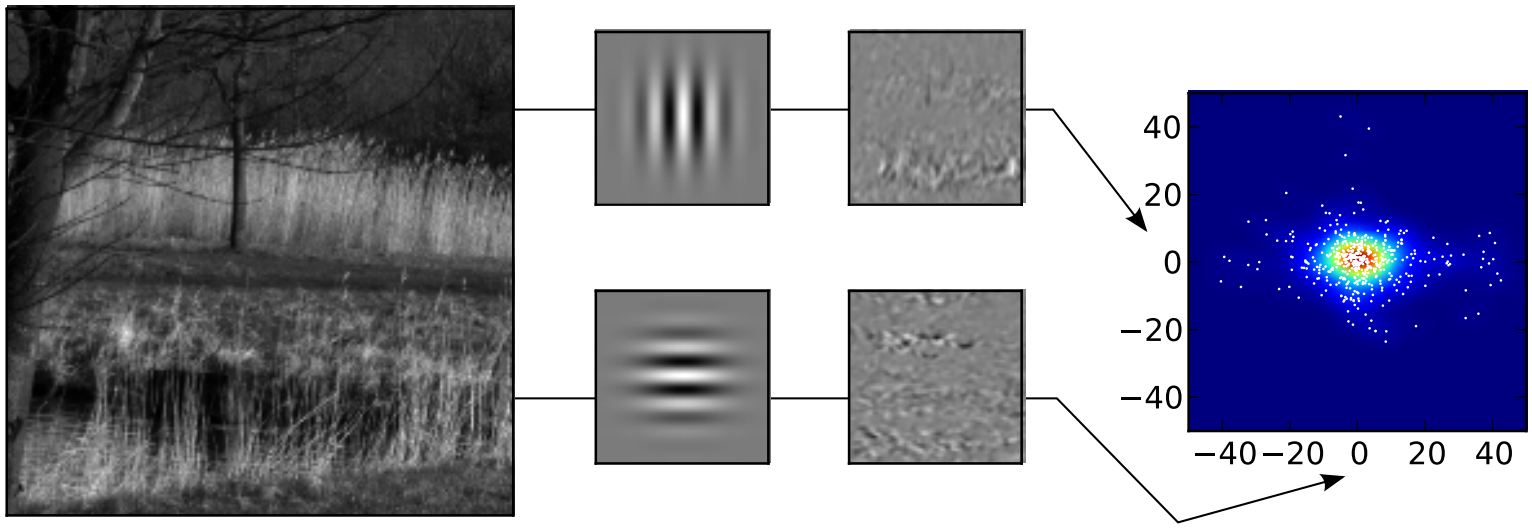
Natural Image Statistics

Models of Higher Brain Function

April 29, and May 20, 2010



Keywords I assume to be known

- Matrix multiplication, Matrix inverse

$$(\mathbf{Ax})_j = \sum_{i=1}^{n} A_{ji} x_i, \quad \mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{1}.$$

- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution
- random variable, expectation and variance
- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform

$$\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{1}$$

- probability distribution, probability density function, normal/gaussian distribution
- random variable, expectation and variance
- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform

$$\mathbf{A} = \mathbf{A}^T \implies \exists \mathbf{U} \text{ orthogonal}, \exists \mathbf{\Sigma} \text{ diagonal} : \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T.$$

- probability distribution, probability density function, normal/gaussian distribution
- random variable, expectation and variance
- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution

$$p(\mathcal{A})$$

- random variable, expectation and variance
- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution

$$p(\mathcal{A}) = \int_{\mathcal{A}} f(x)\,\mathrm{d}x.$$

- random variable, expectation and variance
- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution

$$f_{\text{normal}}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right).$$

- random variable, expectation and variance
- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution

$$f_{\text{normal}}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right).$$

- random variable, expectation and variance
- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution
- random variable, expectation and variance

$$X : \Omega \to \mathbb{R}$$

- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution
- random variable, expectation and variance

$$\mathbb{E}(X) = \int X \, \mathrm{d}p \triangleq \int_{-\infty}^{\infty} xp(x) \, \mathrm{d}x.$$

- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution
- random variable, expectation and variance

$$\mathrm{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

- complex numbers, fourier transform

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution
- random variable, expectation and variance
- complex numbers, fourier transform

$$z = x + iy, \quad i = \sqrt{-1}.$$

Keywords I assume to be known

- Matrix multiplication, Matrix inverse
- Orthogonality, Principal component transform
- probability distribution, probability density function, normal/gaussian distribution
- random variable, expectation and variance
- complex numbers, fourier transform

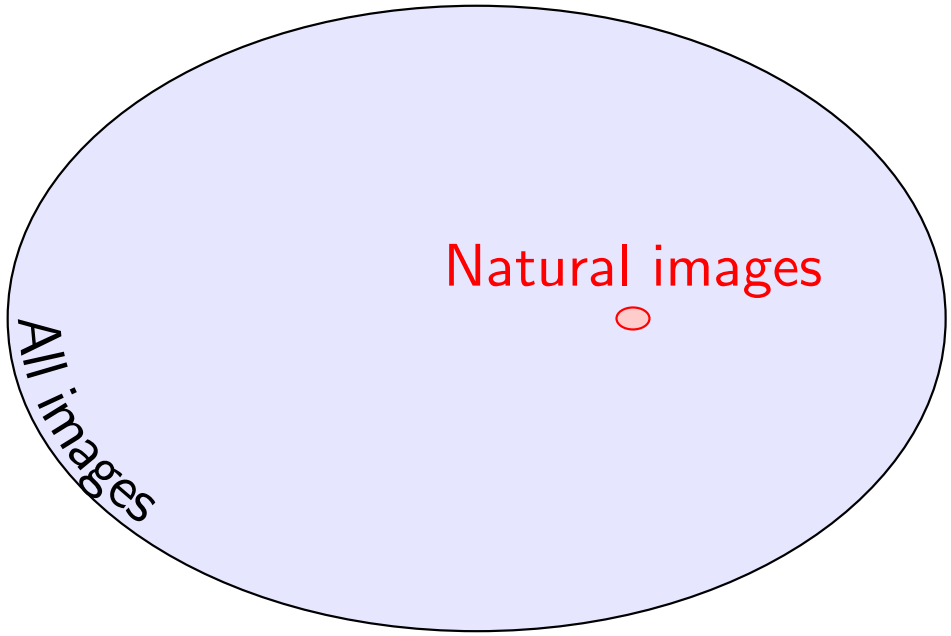
$$\mathcal{F}f(\xi) = \int f(x) \exp(-2\pi i \langle x, \xi \rangle) \, \mathrm{d}x$$

Outline

Why should we care?

Experimental perspective

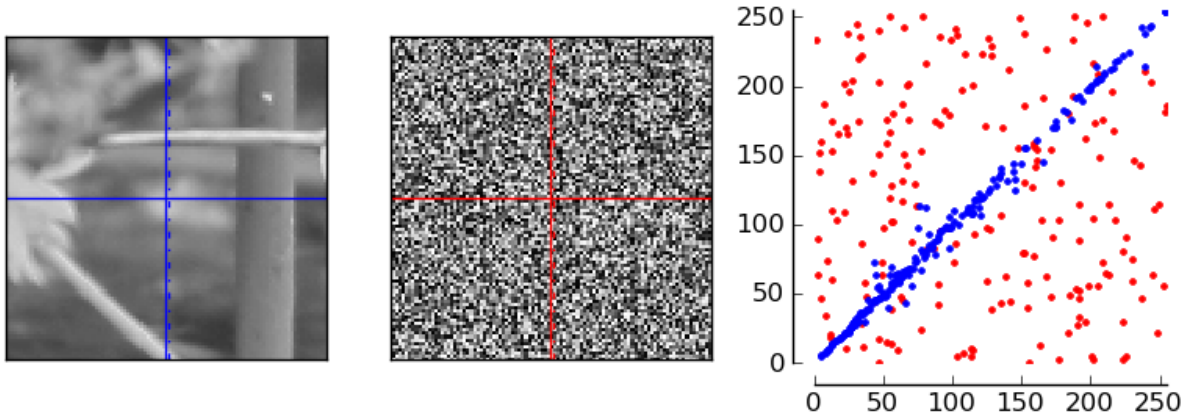
- The number of possible images is incredibly large, natural images form a very small part of these images.



Why should we care?

Experimental perspective

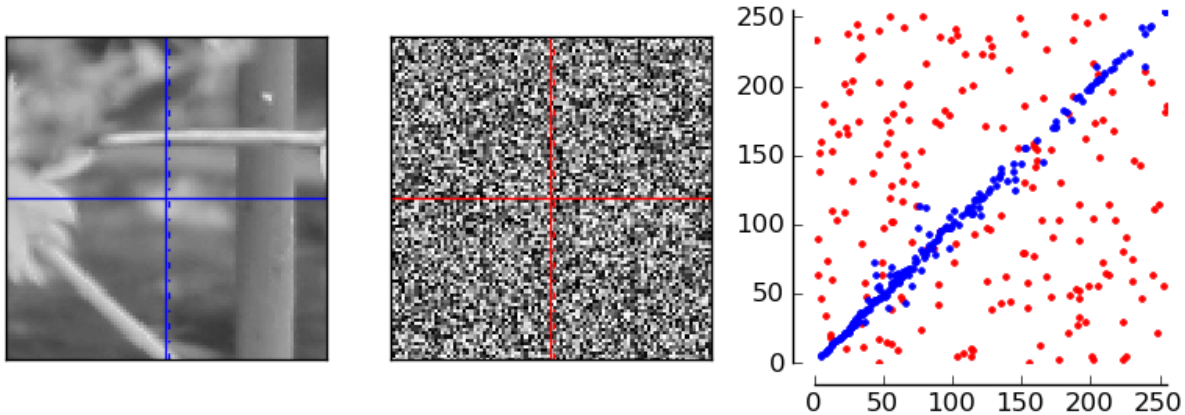
- The number of possible images is incredibly large, natural images form a very small part of these images.



Why should we care?

Experimental perspective

- The number of possible images is incredibly large, natural images form a very small part of these images.

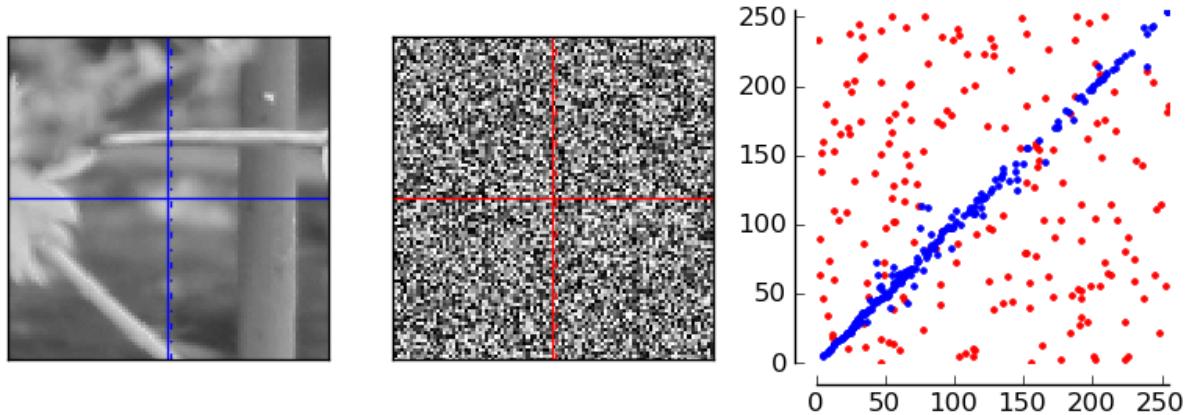


- It is impossible to study the brain's responses for all possible images.

Why should we care?

Experimental perspective

- The number of possible images is incredibly large, natural images form a very small part of these images.



- It is impossible to study the brain's responses for all possible images.

$\rightsquigarrow$ Restrict research to those images that are relevant, i.e. natural images.

Why should we care?

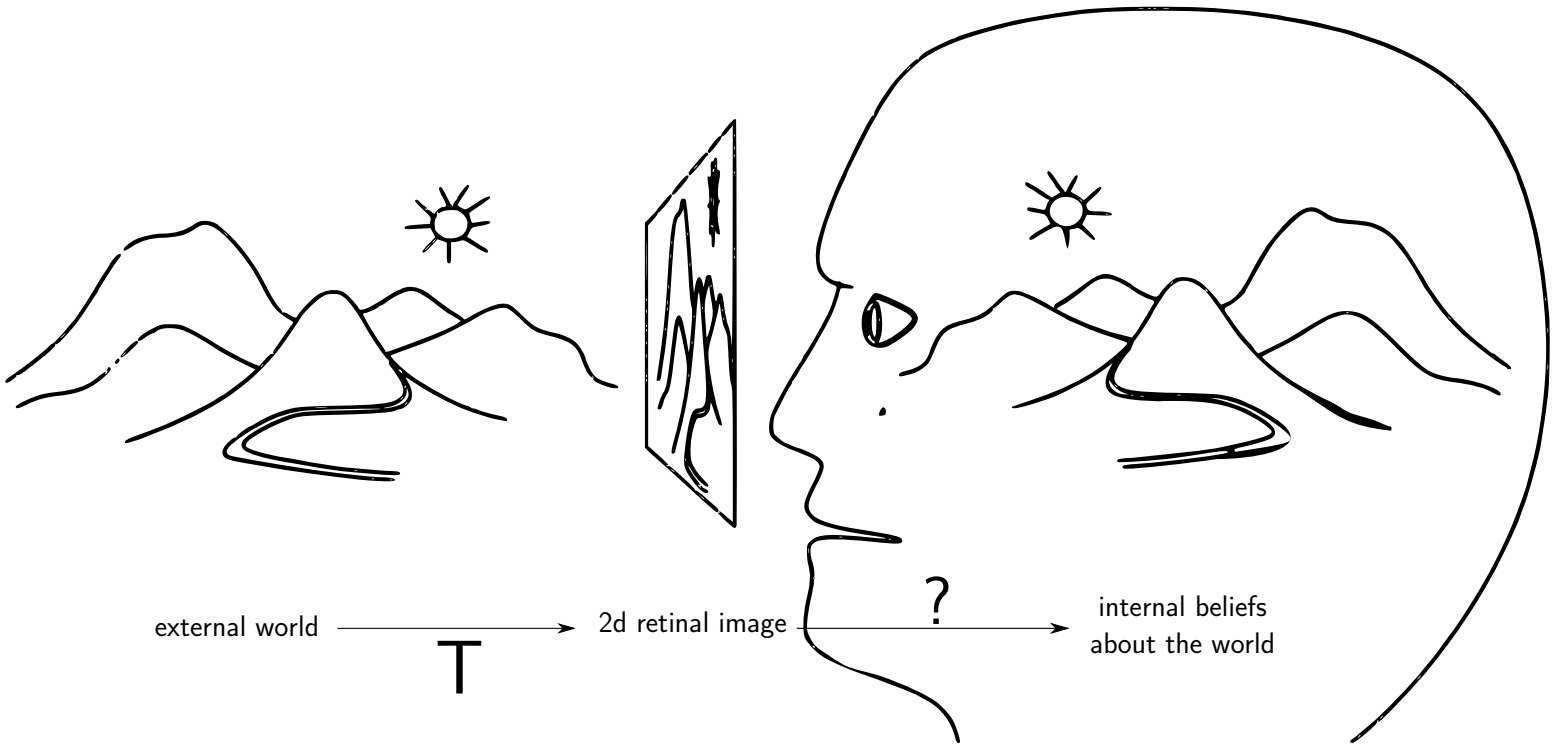
Experimental perspective

Aims from an experimental perspective

- "Coordinate system" for natural images to give quantitative descriptions of natural image
- Generative model: Sampling from the model generates images that can be considered "natural".

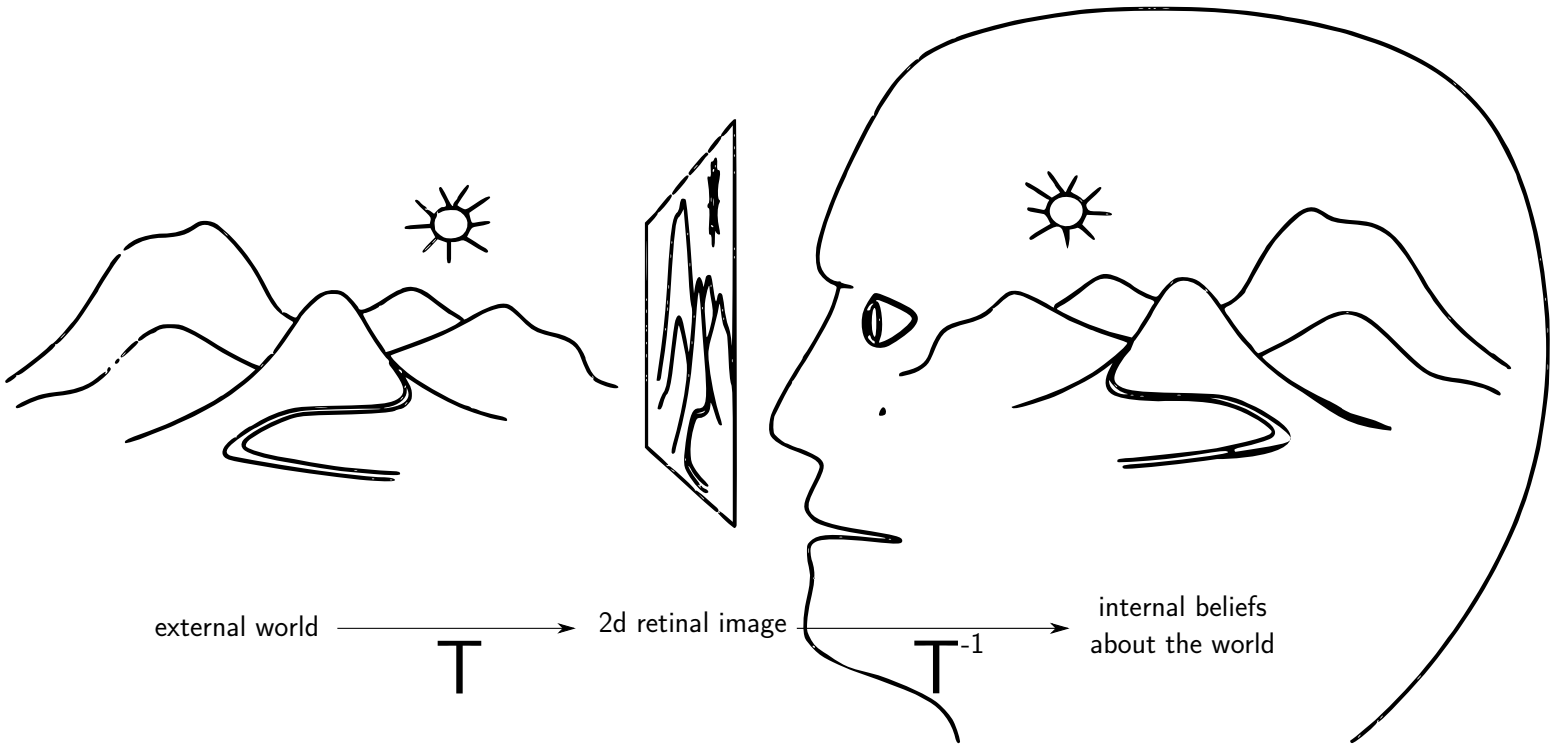
Why should we care?

Theoretical perspective



Why should we care?

Theoretical perspective



Why should we care?

Theoretical perspective

Normative modeling

How would an *optimal* system perform what the brain does with natural images? This is typically about optimal *coding, transmission, or memory*.

- What does optimal coding, transmission, or memory mean?

Why should we care?

Theoretical perspective

Normative modeling

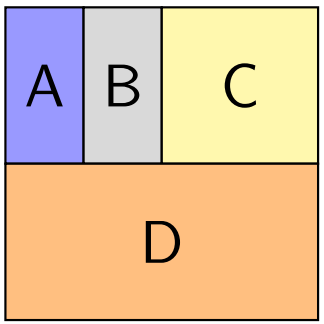
How would an *optimal* system perform what the brain does with natural images? This is typically about optimal *coding*, *transmission, or memory*.

- What does optimal coding, transmission, or memory mean?

$\rightsquigarrow$ information theory

Why should we care?

Theoretical perspective



Trivial code

$A \mapsto 00$

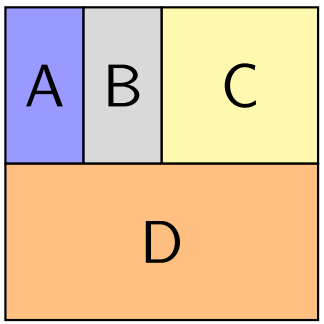
$B \mapsto 01$

$C \mapsto 10$

$D \mapsto 11$

Why should we care?

Theoretical perspective



Trivial code

$A \mapsto 00$

$B \mapsto 01$

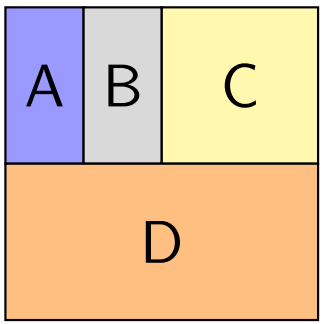
$C \mapsto 10$

$D \mapsto 11$

Average number of digits per letter: 2

Why should we care?

Theoretical perspective



Trivial code

$A \mapsto 00$

$B \mapsto 01$

$C \mapsto 10$

$D \mapsto 11$

Average number of digits per letter: 2

Better code

$A \mapsto 111$

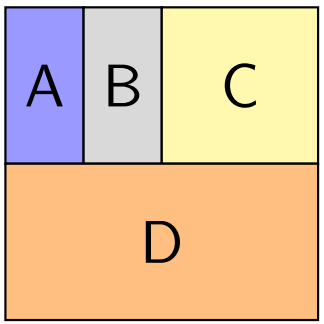
$B \mapsto 110$

$C \mapsto 10$

$D \mapsto 0$

Why should we care?

Theoretical perspective



Trivial code

$A \mapsto 00$

$B \mapsto 01$

$C \mapsto 10$

$D \mapsto 11$

Average number of digits per letter: 2

Better code

$A \mapsto 111$

$B \mapsto 110$

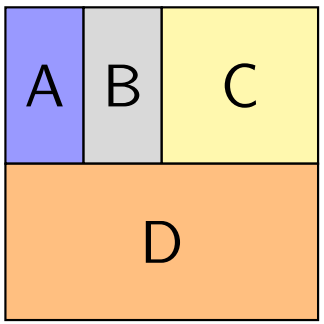
$C \mapsto 10$

$D \mapsto 0$

Average number of digits per letter: $3p(A) + 3p(B) + 2p(C) + p(D)$

Why should we care?

Theoretical perspective



Trivial code

$A \mapsto 00$

$B \mapsto 01$

$C \mapsto 10$

$D \mapsto 11$

Average number of digits per letter: 2

Better code

$A \mapsto 111$

$B \mapsto 110$

$C \mapsto 10$

$D \mapsto 0$

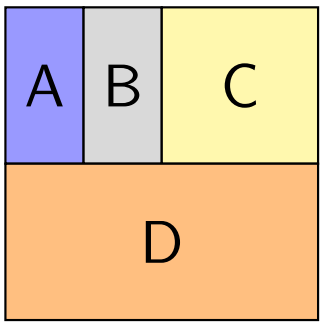
Average number of digits per letter:

$3p(A) + 3p(B) + 2p(C) + p(D)$

$= 3/8 + 3/8 + 2/4 + 1/2$

Why should we care?

Theoretical perspective



Trivial code

$A \mapsto 00$

$B \mapsto 01$

$C \mapsto 10$

$D \mapsto 11$

Average number of digits per letter: 2

Better code

$A \mapsto 111$

$B \mapsto 110$

$C \mapsto 10$

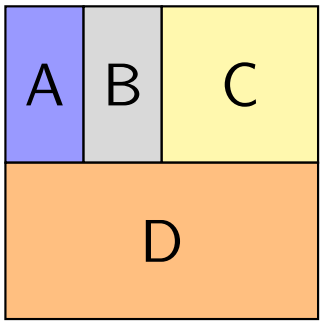
$D \mapsto 0$

Average number of digits per letter:

$$3p(A) + 3p(B) + 2p(C) + p(D)$$
$$= 3/8 + 3/8 + 2/4 + 1/2$$
$$= 14/8 = 1.75$$

Why should we care?

Theoretical perspective



Trivial code

$A \mapsto 00$

$B \mapsto 01$

$C \mapsto 10$

$D \mapsto 11$

Better code

$A \mapsto 111$

$B \mapsto 110$

$C \mapsto 10$

$D \mapsto 0$

Average number of digits per letter: 2

Average number of digits per letter:

$$3p(A) + 3p(B) + 2p(C) + p(D)$$
$$= 3/8 + 3/8 + 2/4 + 1/2$$
$$= 14/8 = 1.75$$

$$= \frac{1}{\log_2 p(A)} p(A) + \frac{1}{\log_2 p(B)} p(B) + \frac{1}{\log_2 p(C)} p(C) + \frac{1}{\log_2 p(D)} p(D)$$

Why should we care?

Theoretical perspective

Average code length of optimal code: Entropy

For a discrete random variable X with states $\{k\} =: \mathcal{A} \subset \mathbb{N}$ we define the *entropy*

$$H(X) := -\sum_{k \in \mathcal{A}} p(k) \log_2(p(k)) = -\mathbb{E}[\log_2(p(k))]$$

Entropy can be interpreted as the average code length needed by an optimal code.

Why should we care?

Theoretical perspective

If X is defined for continuous values, the second part of the definition of entropy is still valid. We then obtain differential entropy, which is

$$H(X) := -\mathbb{E}[\log_2(f)] = -\int f(x)\log_2(f(x))\,\mathrm{d}x.$$

Differential entropy is related to the shortest code length of a discretized version of X.

Why should we care?

Theoretical perspective

Redundancy

For natural images, we will typically have a suboptimal code. This code will have an average code length of

$$H(X) + R.$$

Here, R is called *Redundancy*.

Why should we care?

Theoretical perspective

- Finding an optimal code is related to finding a density model with minimum redundancy.
- This can be achieved by transforming X.
- If X has multiple dimensions, transformations that recombine the information from the different dimensions.

Practical considerations

Models for natural images are models for image patches

To fit a density model, we need lots of samples. For typical image sizes, this soon becomes computationally impossible. Instead of using complete images, models typically work on image patches.

Practical considerations

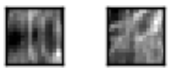
Models for natural images are models for image patches

To fit a density model, we need lots of samples. For typical image sizes, this soon becomes computationally impossible. Instead of using complete images, models typically work on image patches.

Practical considerations

Models for natural images are models for image patches

To fit a density model, we need lots of samples. For typical image sizes, this soon becomes computationally impossible. Instead of using complete images, models typically work on image patches.


Practical considerations

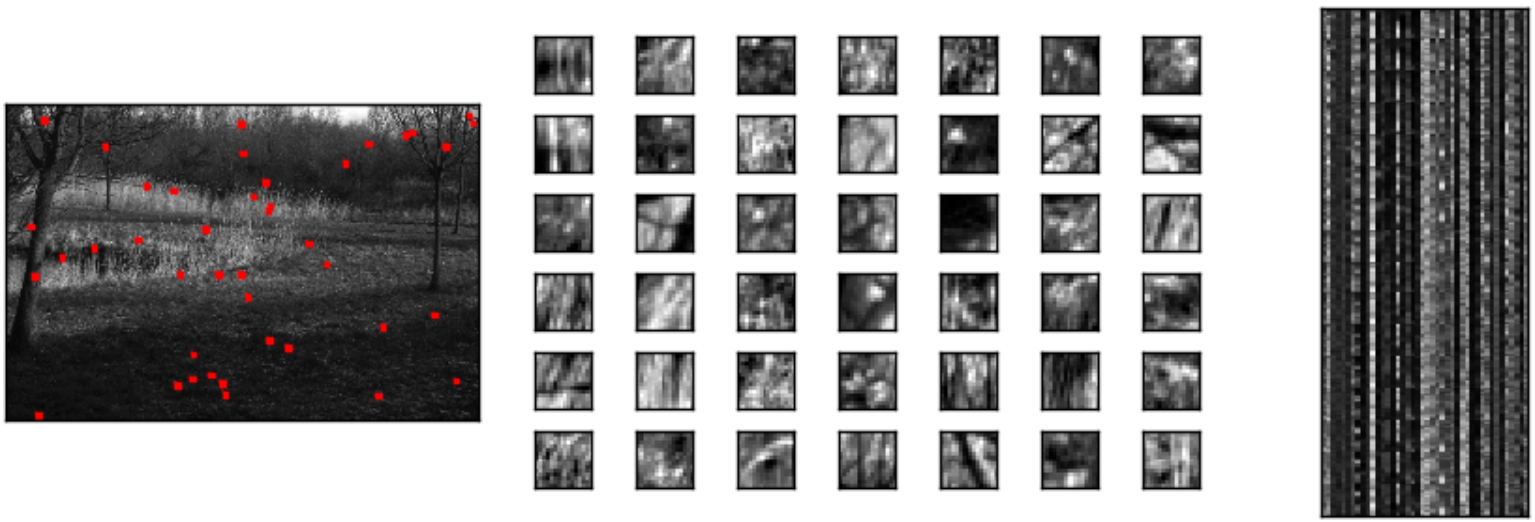
Models for natural images are models for image patches

To fit a density model, we need lots of samples. For typical image sizes, this soon becomes computationally impossible. Instead of using complete images, models typically work on image patches.

Practical considerations

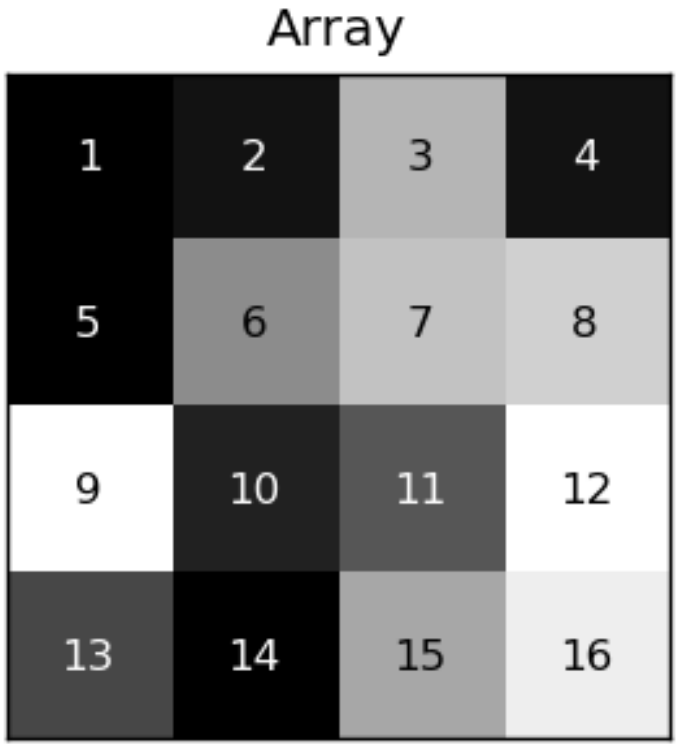
Models for natural images are models for image patches

To fit a density model, we need lots of samples. For typical image sizes, this soon becomes computationally impossible. Instead of using complete images, models typically work on image patches.



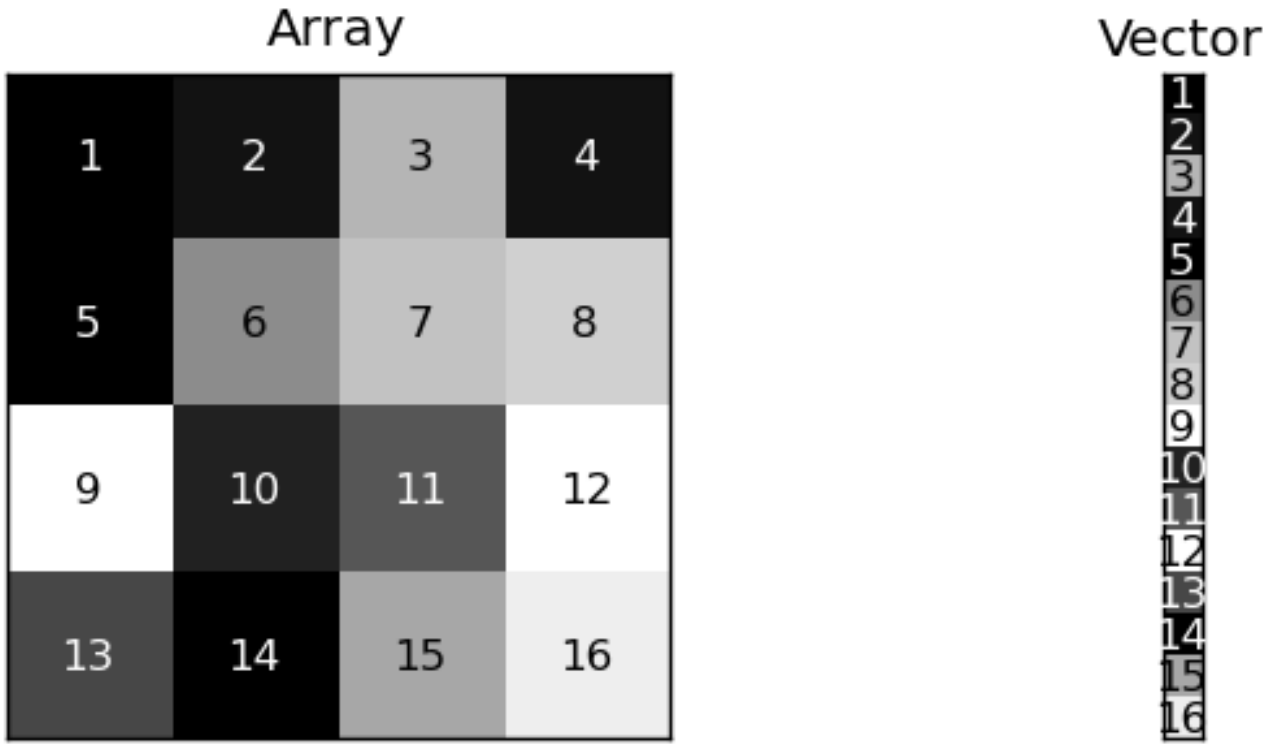
Practical considerations

Two views for image patches



Practical considerations

Two views for image patches



Typically we look at the "Array" view to get some visual intuition, but we perform algebraic manipulations in the "Vector" view.

Classical findings

Power and phase spectra

Covariance and correlation

Consider random variables X and Y. The quantity

$$\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

is called covariance of X and Y. If the covariance of two random variables X and Y is not zero, we can write

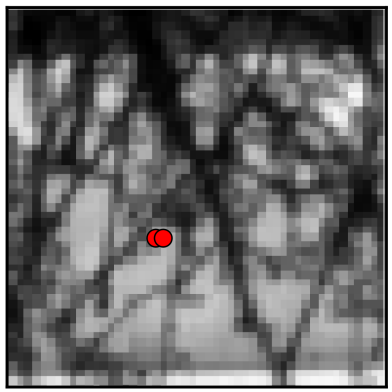
$$Y = aX + b + \xi,$$

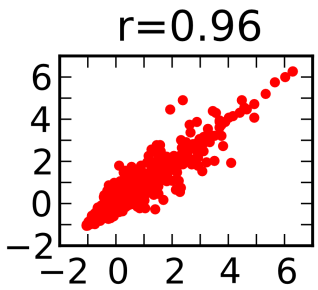
with a random variable ξ that has expectation $\mathbb{E}(\xi) = 0$.
The normalized covariance $\rho(X, Y) = \mathrm{cov}(X, Y)/\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}$ is called correlation.

Classical findings

Power and phase spectra

What information is contained in the linear structure of natural images?

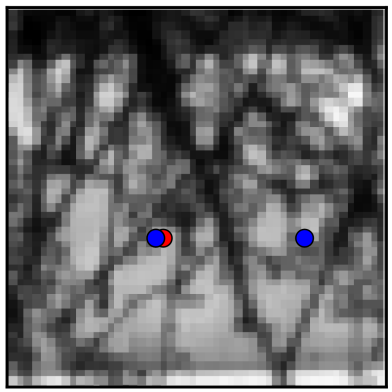


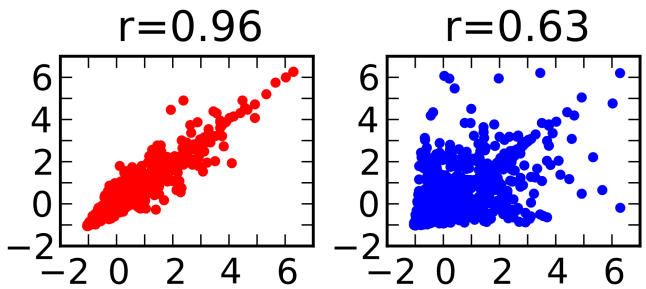


Classical findings

Power and phase spectra

What information is contained in the linear structure of natural images?

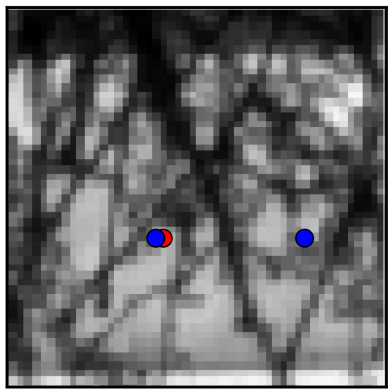


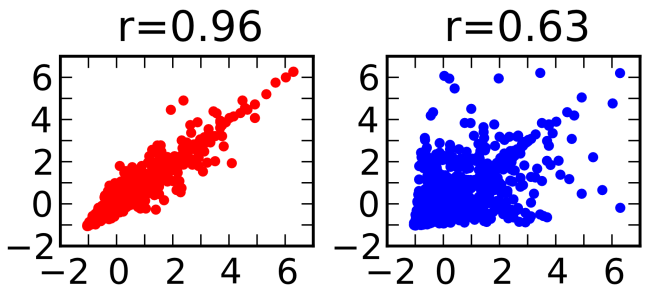


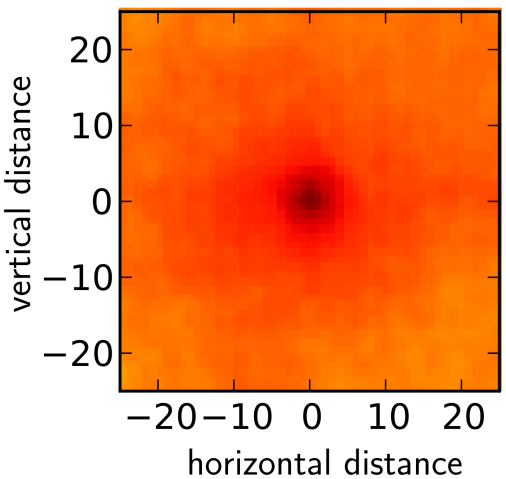
Classical findings

Power and phase spectra

What information is contained in the linear structure of natural images?



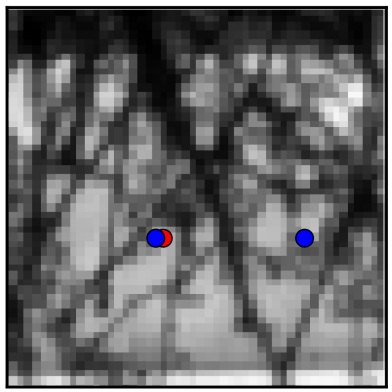


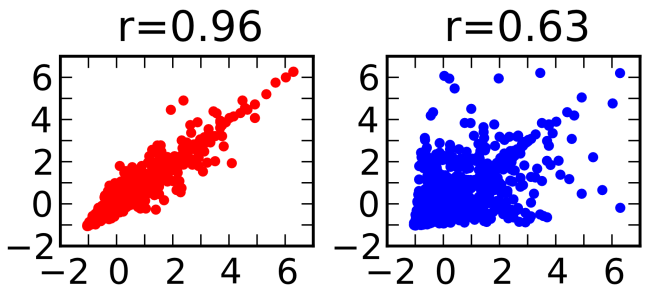


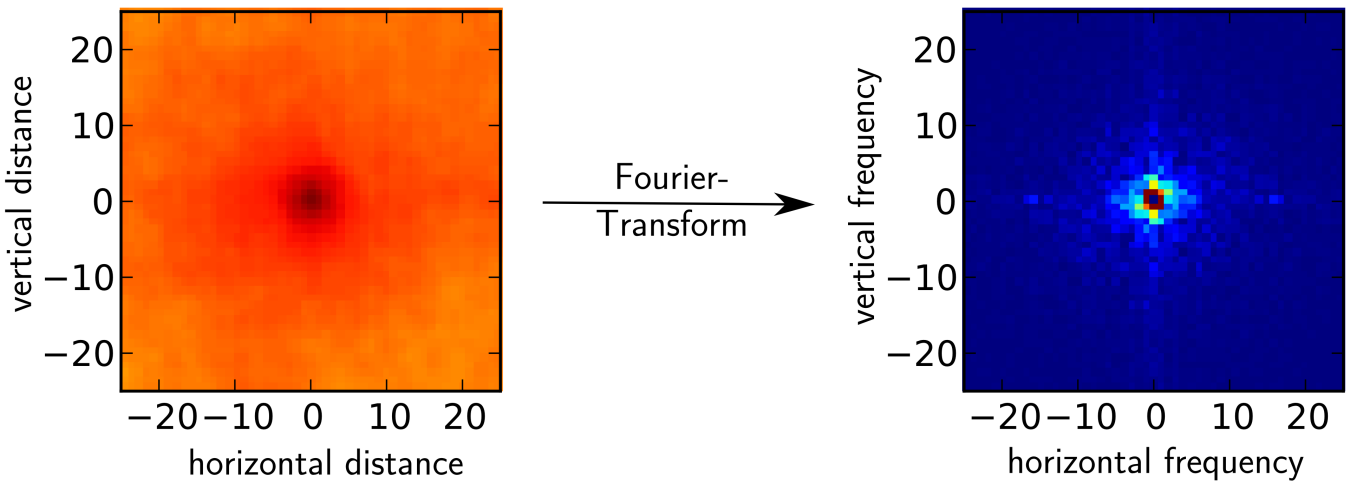
Classical findings

Power and phase spectra

What information is contained in the linear structure of natural images?



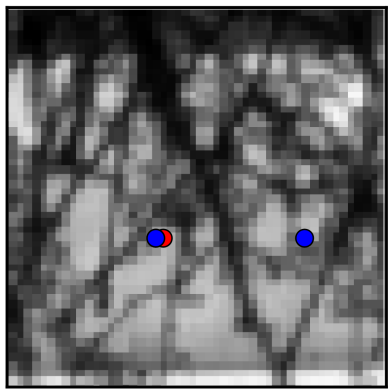


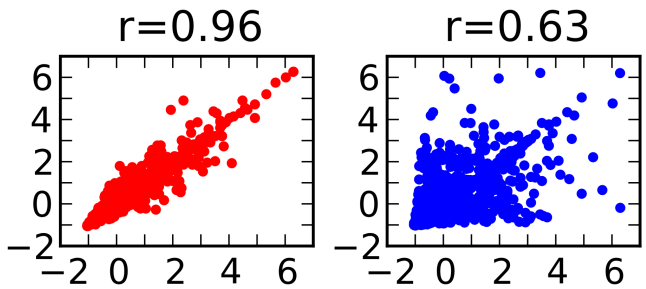


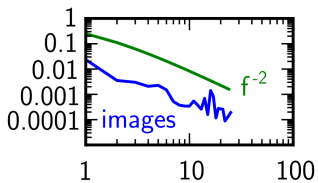
Classical findings

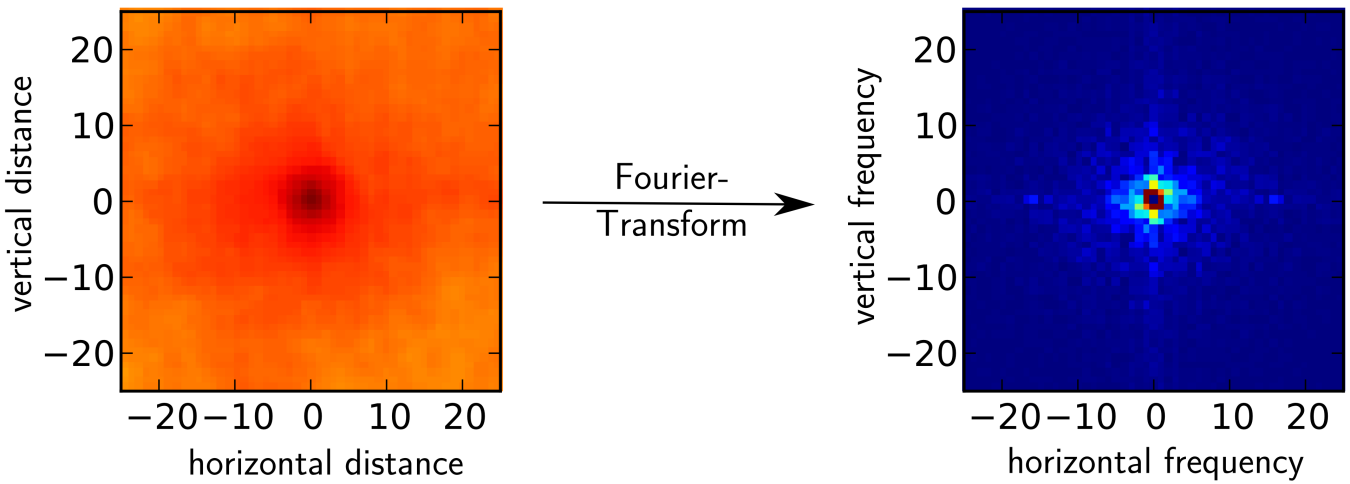
Power and phase spectra

What information is contained in the linear structure of natural images?





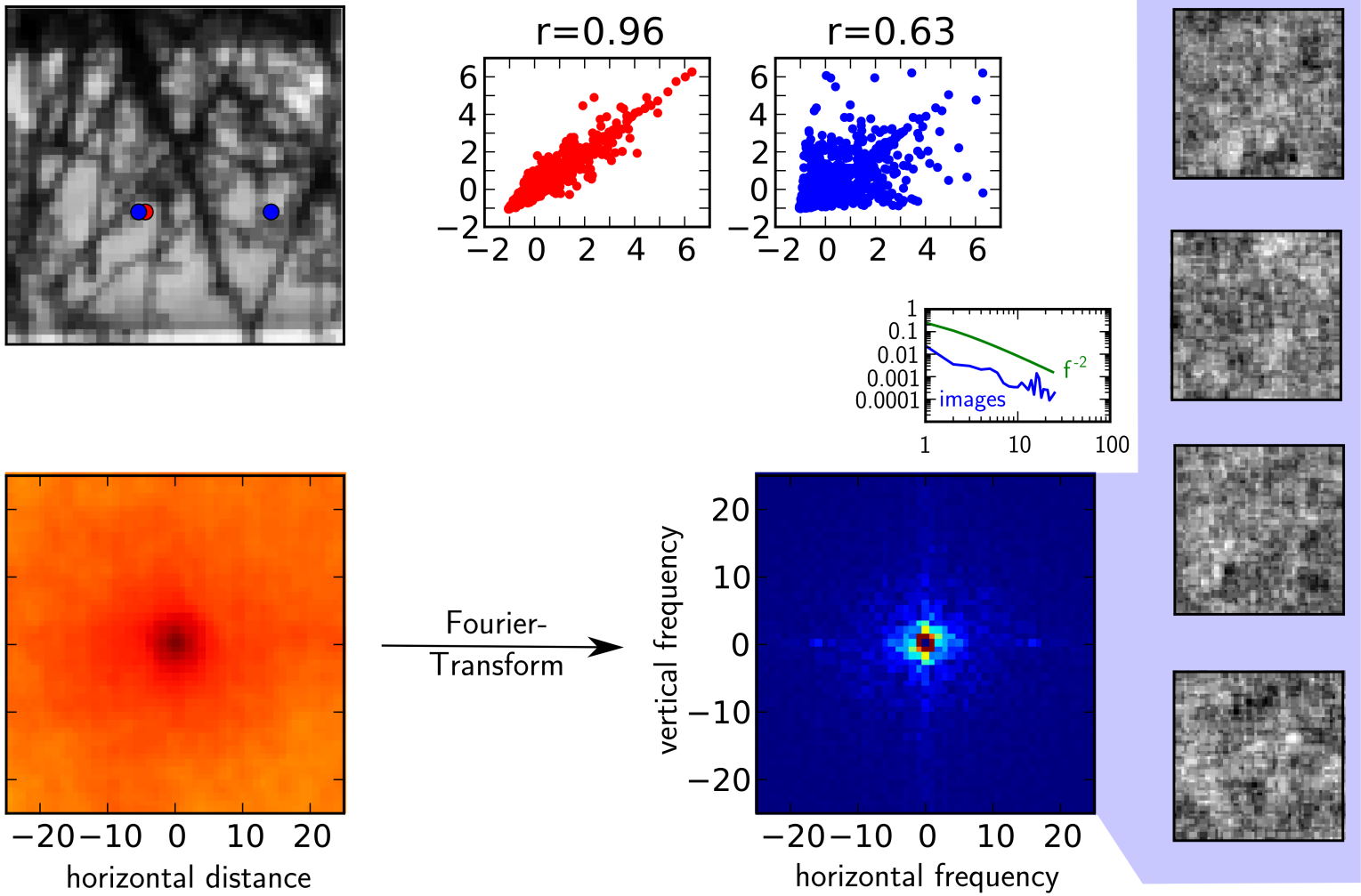




Classical findings

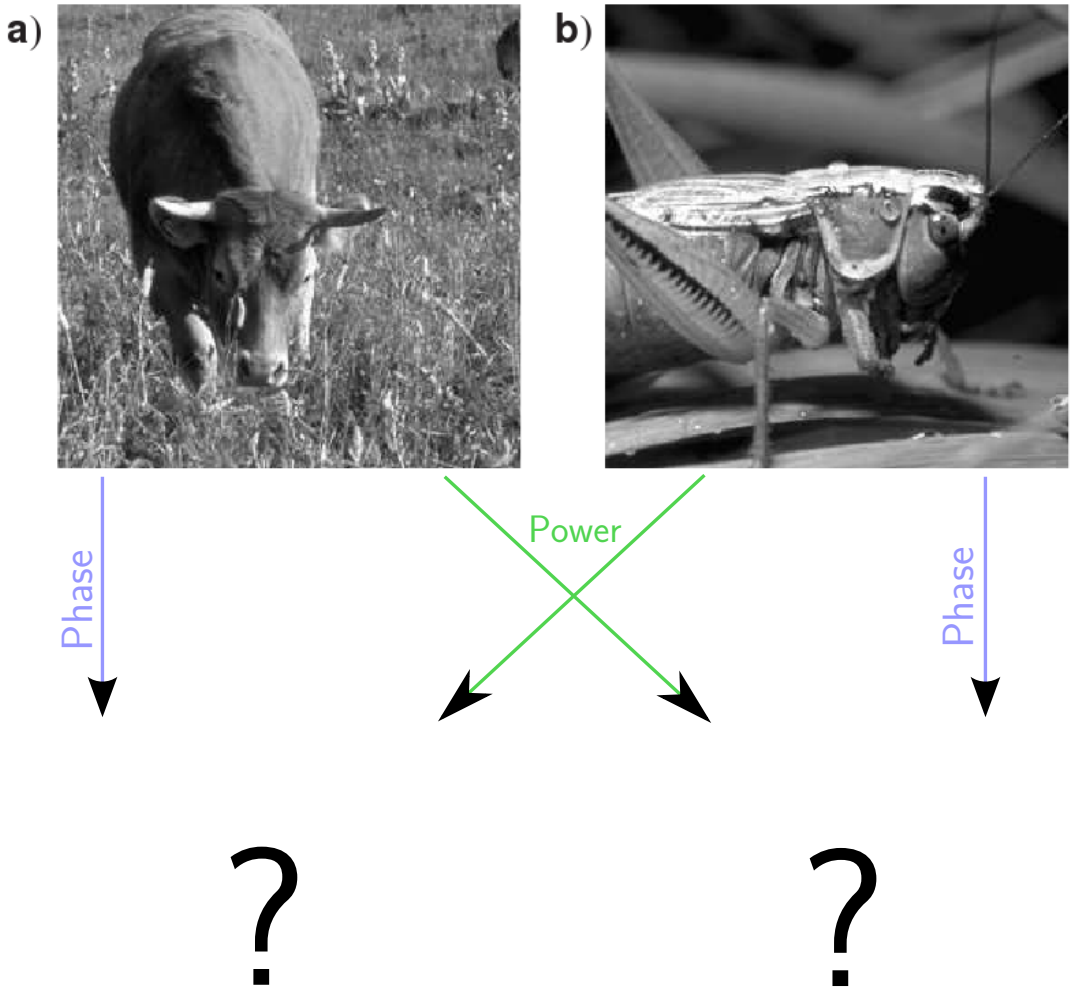
Power and phase spectra

What information is contained in the linear structure of natural images?



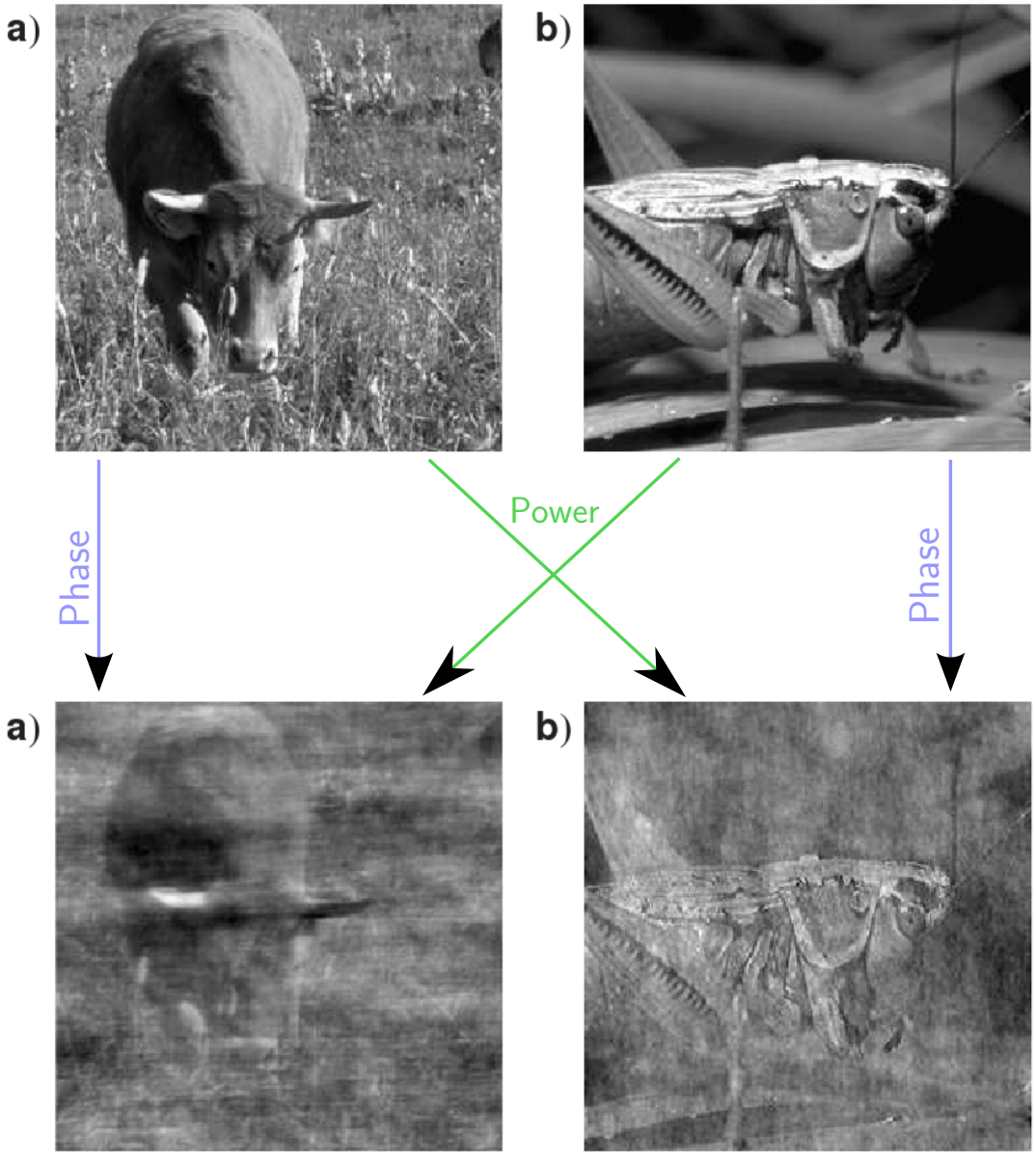
Classical findings

Power and phase spectra



Classical findings

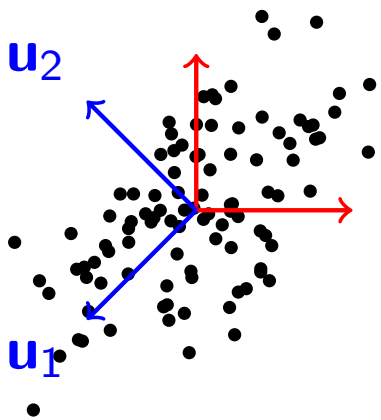
Power and phase spectra



Classical findings

Whitening filters

Removing redundancy introduced by linear dependencies

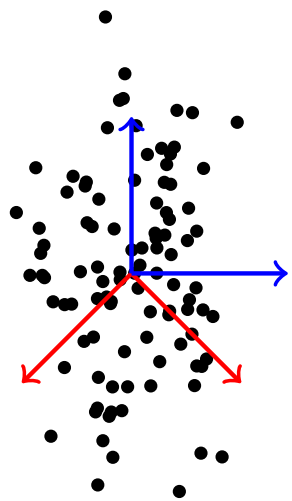


- $(0,1),(1,0)$ standard unit vectors and $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2)$, principal components

Classical findings

Whitening filters

Removing redundancy introduced by linear dependencies

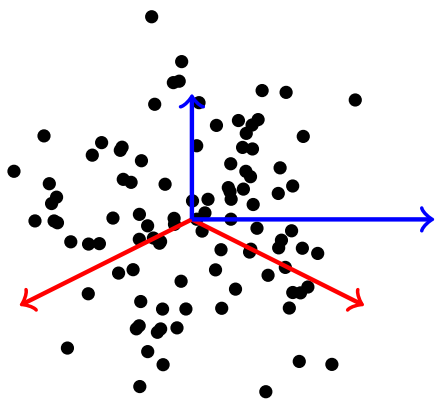


- $(0,1),(1,0)$ standard unit vectors and $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2)$, principal components
- rotate data to principal component space: $\mathbf{y} = \mathbf{U}^T \mathbf{x}$.

Classical findings

Whitening filters

Removing redundancy introduced by linear dependencies

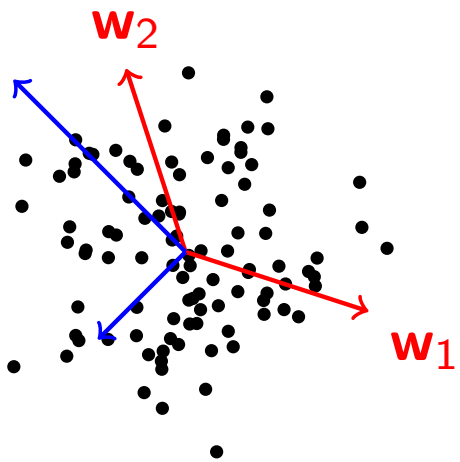


- $(0,1),(1,0)$ standard unit vectors and $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2)$, principal components
- rotate data to principal component space: $\mathbf{y} = \mathbf{U}^T \mathbf{x}$.
- scale the data to unit variance in principal component space: $\mathbf{s} = (s_k)_k, \;\; s_k = \frac{y_k}{\sqrt{\mathrm{var}(y_k)}}$.

Classical findings

Whitening filters

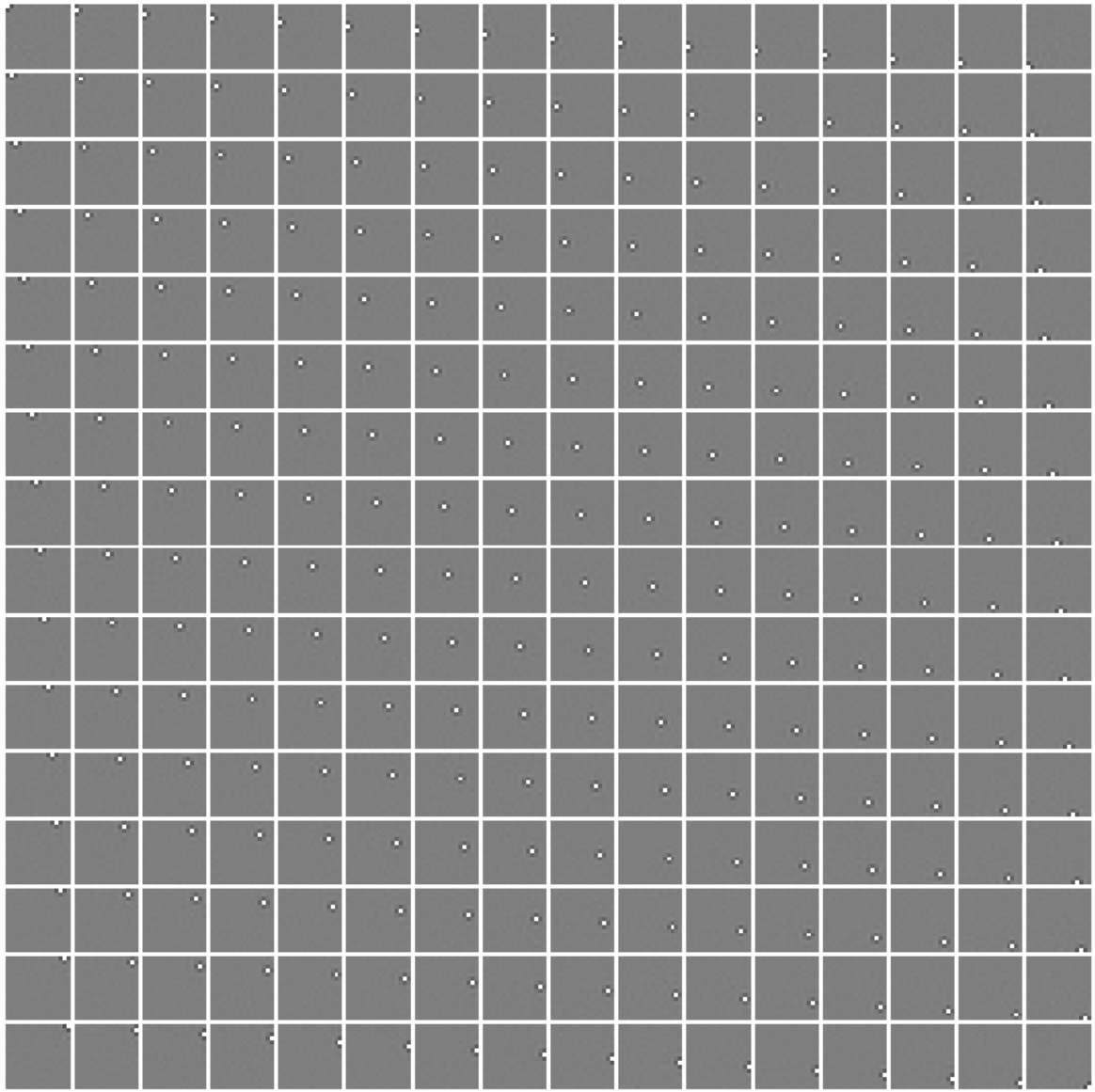
Removing redundancy introduced by linear dependencies



- $(0,1),(1,0)$ standard unit vectors and $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2)$, principal components
- rotate data to principal component space: $\mathbf{y} = \mathbf{U}^T\mathbf{x}$.
- scale the data to unit variance in principal component space: $\mathbf{s} = (s_k)_k, \;\; s_k = \frac{y_k}{\sqrt{\mathrm{var}(y_k)}}$.
- rotate data back to original space: $\mathbf{z} = \mathbf{U}\mathbf{s} = \mathbf{W}\mathbf{x}$.

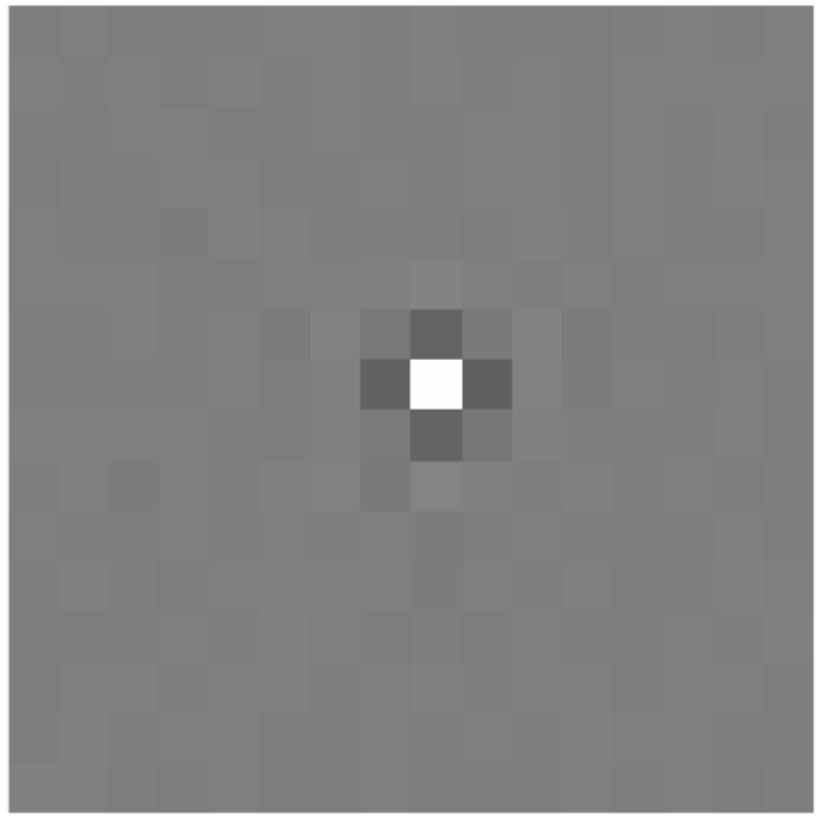
Classical findings

Whitening filters



Classical findings

Whitening filters



Classical findings

Orientation selective filters

Stochastic independence and correlation

Two random variables X and Y are *independent*, if their joint probability distribution can be decomposed as a product of the two marginal distributions:

$$p(X, Y) = p(X)p(Y).$$

- So far we looked at transformations that exploited linear dependencies, i.e. covariances to reduce redundancy.
- Can we also reduce redundancy by exploiting stochastic independence?

Classical findings

Orientation selective filters

Assume that each image results from a superposition of a number of features A_i,

$$I_j = \sum_{i=1}^{m} \mathbf{A}_{ji} s_i .$$

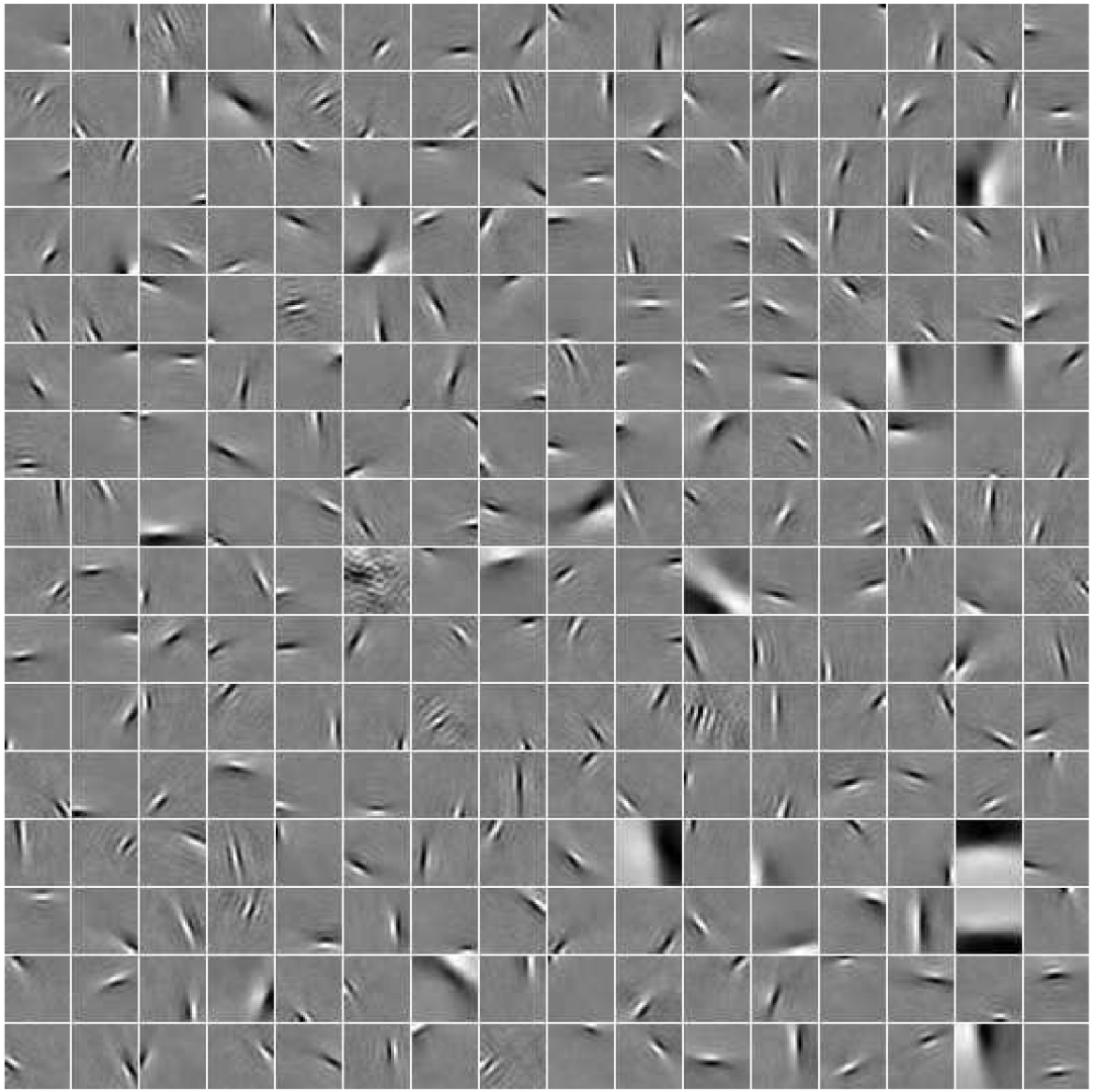
Assumptions about the s_i

- s_i are stochastically independent.
- s_i are non-Gaussian.
- $\mathbf{A}$ is invertible.

$\rightsquigarrow$ independent component analysis (ICA) maximizes stochastic independence of components

Classical findings

Orientation selective filters



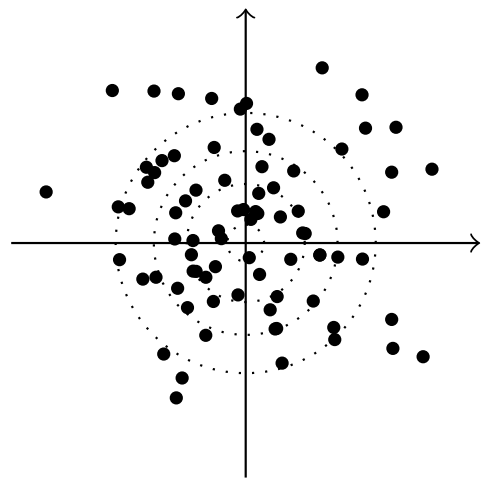
Classical findings

Orientation selective filters

Non gaussian source distributions

After whitening, the ICA just consists of finding the right rotation of the principal axes that maximizes independence.
However, the distribution of whitened gaussian data is circular symmetric

$$p(X_1, \ldots, X_n) = \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2} \sum_i x_i^2 \Big)$$



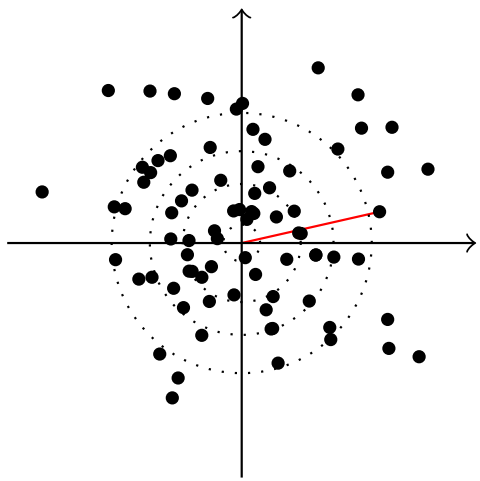
Classical findings

Orientation selective filters

Non gaussian source distributions

After whitening, the ICA just consists of finding the right rotation of the principal axes that maximizes independence.
However, the distribution of whitened gaussian data is circular symmetric

$$p(X_1, \ldots, X_n) = \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_i x_i^2\Big)$$
$$= \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\|\mathbf{x}\|^2\Big).$$



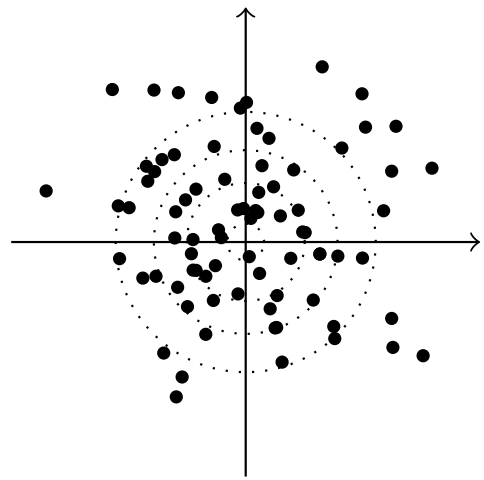
Classical findings

Orientation selective filters

Non gaussian source distributions

After whitening, the ICA just consists of finding the right rotation of the principal axes that maximizes independence.
However, the distribution of whitened gaussian data is circular symmetric

$$p(X_1, \ldots, X_n) = \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_i x_i^2\Big)$$
$$= \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\|\mathbf{x}\|^2\Big).$$



There is no information about rotations in whitened gaussian data!

Classical findings

Orientation selective filters

Non guassian source distributions

Uncorrelated gaussian variables are already independent!

$$
\begin{aligned}
p(X_1, \ldots, X_n) &= \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\Big) \\
&= \frac{1}{(2\pi)^{n/2}} \prod_{i=1}^{n} \exp\Big(-\frac{1}{2}x_i^2\Big) \\
&= \Big(\frac{1}{(2\pi)^{1/2}}\Big)^n \prod_{i=1}^{n} \exp\Big(-\frac{1}{2}x_i^2\Big) \\
&= \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\Big(-\frac{1}{2}x_i^2\Big).
\end{aligned}
$$

So maximizing independence is no use!

Classical findings

Orientation selective filters

What is the distribution of the s_i?

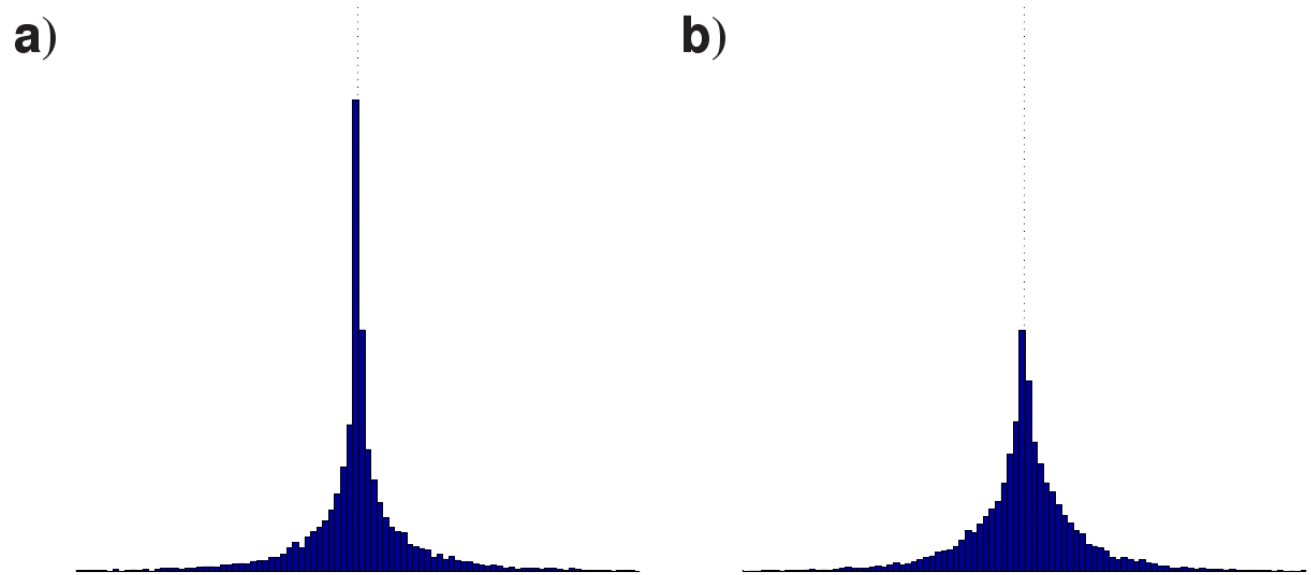
It is of interest to know the distribution of the s_i. We can rewrite the image model,

$$\mathbf{I} = \mathbf{A}\mathbf{s} \implies \mathbf{s} = \mathbf{A}^{-1}\mathbf{I},$$

and see that we can associate the s_i with responses of neurons! In this case the rows of $\mathbf{A}^{-1}$ correspond to the receptive fields of these neurons.

Classical findings

Orientation selective filters

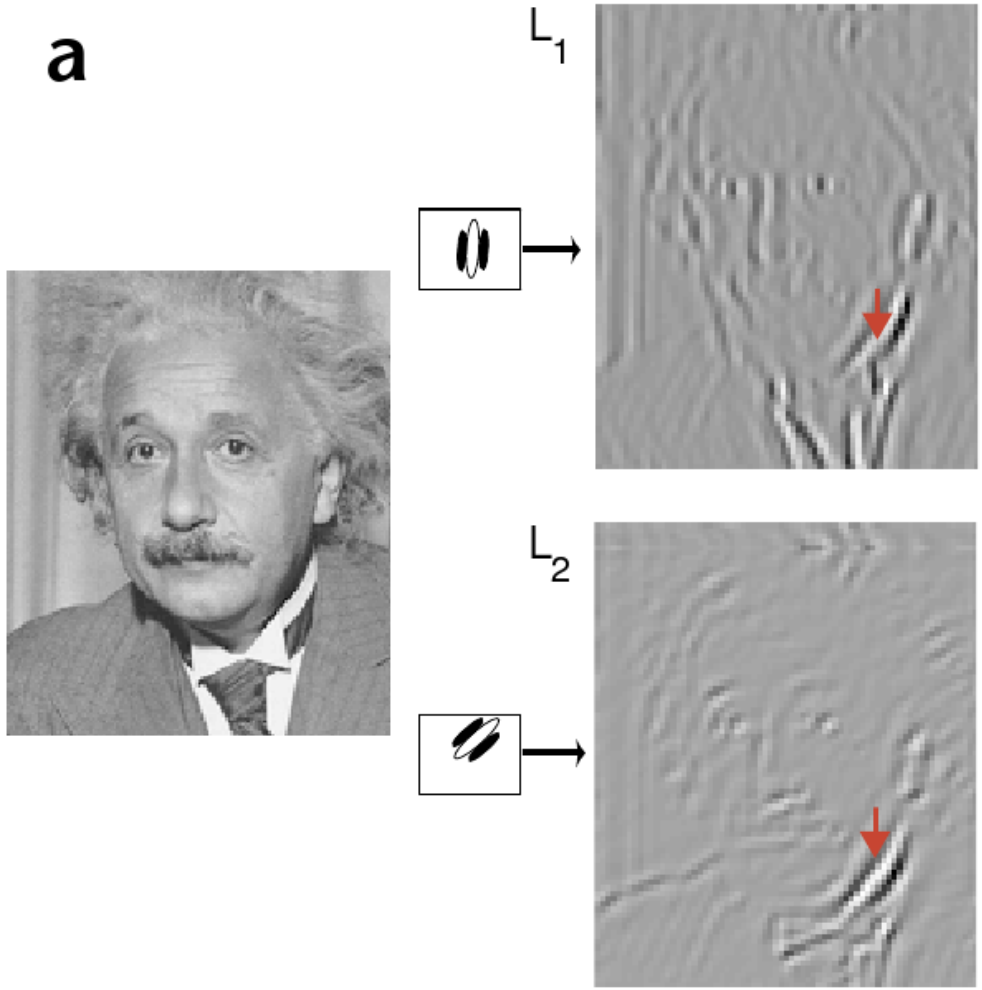


a) Output from an ICA filter

b) Output from a whitening filter

Classical findings

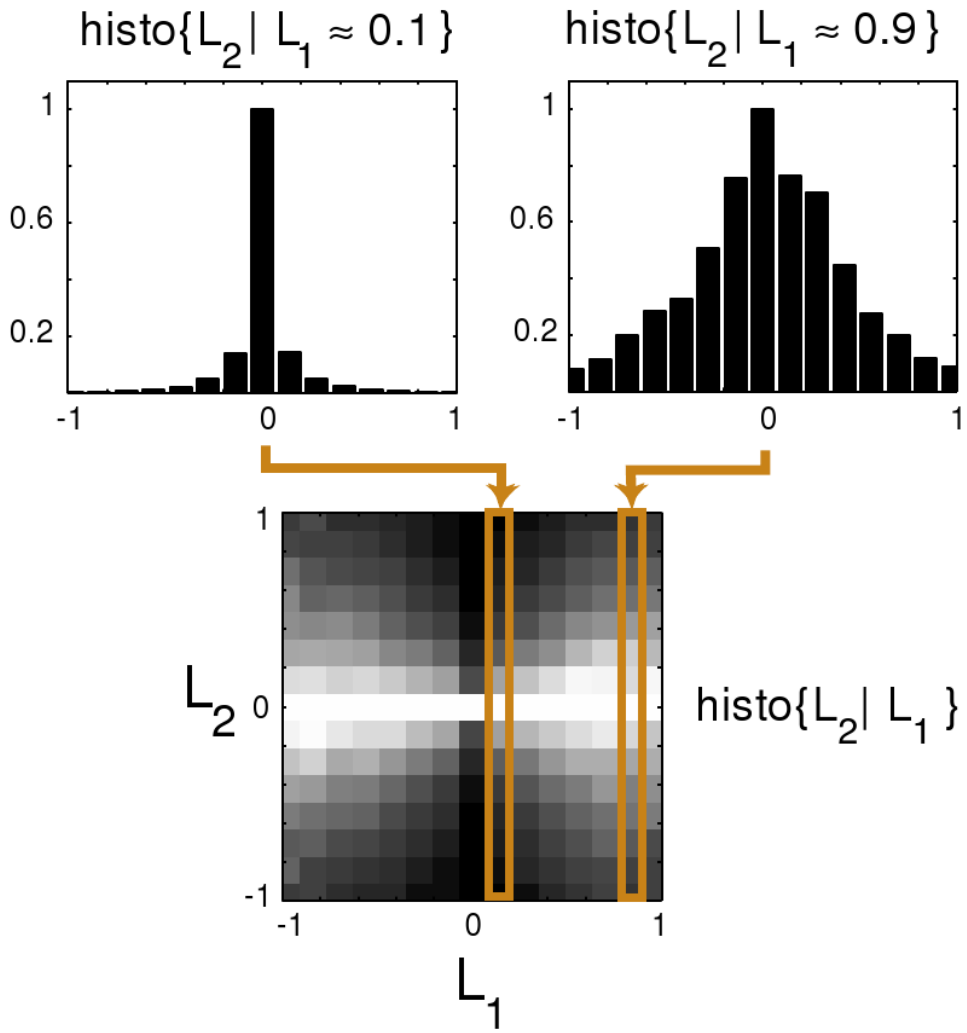
Variance correlations and contrast gain control



see Schwarz & Simoncelli, 2001, Nature Neuroscience

Classical findings

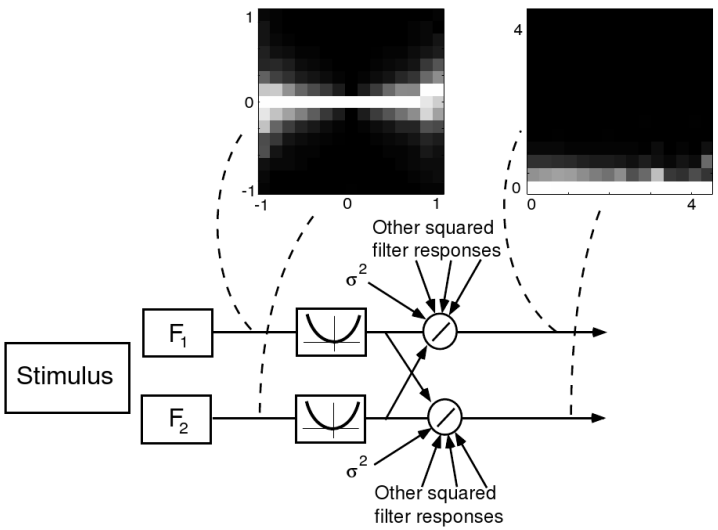
Variance correlations and contrast gain control



see Schwarz & Simoncelli, 2001, Nature Neuroscience

Classical findings

Variance correlations and contrast gain control



Assume two filter outputs L_1, L_2. Model the variance of one filter by

$$\text{var}(L_1|L_2) = wL_2^2 + \sigma^2.$$

Then we can eliminate this dependency by

$$R_1 = \frac{L_1^2}{wL_2^2 + \sigma^2}$$

See details in analytical exercise!

Classical findings

Variance correlations and contrast gain control

Why might gain control be useful?

- The illuminance arriving at the retina is given by

$$I = L \cdot R.$$

L is illuminantion, R is object reflectance. Typically, we are interested in object reflectance because that's relevant for object recognition.

Classical findings

Variance correlations and contrast gain control

Why might gain control be useful?

- Natural neurons have a limited range. Gain control might allow to use this range more efficiently.
- Gain control actually allows for a considerable redundancy reduction.