MC 2019 Part 3 Basics of Probability theory and statistical signal processing

Hirokazu Tanaka

Topics to be covered today.

1. Entropy – it's origin in physics and information theory.
2. Markov process, H-theorem and renormalization group.
3. Random variable theorem and central limit theorem.
4. Independent component analysis as inverse of CLT.
5. Matrix decompositions and non-negative matrix factorization.
6. Tensor decompositions: PARAFAC and Tucker decompositions.
7. Task-related component analysis and its extensions.

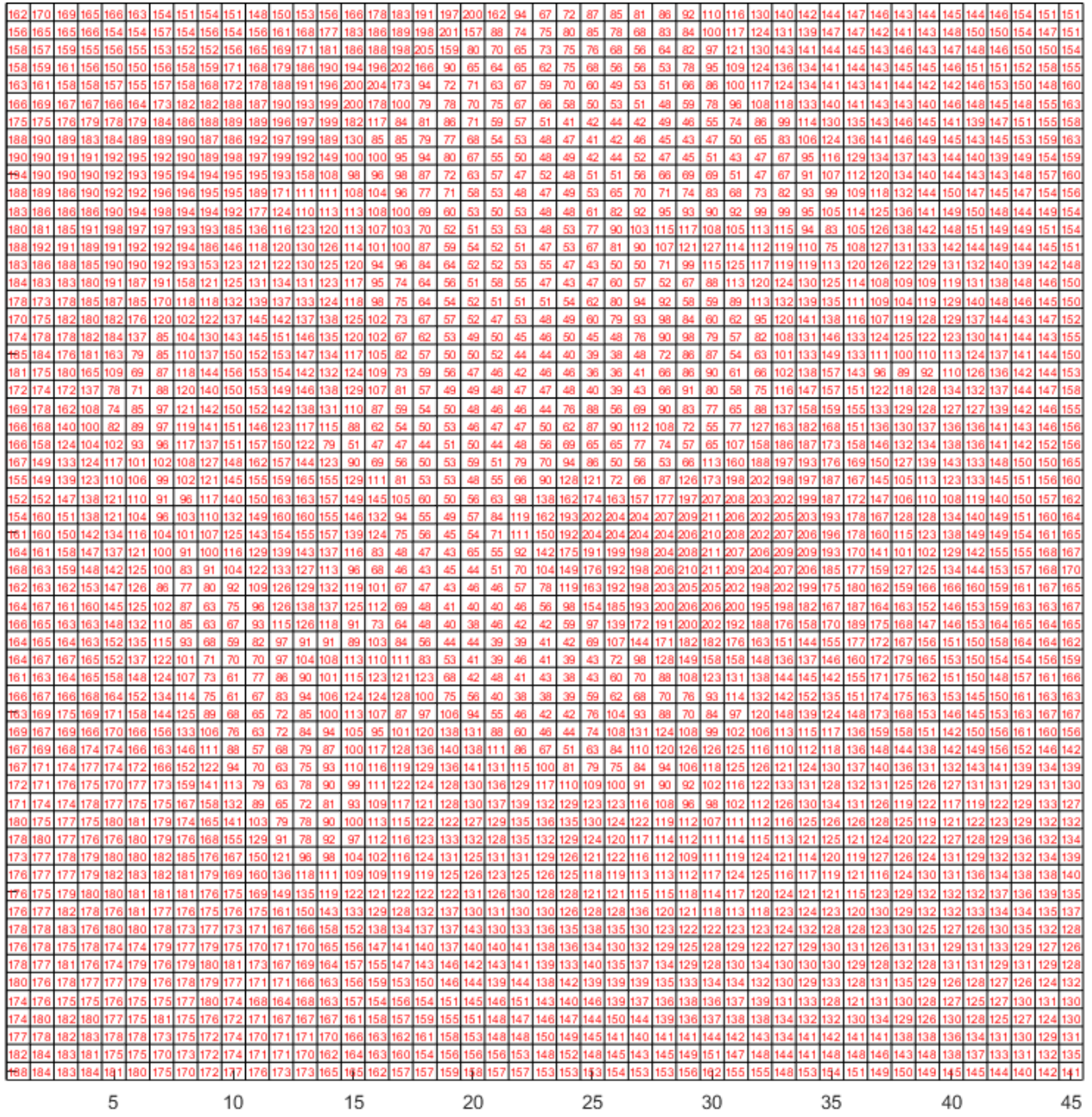
Topics NOT covered today.

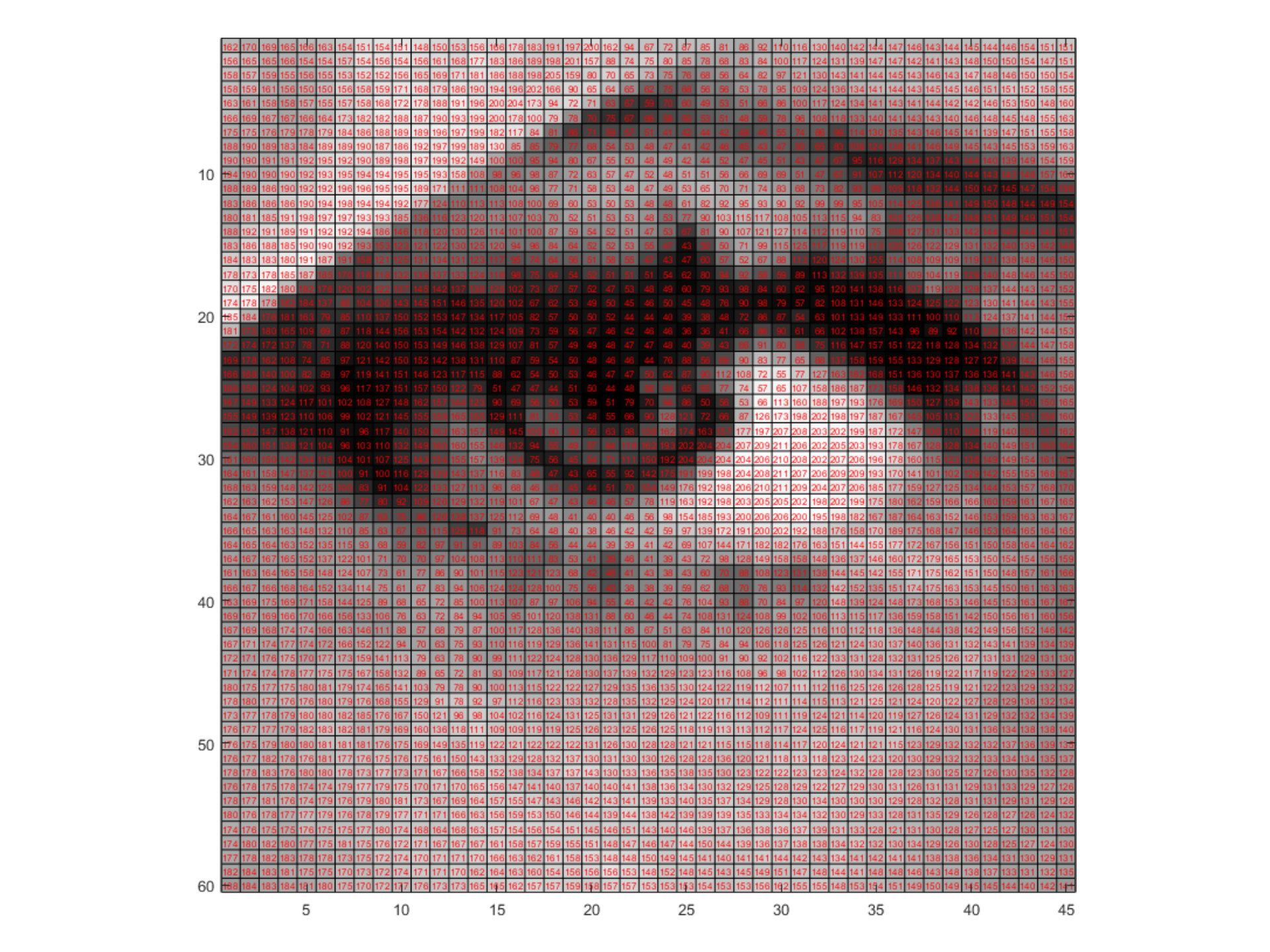
- Linear causality analysis methods:
  - Granger causality (Granger, 1967; Geweke, 1982).
  - Partial directed coherence (PDC) (Baccala & Sameshima, 2001).
  - Directed transfer function (DTF) (Kaminski et al., 2001).

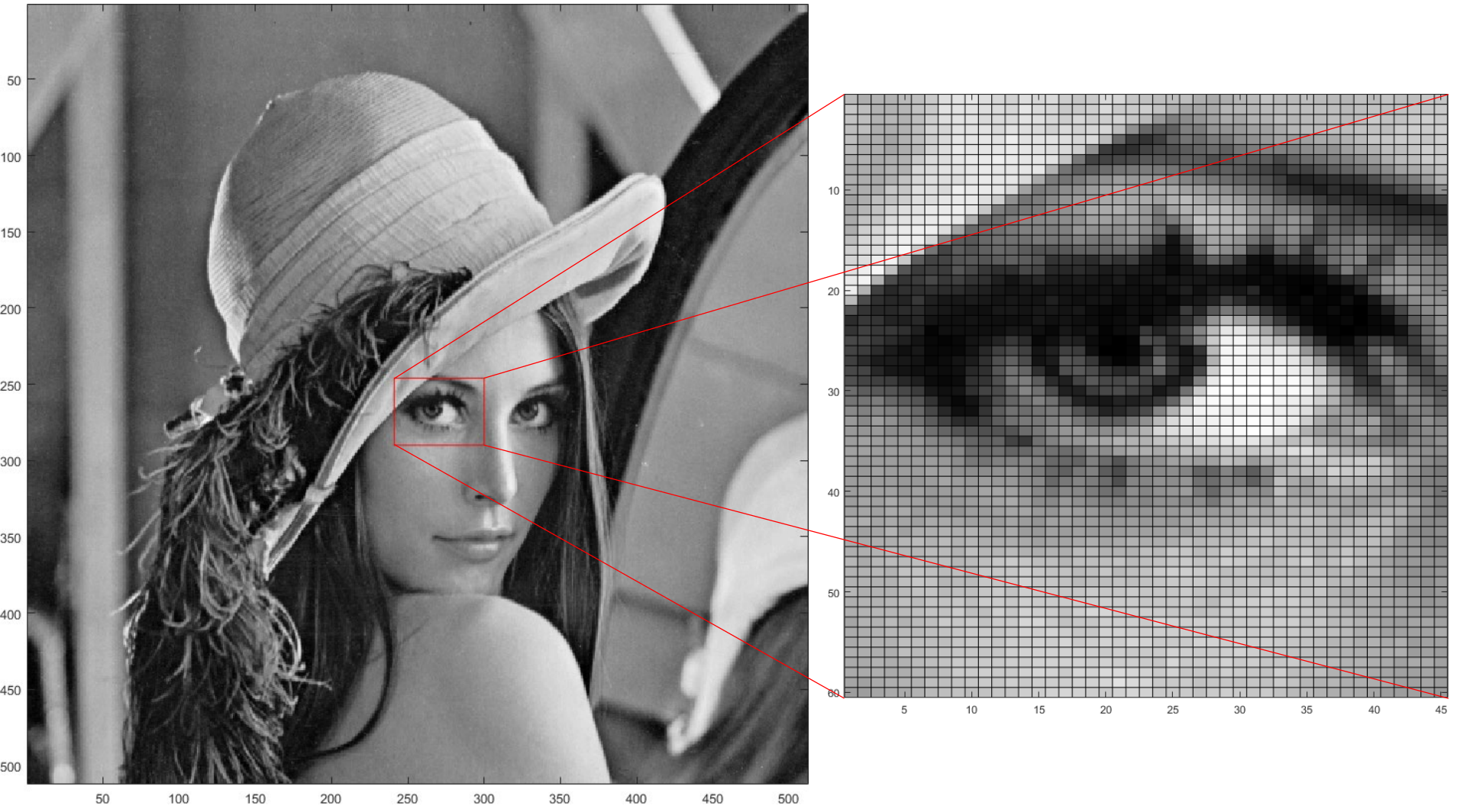
- Nonlinear dynamic analysis methods:
  - Delay differential embedding (DDM) (Lainscsek et al. 2015).
  - Convergent cross mapping (CCM) (Sugihara et al. 2012).

- Non-additive, multiplicative data decomposition:
  - Holo-Hilbert spectral analysis (Huang et al. 2016).

- Dictionary learning
  - Matching pursuit (MP) (Mallat & Zhang, 1993).
  - K-SVD (Aharon et al., 2006).


10
20
30
40
50
60
5
10
15
20
25
30
35
40
45


10
20
30
40
50
60
5
10
15
20
25
30
35
40
45


50
100
150
200
250
300
350
400
450
500
50
100
150
200
250
300
350
400
450
500
10
20
30
40
50
60
5
10
15
20
25
30
35
40
45

Topics to be covered today.

1. Entropy – it's origin in physics and information theory.
2. Markov process, H-theorem and renormalization group.
3. Random variable theorem and central limit theorem.
4. Independent component analysis as inverse of CLT.
5. Matrix decompositions and non-negative matrix factorization.
6. Tensor decompositions: PARAFAC and Tucker decompositions.
7. Task-related component analysis and its extensions.

Many facets of entropy.

- Clausius entropy (1850?)

$$dS(E,V) = \frac{1}{T}dE - \frac{V}{T}dV$$

- Boltzmann entropy (1872)

$$S(\Omega) = k_{\mathrm{B}} \log W(\Omega)$$

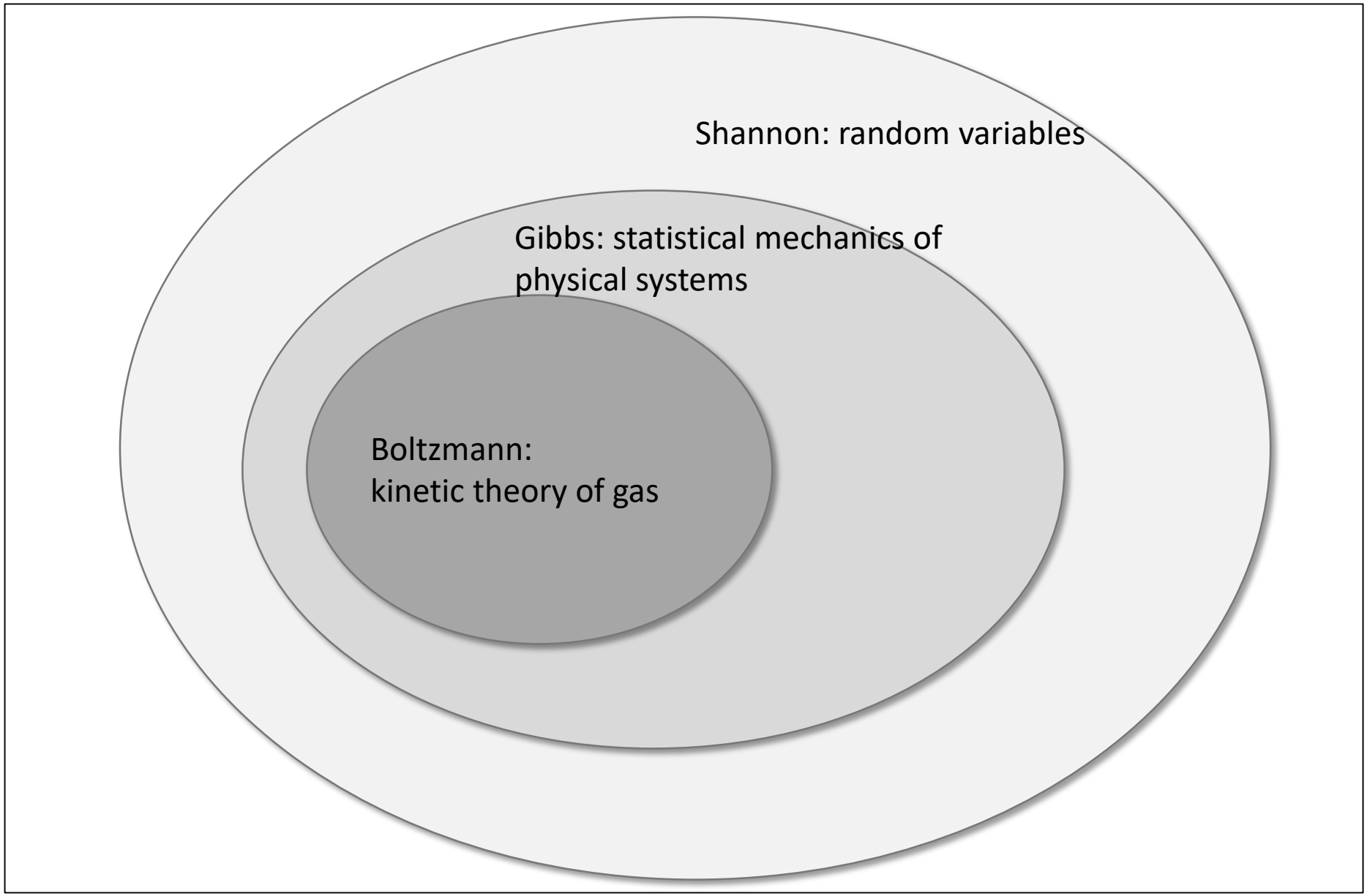
- Gibbs entropy (1902)

$$S(\Omega) = -k_{\mathrm{B}} \sum_{\omega \in \Omega} p(\omega) \log p(\omega)$$

- Shannon entropy (1949)

$$\mathrm{H}(X) = -\sum_{i} p(x_i) \log_2 p(x_i)$$

Entropy: from gas theory to probability.

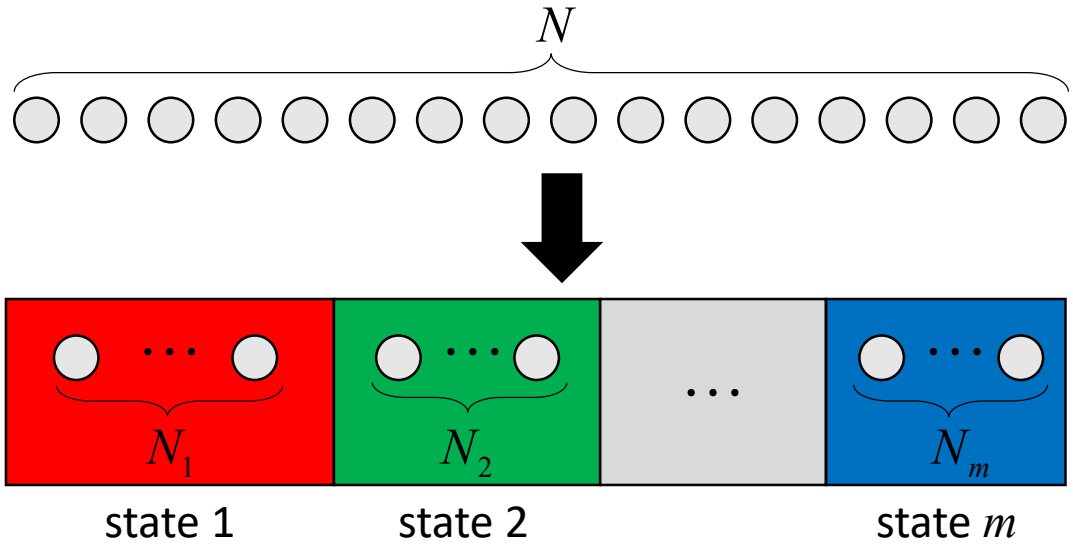


Boltzmann entropy as a number of microstates.

- Ludwig Boltzmann (1844-1906)

$$S_{\mathrm{B}} = k_b \log W$$

$$W = (\#\ \text{micro states}) = \frac{N!}{\prod_{i=1}^{m} N_i!}$$



Multinomial coefficients.

- Multinomial theorem for a positive integer m and a non-negative integer n:

$$\left(x_1 + x_2 + \cdots + x_m\right)^n = \sum_{k_1+k_2+\cdots+k_m=n} \binom{n}{k_1, k_2, \cdots, k_m} x_1^{k_1} x_2^{k_2} \cdots x_m^{k_m}$$

- Multinomial coefficient:

$$\binom{n}{k_1, k_2, \cdots, k_m} = \frac{n!}{k_1!\, k_2! \cdots k_m!}$$

$$\underbrace{\binom{n}{k_1, k_2, \cdots, k_m, k_{m+1}}}_{m+1} = \underbrace{\binom{n}{k_1, k_2, \cdots, k_m + k_{m+1}}}_{m} \underbrace{\binom{k_m + k_{m+1}}{k_m, k_{m+1}}}_{2}$$

Multinomial coefficients.

- The Boltzmann Entropy is defined as a normalized log of multinomial coefficient:

$$\mathrm{H}\left(p_1, p_2, \cdots, p_m\right) \equiv \frac{1}{n} \log \binom{n}{k_1, k_2, \cdots, k_m} \qquad p_i = \frac{k_i}{n_i} \ \left(i = 1, \cdots, m\right)$$

The identity of multinomial coefficients is translated in one of entropies:

$$\mathrm{H}\left(p_1, p_2, \cdots, p_m, p_{m+1}\right) = \mathrm{H}\left(p_1, p_2, \cdots, p_m + p_{m+1}\right) + \left(p_m + p_{m+1}\right) \mathrm{H}\left(\frac{p_m}{p_m + p_{m+1}}, \frac{p_{m+1}}{p_m + p_{m+1}}\right)$$

Shannon entropy based on probability.

- Shannon (1948) A mathematical theory of communication.
- Suppose that X is a random variable and that its sample space is $\{x_1, x_2, \cdots, x_n\}$. Then the entropy of the random variable is defined as:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

$$0 \leq \mathrm{H}(X) \leq \log_2 n$$

- Note: the Gibbs entropy and the Shannon entropy have the same mathematical form. However, they differ in their implications: the Gibbs entropy assumes some physical systems while the Shannon entropy pertains to abstract random variables.

Gibbs-Shannon entropy based on probability.

- Gibbs-Shannon entropy as a thermodynamic limit of Boltzmann entropy:

$$\log\frac{N!}{N_1!\cdots N_m!}\approx\log\frac{\sqrt{2\pi N}\left(N/e\right)^N}{\sqrt{2\pi N_1}\left(N_1/e\right)^{N_1}\cdots\sqrt{2\pi N_m}\left(N_m/e\right)^{N_m}}$$

$$\rightarrow\underbrace{-N\sum_{i=1}^{m}p_i\log p_i}_{\mathcal{O}(N)}+\underbrace{\frac{1-m}{2}\log N}_{\mathcal{O}(\log N)}+\underbrace{\frac{1-m}{2}\log(2\pi)-\frac{1}{2}\sum_{i=1}^{m}\log p_i}_{\mathcal{O}(1)}$$

- In the limit of infinite N:

$$\frac{1}{N}\log\frac{N!}{N_1!\cdots N_m!}\xrightarrow[N\rightarrow\infty]{}-\sum_{i=1}^{m}p_i\log p_i$$

Stirling's formula for approximating factorial.

- Factorial of *N* is approximated when *N* is large:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\frac{1}{12n} - \frac{1}{360n^3} + \frac{1}{1260n^5} + \cdots\right)$$

- Derivation up to the leading order:

$$\log n! = \sum_{k=1}^{n} \log k$$

$$\approx \int_1^n dx \log x$$

$$= \left[x \log x\right]_1^n - (n-1) \approx \log\left(\frac{n}{e}\right)^n$$

$$n! \approx \left(\frac{n}{e}\right)^n$$

- More systematic derivation to higher order… (see, e.g., Mermin 1984)

Digression: Wallis' product formula for π.

- Why "π" in approximating a product of integers?

- Factorial of N is approximated when N is large:

$$\frac{2\cdot 2}{1\cdot 3}\times\frac{4\cdot 4}{3\cdot 5}\times\frac{6\cdot 6}{5\cdot 7}\times\cdots\times\frac{2k\cdot 2k}{(2k-1)\cdot(2k+1)}\times\cdots$$

$$=\prod_{k=1}^{\infty}\frac{2k\cdot 2k}{(2k-1)\cdot(2k+1)}$$

$$=\lim_{k\to\infty}\frac{2^{4k}(k!)^4}{(2k)!\cdot(2k+1)!}=\frac{\pi}{2}$$

Digression: Wallis' product formula for π.

- Derivation: Consider an following integral with large n:

$$\int_{-\pi}^{\pi} \cos^{2n} x \, dx$$

This integral is evaluated *exactly* by expanding the cosine:

$$\int_{-\pi}^{\pi} \cos^{2n} x \, dx = 2\int_{-\pi/2}^{\pi/2} \cos^{2n} x \, dx = 2\int_{-\pi/2}^{\pi/2} \left(\frac{e^{x} + e^{-x}}{2} \right)^{2n} dx$$

$$= 2\binom{2n}{n} 2^{-2n} = 2\frac{(2n)!}{(n!)^2} 2^{-2n}$$

This integral is evaluated *approximately* by replacing the cosine with a Gaussian:

$$\int_{-\pi}^{\pi} \cos^{2n} x \, dx = 2\int_{-\pi/2}^{\pi/2} \cos^{2n} x \, dx \approx 2\int_{-\pi/2}^{\pi/2} e^{-nx^2} \, dx = 2\sqrt{\frac{\pi}{n}}$$

Then we obtain:

$$\boxed{\frac{(2n)!}{(n!)^2} 2^{-2n} \approx \frac{1}{\sqrt{n\pi}}}$$

Digression: Wallis' product formula for π.

- Derivation: Consider an following integral with large n:

$$\int_{-\pi}^{\pi} \cos^{2n} x \, dx$$

This integral is evaluated *exactly* by expanding the cosine:

$$\int_{-\pi}^{\pi} \cos^{2n} x \, dx = 2\int_{-\pi/2}^{\pi/2} \cos^{2n} x \, dx = 2\int_{-\pi/2}^{\pi/2} \left(\frac{e^{x} + e^{-x}}{2} \right)^{2n} dx$$

$$= 2\binom{2n}{n} 2^{-2n} = 2\frac{(2n)!}{(n!)^2} 2^{-2n}$$

This integral is evaluated *approximately* by replacing the cosine with a Gaussian:

$$\int_{-\pi}^{\pi} \cos^{2n} x \, dx = 2\int_{-\pi/2}^{\pi/2} \cos^{2n} x \, dx \approx 2\int_{-\pi/2}^{\pi/2} e^{-nx^2} \, dx = 2\sqrt{\frac{\pi}{n}}$$

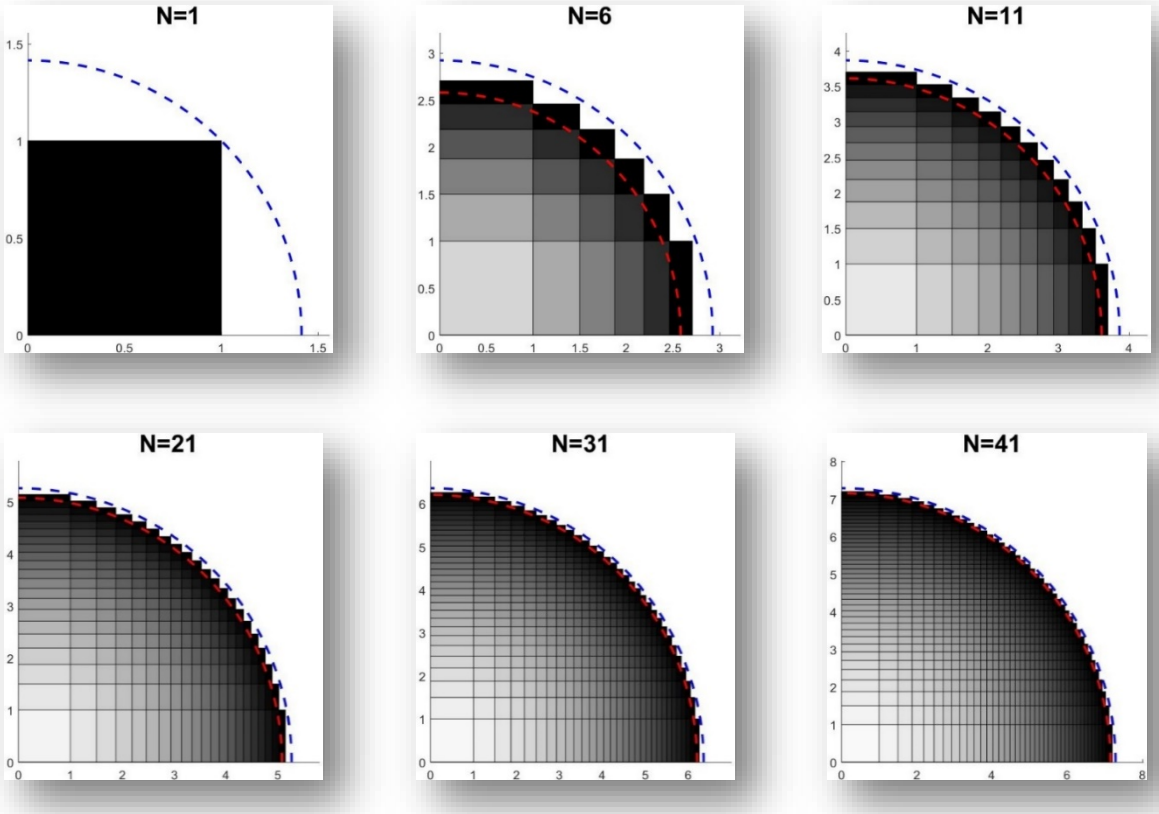
Then we obtain:

$$\boxed{\frac{(2n)!}{(n!)^2} 2^{-2n} \approx \frac{1}{\sqrt{n\pi}}}$$

Digression: Wallis' product formula for π.

$$\frac{(n-1)\pi}{2n} < \prod_{k=1}^{n-1} \frac{2k}{2k-1} \cdot \frac{2k}{2k+1} < \frac{(n+1)\pi}{2n}$$

$$\therefore \lim_{n\to\infty} \prod_{k=1}^{n-1} \frac{2k}{2k-1} \cdot \frac{2k}{2k+1} = \lim_{n\to\infty} \frac{(n-1)\pi}{2n} = \lim_{n\to\infty} \frac{(n+1)\pi}{2n} = \frac{\pi}{2}$$



Wästlund, J. (2007) Am Math Month

Digression: Numerical formulae for the Napier number.

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$$

$$e = \sum_{k=1}^{\infty}\frac{2k+2}{(2k+1)!}$$

$$\frac{e}{2} = \left(\frac{2}{1}\right)^{1/2}\cdot\left(\frac{2}{3}\cdot\frac{4}{3}\right)^{1/4}\cdot\left(\frac{4}{5}\cdot\frac{6}{5}\cdot\frac{6}{7}\cdot\frac{8}{7}\right)^{1/8}\cdot\left(\frac{8}{9}\cdot\frac{10}{9}\cdot\frac{10}{11}\cdot\frac{12}{11}\cdot\frac{12}{13}\cdot\frac{14}{13}\cdot\frac{14}{15}\cdot\frac{16}{15}\right)^{1/16}\cdots$$

$$u_1 = 1;\ u_{n+1} = (n+1)(u_n+1)$$

$$e = \prod_{n=1}^{\infty}\frac{u_n+1}{u_n} = \frac{2}{1}\cdot\frac{5}{4}\cdot\frac{16}{15}\cdot\frac{65}{64}\cdot\frac{326}{325}\cdots$$

$$p_n^* = \prod_{k=1}^{n} p_n$$

$$e = \lim_{n\to\infty}\left(p_n^*\right)^{1/p_n}$$

$$A_n = \frac{1}{n}\sum_{k=1}^{n} k = \frac{n+1}{2},\ G_n \equiv (n!)^{1/n}$$

$$e = \lim_{n\to\infty}\frac{A_n}{G_n}$$

McCartin (2006) Math Intelli

Axiomatic derivation of Shannon entropy.

Shannon's definition of entropy:

$$\mathrm{H}_m = \mathrm{H}_m\left(p_1, \cdots, p_m\right)$$

1. Continuity: the entropy is a continuous function of $\{p_i\}$.

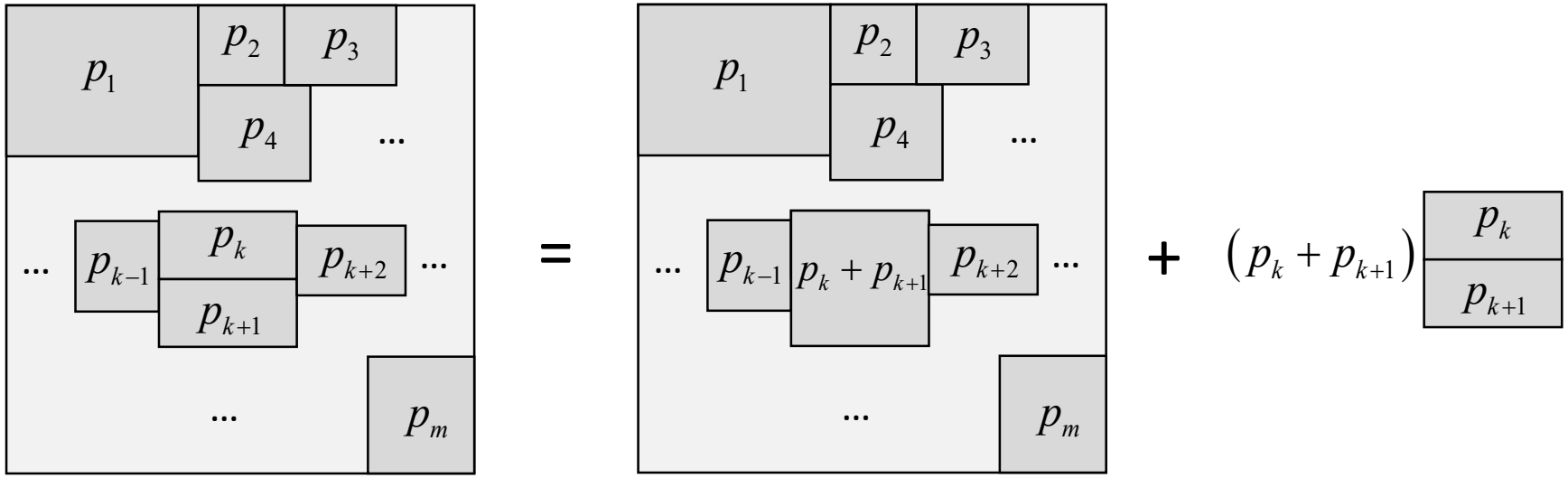
2. Monotonicity: if all n states are equiprobable with the probability $1/n$, the entropy depends only on n and is a monotonically increasing function of n.

3. Compositionality:

$$\begin{aligned}
&\mathrm{H}_n\left(p_1, \cdots, p_{k-1}, p_k, p_{k+1}, p_{k+2}, \cdots, p_n\right) \\
&\quad = \mathrm{H}_{n-1}\left(p_1, \cdots, p_{k-1}, p_k + p_{k+1}, p_{k+2}, \cdots, p_n\right) \\
&\qquad + \left(p_k + p_{k+1}\right)\mathrm{H}_2\left(\frac{p_k}{p_k + p_{k+1}}, \frac{p_{k+1}}{p_k + p_{k+1}}\right)
\end{aligned}$$

Compositionality: Coarse graining + microscopic detail.

- Entropy of m possible states is decomposed into a sum of entropy of (m-1) states and entropy of two states.



$$\mathrm{H}_m(p_1,\cdots,p_{k-1},p_k,p_{k+1},p_{k+2},\cdots,p_m) = \mathrm{H}_{m-1}(p_1,\cdots,p_{k-1},p_k+p_{k+1},p_{k+2},\cdots,p_n) + (p_k+p_{k+1})\,\mathrm{H}_2\left(\frac{p_k}{p_k+p_{k+1}},\frac{p_{k+1}}{p_k+p_{k+1}}\right)$$

Entropy of m-states

Entropy of $(m\text{-}1)$-states

k- and $(k+1)$-th states averaged

Entropy of 2-states

Microscopic details of k- and $(k+1)$-th states

- In other words, the entropy is decomposed into a coarse-grained part and a microscopic detail.

Compositionality: Coarse graining + microscopic detail.

- Entropy of m possible states is decomposed into a sum of entropy of (m-1) states and entropy of two states (compositionality).

$$f\left(m\right)=\mathrm{H}_m\left(\frac{1}{m},\cdots,\frac{1}{m}\right)$$

- Using the compositionality,

$$f\left(mn\right)=f\left(m\right)+f\left(n\right)$$

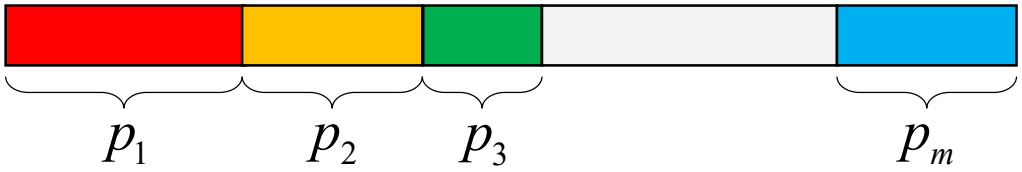
  is satisfied by logarithmic function:

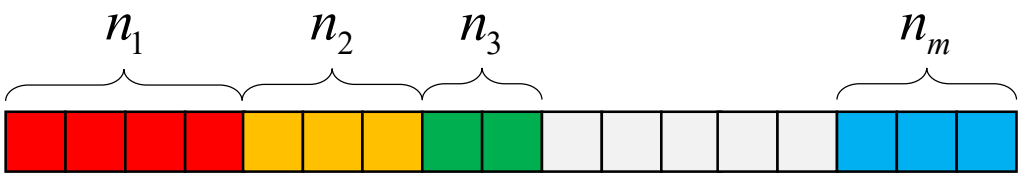
$$f\left(n\right)=K\log n.$$

Axiomatic derivation of Shannon entropy.

Shannon's definition of entropy:

$$H_m\left(p_1, p_2, p_3, \cdots, p_m\right)$$





$$\mathrm{H}_n\left(\frac{1}{n},\cdots,\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n},\cdots,\frac{1}{n},\cdots,\frac{1}{n}\right) = K\log n$$

Axiomatic derivation of Shannon entropy.

Shannon's definition of entropy:

$$\mathrm{H}_n\left(\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n_1},\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n_2},\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n_3},\cdots,\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n_m}\right)$$

$$=\mathrm{H}_{n-n_1+1}\left(\frac{n_1}{n},\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n_2},\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n_3},\cdots,\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n_m}\right)+\frac{n_1}{n}f\left(n_1\right)$$

$$=\mathrm{H}_{n-n_1-n_2+2}\left(\frac{n_1}{n},\frac{n_2}{n},\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n_3},\cdots,\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n_m}\right)+\frac{n_1}{n}f\left(n_1\right)+\frac{n_2}{n}f\left(n_2\right)$$

$$=\mathrm{H}_m\left(p_1,p_2,p_3,\cdots,p_m\right)+\sum_{i=1}^{m}p_i f\left(n_i\right)$$

Axiomatic derivation of Shannon entropy.

Shannon's definition of entropy:

$$f(n) = \mathrm{H}_m(p_1, p_2, p_3, \cdots, p_m) + \sum_{i=1}^{m} p_i f(n_i)$$

$$\mathrm{H}_m(p_1, p_2, p_3, \cdots, p_m) = f(n) - \sum_{i=1}^{m} p_i f(n_i)$$

$$= -K \log n + K \sum_{i=1}^{m} p_i \log n_i$$

$$= -K \sum_{i=1}^{m} p_i \log p_i$$

$$\mathrm{H}_m(\{p\}) = \mathrm{H}_m(p_1, \cdots, p_m) = -K \sum_{i=1}^{m} p_i \log p_i$$

Entropy as a measure of information coding.

- Consider a random variable with the sample space Ω and the probability:

$$\Omega_X = \{x_1, x_2, \cdots, x_n\}$$

$$p_i = \Pr(X = x_i)$$

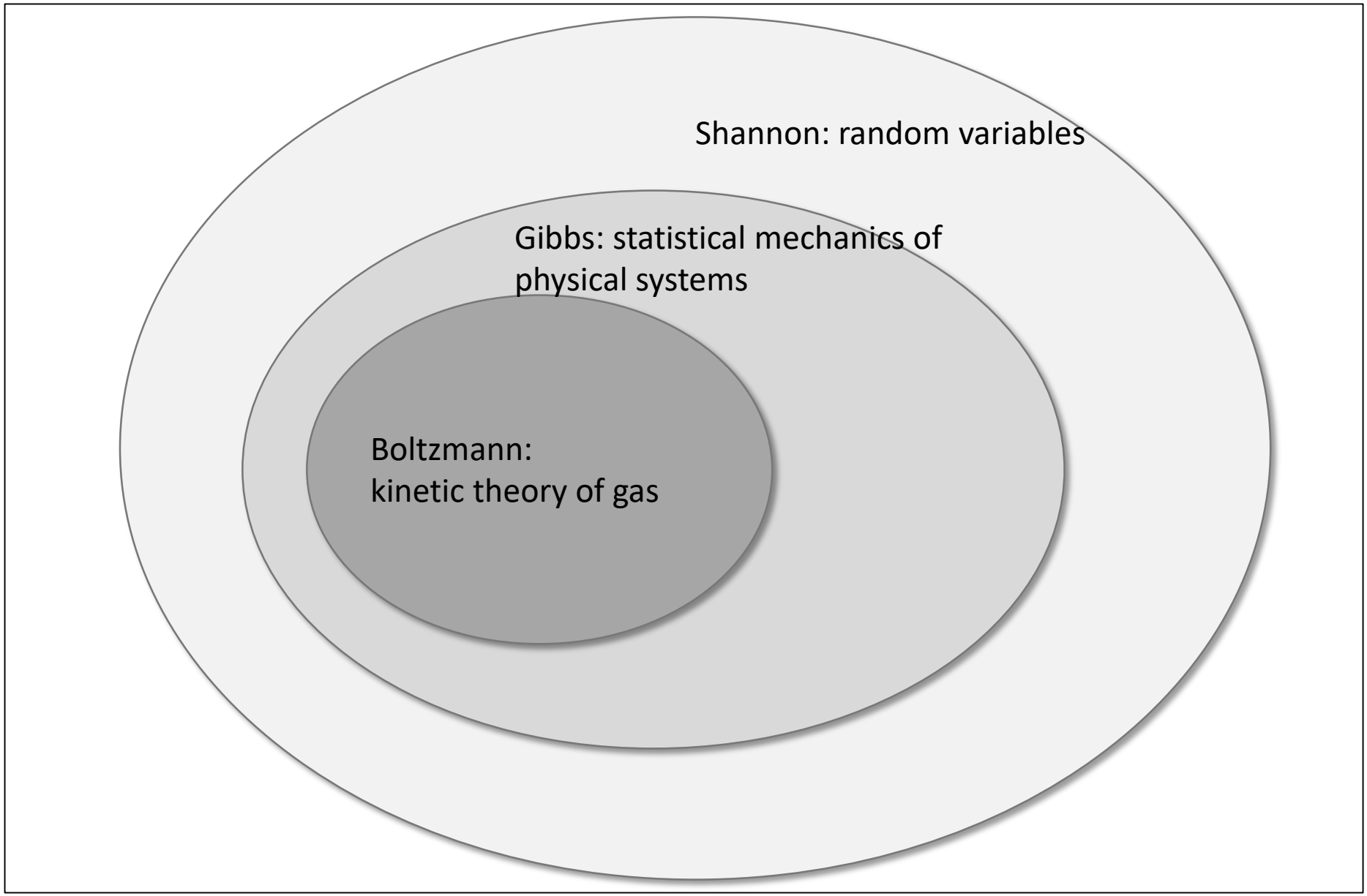
- How many bits are necessary to code a sequence of random numbers:

$$\exp\left(-N\mathrm{H}(X)\right) \approx W = \frac{N!}{\prod_{i=1}^{I} N_i}$$

- For an unbiased dice, the Shannon entropy is $\log_2 6$.

- For a biased dice that shows up only 1 and 2 with probability ½ each, the Shannon entropy is $\log_2 2 = 1$.

Entropy: from gas theory, physical systems to probability.



Differential entropy: entropy for a continuous variable.

- Shannon entropy for a *discrete* random variable X:

$$p_i = \Pr\left(X = x_i\right)$$

$$\boxed{\mathrm{H}\left(X\right) = -\sum_{i=1}^{n} p_i \log_2 p_i}$$

- Differential Shannon entropy for a *continuous* random variable X:

$$p\left(x\right)dx = \Pr\left(x \le X \le x + dx\right)$$

$$\boxed{\mathrm{H}\left(X\right) = -\int dx\ p\left(x\right)\log p\left(x\right)}$$

Note: The differential entropy is not invariant to a variable transformation

$$x \to y = f\left(x\right)$$

Differential entropy: entropy for a continuous variable.

- Differential Shannon entropy for a *continuous* random variable X:

$$\mathrm{H}(X) = -\int dx\ p_X(x) \log p_X(x)$$

Note: The differential entropy is not invariant to a variable transformation:

$$x \to y = f(x)$$

The probability density of y:

$$dx\ p_X(x) = dy\ p_Y(y) \to p_Y(y) = \left|\frac{\partial y}{\partial x}\right|^{-1} p_X(x)$$

$$\begin{aligned}
\mathrm{H}(X) &= -\int dx\ p_X(x) \log p_X(x) \\
&= -\int dy\ p_Y(y) \log p_Y(y) - \int dy\ p_Y(y) \log \left|\frac{\partial y}{\partial x}\right| \\
&= \mathrm{H}(Y) - \int dy\ p_Y(y) \log \left|\frac{\partial y}{\partial x}\right| \\
&\neq \mathrm{H}(Y)
\end{aligned}$$

Conditional entropy:

- Consider (not necessarily independent) two random variables X and Y:

$$p_X(x),\ \ p_Y(y),\ \ p_{XY}(x,y)$$

- Entropy of the variable X given that Y takes a particular value y:

$$\mathrm{H}(X \mid y) = -\int dx p(x \mid y) \log p(x \mid y)$$

- Taking the mean over Y defines the conditional entropy of X given Y:

$$\mathrm{H}(X \mid Y) = \int dy \mathrm{H}(X \mid y) = -\int dy dx p(x,y) \log p(x \mid y)$$

- If X and Y are dependent on each other, knowing about Y reduces the uncertainty of X. Therefore,

$$\mathrm{H}(X) \geq \mathrm{H}(X \mid Y)$$

Conditional entropy.

- Conditional entropy

$$\boxed{\mathrm{H}(X) \geq \mathrm{H}(X \mid Y)}$$

[Proof]:

$$\begin{aligned}
\mathrm{H}(X) - \mathrm{H}(X \mid Y) &= -\int dx p_X(x) \log p_X(x) + \int dx dy p_{XY}(x,y) \log p_{X|Y}(x \mid y) \\
&= \int dx dy p_{XY}(x,y) \log \frac{p_{X|Y}(x \mid y)}{p_X(x)} \\
&= \int dx dy p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)\, p_Y(y)} \\
&= -\int dx dy p_{XY}(x,y) \log \frac{p_X(x)\, p_Y(y)}{p_{XY}(x,y)} \\
&\geq -\log \int dx dy p_{XY}(x,y) \frac{p_X(x)\, p_Y(y)}{p_{XY}(x,y)} \\
&= -\log \int dx dy p_X(x)\, p_Y(y) \\
&= 0.
\end{aligned}$$

Kullback-Leibler divergence.

- KL divergence for a discrete random variable:

$$D_{KL}(q;p) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} \geq 0$$

- KL divergence for a continuous random variable:

$$D_{KL}(q;p) = \int dx p(x) \log \frac{p(x)}{q(x)} \geq 0$$

- Question: why not using an ordinary Euclidean distance?

$$\sum_{i=1}^{N} (p_i - q_i)^2 \qquad \text{or} \qquad \int dx \left(p(x) - q(x)\right)^2$$

Kullback-Leibler divergence.

- How to determine the "distance" between two density functions, $p(x)$ and $q(x)$?

$$D_{KL}(q;p) \equiv \int dx p(x) \log \frac{p(x)}{q(x)} \geq 0$$

- Example 1: two normal densities with an equal variance.

$$p(x) = \mathcal{N}(\mu_p, \sigma^2), \quad q(x) = \mathcal{N}(\mu_q, \sigma^2)$$

$$D_{KL}(q;p) = \frac{(\mu_p - \mu_q)^2}{2\sigma^2}$$

Kullback-Leibler divergence: multivariate normal distribution.

- Example 2: two multivariate normal densities.

$$p(\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\mu}_p, \Sigma_p\right), \quad q(x) = \mathcal{N}\left(\boldsymbol{\mu}_q, \Sigma_q\right)$$

$$D_{KL}(q;p) = \mathrm{E}_p\left[\log\frac{\left|\Sigma_p\right|^{-1/2}}{\left|\Sigma_q\right|^{-1/2}} - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_p\right)^T \Sigma_p^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_p\right) + \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_q\right)^{\mathrm{T}} \Sigma_q^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_q\right)\right]$$

$$= \frac{1}{2}\log\frac{\left|\Sigma_q\right|}{\left|\Sigma_p\right|} + \frac{1}{2}\mathrm{E}_p\left[-\left(\mathbf{x}-\boldsymbol{\mu}_p\right)^T \Sigma_p^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_p\right) + \left(\mathbf{x}-\boldsymbol{\mu}_q\right)^{\mathrm{T}} \Sigma_q^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_q\right)\right]$$

$$= \frac{1}{2}\left\{\log\frac{\left|\Sigma_q\right|}{\left|\Sigma_p\right|} + \mathrm{tr}\left(\Sigma_q^{-1}\Sigma_p\right) + \left(\mu_p-\mu_q\right)^{\mathrm{T}} \Sigma_q^{-1}\left(\mu_p-\mu_q\right) - n\right\}$$

- Note:

$$\mathrm{E}_p\left[\left(\mathbf{x}-\boldsymbol{\mu}_p\right)^{\mathrm{T}} \Sigma_p^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_p\right)\right] = \mathrm{E}_p\left[\mathrm{tr}\left\{\Sigma_p^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_p\right)\left(\mathbf{x}-\boldsymbol{\mu}_p\right)^{\mathrm{T}}\right\}\right] = \mathrm{tr}\left\{\Sigma_p^{-1}\Sigma_p\right\} = \mathrm{tr}\mathbf{I} = n.$$

Kullback-Leibler divergence.

- How to determine the "distance" between two probabilities, $\{p_i\}$ and $\{q_i\}$?

$$\mathrm{H}(\{p\}) = \lim_{N\to\infty} \log \frac{N!}{N_1!\cdots N_n!}, \quad \mathrm{H}(\{q\}) = \lim_{N\to\infty} \log \frac{N!}{M_1!\cdots M_n!},$$

$$\lim_{N\to\infty} \frac{\mathrm{H}(\{p\}) - \mathrm{H}(\{q\})}{N} = \lim_{N\to\infty} \frac{1}{N} \log \frac{M_1!\cdots M_n!}{N_1!\cdots N_n!}$$

$$= \lim_{N\to\infty} \frac{1}{N} \log \frac{M_1^{M_1}\cdots M_n^{M_n}}{N_1^{N_1}\cdots N_n^{N_n}}$$

$$= \lim_{N\to\infty} \frac{1}{N} \log \left(\frac{M_1}{N_1}\right)^{M_1} \cdots \left(\frac{M_n}{N_n}\right)^{M_n} N_1^{M_1-N_1}\cdots N_n^{M_n-N_n}$$

$$= \lim_{N\to\infty} \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i} + \frac{1}{N}\sum_{i=1}^{n} (M_i - N_i)\log N_i$$

$$= \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i}$$

$$\lim_{N\to\infty} \frac{\mathrm{H}(\{p\}) - \mathrm{H}(\{q\})}{N} = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i}$$

Kullback-Leibler divergence

- KL divergence for a continuous random variable is reparametrization invariant. Consider a transformation:

$$\mathbf{x} \to \mathbf{x}' = \mathbf{f}(\mathbf{x})$$

Under this transformation, the probability is conserved

$$p'_X(\mathbf{x}')\,d\mathbf{x}' = p_X(\mathbf{x})\,d\mathbf{x},\;\; q'_X(\mathbf{x}')\,d\mathbf{x}' = q_X(\mathbf{x})\,d\mathbf{x}$$

and the density functions are transformed as:

$$p'_X(\mathbf{x}') = p_X(\mathbf{x})\left|\frac{\partial \mathbf{x}'}{\partial \mathbf{x}}\right|^{-1},\; q'_X(\mathbf{x}') = q_X(\mathbf{x})\left|\frac{\partial \mathbf{x}'}{\partial \mathbf{x}}\right|^{-1}$$

Therefore, the ratio of p to q remains invariant

$$\frac{p'_X(\mathbf{x}')}{q'_X(\mathbf{x}')} = \frac{p_X(\mathbf{x})}{q_X(\mathbf{x})}$$

while the square needs an additional factor,

$$\left(p'_X(\mathbf{x}') - q'_X(\mathbf{x}')\right)^2 \left|\frac{\partial \mathbf{x}'}{\partial \mathbf{x}}\right|^{-2} = \left(p_X(\mathbf{x}) - q_X(\mathbf{x})\right)^2$$

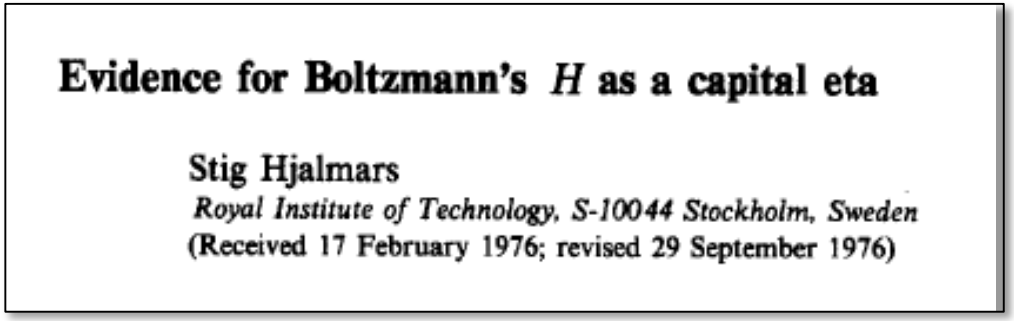
Therefore, the square norm depends particular parametrizations.

Digression: why “H” for entropy?

- From German Entropie, coined in 1865 by Rudolph Clausius, from Ancient Greek ἐντροπία (entropía, “a turning towards”), from ἐν (en, “in”) + τροπή (tropḗ, “a turning”).

  https://en.wiktionary.org/wiki/entropy

- Evidence for Boltzmann’s H as a capital eta



Evidence for Boltzmann's *H* as a capital eta

Stig Hjalmars
Royal Institute of Technology, S-10044 Stockholm, Sweden
(Received 17 February 1976; revised 29 September 1976)

“… this H, like all other Greek letters in the book, is printed vertical (nonslanted), while capital Latin letters are printed in italics (slanted types).”

Hjalmars (1977) Am J Phys

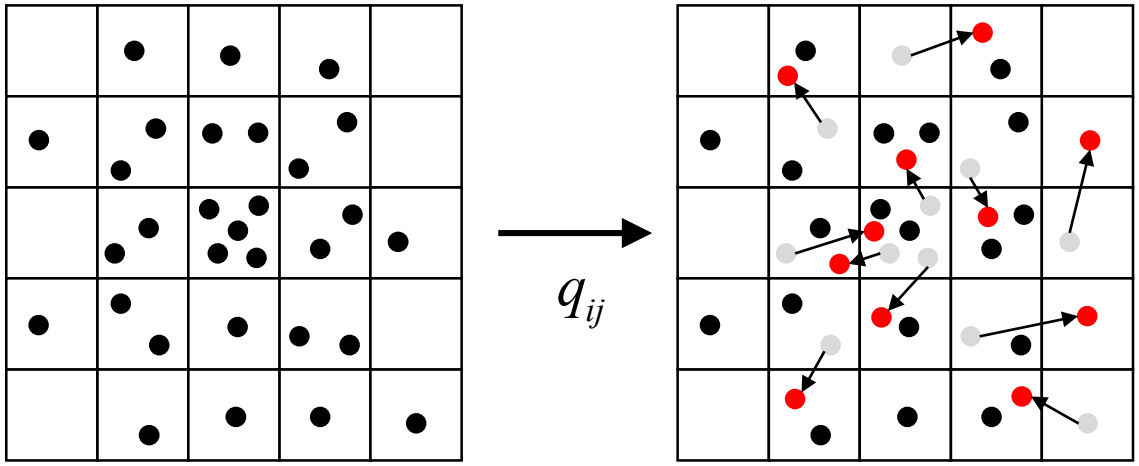
Why is it called “entropy?”

- What's in a name? In the case of Shannon's measure the naming was not accidental. In 1961 one of us (Tribus) asked Shannon what he had thought about when he had finally confirmed his famous measure. Shannon replied: "My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.' “ (Tribus & McIrvine (1971) Sci Am)

Topics to be covered today.

1. Entropy – it's origin in physics and information theory.
2. Markov process, H-theorem and renormalization group.
3. Random variable theorem and central limit theorem.
4. Independent component analysis as inverse of CLT.
5. Matrix decompositions and non-negative matrix factorization.
6. Tensor decompositions: PARAFAC and Tucker decompositions.
7. Task-related component analysis and its extensions.

Markov process: a process that depends on only one-step previous state.



$\{p_i(t)\}$ $\{p_i(t+1)\}$

$$p_i(t+1) = \sum_{i=1}^{N} q_{ij} p_j(t)$$

$$p_i(t) = \Pr\left(s(t) = s_i\right)$$

$$q_{ij} = \Pr\left(s(t+1) = s_i \mid s(t) = s_j\right) \qquad \sum_{i=1}^{N} q_{ij} = 1,\ q_{ij} \geq 0$$

H-theorem: non-decreasing entropy along time.

- In a Markov process, a non-increasing function can be defined.
- Let us introduce a concave function φ of probability and see how it behaves in a Markov process. Using Jensen's inequality,

$$\varphi\left(p_i(t+1)\right)=\varphi\left(\sum_j q_{ij}p_j(t)\right)\geq\sum_j q_{ij}\varphi\left(p_j(t)\right)$$

Then, summing over the indices gives

$$\sum_i\varphi\left(p_i(t+1)\right)\geq\sum_i\sum_j q_{ij}\varphi\left(p_j(t)\right)=\sum_j\varphi\left(p_j(t)\right)$$

Therefore, the function $\sum_i\varphi\left(p_i(t)\right)$ is a monotonically non-decreasing function. A particular choice of φ:

$$\varphi(x)=-x\log x$$

$$H\left(\{p(t)\}\right)=-\sum_i p_i(t)\log p_i(t)$$

H-theorem: why does entropy increase?

- Intuition: entropy takes a large value when the probability is uniform.
- A time step in a Markov process is a weighted average over all probabilities.

$$p_i(t+1) = \sum_{i=1}^{N} q_{ij} p_j(t) = q_{i1} p_1(t) + q_{i2} p_2(t) + \cdots + q_{iN} p_N(t)$$

Therefore, the minimum and maximum values of probability,

$$p^{\min}(t) \equiv \min_i p_i(t), \quad p^{\max}(t) \equiv \max_i p_i(t)$$

are non-decreasing and non-increasing functions of time, respectively.

$$p^{\min}(t-1) \le p^{\min}(t) \le p^{\min}(t+1) \le \cdots \le p^{\max}(t+1) \le p^{\max}(t) \le p^{\max}(t-1)$$

- The distribution becomes more uniform and the entropy hence increases as time goes by.

H-theorem: How can the entropy be reduced over time?

- As seen in the previous slide, in a Markov process,

$$p_i(t+1) = \sum_{i=1}^{N} q_{ij} p_j(t)$$

  the entropy is a non-decreasing function.

- There, all transition probabilities $\{q\}$ are *non-negative*.

- Observation: If we consider "unphysical" transition probabilities that take both positive and negative values, the entropy is no longer a non-decreasing function!

- This is an idea behind independent component analysis (we will come back later).

A Markov process distinguishes future and past.

- A Markov process forward in time:

$$p_i(t+1) = \sum_{j=1}^{N} q_{ij} p_j(t)$$

$$\sum_{i=1}^{N} q_{ij} = 1, \;\; q_{ij} \geq 0$$

- A Markov process *backward* in time:

$$p_i(t) = \sum_{j=1}^{N} \left(q^{-1}\right)_{ij} p_j(t+1)$$

$$\sum_{i=1}^{N} \left(q^{-1}\right)_{ij} = 1, \;\; \left(q^{-1}\right)_{ij} \gtrless 0$$

A Markov process distinguishes the directionality of arrow of time!

Example: diffusion process and random walk.

- Diffusion equation:

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2} D \frac{\partial^2 p(x,t)}{\partial x^2}$$

Discretize the space into bins of Δx and the time into bins of Δt:

$$\frac{p(x,t+\Delta t) - p(x,t)}{\Delta t} = \frac{1}{2} D \frac{p(x+\Delta x,t) - 2p(x+\Delta x,t) + p(x-\Delta x,t)}{(\Delta x)^2}$$

$$p(x,t+\Delta t) = \alpha p(x+\Delta x,t) + (1-2\alpha) p(x,t) + \alpha p(x-\Delta x,t)$$

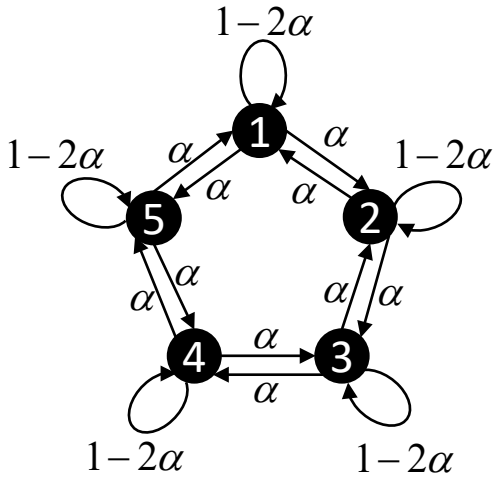
$$\alpha = \frac{D\Delta t}{2(\Delta x)^2}$$

$$p_i(t+1) = \sum_{j=1}^{N} q_{ij} p_j(t), \;\; q_{ii} = 1-2\alpha, \;\; q_{i,i+1} = q_{i-1,i} = \alpha$$

Example: random walk (or diffusion).

- Random walk: transition matrix.

$$\mathbf{Q} = \begin{pmatrix} 1-2\alpha & \alpha & 0 & 0 & \alpha \\ \alpha & 1-2\alpha & \alpha & 0 & 0 \\ 0 & \alpha & 1-2\alpha & \alpha & 0 \\ 0 & 0 & \alpha & 1-2\alpha & \alpha \\ \alpha & 0 & 0 & \alpha & 1-2\alpha \end{pmatrix}$$

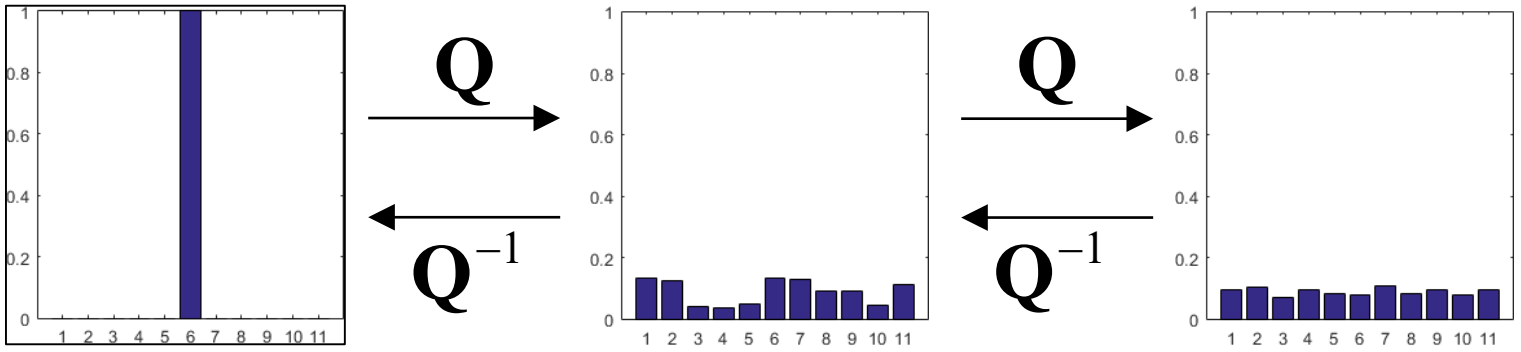


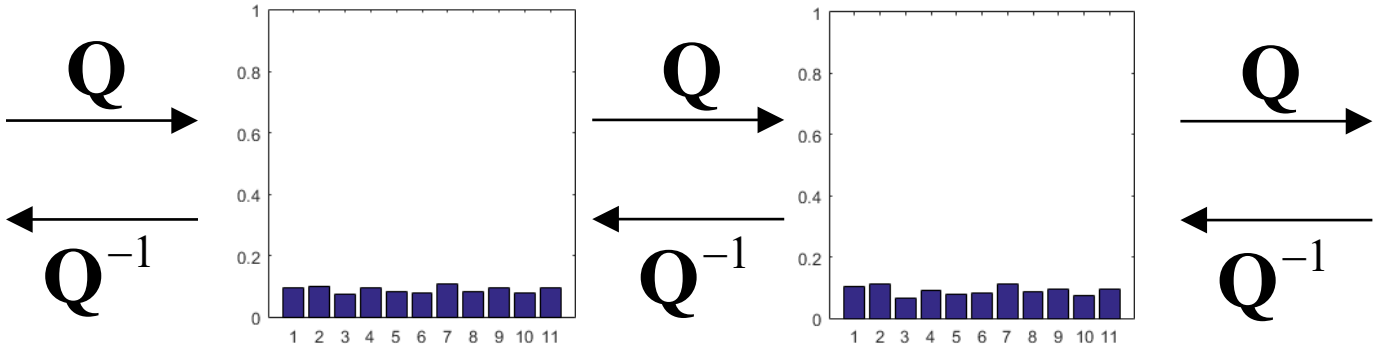
-> Inverse of transition matrix.

$$\mathbf{Q}^{-1} = \frac{1}{1-5\alpha+5\alpha^2} \begin{pmatrix} 1-3\alpha+\alpha^2 & -\alpha(1-\alpha) & \alpha^2 & \alpha^2 & -\alpha(1-\alpha) \\ -\alpha(1-\alpha) & 1-3\alpha+\alpha^2 & -\alpha(1-\alpha) & \alpha^2 & \alpha^2 \\ \alpha^2 & -\alpha(1-\alpha) & 1-3\alpha+\alpha^2 & -\alpha(1-\alpha) & \alpha^2 \\ \alpha^2 & \alpha^2 & -\alpha(1-\alpha) & 1-3\alpha+\alpha^2 & -\alpha(1-\alpha) \\ -\alpha(1-\alpha) & \alpha^2 & \alpha^2 & -\alpha(1-\alpha) & 1-3\alpha+\alpha^2 \end{pmatrix}$$

$$\alpha < \frac{1}{2} - \frac{\sqrt{5}}{10}$$

H-theorem: How can the entropy be reduced over time?





Non-negative matrix: its inverse is not non-negative.

- For any irreducible non-negative square matrix $\mathbf{A}=(a_{ij})$, its inverse is not generally non-negative (i.e., can be positive or negative).

$$\left(\mathbf{A}^{-1}\right)_{ij} \gtrless 0$$

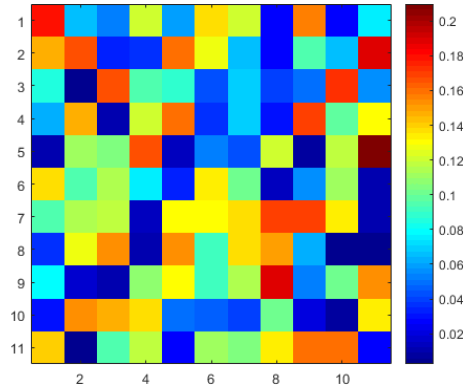
- Proof: $\mathbf{A}^{-1}$ must satisfy the following for different I and j, and if A is non-negative, $\mathbf{A}^{-1}$ must have some negative components.

$$0 = \sum_{j} \mathbf{A}^{-1}_{ij} \mathbf{A}_{jk}$$

- Exception is a monomial matrix (a generalized permutation matrix).

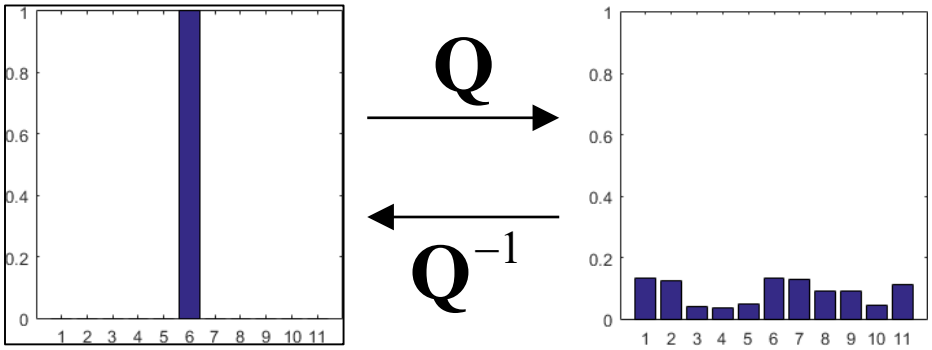
- Mixing (non-negative matrix) and un-mixing (its inverse) separate the arrow of time.

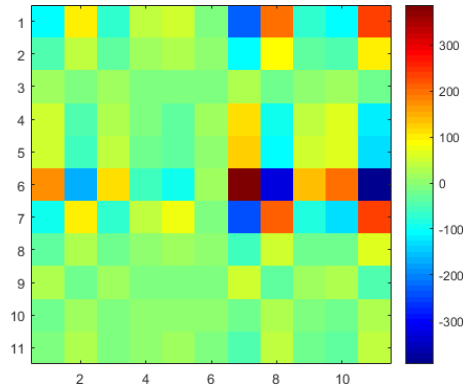
H-theorem: How can the entropy be reduced over time?



$$q_{ij} \geq 0$$

$$\sum_{i=1}^{N} q_{ij} = 1$$





$$\left(q^{-1}\right)_{ij} \gtrless 0$$

$$\sum_{i=1}^{N} \left(q^{-1}\right)_{ij} = 1$$

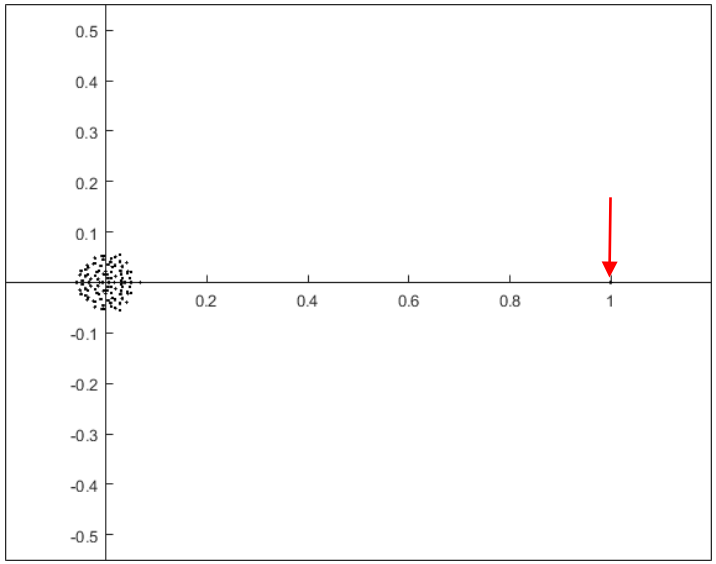
Non-negative matrix: The Perron-Frobenius theorem.

For any irreducible non-negative square matrix $\mathbf{A}=(a_{ij})$:

- There is a unique largest real eigenvalue r.
- The corresponding eigenvector has strictly positive components.
- The eigenvalue r satisfies the inequalities:

$$\min_i \sum_j a_{ij} \le r \le \max_i \sum_j a_{ij}$$

Therefore, for a Markov transition matrix, the Perron-Frobenius eigenvalue is 1.



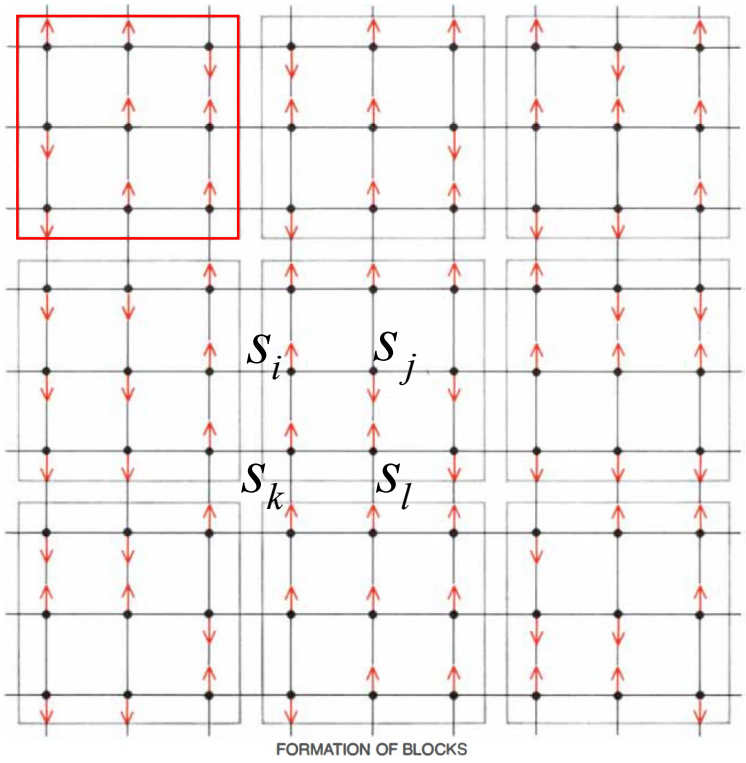
Renormalization group: a spin-system example.

- Renormalization group describes how a physical system behaves when microscopic details of interaction are integrated out scale by scale.
- The original idea was introduced in particle physics in 1950s (Gell-mann & Low) and was extended to statistical physics in 1960s and 1970s (Kadanoff; Wilson).
- Particularly, renormalization group is a powerful method for understanding long-range behaviors at a critical point (i.e., phase transition).

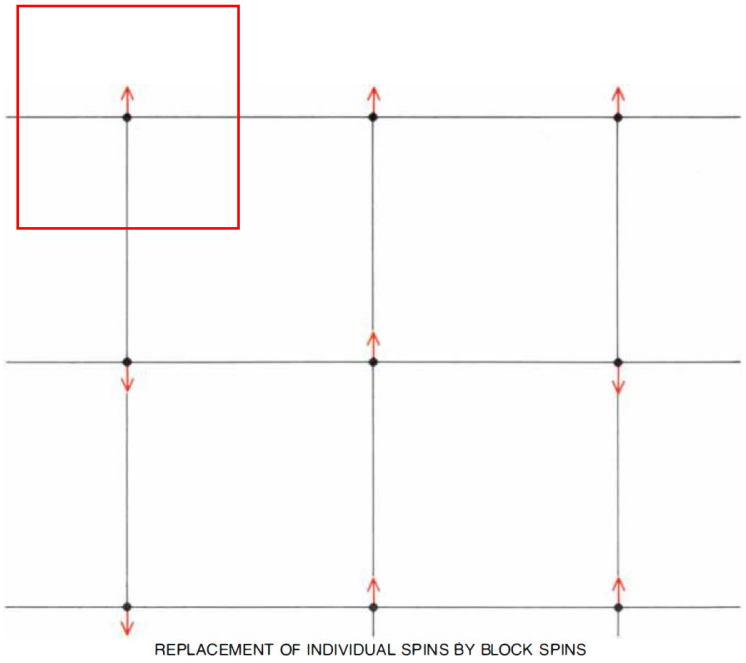
Kenneth G. Wilson (1936-2013)

Renormalization group: a spin-system example.

- Step 1: Replace a group of local spins with their representative value (usually their average):

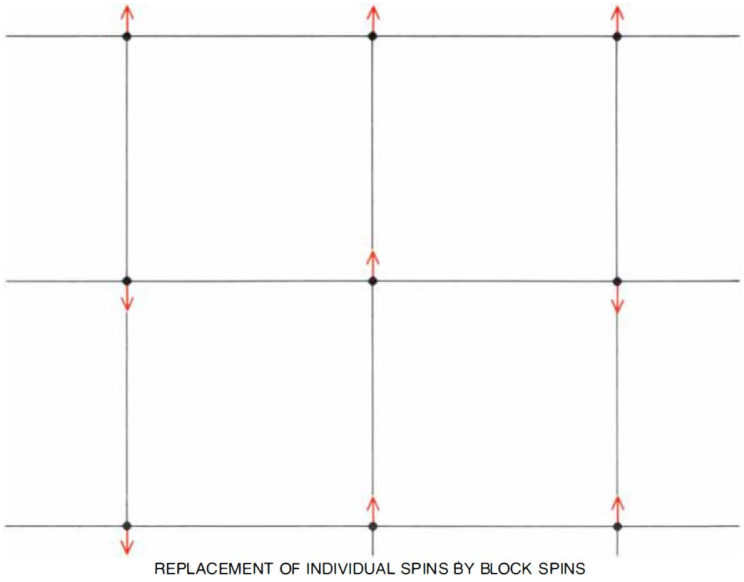


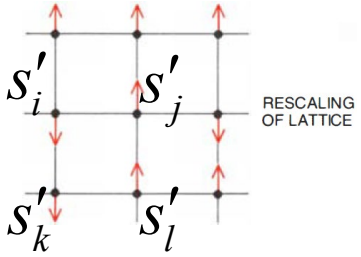
FORMATION OF BLOCKS


REPLACEMENT OF INDIVIDUAL SPINS BY BLOCK SPINS

Renormalization group: a spin-system example.

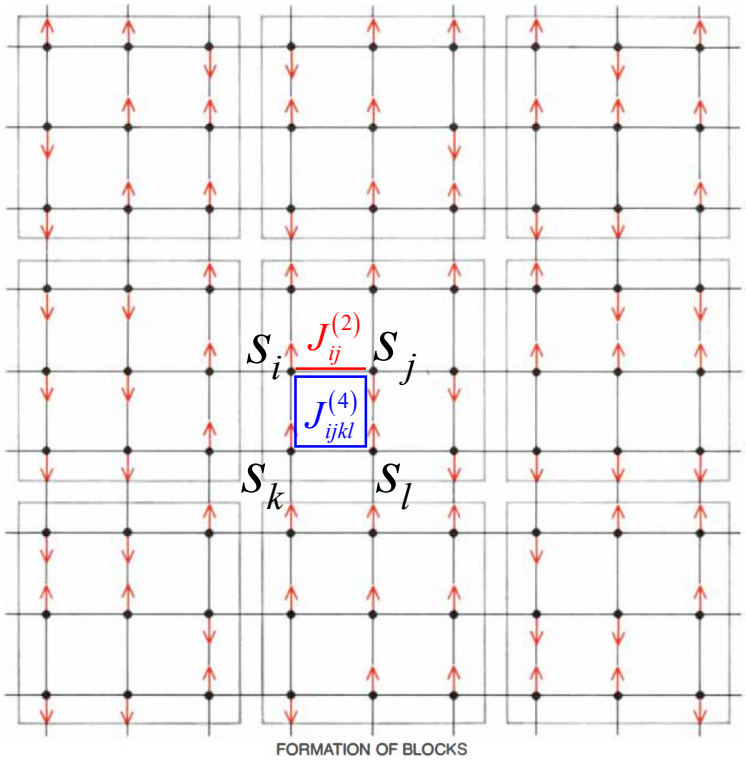
- Step 2: Rescale the spacing distance so that the transformed lattice has the same spacing of the original lattice.

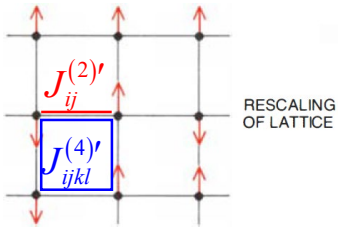




Renormalization group: a spin-system example.

- Step 3: Compute the Hamiltonian (or the interactions) of the transformed system from the Hamiltonian of the original system.





$$H\left(\{s_i\}\right)=\sum_{i,j}J_{ij}^{(2)}s_i s_j+\sum_{i,j}J_{ijkl}^{(4)}s_i s_j s_k s_l+\cdots$$

$$H\left(\{s_i'\}\right)=\sum_{i,j}J_{ij}^{(2)\prime}s_i' s_j'+\sum_{i,j}J_{ijkl}^{(4)\prime}s_i' s_j' s_k' s_l'+\cdots$$

Renormalization group: a spin-system example.

- Renormalization group describes how the system behaves when the microscopic details are integrated out.

$$\left\{J^{(2)}, J^{(4)}, \ldots\right\} \to \left\{J^{(2)\prime}, J^{(4)\prime}, \ldots\right\}$$
$$\to \left\{J^{(2)''}, J^{(4)\prime\prime}, \ldots\right\}$$
$$\to \cdots$$

- However, the explicit computation of this transformation is usually intractable.

- We will see that the central limit theorem is a simplest but non-trivial example of renormalization group.

Mechanical computation of matrix inversion in 1944.

九元連立方程式求解機: 1944年ころ 航空研究所製

アメリカのウィルバーは、1936（昭和11）年に土木の構造解析や経済学上の計算を行える機会を考案、制作した。本機はその情報を元に東京帝国大学航空研究所の佐々木達治郎や志賀亮、三井田純一らが1944（昭和19）年に作成した国内初の大型計算機械である。

Mechanical computation of differential equations.

Topics to be covered today.

1. Entropy – it's origin in physics and information theory.
2. Markov process, H-theorem and renormalization group.
3. Random variable theorem and central limit theorem.
4. Independent component analysis as inverse of CLT.
5. Matrix decompositions and non-negative matrix factorization.
6. Tensor decompositions: PARAFAC and Tucker decompositions.
7. Task-related component analysis and its extensions.

Random variable transformation (RVT) theorem.

- Suppose that n-dimensional random variables

$$\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^{\mathrm{T}} \in \mathbb{R}^n$$

  have a joint density function:

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}}(x_1 \quad \cdots \quad x_n)$$

- If m-dimensional variables

$$\mathbf{y} = \begin{pmatrix} y_1 & \cdots & y_m \end{pmatrix}^{\mathrm{T}} \in \mathbb{R}^m$$

  are uniquely determined from $\mathbf{x}$ by a mapping

$$y_i = f_i(x_1, \cdots, x_n) \text{ or } \mathbf{y} = \mathbf{f}(\mathbf{x})$$

  derive a joint density function for $\mathbf{y}$.

Random variable transformation (RVT) theorem.

- A joint density function for $\mathbf{y}$:

$$p_{\mathbf{y}}(\mathbf{y}) = \int d\mathbf{x} p_{\mathbf{x}}(\mathbf{x}) \prod_{i=1}^{m} \delta\left(y_i - f_i(\mathbf{x})\right)$$

  or in an element form:

$$p_{\mathbf{y}}(y_1, \cdots, y_m) = \int dx_1 \cdots \int dx_n p_{\mathbf{x}}(x_1, \cdots, x_n) \prod_{i=1}^{m} \delta\left(y_i - f_i(\mathbf{x})\right)$$

- From this theorem, a number of important results are derived: the central limit theorem, common probability density functions (lognormal, chi-square, and Student's *t*), the imposition of constraints, sampling of normal distribution, and the chi-square goodness-of-fit test.

Gillespie (1985) Am J Phys

Random variable transformation (RVT) theorem.

- A simple transformation: a variable y is a sum of x_1 and x_2.

$$y = \frac{x_1 + x_2}{\sqrt{2}}$$

then, the density of y is given by

$$p_{\mathbf{y}}(y) = \int dx_1 \int dx_2 \, p_{\mathbf{x}}(x_1, x_2) \, \delta\left(y - \frac{x_1 + x_2}{\sqrt{2}} \right).$$

If x_1 and x_2 are i.i.d., then

$$\begin{aligned} p_{\mathbf{y}}(y) &= \int dx_1 \int dx_2 \, p_{\mathbf{x}}(x_1) \, p_{\mathbf{x}}(x_2) \, \delta\left(y - \frac{x_1 + x_2}{\sqrt{2}} \right) \\ &= \sqrt{2} \int dx_1 \int dx_2 \, p_{\mathbf{x}}(x_1) \, p_{\mathbf{x}}(x_2) \, \delta\left(x_2 - \sqrt{2} y + x_1 \right) \\ &= \sqrt{2} \int dx_1 \, p_{\mathbf{x}}(x_1) \, p_{\mathbf{x}}\left(\sqrt{2} y - x_1 \right) \end{aligned}$$

Random variable transformation (RVT) theorem.

- A simple transformation: a variable y is a sum of x_1 and x_2.

$$p_{\mathbf{y}}(y) = \int dx q(y,x)\, p_{\mathbf{x}}(x)$$

$$q(y,x) \geq 0, \quad \int dy q(y,x) = 1.$$

- Following the same reasoning that derived the H-theorem, for any concave function, we have:

$$\varphi\left(p_{\mathbf{y}}(y)\right) = \varphi\left(\int dx q(y,x)\, p_{\mathbf{x}}(x)\right) \geq \int dx q(y,x) \varphi\left(p_{\mathbf{x}}(x)\right)$$

By integrating over y, we now have:

$$\int dy \varphi\left(p_{\mathbf{y}}(y)\right) \geq \int dy \int dx q(y,x) \varphi\left(p_{\mathbf{x}}(x)\right) = \int dx \varphi\left(p_{\mathbf{x}}(x)\right)$$

Random variable transformation (RVT) theorem.

- The inequality is now

$$\int dy \varphi\left(p_{\mathbf{y}}(y)\right) \geq \int dx \varphi\left(p_{\mathbf{x}}(x)\right)$$

- If we define the convex function as

$$\varphi(x) = -x \log x$$

and the entropy as

$$\mathrm{H}(X) = \int dx \varphi\left(p_X(x)\right)$$

then, we see that the entropy is a non-decreasing function.

$$\boxed{\mathrm{H}(Y) \geq \mathrm{H}(X)}$$

Law of large number.

- Consider a set of n i.i.d. random variables X_i with mean μ and variance σ^2:

$$X_1, X_2, \cdots, X_n$$

- Let us introduce a new random variable defined as:

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- The law of large number: In the limit of infinite n, the distribution of the new variable becomes Dirac's delta function peaked at μ :

$$\overline{X}_n \sim \delta\left(\overline{x}_n - \mu\right)$$

Law of large number: derivation.

- Let us consider a variable Y:

$$Y = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Applying the RVT:

$$p_Y(y) = \int dx_1 \cdots \int dx_n p_X(x_1) \cdots p_X(x_n) \delta\left(y - \frac{x_1 + \cdots + x_n}{n} \right)$$

$$= \int dx_1 \cdots \int dx_n p_X(x_1) \cdots p_X(x_n) \int \frac{dt}{2\pi} e^{-it\left(y - \frac{x_1 + \cdots + x_n}{n} \right)}$$

$$= \int \frac{dt}{2\pi} e^{-ity} \int dx_1 p_X(x_1) e^{\frac{itx_1}{n}} \cdots \int dx_n p_X(x_1) e^{\frac{itx_n}{n}}$$

$$= \int \frac{dt}{2\pi} e^{-ity} \left(\varphi_X\left(\frac{t}{n} \right) \right)^n$$

Therefore, the characteristic function of Y is a power of the characteristic function of X:

$$\varphi_Y(t) = \left(\varphi_X\left(\frac{t}{n} \right) \right)^n$$

Law of large number: derivation.

- In the limit of infinite n, only the first moment (mean) survives and the higher moments will vanish:

$$\varphi_Y(t) = \left(\varphi_X\left(\frac{t}{n}\right)\right)^n$$

$$= \left(\sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{it}{n}\right)^k \left\langle X^k \right\rangle\right)^n$$

$$= \left(1 + \frac{it}{n}\left\langle X\right\rangle + \frac{1}{2!}\left(\frac{it}{n}\right)^2 \left\langle X^2\right\rangle + \frac{1}{3!}\left(\frac{it}{n}\right)^3 \left\langle X^3\right\rangle + \cdots\right)^n$$

$$\to e^{it\langle X\rangle}$$

$$\therefore p_Y(y) \to \delta\left(y - \left\langle X\right\rangle\right)$$

- Therefore, the density of the variable Y approaches to the delta function peaked at <X>.

Central limit theorem.

- Consider a set of n i.i.d. random variables X_i with mean 0 and variance σ^2:

$$X_1, X_2, \cdots, X_n$$

- Let us introduce a new random variable defined as:

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}$$

- The central limit theorem (CLT): In the limit of infinite n, the distribution of the new variable becomes a Gaussian distribution with mean 0 and variance σ^2:

$$\bar{X}_n \sim \mathcal{N}\left(0, \sigma^2\right)$$

Central limit theorem: sketch of derivation.

- As in the case of law of large number:

$$\varphi_Y(t) = \left(\varphi_X\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

In the limit of large n, only the second moment survives and the higher moments vanish:

$$\varphi_Y(t) = \left(\varphi_X\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

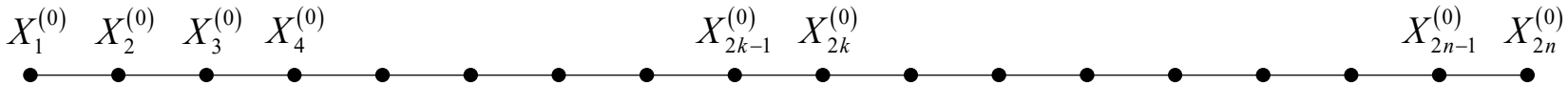
$$= \left(1 + \frac{1}{2!}\left(\frac{it}{\sqrt{n}}\right)^2 \langle X^2 \rangle + \frac{1}{3!}\left(\frac{it}{\sqrt{n}}\right)^3 \langle X^3 \rangle + \cdots\right)^n \to e^{-\frac{\langle X \rangle^2}{2} t^2}$$

$$\therefore p_Y(y) \to \mathcal{N}\left(0, \langle X \rangle^2\right)$$

- Therefore, the density function of Y approaches to a normal density.

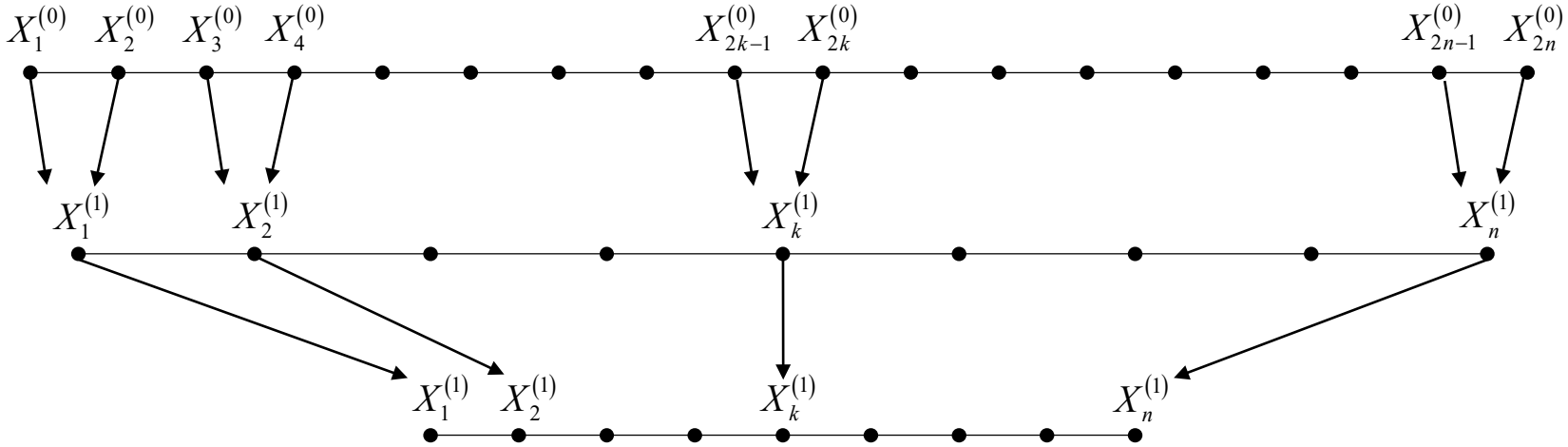
Central limit theorem as renormalization group.

- Consider a set of i.i.d. $2n$ random variables $X^{(0)}$.



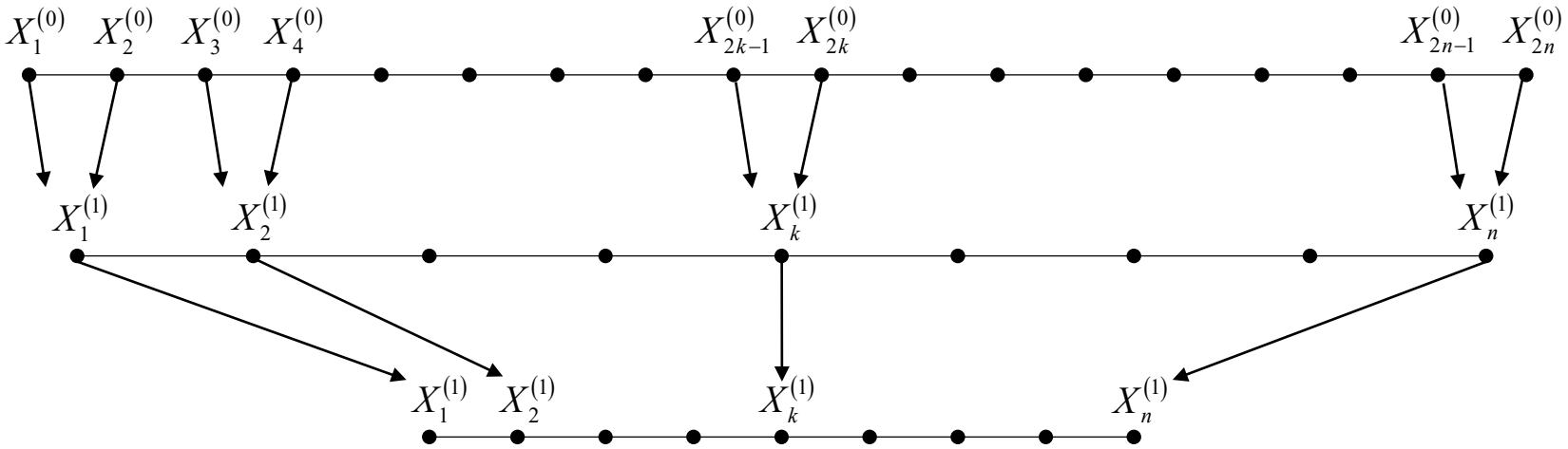
"Block-spin" transformation of random variables:

$$X_k^{(1)} = \frac{X_{2k-1}^{(0)} + X_{2k}^{(0)}}{\sqrt{2}}$$



Central limit theorem as renormalization group.

- Renormalization group of moments:



The Renormalization Group: A Probabilistic View (*).

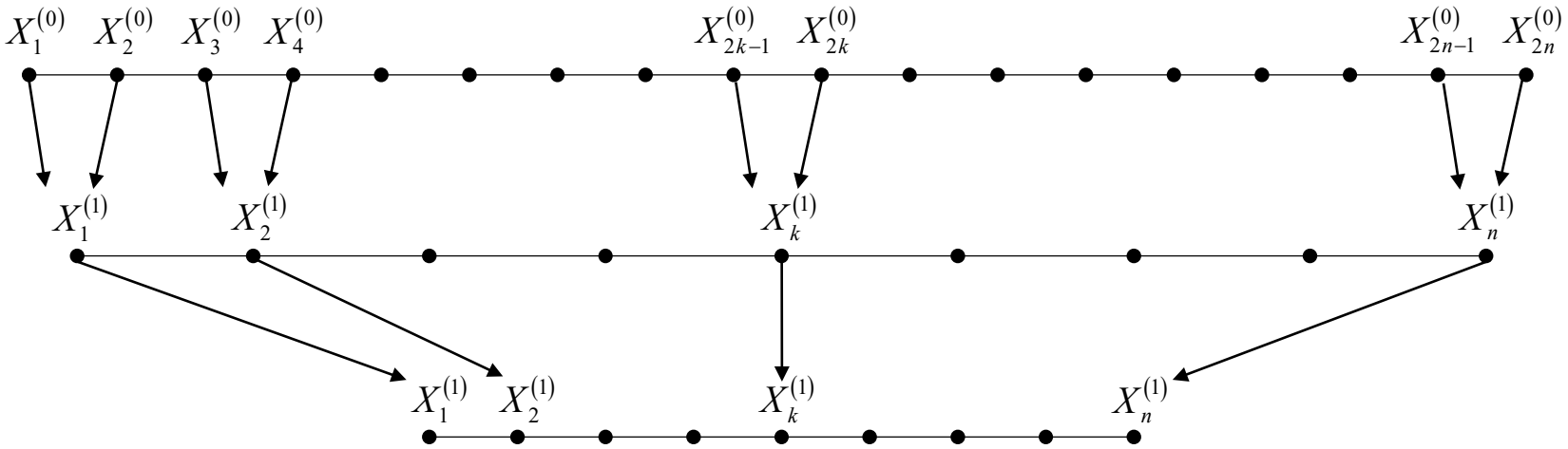
G. JONA-LASINIO (**)

Istituto di Fisica dell'Università - Padova
Gruppo Nazionale di Struttura della Materia - Padova

(ricevuto il 7 Agosto 1974)

Central limit theorem as renormalization group.

- Renormalization group of moments:



$$\kappa_k^{(1)} = \frac{\kappa_k^{(0)}}{2^{\frac{k}{2}-1}}$$

$$\kappa_k^{(1)} = \frac{\kappa_k^{(0)}}{2^{\frac{k}{2}-1}}$$

$$\left\{\kappa_2^{(0)}, \kappa_3^{(0)}, \kappa_4^{(0)}, \kappa_5^{(0)} \cdots\right\} \to \left\{\kappa_2^{(1)}, \kappa_3^{(1)}, \kappa_4^{(1)}, \kappa_5^{(1)} \cdots\right\} = \left\{\kappa_2^{(0)}, \frac{\kappa_3^{(0)}}{\sqrt{2}}, \frac{\kappa_4^{(0)}}{2}, \frac{\kappa_5^{(0)}}{2^{3/2}}, \cdots\right\}$$

CLT as renormalization group transformation.

- Higher-order cumulants decrease after each iteration of the transformation:

$$\left\{\kappa_2^{(0)},\kappa_3^{(0)},\kappa_4^{(0)},\kappa_5^{(0)}\cdots\right\}\to\left\{\kappa_2^{(1)},\kappa_3^{(1)},\kappa_4^{(1)},\kappa_5^{(1)}\cdots\right\}=\left\{\kappa_2^{(0)},\frac{\kappa_3^{(0)}}{\sqrt{2}},\frac{\kappa_4^{(0)}}{2},\frac{\kappa_5^{(0)}}{2^{3/2}},\cdots\right\}$$

$$\to\left\{\kappa_2^{(2)},\kappa_3^{(2)},\kappa_4^{(2)},\kappa_5^{(2)}\cdots\right\}=\left\{\kappa_2^{(0)},\frac{\kappa_3^{(0)}}{2},\frac{\kappa_4^{(0)}}{2^2},\frac{\kappa_5^{(0)}}{2^3},\cdots\right\}$$

$$\to\cdots$$

$$\to\left\{\kappa_2^{(0)},0,0,0,\cdots\right\}$$

- Therefore, “the microscopic details” of the original distribution forget iteration by iteration, approaching to a Gaussian distribution.

Maximum entropy principle

- Principle of insufficient reasoning (Laplace)
  Two events are to be assigned equal probabilities of there is no reason to think otherwise.

- Principle of maximum entropy (Janes)
  Distributions are determined so as to maximize the entropy (as a measure of uncertainty) in a way consistent with given measurements.

$$H\left(\{p_i\}\right) = -\sum_i p_i \log p_i$$

$$\sum_i p_i = 1$$

$$\sum_i p_i f\left(x_i\right) = f_0$$

Maximum entropy derivation of Boltzmann distribution.

- Find the maximum entropy distribution

$$H\left(\{p_i\}\right) = -\sum_i p_i \log p_i$$

with the normalization and the energy constrains:

$$\sum_i p_i = 1, \left\langle E \right\rangle = \sum_i p_i \varepsilon_i = E_0$$

- This problem is solved by introducing an augmented Lagrangian:

$$J\left(\{p_i\}, \alpha, \beta\right) = H\left(\{p_i\}\right) - \alpha\left(\sum_i p_i - 1\right) - \beta\left(\sum_i p_i \varepsilon_1 - \left\langle E \right\rangle\right)$$

This gives the Boltzmann distribution of canonical states:

$$p_i = \frac{e^{-\beta\varepsilon_i}}{\sum_j e^{-\beta\varepsilon_j}}$$

Topics to be covered today.

1. Entropy – it's origin in physics and information theory.
2. Markov process, H-theorem and renormalization group.
3. Random variable theorem and central limit theorem.
4. Independent component analysis as inverse of CLT.
5. Matrix decompositions and non-negative matrix factorization.
6. Tensor decompositions: PARAFAC and Tucker decompositions.
7. Task-related component analysis and its extensions.

Independent component analysis as an inverse of central limit theorem.

- Physical processes that mix independent signal sources with positive coefficients increase the entropy of observed signals.
- In other words, observed signals are more “Gaussian-like” than original signal sources.
- We don’t know either of mixing coefficients and distributions of original sources; reconstructing original sources from observed signal is an ill-posed problem.
- However, if it is possible to construct variables from observed data that differ as much from Gaussian as possible, they should be simila to independent sources.

Independent component analysis: A linear formulation.

$$\mathbf{x} = \left(x_1, \cdots, x_n\right)^{\mathrm{T}} \in \mathbb{R}^n, \ \mathbf{s} = \left(s_1, \cdots, s_n\right)^{\mathrm{T}} \in \mathbb{R}^n, \ \mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

- Generative model

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

- Linear unmixing

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

Independent component analysis: A linear formulation.

- The goal of ICA is to estimate the unmixing matrix W and the source signals s:

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s}$$

- The signals are recovered if

$$\mathbf{W}\mathbf{A} = \mathbf{P}_{\pi}$$

  where $\mathbf{P}_{\pi}$ is a permutation matrix.

- The Amari index:

Rank-one matrix decomposition

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1t} & \cdots & x_{1T} \\ \vdots & \ddots & & & \vdots \\ x_{n1} & & x_{nt} & & x_{nT} \\ \vdots & & & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nt} & \cdots & x_{NT} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_t & \cdots & \mathbf{x}_T \end{bmatrix} \in \mathbb{R}^{N \times T}$$

$$\mathbf{S} = \begin{bmatrix} s_{11} & \cdots & s_{1t} & \cdots & s_{1T} \\ \vdots & \ddots & & & \vdots \\ s_{n1} & & s_{nt} & & s_{nT} \\ \vdots & & & \ddots & \vdots \\ s_{N1} & \cdots & s_{Nt} & \cdots & s_{NT} \end{bmatrix} = \begin{bmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_t & \cdots & \mathbf{s}_T \end{bmatrix} \in \mathbb{R}^{N \times T}$$

$$\left(\mathbf{x}_t\right)_n = x_{nt} = \sum_{n'=1}^{N} a_{nn'} s_{n't} = \left(\mathbf{A}\mathbf{s}_t\right)_n$$

$$\mathbf{X} = \mathbf{A}\mathbf{S}$$

Rank-one matrix decomposition of a matrix.

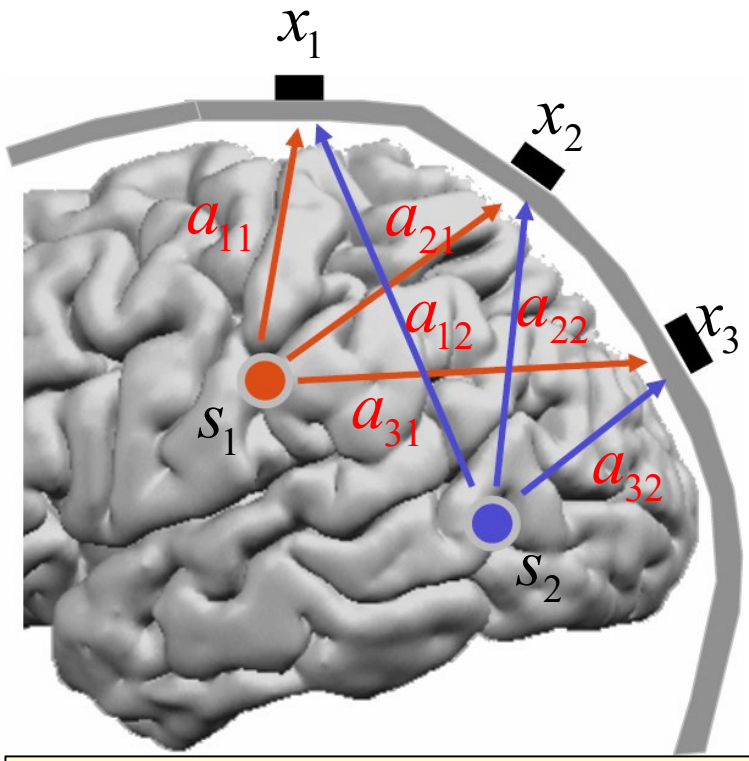
$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1t} & \cdots & x_{1T} \\ \vdots & \ddots & & & \vdots \\ x_{n1} & & x_{nt} & & x_{nT} \\ \vdots & & & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nt} & \cdots & x_{NT} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_N^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{N \times T}, \quad \mathbf{S} = \begin{bmatrix} s_{11} & \cdots & s_{1t} & \cdots & s_{1T} \\ \vdots & \ddots & & & \vdots \\ s_{n1} & & s_{nt} & & s_{nT} \\ \vdots & & & \ddots & \vdots \\ s_{N1} & \cdots & s_{Nt} & \cdots & s_{NT} \end{bmatrix} = \begin{bmatrix} \mathbf{s}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{s}_n^{\mathrm{T}} \\ \vdots \\ \mathbf{s}_N^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{N \times T}$$

$$\mathbf{X} = \mathbf{AS} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n & \cdots & \mathbf{a}_N \end{bmatrix} \begin{bmatrix} \mathbf{s}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{s}_n^{\mathrm{T}} \\ \vdots \\ \mathbf{s}_N^{\mathrm{T}} \end{bmatrix} = \sum_{n=1}^{N} \mathbf{a}_n \mathbf{s}_n^{\mathrm{T}}$$

$$= \mathbf{a}_1 \mathbf{s}_1^{\mathrm{T}} + \mathbf{a}_2 \mathbf{s}_2^{\mathrm{T}} + \cdots + \mathbf{a}_N \mathbf{s}_N^{\mathrm{T}}$$

Positive weight mixing in physical phenomena.

- Physical processes mix source signals with *positive* coefficients.
- In other words, observed signals are weighted averages of source signals.



mixing process (physical)

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t)$$
$$x_2(t) = a_{21}s_1(t) + a_{32}s_2(t) \qquad a_{ij} \geq 0$$
$$x_3(t) = a_{31}s_1(t) + a_{32}s_2(t)$$

unmixing process (unphysical)

$$s_1(t) = w_{11}x_1(t) + w_{12}x_2(t) + w_{13}x_3(t)$$
$$s_2(t) = w_{21}x_1(t) + w_{22}x_2(t) + w_{23}x_3(t) \qquad w_{ij} \gtrless 0$$

The observed signals $\{x_i\}$ are more "Gaussian-like" than the source signals $\{s_i\}$.

ICA: various approaches toward "non-Gaussianity."

- Mutual information as a measure of independence (InfoMax, extended InfoMax, AMICA) (Bell & Sejnowski, 1995; Lee et al. 1999)

$$\sum_{i=1}^{n} \mathrm{H}_i\left(y_i\right) - \mathrm{H}\left(y_1, \cdots, y_n\right) = \int dy_1 \cdots dy_n p_{\mathbf{y}}\left(y_1, \cdots, y_n\right) \log \frac{p_{\mathbf{y}}\left(y_1, \cdots, y_n\right)}{p_1\left(y_1\right) \cdots p_n\left(y_n\right)} \geq 0$$

- Moment-based method: Kurtosis maximization (FastICA) (Hyvarinen & Oja, 1997).

$$\mathrm{H}\left(y\right) = \mathrm{H}_{\mathrm{Gauss}} - \frac{\kappa_3^2}{2 \cdot 3!} - \frac{\kappa_4^2}{2 \cdot 4!} + \cdots$$

- Joint-diagonalization (JADE, SOBI) (Cardoso, 1997; Belouchrani et al. 1997)

$$C\left(\tau_1\right) = C\left(\tau_2\right) = \cdots = C\left(\tau_T\right) = 0$$

Infomax learning algorithm (1/3): mutual information.

- Find mutually independent components by minimizing the KL divergence:

$$I(\mathbf{W}) = \sum_{i=1}^{n} \mathrm{H}_i(y_i) - \mathrm{H}(y_1, \cdots, y_n) = \int dy_1 \cdots dy_n \, p_{\mathbf{y}}(y_1, \cdots, y_n) \log \frac{p_{\mathbf{y}}(y_1, \cdots, y_n)}{p_1(y_1) \cdots p_n(y_n)}$$

- Applying a formula,

$$p_{\mathbf{y}}(y_1, \cdots, y_n) = |\mathbf{W}|^{-1} \, p_{\mathbf{x}}(x_1, \cdots, x_n)$$

$$I(\mathbf{W}) = H(X) - \log|\mathbf{W}| - \sum_{i=1}^{n} \int dx_1 \cdots dx_n \, p_{\mathbf{x}}(x_1, \cdots, x_n) \log p_i(y_i)$$

Infomax learning algorithm (2/3): Matrix derivative:

- Taking the derivative with respect to $\mathbf{W}$:

$$\frac{\partial I(\mathbf{W})}{\partial W_{ij}} = -\left(\mathbf{W}^{-\mathrm{T}}\right)_{ij} + \int dx_1 \cdots dx_n p_{\mathbf{x}}(x_1, \cdots, x_n) \left\{ -\frac{1}{p_i(y_i)} \frac{\partial p_i(y_i)}{\partial y_i} x_j \right\}$$

$$= -\left(\mathbf{W}^{-\mathrm{T}}\right)_{ij} + \left\langle \varphi_i(y_i) x_j \right\rangle_{\mathbf{x}}$$

or in a matrix form:

$$\frac{\partial I(\mathbf{W})}{\partial \mathbf{W}} = -\mathbf{W}^{-\mathrm{T}} + \left\langle \boldsymbol{\varphi}(\mathbf{y}) \mathbf{x}^{\mathrm{T}} \right\rangle_{\mathbf{x}}$$

- The infomax learning algorithm (Bell & Sejnowski, 1995):

$$\Delta \mathbf{W} = -\eta \left[\mathbf{W}^{-\mathrm{T}} - \left\langle \boldsymbol{\varphi}(\mathbf{y}) \mathbf{x}^{\mathrm{T}} \right\rangle_{\mathbf{x}} \right]$$

Infomax learning algorithm (3/3): Natural gradient.

- The gradient itself is not a steepest direction – it has to be multiplied by $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ from the right:

$$(\Delta\mathbf{W})\mathbf{W}^{\mathrm{T}}\mathbf{W} = -\eta\left[\mathbf{W}^{-\mathrm{T}} - \left\langle\boldsymbol{\varphi}(\mathbf{y})\mathbf{x}^{\mathrm{T}}\right\rangle_{\mathbf{x}}\right]\mathbf{W}^{\mathrm{T}}\mathbf{W}$$
$$= -\eta\left[\mathbf{I} - \left\langle\boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^{\mathrm{T}}\right\rangle_{\mathbf{x}}\right]\mathbf{W}$$

- The natural-gradient infomax algorithm (Amari 1998):

$$\Delta\mathbf{W} = -\eta\left[\mathbf{I} - \left\langle\boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^{\mathrm{T}}\right\rangle_{\mathbf{x}}\right]\mathbf{W}$$

Here the factor in the parenthesis is regarded as nonlinear decorrelation.

$$\mathbf{I} = \left\langle\boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^{\mathrm{T}}\right\rangle_{\mathbf{x}}$$

Choice of the score function: super- or sub-Gaussian.

- The score function realizes our prior expectation about the source distribution:

$$\varphi(y) = -\frac{\partial \log p_y(y)}{\partial y}$$

Originally, $p_y(y)$ is assumed to be super-Gaussian (Bell & Sejnowski):

$$p_y(y) \propto \frac{1}{\cosh(y)} \Rightarrow \varphi(y) = \tanh(y)$$

- Later, $p_y(y)$ is assumed to be super- or sub-Gaussian (Lee et al.):

$$p_y^{\text{super}}(y) \propto \frac{e^{-\frac{y^2}{2}}}{\cosh^2(y)} \Rightarrow \varphi(y) = 1 + \tanh(y)$$

$$p_y^{\text{sub}}(y) = \frac{1}{2}\left[\mathcal{N}(y;-1,1) + \mathcal{N}(y;+1,1)\right] \Rightarrow \varphi(y) = 1 - \tanh(y)$$

Often called as “extended infomax algorithm.”

Amari index: Performance measure of ICA.

- In general, the unmixing matrix is not determined uniquely up to scaling and permutations.

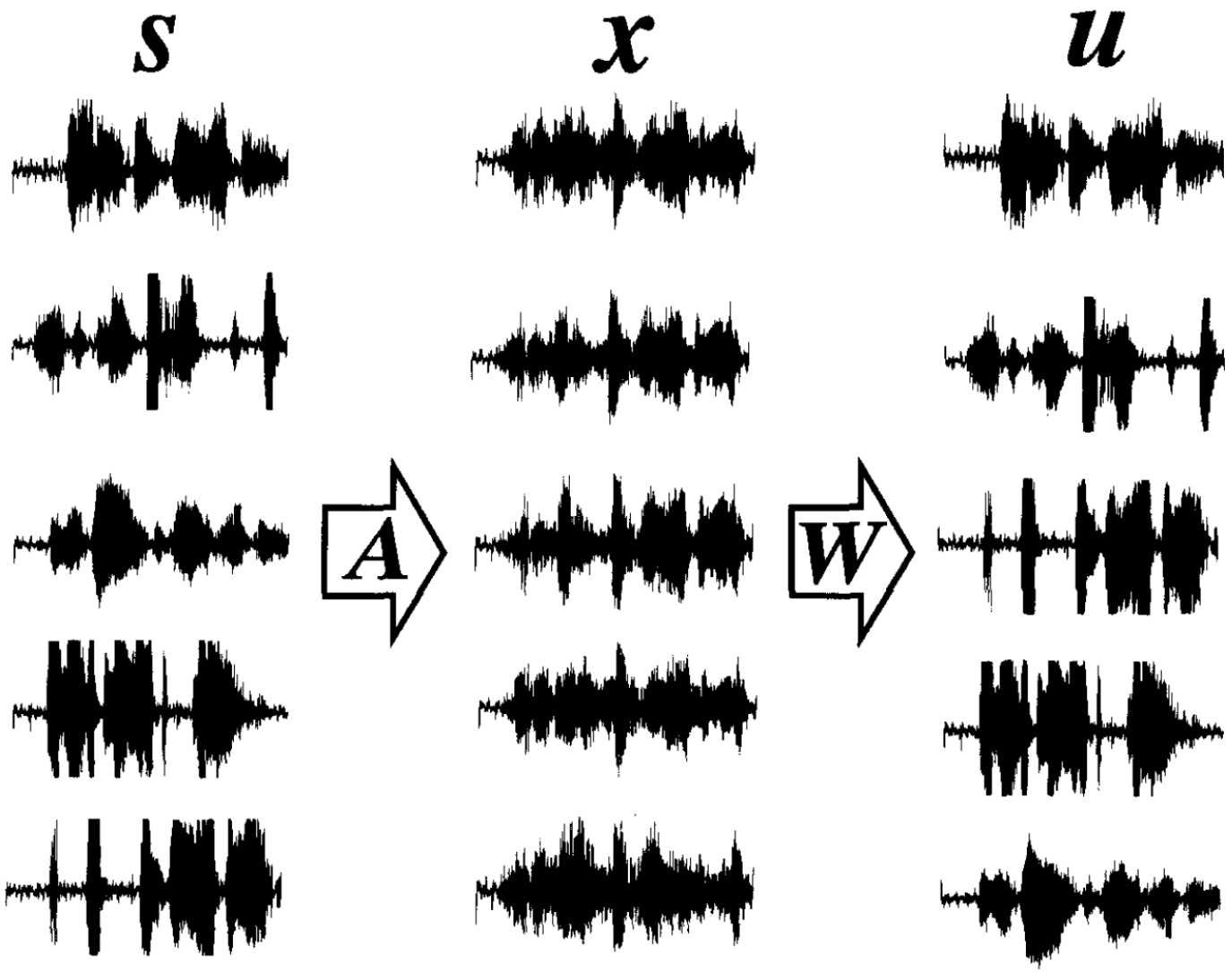
$$\mathbf{WA} = \mathbf{DP}_{\pi}$$

- Therefore, the Amari index for a matrix G is:

$$\frac{1}{2n(n-1)}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{|G_{ij}|}{\max_{k}|G_{ik}|}-1\right)+\frac{1}{2n(n-1)}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}\frac{|G_{ij}|}{\max_{k}|G_{kj}|}-1\right)$$

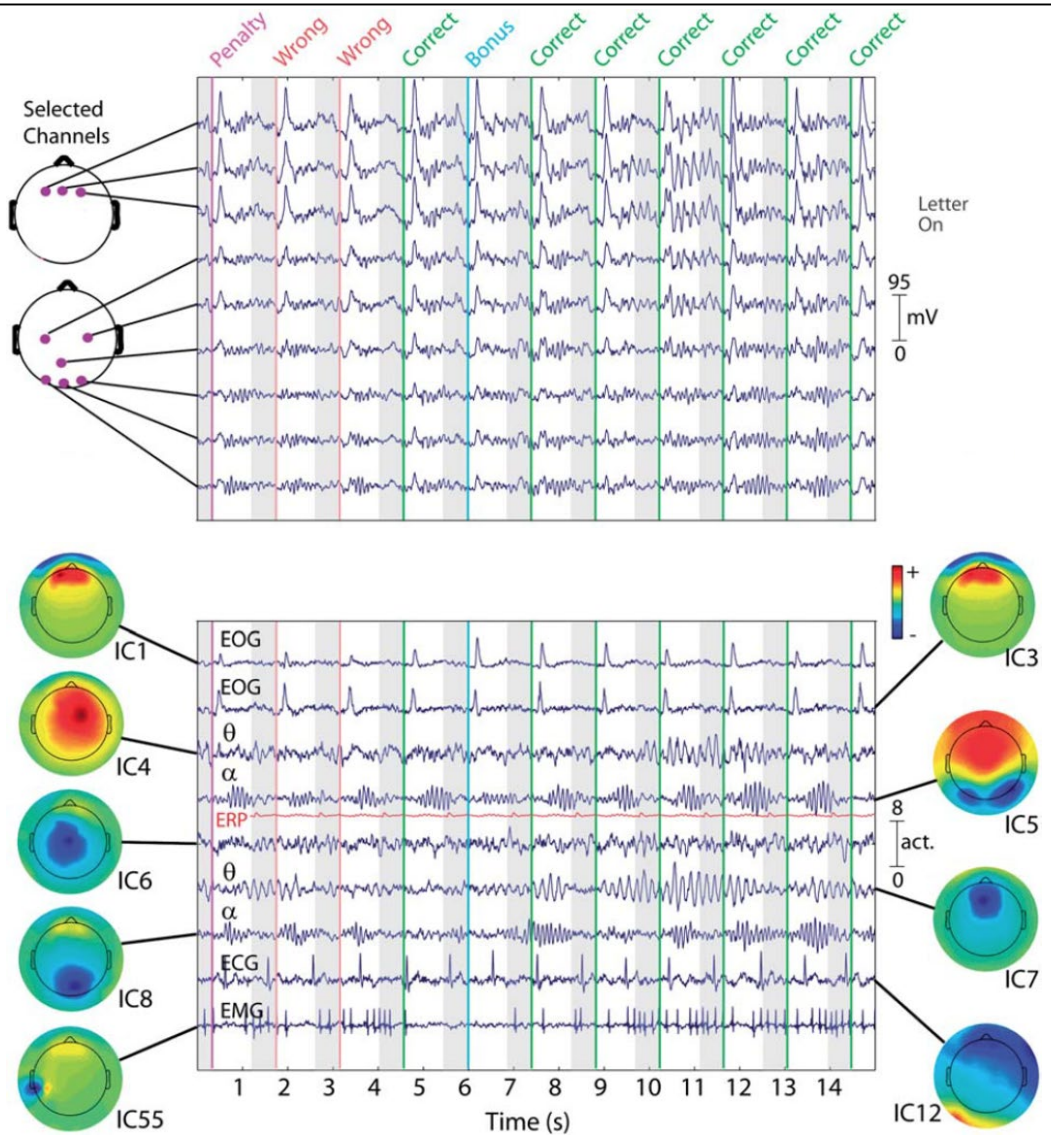
The index takes a maximum value of 1 when G is a generalized permutation matrix.

ICA Example 1: Sound source separation.



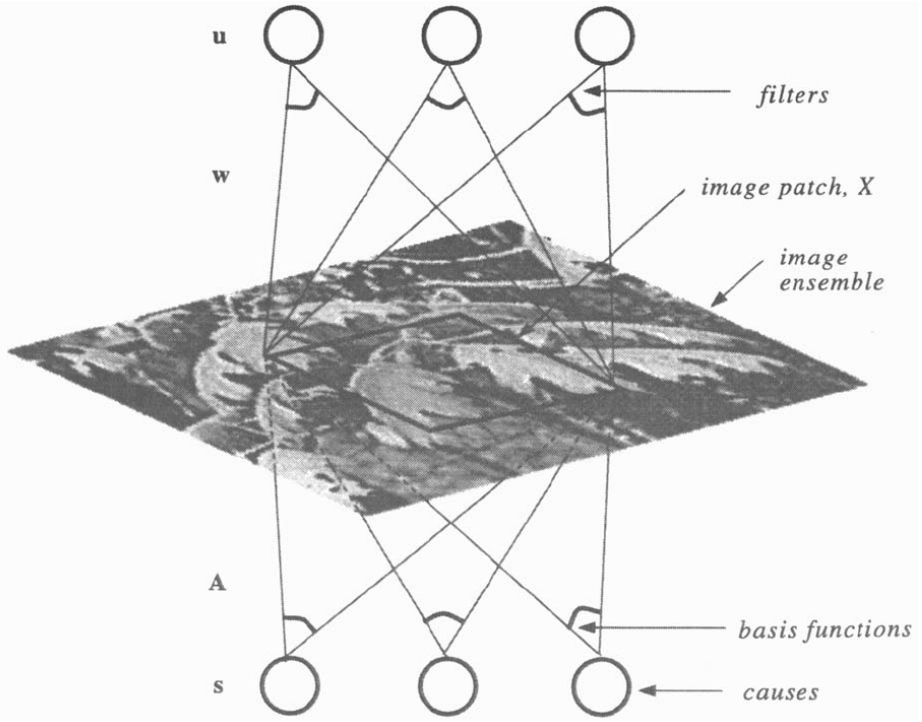
Bell & Sejnowski (1995) Neural Comput

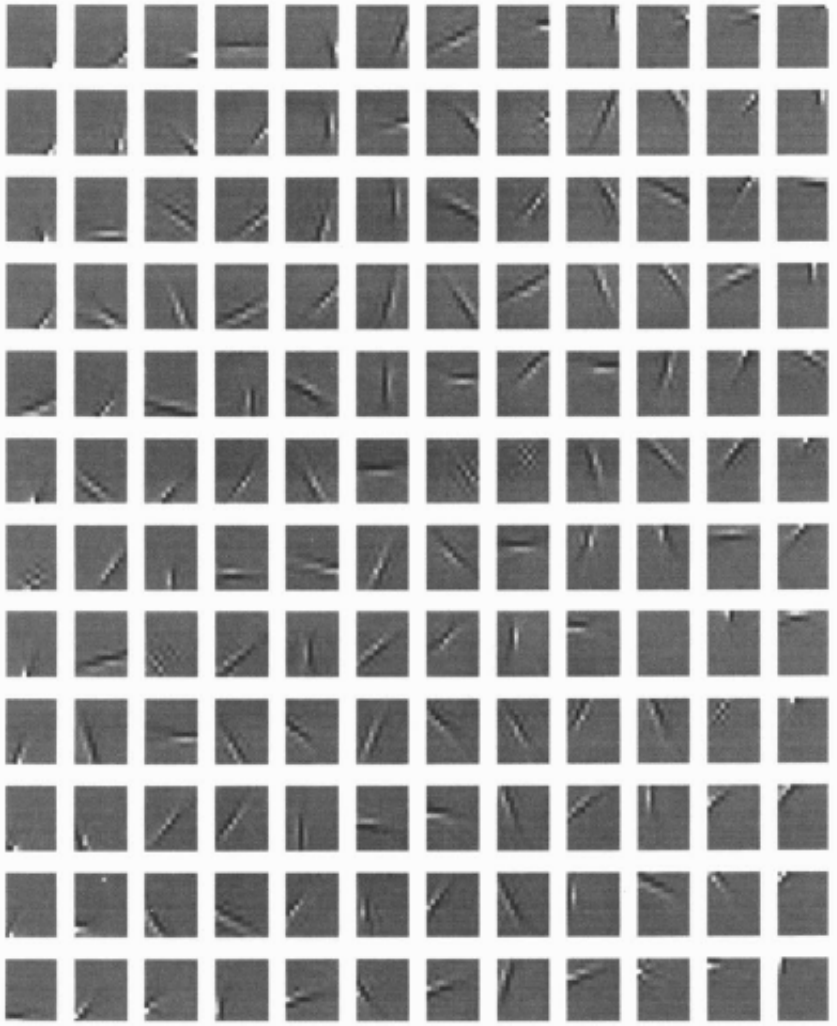
ICA Example 2: EEG source separation.



Onton & Makeig (2006) Prog Neurobiol

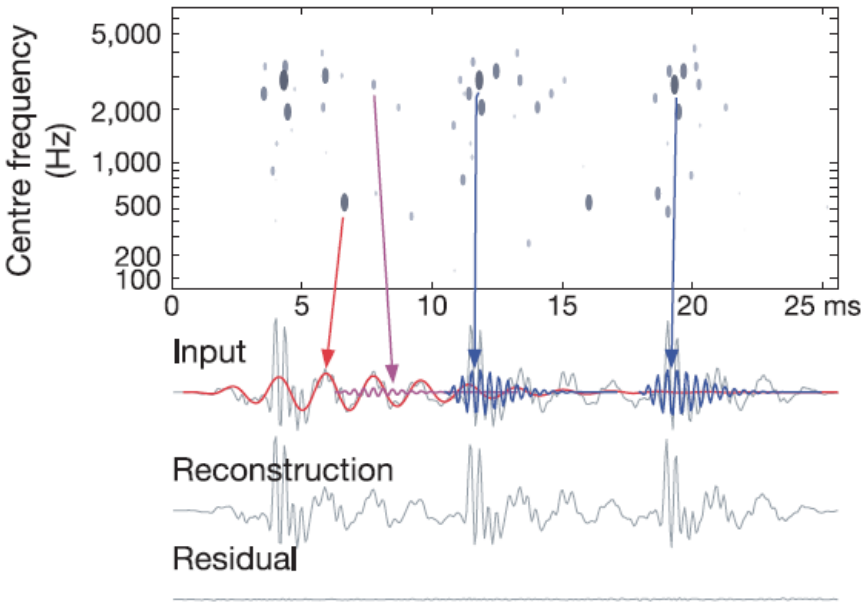
ICA Example 4: Independent components of natural scenes.

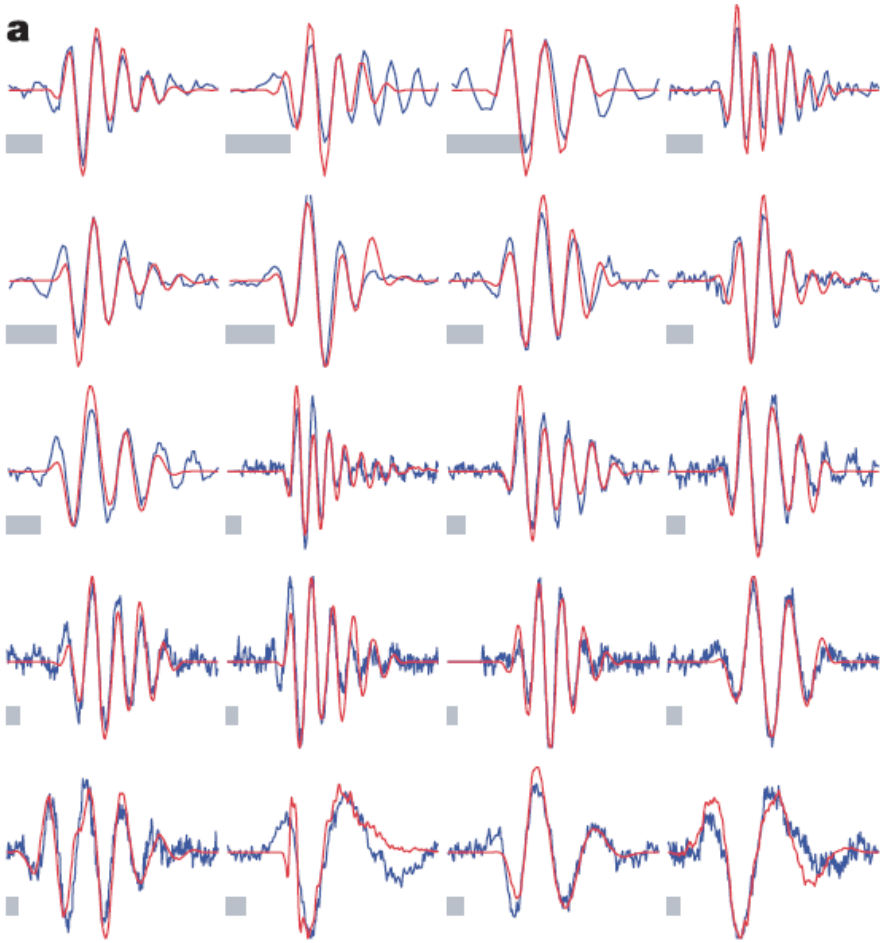




Bell & Sejnowski (1997) Vis Res.

ICA Example 5: Cochlear filters as independent sound sources.





Smith & Lewicki (2006) Nature

Topics to be covered today.

1. Entropy – it's origin in physics and information theory.
2. Markov process, H-theorem and renormalization group.
3. Random variable theorem and central limit theorem.
4. Independent component analysis as inverse of CLT.
5. Matrix decompositions and non-negative matrix factorization.
6. Tensor decompositions: PARAFAC and Tucker decompositions.
7. Task-related component analysis and its extensions.

Matrix decompositions in linear algebra.

- Eigenvalue decomposition.
- Singular value decomposition.
- Cholesky decomposition.
- LU decomposition.
- QR decomposition.
- Schur decomposition.

Non-negative matrix factorization (NMF).

- Many kinds of data are non-negative (e.g., images, texts, spectrum, EMG, …).
- Approximating an $M \times N$ non-negative matrix $\mathbf{Y}$

$$y_{nm} \geq 0$$

as a product of $M \times K$ non-negative matrix $\mathbf{A}$ and $K \times N$ non-negative matrix $\mathbf{X}$:

$$\mathbf{Y} \approx \mathbf{AX} \quad \text{subject to } \mathbf{A} \geq 0, \mathbf{X} \geq 0$$

- Solve a constrained optimization problem:

$$\min_{\mathbf{A} \geq 0, \mathbf{X} \geq 0} \frac{1}{2} \left\| \mathbf{Y} - \mathbf{AX} \right\|_{\mathrm{F}}^2$$

Frobenius norm of matrix.

- The Frobenius norm of a matrix is defined

$$\|\mathbf{X}\|_{\mathrm{F}}^2 = \sum_{n,m} x_{nm}^2 = \mathrm{tr}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathrm{tr}\mathbf{X}\mathbf{X}^{\mathrm{T}}.$$

- A derivative of the norm is

$$\frac{\partial}{\partial \mathbf{X}}\|\mathbf{X}\|_{\mathrm{F}}^2 = 2\mathbf{X}.$$

- Using this property, the derivative of the squared-error is

$$\frac{\partial}{\partial \mathbf{X}}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_{\mathrm{F}}^2 = -2\mathbf{A}^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{A}\mathbf{X}\right),$$

$$\frac{\partial}{\partial \mathbf{A}}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_{\mathrm{F}}^2 = -2\left(\mathbf{Y}-\mathbf{A}\mathbf{X}\right)\mathbf{X}^{\mathrm{T}}.$$

Least squares solution.

- Thus, if no non-negative constrains are imposed,

$$0 = -2\mathbf{A}^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{A}\mathbf{X}\right),$$
$$0 = -2\left(\mathbf{Y} - \mathbf{A}\mathbf{X}\right)\mathbf{X}^{\mathrm{T}}.$$

  then, the solution simply becomes:

$$\mathbf{X} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Y} = \mathbf{A}^{\dagger}\mathbf{Y},$$
$$\mathbf{A} = \mathbf{Y}\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\mathbf{X}^{T}\right)^{-1} = \mathbf{Y}\mathbf{X}^{\dagger}.$$

- However, when A and X are required to be non-negative, the algorithm becomes a bit complicated…

Three approaches to non-negativity.

1. Alternate least squares (ALS) method.
2. Karush-Kuhn-Tucker (KKT) method.
3. Auxiliary variable method.

Solution 1: Alternating Least squares (ALS) solution.

- A least-squares solution with no non-negativity constraints:

$$\mathbf{X} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Y} = \mathbf{A}^{\dagger}\mathbf{Y},$$

$$\mathbf{A} = \mathbf{Y}\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\mathbf{X}^{T}\right)^{-1} = \mathbf{Y}\mathbf{X}^{\dagger}.$$

- The ALS algorithm chops off negative components and leaves only non-negative components

$$\mathbf{X} \leftarrow \left[\mathbf{A}^{\dagger}\mathbf{Y}\right]_{+},$$

$$\mathbf{A} \leftarrow \left[\mathbf{Y}\mathbf{X}^{\dagger}\right]_{+}.$$

and iterate until convergence.

Solution 2: KKT conditions of constrained optimization (1/5).

- Minimize

$$\frac{1}{2}\left\|\mathbf{Y}-\mathbf{A}\mathbf{X}\right\|_{\mathrm{F}}^{2}$$

subject to

$$\mathbf{A}\geq 0,\ \mathbf{X}\geq 0$$

- Lagrange multiplier method: an augmented Lagrangian

$$J\left(\mathbf{A},\mathbf{X};\boldsymbol{\Lambda},\boldsymbol{\Xi}\right)=\frac{1}{2}\left\|\mathbf{Y}-\mathbf{A}\mathbf{X}\right\|_{\mathrm{F}}^{2}-\mathrm{tr}\left(\boldsymbol{\Lambda}^{\mathrm{T}}\mathbf{A}\right)-\mathrm{tr}\left(\boldsymbol{\Xi}^{\mathrm{T}}\mathbf{X}\right)$$

with non-negative Lagrange multiplier

$$\boldsymbol{\Lambda}\geq 0,\ \boldsymbol{\Xi}\geq 0$$

Digression: Karush-Kuhn-Tucker (KKT) condition (2/5).

- Constrained optimization problem:

$$\text{Minimize } f(x) \text{ subject to } g(x) \leq 0.$$

- Lagrange multiplier method: an augmented Lagrangian

$$J(x;\lambda) = f(x) - \lambda g(x)$$

- KKT condition for optimality:

$$0 = \frac{\partial}{\partial x} J(x;\lambda) = \frac{\partial}{\partial x} f(x) - \lambda \frac{\partial}{\partial x} g(x)$$

$$0 \leq \lambda$$

$$0 \geq g(x)$$

$$0 = \lambda \cdot g(x)$$

Solution 2: Multiplicative learning algorithm (3/5).

- The KKT condition

$$J\left(\mathbf{A},\mathbf{X};\mathbf{\Lambda},\mathbf{\Xi}\right)=\frac{1}{2}\left\|\mathbf{Y}-\mathbf{AX}\right\|_{\mathrm{F}}^{2}-\mathrm{tr}\left(\mathbf{\Lambda}^{\mathrm{T}}\mathbf{A}\right)-\mathrm{tr}\left(\mathbf{\Xi}^{\mathrm{T}}\mathbf{X}\right)$$

$$\frac{\partial J}{\partial \mathbf{A}}=-\left(\mathbf{Y}-\mathbf{AX}\right)\mathbf{X}^{\mathrm{T}}-\mathbf{\Lambda}=0, \qquad \frac{\partial J}{\partial \mathbf{X}}=-\mathbf{A}^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{AX}\right)-\mathbf{\Xi}=\mathbf{0},$$

$$\mathbf{A}\geq\mathbf{0} \qquad \mathbf{X}\geq\mathbf{0},$$

$$\mathbf{\Lambda}\geq\mathbf{0} \qquad \mathbf{\Xi}\geq\mathbf{0},$$

$$\mathbf{A}\circledast\mathbf{\Lambda}=\mathbf{0}. \qquad \mathbf{X}\circledast\mathbf{\Xi}=\mathbf{0}.$$

$$\mathbf{A}\circledast\left(\mathbf{AXX}^{\mathrm{T}}-\mathbf{YX}^{\mathrm{T}}\right)=\mathbf{0},$$

$$\mathbf{X}\circledast\left(\mathbf{A}^{\mathrm{T}}\mathbf{AX}-\mathbf{A}^{\mathrm{T}}\mathbf{X}\right)=\mathbf{0}.$$

Solution 2: Multiplicative learning algorithm (4/5).

- A multiplicative learning algorithm

$$\mathbf{A} \circledast \mathbf{A}\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{A} \circledast \mathbf{Y}\mathbf{X}^{\mathrm{T}},$$
$$\mathbf{X} \circledast \mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{X} = \mathbf{X} \circledast \mathbf{A}^{\mathrm{T}}\mathbf{X}.$$

  or elementwise

$$a_{nk}\left(\mathbf{A}\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)_{nk} = a_{nk}\left(\mathbf{Y}\mathbf{X}^{\mathrm{T}}\right)_{nk},$$
$$x_{km}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{X}\right)_{km} = x_{km}\left(\mathbf{A}^{\mathrm{T}}\mathbf{X}\right)_{km}.$$

- To solve the simultaneous equations, a fixed-point method is used as

$$a_{nk} \leftarrow a_{nk}\frac{\left(\mathbf{Y}\mathbf{X}^{\mathrm{T}}\right)_{nk}}{\left(\mathbf{A}\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)_{nk}},\quad x_{km} \leftarrow x_{km}\frac{\left(\mathbf{A}^{\mathrm{T}}\mathbf{X}\right)_{km}}{\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{X}\right)_{km}}.$$

Solution 2: Multiplicative learning algorithm (5/5).

- A multiplicative learning algorithm (elementwise form):

$$a_{nk} \leftarrow a_{nk} \frac{\left(\mathbf{Y}\mathbf{X}^{\mathrm{T}}\right)_{nk}}{\left(\mathbf{A}\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)_{nk}}, \quad x_{km} \leftarrow x_{km} \frac{\left(\mathbf{A}^{\mathrm{T}}\mathbf{X}\right)_{km}}{\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{X}\right)_{km}}.$$

or in matrix form:

$$\mathbf{A} \leftarrow \mathbf{A} \circledast \left(\mathbf{Y}\mathbf{X}^{\mathrm{T}}\right) \oslash \left(\mathbf{A}\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)$$
$$\mathbf{X} \leftarrow \mathbf{X} \circledast \left(\mathbf{A}^{\mathrm{T}}\mathbf{X}\right) \oslash \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{X}\right)$$

where $\circledast$ and $\oslash$ stand for elementwise multiplication and division, respectively.

$$\left(\mathbf{A} \circledast \mathbf{B}\right)_{ij} = a_{ij} b_{ij}$$

$$\left(\mathbf{A} \oslash \mathbf{B}\right)_{ij} = a_{ij} \Big/ b_{ij}$$

Solution 3: Auxiliary function method.

- Definition: a function $g(x, \lambda)$ is called an auxiliary function of $f(x)$ if

$$f\left(x\right)=\min_{\lambda} g\left(x,\lambda\right)$$

- A local extrema of $f(\theta)$ can be found by iteratively updating θ and λ:

$$\lambda^{(n+1)}=\arg\min_{\lambda} g\left(x^{(n)},\lambda\right),$$
$$x^{(n+1)}=\arg\min_{x} g\left(x,\lambda^{(n+1)}\right).$$

- Proof:

$$f\left(x^{(n)}\right)=g\left(x^{(n)},\lambda^{(n+1)}\right)\geq g\left(x^{(n+1)},\lambda^{(n+1)}\right)\geq\min_{\lambda} g\left(x^{(n+1)},\lambda\right)=f\left(x^{(n+1)}\right)$$

Solution 3: Auxiliary function method.

- Jensen's inequality for a convex function:

$$f\left(\sum_{i=1}^{N}\lambda_i x_i\right)\le\sum_{i=1}^{N}\lambda_i f\left(x_i\right),\quad \sum_{i=1}^{N}\lambda_i=1,\ \lambda_i\ge 0.$$

The l.h.s. is a nonlinear function of the sum (therefore {xi} are coupled); the r.h.s is a sum of nonlinear functions (therefore {xi} are decoupled).

- A choice of auxiliary function may be constructed by:

$$f\left(\sum_{i=1}^{N}x_i\right)=f\left(\sum_{i=1}^{N}\lambda_i\frac{x_i}{\lambda_i}\right)\le\sum_{i=1}^{N}\lambda_i f\left(\frac{x_i}{\lambda_i}\right)=g\left(\{x_i\},\{\lambda_i\}\right)$$

Solution 3: Auxiliary function method.

- The original cost function:

$$f(\mathbf{A},\mathbf{X}) = \frac{1}{2}\left\|\mathbf{Y}-\mathbf{A}\mathbf{X}\right\|_{\mathrm{F}}^{2} = \frac{1}{2}\sum_{m,n}\left(y_{mn}-\sum_{k}a_{mk}x_{kn}\right)^{2}$$

$$= \frac{1}{2}\sum_{m,n}y_{mn}^{2} - \sum_{m,n,k}y_{mn}a_{nk}x_{kn} + \frac{1}{2}\sum_{m,n}\left(\sum_{k}a_{mk}x_{kn}\right)^{2}$$

- Using Jensen's inequality:

$$\left(\sum_{k}a_{mk}x_{kn}\right)^{2} = \left(\sum_{k}\lambda_{mnk}\frac{a_{mk}x_{kn}}{\lambda_{mnk}}\right)^{2} \le \sum_{k}\lambda_{mnk}\left(\frac{a_{mk}x_{kn}}{\lambda_{mnk}}\right)^{2} = \sum_{k}\frac{a_{mk}^{2}x_{kn}^{2}}{\lambda_{mnk}}$$

An auxiliary function:

$$g\left(\mathbf{A},\mathbf{X};\{\lambda\}\right) = \frac{1}{2}\sum_{m,n}y_{mn}^{2} - \sum_{m,n,k}y_{mn}a_{nk}x_{kn} + \frac{1}{2}\sum_{m,n}\frac{a_{mk}^{2}x_{kn}^{2}}{\lambda_{mnk}}$$

Solution 3: Auxiliary function method.

- Auxiliary function:

$$g\left(\mathbf{A},\mathbf{X};\{\lambda\}\right)=\frac{1}{2}\sum_{m,n}y_{mn}^{2}-\sum_{m,n,k}y_{mn}a_{nk}x_{kn}+\frac{1}{2}\sum_{m,n}\frac{a_{mk}^{2}x_{kn}^{2}}{\lambda_{mnk}}$$

- For the auxiliary variables, with the normalization constraint,

$$\lambda_{mnk}=\frac{a_{mk}x_{kn}}{\overline{y}_{mn}}$$

- By taking the derivatives of matrix components:

$$\sum_{n}y_{mn}x_{kn}=\sum_{n}\frac{a_{mk}x_{kn}^{2}}{\lambda_{mnk}}$$

$$\sum_{m}y_{mn}a_{mk}=\sum_{n}\frac{a_{mk}^{2}x_{kn}}{\lambda_{mnk}}$$

Solution 3: Auxiliary function method.

- Eliminating the auxiliary variables,

$$a_{mk}\sum_n \bar{y}_{mn}x_{kn} = a_{mk}\sum_n y_{mn}x_{kn}$$

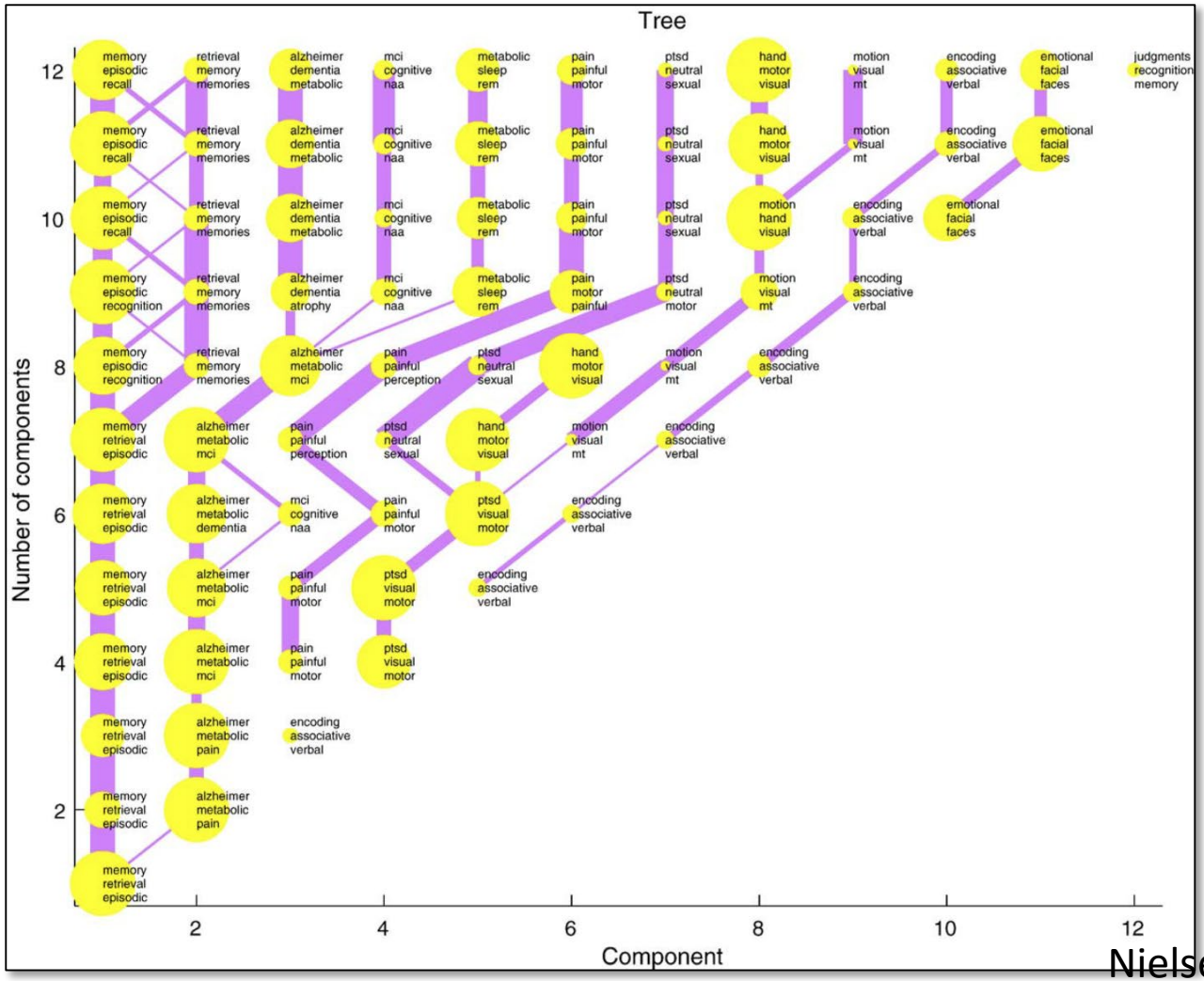
$$x_{kn}\sum_m \bar{y}_{mn}a_{mk} = x_{kn}\sum_m y_{mn}x_{mk}$$

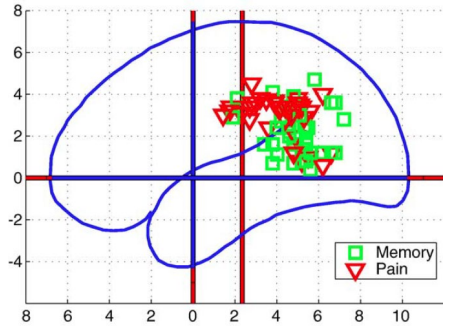
- Then, the multiplicative update rule is:

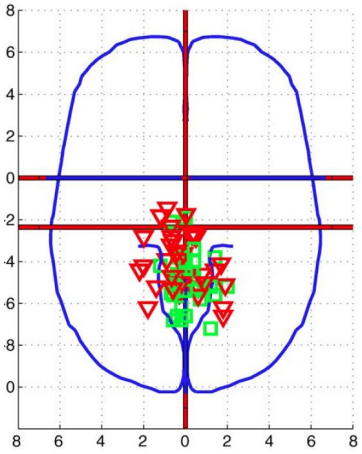
$$a_{mk} \leftarrow a_{mk}\frac{\sum_n y_{mn}x_{kn}}{\sum_n \bar{y}_{mn}x_{kn}}, \quad x_{kn} \leftarrow x_{kn}\frac{\sum_m y_{mn}x_{mk}}{\sum_m \bar{y}_{mn}a_{mk}}$$

NMF Example: Text mining of neuroimaging literature.

- 272 papers on posterior cingulate cortex (PCC); NMF automatically finds the topics related to PCC, such as memory and pain.



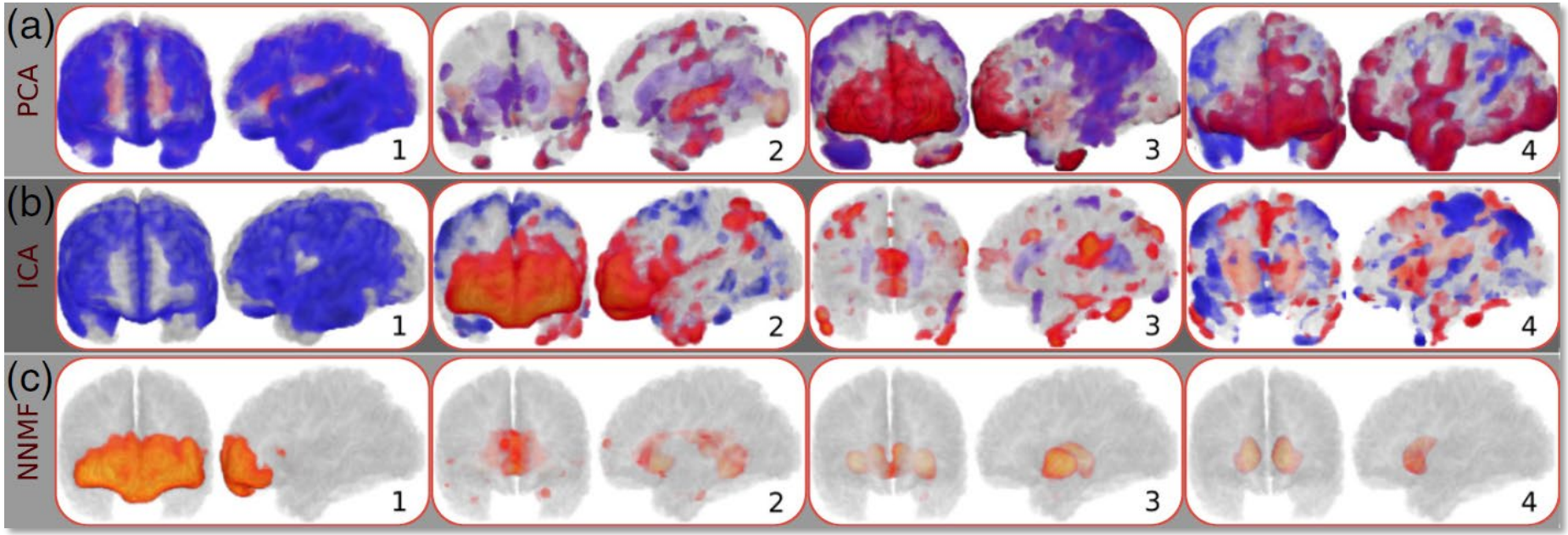




Nielsen et al. (2005) NeuroImage

NMF Example: Structural MRI.

- Comparison of PCA, ICA, and NNMF applied to structural MRI data.
- The results of PCA and ICA are non-sparse and difficult to interpret; the results of NNMF is sparse and easy to interpret.



Sotiras et al. (2015) NeuroImage

Topics to be covered today.

1. Entropy – it's origin in physics and information theory.
2. Markov process, H-theorem and renormalization group.
3. Random variable theorem and central limit theorem.
4. Independent component analysis as inverse of CLT.
5. Matrix decompositions and non-negative matrix factorization.
6. Tensor decompositions: PARAFAC and Tucker decompositions.
7. Task-related component analysis and its extensions.

Rank-one matrix and tensor decompositions

- A rank-R $I \times J$ matrix X may be decomposed into a sum of rank-one matrices as:

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{a}_r \mathbf{b}_r^{\mathrm{T}} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r .$$

  Here $\mathbf{a}_r \in \mathbb{R}^I, \mathbf{b}_r \in \mathbb{R}^J \ (r = 1, \cdots, R)$

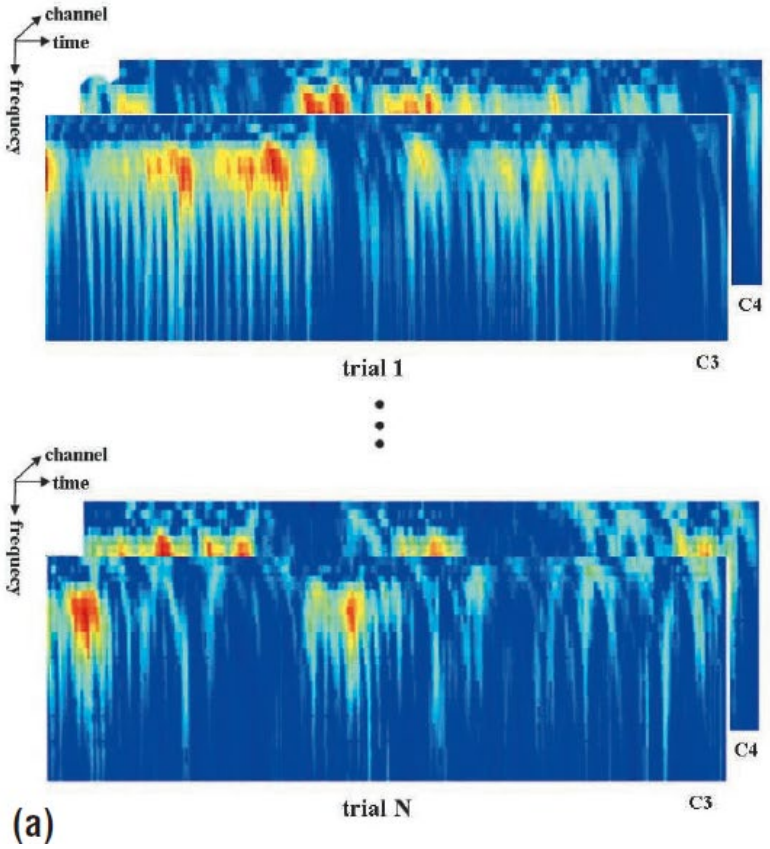
$$x_{ij} = \sum_{r=1}^{R} a_{ir} b_{ir}$$

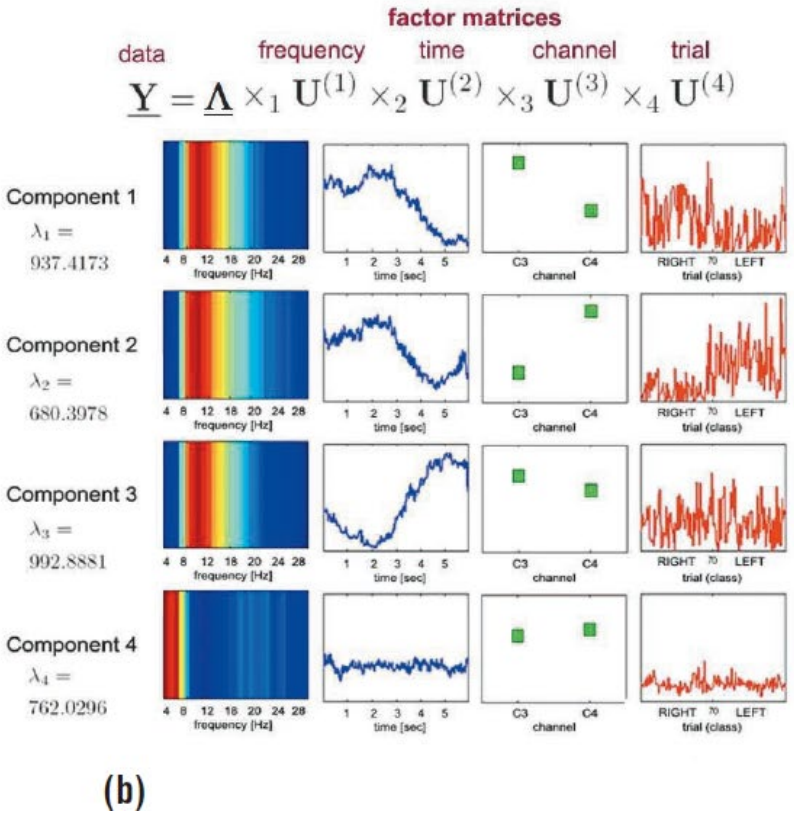
- Similarly, a three-way tensor may be decomposed into a sum of rank-one tensors as:

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r .$$

Tensors as a natural extension of vectors and matrices.

- Neuroimaging data often consist of multi-dimensional arrays more than two-dimensions.
- EEG: frequency, time, channels, trials, subjects, conditions, …





Cichocki et al. (2008) IEEE Computer

Tensors as a natural extension of vectors and matrices.

- Vector:

$$\mathbf{v} = \begin{pmatrix} v_1 & \cdots & v_i & \cdots & v_I \end{pmatrix}^{\mathrm{T}} \in \mathbb{R}^I,$$

- Matrix:

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1J} \\ \vdots & \ddots & \vdots & & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{iJ} \\ \vdots & & \vdots & \ddots & \vdots \\ a_{I1} & \cdots & a_{Ii} & \cdots & a_{IJ} \end{pmatrix}^{\mathrm{T}} \in \mathbb{R}^{I \times J},$$

- N-way Tensor:

$$\underline{\mathbf{X}} = \left(x_{i_1 i_2 \cdots i_N} \right) \in \mathbb{R}^{I_1 \times I_2 \cdots \times I_N}.$$

Mode-*n* tensor-matrix multiplication.

- *N*-way Tensor:

$$\underline{\mathbf{X}} = \left(x_{i_1 i_2 \cdots i_N} \right) \in \mathbb{R}^{I_1 \times I_2 \cdots \times I_N}.$$

- $I_n \times J_n$ matrix

$$\mathbf{A} = \left(a_{i_n j_n} \right) \in \mathbb{R}^{I_n \times J_n}.$$

- n-mode tensor-matrix multiplication:

$$\left(\underline{\mathbf{X}} \times_n \mathbf{A} \right)_{i_1 \cdots j_n \cdots i_N} = \sum_{i_n = 1}^{I_n} x_{i_1 \cdots i_n \cdots i_N} a_{i_n j_n} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times J_n \times \cdots \times I_N}.$$

Tensor approximation (1): Canonical polyadic (CP) decomposition.

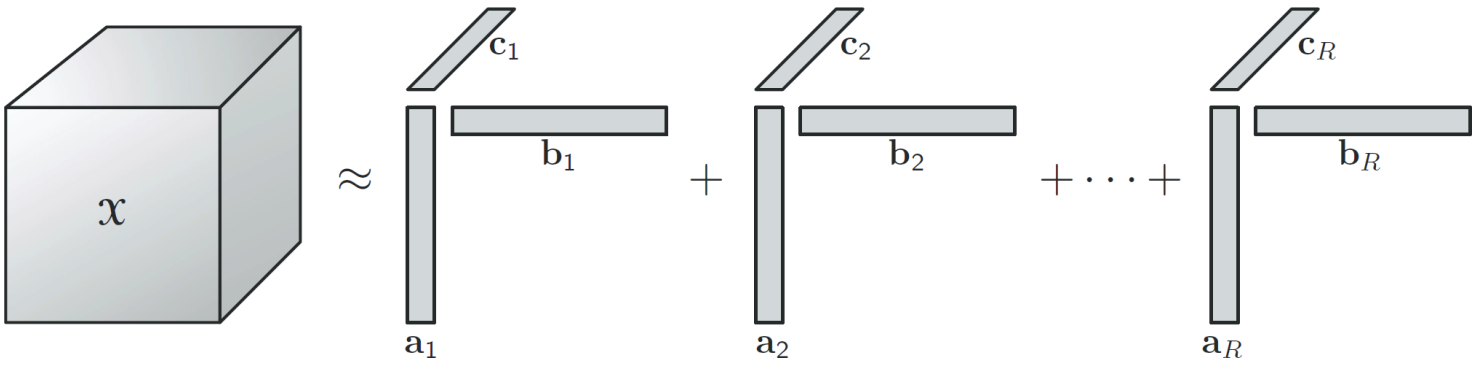
- three-way Tensor:

$$\underline{\mathbf{X}} = \left(x_{ijk} \right) \in \mathbb{R}^{I \times J \times K}.$$

- CP decomposition:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk}$$

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \underline{\mathbf{I}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$$



Tensor approximation (2): Tucker decomposition.

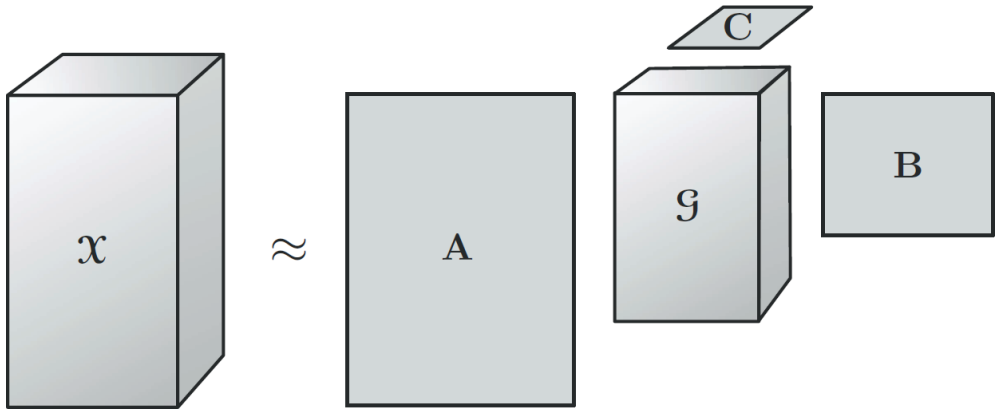
- three-way Tensor:

$$\underline{\mathbf{X}} = \left(x_{ijk} \right) \in \mathbb{R}^{I \times J \times K}.$$

- Tucker decomposition.

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_{ip} b_{jq} c_{kr} + e_{ijk}$$

$$\underline{\mathbf{X}} \approx \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = [\![\underline{\mathbf{G}}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \underline{\mathbf{G}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$$



CP decomposition is unique under mild conditions.

- 3-way Tensor case:

$$\mathbf{A} = \left(\mathbf{a}_1, \cdots, \mathbf{a}_{N_\mathbf{A}}\right), \mathbf{B} = \left(\mathbf{b}_1, \cdots, \mathbf{b}_{N_\mathbf{B}}\right), \mathbf{C} = \left(\mathbf{c}_1, \cdots, \mathbf{c}_{N_\mathbf{B}}\right).$$

$$k_\mathbf{A} = \text{rank}\left(\mathbf{A}\right), k_\mathbf{B} = \text{rank}\left(\mathbf{B}\right), k_\mathbf{C} = \text{rank}\left(\mathbf{B}\right)$$

- CP decomposition

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \underline{\mathbf{I}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$$

is unique if

$$\boxed{k_\mathbf{A} + k_\mathbf{B} + k_\mathbf{C} \geq 2R + 2}$$

Least squares fitting of CP decomposition.

- CP decomposition: Minimize the squared error:

$$\frac{1}{2}\left\|\underline{\mathbf{X}}-[\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]\right\|^2=\frac{1}{2}\sum_{i,j,k}\left(x_{ijk}-\sum_r a_{ir}b_{jr}c_{kr}\right)^2$$

- Tucker decomposition: Minimize the squared error:

$$\frac{1}{2}\left\|\underline{\mathbf{X}}-[\![\underline{\mathbf{G}};\mathbf{A},\mathbf{B},\mathbf{C}]\!]\right\|^2=\frac{1}{2}\sum_{i,j,k}\left(x_{ijk}-\sum_p\sum_q\sum_r g_{pqr}a_{ip}b_{jq}c_{kr}\right)^2$$

mode-*n* Matricization of three-way tensor.

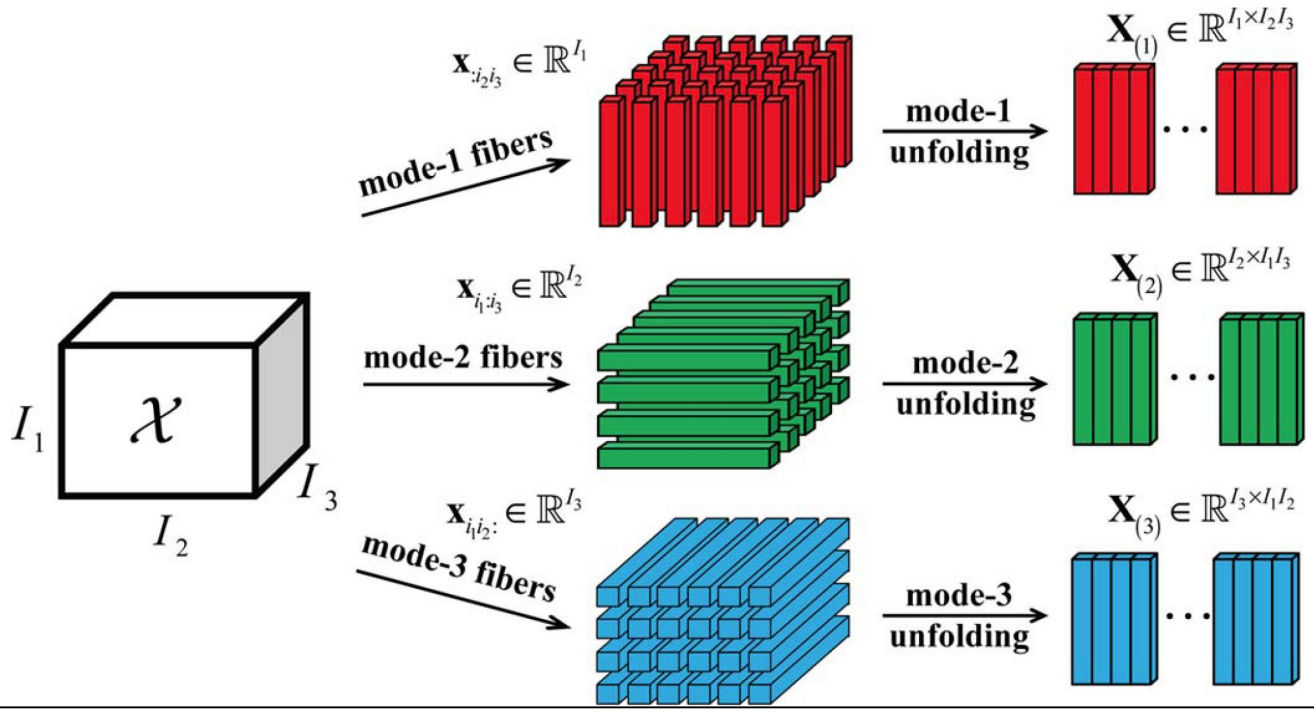
- three-way Tensor:

$$\underline{\mathbf{X}} = \left(x_{ijk} \right) \in \mathbb{R}^{I_1 \times I_2 \times I_3}.$$

$$\mathbf{X}_{(1)} = \left(x_{i(jk)} \right) \in \mathbb{R}^{I \times JK}.$$

$$\mathbf{X}_{(2)} = \left(x_{j(ik)} \right) \in \mathbb{R}^{J \times IK}.$$

$$\mathbf{X}_{(3)} = \left(x_{k(ij)} \right) \in \mathbb{R}^{K \times IJ}.$$



Least squares fitting of CP decomposition.

- CP decomposition:

$$\underline{\mathbf{X}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] + \underline{\mathbf{E}} \in \mathbb{R}^{I \times J \times K}$$

$$\mathbf{X}_{(1)} = \mathbf{A}\left(\mathbf{C} \odot \mathbf{B}\right)^{\mathrm{T}} + \mathbf{E}_{(1)} \in \mathbb{R}^{I \times JK}$$

$$\mathbf{X}_{(2)} = \mathbf{B}\left(\mathbf{C} \odot \mathbf{A}\right)^{\mathrm{T}} + \mathbf{E}_{(2)} \in \mathbb{R}^{J \times KI}$$

$$\mathbf{X}_{(3)} = \mathbf{C}\left(\mathbf{B} \odot \mathbf{A}\right)^{\mathrm{T}} + \mathbf{E}_{(3)} \in \mathbb{R}^{K \times JI}$$

- Squared-error cost function:

$$\begin{aligned} \left\| \underline{\mathbf{X}} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \right\|^2 &= \left\| \mathbf{X}_{(1)} - \mathbf{A}\left(\mathbf{C} \odot \mathbf{B}\right)^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \\ &= \left\| \mathbf{X}_{(2)} - \mathbf{B}\left(\mathbf{C} \odot \mathbf{A}\right)^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \\ &= \left\| \mathbf{X}_{(3)} - \mathbf{C}\left(\mathbf{B} \odot \mathbf{A}\right)^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \end{aligned}$$

Alternative least-squares (ALS) algorithm.

- CP decomposition:

$$\frac{\partial}{\partial \mathbf{A}}\left\|\mathbf{X}_{(1)}-\mathbf{A}\left(\mathbf{C}\odot\mathbf{B}\right)^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}=-2\left\{\mathbf{X}_{(1)}-\mathbf{A}\left(\mathbf{C}\odot\mathbf{B}\right)^{\mathrm{T}}\right\}\left(\mathbf{C}\odot\mathbf{B}\right)$$

$$=2\mathbf{A}\left(\mathbf{C}\odot\mathbf{B}\right)^{\mathrm{T}}\left(\mathbf{C}\odot\mathbf{B}\right)-2\mathbf{X}_{(1)}\left(\mathbf{C}\odot\mathbf{B}\right)$$

$$=2\mathbf{A}\left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)\circledast\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)-2\mathbf{X}_{(1)}\left(\mathbf{C}\odot\mathbf{B}\right)$$

- If there is no constraint on the matrices, then

$$\mathbf{A}=\mathbf{X}_{(1)}\left(\mathbf{C}\odot\mathbf{B}\right)\left\{\left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)\circledast\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)\right\}^{-1}$$

- When non-negativity constraints are imposed:

$$\mathbf{A}\leftarrow\left[\mathbf{X}_{(1)}\left(\mathbf{C}\odot\mathbf{B}\right)\left\{\left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)\circledast\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)\right\}^{-1}\right]_{+}$$

Alternative least-squares (ALS) algorithm.

- CP decomposition:

$$\frac{\partial}{\partial \mathbf{A}}\left\|\mathbf{X}_{(1)}-\mathbf{A}\left(\mathbf{C}\odot\mathbf{B}\right)^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}=-2\left\{\mathbf{X}_{(1)}-\mathbf{A}\left(\mathbf{C}\odot\mathbf{B}\right)^{\mathrm{T}}\right\}\left(\mathbf{C}\odot\mathbf{B}\right)$$

$$=2\mathbf{A}\left(\mathbf{C}\odot\mathbf{B}\right)^{\mathrm{T}}\left(\mathbf{C}\odot\mathbf{B}\right)-2\mathbf{X}_{(1)}\left(\mathbf{C}\odot\mathbf{B}\right)$$

$$=2\mathbf{A}\left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)\circledast\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)-2\mathbf{X}_{(1)}\left(\mathbf{C}\odot\mathbf{B}\right)$$

$$\frac{\partial}{\partial \mathbf{B}}\left\|\mathbf{X}_{(2)}-\mathbf{B}\left(\mathbf{C}\odot\mathbf{A}\right)^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}=2\mathbf{B}\left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)\circledast\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)-2\mathbf{X}_{(2)}\left(\mathbf{C}\odot\mathbf{A}\right)$$

$$\frac{\partial}{\partial \mathbf{C}}\left\|\mathbf{X}_{(3)}-\mathbf{C}\left(\mathbf{B}\odot\mathbf{A}\right)^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}=2\mathbf{C}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)\circledast\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)-2\mathbf{X}_{(3)}\left(\mathbf{B}\odot\mathbf{A}\right)$$

Alternative least-squares (ALS) algorithm.

- When non-negativity constraints are imposed:

1. Given $\mathbf{B}$ and $\mathbf{C}$, update the matrix $\mathbf{A}$:

$$\mathbf{A} \leftarrow \left[\mathbf{X}_{(1)} \left(\mathbf{C} \odot \mathbf{B} \right) \left\{ \left(\mathbf{C}^{\mathrm{T}} \mathbf{C} \right) \circledast \left(\mathbf{B}^{\mathrm{T}} \mathbf{B} \right) \right\}^{-1} \right]_{+}$$

2. Given $\mathbf{C}$ and $\mathbf{B}$, update the matrix $\mathbf{B}$:

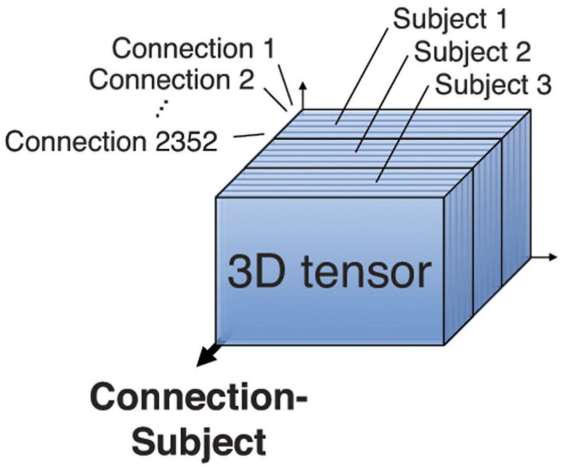
$$\mathbf{B} \leftarrow \left[\mathbf{X}_{(2)} \left(\mathbf{C} \odot \mathbf{A} \right) \left\{ \left(\mathbf{C}^{\mathrm{T}} \mathbf{C} \right) \circledast \left(\mathbf{A}^{\mathrm{T}} \mathbf{A} \right) \right\}^{-1} \right]_{+}$$

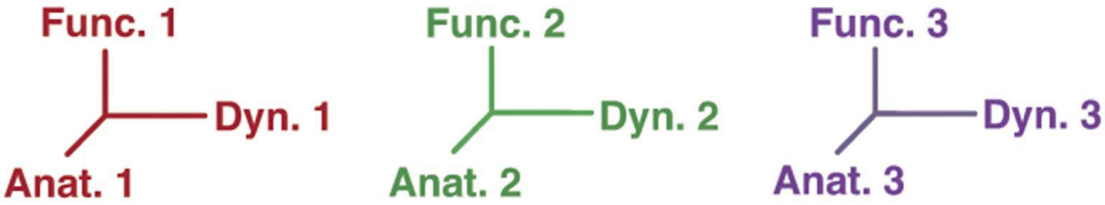
3. Given $\mathbf{A}$ and $\mathbf{B}$, update the matrix $\mathbf{A}$:

$$\mathbf{C} \leftarrow \left[\mathbf{X}_{(3)} \left(\mathbf{B} \odot \mathbf{A} \right) \left\{ \left(\mathbf{B}^{\mathrm{T}} \mathbf{B} \right) \circledast \left(\mathbf{A}^{\mathrm{T}} \mathbf{A} \right) \right\}^{-1} \right]_{+}$$

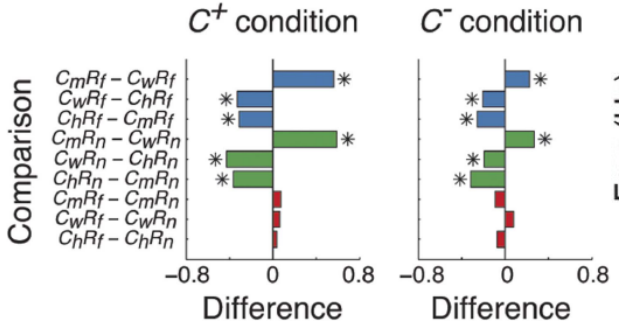
4. Repeat 1-3 until convergence.

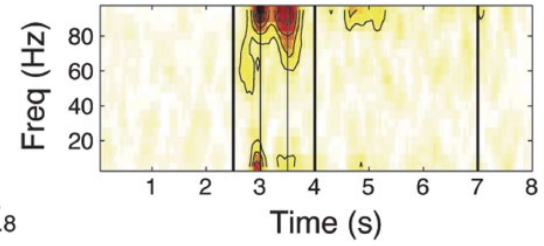
CP decomposition example: causal network analysis.

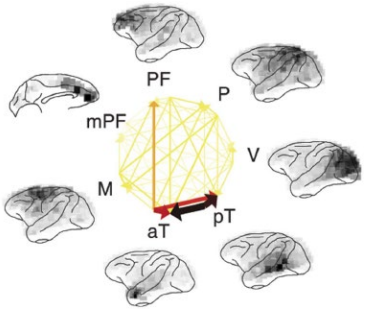




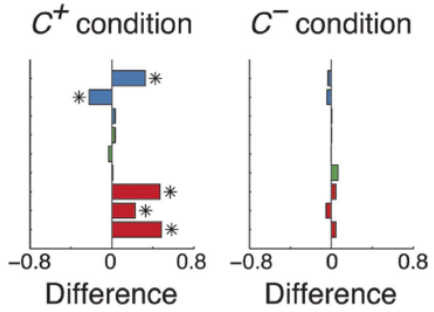
comp 1

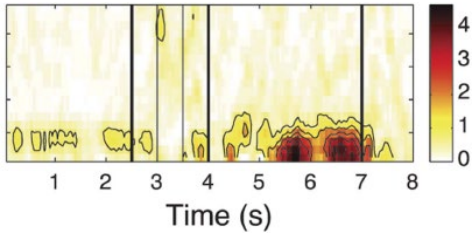


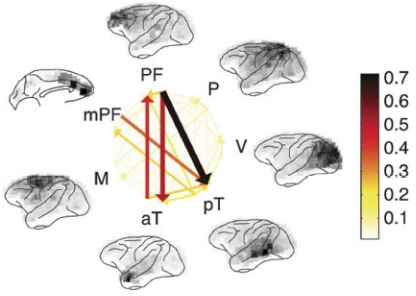




comp 5







Chao et al. (2015) eLife

Tensor ICA for group analysis.

- Neuroimaging data often has more than two indicies.
  - (1) time, (2) space, and (3) subject:

$$\underline{\mathbf{X}} = \underline{\mathbf{I}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} + \underline{\mathbf{E}}$$

- For fMRI data, spatial maps are often assumed to be mutually independent.

$$\mathbf{X}_{(2)}^{\mathrm{T}} = \left(\mathbf{C} \odot \mathbf{A}\right)\mathbf{B}^{\mathrm{T}} + \mathbf{E}_{(2)}^{\mathrm{T}}$$

- Problem: given the data tensor X, find the independent components B and the unmixing matrix $\mathbf{W}$

$$\mathbf{X}_{(2)}^{\mathrm{T}} = \mathbf{W}\mathbf{B}^{\mathrm{T}}$$

  so that the matrix W has a form of Khatri-Rao product:

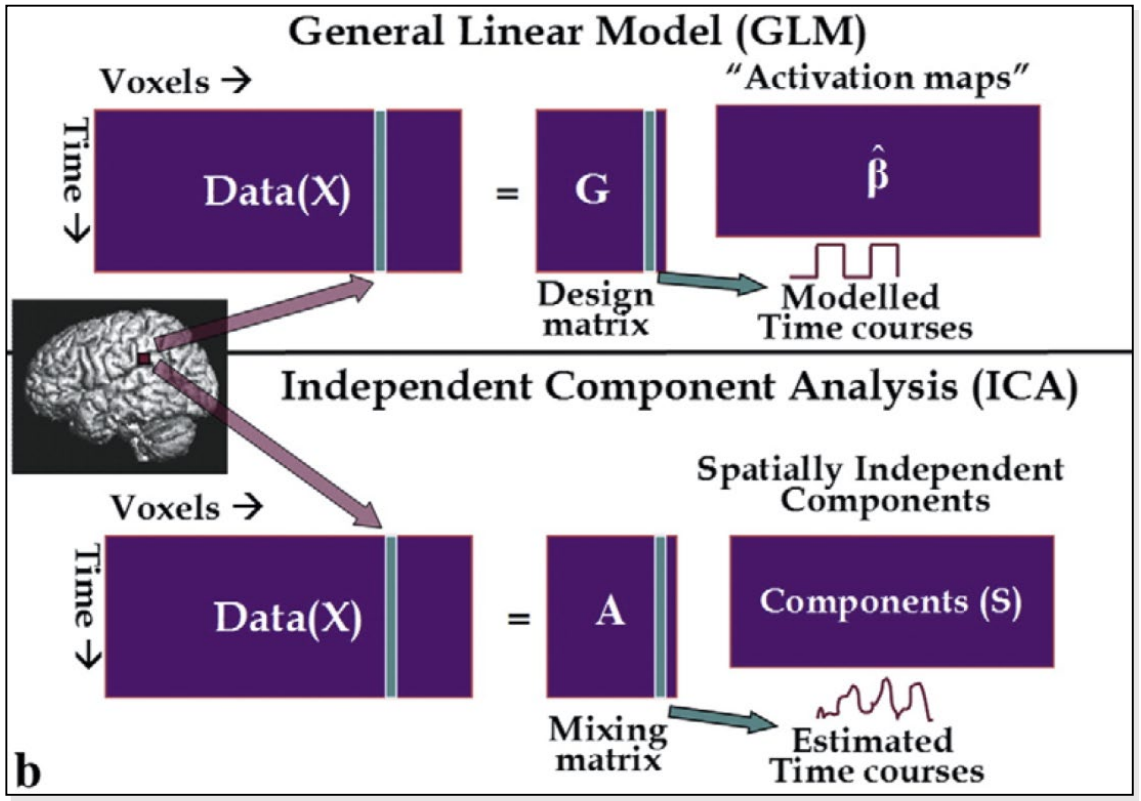
$$\mathbf{W} \approx \left(\mathbf{C} \odot \mathbf{A}\right)$$

Topics to be covered today.

1. Entropy – it's origin in physics and information theory.
2. Markov process, H-theorem and renormalization group.
3. Random variable theorem and central limit theorem.
4. Independent component analysis as inverse of CLT.
5. Matrix decompositions and non-negative matrix factorization.
6. Tensor decompositions: PARAFAC and Tucker decompositions.
7. Task-related component analysis and its extensions.

Two approaches in neuroimaging data analysis.

- Hypothesis-driven approach
  - General linear models (GLMs).
- Data-driven approach
  - Principal component analysis (PCA), independent component analysis (ICA), factor analysis (FA).



Task-related reproducibility as an criterion.

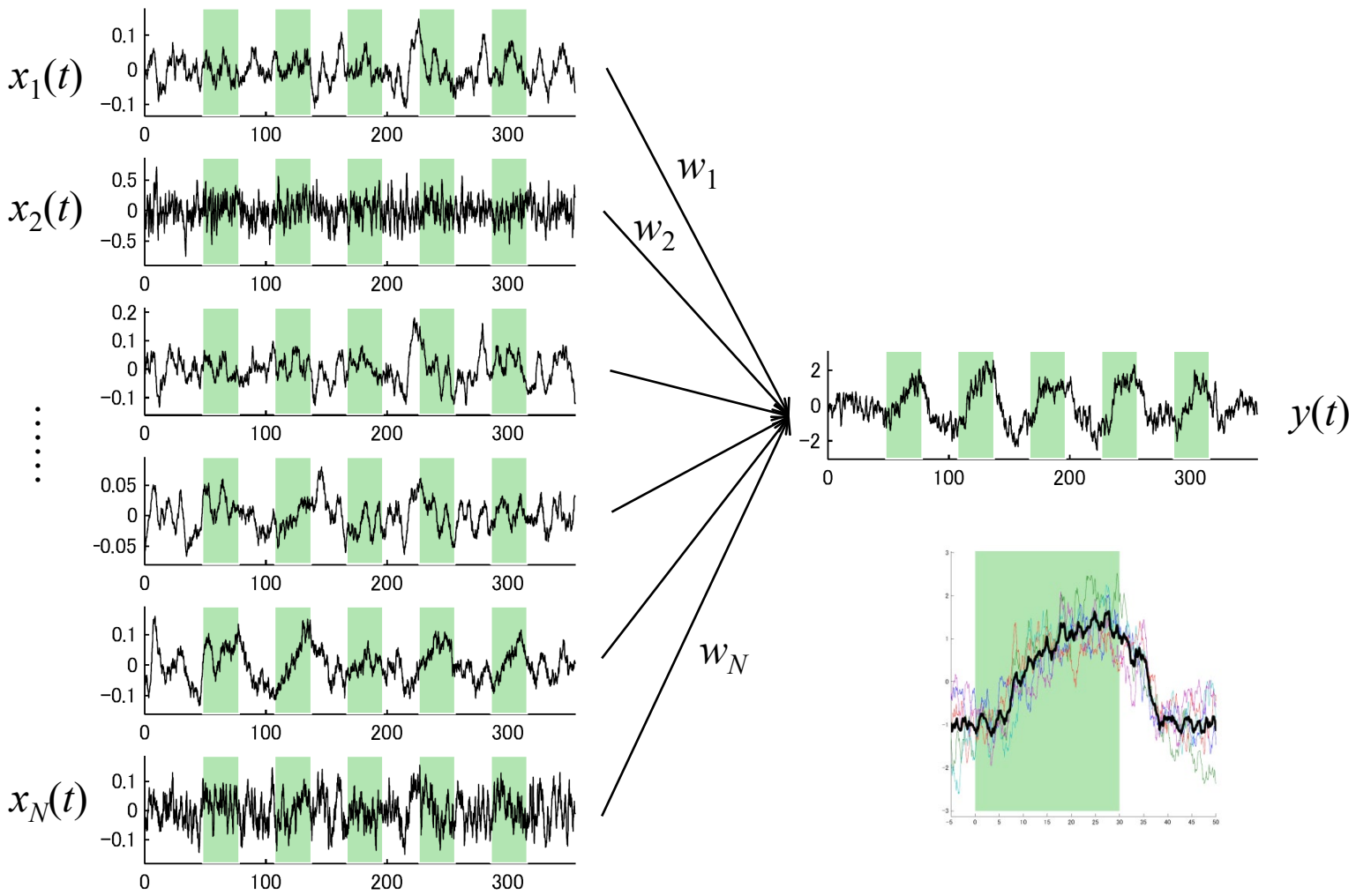
- Hypothesis-driven approach (Friston et al. 1994)
  - General linear models (GLMs).
  -- Pros: Explicit testing of a hypothesis
  "Popperian" (hypothesis generation, refutation and design)
  -- Cons: Scientifically valid hypothesis is needed; only modelled hypotheses are examined.

- Data-driven approach (Makeig et al. 1997; McKeown et al. 1998)
  - Principal component analysis (PCA), independent component analysis (ICA), factor analysis (FA).
  -- Pros: No a priori assumption about hypotheses.
  -- Cons: Expert interpretations and expertise needed.

Friston, K. J. (1998). Modes or models: a critique on independent component analysis for fMRI. *Trends in cognitive sciences*, *2*(10), 373-375.

Makeig et al. (1998). Response from McKewon, Makeig, Brown, Jung, Kindermann, Bell and Sejnowki.

Brain timing cannot be estimated from external events.



Task-related component analysis: Formulation (1/2).

- Construct a time series as a weighted sum:

$$y(t) = \sum_i w_i x_i(t) = \mathbf{w}^T \mathbf{x}(t)$$

- TRCA proposes to maximize a sum of inter-block covariances:

$$\sum_{\substack{k,l=1 \\ k \neq l}}^{K} \hat{C}_{kl} = \sum_{\substack{k,l=1 \\ k \neq l}}^{K} \mathrm{Cov}\left[y^{(k)}(t), y^{(l)}(t) \right]$$

$$= \sum_{\substack{k,l=1 \\ k \neq l}}^{K} \sum_{i,j} w_i w_j \mathrm{Cov}\left[x_i^{(k)}(t), x_j^{(l)}(t) \right] = \mathbf{w}^T \mathbf{S} \mathbf{w}$$

under the condition that the variance is constrained to one:

$$\mathrm{Var}\left[y(t) \right] = \sum_{i,j} w_i w_j \mathrm{Cov}\left[x_i(t), x_j(t) \right] = \mathbf{w}^T \mathbf{Q} \mathbf{w} = 1$$

Task-related component analysis: Formulation (2/2).

- TRCA is equivalent to maximizing the Rayleigh-Ritz quotient:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S} \mathbf{w}}{\mathbf{w}^T \mathbf{Q} \mathbf{w}}$$

- A solution of this maximization problem is given as eigenvectors of the matrix:

$$\mathbf{Q}^{-1}\mathbf{S}$$

Then the solutions are:

$$\left\{\hat{\mathbf{w}}_1, \cdots, \hat{\mathbf{w}}_N\right\}$$

$$\left\{\lambda_1, \cdots, \lambda_N\right\} \quad \left(\lambda_1 \geq \cdots \geq \lambda_N\right)$$

Task-related component analysis: Matlab code.

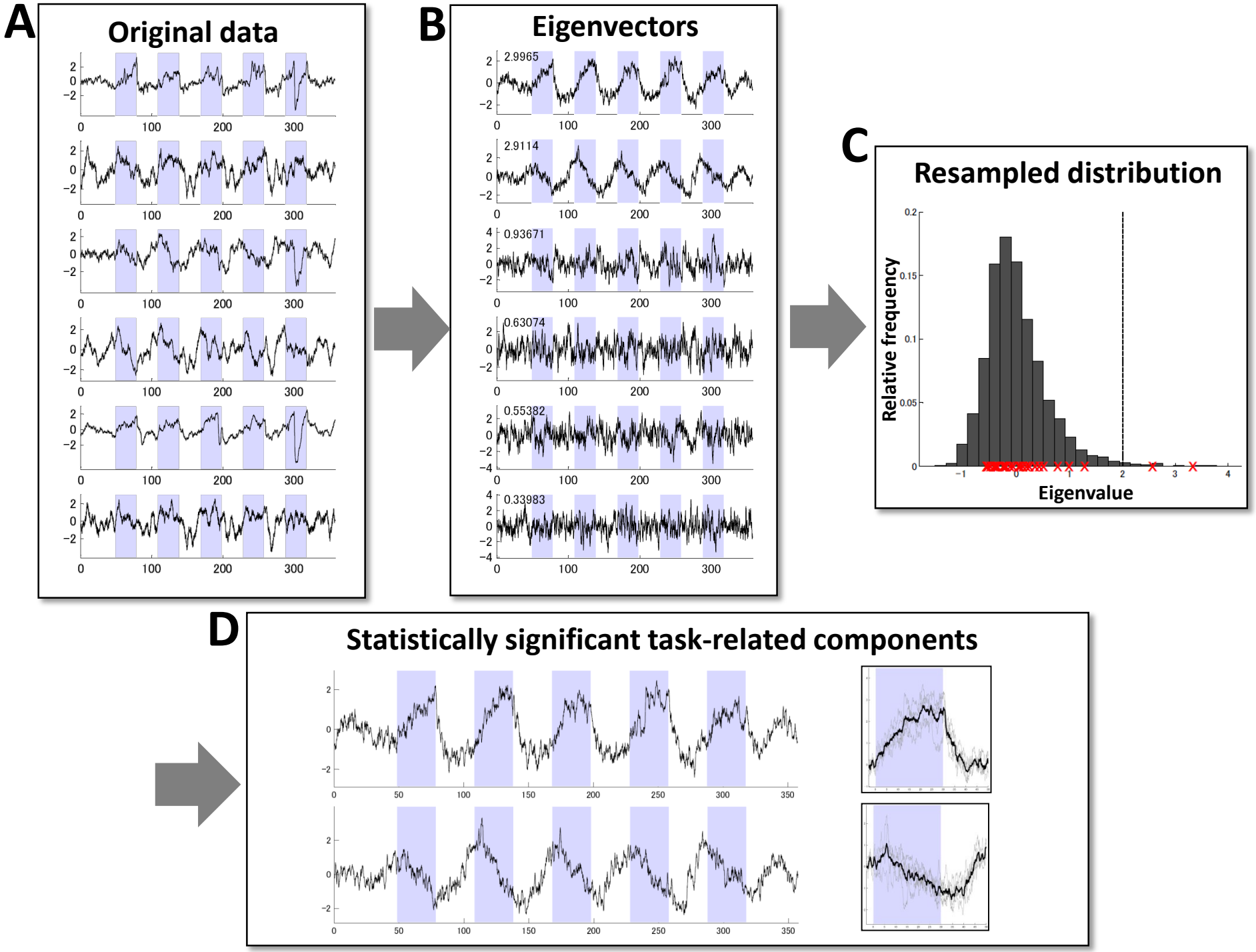
```
function [Y, V, D, S, Q] = TRCA(X, t1, Nexp)

Nchannel = size(X, 1);
Nblock = length(t1);

% data of task duration
Xb = zeros(Nchannel, Nexp+1, Nblock);
for i=1:Nblock
     Xb(:, :, i) = X(:, t1(i):t1(i)+Nexp) - repmat(mean(X(:, t1(i):t1(i)+Nexp),2),1,Nexp+1);
end

% computation of correlation matrices:
S = zeros(Nchannel);
for i=1:Nblock-1
   for j=i+1:Nblock
      S = S + Xb(:, :, i)*Xb(:, :, j)';
   end
end
S = S+S';

X = X -repmat(mean(X,2),1,size(X,2));
Q = X*X';

% TRCA eigenvalue algorithm
[V, D] = eig(Q\S);
Y = V'*X;
```

"Blocking" of continuous data

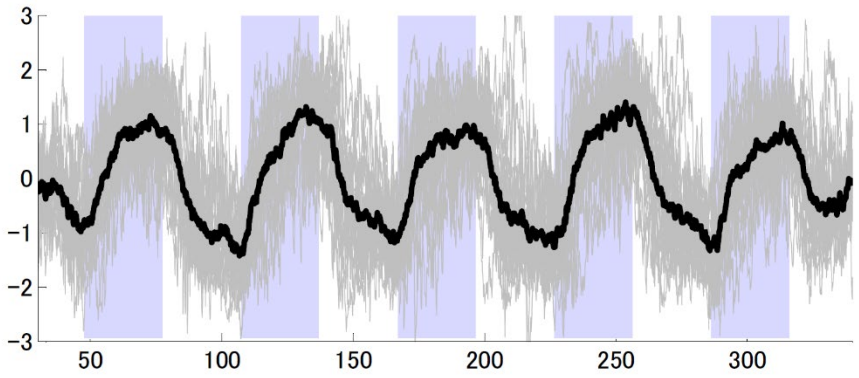
Computation of **S** matrix

Computation of **Q** matrix

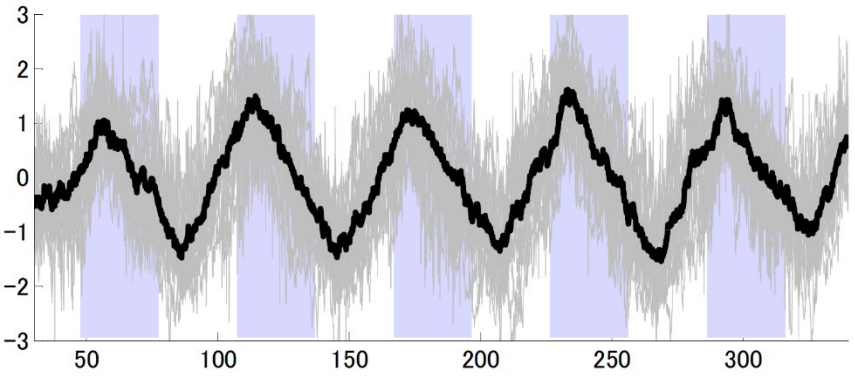
Eigenvalue computation


A
Original data
B
Eigenvectors
2.9965
2.9114
0.93671
0.63074
0.55382
0.33983
C
Resampled distribution
Relative frequency
Eigenvalue
D
Statistically significant task-related components

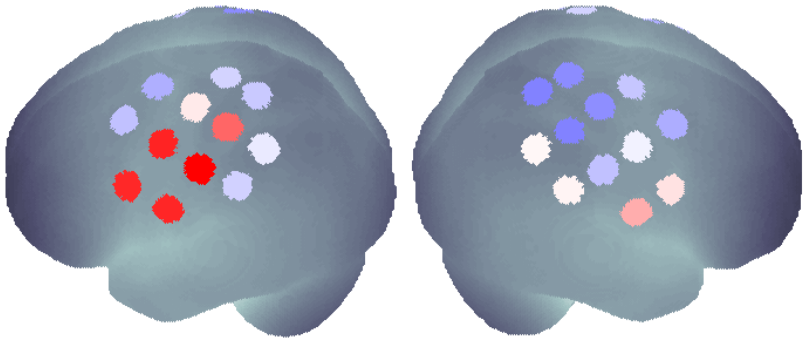
Task-related component analysis: right finger tapping.

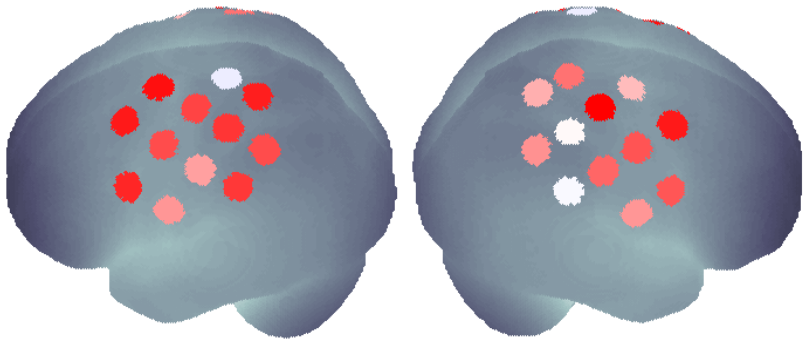


Time

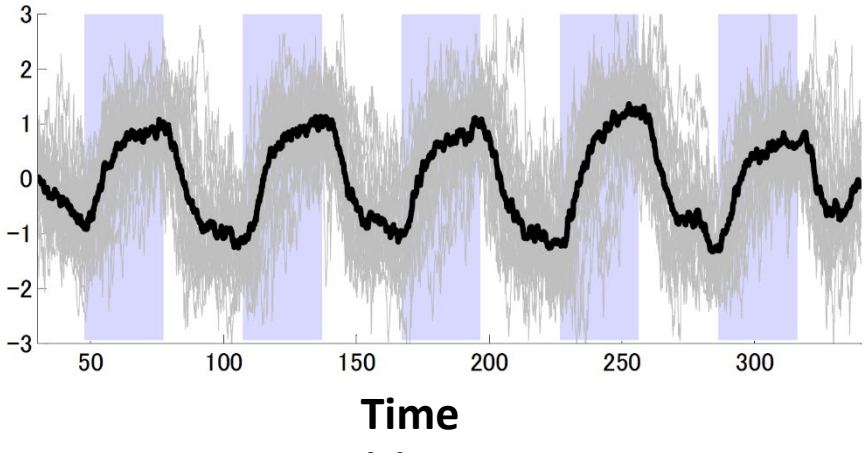


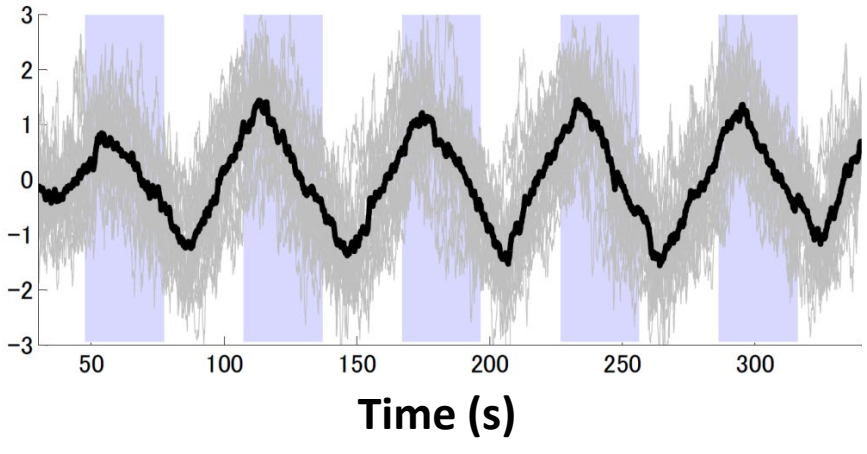
Time (s)

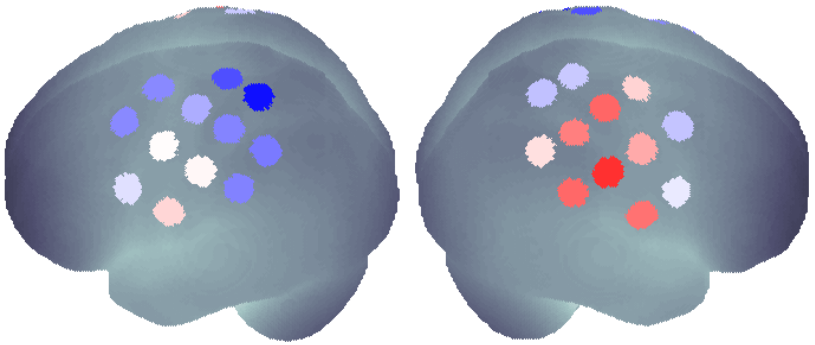


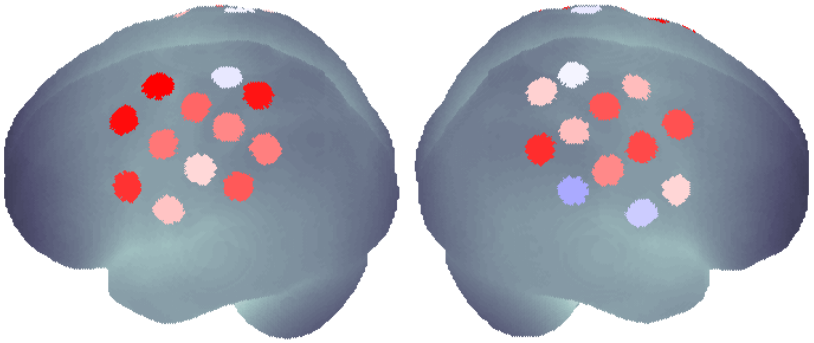


Task-related component analysis: left finger tapping.

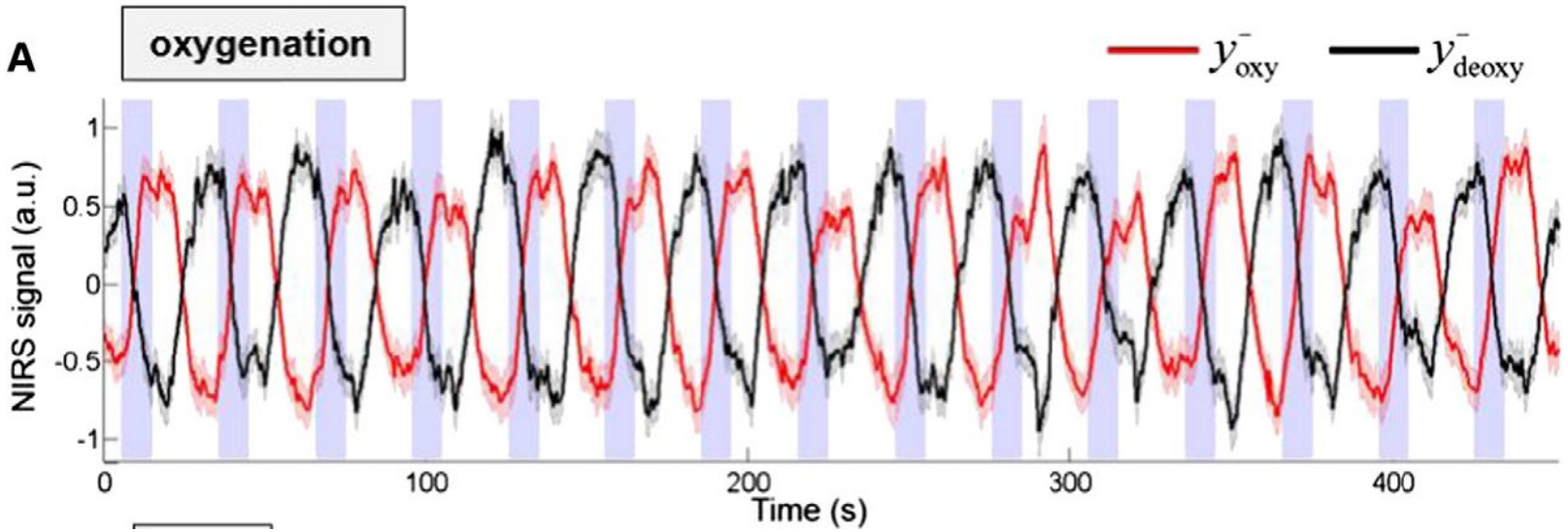


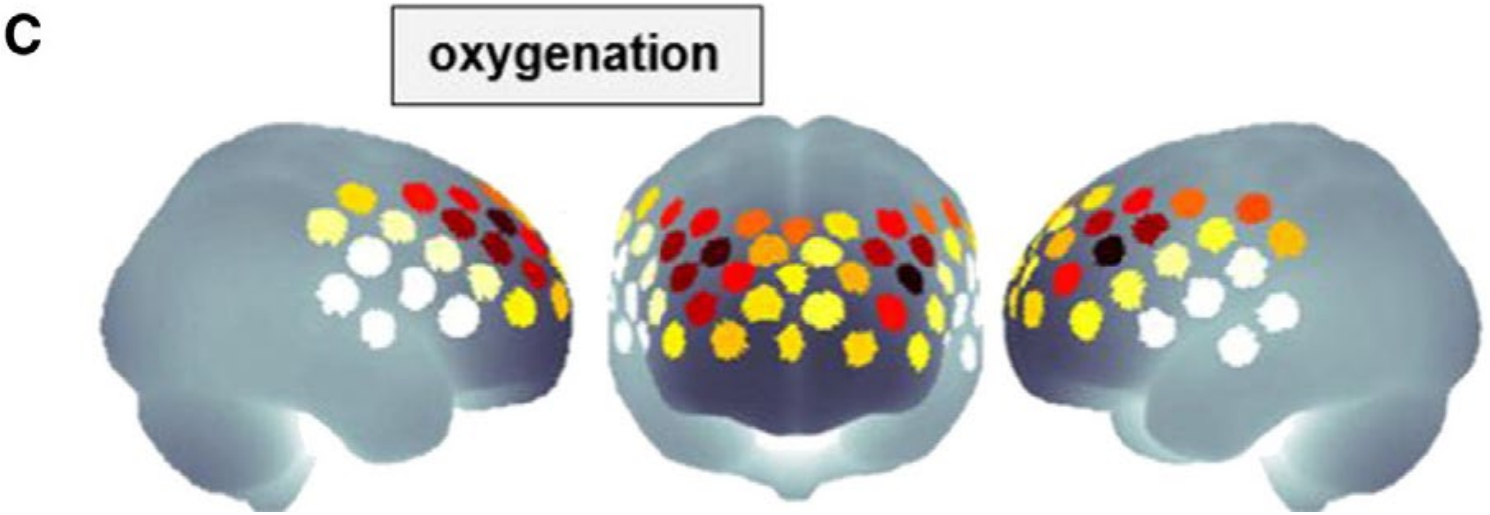




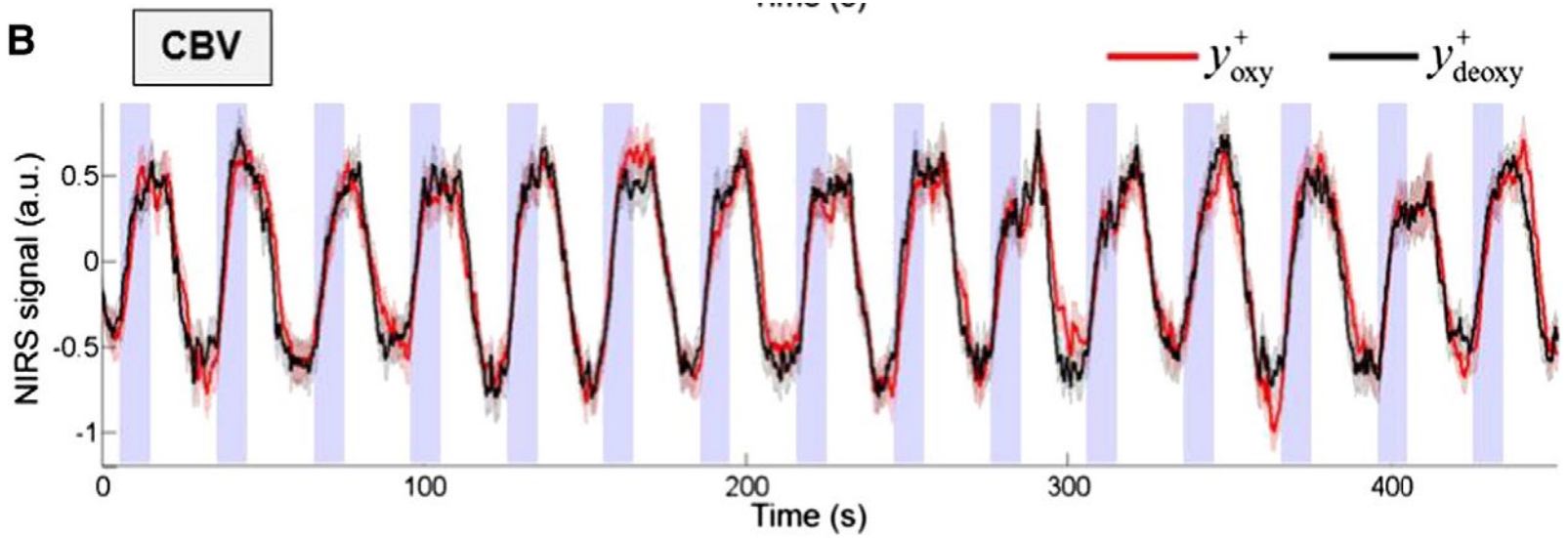


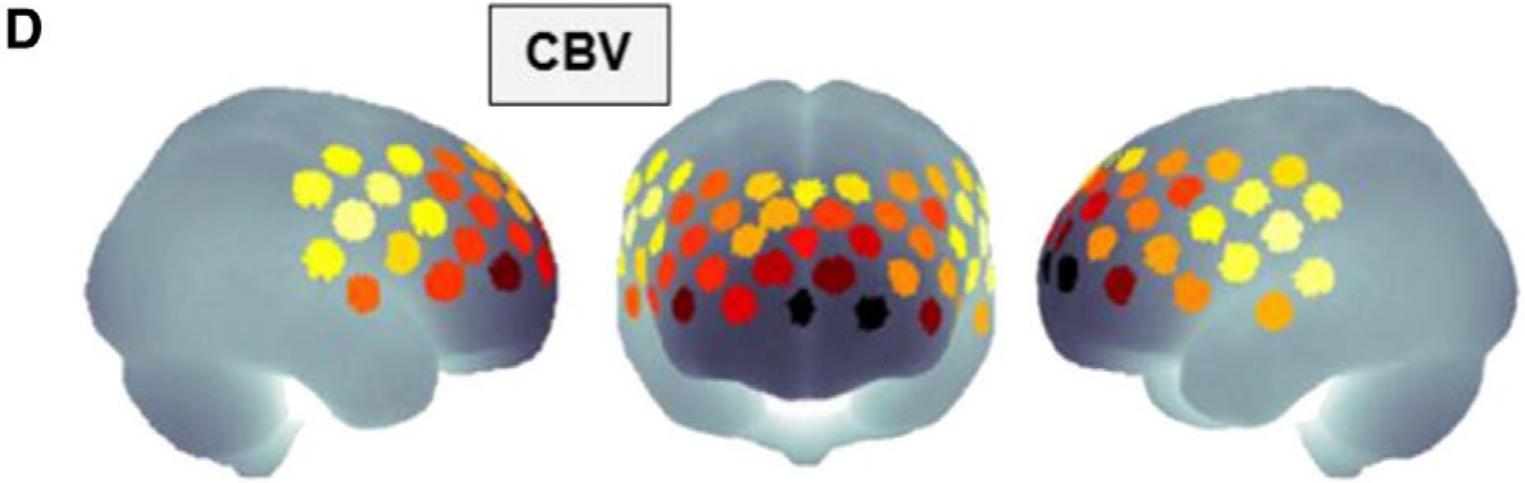
TRCA discriminates oxygenation and CBV (1/2).





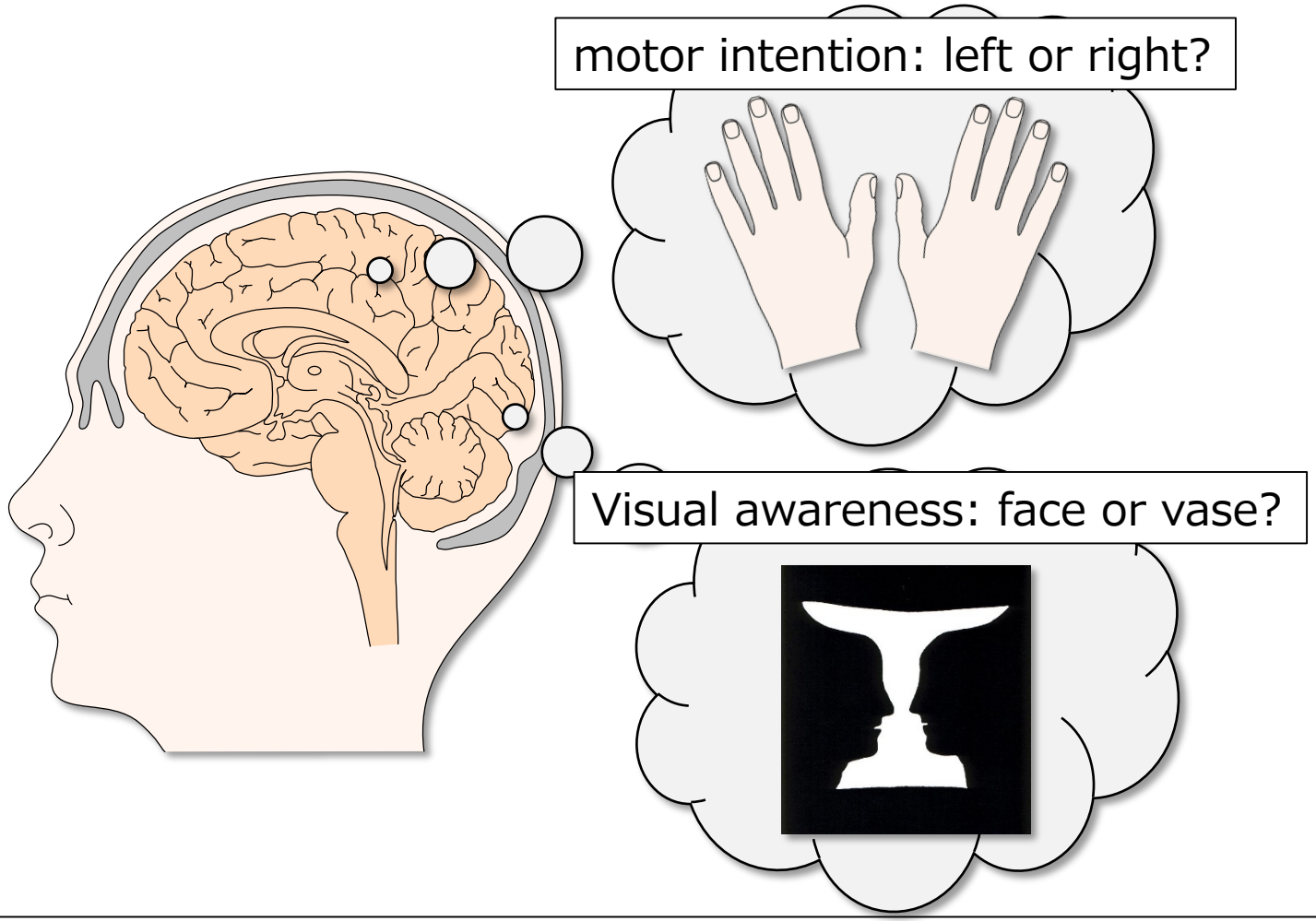
TRCA discriminates oxygenation and CBV (2/2).





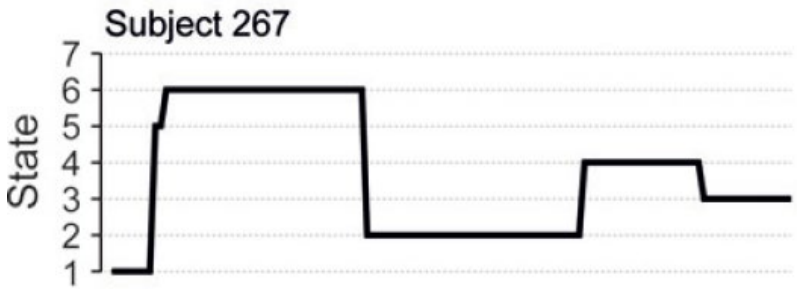
Brain timing cannot be estimated from external events.

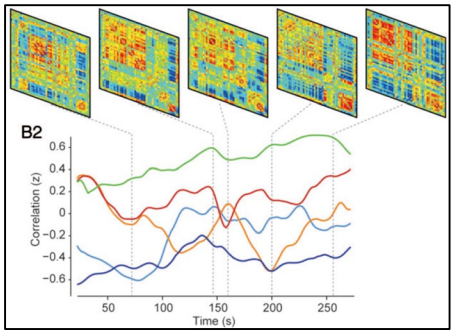
- A number of important cognitive functions (e.g. motor intention and visual awareness) are not necessarily time-locked to an external event.



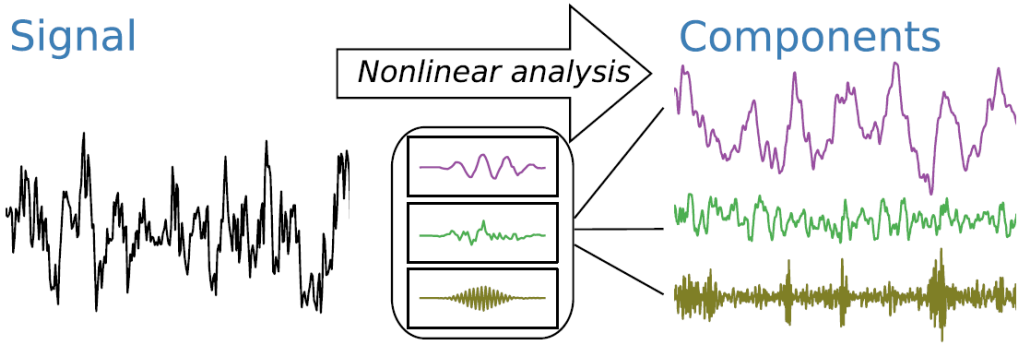
Estimation of event timings from EEG series *per se*.

- Change detection of dynamic connectivity.
  - Assumption: Brain network transits from one state to another.
  -- functional connectivity (temporal correlation) (Allen et al. 2014)
  -- phase synchronization

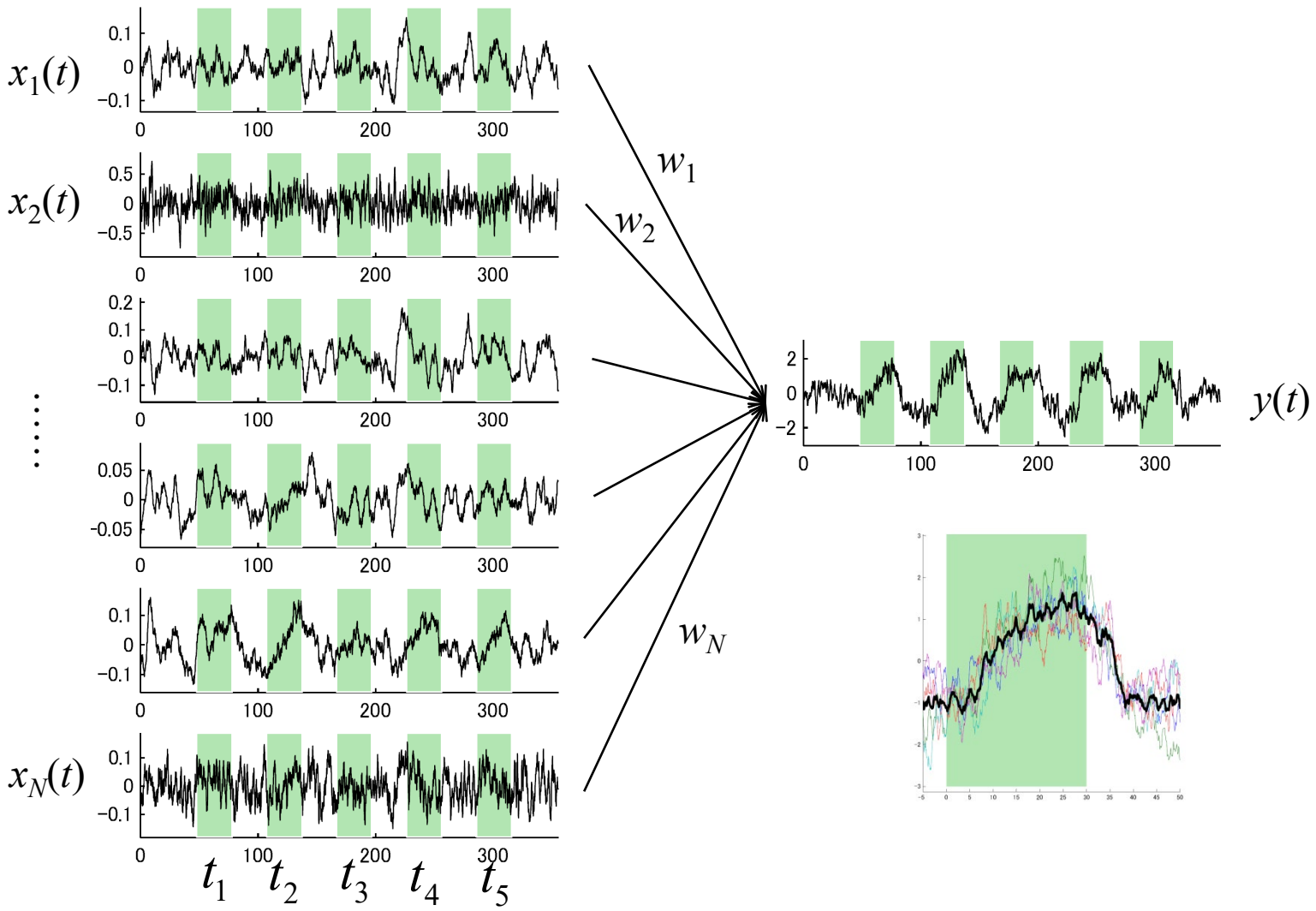




- Detection of recurring wave forms.
  - Assumption: a certain brain event is associated with same EEG patterns.
  -- dictionary learning (Brockmeier & Principe ,2016)
  -- sliding window matching (Gips et al., 2016)



Spatio-temporal filtering for finding recurrent waves.



- Not only the spatial weights $\{\mathbf{w}\}$ but also the timings $\{t_k\}$ are optimized.

Spatio-temporal filtering for finding recurring waves.

- So far, the block timings have been assumed to be known a priori. TRCA optimizes a spatial filter when block timing are provided.

$$\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{Q}\mathbf{w}}$$

- Here, the experimental timings,

$$\mathbf{t} = \begin{pmatrix} t_1 & t_2 & \cdots & t_k \end{pmatrix}^{\mathrm{T}}$$

  are also optimized:

$$\max_{\mathbf{w},\mathbf{t}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}(\mathbf{t})\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{Q}\mathbf{w}}$$

TRCA Sliding Window Matching (TRCA-SWM) algorithm.

1. Initialize the timing vector $\mathbf{t}$ by randomization or according to experimental data.
2. With the initial timing vector, find an optimal weight vector by solving

$$\mathbf{w}(\mathbf{t}) = \max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}(\mathbf{t})\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{Q}\mathbf{w}}$$

3. Perturb the timing vector and compute a new weight vector.
4. Accept the perturbed timing vector according to MCMC algorithm; otherwise reject.
5. Repeat Step 3 and Step 4 until some convergence criteria are met.

TRCA-SWM: Matlab code.

```
while iter<maxIterations
    % generate trial move:
    k = randi(K);   % randomly picked timing.
    Ttrial = T;
    Ttrial(k) = Ttrue(k) + randi([-50 50]); % perturb the picked timing.

    % evaluate the trial move:
    Jtrial = TRCAs(X, Ttrial, L);

    % accept or not:
    Pacc = exp(-(Jtrial-J)/Temp);
    if rand<Pacc
        T = Ttrial; J = Jtrial;
        disp([num2str(iter) '-th trial: move accepted, J: ' num2str(J)]);
    else
        disp([num2str(iter) '-th trial: move rejected, J: ' num2str(J)]);
    end

    % update the optimum:
    if J<J0
        iter0 = iter; J0 = J; T0 = T;
    end

    Jall(iter) = J; iter = iter + 1;
end
```

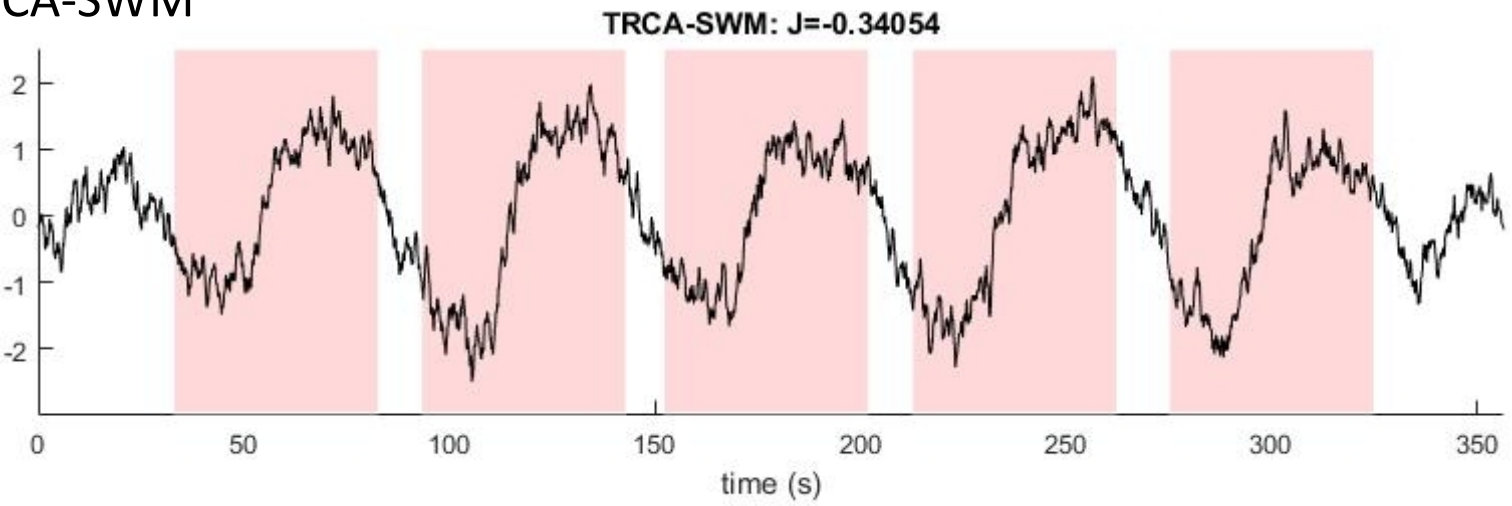
Perturb timing vector.

Accept or reject the perturbation.

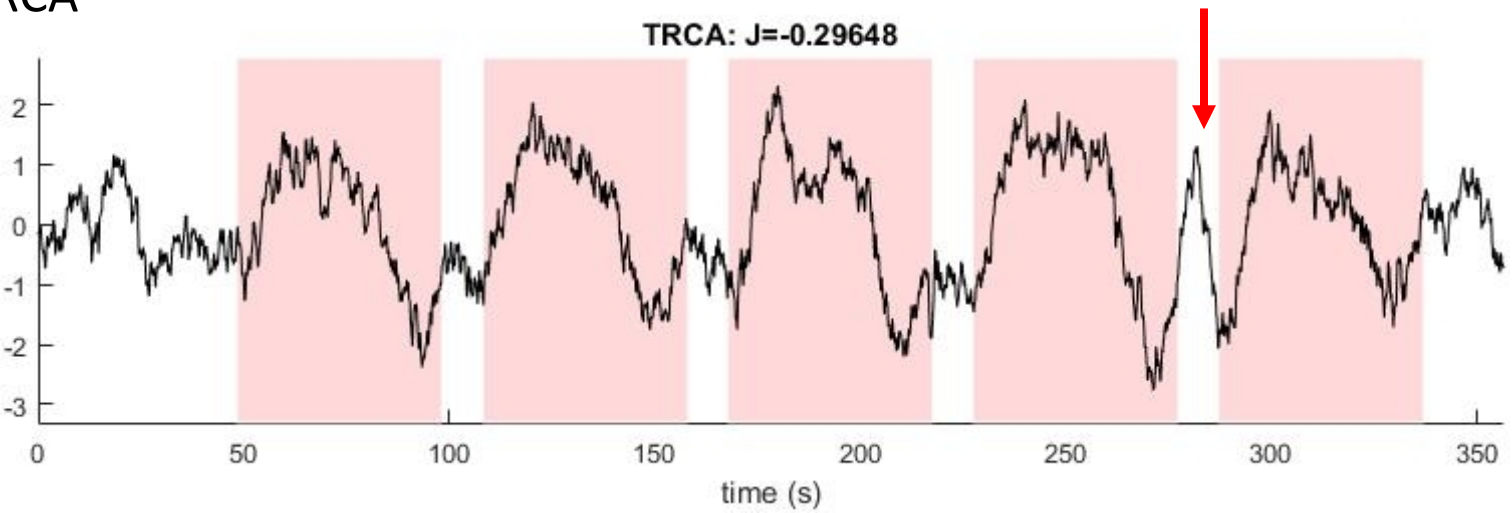
If accepted, update the optimum.

TRCA vs TRCA-SWM: more reproducible components are found.

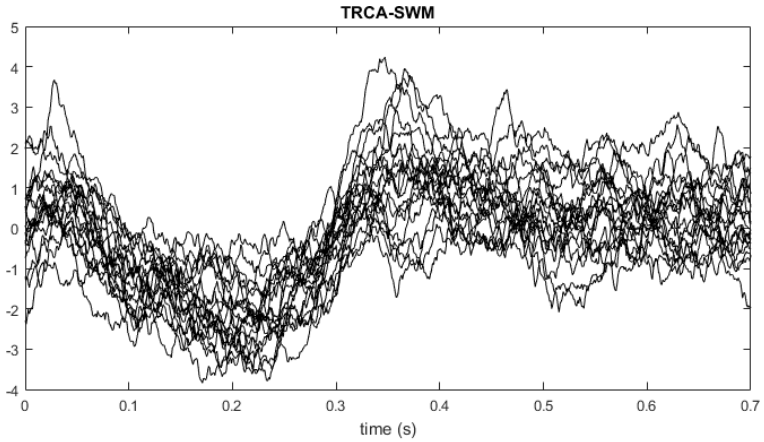
TRCA-SWM

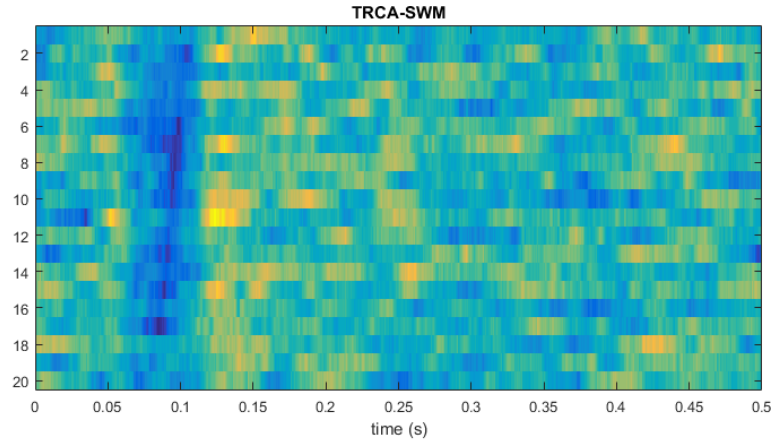


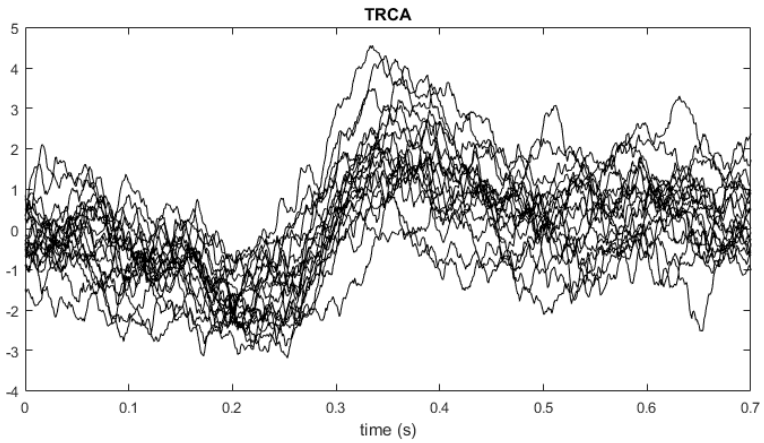
TRCA

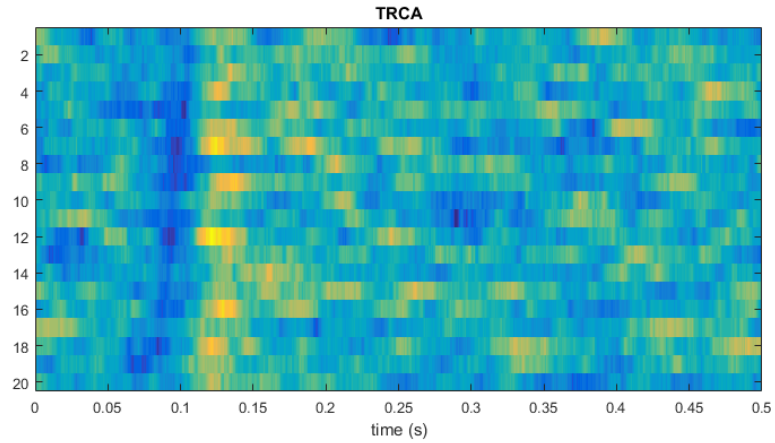


TRCA vs TRCA-SWM: more reproducible components are found.









Topics to be covered today.

1. Entropy – it's origin in physics and information theory.
2. Markov process, H-theorem and renormalization group.
3. Random variable theorem and central limit theorem.
4. Independent component analysis as inverse of CLT.
5. Matrix decompositions and non-negative matrix factorization.
6. Tensor decompositions: PARAFAC and Tucker decompositions.
7. Task-related component analysis and its extensions.

To be continued: topics NOT covered today.

- Linear causality analysis methods:
  - Granger causality
  - Partial directed coherence (PDC)
  - Directed transfer function (DTF)

- Nonlinear dynamic analysis methods:
  - Delay differential embedding (DDM)
  - Convergent cross mapping (CCM)

- Non-additive, multiplicative data decomposition:
  - Holo-Hilbert transform.

- Dictionary learning
  - Matching pursuit (MP)
  - K-SVD

References (1/2)

- Gillespie, D. T. (1983). A theorem for physicists in the theory of random variables. American Journal of Physics, 51(6), 520-533.
- Jaynes, E. T. (1957). Information theory and statistical mechanics. Physical Review, 106(4), 620.
- Jona-Lasinio, G. (1975). The renormalization group: A probabilistic view. Il Nuovo Cimento B (1971-1996), 26(1), 99-119.
- Wilson, K. G. (1979). Problems in physics with many length scales. Scientific American, 241(2), 158-179
- Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural computation, 7(6), 1129-1159..
- Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. SIAM Review, 51(3), 455-500.

- Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S. I. (2009). Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons.
- Hyvärinen, A., Karhunen, J., & Oja, E. (2004). Independent component analysis (Vol. 46). John Wiley & Sons.
- Morimoto, T. (1963). Markov processes and the H-theorem. Journal of the Physical Society of Japan, 18(3), 328-331.
- Hjalmars, S. (1977). Evidence for Boltzmann's H as a capital eta. American Journal of Physics, 45(2), 214-215.
- Makeig, S., Jung, T. P., Bell, A. J., Ghahremani, D., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences*, *94*(20), 10979-10984.

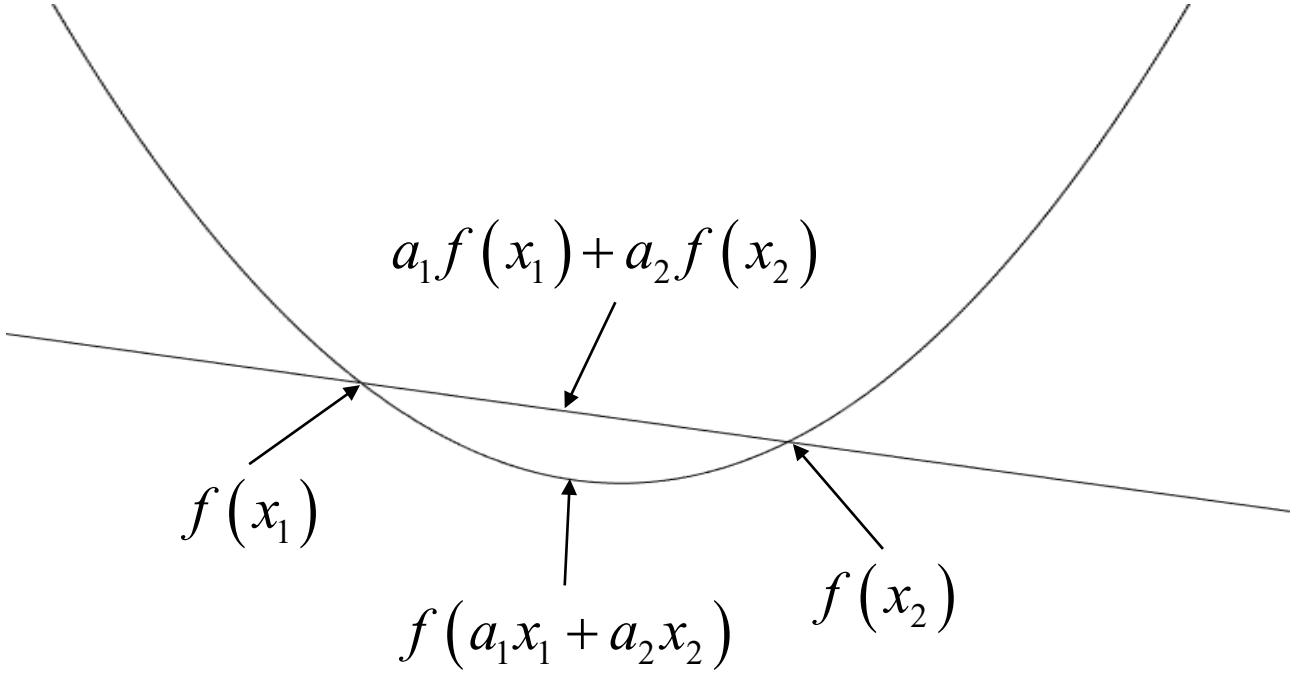
References (2/2)

- McKeown, M. J., Jung, T. P., Makeig, S., Brown, G., Kindermann, S. S., Lee, T. W., & Sejnowski, T. J. (1998). Spatially independent activity patterns in functional MRI data during the Stroop color-naming task. Proceedings of the National Academy of Sciences, 95(3), 803-810.
- Bell, A. J., & Sejnowski, T. J. (1997). The “independent components” of natural scenes are edge filters. Vision research, 37(23), 3327-3338.
- Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. Nature, 439(7079), 978-982.
- McCartin, B. J. (2006). e: The master of all. The Mathematical Intelligencer, 28(2), 10-21
- Wästlund, J. (2007). An elementary proof of the Wallis product formula for pi. The American Mathematical Monthly, 114(10), 914-917.
- Mermin, N. D. (1984). Stirling’s formula!. American Journal of Physics, 52(4), 362-365.
- Nielsen, F. Å., Balslev, D., & Hansen, L. K. (2005). Mining the posterior cingulate: segregation between memory and pain components. Neuroimage, 27(3), 520-532.
- Sotiras, A., Resnick, S. M., & Davatzikos, C. (2015). Finding imaging patterns of structural covariance via non-negative matrix factorization. NeuroImage, 108, 1-16.0
- Chao, Z. C., Nagasaka, Y., & Fujii, N. (2015). Cortical network architecture for context processing in primate brain. eLife, 4, e06121.
- Tanaka, H., Katura, T., & Sato, H. (2013). Task-related component analysis for functional neuroimaging and application to near-infrared spectroscopy data. Neuroimage, 64, 308-327.
- Tanaka, H., Katura, T., & Sato, H. (2014). Task-related oxygenation and cerebral blood volume changes estimated from NIRS signals in motor and cognitive tasks. Neuroimage, 94, 107-119.

Appendix A: Jensen's inequality for a convex function (1/2).

- For a convex function f : the value of function at intermediate point is smaller than or equal to the average of the function.

$$f\left(a_1x_1+a_2x_2\right)\le a_1f\left(x_1\right)+a_2f\left(x_2\right),\quad a_1+a_2=1, a_1\ge 0, a_2\ge 0$$



Appendix A: Jensen's inequality for a convex function (2/2).

- For a convex function f :

$$f\left(\sum_{i=1}^{N} a_i x_i\right) \le \sum_{i=1}^{N} a_i f\left(x_i\right), \quad \sum_{i=1}^{N} a_i = 1, a_i \ge 0.$$

[Proof] For three points:

$$f\left(a_1x_1 + a_2x_2 + a_3x_3\right) = f\left(a_1x_1 + \left(a_2 + a_3\right)\left(\frac{a_2}{a_2 + a_3}x_2 + \frac{a_3}{a_2 + a_3}x_3\right)\right)$$

$$\le a_1 f\left(x_1\right) + \left(a_2 + a_3\right) f\left(\frac{a_2}{a_2 + a_3}x_2 + \frac{a_3}{a_2 + a_3}x_3\right)$$

$$\le a_1 f\left(x_1\right) + a_2 f\left(x_2\right) + a_3 f\left(x_3\right)$$

The same proof for N-point case.

Appendix B: Log-sum inequality.

- Logarithm is a concave function.

$$\sum_{i=1}^{N} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{N} a_i \log \frac{\sum_{j=1}^{N} a_j}{\sum_{k=1}^{N} b_k}$$

- [proof]: Use

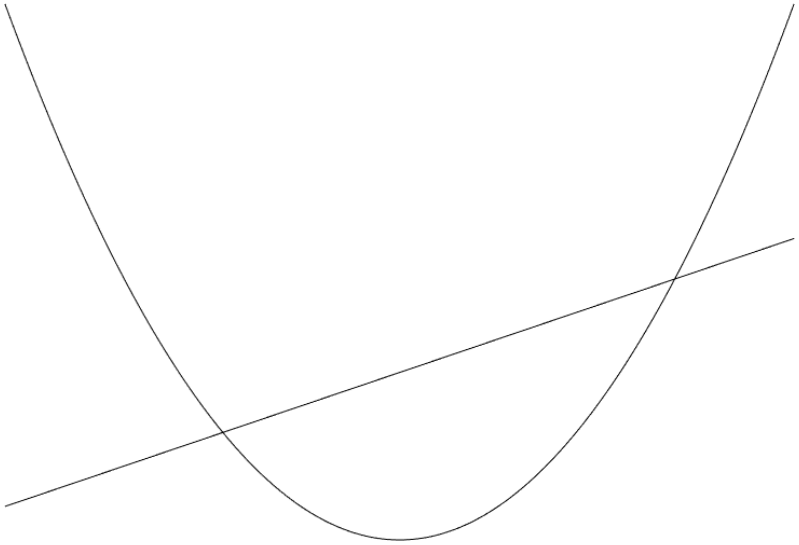
$$\sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} \geq 0, \ \sum_i p_i = \sum_i q_i = 1$$

$$p_i = \frac{a_i}{A}, q_i = \frac{b_i}{B}, \ A = \sum_j a_j, B = \sum_j b_j$$

$$0 \leq \sum_{i=1}^{N} \frac{a_i}{A} \log \frac{a_i / A}{b_i / B} = \frac{1}{A}\left(\sum_{i=1}^{N} a_i \log \frac{a_i}{b_i} - \sum_{i=1}^{N} a_i \log \frac{A}{B} \right) \qquad \therefore \sum_{i=1}^{N} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{N} a_i \log \frac{A}{B}$$

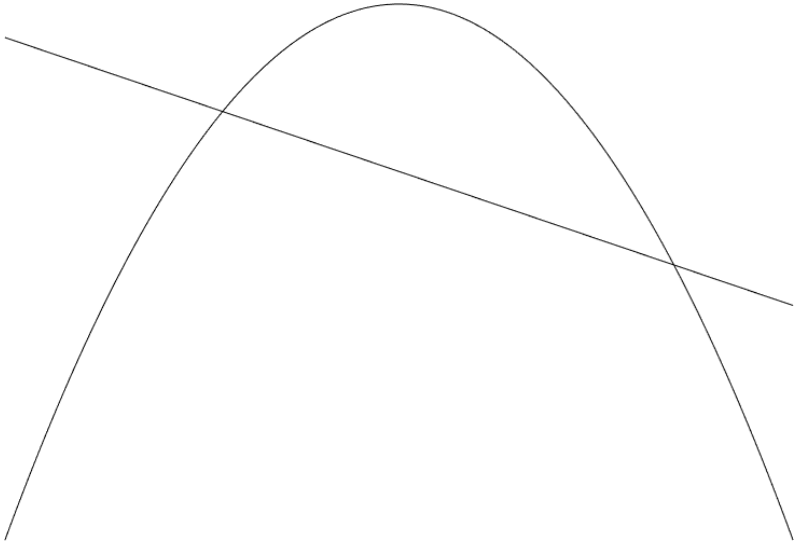
Digression: Convex (凸) and concave (凹) functions.

- "Mathematical Stroop":



$$f\left(\alpha x_1+(1-\alpha)x_2\right)\leq \alpha f\left(x_1\right)+(1-\alpha)f\left(x_2\right)$$

- ConVex function
- 凸関数



$$f\left(\alpha x_1+(1-\alpha)x_2\right)\geq \alpha f\left(x_1\right)+(1-\alpha)f\left(x_2\right)$$

- ConCave function
- 凹関数

Appendix: Matrix derivative of a determinant.

- Derivative of determinant of matrix A:

$$\frac{\partial}{\partial \mathbf{A}}\left|\mathbf{A}\right| = \mathbf{A}^{-\mathrm{T}}\left|\mathbf{A}\right|$$

- [Proof]: The determinant of A is expanded by its cofactors:

$$\left|\mathbf{A}\right| = \sum_i A_{ij} C_{ij}$$

The inverse of A is also expressed by the cofactors:

$$A_{ij}^{-1} = \frac{1}{\left|\mathbf{A}\right|} C_{ji}$$

$$\frac{\partial \left|\mathbf{A}\right|}{\partial A_{ij}} = C_{ij} = A_{ji}^{-1}\left|\mathbf{A}\right| = \left(\mathbf{A}^{-\mathrm{T}}\right)_{ij}\left|\mathbf{A}\right|$$

Appendix: Measures for matrix distance.

- Frobenius norm:

$$D_{\mathrm{EU}}\left(\mathbf{Y},\mathbf{X}\right)=\frac{1}{2}\left\|\mathbf{Y}-\mathbf{X}\right\|_{\mathrm{F}}^{2}=\frac{1}{2}\sum_{i,j}\left(y_{ij}-x_{ij}\right)^{2}=\frac{1}{2}\mathrm{tr}\left(\mathbf{Y}-\mathbf{X}\right)\left(\mathbf{Y}-\mathbf{X}\right)^{\mathrm{T}}$$

- Kullback-Leibler divergence:

$$D_{\mathrm{KL}}\left(\mathbf{Y},\mathbf{X}\right)=\sum_{i,j}\left(y_{ij}\log\frac{y_{ij}}{x_{ij}}-y_{ij}+x_{ij}\right)$$

- Itakura-Saito divergence:

$$D_{\mathrm{IS}}\left(\mathbf{Y},\mathbf{X}\right)=\sum_{i,j}\left(\frac{y_{ij}}{x_{ij}}-\log\frac{y_{ij}}{x_{ij}}-1\right)$$

Appendix: Beta matrix divergence.

- Beta divergence:

$$D_\beta\left(\mathbf{Y},\mathbf{X}\right)=\sum_{i,j}\left(y_{ij}\frac{y_{ij}^{\beta-1}-x_{ij}^{\beta-1}}{\beta-1}-\frac{y_{ij}^{\beta}-x_{ij}^{\beta}}{\beta}\right)$$

$$\lim_{\beta\to 0}D_\beta\left(\mathbf{Y},\mathbf{X}\right)=D_{\mathrm{IS}}\left(\mathbf{Y},\mathbf{X}\right),$$

$$\lim_{\beta\to 1}D_\beta\left(\mathbf{Y},\mathbf{X}\right)=D_{\mathrm{KL}}\left(\mathbf{Y},\mathbf{X}\right),$$

$$\lim_{\beta\to 2}D_\beta\left(\mathbf{Y},\mathbf{X}\right)=D_{\mathrm{EU}}\left(\mathbf{Y},\mathbf{X}\right).$$

- Note that:

$$\lim_{\beta\to 0}\frac{x^\beta-1}{\beta}=\lim_{\beta\to 0}\frac{e^{\beta\log x}-1}{\beta}=\lim_{\beta\to 0}\frac{\beta\log x}{\beta}=\log x$$

Appendix: Kronecker product.

- The Kronecker product of matrices

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_J \end{pmatrix} = \begin{pmatrix} a_{ij} \end{pmatrix} \in \mathbb{R}^{I \times J}, \mathbf{B} = \begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_J \end{pmatrix} = \begin{pmatrix} b_{kl} \end{pmatrix} \in \mathbb{R}^{K \times L}$$

is defined as

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1J}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2J}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}\mathbf{B} & a_{I2}\mathbf{B} & \cdots & a_{IJ}\mathbf{B} \end{pmatrix} \in \mathbb{R}^{IK \times JL}$$

also written as:

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 & \mathbf{a}_1 \otimes \mathbf{b}_2 & \mathbf{a}_1 \otimes \mathbf{b}_3 & \cdots & \mathbf{a}_J \otimes \mathbf{b}_{L-1} & \mathbf{a}_J \otimes \mathbf{b}_L \end{pmatrix}$$

Appendix: Khatri-Rao and Hadamard product.

- The Khatri-Rao product of matrices

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_K \end{pmatrix} = \left(a_{ik}\right) \in \mathbb{R}^{I \times K}, \mathbf{B} = \begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_K \end{pmatrix} = \left(b_{jk}\right) \in \mathbb{R}^{J \times K}$$

  is defined as

$$\mathbf{A} \odot \mathbf{B} = \begin{pmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 & \mathbf{a}_2 \otimes \mathbf{b}_2 & \cdots & \mathbf{a}_K \otimes \mathbf{b}_K \end{pmatrix} \in \mathbb{R}^{IJ \times K}$$

- The Hadamard product of matrices

$$\mathbf{A} = \left(a_{ij}\right) \in \mathbb{R}^{I \times J}, \mathbf{B} = \left(b_{ij}\right) \in \mathbb{R}^{I \times J}$$

  is defines as

$$\mathbf{A} \circledast \mathbf{B} = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} & \cdots & a_{1J}b_{1J} \\ a_{21}b_{21} & a_{22}b_2 & \cdots & a_{2J}b_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}b_{I1} & a_{I2}b_{I2} & \cdots & a_{IJ}b_{IJ} \end{pmatrix} \in \mathbb{R}^{I \times J}$$

Appendix: Some properties of matrix products.

$$\left(\mathbf{A}\otimes\mathbf{B}\right)\left(\mathbf{C}\otimes\mathbf{D}\right)=\mathbf{AC}\otimes\mathbf{BD}$$

$$\left(\mathbf{A}\otimes\mathbf{B}\right)^{\dagger}=\mathbf{A}^{\dagger}\otimes\mathbf{B}^{\dagger}$$

$$\mathbf{A}\odot\mathbf{B}\odot\mathbf{C}=\left(\mathbf{A}\odot\mathbf{B}\right)\odot\mathbf{C}=\mathbf{A}\odot\left(\mathbf{B}\odot\mathbf{C}\right)$$

$$\left(\mathbf{A}\odot\mathbf{B}\right)^{\mathrm{T}}\left(\mathbf{A}\odot\mathbf{B}\right)=\mathbf{A}^{\mathrm{T}}\mathbf{A}\circledast\mathbf{B}^{\mathrm{T}}\mathbf{B}$$

$$\left(\mathbf{A}\odot\mathbf{B}\right)^{\dagger}=\left(\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)\circledast\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)\right)^{\dagger}\left(\mathbf{A}\odot\mathbf{B}\right)^{\mathrm{T}}$$