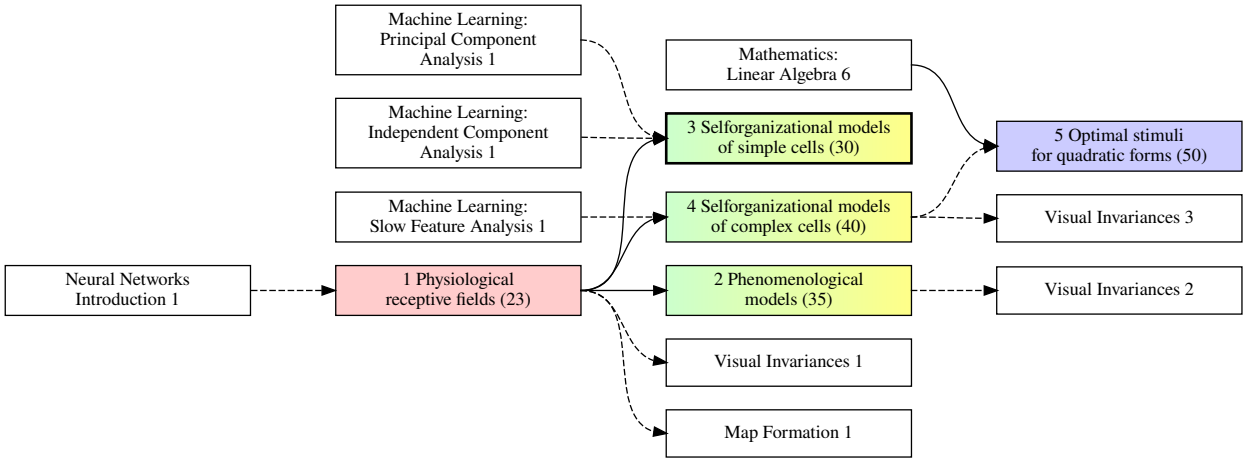
Visual Receptive Fields

— Lecture Notes —

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

27 March 2021

— Summary —



3 Selforganizational models of simple cells (→ slides) are usually linear and based on various objectives, which yield certain receptive fields as optimal solutions. If these receptive fields are similar to the physiological ones, this provides support for the objective used to be the underlying reason for the shape of the physiological receptive fields. This section shows that sparseness and statistical independence are plausible objectives, while linear compression is not.

Contents

3 Selforganizational models of simple cells (→ slides)

While the what-question can be addressed rather directly, because one can measure the responses of the cells and make a model of the responses as a function of the stimulus, the why-question is more difficult and needs to be adressed indirectly. One approach is to make a hypothesis about why the cells have developed their response properties, formulate that as an optimization problem, solve the optimization problem, and then see whether the result bares similarity with the physiological response properties. If it does, it supports the hypothesis, if not, it discredits the hypothesis. For simple cells we consider here the following three hypotheses:

- Simple cells are there to (linearly) compress the visual input.
- Simple cells are there to decompose the visual input into statistically independent components.
- Simple cells are there to yield a sparse representation of the visual input.

3.1 Principal component analysis does not lead to simple cells

Learning material:[1]

☐ 6 min video ~~3.1 Principal Component Analysis does not Lead to Simple Cells~~

☐ Text below

[1] Generic instruction: Consider the (possibly nested) list of resources like a horizontal tree with an invisible root on the very left, and decide from left to right what you want to select to work through. The invisible root node has to be selected. For any selected parent node all children nodes marked with ■ or ● are mandatory and have to be selected. Children nodes marked with □ or ○ are optional and may be selected in addition to get a better understanding of the material. If a parent node has no mandatory child, then at least one optional child has to be selected. Children marked with + provide additional voluntary material that can be safely ignored, typically going beyond the scope of the section. Children of non-selected parents may be ignored. ■ and □ indicate children that cover (almost) the whole material of the section. Missing content might then be indicated by struck through references to the corresponding learning objectives. Items tend to be ordered by precedence and/or recommended temporal order from top to bottom, assuming that you prefer to first watch a video before reading through lecture notes. If a detailed table of content for videos or lecture notes is given, references to learning objectives might be provided in green, 1:30 should be read as 1 min and 30 seconds, and 1'30 should be read as page 1 at about 30% of the page. Video times may be linked directly to the indicated position in the video, but be aware that the video might be downloaded anew each time you click on a time. Resources without author name are usually authored by Laurenz Wiskott and his team.

A common way to linearly compress data is *principal component anaylsis (PCA)* (D: Hauptkomponentenanalyse) (see Wiskott, 2016b, for an introduction). The data is considered as points in a vector space, and PCA finds an ordered set of orthogonal directions, called principal components (PC) (D: Hauptkomponenten), such that the variance of the data along the first PC (or projected onto the first PC) is maximal, along the second PC it is maximal under the constraint of being uncorrelated to the first one, along the third PC it is maximal under the constraint of being uncorrelated to the first and second one, ect. For optimal linear compression one keeps the first few principal components and discards the other. How many PCs to keep depends on several factors such as how much compression one needs and how much variance there is along

the individual PCs.

Principal Components of Natural Images

(Hancock, Baddeley, & Smith, 1992, Network 3(1):61–70, Fig. 1)

15 natural images of size 256×256 pixels.

20,000 random samples of size 64×64 pixels.

For each pixel the mean gray value over the 20,000 samples was removed.

The samples were windowed with a Gaussian with std. dev. 10 pixels.

Sanger's rule was applied to the samples.

25/56

An early hypothsis was that the purpose of simple cells is to compress images for further processing. Hancock et al. (1992) have tested this by taking 15 natural images of size 256×256 pixels, from these cutting out 20,000 random samples of size 64×64 pixels, removing the mean from each pixel across the 20,000 samples, windowing the samples with a Gaussian, and finally calculating the first principal components with Sanger's rule, which is a neural learning rule for performing PCA. The image patches of size 64 × 64 are cast into vectors by simply concatenating the rows (or columns) into a vector of length 4096, a transformation that can be easily inverted by rearranging the components of the vector back into a matrix.

Principal Components of Natural Images

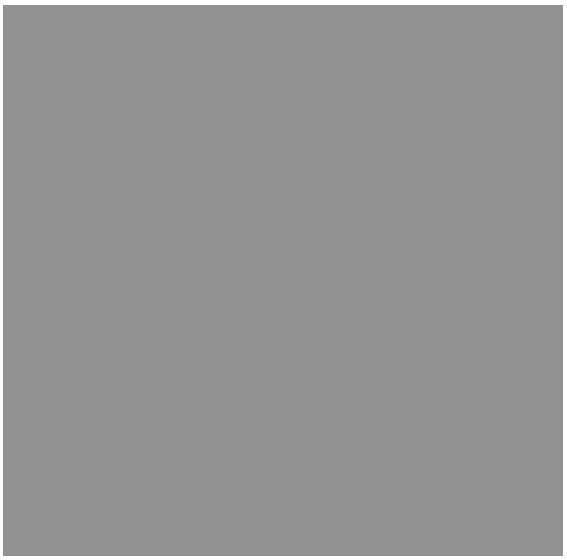
(Hancock, Baddeley, & Smith, 1992, Network 3(1):61–70, Fig. 1)

The first principal components resemble simple-cell receptive fields in the primary visual cortex, the later ones do not.

26/56

The first principal components extracted from natural image patches windowed with a Gaussian somewhat resemble simple cells (Hancock et al., 1992). However, later ones do not, and the Gaussian window plays an important role in making the filters look plausible at all.

Principal Components of Natural Images



(Olshausen & Field, 1996, Nature 381:607–9, Fig. 1)

27/56

Olshausen and Field (1996) have applied principal component analysis (PCA) to natural image patches of size 8×8 and have found filters as shown here.

One can understand this result, if one resorts to Fourier theory and considers the image patches as a linear superposition of sine waves of different frequency, orientation, and phase. Since natural images are known to have a $1/f^2$ power spectrum, i.e. low frequencies f are stronger and thus carry more variance, it is clear that the early principal components (PCs) should focus on low frequencies and the later ones on high frequencies. If one furthermore assumes that the statistics of natural images is translation and rotation invariant (which is at least approximately true), one can see that sine waves of different phase (related by translation) and orientation (related by rotation) but same frequency can be randomly mixed, since they carry identical variance. Taking this together yields the PCs shown here.

3.2 Sparseness leads to simple cells

Learning material:

- ☐ 14 min video 3.2 Sparseness Leads to Simple Cells
- ☐ Text below

Sparseness Principle

(Olshausen & Field, 2004, Curr. Opp. Neurobiol 14:481)

A sparse representation

- can reduce metabolic costs, because fewer units are active,
- can reduce wiring, because fewer units need to be connected,
- can be more robust, because units tend to be more binary,
- can simplify learning and processing, because relevant information is more localized,
- ...

28/56

Olshausen and Field (1996) have argued that the goal of sensory coding is to yield a *sparse* (D: spärliche(?)) representation. A sparse representation is one, where for any given input only few units are strongly active, all others are close to zero. This code might have various advantages for the brain.
The figure (Olshausen and Field, 2004) shows a non-sparse representation at the top and a sparse representation at the bottom.

Sparse Coding

Assumption: Images can be written as a superposition of basis functions,

$$I(\mathbf{x}) = \sum_i a_i \phi_i(\mathbf{x}) \,, \tag{1}$$

with fixed functions $\phi_i(\mathbf{x})$ and variable coefficients a_i.

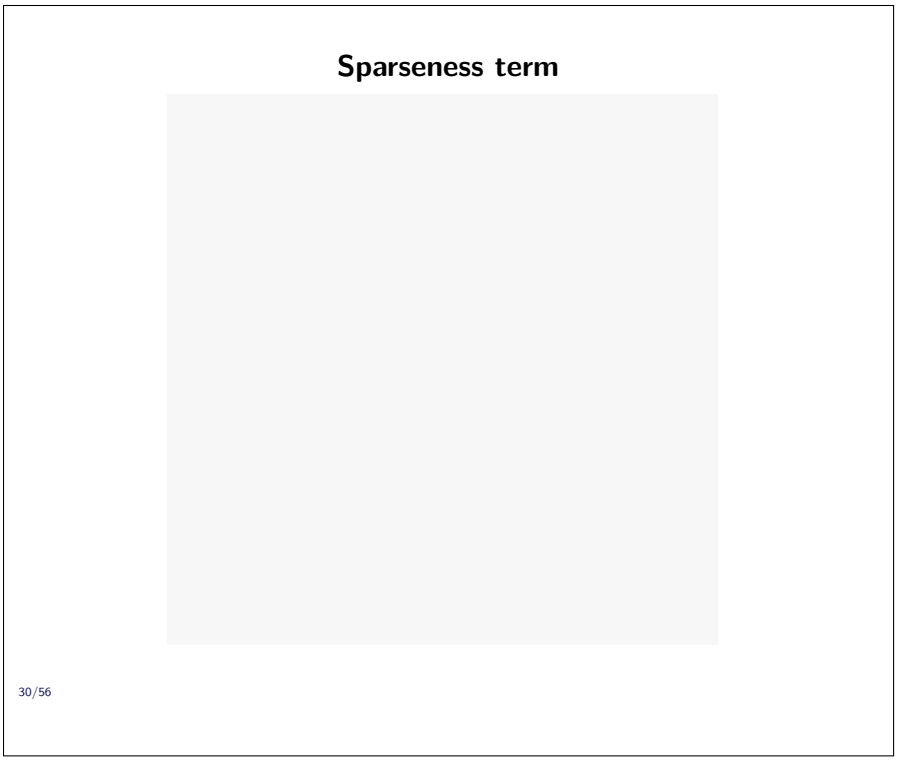
Objective: Choose the (probably normalized) functions such that the reconstruction error is small and the distribution of coefficients sparse, i.e.

$$\text{minimize} \quad E := \underbrace{\int_{\mathbf{x}} (I(\mathbf{x}) - \sum_i a_i \phi_i(\mathbf{x}))^2 \, \mathrm{d}^2\mathbf{x}}_{\text{reconstruction term}} + \lambda \underbrace{\sum_i |a_i|}_{\text{sparseness term}} \,. \tag{2}$$

(Olshausen & Field, 1996, Nature 381:607–9)

29/56

The model by Olshausen and Field (1996) assumes that images $I(\mathbf{x})$ can be represented by a linear superposition of some fixed basis functions $\phi_i(\mathbf{x})$, wich leads to the first term in the cost function E. The basis functions may be overcomplete, i.e. there may be more functions than pixels in the image, and non-orthogonal, which they must be in cast of an overcomplete set.
The weighting coefficients a_i vary from image to image and should be sparsely distributed, i.e. should be near zero most of the time and only occasionally have a large positive or negative value. The second term in the cost function E formalizes the sparseness objective.
An optimization procedure optimizes both, the basis functions across all images as well as the weighting coefficients for each image individually.



Consider the case where we want to represent a vector $\boldsymbol{I}$ (the image) as a linear combination of some basis vectors $\boldsymbol{\phi}_i$ with weighting factors a_i, i.e. $\boldsymbol{I} = \sum_i \boldsymbol{\phi}_i a_i$. If we combine the basis vectors in a matrix $\boldsymbol{\Phi} := (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_N)$ and the weighting factors in a vector $\boldsymbol{a} := (a_1, ..., a_N)^T$, we can write $\boldsymbol{I} = \boldsymbol{\Phi}\boldsymbol{a}$. With any orthogonal (rotation) matrix $\boldsymbol{U}$ we can define a new $\boldsymbol{\Phi}' := \boldsymbol{\Phi}\boldsymbol{U}^T$ and $\boldsymbol{a}' := \boldsymbol{U}\boldsymbol{a}$, so that the image is preserved,

$$\boldsymbol{\Phi}'\boldsymbol{a}' = \boldsymbol{\Phi}\underbrace{\boldsymbol{U}^T\boldsymbol{U}}_{=\mathbf{1}}\boldsymbol{a} = \boldsymbol{I}\,,$$

but the basis vectors as well as the weighting factors change. Thus, we can rotate the representation without compromising the qualitiy of the represented image, which leaves room to optimize sparseness in addition.

The figure illustrates with a dashed circle all the weight vectors with length 3, which can be realized by rotating one weight vector of same length. The solid lines represent the level lines of the sparseness term $|a_1| + |a_2|$. One can see that on the circle the points on the axes, namely $(0, 3), (3, 0), (0, -3), (-3, 0)$, have the smallest value for the sparseness term, which corresponds to the intuition that the coefficients should be either close to zero or large.

In the model (Olshausen and Field, 1996), the solution is not as clean, since the code is optimized for many images simultaneously. Also, some normalization must be imposed on either the basis functions or the weighting factors, because otherwise the latter could be made arbitrarily small while the former grow larger and larger.

Filters Generating a Sparse Code of Natural Images

(Olshausen & Field, 2004, Curr. Opp. Neurobiol 14:481, Fig. 1a)

31/56

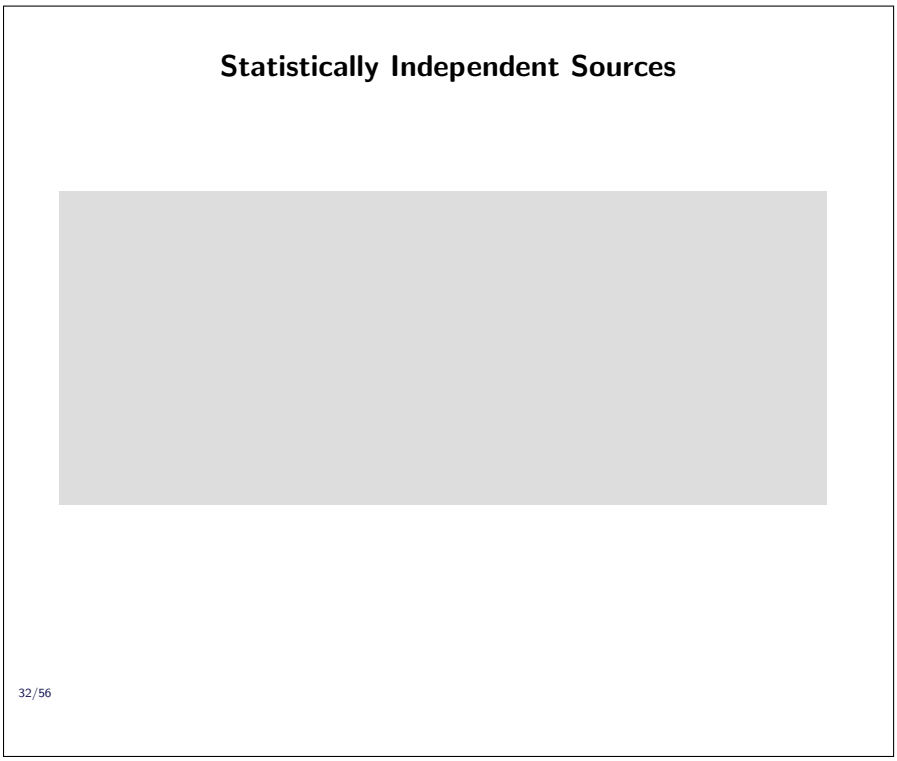
The filters obtained by optimizing the sparseness of the code in the model by Olshausen and Field (1996) resemble simple cell receptive fields fairly well (figure from Olshausen and Field, 2004).

3.3 Statistical independence leads to simple cells

Learning material:

- ☐ 8 min video ~~3.3 Statistical Independence Leads to Simple Cells~~
- ☐ Text below

For a more in depth introduction into independent component analysis see (Wiskott, 2016a).



Assume two stastically independent sources s_1 and s_2 are given, in this case sound sources (left). If one plots samples from the two sources in a common coordinate system such that one component always comes from one source and the other component from the other source, then one gets a two-dimensional data distribution with two statistically independent components. Please notice that the time structure of the signal is now gone and actually irrelevant for what follows.
Intuitively *statistical independence* (D: statistische Unabhängigkeit) means that knowing the value of one component does not tell you anything about the other component of that sample. Visually this roughly means that there may not be any diagonal structures in the plot.
Formaly statistical independence means that the joint probability density function (pdf) equals the product of its marginal pdfs $p(s_1, s_2) = p(s_1)p(s_2)$. This implies that if you cut through the distribution horizontally anywhere, you always get the same 1D curve (namely $p(s_2)$) just scaled differently (by $p(s_1)$), and the same holds for the vertical dimension.

Linear Blind Source Separation

Whitening can be done with PCA.

Rotation can be done based on the objective that the components y_i be statistically independent, here $p(y_1, y_2) = p(y_1)p(y_2)$.

33/56

If one takes samples $\boldsymbol{s}$ from two statistically independent sources and mixes them linearly with an invertible matrix $\boldsymbol{A}$, one gets a mixed signal $\boldsymbol{x}$. If one knew $\boldsymbol{A}$ it would be easy to unmix the data again. One would simply calculate the inverse of $\boldsymbol{A}$ and multiply the data vectors with it. However, even if $\boldsymbol{A}$ is unknown can one unmix the data, up to permutation and scaling, a process called *linear blind source separation*, 'blind' because neither the mixing matrix $\boldsymbol{A}$ nor the sources s_i are known (except that at most one may be Gaussian). The linear algorithm is usually referred to as *Independent Component Analysis (ICA)*.
The first step is whitening, with the argument that statistically independent components must at least be uncorrelated, and that is what whitening gives us. The second step is a rotation, because any skewing or stretching would ruin our whitening again. The rotation angle is dertermined such that some measure of statistical independence or non-Gaussianity is optimized.
It is interesting that making the individual components as non-Gaussian as possible is equivalent to making them as statistically independent as possible. The converse is known from the central limit theorem, if one mixes (adds) random variables, the resulting distribution gets more Gaussian.
The statistically independent components being extracted are all normalized to unit variance and their assignment to the components as well as their sign is arbitrary. This is why I stated above 'up to permutation and scaling'.

Independent Component Analysis on Natural Images

(Bell & Sejnowski, 1997, Vision Research 37:3327–38, Fig. 1)

$\mathbf{y} = \mathbf{Rx} = \mathbf{RAs}$ (with whitened $\mathbf{y}$, i.e. $\langle \mathbf{y}\mathbf{y}^T \rangle = \mathbf{I}$)

34/56

When applying ICA to natural images, the view is that each image itself is a mixture, i.e. a linear superposition, of some statistically independent sources in the real world, and the task of the visual system is to extract these underlying sources from the image (Bell and Sejnowski, 1997).

ICA-Filters for Natural Images

(Bell & Sejnowski, 1997, Vision Research 37:3327–38, Fig. 4)

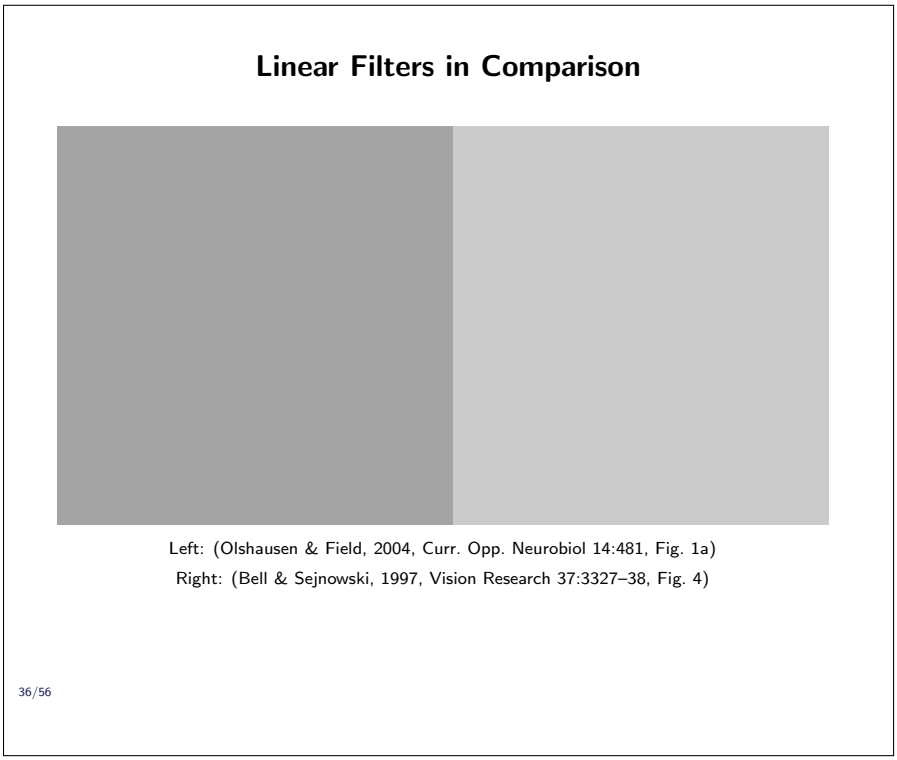
35/56

When one applies ICA to natural images, one gets filters that resemble simple cell receptive fields fairly well (Bell and Sejnowski, 1997).

3.4 Sparseness vs statistical independence

Learning material:

- ☐ 8 min video ~~3.4 Sparseness vs Statistical Independence~~
- ☐ Text below



The filters obtained by the sparseness objective (left) (Olshausen and Field, 2004) and by ICA (right) (Bell and Sejnowski, 1997) look very similar. The reason is that in the linear and complete case and if the underlying sources are sparse the two objectives are equivalent.

Relation Between Sparseness and Independence

37/56

The left figure shows two sparse and statistically independent sources plotted in a common graph with their joint pdf. If one rotates the joint pdf, thereby mixing the components, two things happen: (i) The individual components get less sparse; (ii) The components get more statistically dependent on each other. Thus optimizing sparseness as well as statistical independence both lead to an unmixing of the data; the objectives are equivalent. This would not be true if the sources were non-sparse to begin with.

Linear Models of Visual Receptive Fields - Summary

- Linear filters resulting from principal component analysis on natural images do not resemble simple cell receptive fields.
- Linear filters optimized for sparseness on natural images resemble simple cell receptive fields.
- Linear filters optimized for statistical independence on natural images also resemble simple cell receptive fields.
- Sparseness and statistical independence lead to similar results in linear systems if the underlying sources are sparse.

38/56

References

Bell, A. J. and Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Res.*, 37(23):3327–3338.

Hancock, P. J. B., Baddeley, R. J., and Smith, L. S. (1992). The principal components of natural images. *Network*, 3(1):61–70.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–7.

Wiskott, L. (2016a). Lecture notes on independent component analysis. Available at https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/index.html.

Wiskott, L. (2016b). Lecture notes on principal component analysis. Available at https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/index.html.