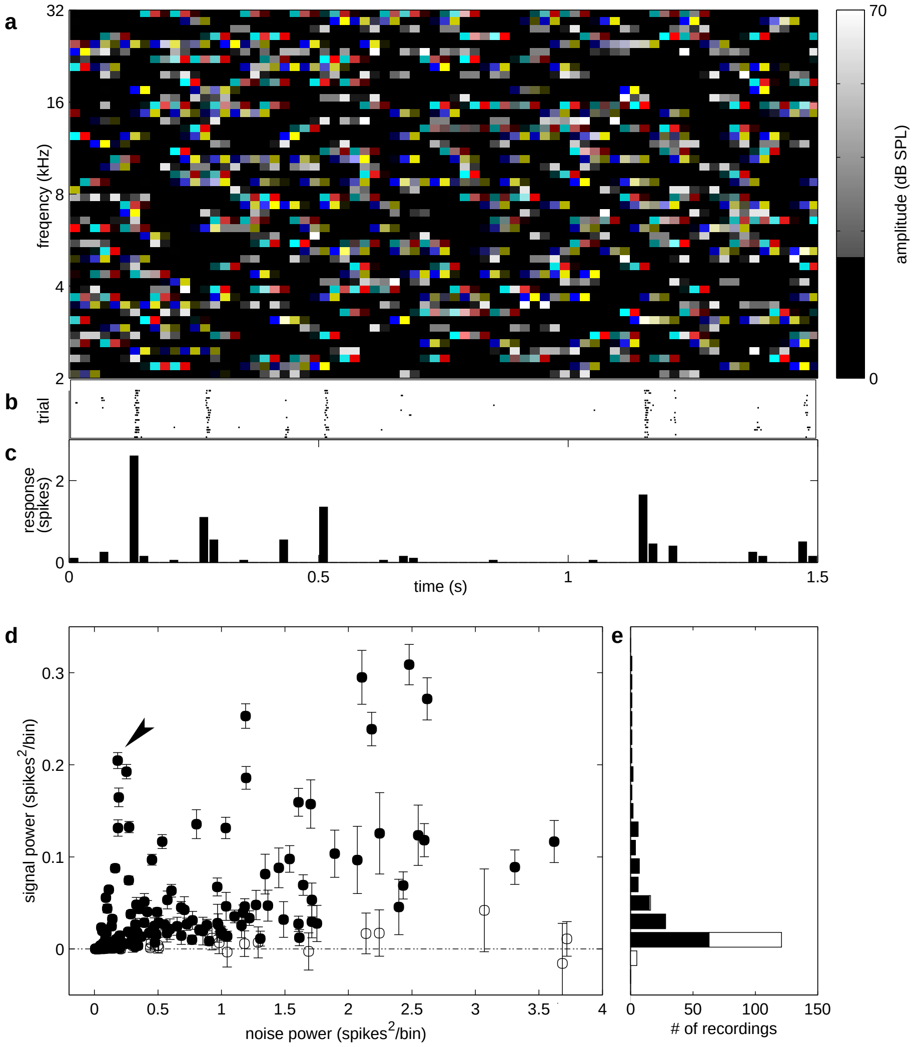

Figure 1: **a**: Spectrographic representation of a 1.5s segment of the auditory stimulus. Each coloured rectangle represents a short tone pulse centred on the indicated frequency and time, with a maximal intensity given by the colour. Pulses were 20ms in length and ramped up and down with 5-ms cosine-gates. The times, frequencies and sound intensities of all pulses were chosen randomly and independently within the discretizations of those variables (20 ms bins in time, 1/12 octave bins covering 2-32 kHz in frequency, and 5 dB SPL bins covering 25-70 dB SPL in amplitude). At any time point, the stimulus averaged two tone pips per octave, with an expected loudness of approximately 73 dB SPL. For each recording site the same 60s long stimulus was repeated either 10 or 20 times. **b–c**: Measured response to the stimulus segment in **a** for one recording site, shown as a spike rastergram (**b**), and as a histogram (**c**) obtained by binning the recorded spike train with a precision of 20ms. **d–e**: Distribution of signal powers for all recordings, shown as a function of noise power (**d**), and as a histogram (**e**). The point corresponding the recording in **c** is indicated by the arrowhead. Error bars show two standard errors of the signal power estimate. Filled circles and bars indicate recordings where the signal power was significantly greater than zero, open circles and bars indicate those with signal power indistinguishable from zero.

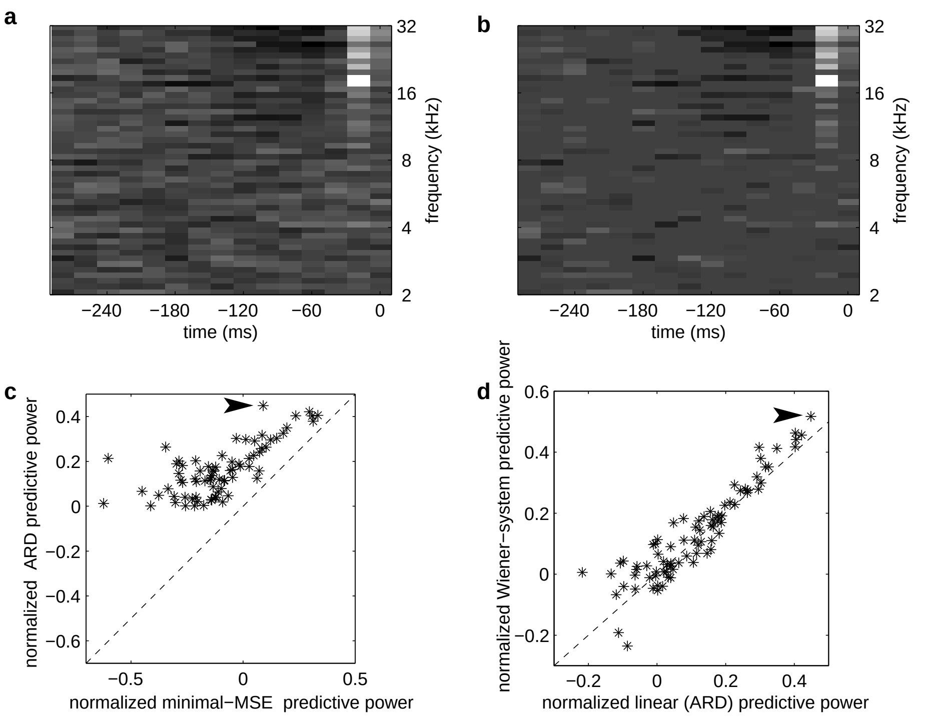

Figure 2: RF model estimation. **a**: minimum-MSE estimate of the STRF for one recording. **b**: Bayesian ARD-derived estimate of the STRF for the same recording. **c**: comparison of cross-validation predictive powers for minimum-MSE and ARD models. **d**: comparison of predictive powers for ARD-derived linear model and Wiener-system model, which incorporates a static output nonlinearity. In **c** and **d** the predictive powers have been normalized by the estimated signal power. The arrowhead indicates the point corresponding to the STRFs in **a** and **b**.

for the time-series (Fig. 2**a**). This solution yielded the estimated upper bounds on predictive power.

The minimum-MSE solution generalizes poorly, and so generates overly pessimistic estimated lower bounds. However, the linear regression literature provides alternative parameter estimation techniques with improved generalization ability. In particular, we used a Bayesian hyperparameter optimization technique known as Automatic Relevance Determination[3] (ARD) to find an optimized prior on the regression parameters, and then chose parameters which optimized the posterior distribution under this prior and the training data (Fig. 2**b**). Fig. 2**c** shows the cross-validation predictive power of the ARD-derived parameters for each recording compared to those obtained by minimizing the MSE, and clearly demonstrates the improved generalization of the Bayesian approach. The ARD-derived cross-validation predictive powers provided the estimated lower bounds on the predictive power of the spectrogram-linear class of models.

The linear model is not constrained to predict non-negative firing rates. To test whether including this additional constraint could improve predictions, we also fit models in which the output of an FIR filter on the spectogram is non-linearly transformed time-point by time-point to obtain the firing rate prediction — a model known as a “Wiener system”. The ARD-derived estimates of the spectogram-linear model were used for the filter component of the Wiener system, while the one-dimensional non-linearity was estimated by a non-parametric kernel regression technique. The resulting cross-validation predictive powers are compared to those of the spectogram-linear model in Fig. 2**d**; the addition of a non-linear stage contributes very little to the predictive power of the basic STRF model class.

Fig. 3 shows the estimated upper and lower bounds on predictive power of the spectogram-linear model class for our population of rodent auditory cortex cells, as a function of the estimated noise

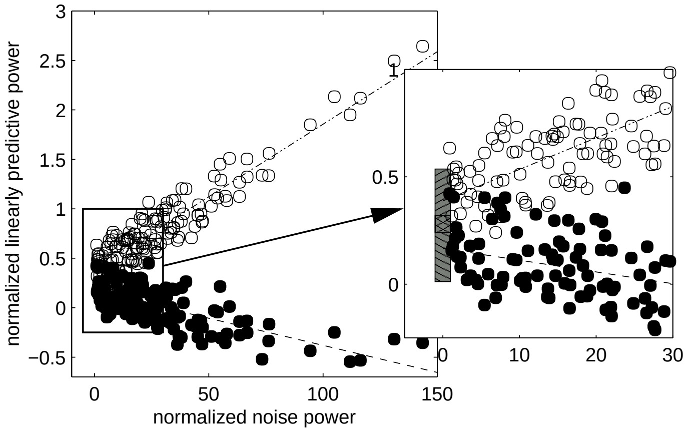

Figure 3: Estimated bounds on predictive power as a function of noise power. For each analyzed recording, the open circle represents the training predictive power (estimated upper bounds) and the filled circle shows the cross-validation predictive power (estimated lower bounds). The dotted line is fit to the upper bounds, while the dashed line is fit to the lower bounds. In the expanded axes the boxes indicate the regions containing at least 50% of the population scatter, extrapolated to zero noise.

level in each recording. The divergence of the bounds at higher noise levels, as was described above, is evident. At low noise levels, the bounds do not converge perfectly. Indeed, even with no noise the models would overfit to the training stimulus sequence, and so the extrapolated population-average bounds are separated even at zero: 0.3878 ± 0.0076 for the upper bound and 0.1663 ± 0.0079 for the lower (intervals are standard errors).

Some portion of the scatter of the points about the population average lines reflects genuine variability in the population, and so the extrapolated scatter at zero noise is also of interest (Fig. 3 inset). Intervals containing at least 50% of the population are: $(0.239, 0.536)$ for the upper estimate and $(0.012, 0.320)$ for the lower estimate (assuming normal scatter). These will be overestimates of the spread in the underlying population distribution because of additional scatter from estimation noise.

Thus models that are linear in the stimulus spectrogram are unlikely to be able to account for more than half of the predictable stimulus-related power in the responses of these neurons from the thalamo-receipient layers of rodent primary auditory cortex. Further, elaborated models that append a static non-linearity to the linear filter are no more effective at predicting responses to novel stimuli than is the linear model class alone. Current and future work will need to be directed towards the development of better classes of models for the responses of such cells. The statistical apparatus described here will provide a way to assess the success of such models against an absolute benchmark.

[1] A. M. Aertsen and P. I. Johannesma. The spectro-temporal receptive field. a functional characteristic of auditory neurons. *Biol Cybern*, 42(2):133–43, 1981.

[2] M. S. Lewicki. Bayesian modeling and classification of neural signals. *Neural Comp*, 6(5):1005–1030, 1994.

[3] D. J. C. MacKay. Bayesian non-linear modelling for the prediction competition. In *ASHRAE Transactions, V.100, Pt.2*, pages 1053–1062, Atlanta Georgia, 1994. ASHRAE.

[4] B. D. Ripley. *Pattern recognition and Neural networks.* CUP, Cambridge, 1996.

[5] S. M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists.* Academic Press, London, 2nd edition, 1999.

[6] M. Sahani. *Latent Variable Models for Neural Data Analysis.* PhD thesis, California Institute of Technology, Pasadena, California, 1999.