# Application of information theory to early visual coding

Zhaoping Li, Feb. 2005.

This lecture notes are modified from my lecture at the Gatsby Tutorial 1999, and my lecture notes for the computational neuroscience course in 1998. See http://www.gatsby.ucl.ac.uk/~zhaoping/visiontutorial.html for more information, and a short and easy to read review article is "Optimal sensory encoding" by L. Zhaoping in *The Handbook of Brain Theory and Neural Networks:* page 815-819, The Second Edition. Michael A. Arbib, Editor MIT Press 2002

**Question**: Signal $S$ + noise $N$ $\overset{\text{coding transform}}{\rightarrow}$ Output $O$ with output noise $N^o$

What should be the coding transform? Why?

e.g.: $O = K(S + N) + N^o$ —— $K$ =?

**Infomax** Optimal $K$ for $O$ to contain most information about $S$ given cost constraints on output $O$.

Cost contraints:,

E.g., (1) optic nerve is long, but has limited thickness or transmission capacity.

(2) The neurons have limited maximum firing frequency.

e.g., $S = \{S_1, ..., S_{1000\times1000}\}$ an image of 1000x1000 pixels, one byte per pixel, with 1 Mbyte date.

Find $K$ s.t. $O = \{O_1, ..., O_{1000\times1000}\}$ 1000 x 1000 pixels with one bit per pixel $\rightarrow\sim$ 0.12 Mbyte (the output cost), with $O$ containing most info. about $S$.

$K_{ij}$ is the receptive field for the output cell $i$.

$I(O;S)$ — information in $O$ about $S$.

e.g., $O = S + N$, $S \in (0, 255)$, $N \in (-0.5, 0.5)$
$I(O;S) \sim \log_2 \frac{256}{1} = 8$ bits.

If $S$, $N$, gaussian with variances $\sigma_S^2$ and $\sigma_N^2$, $\rightarrow O$ gaussian with variance $\sigma_O^2 = \sigma_S^2 + \sigma_N^2$. $I(O;S) \sim \frac{1}{2} \log_2 \frac{\sigma_O^2}{\sigma_N^2}$.

In general $S$, $N$, $O$ are vectors with covariance matrices
$R_{ij}^S \equiv < S_i S_j >$, $R_{ij}^N \equiv < N_i N_j >$ $R_{ij}^O \equiv < O_i O_j >$,
$I(O;S) \sim \frac{1}{2} \log_2 \frac{\det(R^o)}{\det(R^N)}$.

e.g.: $O = K(S + N) + N^o$, $S, N, N^o$ are gaussians and independent from each other.

$$K R^N K^T + R^{N^o} \rightarrow R^N$$

$$K(R^S + R^N) K^T + R^{N^o} \rightarrow R^O$$

Cost — output power $\sum_i < O_i^2 >$

or output channel capacity $\sum_i H(O_i)$ (H: entropy)

For gaussian signals, $H(O_i) \propto \log < O_i^2 > +$ constant.

Infomax: Minimize

$$E = \sum_i < O_i^2 > -\lambda I(O;S) = Tr(R^o) - \frac{\lambda}{2} \log_2 \frac{\det(R^o)}{\det(R^N)}.$$

Infomax gives **efficient coding (data compression)**

Compression possible due to redundancy in $S$, $S_i$ and $S_j$ correlated in images.

$I(O;S) = H(O) - H(O|S)$.

$H(O) \leq \sum_i H(O_i)$,with $=$ when $O_i$ and $O_j$ are independent or non-correlated. Hence, given output cost $\sum_i H(O_i)$, $H(O)$ or $I(O;S)$ can be maximized when $O$ is decorrelated by a coding $K$.

An efficient coding $K$ captures the statistical (correlation) structure in $S$, and gives **cognitive advantages**.

**E.g., stereo coding**:

**Input** $S^L, S^R$ from the left and right eye.

$S^L$ $S^R$

Correlation $R^S = \begin{pmatrix} < S^L S^L > & < S^L S^R > \\ < S^R S^L > & < S^R S^R > \end{pmatrix} = < S^L S^L > \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$

$0 \leq r \leq 1$. (Assuming $< S^L S^L > = < S^R S^R >$).

**Output** $O_1, O_2$ in the visual cortex, $O_i = C_i^L S^L + C_i^R S^R$.

If $C^L = C^R$, the cell is **binocular**.

If $C^L \gg$ or $\ll C^R$, the cell is **monocular**.

Eigenvectors of $R^S$ are $S^{\pm} = S^L \pm S^R$, with eigenvalues $< S^L S^L > (1 \pm r)$.

$S^L, S^R \rightarrow S^+, S^-$, $S^+$ is decorrelated with $S^-$.

The signal power in $\pm$ channels are the eigenvalues of $R^S$, i.e., $(\sigma_S^{\pm})^2 \propto (1 \pm r)$.

$S^+$ is stereo blind, has larger signals, sums from two eyes.

$S^-$ is the stereo informative "edge" signal, difference from two eyes.

Let output $O = K^+ S^+ + K^- S^-$ + noise.

If $|K^+| \gg |K^-|$, the cell is binocular.

If $|K^-| \gg |K^+|$, the cell is monocular.

## Gains for the + and - channels

Let $O^{\pm} = K^{\pm}(S^{\pm} + N^{\pm}) + N^{o\pm}$. For either + or − channel, minimize

$$E = < O^2 > -\lambda I(O;S) = < O^2 > -\frac{\lambda}{2}\log_2 \frac{K^2(\sigma_S^2+\sigma_N^2)+\sigma_{No}^2}{K^2\sigma_N^2+\sigma_{No}^2}$$

Hence $(\sigma_N^2/\sigma_{N^o}^2)K^2 = \begin{cases} \frac{1}{2}\frac{\sigma_S^2}{\sigma_S^2+\sigma_N^2}(1+\sqrt{1+\frac{2\lambda\sigma_N^2}{\sigma_S^2}})-1 & \text{if } K^2 \geq 0 \\ 0 & \text{otherwise} \end{cases}$

For large signal-to-noise ($\sigma_S/\sigma_N \gg 1$), $K \propto 1/\sigma_S$, lower gain to larger signals to save output power $< O^2 >$.
For small signal-to-noise ($\sigma_S/\sigma_N \ll 1$), $K \sim 0$, avoid wasting output power on input noises.

**Signal-to-noise and Scale dependence**

In the $S^+$ and $S^-$ channels, $(\sigma_S^2/\sigma_N^2)_\pm \propto (1 \pm r)$.

$(\sigma_S^2/\sigma_N^2)_+ > (\sigma_S^2/\sigma_N^2)_-$

In the cortex, receptive fields come in different sizes or scales (see later). Input signals are stronger for larger scales.

For large scale, when $(\sigma_S^2/\sigma_N^2)_\pm \gg 1$, $K^- > K^+$ — monocular cells.

For small scale, when $(\sigma_S^2/\sigma_N^2)_\pm \sim< 1$, $K^- < K^+$ — binocular cells.

**Multiplexing the channels**

Most cortical cells $O = C^L S^L + C^R S^R$ have $C^L \neq \pm C^R$, i.e., not strictly binocular or monocular.

The $+$ and $-$ channels are multiplexed, i.e.,

$$\begin{pmatrix} S^L \\ S^R \end{pmatrix} \rightarrow \begin{pmatrix} S^+ \\ S^- \end{pmatrix} \rightarrow \begin{pmatrix} K^+ S^+ \\ K^- S^- \end{pmatrix} \rightarrow$$
$$\begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} K^+ S^+ \\ K^- S^- \end{pmatrix}$$

for arbitrary $\alpha$.

The matrix $U = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}$ is an unitary (orthonormal) matrix $UU^\dagger = 1$ in 2-dimension.

**Infomax for gaussian signals** $E = \sum_i < O_i^2 > -\lambda I(O;S)$
$= Tr(R^o) - \frac{\lambda}{2} \log_2 \frac{\det(R^o)}{\det(R^N)}$

**is invariant to such transform** $K \rightarrow UK$, since trace and determinant of matrices are invariant to this transform.

Since

$R^o = K(R^S + R^{N_{input}})K^T + R^{N_{output}}$.

$R^N = KR^{N_{input}}K^T + R^{N_{output}}$.

Many different coding $K$ can minimize $E$ to the same degree. There could be different sets of receptive fields employed by the visual system for efficient coding purposes. (The visual system must choose one particular set for one particular purpose).

(There is no unique seperation of blind sources if the source signals are gaussian).

Mixing the + and − channels:

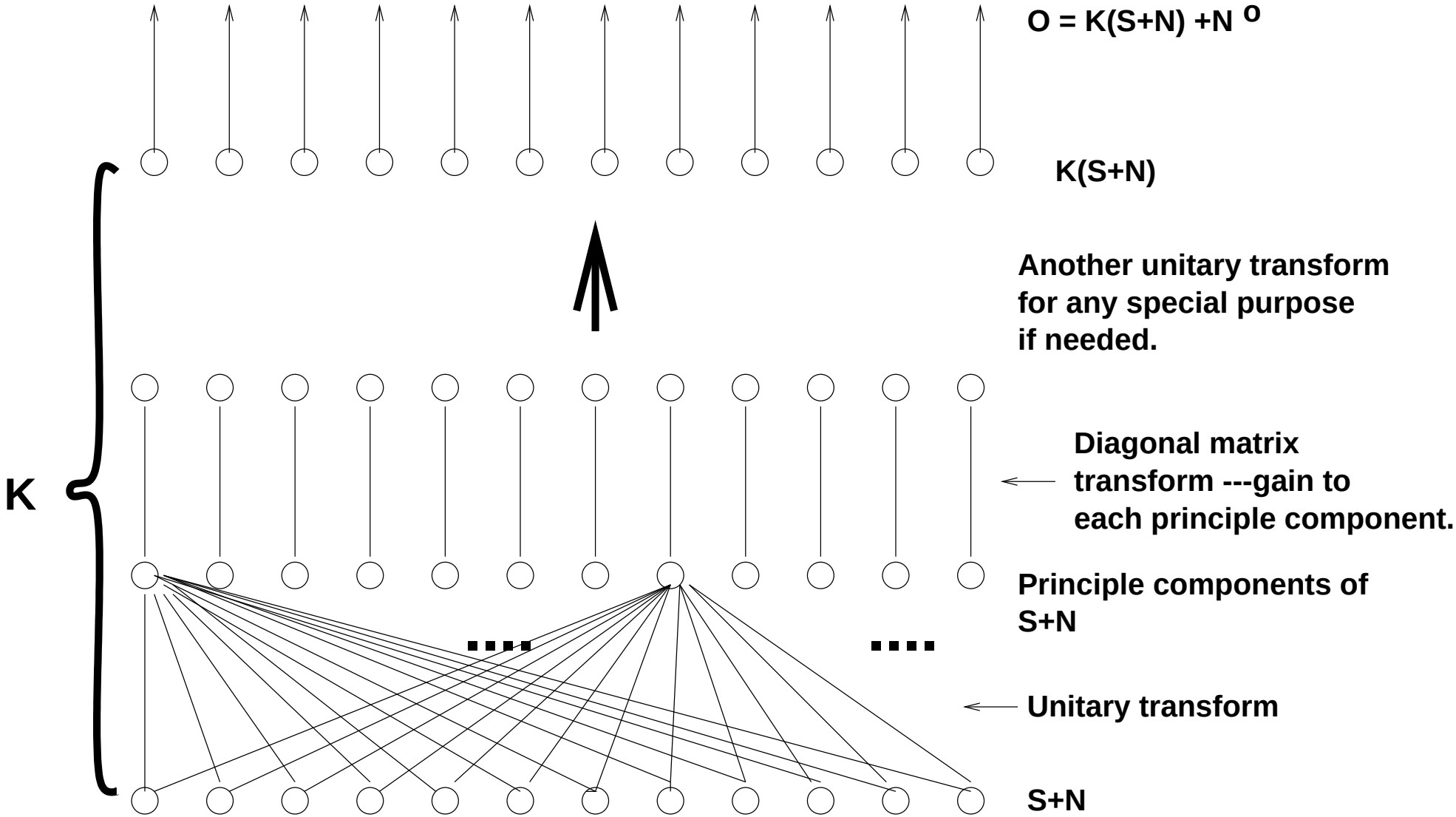
$$\begin{aligned} O &= \cos(\alpha)K^{+}S^{+} + \sin(\alpha)K^{-}S^{-} \\ &= (\cos(\alpha)K^{+} + \sin(\alpha)K^{-})\mathbf{S^{L}} \\ &+ (\cos(\alpha)K^{+} - \sin(\alpha)K^{-})\mathbf{S^{R}} \end{aligned}$$

In general, all ocularities (combinations of $S^L, S^R$) are possible.

When $K^+ \gg K^-$, $O \sim\propto (S^L + S^R)$ most cells are binocular.
When $K^+ \ll K^-$, $O \sim\propto (S^L - S^R)$, leads to many monocular (often nonlinear) cells.

As observed in the cortex.

# Schematic of the steps to obtain infomax code



**Color coding** — analogous to stereo coding

Input $S^{red}, S^{green}, S^{blue}$ are correlated (3 x 3 correlation matrix).

$$R^S = \begin{pmatrix} < S^r S^r > & < S^r S^g > & < S^r S^b > \\ < S^g S^r > & < S^g S^g > & < S^g S^b > \\ < S^b S^r > & < S^b S^g > & < S^b S^b > \end{pmatrix}$$

$$\begin{pmatrix} S^r \\ S^g \\ S^b \end{pmatrix} \overset{decorrelate}{\rightarrow} \sim \begin{pmatrix} a^r S^r + a^g S^g + a^b S^b \\ b^r S^r + b^g S^g - b^b S^b \\ c^r S^r - c^g S^g \end{pmatrix} = \begin{pmatrix} luminance \\ chrominance_1 \\ chrominance_2 \end{pmatrix}$$

The luminance channel is gray level (black-white, or color-blind), has highest signal power and most information (see later).
The two chrominance channels (color selective) are (1) yellow - blue , (2) red - green. They have lower signal-to-noise.

**Different gains to Luminance/Chrominance channels from Infomax**

$$\begin{pmatrix} luminance \\ chrominance_1 \\ chrominance_2 \end{pmatrix} \overset{gain}{\rightarrow} \begin{pmatrix} K^l \cdot luminance \\ K^{c1} \cdot chrominance_1 \\ K^{c2} \cdot chrominance_2 \end{pmatrix}$$

Since $\sigma_S/\sigma_N(Chrominance) < \sigma_S/\sigma_N(Luminance)$,

$K^l < K^{c1}, K^{c2}$ — at large scale when $\sigma_S/\sigma_N \gg 1$,
$K^l > K^{c1}, K^{c2}$ — at small scale when $\sigma_S/\sigma_N \ll 1$,

Multiplex $\rightarrow$ various color selective or color blind cells.

In cortex: color blind cells for small scales, and color selective cells for large scales.
In retina, — e.g., red-center-green-surround receptive fields (later).

**Retinal space coding**

**Input statistics**:

Input $S = \{S_1, S_2, ..., S_N\}$ from photoreceptors.

Output $O = \{O_1, O_2, ..., O_N\}$ at ganglion cells.

$R^S_{ab} \equiv < S_a S_b >$ decays with distance $a - b$, depends only on $a - b$. Denote $R^S(a - b) \equiv R^S_{ab}$.

Eigenvectors of $R^S$ are Fourier waves.

i.e., the Fourier component $S'_k \equiv \sum_a S_a e^{ika}$, for different frequency $k$ are decorrelated.

Eigenvalues of $R^S$ are Fourier transform of $R^S(a)$, i.e., $\sigma^2_S(k) = < (S'_k)^2 > \propto \sum_a R^S(a) e^{ika}$.

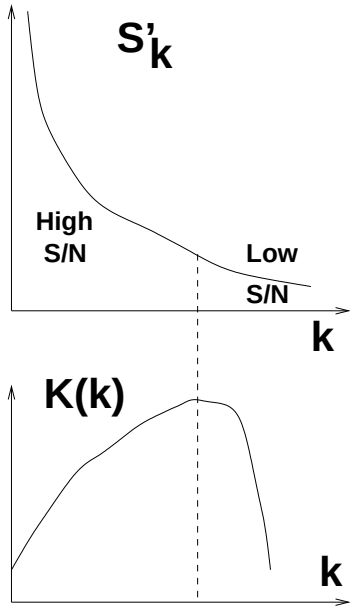
$< (S'_k)^2 > \propto 1/k^2$, decays with frequency $k$.

$\rightarrow$ The gain $K(k)$ for $S'_k$:

$K(k) \propto k$ for small $k$ where signal-to-noise is large
$(< (S'_k)^2 > /\sigma_N^2 \gg 1)$

$K(k)$ decreases with $k$ for large $k$ where signal-to-noise is small
$(< (S'_k)^2 > /\sigma_N^2 <\sim 1)$

$\rightarrow$ $K(k)$ band pass in $k$.



$$
\begin{aligned}
S_a &\rightarrow S'_k = \sum_a e^{ika} S_a \\
&\rightarrow O(k) \equiv K(k) S'_k \\
&= K(k) \sum_a e^{ika} S_a
\end{aligned}
$$

Hence, one could achieve infomax by constructing $k = 1, 2, ..., N$ different receptive fields, each is a Fourier wave shaped, infinitely large, i.e., for the $k^{th}$ output $O(k)$, the synaptic weight to the $a^{th}$ input pixel $S_a$ is:

$K_{ka} = K(k) e^{ika}$

However, the visual system does not choose this code.

Ganglion cell receptive fields are not large Fourier waves, but are small sized, center-surround shaped. They arise from multiplexing the large Fourier fields with gains $K(k)$. The multiplexing unitary transform is the inverse Fourier transform.

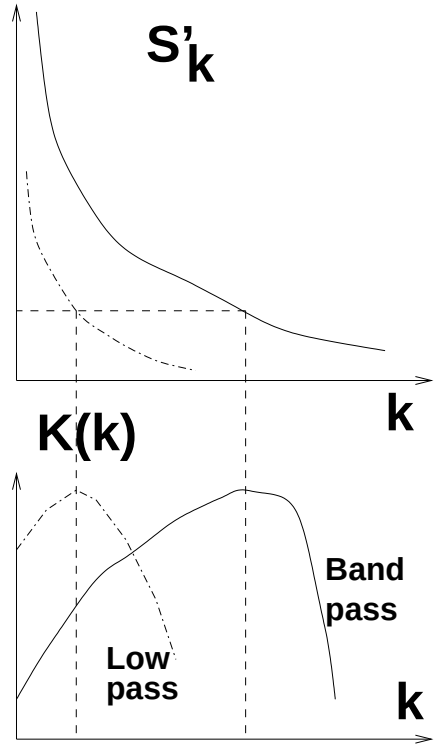
Hence, the coding transform is:

$$O_b \equiv \sum_k e^{-ikb} O(k) = \sum_a (\sum_k e^{-ik(b-a)} K(k)) S_a$$

Hence, the receptive field for $b^{th}$ output cell is is $\sum_k e^{-ik(b-a)} K(k)$ — a band-pass spatial filter, center-surround shaped. It only depends on $b - a$ — receptive fields for all cells $b = 1, 2, ..., N$ has the same shape except for a shift in center position $b$.

## Adapting the receptive field to input intensity

When the signal power $< (S'(k))^2 >= S^2/k^2$, or $S^2$, decreases, the band-pass becomes low-pass, the center-surround (which is a band-pass filter) becomes gaussian shaped (a low-pass filter). This indeed happens to the retinal receptive fields.



## Temporal coding in early vision

Input $S_t$ for time $t = -\infty, ..., 0, 1, 2, ...$ Output $O_t = \sum_{t' \leq t} K(t - t') S_{t'}$

— $K(t - t')$, causal, i.e., $K(t - t') = 0$ if $t' > t$.

Coding transform steps:

$$
\begin{array}{lll}
S_t & \rightarrow & S'_\omega = \sum_t e^{i\omega t} S_t \rightarrow K(\omega) S'_\omega \\
 & \overset{\text{multiplex}}{\rightarrow} & \sum_\omega C(\omega) K(\omega) S'_\omega \\
 & \overset{C(\omega) = e^{-i\omega t + \phi(\omega)}}{\rightarrow} & \sum_{t'} (\sum_\omega K(\omega) e^{i\omega(t - t') + \phi(\omega)}) S_{t'}
\end{array}
$$

Hence $O(t) = \sum_{t'} (\sum_\omega K(\omega) e^{i\omega(t-t') + \phi(\omega)}) S_{t'}$ Choose $\phi(\omega)$ such that $K(t - t') \equiv \sum_\omega K(\omega) e^{i\omega(t-t') + \phi(\omega)}$ is causual and has a finite duration.

**Transiant or sustained impulse responses**

$K(t-t')$ is also the impulse response of the cell.

$S'_\omega$ has a signal power $< (S'_\omega)^2 >$ that also decays with $\omega$. Hence $K(\omega)$ is a band-pass or low pass in $\omega$ depending on the overall ensemble input intensity.

Under high input intensity (large receptive fields), $K(\omega)$ is band-pass, impulse response $K(t-t')$ is transient.

Under low input intensity (smaller receptive fields), $K(\omega)$ is band-pass, impulse response $K(t-t')$ is more sustained.

## Multiscale coding in V1

- Alternatives to the spatial coding $K = U_s^T V U_s$, is of course, first the $K = V U_s$.

- Between the two extremes: $K = U_s^T V U_s$ and $K = V U_s$, there is a multiscale coding.

- Related to the wavelet coding, with translation and scale invariance (under noiseless case, when $K(k) \propto k$).

  If $S'(x) = S(x + \delta x) \rightarrow O_n^a[S] = O_{n+\delta n}^a[S']$.

  if $S'(x) = S(\lambda x)$, $\rightarrow O_n^a[S] = O_n^{a+\delta a}[S']$

  Receptive fields in different scales are scaled versions of each other. $K^a(x) \propto \int_{k_a}^{k_a+1} dk k \cos(kx + \phi_a)$.

$$K^a(x_n^a - \lambda x) = \frac{1}{\lambda^2} K^{a+1}(x_n^{a+1} - x).$$

- The phase of the receptive field $\phi_a$ is another freedom allowed by the $U$ symmetry.

- One of the choices for spatial frequency bandwidth is $\log 3 \sim 1.5$ octaves. – $k_{a+1}/k_a = 3$.

- Orientation selectivity is a consequence.

- Quadrature structure: both edge and bar detectors, is also a consequence, in order that the receptive fields within a scale is roughly "translation invariant".

Zhaoping Li and J. J. Atick "Towards a theory of striate cortex" *Neural Computation* **6, 127-146** (1994). Available on my web page

**Color coding in V1, the multiscale representation**

- Only cells selective to low spatial frequencies are color selective.

- the Double opponency color selective cells in V1, rather than the single opponency color selective cells in retina.

## Stereo coding in V1, the multiscale representation

- The signals from the two eyes only start to mix in V1.

- Input signal $S(x, y, e)$, where $e = left, right$.

  $< S(x, y, e)S(x', y', e') > \neq R(x - x', y - y')R^{e,e'}$

  $< S(x, y, e)S(x', y', e') > \equiv R^{e,e'}(x - x', y - y')$

  Correlation matrix $R^{e,e'}(k_i)$.

- Stereo edge signal $L - R$, the '-' channel, stereo-blind signal $L + R$, the '+' channel. The 2 receptive fields

  $K^{1,2} = K^{+} \pm K^{-} = L(K^{+} \pm K^{-}) + R(K^{+} \mp K^{-})$,

$K^L = K^+ \pm K^- = |K^L| e^{i\phi_L}$
$K^R == K^+ \mp K^- = |K^R| e^{i\phi_R}$

Disparity $\Delta\phi = \phi_L - \phi_R$

$\Delta\phi > 90^o$, if $K^- > K^+$ $\Delta\phi < 90^o$, if $K^- < K^+$

- Stereo coding: ocular dominance and disparity selectivity.

- $K^- > K^+$ when spatial freq. $k$ is small, where S/N is high.

  $K^- < K^+$ when spatial freq. $k$ is large, where S/N is low.

- Correlation of stereo coding with the preferred spatial frequency and orientation of the V1 cell.

Zhaoping Li and J. J. Atick "Efficient stereo coding in the multiscale representation" *Network* Vol.5 1-18. (1994). Available on my web site

## Motion coding in V1, the multiscale representation

- Directional selectivity as consequence:

  $K(x, y, t) =$
  $\int d\omega dk k(k, \omega)(A^{+} \cos(kx + \omega t + \phi^{+}) + A^{-} \cos(kx - \omega t + \phi^{-}))$

  When $A^{\pm} = 0$, there is preferred directions of motion for the stimulus: $x = \pm(\omega/k)t$

- $A^{+}/A^{-}$ can take many different choices allowed by $U$ in $K = UVU_s$.

- When $A^{+} = A^{-}$, the receptive field can be almost space-time seperable.

- Cells selective to higher spatial frequency $k$ is selective to lower temporal frequency $\omega$, making motion or direction sensitivity less important for these cells.

- Cells can be selective to motion in depth, mostly for cells with large receptive fields.

$$K_{L,R} \propto \int dk \int d\omega K_{L,R}(k,\omega)(A^{+}_{L,R}\cos(kx+\omega t+\phi^{+}_{L,R}) + A^{-}_{L,R}\cos(kx-\omega t+\phi^{-}_{L,R}))$$

Zhaoping Li "A theory of the visual motion coding in the primary visual cortex" *Neural Computation* vol. 8, no.4, p705-30, May, 1995,. Available on my web page

**Coupling the space and color coding**

Retina receptive fields are red-center-green-surround, how come?

Input $S_{ci}$, $c = red, green$ (ignore blue for simplicity), $i = 1, 2, ..., N$ (spatial location).

The transforms:

$S_{ci} \rightarrow S_{Ci}$, $C =$ red+green (luminance) , $C =$ red-green (chrominance), $i = 1, 2, ..., N$.

For each channel $C$,

$S_{Ci} \rightarrow O_{Ci}$, where $O_{Ci}$ has receptive field in space for channel $C$.

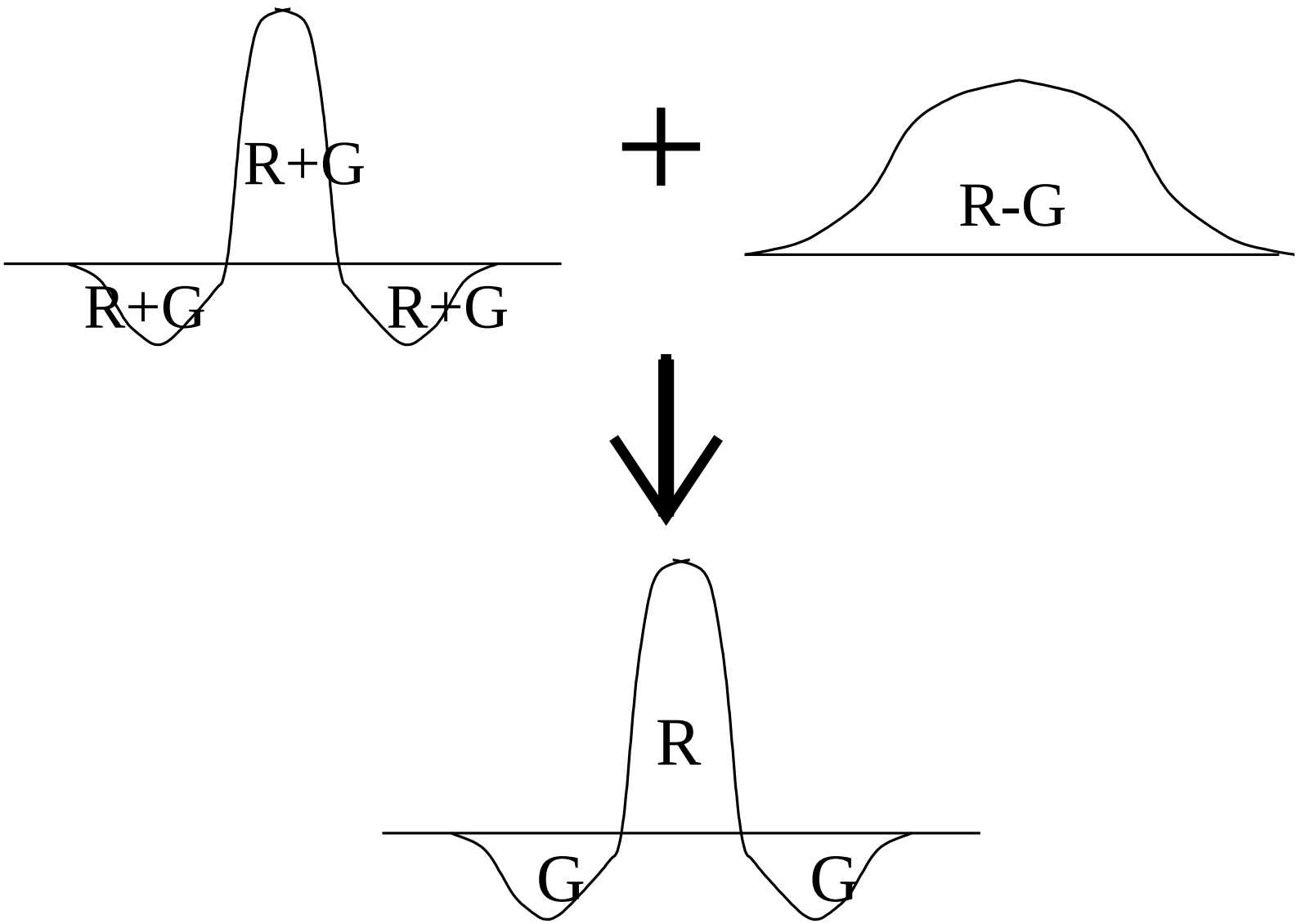
If $C =$ Luminance, (red+green), signal-to-noise is large, $O_{Ci}$ has center-surround receptive field.

If $C =$ Chrominance (red-green), signal-to-noise is small, $O_{Ci}$ has gaussian shaped receptive field.

Multiplex $O_{Luminance,i}$ and $O_{Chrominance,i} \rightarrow$

$O_{Luminance,i} + O_{Chrominance,i}$ $O_{Luminance,i} - O_{Chrominance,i}$

$\rightarrow$ red-center-green-surround (or green-center-red-surround) receptive fields.


R+G
R+G
R+G
+
R-G
R
G
G

**Coupling visual coding in space, time, color, stereo**

One could generalize, and obtain cells with receptive field properties in all these dimensions. E.g., cells tuned to motion in depth, tuned to orientation and direction, to color and scale, etc. A whole spectrum of tunings can be found in the cortex.

One also finds the correlations between tunings to different dimensions. E.g., correlation between

color-selectivity and orientation non-selectivity color-selectivity and motion insensitvity ocularity and orientation selectivity receptive field size and motion sensitivity

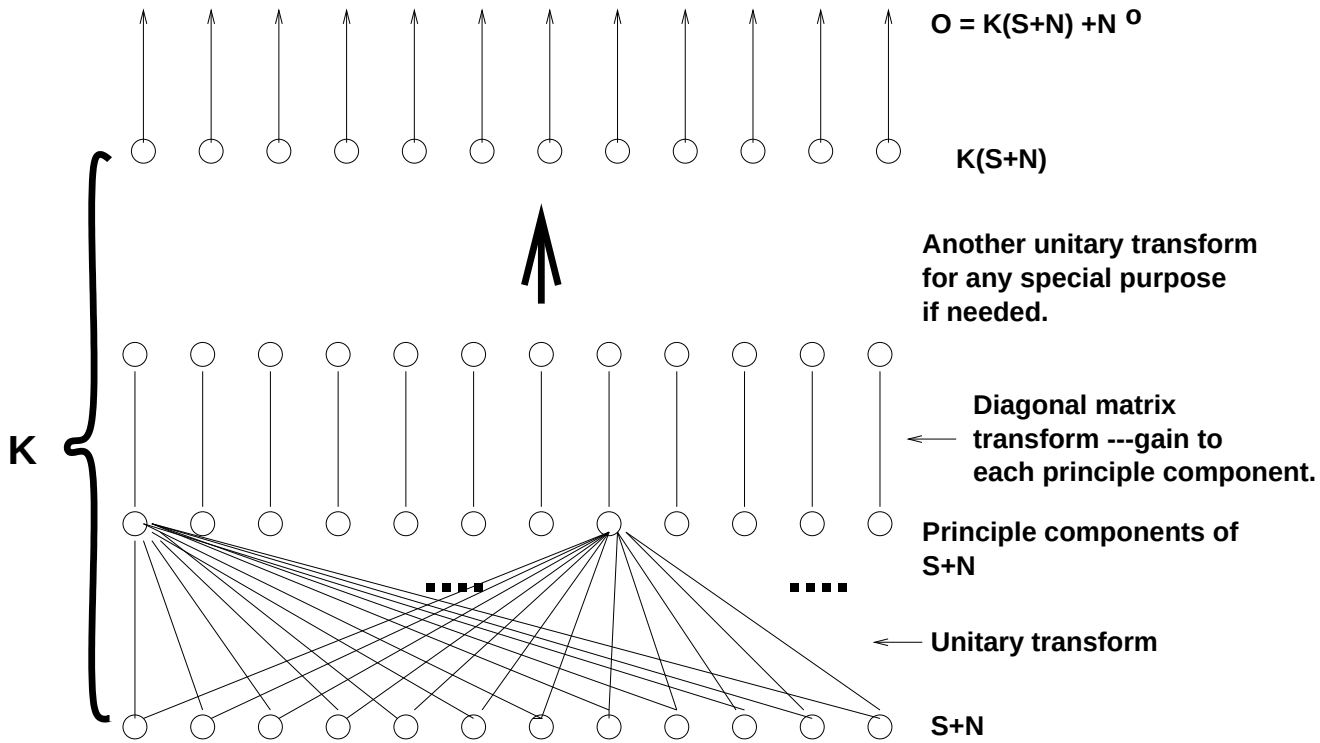
etc.

## Over-complete sampling in V1

Why?

## Some frequently used terms and concepts

- **Whitening** – $K = UVU_s$, with $V_{ii}$ as the gain for the $i^{th}$ principle component.



$$(\sigma_N^2/\sigma_{N^o}^2)v_{ii}^2 = \begin{cases} \frac{1}{2}\frac{R_i}{R_i+\sigma_N^2}(1+\sqrt{1+\frac{2\lambda\sigma_N^2}{R_i}})-1 & \text{if } v_{ii}^2 \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $R_i$ is the eigenvalue or the signal power of the $i^{th}$ principle component of $S$. When $R_i/\sigma_N^2 \to \infty$, $V_{ii} \propto (R_i)^{1/2}$, hence before the transform $U$, $< O_i O_j >= \delta_{ij}$ constant. The outputs are not only uncorrelated, but of equal variance — like **White noise**. Each channel is filled to equal capacity or dynamic range.

- **Decorrelation** — This means rigorously $P(\mathbf{X}) = \Pi_i P(X_i)$, hence $< (X_i - < X_i >)^a (X_j - < X_j >)^b \propto \delta_{ij}$ for all $a, b \neq 0$. Sometimes, people also say decorrelation for second order decorrelation only, when $a = b = 1$.

- **Redundancy** only means $\sum_i H(X_i) > H(X)$. Hence, it is equivalent to decorrelation. However, in the literature,

people only use the terms confusingly, as if redundancy reduction and decorrelation are two different things. Of course, if decorrelation is only restricted to second order decorrelation, then redundancy reduction and decorrelation are *not* exactly the same.

- **Efficient coding** — they can be different definition of efficient coding. For instance, some efficient coding means to minimize $\sum_i < O_i^2 > -\lambda I(O;S)$, others may mean to minimize $\sum_i H(O_i) - \lambda I(O;S)$. These differ in the definition of the cost to coding. And they may be other definition. Hence, sometimes, efficient coding is equivalent to redundancy reduction or to decorrelation, sometimes it is not. It is not uncommon to find statements and claims in literature without clear definitions of what the meanings are.

- **Informax** — often means maximizing $I(O;S)$, subjecting to some constraints. Hence, it can mean the same as (or different from) efficient coding or redundancy reduction or decorrelation depending on the definitions.

- **Minimum predictability or least mutual information between output units** — often this means to minimize $I(O_i;O_j)$. Hence, this is equivalent to minimize $\sum_i H(O_i)/H(O)$, or to reduce redundancy.

- **Sparse coding** — often not clearly defined, but it is often understood as meaning to minimize the number of coding units $O_i$ to be active. Let us make it more precise. Let $O_i = 0$ (or any discrete or definite token) as the usual definition of being inactive, and $O_i \neq 0$ meaning to be active. Given an input $S$, let $P(O_i \neq 0|S)$ be the prob. of $O_i$

active. Sparse coding can mean to minimize $\sum_S P(S) \sum_i P(O_i \neq 0|S) = \sum_i P(O_i \neq 0)$. However, the code should also convey input information, meaning $I(O;S)$ should be maintained, otherwise, the sparsiest code is $P(O_i \neq 0|S) = 0$ for all $S$. When $P(O_i \neq 0) < 1/2$, minimizing $P(O_i \neq 0)$ is the same as minimizing $H(O_i)$. Hence, minimizing $\sum_i P(O_i \neq 0)$ is not too far from minimizing $\sum_i H(O_i)$ (subjecting to a constraint on $I(O;S)$). Hence, sparse coding and redundancy reduction is often related or even equivalent depending on the definition.

**Some related terms**

Efficient codes, decorrelation, independence, non-redundant codes, predictive codes, sparse codes, PCA, factor analysis, ICA, MDL, Whitening, minimum mutual information between

output units, second order vs. higher order correlations, degeneracy or non-unique-ness in blind source seperation.

In the literature of last 10 years or more, there have been many different proposals for principles of coding. The terms above have been among some of the "principles", even though many of them are equivalent or special cases of one another.

## What are the M cells for?

One proposal: To extract information fast, not in terms of information rate (bits/per second), but how long $\tau$ one has to wait before one can extract enough information from signals $O(t \leq t + \tau)$ about input stimulus $S(t)$.

Define $O^t = \{O(t), O(t-1), O(t-2), ...\}$, and similarly $S^t$. Then $I(t, t') \equiv I(t - t') \equiv I(O^t; S^{t'}) - I(O^t; S^{t'-1})$ is the amount of information in $O^t$ about stimulus $S(t')$ that is unpredictable from previous stimulus $S^{t'-1}$.

Maximize $F \equiv \sum_{t \geq 0} \frac{dI(t)}{dt} e^{-t/T}$ subject to cost constraints.

Implications for top-down feedback and top-down bottom up iterations, from M pathway to P pathway feedback.

see Zhaoping Li "Different retinal ganglion cells have different functional goals"

*International J. of Neural Systems* Vol. 3, No.3 (1992) 237-248.

## Other things I have not talked about

- Information rate per neuron, per spike, noise, reproducibility.
- Attentional bottle neck.
- Similar approaches to other sensory systems, audition, olfaction, other animals.

## References


Barlow H. B. " Possible principles underlying the transformation of sensory messages" *Sensory communication* Ed. W. A. Rosenblith, Cambridge, MA, MIT Press, 1961.

Linsker R. "Perceptual neural organization: some approaches based on network models and information theory." *Annual Rev. Neurosci.* 1990: 13: 257-81.

J.J. Atick, Zhaoping Li, and A.N. Redlich "Understanding retinal color coding from first principles" *Neural Computation* 4. 559-572 (1992).

Zhaoping Li and J. J. Atick "Towards a theory of striate cortex" *Neural Computation* **6, 127-146** (1994).

Zhaoping Li and J. J. Atick "Efficient stereo coding in the multiscale representation" *Network* Vol.5 1-18. (1994).

Zhaoping Li "A theory of the visual motion coding in the primary visual cortex" *Neural Computation* vol. 8, no.4, p705-30, May, 1995,.


Some of the papers are available on webpage: http:// www.gatsby.ucl.ac.uk/~ zhaoping