Lecture Notes on Independent Component Analysis

Laurenz Wiskott
Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany, EU

11 December 2016

Contents

LECTURE 1/2 **2**

1 Intuition **2**

1.1 Mixing and unmixing . . . 2
1.2 How to find the unmixing matrix? . . . 4
1.3 Sources can only be recovered up to permutation and rescaling . . . 5
1.4 Whiten the data first . . . 5
1.5 A generic ICA algorithm . . . 5

2 Formalism based on cumulants **6**

2.1 Moments and cumulants . . . 6
2.2 Cross-cumulants of statistically independent components are zero . . . 7

LECTURE 2/2 **8**

2.3 Components with zero cross-cumulants are statistically independent . . . 8
2.4 Rotated cumulants . . . 9
2.5 Contrast function . . . 9
2.6 Givens-rotations . . . 10
2.7 Optimizing the contrast function . . . 10
2.8 The algorithm . . . 11

© 2009, 2011–2013, 2016 Laurenz Wiskott (homepage https://www.ini.rub.de/PEOPLE/wiskott/). This work (except for all figures from other sources, if present) is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/4.0/. Figures from other sources have their own copyright, which is generally indicated. Do not distribute parts of these lecture notes showing figures with non-free copyrights (here usually figures I have the rights to publish but you don't, like my own published figures). Figures I do not have the rights to publish are grayed out, but the word 'Figure', 'Image', or the like in the reference is often linked to a pdf.
More teaching material is available at https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/.

3 Other resources **11**
3.1 Written material 11
3.2 Videos 11
3.3 Software 12
3.4 Exercises 12

These lecture notes depend on my lecture notes on principal component analysis and are largely based on (Hyvärinen et al., 2001; Blaschke and Wiskott, 2004).

LECTURE 1/2

1 Intuition

1.1 Mixing and unmixing

In contrast to principal component analysis, which deals with the second-order moments of a data distribution, independent component analysis focuses on higher-order moments, which can, of course, be of very diverse and very complex nature. **In (linear) independent component analysis (ICA) one assumes**[1] a very simple model of the data, namely that it is **a linear mixture** (D: Mischung) **of some statistically independent sources** (D: Quellen) s_I**, and** one often even assumes **that the number of sources** I **is the same as the dimensionality** N **of the data.** Each source is characterized by a probability density function (pdf) (D: Wahrscheinlichkeitsdichtefunktino) $p_{s_i}(s_i)$ and **the joint pdf of the sources is simply the product of its individual pdfs,** for two sources we have

$$\blacklozenge \quad p_{\mathbf{s}}(s_1, s_2) \quad = \quad p_{s_1}(s_1)\, p_{s_2}(s_2)\,. \tag{1}$$

For an example see figure 1.

It is assumed that the data is generated by mixing the sources linearly like

$$\blacklozenge \quad \mathbf{x} \quad := \quad \mathbf{M}\mathbf{s}\,, \tag{2}$$

with an invertible square mixing matrix $\mathbf{M}$. The task of independent component analysis then is to find a square matrix $\mathbf{U}$ that inverts this mixture, so that

$$\blacklozenge \quad \mathbf{y} \quad := \quad \mathbf{U}\mathbf{x} \tag{3}$$

recovers the sources. Consider first two examples where we know the mixing.

Example 1: Figure 2 shows an example of the two-dimensional mixture

$$\lozenge \quad \mathbf{x} \quad := \quad s_1\mathbf{d}_1 + s_2\mathbf{d}_2 \;=\; \underbrace{\begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 \end{pmatrix}}_{=:\,\mathbf{M}} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \tag{4}$$

of the sources of figure 1, where the mixing is orthogonal. This means that matrix $\mathbf{M}$ is orthogonal if we assume the vectors $\mathbf{d}_i$ are normalized. Notice that the vectors $\mathbf{d}_i$ indicate how a single source is distributed over the different data vector components. I therefore call them *distribution vectors* (D: Verteilungsvektoren),

[1] Important text (but not inline formulas) is set in bold face; ♦ marks important formulas or items worth remembering; ◊ marks less important formulas or items, which I also discuss in the lecture; + marks sections that I typically skip during my lectures.

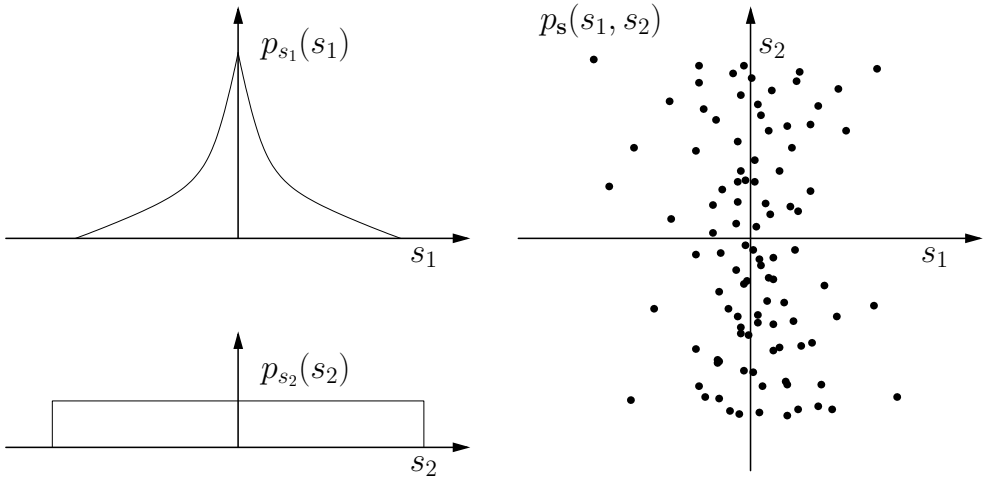

Figure 1: Individual and joint pdfs of two sources.
© CC BY-SA 4.0

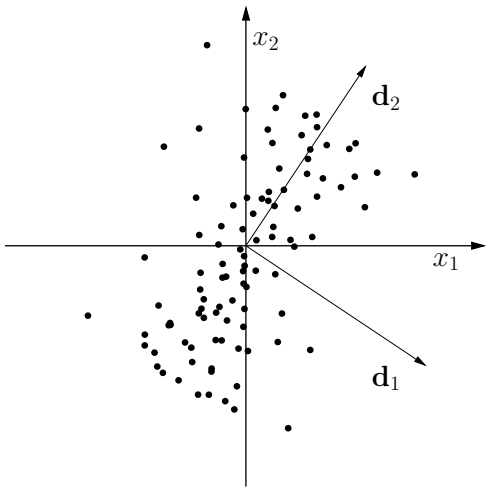

Figure 2: A two-dimensional orthogonal mixture of the two sources given above.
© CC BY-SA 4.0

but this is not a standard term. They do not indicate the mixing of all sources on one component of the data vector, thus calling them mixture vectors would be misleading. Notice also that the fact that the data is concentrated around $\mathbf{d_2}$ is a consequence of s_1 (not s_2!) being concentrated around zero.

Since the vectors $\mathbf{d}_i$ are orthogonal, extracting the sources from the mixture can be done by mulitplying with these vectors, for instance

$$\begin{aligned} \lozenge \quad y_1 &:= \mathbf{d}_1^T \mathbf{x} && (5)\\ \lozenge &\overset{(4)}{=} \mathbf{d}_1^T (s_1 \mathbf{d}_1 + s_2 \mathbf{d}_2) && (6)\\ \lozenge &= s_1 \underbrace{\mathbf{d}_1^T \mathbf{d}_1}_{=1} + s_2 \underbrace{\mathbf{d}_1^T \mathbf{d}_2}_{=0} && (7)\\ \lozenge &= s_1 \,, && (8) \end{aligned}$$

and likewise for y_2.

The unmixing matrix therefore is

$$\mathbf{U} := \begin{pmatrix} \mathbf{d}_1^T \\ \mathbf{d}_2^T \end{pmatrix}, \qquad (9)$$

since

$$\mathbf{UM} \overset{(9,4)}{=} \begin{pmatrix} \mathbf{d}_1^T \\ \mathbf{d}_2^T \end{pmatrix} \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 \end{pmatrix} \tag{10}$$

$$= \begin{pmatrix} \mathbf{d}_1^T\mathbf{d}_1 & \mathbf{d}_1^T\mathbf{d}_2 \\ \mathbf{d}_2^T\mathbf{d}_1 & \mathbf{d}_2^T\mathbf{d}_2 \end{pmatrix} \tag{11}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{12}$$

Example 2: If the distribution vectors are not orthogonal, matters become a bit more complicated, see figure 3. **It is somewhat counterintuitive that** now **the vectors $\mathbf{e}_i$ to extract the sources from the**

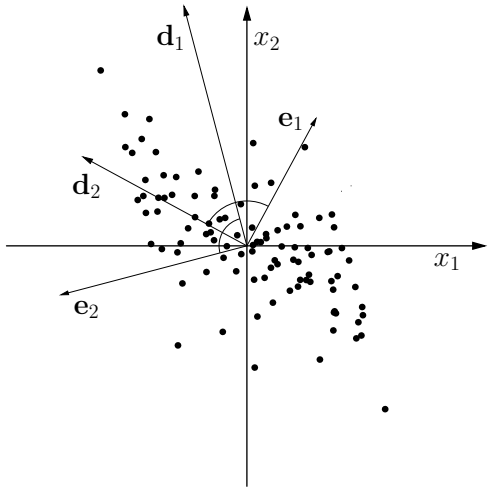

Figure 3: A two-dimensional non-orthogonal mixture of the two sources given above.
© CC BY-SA 4.0

mixture do not have to point in the direction of the corresponding distribution vectors but rather must be orthogonal to all other distribution vectors, so that

$$\Diamond \quad \mathbf{e}_i^T\mathbf{d}_j = \delta_{ij}\,. \tag{13}$$

Notice that in the figure $\mathbf{e}_1$ has the same angle to $\mathbf{d}_1$ as $\mathbf{e}_2$ has to $\mathbf{d}_2$ and that $\mathbf{e}_1$ must be shorter than $\mathbf{e}_2$, because $\mathbf{d}_1$ is longer than $\mathbf{d}_2$, to keep the inner products $\mathbf{e}_i^T\mathbf{d}_i$ equal.

With the *extraction vectors* (D: Extraktionsvektoren) $\mathbf{e}_i$ we get the unmixing matrix

$$\mathbf{U} := \begin{pmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{pmatrix} \tag{14}$$

and verify

$$\mathbf{UM} \overset{(14,4)}{=} \begin{pmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{pmatrix} \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 \end{pmatrix} \tag{15}$$

$$= \begin{pmatrix} \mathbf{e}_1^T\mathbf{d}_1 & \mathbf{e}_1^T\mathbf{d}_2 \\ \mathbf{e}_2^T\mathbf{d}_1 & \mathbf{e}_2^T\mathbf{d}_2 \end{pmatrix} \tag{16}$$

$$\overset{(13)}{=} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{17}$$

1.2 How to find the unmixing matrix?

So far we have only derived unmixing matrices if the mixing matrix was known. But we have not said anything about how to find the unmixing matrix if only the data is given and nothing is known about the mixing (or

the sources) apart from it being linear. So **we need some statistical criteria to judge whether an unmixing matrix is good or not.** There are two fundamental approaches, either one compares the joint pdf of the unmixed data with the product of its marginals, or one maximizes the non-Gaussianity of the marginals.

Make the output signal components statistically independent: A fundamental assumption of ICA is that the data is a linear mixture of statistically independent sources. If we unmix the data the resulting output signal components should therefore again be statistically independent. Thus a possible criterion for whether the unmixing is good or not is whether

$$p_{\mathbf{y}}(y_1, y_2) \quad = \quad p_{y_1}(y_1)\, p_{y_2}(y_2)\,. \tag{18}$$

or not. In practice one measures the difference between the joint distribution and the product of marginals and tries to minimize it. For instance **one can use the *Kullback-Leibler divergence* between $p_{\mathbf{y}}(y_1, y_2)$ and $p_{y_1}(y_1)\, p_{y_2}(y_2)$.** In this lecture **I will focus on cross-cumulants as a measure of statistical independence.**

This approach necessarily optimizes the unmixing matrix as a whole.

Make the output signal components non-Gaussian: Another approach is based on the observation that, **if you mix two sources, the mixture tends to be more Gaussian than the sources.** Applied over and over again this culminates in the central limit theorem of statistics, which basically states that a mixture of infinitely many variables has a Gaussian distribution. **Turning this argument around** it might be a good strategy to **search for output signal components that are as different from a Gaussian as possible.** This are then most likely sources. To measure non-Gaussianity, one often uses *kurtosis* (D: Kurtosis), see below.

This approach permits to extract one source after the other. One then typically first extracts the most non-Gaussian signal, eliminates the corresponding dimension from the data and then finds the second non-Gaussian signal.

1.3 Sources can only be recovered up to permutation and rescaling

It is clear that if y_1 and y_2 are statistically independent, $3y_2$ and $0.5y_1$ are also statistcally independent. Thus, there is no way to tell the order of the sources and their scale. A similar argument holds for the approach based on non-Gaussianity. Thus, the **sources can principally be only recovered up to a permutation and scaling.**

1.4 Whiten the data first

The least one can expect from the estimated sources is that they are uncorrelated. To fix the arbitrary scaling factor, it is also common to require the output data components to have unit variance (zero mean is assumed in any case). It is therefore a good idea to **whiten or sphere the data first, because then** we know that the data projected onto any normalized vector has zero mean and unit variance and that the data projected onto two orthogonal vectors are uncorrelated. Thus, **the unmixing matrix must be orthogonal** and the unmixing problem reduces to finding the right rotation, which is still difficult enough. Intuitively, whitening brings us from the situation in figure 3 to that in figure 2.

1.5 A generic ICA algorithm

Summarizing what we have learned so far, **a generic ICA algorithm is** conceptually relatively simple and works as follows:

1. **remove the mean and whiten the data,**
2. **rotate the data such that either**
   (a) **the output signal components are as statistically independent as possible, or**
   (b) **the output signal components are most non-Gaussian.**

2 Formalism based on cumulants

There is a whole zoo of different ICA algorithms. I focus here on a method based on higher-order cumulants.

2.1 Moments and cumulants

Moments and cumulants are a convenient way of describing statistical distributions. If all moments or all cumulants are known, the distribution is uniquely determined. Thus, moments and cumulants are equivalent descriptions of a distribution, although one usually only uses the lower moments or cumulants. If $\langle\cdot\rangle$ indicates the average over all data points, then the **moments of some vectorial data y** written as single components **are defined as**

$$\begin{aligned}
\lozenge \quad \text{first moment} \quad & \langle y_i \rangle && (19)\\
\lozenge \quad \text{second moment} \quad & \langle y_i y_j \rangle && (20)\\
\lozenge \quad \text{third moment} \quad & \langle y_i y_j y_k \rangle && (21)\\
\lozenge \quad \text{fourth moment} \quad & \langle y_i y_j y_k y_l \rangle && (22)\\
\lozenge \quad \text{higher moments} \quad & \ldots .
\end{aligned}$$

A disadvantage of moments is that higher moments contain information that can already be expected from lower moments, for instance

$$\lozenge \quad \langle y_i y_j \rangle \quad = \quad \langle y_i \rangle \langle y_j \rangle + ? \,, \tag{23}$$

or for zero mean data

$$\lozenge \quad \langle y_i y_j y_k y_l \rangle \quad = \quad \langle y_i y_j \rangle \langle y_k y_l \rangle + \langle y_i y_k \rangle \langle y_j y_l \rangle + \langle y_i y_l \rangle \langle y_j y_k \rangle + ? \,. \tag{24}$$

It would be nice to have a definition for only the part that is not known yet, much like the variance for the second moment of a scalar variable,

$$\lozenge \quad \left\langle (y - \langle y \rangle)^2 \right\rangle \quad = \quad \langle yy \rangle - \langle y \rangle \langle y \rangle \;, \tag{25}$$

where one has removed the influence of the mean from the second moment. This idea can be generalized to mixed and higher moments. **If one simply subtracts off from the higher moments what one would expect from the lower moments already, one gets corrected moments that are called *cumulants*** (D: Kumulanten)**. For** simplicity we assume **zero mean data,** then

$$\begin{aligned}
\lozenge \quad C_i \quad &:= \quad \langle y_i \rangle \stackrel{!}{=} 0 \,, && (26)\\
\lozenge \quad C_{ij} \quad &:= \quad \langle y_i y_j \rangle \;, && (27)\\
\lozenge \quad C_{ijk} \quad &:= \quad \langle y_i y_j y_k \rangle \;, && (28)\\
\lozenge \quad C_{ijkl} \quad &:= \quad \langle y_i y_j y_k y_l \rangle - \langle y_i y_j \rangle \langle y_k y_l \rangle - \langle y_i y_k \rangle \langle y_j y_l \rangle - \langle y_i y_l \rangle \langle y_j y_k \rangle \;. && (29)
\end{aligned}$$

Notice that there are no terms to be subtracted off from C_{ij} and C_{ijk} due to the zero-mean constraint. Any term that one might extract would have a factor of the form $\langle y_i \rangle$ included, which is zero.

For scalar variables these four cumulants have nice intuitive interpretations. We know already that C_i **and** C_{ii} **are the mean and variance** of the data, respectively. C_{iii} **(often normalized by** $C_{ii}^{3/2}$**) is the *skewness*** (D: Schiefe) of the distribution, which indicates how much tilted to the left or to the right the distribution is. C_{iiii} **(often normalized by** C_{ii}^2**) is the *kurtosis*** (D: Kurtosis) of the distribution, which indicates how peaky the distribution is. A uniform distribution has negative kurtosis and is called sub-Gaussian; a peaky distribution with heavy tails has positive kurtosis and is called super-Gaussian. A Gaussian distribution has zero kurtosis, and also all its other cumulants higher than the second cumulant vanish.

Cumulants have the important property that if you add two (or more) statistically independent variables, the cumulants of the sum equals the sum of the cumulants of the variables.

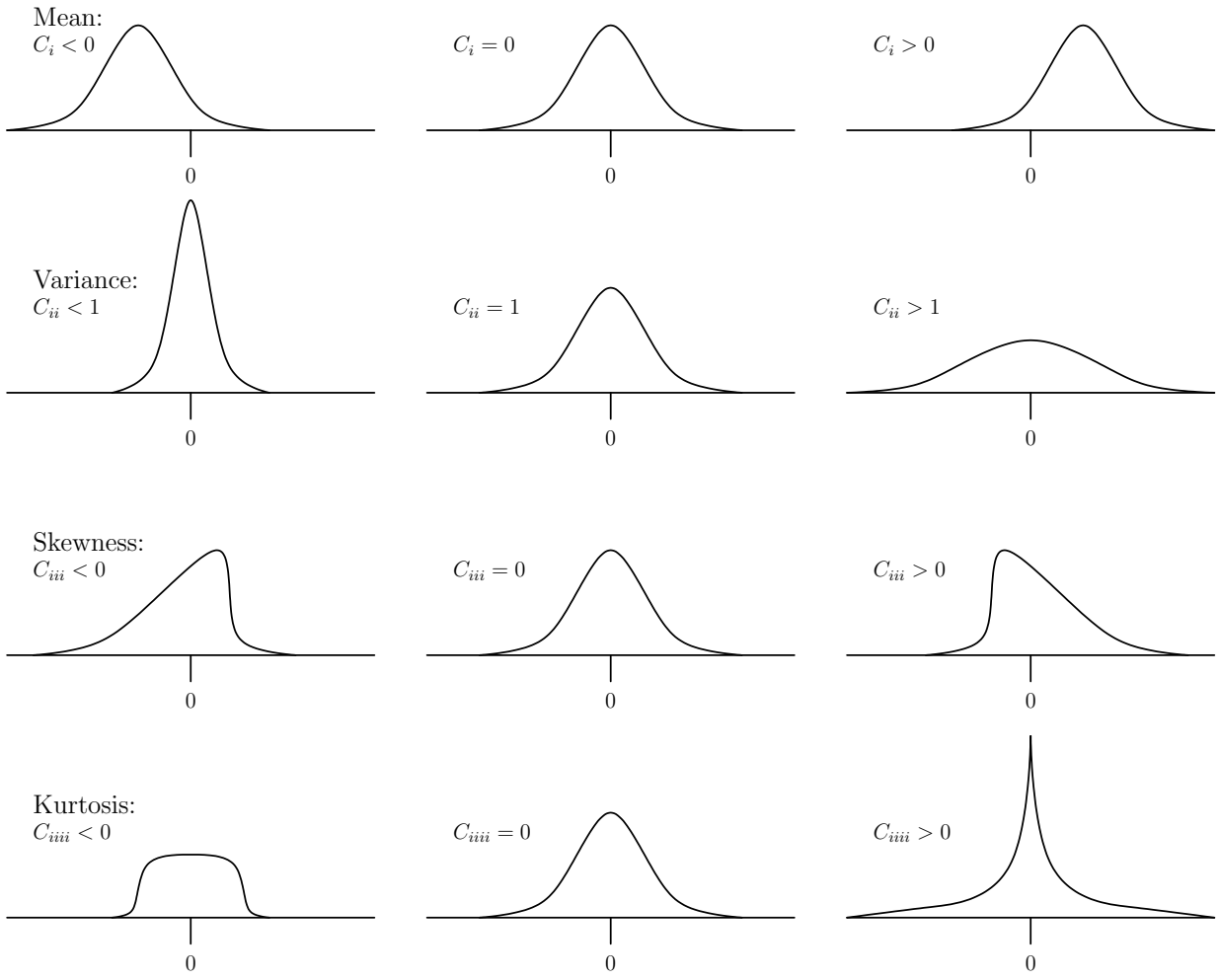

Figure 4: Illustration of the effect of the first four cumulants in one variable, i.e. mean, variance, skewness, and kurtosis.
© CC BY-SA 4.0

2.2 Cross-cumulants of statistically independent components are zero

First notice that **if the random variables of a higher moment can be split into two statistically independent groups, then the moment can be written as a product of the two lower moments of the groups.** For instance, if y_i and y_j are statistically independent of y_k, then

$$\begin{aligned}
\lozenge \quad \langle y_i y_j y_k \rangle &= \int_{y_i}\int_{y_j}\int_{y_k} y_i y_j y_k \, p(y_i, y_j, y_k) \, \mathrm{d}y_k \, \mathrm{d}y_j \, \mathrm{d}y_i && (30)\\
\lozenge &= \int_{y_i}\int_{y_j}\int_{y_k} y_i y_j y_k \, p(y_i, y_j) \, p(y_k) \, \mathrm{d}y_k \, \mathrm{d}y_j \, \mathrm{d}y_i \quad \text{(due to statistical independence)} && (31)\\
\lozenge &= \int_{y_i}\int_{y_j} y_i y_j \, p(y_i, y_j) \, \mathrm{d}y_j \, \mathrm{d}y_i \cdot \int_{y_k} y_k \, p(y_k) \, \mathrm{d}y_k && (32)\\
\lozenge &= \langle y_i y_j \rangle \, \langle y_k \rangle \; . && (33)
\end{aligned}$$

This has an important implication for cumulants of statistically independent variables. We have argued above that a cumulant can be interpreted as the corresponding moment of identical structure minus all what can be expected from lower-order moments (or cumulants) already. If the random variables can be split into two statistically independent groups then the corresponding moment can be written as a product of two lower moments, which means that it can be completely predicted from these lower moments. As a consequence

the cumulant is zero, because there is nothing left, that could not be predicted. Thus we can state (without proof) that **if a set of random variables are statistically independent then all cross-cumulants vanish.** For instance, for statistically independent random variables y_i and y_j with zero mean we get

$$\begin{aligned}
\lozenge \quad C_{iij} &= \langle y_i y_i y_j \rangle && (34)\\
\lozenge &\overset{(33)}{=} \langle y_i y_i \rangle \underbrace{\langle y_j \rangle}_{=0} \quad \text{(since } y_i \text{ and } y_j \text{ are statistically independent)} && (35)\\
\lozenge &= 0 \quad \text{(since the data is zero-mean)}\,, && (36)\\
\lozenge \quad C_{iijj} &= \langle y_i y_i y_j y_j \rangle - \langle y_i y_i \rangle \langle y_j y_j \rangle - \langle y_i y_j \rangle \langle y_i y_j \rangle - \langle y_i y_j \rangle \langle y_i y_j \rangle && (37)\\
\lozenge &\overset{(33)}{=} \langle y_i y_i \rangle \langle y_j y_j \rangle - \langle y_i y_i \rangle \langle y_j y_j \rangle - 2 \underbrace{\langle y_i \rangle}_{=0} \langle y_j \rangle \langle y_i \rangle \langle y_j \rangle && (38)\\
&\qquad \text{(since } y_i \text{ and } y_j \text{ are statistically independent)}\\
\lozenge &= 0 \quad \text{(since the data is zero-mean)}\,. && (39)
\end{aligned}$$

LECTURE 2/2

2.3 Components with zero cross-cumulants are statistically independent

The converse is also true (again without proof). **If all cross-cumulants vanish then the random variables are statistically independent.** Notice that pairwise statistical independence does not generally suffice for the overall statistical independence of the variables. Consider, for instance, the three binary varibles A, B, and $C = A \operatorname{xor} B$.

Thus, the cross-cumulants can be used to measure the statistical dependence between the output signal components in ICA. Of course, one cannot use all cross-cumulants, but for instance one can require

$$\begin{aligned}
\blacklozenge \quad C_{ij} &= \delta_{ij} \quad \text{(unit covariance matrix)}\,, && (40)\\
\lozenge \quad \text{and} \quad \sum_{ijk \neq iii} C_{ijk}^2 & \quad \text{minimal}, && (41)\\
\blacklozenge \quad \text{or} \quad \sum_{ijkl \neq iiii} C_{ijkl}^2 & \quad \text{minimal}, && (42)\\
\text{or} \quad \sum_{ijk \neq iii} C_{ijk}^2 + \sum_{ijkl \neq iiii} C_{ijkl}^2 & \quad \text{minimal}. && (43)
\end{aligned}$$

Usually one uses the fourth-order cumulants because signals are often symmetric and then the third-order cumulants vanish in any case. Assuming we disregard third-order cumulants the optimiziation can be stated as follows: Given some multi-dimensional input data $\mathbf{x}$, find the matrix $\mathbf{U}$ that produces output data

$$\mathbf{y} = \mathbf{U}\mathbf{x} \tag{44}$$

that minimize (42) under the constraint (40). The constraint is trivial to fulfill by whitening the data. Once the data has been whitened with some matrix $\mathbf{W}$ to obtain the whitened data $\hat{\mathbf{x}}$, the only transformation that is left to do is a rotation by a rotation matrix $\mathbf{R}$, as we have discussed earlier. Thus, we have

$$\begin{aligned}
\mathbf{y} &= \mathbf{R}\hat{\mathbf{x}} && (45)\\
&= \underbrace{\mathbf{R}\mathbf{W}}_{=\mathbf{U}}\mathbf{x} && (46)
\end{aligned}$$

2.4 Rotated cumulants

Let $\hat{\mathbf{x}}$ and $\mathbf{y}$ indicate the whitened and the rotated data, respectively, and $C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon}$ and $C^{\mathbf{y}}_{ijkl}$ the corresponding cumulants. Then we write the cumulants of $\mathbf{y}$ in terms of the cumulants of $\hat{\mathbf{x}}$.

$$\begin{aligned}
\lozenge \quad C^{\mathbf{y}}_{ijkl} &= \langle y_i y_j y_k y_l\rangle - \langle y_i y_j\rangle \langle y_k y_l\rangle - \langle y_i y_k\rangle \langle y_j y_l\rangle - \langle y_i y_l\rangle \langle y_j y_k\rangle && (47)\\
\lozenge &= \left\langle \sum_\alpha R_{i\alpha}\hat{x}_\alpha \sum_\beta R_{j\beta}\hat{x}_\beta \sum_\gamma R_{k\gamma}\hat{x}_\gamma \sum_\epsilon R_{l\epsilon}\hat{x}_\epsilon \right\rangle \\
&\quad - \left\langle \sum_\alpha R_{i\alpha}\hat{x}_\alpha \sum_\beta R_{j\beta}\hat{x}_\beta \right\rangle \left\langle \sum_\gamma R_{k\gamma}\hat{x}_\gamma \sum_\epsilon R_{l\epsilon}\hat{x}_\epsilon \right\rangle - ... - ... && (48)\\
&\quad \text{(since (45)} \Leftrightarrow y_i = \sum_\alpha R_{i\alpha}\hat{x}_\alpha) \\
\lozenge &= \sum_{\alpha\beta\gamma\epsilon} R_{i\alpha}R_{j\beta}R_{k\gamma}R_{l\epsilon} \underbrace{\left(\langle \hat{x}_\alpha\hat{x}_\beta\hat{x}_\gamma\hat{x}_\epsilon\rangle - \langle \hat{x}_\alpha\hat{x}_\beta\rangle\langle \hat{x}_\gamma\hat{x}_\epsilon\rangle - ... - ...\right)}_{C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon}} && (49)\\
\lozenge &= \sum_{\alpha\beta\gamma\epsilon} R_{i\alpha}R_{j\beta}R_{k\gamma}R_{l\epsilon}\, C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon}\,. && (50)
\end{aligned}$$

The fact that the new cumulants are simply linear combinations of the old cumulants reflects the multilinearity of cumulants.

2.5 Contrast function

It is interesting to note that **the square sum over all cumulants of a given order does not change under a rotation,** as can be easily verified for fourth order.

$$\begin{aligned}
\lozenge \quad \sum_{ijkl}\left(C^{\mathbf{y}}_{ijkl}\right)^2 &\overset{(50)}{=} \sum_{ijkl}\left(\sum_{\alpha\beta\gamma\epsilon} R_{i\alpha}R_{j\beta}R_{k\gamma}R_{l\epsilon}\, C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon}\right)^2 && (51)\\
\lozenge &= \sum_{ijkl}\left(\sum_{\alpha\beta\gamma\epsilon} R_{i\alpha}R_{j\beta}R_{k\gamma}R_{l\epsilon}\, C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon}\right)\left(\sum_{\kappa\lambda\mu\nu} R_{i\kappa}R_{j\lambda}R_{k\mu}R_{l\nu}\, C^{\hat{\mathbf{x}}}_{\kappa\lambda\mu\nu}\right) && (52)\\
\lozenge &= \sum_{\alpha\beta\gamma\epsilon}\sum_{\kappa\lambda\mu\nu} C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon}C^{\hat{\mathbf{x}}}_{\kappa\lambda\mu\nu} \underbrace{\sum_i R_{i\alpha}R_{i\kappa}}_{\delta_{\alpha\kappa}} \underbrace{\sum_j R_{j\beta}R_{j\lambda}}_{\delta_{\beta\lambda}} \underbrace{\sum_k R_{k\gamma}R_{k\mu}}_{\delta_{\gamma\mu}} \underbrace{\sum_l R_{l\epsilon}R_{l\nu}}_{\delta_{\epsilon\nu}} && (53)\\
\lozenge &= \sum_{\alpha\beta\gamma\epsilon}\left(C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon}\right)^2\,. && (54)
\end{aligned}$$

This implies that instead of minimizing the square sum over all cross-cumulants of order four it is equivalent to maximize the square sum over the kurtosis of all components, i.e.

$$\begin{aligned}
\blacklozenge \quad &\text{minimize} \quad \sum_{ijkl\neq iiii} C^2_{ijkl} && (55)\\
\blacklozenge \quad \Longleftrightarrow \quad &\text{maximize} \quad \Psi_4 := \sum_i C^2_{iiii}\,, && (56)
\end{aligned}$$

since the sum over these two sums is constant. This is one way of formalizing our intuition that making all components statistically indendent is equivalent to making them as non-Gaussian as possible. Maximizing Ψ_4 is obviously much easier than minimizing the square sum over all cross cumulants. Thus, we will use that as our objective function or contrast function.

A corresponding relationship also holds for third-order cumulants. However, if the sources are symmetric, which they might well be, then their skewness is zero in any case and maximizing their square sum is of little use. One therefore usually considers fourth- rather than third-order cumulants for ICA. Considering both simulataneously, however, might even be better.

2.6 Givens-rotations

A rotation matrix in 2D is given by

$$\lozenge \quad \mathbf{R} \quad = \quad \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} . \tag{57}$$

In higher dimensions a rotation matrix can be quite complex. To keep things simple, we use so-called ***Givens rotations*** (D: Givens-Rotation), which **are defined as a rotation within the 2D subspace spanned by two axes n and m.** For instance a rotation within the plane spanned by the second and the fourth axis $(n = 2, m = 4)$ of a four-dimensional space is given by

$$\blacklozenge \quad \mathbf{R} \quad = \quad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\phi & 0 & -\sin\phi \\ 0 & 0 & 1 & 0 \\ 0 & \sin\phi & 0 & \cos\phi \end{pmatrix} . \tag{58}$$

It can be shown that **any general rotation**, i.e. any orthogonal matrix with positive determinant, **can be written as a product of Givens-rotations.** Thus, we can find the general rotation matrix of the ICA problem by applying a series of Givens rotations and each time improve the objective function (42) a bit.

2.7 Optimizing the contrast function

We have argued above that ICA can be performed by repeatedly applying a Givens-rotation to the whitened data. Each time two axes (n and m) that define the rotation plane are selected at random and then the rotation angle is optimized to maximize the value of the contrast function. **By applying enough** of such **Givens-rotations, the algorithm should eventually converge to the globally optimal solution** (although we have no guarantee for that). Writing the constrast function as a function of the rotation angle ϕ yields

$$\lozenge \quad \Psi_4(\phi) \quad := \quad \sum_i \left(C^{\mathbf{y}}_{iiii}\right)^2 \tag{59}$$

$$\lozenge \quad \overset{(50)}{=} \quad \sum_i \left(\sum_{\alpha\beta\gamma\epsilon} R_{i\alpha} R_{i\beta} R_{i\gamma} R_{i\epsilon}\, C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon} \right)^2 \tag{60}$$

$$\lozenge \quad = \quad K + \sum_{i=n,m} \left(\sum_{\alpha\beta\gamma\epsilon=n,m} R_{i\alpha} R_{i\beta} R_{i\gamma} R_{i\epsilon}\, C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon} \right)^2 . \tag{61}$$

For $i \neq n, m$ the entries of the rotation matrix are $R_{i\omega} = \delta_{i\omega}$ for $\omega = \alpha, \beta, \gamma, \epsilon$ and do not depend on ϕ. Thus, these terms are constant and are contained in K.

For $i = n, m$ the entries of the rotation matrix are $R_{i\omega} = 0$ for $\omega \neq n, m$ and $R_{i\omega} \in \{\cos(\phi), \pm\sin(\phi)\}$ for $\omega = n, m$ with $\omega = \alpha, \beta, \gamma, \epsilon$. Thus, the sums over $\alpha, \beta, \gamma, \epsilon$ can be restricted to n, m and each resulting term contains eight cosine- or sine-factors, because $R_{i\omega} \in \{\cos(\phi), \pm\sin(\phi)\}$ and due to the squaring. Thus, Ψ_4 can always be written as

$$\Psi_4(\phi) \quad = \quad K + \sum_{p=0}^{8} k'_p \cos(\phi)^{(8-p)} \sin(\phi)^p . \tag{62}$$

The old cumulants $C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon}$ are all contained in the constants k'_p.

We can simplify $\Psi_4(\phi)$ even further with the following two considerations: Firstly, a rotation by multiples of 90° should have no effect on the contrast function, because that would only flip or exchange the components of $\mathbf{y}$. Thus, Ψ_4 must have a 90° periodicity and can always be written like

$$\Psi_4(\phi) \quad = \quad A_0 + A_4 \cos(4\phi + \phi_4) + A_8 \cos(8\phi + \phi_8) + A_{12} \cos(12\phi + \phi_{12}) + A_{16} ... \tag{63}$$

Secondly, products of two sin- and cos-functions produce constants and frequency doubled terms, e.g.

$$\sin(\phi)^2 = (1-\cos(2\phi))/2 \quad (64)$$
$$\cos(\phi)^2 = (1+\cos(2\phi))/2 \quad (65)$$
$$\sin(\phi)\cos(\phi) = \sin(2\phi)/2\,. \quad (66)$$

Products of eight sin- and cos-functions produce at most eightfold frequencies (three time frequency doubling). This limits equation (63) to terms up to 8. Thus, **finally we get the following contrast function for one Givens-rotation:**

$$\Diamond \quad \Psi_4(\phi) = A_0 + A_4\cos(4\phi+\phi_4) + A_8\cos(8\phi+\phi_8)\,. \quad (67)$$

Again the constants contain the old cumulants $C^{\hat{\mathbf{x}}}_{\alpha\beta\gamma\epsilon}$ in some complicated but computable form.

It is relatively simple to find the maximium of this function once the constants are known.

2.8 The algorithm

A cumulant based algorithm for independent component analysis could now look as follows:

1. Whiten the data $\mathbf{x}$ with whitening matrix $\mathbf{W}$ and create $\mathbf{y} = \mathbf{W}\mathbf{x}$.
2. Select two axes/variables y_n and y_m randomly.
3. Rotate $\mathbf{y}$ in the plane spanned by y_n and y_m such that $\Psi_4(\phi)$ is maximized. This leads to a new $\mathbf{y}$.
4. Go to step 2 unless a suitable convergence criterion is fulfilled, for instance, the rotation angle has been smaller than some threshold for the last 1000 iterations.
5. Stop.

3 Other resources

Numbers in square brackets indicate sections of these lecture notes to which the corresponding item is related.

3.1 Written material

- ICA on Wikipedia
  https://en.wikipedia.org/wiki/Independent_component_analysis

3.2 Videos

- Abstract conceptual introduction to ICA from Georgia Tech
  1. ICA objective
     https://www.youtube.com/watch?v=2WY7wCghSVI (2:13)
  2. Mixing and unmixing / blind source separation / cocktail party problem
     https://www.youtube.com/watch?v=wIlrddNbXDo (4:17)
  3. How to formulate this mathematically
     https://www.youtube.com/watch?v=pSwRO5d266I (5:15)
  4. Application to artificially mixed sound
     https://www.youtube.com/watch?v=T0HP9cxri0A (3:25)
  5. Quiz Questions PCA vs ICA
     https://www.youtube.com/watch?v=TDW0vMz_3ag (0:32)

6. Quiz Answers PCA vs ICA with further comments
   https://www.youtube.com/watch?v=SjM2Qm7N9CU (10:04)
7. More on differences between PCA and ICA
   https://www.youtube.com/watch?v=e4woe8GRjEI (7:17)
   - 01:23–01:34 I find this statement confusing. Transposing a matrix is a very different operation than rotating data, and I find it highly non-trivial that PCA works also on the transposed data matrix, see the section on singular value decomposition in the lecture notes on PCA.
   - 02:49–02:59 Hm, this is not quite true. Either the mean has been removed from the data, which is normally the case, then the average face is at the origin and cannot be extracted with PCA. Or the mean has not been removed, then it is typically the first eigenvector (not the second), that goes through the mean, although not exactly, as one can see in one of the python exercises. There are marked differences around the eyes.

- Lecture on ICA based on cumulants by Santosh Vempala, Georgia Institute of Technology.
  https://www.youtube.com/watch?v=KSIA908KNiw (52:22)
  - 00:45–06:31 Introducing the ICA problem definition
  - 06:31–07:50 Would PCA solve the problem? No!
  - 07:50–15:04 Formulation of a deflation algorithm based on cumulants
  - 15:04–20:35 (Reformulation with tensor equations)
  - 20:35– (Two problems in solving the ICA problem)
    - 21:32– (What if the second and forth order cumulants are zero?)
      - 25:26– (Algorithm: Fourier PCA)
    - ...
  - ...

3.3 Software

- FastICA in scikit-learn, a python library for machine learning
  http://scikit-learn.org/stable/modules/decomposition.html#independent-component-analysis-ica
- Examples using FastICA in scikit-learn
  http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html#sklearn.decomposition.FastICA

3.4 Exercises

- Analytical exercises by Laurenz Wiskott
  https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/IndependentComponentAnalysis-ExercisesPublic.pdf
  https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/IndependentComponentAnalysis-SolutionsPublic.pdf
- Python exercises by Laurenz Wiskott
  https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/IndependentComponentAnalysis-PythonExercisesP
  zip
  https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/IndependentComponentAnalysis-PythonSolutionsP
  zip

References

Blaschke, T. and Wiskott, L. (2004). CuBICA: Independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. *IEEE Trans. on Signal Processing*, 52(5):1250–1256. 2

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons. 2