Encoding and Decoding from populations of neurons

Peggy Seriès
ANC, University of Edinburgh
pseries@inf.ed.ac.uk
Neuroinformatics lecture -- 14/2/2007

Outline

Part 1 (14/2/07):
- Questions that we would like to answer
  - Encoding side
  - Decoding side
- Fisher Information in simple models
- Decoding in detailed models : application to the problem of orientation selectivity

Part 2 (16/02/07):
Matlab session. Simple population encoding models.
Decoding and computing Fisher information.
Cracking the code

We would like to understand the link between the activity of neurons, and (visual) perception. For example, neurons in V1 and estimation/discrimination of orientation.

Primary Visual Cortex (V1)
Orientation selectivity in V1

Optical imaging

Electrophysiology

Orientation discrimination

Is the second grating of the same orientation as the first grating, or a different orientation?

Performances is not fixed, but vary with conditions
Orientation (mis)estimation

Tilt after-effect

Stare at this for 20 sec

Then look at that

Zollner Illusion

Fraser Illusion

Here too things can go wrong!!

Poggendorf Illusion

Orientation (mis)estimation

Tilt after-effect

Stare at this for 20 sec

Then look at that

Zollner Illusion

Fraser Illusion

Poggendorf Illusion
Cracking the code

We would like to understand the link between the activity of neurons, and (visual) perception. For example, neurons in V1 and estimation / discrimination of orientation. And how things change depending on the conditions.

But

- Neurons are noisy
- The code relevant to a given task or object is distributed in (probably very large) populations of neurons, and may be different areas.
- ...Lots of unknown in encoding and decoding

Encoding questions

1. What information is encoded?
2. What is the format of the representation? What is the signal, what is the noise?
3. How can we quantify the information?
4. How precise is the representation / How much information?
5. What are the factors that control the amount of information?
   - shape of the tuning curve
   - variance
   - correlations
6. How redundant is the information / what is the benefit of pooling? (in space, or integrating in time)
What information is encoded? (in V1)

1. Spatial position
2. Orientation
3. Contrast
4. Spatial frequency
5. Temporal frequency
6. Direction
7. Color

Fairly well understood, but also:
- Temporal context (e.g., adaptation)
- Spatial context (center-surround modulation)
- Natural scenes …
- Eye position?
- Reward?
- Auditory inputs?
- Attention?

We don’t understand more than 15% of what V1 is doing …

[Sceniak et al., 2002]

Encoding questions

1. What information is encoded?
2. What is the format of the representation? What is the signal, what is the noise?
3. How can we quantify the information? How precise is the representation / How much information?
4. What are the factors that control the amount of information?
   - shape of the tuning curve
   - variance
   - correlations
5. How redundant is the information / what is the benefit of pooling? (in space, or integrating in time)

[Olshausen & Field, 2005]
Format of the representation

- Rate or # spikes  -- (over which time window?)

But
- Latency
- First spike rank coding
- Precise spike times?
- Correlations? Synchrony?
- How many neurons involved? How sparse the code?

{Tuning curve + Noise} model

\[ a_i(\theta) = f_i(\theta) + \eta_i(\theta) \]

- Noise – random variable
  \[ \eta_i(\theta) \sim N(0, \sigma(\theta)) \]

Mean activity
  = tuning curve

Abstraction : Encoding model

Poisson:

\[ p(a_i = k \mid \theta) = f_i(\theta)^k \frac{e^{-f_i(\theta)}}{k!} \]

- \[ E[a_i \mid \theta] = f_i(\theta) \]
- \[ \text{var}[a_i \mid \theta] = f_i(\theta) \]
Abstraction : Encoding model

**Poisson:**
\[
p(a_i = k \mid \theta) = f_i(\theta)^k \frac{e^{-f_i(\theta)}}{k!}
\]
\[
E[a_i \mid \theta] = f_i(\theta)
\]
\[
\text{var}[a_i \mid \theta] = f_i(\theta)
\]

**Gaussian:**
\[
p(a_i \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2(\theta)}} e^{-\frac{(a_i - f_i(\theta))^2}{2\sigma^2(\theta)}}
\]
\[
E[a_i \mid \theta] = f_i(\theta)
\]
\[
\text{var}[a_i \mid \theta] = \sigma_i^2(\theta)
\]
Encoding questions

1. What information is encoded?

2. What is the format of the representation? What is the signal, what is the noise?

3. How can we quantify the information? How precise is the representation / How much information?

4. What are the factors that control the amount of information?
   - shape of the tuning curve
   - variance
   - correlations

5. How redundant is the information / what is the benefit of pooling? (in space, or integrating in time)

How can we quantify the information?

1. **Information Theory approach**
   Entropy and mutual information -- single neurons or pairs

   \[
   H = -\sum r[i] \log_2 p[r[i]] \\
   H_s = -\sum r[i] \log_2 p[r[i] | s] \\
   L_s = H - H_s = K L_s (p[r[i] | s]) \cdot (p[r[i]])
   \]

   A measure of statistical dependency between the stimulus and the response.

2. **Decoding**
   Estimation Theory / Detection theory / Classification approach
   Populations of neurons

   **Fisher information** (a bound on how well you can decode the stimulus)
Say you observe a response $r = 5$ spikes. You can form the likelihood $p(r | \theta)$ which is going to tell you how probable it is to observe 5 spk for each possible stimulus.

How “peaky” the (mean) likelihood? The more peaky it is, the easier it will be to “guess” the stimulus – the higher Fisher information.

Encoding questions

1. What information is encoded?
2. What is the format of the representation? What is the signal, what is the noise?
3. How can we quantify the information? How precise is the representation / How much information?
4. What are the factors that control the amount of information?
   - shape of the tuning curve (slope, width, amplitude)
   - variance (fixed? Proportional to the mean?)
   - correlations (hurt? Help?)
5. How redundant is the information / what is the benefit of pooling? (in space, or integrating in time)
The factors that control accuracy

- shape of the tuning curve (slope, width, amplitude)
- number of neurons involved
- variance (fixed? proportional to the mean?)
- correlations (hurt? help?)

If we understand those, we can understand how the accuracy of the code / precision of the sensation / psychophysical performance can change

In:
Perceptual Learning (what changes in your brain when you train)
Attention (-------------------------- when you focus)
Effects of Adaptation (---------------------- when you adapt)
Effects of Aging (--------------------------------- when you age)
Lack of sleep 😔
Disease
Drugs
Etc..

One neuron: the slope

![Diagram showing the firing rates of a neuron in response to different orientations](image-url)
One neuron: the slope

How can the efficiency of the code change?
How can the efficiency of the code change?

- Sharpening the tuning curve or increasing its amplitude.
- Orientation selectivity: Attention, Perceptual learning?

Factors that control the amount of information: slope, noise, the number of neurons

**Poisson noise**

\[ I_i(\theta) = \frac{[f_i'(\theta)]^2}{f_i(\theta)} \]

\[ I_F(\theta) = \sum_i \frac{[f_i'(\theta)]^2}{f_i(\theta)} \]

**Gaussian noise**

\[ I_i(\theta) = \frac{[f_i'(\theta)]^2}{\sigma_i^2(\theta)} \]

\[ I_F(\theta) = \mathbf{r}^T \mathbf{Q}^{-1} \mathbf{r} \]

Slope squared variance

For independent neurons, FI of the population is the sum of each neuron's FI.
Factors that control the amount of information: The noise correlations

We look at the trial to trial fluctuations of activity of pairs of neurons, around their mean. When neuron 1 is above its mean, is neuron 2 also ?, or are their fluctuations independent?

Small positive noise correlations are found in cortex.

Factors that control the amount of information: The noise correlations

• Noise correlations have a strong impact on the accuracy of the code.

• Whether they hurt or help depends on their relationship with “signal” correlations (overlap in tuning curves)

• To assess the influence of correlations, one way is to artificially destroy them in the data (shuffle the trials) and recompute information.

• Note that population representation inferred from single electrode recordings corresponds to this “shuffled” information. Understanding how different it is from the true representation is thus critical.

[averbeck et al, 2006]
Encoding questions

1. What information is encoded?

2. What is the format of the representation? What is the signal, what is the noise?

3. How can we quantify the information? How precise is the representation / How much information?

4. What are the factors that control the amount of information?
   - shape of the tuning curve
   - variance
   - correlations

5. How redundant is the information / what is the benefit of pooling? (in space, or integrating in time)

The Benefit of Pooling

- Pooling from large populations of neurons was commonly thought to be a way to average out the noise.

- Since neurons are correlated, a critical question (still debated) is whether pooling more and more neurons increases (linearly) the accuracy of the representation (as would be the case if neurons were independent)

- Is information saturating over a certain number of neurons? [Zohary et al 1994]

- = What are the “units” of representation in the brain? 10 neurons? 100? 1000? 10 000?

- The answer to this question depends on i) the structure of the noise in the brain ii) how “decoding” is done (= whether the brain is optimal or not).
Decoding questions

1. What’s decoding
3. Does the brain “decode”? What are the “neural correlates of consciousness”?
4. Does the brain explicitly encode the likelihood, the prior.
5. How to build Brain Machines Interfaces?

Decoding populations of neurons

The nature of the problem
In response to a stimulus with unknown orientation $\theta$, we observe a pattern of activity $r$. What can we say about $\theta$ given $r$?

Estimation theory: come up with a single value (a guess) for $\theta$.

Bayesian Approach: recover $P(\theta | r)$ (the posterior distribution).
Decoding populations of neurons

Possible decoding algorithms.
(statistical optimality, optimality within a class, or arbitrary choice)

• Maximum likelihood estimation. The ML estimate is the value of $\theta$
maximizing the likelihood $p(r|\theta)$. We seek:

$$\hat{\theta} = \arg \max_{\theta} P(r | \theta)$$
Decoding populations of neurons

Possible decoding algorithms.

- Maximum likelihood estimator. The ML estimate is the value of $\theta$ maximizing the likelihood $p(r|\theta)$. We seek:

  $\hat{\theta} = \arg\max_\theta P(r | \theta)$

- Population vector

  $\hat{P} = \sum_{i=1}^{N} r_i \tilde{P}_i$

  $\hat{\theta} = \text{angle}(\hat{P})$

- Winner Take all. Choose the preferred orientation of the neuron with largest response.
Decoding populations of neurons

Possible decoding algorithms. (statistical optimality or simplicity?)

- Maximum likelihood estimation. The ML estimate is the value of $\theta$ maximizing the likelihood $p(r|\theta)$. We seek:
  $$\hat{\theta} = \arg \max_{\theta} P(r | \theta)$$

- Population vector
  $$\bar{P} = \sum_{i=1}^{N} r_i \hat{P}_i$$
  $$\hat{\theta} = \text{angle}(\bar{P})$$

- Winner Take all. Choose the preferred orientation of the neuron with largest response.

Non-optimal Decoders: reducing complexity

Using optimal decoders often requires much too much data.

The question then is what is the cost of using non optimal decoders.

The class of decoders that people work with are:

- Linear Decoders
- Decoders that ignore the correlations (decode with the "wrong model" which assumes independence)

Still very much debated.

(PV and WTA are non optimal but we know they are bad).
Comparing different estimators

- **Bias of the estimator.**
  
  \[ b(\theta) = E[\hat{\theta} | \theta] - \theta \]
  
  If \( E[\hat{\theta} | \theta] = \theta \) the estimator is said to be unbiased.

- **Variance**
  
  \[ \sigma^2_{\hat{\theta}} = E[(\hat{\theta} - \theta)^2 | \theta] - b^2(\theta) \]

- **Fisher Information.** The Crâmer-Rao bound states that for an unbiased estimator:
  
  \[
  I_f(\theta) = -\int P(r | \theta) \frac{\partial^2 \ln P(r | \theta)}{\partial \theta^2} dr \\
  \text{var}(\hat{\theta}) \geq \frac{1}{I_f(\theta)}
  \]
Decoding questions: the philosophical, the bayesian, and the practical

1. Non-optimal decoders.
   The particular case of decoders that ignore correlations.
   Using linear decoders.

2. Does the brain "decode"?
   What are the "neural correlates of consciousness"? [cf Koch's research]

3. Does the brain explicitly encode the likelihood, the prior?
   [Pouget, Dayan, Zemel, Stocker & Simoncelli …]

4. Application: How to build Brain Machines Interfaces?
Revisiting V1 orientation selectivity models:

Coding efficiency
and the Impact of Correlations


Origin of orientation selectivity?

Two classes of models:

1. “Feed-forward” model or “No Sharpening” model
   (Hubel and Wiesel, 1962; Troyer, Krukowski, Priebe and Miller, 1998).
Selectivity is fully present in the LGN inputs to the cortex.

"Feedforward Inhibition" for Contrast Invariance

Origin of orientation selectivity?

Two classes of models:

1. “Feed-forward” model or “No Sharpening” model (Hubel and Wiesel, 1962; Troyer, Krukowski, Priebe and Miller, 1998).

2. “Recurrent” model or “Sharpening” model (Somers, Nelson and Sur 1995; Sompolinsky and Shapley, 1997).

V1

input

Orientation

LGN

Bar Stimulus (θ)

output

ON

OFF

Selectivity is fully present in the LGN inputs to the cortex.
• The validity of these models has been discussed in terms of their physiological and anatomical plausibility:
  - cortical connectivity scheme,
  - thalamocortical connectivity,
  - properties of inhibition in Cx (inactivation)
...
(Sompolinsky and Shapley, 1997; Ferster and Miller, 2000).

• They have never been evaluated and compared in terms of coding efficiency.
1. Is information in the **output** the same in both models?
2. Is information in the **input** the same in both models?

No-Sharpening

Sharpening

3. Is the **format** of the code the same in both models?

No-Sharpening

Sharpening
Mean spiking response and variance are matched

Variability in the distribution of the inter-spike intervals (ISI) is matched
Are these models equivalent in terms of information transmission?

Estimate Fisher Information

→ Measure the discrimination threshold of an “ideal observer”

\[ \hat{\theta} = \sum_i w_i r_i + b \]

\[ d' \text{ measure of discriminability: } d' = \frac{|\langle \hat{\theta}_2 \rangle - \langle \hat{\theta}_1 \rangle|}{\sqrt{\sigma^2_{\hat{\theta}_2} + \sigma^2_{\hat{\theta}_1}}} \]

\[ I_{LOLE} = \left( \frac{d'}{\partial \theta} \right)^2 \]

= Lower bound on Fisher Information / an estimate of the first term of Fisher information when the noise is Gaussian

\[ I(\theta) = \Gamma'(\theta)^T Q^{-1}(\theta) \Gamma'(\theta) + \frac{1}{2} \text{Tr}[Q'(\theta)Q^{-1}(\theta)Q'(\theta)Q^{-1}(\theta)] \]
Information is not equivalent in the two models

OUTPUT of V1

> 6 times as much information in no-sharpening model compared to sharpening model

The information difference in the output is not due to a different of information in the input

INPUT to V1

Input information was approx. the same
The difference in information can be explained by the structure of the covariances of the output responses.

- How much information?
  - more information in no-sharpening model

- What is the format of the information?
- Is there information in the correlations?
Complete knowledge of the statistics of the responses, including the correlations ~ 500,000 parameters.

Optimal “read out”

Assumption: knowledge of the correlations is not necessary → ignore them

~ 1000 parameters for 1000 neurons

1000 neurons  V1

Simplified “read-out”

What is the cost?
How much information is encoded in the correlations?

Information is encoded in the correlations in the S model ≠ NS model

75% is recovered

Robustness

Similar results were obtained with different sharpening and no-sharpening models based on different assumptions about:

- The structure of the connectivity: flat, ≈I, ≈EI, Mexican-Hat.
- The amplitude of the thalamic inputs (and subsequent amplification/suppression)
- The correlations in the thalamic inputs

**Conclusions**

The sharpening and no-sharpening models can produce identical tuning curves (mean, variance).

However:
1. They differ in their **covariance structure**
2. The Sharpening model is **less efficient**
3. The Sharpening model encodes information in the correlations, i.e. a complicated **format**

Contrary to commonly held assumptions, it is not always true that:
- wide tuning curves are worse than narrow ones,
- sharpening can improve the quality of a code

Implications for attention, learning.

**Directions**

So how can the efficiency of the code increase?

1. **Sharpening**
   Other forms of sharpening can exist, surround influences from outside the receptive field

   Caveat: noise model.

2. **Gain Modulation**
   How do amplification and suppression affect accuracy? (attention, perceptual learning, surround modulation)
**Directions**

**Gain Modulation**

How do amplification and suppression affect accuracy?

- Evaluate Physiological models of attention, learning, adaptation, center-surround modulation [Seriès, Lorenceau, Frégnac, 2003]

- Compare with Psychophysical data.

**Extensions**

- Other systems (eg. Barrel cortex)
- extension to coding of multiple features.

- Derive models from optimization principles. Bayesian models of attention, adaptation, surround interactions.

**Analytical studies using the framework of Fisher Information in simplified models.**

**Development of Decoding Algorithms.**
Simplified model of a population (rate) code

- N neurons with a preferred orientation \( \theta \),

- Each neuron is described by # of spikes \( r_i \) in some time window

\[
r_i = f(\theta - \theta_i) + \eta_i
\]  

(1)

- Noise is Gaussian, multiplicative, and neurons are correlated:

\[
Q_{ij}(\theta) = f''(\theta - \theta_i) R_i f''(\theta - \theta_j)
\]

(2)

\[
P[r | \theta] = \frac{1}{\sqrt{(2\pi)^n \text{det}(Q(\theta))}} \exp \left( -\frac{1}{2} (r - f(\theta))^T Q^{-1} \theta (r - f(\theta)) \right)
\]

(3)

• Given the population activity \( r \), what’s the stimulus parameter \( \theta \)? (estimation)

• What’s the smallest difference in \( \theta \) that we can detect, given \( r \)? (discrimination)

• How does it depend on the structure of the noise (variance + covariance)

• What is the effect of pooling more and more neurons on information? (saturation?)

• (When) Is a linear decoder efficient?
Background: Fisher Information

- Fisher Information provides a useful measure of the accuracy of a population code.
  \[ I_f(\theta) = -\int P(r|\theta) \frac{\partial^2 \ln P(r|\theta)}{\partial \theta^2} dr \quad (1) \]

- The Crâmer-Rao bound states that for an unbiased estimator: \( \text{var}(\hat{\theta}) \geq \frac{1}{I_f(\theta)} \)

- Fisher information is related to the discriminability \( d' = \Delta \theta \sqrt{I_f(\theta)} \)

- For Gaussian noise:
  \[ I_f(\theta) = \Gamma(\theta)^T Q^{-1}(\theta) \Gamma(\theta) + \frac{1}{2} \text{Trace} \left( Q^{-1}(\theta) Q^*(\theta) Q^{-1}(\theta) Q^*(\theta) \right) \]
  
  \[ I_{\text{mean}}(\theta) \quad \downarrow \quad I_{\text{cov}}(\theta) \]
  
  “Noise” term, depends on covariance matrix, -- in fact mainly on the variance.

The problem: \( I_{\text{mean}} \) vs \( I_{\text{cov}} \)

- A locally optimal linear estimator (LOLE) is able to extract exactly \( I_{\text{mean}}(\theta) \). A recurrent network can also extract \( I_{\text{mean}}(\theta) \) (Deneve et al, 2001). However, these methods fail to extract \( I_{\text{cov}}(\theta) \).

\[ \Rightarrow \text{Aim of this study:} \]

1 - What is the relative importance of \( I_{\text{cov}}(\theta) \)?
2 - How can we decode it?

- \( I_{\text{cov}}(\theta) \) vanishes for additive noise.
- However, noise seems to be multiplicative: variance \~ k \cdot \text{mean}^2
- How \( I_{\text{cov}} \) and \( I_{\text{mean}} \) compare depends on the correlational structure and exact mean-variance relationship \[ \Rightarrow \text{explore different noise models} \]

**$I_{\text{cov}}$ can dominate for exponentially decaying correlations**

- **a)**
  - Activity vs. Correlation plot showing $I_{\text{cov}}$ and $I_{\text{mean}}$.

- **b)**
  - Information (deg$^{-2}$) vs. Number of neurons plot showing $I_{\text{cov}}$ and $I_{\text{mean}}$.

- **c)**
  - Correlation vs. Information (deg$^{-2}$) plot showing $I_{\text{mean}}$ and $I_{\text{cov}}$.

**$I_{\text{cov}}$ dominates at large $N$. Linear decoding is not sufficient.**

---

**$I_{\text{cov}}$ is large in a pool of identically tuned neurons.**

- **a)**
  - Activity vs. Correlation plot.

- **b)**
  - Information vs. Number of neurons plot.

**Contrary to what is often believed (eg. Zohary et al 1994), there is no saturation of information with $N$ despite the correlations. $I_{\text{cov}}$ dominates at large $N$.**
Analytical study (and simulations) show that:

- Except for very special situations, correlations decrease Fisher Information.
- In particular, correlations induce a saturation of $I_{\text{mean}}$ with $N$. However, $I_{\text{cov}}$ continues to scale with $N$. $\Rightarrow I_{\text{cov}}$ can dominate at large $N$.
- $I_{\text{cov}}$ is almost insensitive to properties of the correlations. It reflects the response variance.
- The higher the mean-variance relationship $\alpha$ and the amplitude of correlations, the more important $I_{\text{cov}}$ relative to $I_{\text{mean}}$.

$\Downarrow$ If noise is $\geq$ Poisson, and large populations of neurons, $I_{\text{cov}}$ matters!
A nonlinear decoder is necessary.

A Quadratic Decoder: 1- local optimality.

- We examine an estimator of the form:

$$\hat{\theta} = \sum_{i=1}^{N} w_i r_i + \sum_{i,j=1}^{N} M_{ij} r_i r_j$$
A Quadratic Decoder: 1- local optimality.

• We examine an estimator of the form:

\[ \hat{\theta} = \sum_{i=1}^{N} w_i r_i + \sum_{i,j=1}^{N} M_{ij} r_i r_j \]

• The vector w and the matrix M (NxN) are the parameters of this estimator. We would like to find the best possible estimator, it should:

(i) be unbiased, \( \frac{\partial \hat{\theta}}{\partial \theta} = 1 \)

(ii) have minimum variance under this constraint.

• When parameters are optimized, if noise is Gaussian and multiplicative, the variance of this estimator is:

\[
\text{var}(\hat{\theta}) = \frac{1}{\text{Fisher Information}} = \frac{1}{r^T Q' r + \frac{1}{2} \text{Trace} (Q' Q^T Q'^T)}
\]
A Quadratic Decoder: 2 – How much data do we need to estimate Fisher Information with a precision of 80%, 90%?

LOQE requires a huge amount of data. Look for tricks to reduce the complexity of this decoder.

3- A Simplified Quadratic Decoder?

- Since $I_{\text{cov}}$ depends primarily on the variance (vs the correlations), we investigated the performance of a decoder of the form:

$$\hat{\theta} = \sum_{i=1}^{N} w_i r_i + \sum_{i=1}^{N} m_i r_i^2$$

- The decoder is optimal for uncorrelated neurons, but not for correlated neurons.

- If used on shuffled data, it can be used to estimate information in $I_{\text{cov}}$. If used in conjunction with linear decoder (to estimate $I_{\text{mean}}$), it can be used to estimate Fisher Information.

- Other tricks are possible too.
One set of 20,000 trials

Independent Neurons

Correlated neurons

LOQE requires a huge amount of data.
SQE (with ‘shuffled trick’) provides a precise estimate while being much less data-intensive.
**Conclusions:**

- Better understanding of the impact of noise structure on population accuracy.
- Development of relevant decoding techniques.

**Directions:**

- More complex models: Beyond the Gaussian assumption; inhomogeneous tuning curves; random correlations coefficients.
- Comparison with related quantities in Shannon information ($I_{corr,dep}$) and link with temporal correlations.
- Does the brain extract this information? If so, biological implementation of a globally optimal estimator [Shamir and Sompolinsky, 2004].
- Correlations in other decoding techniques (eg model-based point process [E. Brown]). Learning data (Machine learning).
- Importance of measuring the 2nd order statistics in electrophysiology. Tuning of the variance?
- Application to simulated circuits, and real data.