Lecture II

Mathematical preliminaries

Ivica Kopriva

Ruđer Bošković Institute

e-mail: ikopriva@irb.hr
ikopriva@gmail.com
Course outline

◆ Motivation with illustration of applications (lecture I)
◆ Mathematical preliminaries with principal component analysis (PCA)? (lecture II)
◆ Independent component analysis (ICA) for linear static problems: information-theoretic approaches (lecture III)
◆ ICA for linear static problems: algebraic approaches (lecture IV)
◆ ICA for linear static problems with noise (lecture V)
◆ Dependent component analysis (DCA) (lecture VI)
Course outline

- Underdetermined blind source separation (BSS) and sparse component analysis (SCA) (lecture VII/VIII)
- Nonnegative matrix factorization (NMF) for determined and underdetermined BSS problems (lecture VIII/IX)
- BSS from linear convolutive (dynamic) mixtures (lecture X/XI)
- Nonlinear BSS (lecture XI/XII)
- Tensor factorization (TF): BSS of multidimensional sources and feature extraction (lecture XIII/XIV)
Mathematical preliminaries

- Expectations, moments, cumulants, kurtosis
- Gradient and optimization: vector gradient, Jacobian, matrix gradient – natural and relative gradient, gradient descent, stochastic gradient descent (batch and online)
- Estimation theory
- Information theory: entropy, differential entropy, entropy of a transformation, mutual information, distributions with maximum entropy, negentropy, entropy approximation: Gram-Charlier expansion
- PCA and whitening (batch and online)
Mathematical preliminaries

The r\textsuperscript{th} order moment of the scalar stochastic process \( z \) is defined as

\[
M_r(z) = E\left[ z^r \right] = \int_{-\infty}^{\infty} z^r f(z) dz \approx \frac{1}{T} \sum_{t=1}^{T} z^r(t)
\]

where \( f(z) \) is pdf of random variable \( z \) and \( T \) represents sample size. For real stochastic processes \( z_1 \) and \( z_2 \) the cross-moment is defined as

\[
M_{rp}(z_1, z_2) = E\left[ z_1^r z_2^p \right] \approx \frac{1}{T} \sum_{t=1}^{T} z_1^r(t) z_2^p(t)
\]

Cross-moment can be defined for nonzero time shifts as

\[
M_{r,z}^{i,j,k,l,\ldots}(\tau_1, \tau_2, \ldots, \tau_{r-1}) = E\left[ z_i(t) z_j(t + \tau_1) z_k(t + \tau_2) \ldots z_l(t + \tau_{r-1}) \right]
\]

\[
\approx \frac{1}{T} \sum_{t=1}^{T} z_i(t) z_j(t + \tau_1) z_k(t + \tau_2) \ldots z_l(t + \tau_{r-1})
\]
Mathematical preliminaries

The first characteristic function of the random process $z$ is defined as

$$
\Psi_z(\omega) = E\left[e^{j\omega z}\right] = \int_{-\infty}^{\infty} e^{j\omega z} f(z)dz
$$

Power series expansion of $\psi_z(\omega)$ is obtained as

$$
\Psi_z(\omega) = E\left[1 + j\omega z + \frac{(j\omega)^2}{2!} z^2 + ... + \frac{(j\omega)^n}{n!} z^n + ...\right]
$$

$$
= 1 + E[z] j\omega + E\left[z^2\right] \frac{(j\omega)^2}{2!} + ... + E\left[z^n\right] \frac{(j\omega)^n}{n!} + ...
$$

Moments of the random process $z$ are defined as

$$
E\left[z^n\right] = \frac{1}{j^n} \left[ \frac{d^n \Psi_z(\omega)}{d\omega^n} \right] \text{at } \omega = 0.
$$
Mathematical preliminaries

The second characteristic function of the random process $z$ is defined as

$$K_z(\omega) = \ln \Psi_z(\omega)$$

Power series expansion of $K_z(\omega)$ is obtained as

$$K_z(\omega) = C_1(z)(j\omega) + C_2(z)\frac{(j\omega)^2}{2!} + \ldots + C_n(z)\frac{(j\omega)^n}{n!} + \ldots$$

where $C_1, C_2, \ldots, C_n$ are cumulants of the appropriate order and are defined as

$$C_n(z) = \frac{1}{j^n} \left[ \frac{d^n K_z(\omega)}{d\omega^n} \right] \text{ at } \omega = 0.$$
Mathematical preliminaries

In practice cumulants, as well as moments, must be estimated from data. For that purpose relations between cumulants and moments are important. It follows from the definition of the first and second characteristic function

$$\exp\{K_z(\omega)\} = \Psi_z(\omega)$$

After Taylor series expansion the following is obtained

$$\exp\{K_z(\omega)\} = \exp[C_1(z)(j\omega)] \exp\left[C_2(z)\frac{(j\omega)^2}{2!}\right] \cdots \exp\left[C_n(z)\frac{(j\omega)^n}{n!}\right] \cdots$$

$$= 1 + E[z]\; j\omega + E[z^2]\frac{(j\omega)^2}{2!} + \cdots + E[z^n]\frac{(j\omega)^n}{n!} + \cdots$$

After exponential function is Taylor series expanded relations between cumulants and moments are obtained by equating coefficients with the same power of $j\omega$.  

8/58
Mathematical preliminaries

Relation between moments and cumulants:\textsuperscript{1-3}

\[
E[z] = C_1(z) \\
E[z^2] = C_2(z) + C_1^2(z) \\
E[z^3] = C_3(z) + 3C_2(z)C_1(z) + C_1^3(z) \\
E[z^4] = C_4(z) + 4C_3(z)C_1(z) + 3C_2^2(z) + 6C_2(z)C_1^2(z) + C_1^4(z)
\]

For zero mean process relations simplify to:

\[
E[z^2] = C_2(z) \\
E[z^3] = C_3(z) \\
E[z^4] = C_4(z) + 3C_2^2(z)
\]

Mathematical preliminaries

Relations between cumulants and moments follow from previous expressions as

\[ C_1(z) = E[z] \]
\[ C_2(z) = E[z^2] - E^2[z] \]
\[ C_3(z) = E[z^3] - 3E[z]E[z^2] + 2E^3[z] \]

and for zero mean processes relations simplify to

\[ C_2(z) = E[z^2] \]
\[ C_3(z) = E[z^3] \]
\[ C_4(z) = E[z^4] - 3E^2[z^2] \]
Mathematical preliminaries

As it is the case with cross-moments it is possible to define the cross-cumulants. They are useful in measuring statistical dependence between the random processes. For zero mean processes second order cross-cumulants are defined as

\[ C_z(i, j) = E[z_i(t)z_j(t + \tau)] \]

third order cross-cumulants are defined as

\[ C_z(i, j, k) = E\left[z_i(t)z_j(t + \tau_1)z_k(t + \tau_2)\right] \]

several third order cross-cumulants could be computed such as for example

\[ C_{z_i z_j z_k}^{111}(\tau_1, \tau_2) \quad C_{z_i z_j}^{21}(\tau_1, \tau_2) \quad C_{z_i z_j}^{12}(\tau_1, \tau_2) \quad C_z^{3}(\tau_1, \tau_2) \]
Mathematical preliminaries

fourth order cross-cumulants are defined as

\[
C_z(i, j, k, l) = \text{cum}[z_i(t), z_j(t + \tau_1), z_k(t + \tau_2), z_l(t + \tau_3)]
\]

\[
= E[z_i(t), z_j(t + \tau_1), z_k(t + \tau_2), z_l(t + \tau_3)]
- E[z_i(t)z_j(t + \tau_1)] E[z_i(t + \tau_3)z_k(t + \tau_2)]
- E[z_i(t)z_k(t + \tau_2)] E[z_i(t + \tau_3)z_j(t + \tau_1)]
- E[z_i(t)z_l(t + \tau_3)] E[z_j(t + \tau_1)z_k(t + \tau_2)]
\]

Several fourth order cross-cumulants could be computed such as for example

\[
C_{z_i z_j z_k z_i}^{1111}(\tau_1, \tau_2, \tau_3), \quad C_{z_i z_j}^{22}(\tau_1, \tau_2, \tau_3), \quad C_{z_i z_j}^{13}(\tau_1, \tau_2, \tau_3), \quad C_{z_i z_j}^{31}(\tau_1, \tau_2, \tau_3), \quad C_{z_i}^{4}(\tau_1, \tau_2, \tau_3)
\]
Mathematical preliminaries

If two random processes $z_1(t)$ and $z_2(t)$ are mutually independent then

$$E[g(z_1)h(z_2)] = E[g(z_1)]E[h(z_2)]$$

Proof:

$$E[g(z_1)h(z_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(z_1)h(z_2) p_{z_1z_2}(z_1, z_2)dz_1dz_2$$

$$= \int_{-\infty}^{\infty} g(z_1)p_{z_1}(z_1)dz_1 \int_{-\infty}^{\infty} h(z_2)p_{z_2}(z_2)dz_2$$

$$= E[g(z_1)] E[h(z_2)]$$
Mathematical preliminaries

Operator properties of the cumulants. Cumulants have certain properties that make them very useful in various computations with stochastic processes. Most of these properties will be stated without proofs which can be found in\(^2,4\).

[CP1] If \(\{\alpha_i\}_{i=1}^n\) are constants and \(\{z_i\}_{i=1}^n\) are random variables then

\[
cum(\alpha_1 z_1, \alpha_2 z_2, \ldots, \alpha_n z_n) = \left(\prod_{i=1}^n \alpha_i\right) cum(z_1, z_2, \ldots, z_n)
\]

[CP2] Cumulants are additive in their arguments

\[
cum(x_1 + y_1, z_2, \ldots, z_n) = cum(x_1, z_2, \ldots, z_n) + cum(y_1, z_2, \ldots, z_n)
\]

Mathematical preliminaries

[CP3] – Cumulants are symmetric in their arguments i.e.

\[ \text{cum}(z_1, z_2, \ldots, z_k) = \text{cum}(z_{i_1}, z_{i_2}, \ldots, z_{i_k}) \]

where \((i_1, i_2, \ldots, i_k)\) is any permutation of \((1, 2, \ldots, k)\).

[CP4] – If \(\alpha\) is constant then

\[ \text{cum}(\alpha + z_1, z_2, \ldots, z_k) = \text{cum}(z_1, z_2, \ldots, z_k) \]

[CP5] – If random processes \((z_1, z_2, \ldots, z_k)\) and \((y_1, y_2, \ldots, y_k)\) are independent then

\[ \text{cum}(y_1 + z_1, y_2 + z_2, \ldots, y_k + z_k) = \text{cum}(y_1, y_2, \ldots, y_k) + \text{cum}(z_1, z_2, \ldots, z_k) \]
Mathematical preliminaries

[CP6] – If any subset of $k$ random variables $\{z_i\}_{i=1}^k$ is statistically independent from the rest then

$$\text{cum}(z_1, z_2, ..., z_k) = 0$$

[CP7] – Cumulants of independent identically distributed (i.i.d.) random processes are delta functions i.e.

$$C_{k,z}(\tau_1, \tau_2, ..., \tau_{k-1}) = \gamma_{k,z} \delta_{\tau_1 \tau_2 ... \tau_{k-1}}$$

where $\gamma_{k,z} = C_{k,z}$. 
Mathematical preliminaries

[CP8] - If random process is Gaussian $N(\mu, \sigma^2)$ then $C_{kz}=0$ for $k>2$.

**Proof.** Let $f(z)$ be p.d.f. of the normally distributed random process $z$

$$f(z) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{(z - \mu)^2}{2\sigma^2} \right\}$$

where $\mu=E(z)=C_1(z)$ is the mean value or the first order cumulant of the random process $z$ and $\sigma^2=E[z^2]-E^2[z]=C_2(z)$ is the variance or the second order cumulant of the random process $z$.

According to the definition of the first characteristic function

$$\Psi_z(\omega) = E[e^{i\omega z}] = \int_{-\infty}^{\infty} f(z)e^{i\omega z} dz =$$

$$= \exp \left\{ j\mu \omega + \frac{(\sigma j \omega)^2}{2} \right\} \times \int_{-\infty}^{\infty} \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{(z - \hat{\mu})^2}{2\sigma^2} \right\} dz$$

$$= 1$$
Mathematical preliminaries

where $\hat{\mu} = \mu + \sigma^2 j\omega$. By definition of the second characteristic function

$$K_z(\omega) = \ln \Psi_z(\omega) = \mu(j\omega) + \sigma^2 \frac{(j\omega)^2}{2}$$

when compared with Taylor series expansion of $K_z(\omega)$ it follows that $C_k(z)=0$ for $k>2$.

The complexity for the cumulants of the order higher than 4 is very high. An important question that has to be answered from the application point of view is what statistics of the order higher than 2 should be used? Because symmetrically distributed random processes, which occur very often in engineering applications, have odd order statistics equal to zero the third order statistics are not used. So, the first statistics of the order higher than 2 that are used are the fourth order statistics.
Mathematical preliminaries

Another natural question that should be answered is: why and when cumulants are better than moments?

- Computation with cumulants is simpler than by using moments.
- For independent random variables cumulant of the sum equals the sum of the cumulants.
- Cross-cumulants of the independent random processes are zero what does not apply for moments.
- For normally distributed processes cumulants of order greater than 2 vanish what does not apply for moments.
- Blind identification of the LTI systems (MA, AR and ARMA) is possible to formulate using cumulants provided that input signal is non-Gaussian i.i.d. random process.
Mathematical preliminaries

Density of transformation. Assume that both \( x \) and \( y \) are \( n \)-dimensional random vectors related through the vector mapping \( y = g(x) \). It is assumed that inverse mapping \( x = g^{-1}(y) \) exists and is unique. Density \( p_y(y) \) is obtained from density \( p_x(x) \) as follows\(^5\)

\[
p_y(y) = \frac{1}{\det Jg(g^{-1}(y))} p_x(x)
\]

Where \( Jg \) is the Jacobian matrix representing the gradient of the vector-valued function

\[
Jg(x) = \begin{bmatrix}
\frac{\partial g_1(x)}{\partial x_1} & \frac{\partial g_2(x)}{\partial x_1} & \cdots & \frac{\partial g_n(x)}{\partial x_1} \\
\frac{\partial g_1(x)}{\partial x_2} & \frac{\partial g_2(x)}{\partial x_2} & \cdots & \frac{\partial g_n(x)}{\partial x_2} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial g_1(x)}{\partial x_n} & \frac{\partial g_2(x)}{\partial x_n} & \cdots & \frac{\partial g_n(x)}{\partial x_n}
\end{bmatrix}
\]

For linear nonsingular transformation \( y = Ax \) the relation between densities becomes

\[
p_y(y) = \frac{1}{|\det A|} p_x(x)
\]

Mathematical preliminaries

Kurtosis. Based on the measure called kurtosis it is possible to measure how far certain stochastic process is from Gaussian distribution. The kurtosis is defined as

\[ \kappa(z_i) = \frac{C_4(z_i)}{C_2^2(z_i)} \]

where \( C_4(z_i) \) is the FO and \( C_2(z_i) \) is the SO cumulant of the signal \( z_i \). For the zero mean signals kurtosis becomes

\[ \kappa(z_i) = \frac{E[z_i^4]}{E^2[z_i^2]} - 3 \]
According to cumulant property [CP8] the FO cumulant of the Gaussian process is zero. Signals with positive kurtosis are classified as super-Gaussian and signals with negative kurtosis as sub-Gaussian. Most of the communication signals are sub-Gaussian processes. Speech and music signals are super-Gaussian processes. Uniformly distributed noise is sub-Gaussian process. Impulsive noise (with Laplacian or Lorentz distribution) is highly super-Gaussian process.
Mathematical preliminaries

Gradients and optimization (learning) methods.

• Gradients are used in minimization or maximization of the cost functions. This process is called learning.

• Solution of the ICA problem is also obtained as a solution of the minimization or maximization problem of the scalar cost function w.r.t. matrix argument.

• Scalar function $g$ with vector argument is defined as $g = g(w_1, w_2, ..., w_m) = g(w)$. Vector gradient w.r.t. $w$ is defined as

$$\frac{\partial g}{\partial w} = \begin{pmatrix} \frac{\partial g}{\partial w_1} \\ \vdots \\ \frac{\partial g}{\partial w_m} \end{pmatrix}$$

Another commonly used notation for vector gradient is $\nabla g$ or $\nabla_w g$. 
Mathematical preliminaries

Gradients and optimization (learning) methods.

Vector valued function is defined as

$$g(w) = \begin{pmatrix}
g_1(w) \\
... \\
g_n(w)
\end{pmatrix}$$

Gradient of the vector valued function is called \textit{Jacobian matrix} of $g$ and is defined as

$$Jg(w) = \begin{bmatrix}
\frac{\partial g_1(w)}{\partial w_1} & \frac{\partial g_2(w)}{\partial w_1} & ... & \frac{\partial g_n(w)}{\partial w_1} \\
\frac{\partial g_1(w)}{\partial w_2} & \frac{\partial g_2(w)}{\partial w_2} & ... & \frac{\partial g_n(w)}{\partial w_2} \\
... & ... & ... & ... \\
\frac{\partial g_1(w)}{\partial w_m} & \frac{\partial g_2(w)}{\partial w_m} & ... & \frac{\partial g_n(w)}{\partial w_m}
\end{bmatrix}$$
Mathematical preliminaries

Gradients and optimization (learning) methods.

In solving ICA problem a scalar-valued function $g$ with matrix argument $W$ is defined as $g=g(W)=g(w_{11},...,w_{ij},...,w_{mn})$.

The matrix gradient is defined as

$$\frac{\partial g}{\partial W} = \begin{pmatrix}
\frac{\partial g}{\partial w_{11}} & \ldots & \frac{\partial g}{\partial w_{1n}} \\
\vdots & \ddots & \vdots \\
\frac{\partial g}{\partial w_{m1}} & \ldots & \frac{\partial g}{\partial w_{mn}}
\end{pmatrix}$$
Mathematical preliminaries

Example – matrix gradient of the determinant of a matrix (necessary in some ICA algorithms).

For invertible square matrix $W$ gradient of the determinant of a matrix is obtained as

$$\frac{\partial}{\partial W} \det W = \left( W^T \right)^{-1} \det W$$

**Proof.** The following definition for inverse matrix is used

$$W^{-1} = \frac{1}{\det W} \text{adj}(W)$$

where $\text{adj}(W)$ stands for *adjoint* of $W$ and is defined as

$$\text{adj}(W) = \begin{pmatrix}
W_{11} & \cdots & W_{n1} \\
\vdots & \ddots & \vdots \\
W_{1n} & \cdots & W_{nn}
\end{pmatrix}$$

and scalars $W_{ij}$ are called *cofactors.*
Mathematical preliminaries

Determinant can be expressed in terms of cofactors as

$$\det W = \sum_{k=1}^{n} w_{ik} W_{ik}$$

It follows

$$\frac{\partial \det W}{\partial w_{ij}} = W_{ij} \rightarrow \frac{\partial \det W}{\partial W} = \text{adj}(W)^T = (\det W)(W^T)^{-1}$$

The previous result also implies

$$\frac{\partial \log|\det W|}{\partial w_{ij}} = \frac{1}{|\det W|} \frac{\partial |\det W|}{\partial W} = (W^T)^{-1}$$
Mathematical preliminaries

Natural or relative gradient. Parameter space of the square matrices has Riemannian metric structure and Euclidean gradient does not point into steepest direction of the cost function $J(W)$. It has to be corrected in order to obtain natural or relative gradient i.e. $(\partial J/\partial W) W^T W$ i.e. correction term is given by $W^T W$.

---

Mathematical preliminaries

**Proof.** We shall follow derivation given by Cardoso in\(^7\). Let us write the Taylor series of \(J(W + \delta W)\):

\[
J(W + \delta W) = J(W) + \text{trace} \left[ \left( \frac{\partial J}{\partial W} \right)^T \delta W \right] + ...
\]

Let us require that displacement \(\delta W\) is always proportional to \(W\) itself, \(\delta W = DW\). Then

\[
J(W + DW) = J(W) + \text{trace} \left[ \left( \frac{\partial J}{\partial W} \right)^T DW \right] + ...
\]

\[
= J(W) + \text{trace} \left[ DW \left( \frac{\partial J}{\partial W} \right)^T \right] + ...
\]
Mathematical preliminaries

Using the property of the trace operator \( \text{trace}(M_1M_2) = \text{trace}(M_2M_1) \) we can write

\[
\text{trace} \left[ D \left( \frac{\partial J}{\partial W} W^T \right)^T \right] = \text{trace} \left[ \left( \frac{\partial J}{\partial W} W^T \right)^T D \right]
\]

We seek for the largest decrement in the value of \( J(W+DW)-J(W) \) that is obtained when term \( \text{trace} \left[ \left( \frac{\partial J}{\partial W} W^T \right)^T D \right] \) is minimized. That will happen when \( D \) is proportional to \( -\frac{\partial J}{\partial W} W^T \). Because \( \delta W = DW \) we obtain the gradient descent learning rule

\[
\Delta W \propto - \left( \frac{\partial J}{\partial W} W^T \right) W = - \frac{\partial J}{\partial W} W^T W
\]

implying that correction term is \( W^T W \).
Mathematical preliminaries

Exercise. Let the scalar function with matrix argument be defined as $f(A) = (y-Ax)^T (y-Ax)$ where $y$ and $x$ are vectors.

Euclidean gradient is given with: $\frac{\partial f}{\partial A} = -2(y-Ax)x^T$.

Riemannian gradient is given with $(\frac{\partial f}{\partial A})^T A = -2(y-Ax)x^T A^T A$.

Euclidean gradient based learning equation:
$$A(k+1) = A(k) - \mu \frac{\partial f}{\partial A}$$

Riemannian gradient based learning equation:
$$A(k+1) = A(k) - \mu (\frac{\partial f}{\partial A})^T A$$

Evaluate convergence properties of these two learning equations starting with the initial condition $A^0 = I$. Assume the true values as: $A = [2 1; 1 2]$, $x = [1 1]^T$, $y = Ax = [3 3]^T$. 

31/58
Mathematical preliminaries

**Gradient descent and stochastic gradient descent.** ICA like in many other statistical and neural network techniques is data driven i.e. we need observation data \((x)\) in order to solve the problem. Typical cost function has the form

\[
J(W) = E\{g(W, x)\}
\]

where expectation is defined w.r.t. some unknown density \(f(x)\). The steepest descent learning rule becomes

\[
W(t) = W(t - 1) - \mu(t) \frac{\partial}{\partial W} E\{g(W, x(t))\} \bigg|_{W=W(t-1)}
\]

In practice expectation has to be approximated by the sample mean of function \(g\) over the sample \(x(1), x(2), \ldots, x(T)\). The algorithm where entire training set is used at every iteration step is called *batch learning.*
Mathematical preliminaries

**Gradient descent and stochastic gradient descent.** Sometimes not all the observation could be known in advance but only the latest \( x(t) \) could be used. This is called *on-line* or *adaptive* learning. The *on-line* learning rule is obtained by dropping expectation operator from the batch learning i.e.

$$ W(t) = W(t - 1) - \mu(t) \frac{\partial}{\partial W} g(W, x) \bigg|_{W=W(t-1)} $$

Parameter \( \mu(t) \) is called the learning gain and regulates smoothness and speed of convergence. *On-line* and *batch* algorithms converge toward the same solution that represents fixed point of the first order autonomous differential equation

$$ \frac{dW}{dt} = - \frac{\partial}{\partial W} E \{ g(W, x) \} $$

The learning rate \( \mu(t) \) should be chosen as

$$ \mu(t) = \frac{\beta}{\beta + t} $$

where \( \beta \) is an appropriate constant (e.g. \( \beta = 100 \)).
Mathematical preliminaries

Information theory – entropy, negentropy and mutual information.

Entropy of a random variable can be interpreted as the degree of information that the observation of the variable gives. For discrete random variable $X$ entropy $H$ is defined as

$$H(X) = -\sum_i P(X = a_i) \log P(X = a_i) = \sum_i f(P(X = a_i))$$

Entropy could be generalized for continuous random variables and vectors in which case it is called the differential entropy.

$$H(x) = -\int p_x(\xi) \log p_x(\xi) d\xi = -E[\log p_x(x)]$$

Unlike discrete entropy, differential entropy can be negative. Entropy is small when variable is not very random i.e. when it is contained in some limited intervals with high probabilities.
Mathematical preliminaries

Maximum entropy distributions – finite support case.

Uniform distribution in the interval \([0, a]\). Density is given with

\[
p_x(\xi) = \begin{cases} 
\frac{1}{a} & \text{for } 0 \leq \xi \leq a \\
0 & \text{otherwise}
\end{cases}
\]

Differential entropy is evaluated as

\[
H(x) = -\int_{0}^{a} \frac{1}{a} \log \frac{1}{a} d\xi = \log a
\]

Entropy is small (in a sense of negative numbers) or large in absolute value sense when \(a\) is small.

In limit

\[
\lim_{a \to 0} H(x) = -\infty
\]

Uniform distribution is the maximum entropy distribution among the finite support distributions.
Mathematical preliminaries

Maximum entropy distributions – infinite support case.

What is distributions that is compatible with the measurements \( (c_i = E\{F^i(x)\}) \) and makes minimum number of assumptions on the data → maximum entropy distributions?

\[
\int p(\xi) F^i(\xi) d\xi = c_i \quad i = 1, \ldots, m
\]

It has been shown in⁵ that density \( p_0(\xi) \) which satisfies previous constraint and has maximum entropy is of the form

\[
p_0(\xi) = A \exp \left( \sum_i a_i F^i(\xi) \right)
\]

For set of random variables that take all the values on the real line and have zero mean and fixed variance, say 1, maximum entropy distribution takes the form

\[
p_0(\xi) = A \exp \left( a_1 \xi^2 + a_2 \xi \right)
\]

that is recognized as Gaussian variable. It has the largest entropy among all random variables of finite (unit) variance with infinite support.

Mathematical preliminaries

Entropy of a transformation. Consider an invertible transformation of the random vector \( x \)

\[
y = f(x)
\]

Relation between densities is given by

\[
p_y(\eta) = p_x(f^{-1}(\eta)) \left| \det Jf(f^{-1}(\eta)) \right|^{-1}
\]

Expressing entropy as expectation we obtain

\[
H(y) = -E\left\{ \log p_y(y) \right\} = -E\left\{ \log \left[ p_x(x) \left| \det Jf(x) \right|^{-1} \right] \right\} = E\left\{ \log |\det Jf(x)| \right\} - E\left\{ \log p_x(x) \right\} = H(x) + E\left\{ \log |\det Jf(x)| \right\}
\]

Entropy is increased by transformation. For the linear transform \( y=Ax \) we obtain

\[
H(y) = H(x) + \log |\det A|
\]

Differential entropy is not scale invariant i.e. \( H(\alpha x) = H(x) + \log | \alpha | \).

Scale must be fixed before differential entropy is measured.
Mathematical preliminaries

**Mutual information** is a measure of the information that members of a set of random variables have on the other random variables in the set (vector). One measure of mutual information is a distance between components of the set (vector) measured by Kullback-Leibler divergence:

\[
MI(y) = D \left( p(y), \prod_{i=1}^{N} p(y_i) \right) = \int p(y) \log \frac{p(y)}{\prod_{i=1}^{N} p(y_i)} \, dy
\]

\[
= \sum_{i=1}^{N} H(y_i) - H(y)
\]

Using Jensen’s inequality\(^8\) it can be shown \(MI(y) \geq 0\).

\[
MI(y) = 0 \iff p(y) = \prod_{i=1}^{N} p(y_i)
\]

Mutual information can be used as a measure of statistical (in)dependence.

Mathematical preliminaries

Negentropy. The fact that Gaussian variable has maximum entropy enables to use entropy as a measure of non-Gaussianity. A measure called negentropy is defined to be zero for Gaussian variable and positive otherwise:

\[ J(x) = H(x_{gauss}) - H(x) \]

Where \( x_{gauss} \) is a Gaussian random vector of the same covariance matrix \( \Sigma \) as \( x \). Entropy of Gaussian random vector can be evaluated as

\[ H(x_{gauss}) = \frac{1}{2} \log |\det \Sigma| + \frac{N}{2} \left[ 1 + \log 2\pi \right] \]

Negentropy has an additional property to be \textit{invariant for invertible linear transformation} \( y = Ax \)

\[ J(Ax) = \frac{1}{2} \log |\det (A\Sigma A^T)| + \frac{N}{2} \left[ 1 + \log 2\pi \right] - \left( H(x) + \log |A| \right) \]

\[ = \frac{1}{2} \log |\det \Sigma| + 2 \frac{1}{2} \log |\det A| + \frac{N}{2} \left[ 1 + \log 2\pi \right] - H(x) - \log |\det A| \]

\[ = \frac{1}{2} \log |\det \Sigma| + \frac{N}{2} \left[ 1 + \log 2\pi \right] - H(x) \]

\[ = H(x_{gauss}) - H(x) = J(x) \]
Mathematical preliminaries

Approximation of entropy by cumulants. Entropy and negentropy are important measures of non-Gaussianity and thus very important for the ICA. However, computation of entropy involves evaluation of integral that include density function. This is computationally very difficult and would limit the use of entropy/negentropy in the algorithm design.

To obtain computationally tractable algorithms density function used in entropy integral is approximated using either Edgeworth or Gram-Charlier expansion which are Taylor series-like expansions around Gaussian density.

Gram-Charlie expansion of the standardized random variable \( x \) with non-Gaussian density is obtained as

\[
p_x(\xi) \approx \hat{p}_x(\xi) = \varphi(\xi) \left( 1 + C_3(x) \frac{H_3(\xi)}{3!} + C_4(x) \frac{H_4(\xi)}{4!} + \ldots \right)
\]

where \( C_3(x) \) and \( C_4(x) \) are cumulants of order 3 and 4, \( \varphi(\xi) \) is Gaussian density and \( H_3(\xi), H_4(\xi) \) are Hermite polynomials defined as

\[
H_i(\xi) = \frac{1}{\varphi(\xi)} (-1)^i \frac{\partial^i \varphi(\xi)}{\partial \xi^i}
\]
Mathematical preliminaries

Entropy can now be estimated from

\[ H(x) \approx -\int \hat{p}_x(\xi) \log \hat{p}_x(\xi) \, d\xi \]

Assuming small deviation from Gaussianity we obtain

\[ H(x) \approx -\int \phi(\xi) \left(1 + C_3(x) \frac{H_3(\xi)}{3!} + C_4(x) \frac{H_4(\xi)}{4!}\right) \, d\xi \]

\[ \approx -\int \phi(\xi) \log \phi(\xi) \, d\xi - \frac{C_3(x)^2}{2 \times 3!} - \frac{C_4(x)^2}{2 \times 4!} \]

and computationally simple approximation of negentropy is obtained as

\[ J(x) \approx \frac{1}{12} E\{x^3\}^2 + \frac{1}{48} kurt(x)^2 \]
Mathematical preliminaries

Principal Component Analysis (PCA) and whitening (batch and online). PCA is basically decorrelation based transform used in multivariate data analysis. In connection with ICA it is very often a useful preprocessing step used in the whitening transformation after which multivariate data become uncorrelated with unit variance.

PCA and/or whitening transform are designed on the basis of the eigenvector decomposition of the sample data covariance matrix

$$R_{xx} \approx \frac{1}{K} \sum_{k=1}^{K} x(k)x(k)^T$$

It is assumed data $x$ is zero mean. If not this is achieved by $x \leftarrow x - E\{x\}$. Eigendecomposition of $R_{xx}$ is obtained as

$$R_{xx} = E\Lambda E^T$$

Where $E$ is matrix of eigenvectors and $\Lambda$ is diagonal matrix of eigenvalues of $R_{xx}$. 
Mathematical preliminaries

Batch form of PCA/whitening transform is obtained as

\[ z = Vx = \Lambda^{-1/2}E^Tx \]

and one verifies

\[
E\left[zz^T\right] = E\left[\Lambda^{-1/2}E^Txx^TE\Lambda^{-1/2}\right] = \Lambda^{-1/2}E^TE\left[xx^T\right]E\Lambda^{-1/2} \\
= \Lambda^{-1/2} \underbrace{E^TE}_{I} \Lambda \underbrace{E^TE}_{I} \Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I
\]

an adaptive whitening transform is obtained as

\[ z_k = V_k x_k \]

\[
V_{k+1} = V_k - \mu_k \left[ z_k z_k^T - I \right] V_k
\]

where \( \mu \) is small learning gain and \( V_0 = I \).
Mathematical preliminaries

Scatter plots of two uncorrelated Gaussian signals (left); two correlated signals obtained as linear combinations of the uncorrelated Gaussian signals (center); two signals after PCA/whitening transform (right).

\[ y = \Lambda^{-1/2} E^T z \]

\[ z = [z_1; z_2] \]

\[ x_1 = N(0,4); \quad x_2 = N(0,9) \]

\[ z_1 = x_1 + x_2 \]
\[ z_2 = x_1 + 2x_2 \]
Mathematical preliminaries

\[ x_1 = 2s_1 + s_2 \]
\[ x_2 = s_1 + s_2 \]
\[ y_1 \approx s_1 (?) \]
\[ y_2 \approx s_2 (?) \]
What is ICA?

Imagine situation in which two microphones recording weighted sums of the two signals emitted by the speaker and background noise.

\[
x_1 = a_{11}s_1 + a_{12}s_2 \\
x_2 = a_{21}s_1 + a_{22}s_2
\]

The problems is to estimated the speech signal \((s_1)\) and noise signal \((s_2)\) from observations \(x_1\) and \(x_2\).

If mixing coefficients \(a_{11}, a_{12}, a_{21}\) and \(a_{22}\) are known problem would be solvable by simple matrix inversion.

ICA enables to estimated speech signal \((s_1)\) and noise signal \((s_2)\) without knowing the mixing coefficients \(a_{11}, a_{12}, a_{21}\) and \(a_{22}\). This is why the problem of recovering source signals \(s_1\) and \(s_2\) is called *blind source separation* problem.
What is ICA ?

ICA is the most frequently used method to solve the BSS problem. ICA could be considered as theory for multichannel blind signal recovery requiring minimum of *a priori* information.

**Problem:**

\[ x = As \text{ or } x = A*s \text{ or } x = f(As) \text{ or } x = f(A*s) \]

**Goal:** find \( s \) based on \( x \) only (\( A \) is unknown).

\( W \) can be found such that:

\[ y \cong s = Wx \text{ or } y \cong s = W*x \rightarrow y \cong P\Lambda s \]

---

What is ICA?

Environment

Sources $S_1, S_2, \ldots, S_N$

Sensors

Observations

$x = As$

$y = Wx$

Source Separation Algorithm

$W$
Speech from noise separation
When does ICA work!? 

- Source signals \( s_i(t) \) must be statistically independent.

\[
p(s) = \prod_{i=1}^{N} p_i(s_i)
\]

- Source signals \( s_i(t) \), except one, must be non-Gaussian.

\[
C_n(s_i) \neq 0 \quad n > 2
\]

- Mixing matrix \( A \) must be nonsingular and square.

\[
W \cong A^{-1}
\]

Knowledge of the physical interpretation of the mixing matrix is of crucial importance.
When does ICA work !?

Ambiguities of ICA.

a) Variances (energies) of the independent components can not be determined. This is called *scaling indeterminacy*. The reason is that both $s$ and $A$ being unknown any scalar multiplier in one of the sources can always be canceled by dividing the corresponding column of $A$ by the same multiplier:

$$ x = \sum_i \left( \frac{1}{\alpha_i} a_i \right) (s_i \alpha_i) $$

b) Order of the independent components can not be determined. This is called *permutation indeterminacy*. The reason is that components of the source vector $s$ and columns of the mixing matrix $A$ could be freely changed in such that

$$ x = AP^{-1}Ps $$

where $P$ permutation matrix, $Ps$ is new source vector with original components but in different order and $AP^{-1}$ is a new unknown mixing matrix.
When does ICA work!?

To illustrate ICA model in statistical terms we consider two independent uniformly distribute signals $s_1$ and $s_2$. Scatter plot on the left picture shows that there is no redundancy between them i.e. no knowledge could be gained about $s_2$ from $s_1$.

Right picture shows mixed signals obtained according to $x = As$ where $A = [\begin{bmatrix} 5 & 10 \\ 10 & 2 \end{bmatrix}]$. Obviously there is dependence between $x_1$ and $x_2$. Knowing maximum or minimum of $x_1$ enables to know $x_2 \rightarrow A$ could be estimated. But what if sources signals have different statistics?
When does ICA work!?

Consider two independent source signals $s_1$ and $s_2$ generated by the truck and tank engines. Scatter plot on the left picture shows that there is again no redundancy between them i.e. no knowledge could be gained about $s_2$ from $s_1$. But distribution is different i.e. \textit{sparse}. Right picture shows mixed signals obtained according to $x = As$ where $A = \begin{bmatrix} 5 & 10 \\ 10 & 2 \end{bmatrix}$. Obviously there is dependence between $x_1$ and $x_2$. But edges are at different positions this time. Estimated $A$ from the scatter plot would be very unreliable. ICA can do that for different signal statistics without knowing that in advance.
When does ICA work!?

**Whitening is only half of the ICA.** Whitening transform decorrelates signals. If signals are non-Gaussian it does not make them independent. Whitening transform is usually useful first processing step in ICA. A second rotation stage achieved by an unitary matrix can be obtained by ICA exploiting non-Gaussianity of the signals.
When does ICA work!? 

Why Gaussian variables are forbidden? Suppose two independent components $s_1$ and $s_2$ are Gaussian. Their joint density is given by 

$$p(s_1, s_2) = \frac{1}{2\pi} \exp \left( -\frac{s_1^2 + s_2^2}{2} \right) = \frac{1}{2\pi} \exp \left( -\frac{\|s\|^2}{2} \right)$$

Assume that source signals are mixed with orthogonal mixing matrix i.e. $A^{-1} = A^T$. Their joint density is given by 

$$p(x_1, x_2) = \frac{1}{2\pi} \exp \left( -\frac{\|A^T x\|^2}{2} \right) \left| \det A^T \right| = \frac{1}{2\pi} \exp \left( -\frac{\|x\|^2}{2} \right)$$

Original and mixing distributions are identical $\Rightarrow$ there is no way that we could infer mixing matrix from the mixtures.
When does ICA work!? 

Why Gaussian variables are forbidden? Consider mixture of two Gaussian signals $s_1 = N(0,4)$ and $s_2 = N(0,9)$. After whitening stage scatter plot is the same as for source signals. There is no redundancy left after the whitening stage $\rightarrow$ nothing for ICA to do.

Source signals  
Mixed signals  
Whitened signals
When does ICA work!? 

PCA applied to blind image separation:

MATLAB code:

```
R_x = cov(X');  % estimate of the data covariance matrix
[E,D] = eig(R_x);  % eigen-decomposition of the data covariance matrix
Z = E'*X;       % PCA transform
z_1 = reshape(Z(1,:),P,Q);  % transforming vector into image
figure(1); imagesc(z_1);  % show first PCA image
z_2 = reshape(Z(2,:),P,Q);  % transforming vector into image
figure(2); imagesc(z_2);  % show second PCA image
```
Histograms of source, mixed and PCA extracted images

Source image  Mixed images  PCA extracted images