Tutorial Proposal for ICANN 2011 (June 14–17, 2011, in Espoo, Finland)

—

Cybernetics of Neuron Systems

Heikki Hyötyniemi
neocybernetics.com
heikki.hyotyniemi@tkk.fi

Tel. +358-9-50-3841626

Abstract

In this tutorial, a concise theory of the *cybernetic neuron system* is presented. Starting from very elementary local actions, powerful global level functionalities are found. For example, it turns out that a cybernetic neuron population implements *sparse subspace coding* of data in its *principal subspace.* Because the information presentation in the cybernetic model is *optimal,* further hypotheses can also be drawn: this novel approach may span the whole continuum from the elementary actions to the high-level cognitive functionalities, giving new intuitions on how emergence in complex systems in general can be attacked. As the theme of ICANN 2011 is "machine learning re-inspired by brain and cognition" this topic could be an inspiring contribution at the Conference; and, as Geoffrey Hinton will give a plenary talk, there would be synergy of approaches. — In the final presentation, there will be detailed case studies on how the methodology can be used to efficiently implement practical pattern recognition and regression tasks.

During the tutorial, the theoretical observations will be presented by *docent Heikki Hyötyniemi* and simulations and practical experiments will be presented by *Mr. Petri Lievonen,* both from Aalto University.

Contents

1 Introduction **3**

2 Properties of the information flow **4**
2.1 Neurons — just trying to prosper 4
2.2 Engineering of emergence 5
2.3 Interplay among emergent levels 7

3 Facing the reality and exploiting it **9**
3.1 Feedback through environment 9
3.2 Maximum of variance inherited! 11
3.3 Relation between principal subspaces 13

4 Emergence of new structures **14**
4.1 Details of the neuronal mapping 15
4.2 Adding "inverse noise" 16
4.3 Analysis of the inherited variance 18

5 Neuron systems as models **19**
5.1 Controlling of information 20
5.2 Interpretation of mathematical patterns 21
5.3 "Metamodels" of neuron systems? 23

6 Step aside: an example application **25**
6.1 Implementation of the algorithm 25
6.2 Data and its model 27
6.3 Classification results 29

7 Towards full-scale neuron systems **32**
7.1 Capturing dynamics 32
7.2 Entering the frequency domain 34
7.3 Scenario: higher-level models? 36

8 Cybernetic minds **38**
8.1 Further interpretations 38
8.2 What is consciousness? 40
8.3 Extensions beyond neuronal realms 41

9 Discussion: what lies ahead? **43**

10 About bibliographies **44**

1 Introduction

Artificial neural networks are usually studied in a static setting — input patterns being transformed into some outputs or inner states. However, in reality such static mappings are just superficial reflections of the underlying dynamic phenomena. Concentrating on the observed surface patterns, or looking at the system only from above, cannot easily be changed afterwards; to attack the true essence of the neuronal systems, one has to convert the top-down approach to a bottom-up view right in the beginning. The connection between the two views is supplied through *dynamic feedback loops*, and the study of such structures is the field of *cybernetics.*

This changing of viewpoint is important if one truly wants to mimic natural phenomena or understand them. What is the nature of emergent functional structures, how can they come out from the original nothing? The claim here is that functionalities are based on dynamical, self-sustained *attractors,* and only as seen from outside, in the slower time scale, there are fixed-looking functionalities. In appropriate environments there is convergence rather than divergence at all levels, the higher levels being crucially dependent on the existence of the lower-level balances. The view of ever increasing complexity and chaos can be substituted with principles of self-organization and self-regulation when applying such "inverted" view of a complex system.

But, once more, why there is something instead of nothing, what are the original driving forces? In the domain of neural networks, one can start from the principle of *neuronal activity pursuit*: a neuron wants to receive activation, and there is a large number of such hungry neurons. After that, functionalities and properties start stacking on top of each other, one by one, from bottom to the very top, as shown in this presentation.

Of course, there are always many ways to proceed, and some guidelines are needed. The first principle to follow here is to study what are the constraints and what can be implemented using the already available functionalities; second, as there are no central controls, all operations have to be strictly local. The third guideline — the most challenging one — is the demand of *scalability.* Otherwise the studies can never be extended beyond laboratory-scale toy worlds.

This scalability claim means, for example, that there should be minimum number of adjustable parameters; tuning a hierarchy of interacting parameters soon becomes an unmanageable problem. The other principle related to scalability may sound astonishing: to understand qualitatively what happens in a large-scale system, to be able to copy simpler substructures, *they must be essentially linear.* At least in this presentation, all approaches are basically strictly linear. — Truly, it seems that there still are fresh problem settings in linear theory, starting from the fact that stability in dynamic structures can be achieved also in linear terms using *negative feedback*. The structural complexity, or the traditional nonlinearity, is thus changed to dynamic complexity; but we are not afraid of that, as our working hypothesis is that *everything is dynamics.*

And, after all, the strongest guiding principle in discussions is *intuition.* Reality is respected, and nature's ways to implement its amazing functionalities are appreciated. Perhaps surprisingly, our trained sense of *beauty* suggests if a specific path is worth following; this aesthetics cannot be formalized, but it is based on a deep look at *mathematical patterns.* This key role of intuition as a basis of the "new science" will be studied closer in the end of the presentation.

The cybernetic approach to neuron systems is related to various other established neural network paradigms:

- The adopted starting point results in *Hebbian neurons,* and, as there are the negative feedbacks, the *anti-Hebbian algorithms* with sparse coding property are closely related.
- Being based on linear matrix structures, *subspace methods* and *principal component networks* are discussed, the system acting as an *auto-encoder*, finding codes for patterns.
- As a long sequence of neuron layers becomes "collapsed", a simplified version of *error back-propagation* can be implemented.
- An energy function being iteratively minimized, there is a connection to *Hopfield nets,* and as the input is also iteratively tailored, it is near *(restricted) Boltzmann machines.*
- If one allows some crosstalk among neighboring neurons, the network can be seen as a (distributed) extension of *Kohonen's self-organizing map.*
- Finally, as the overall system can be studied also in terms of frequencies and vibration fields, there is perhaps even a connection to *holographic memories* and the like.

2 Properties of the information flow

A complex system is typically characterized by a fractal hierarchy of emergent levels. Here, we take (in a rather traditional manner) the level of *neuronal activations* as the starting point (however, we go beyond this assumption in 7.3).

2.1 Neurons — just trying to prosper

Available activation sources are denoted here as $\bar{u}_j$, where $1 \leq j \leq m$, and neuronal activities are denoted as $\bar{x}_i$, where $1 \leq i \leq n$. Typically, there holds $n \ll m$. Interpreting the *spread of activation* as being a result of some kind of *generalized diffusion*, it is assumed that the synapse a_{ij}, or the connection between $\bar{u}_j$ and $\bar{x}_i$ is linear; then, the total neuronal activation becomes

$$\bar{x}_i = a_{i1}\bar{u}_1 + \cdots + a_{im}\bar{u}_m = \sum_{j=1}^{m} a_{ij}\,\bar{u}_j. \tag{1}$$

The measure for neuronal success is the average variation level or (uncentered) *variance* $\mathrm{E}\{\bar{x}_i^2\}$. This variance can be interpreted as *(Fisher) information.* An expression for variance can be found, for example, by multiplying both sides in (1) by $\bar{x}_i$, and taking expectation:

$$\mathrm{E}\left\{\bar{x}_i^2\right\} = \sum_{j=1}^{m} a_{ij}\,\mathrm{E}\left\{\bar{x}_i\bar{u}_j\right\}. \tag{2}$$

The neuron's goal is to maximize this quantity by altering the synaptic weights. To define a sound optimization task, one needs to assume that there is some cost for keeping up the synaptic couplings a_{ij}. A practical way to assess the costs is to interpret a_{ij} as characterizing the *relative proximity* between the input and the neuron, and each input can be seen as defining a separate *orthogonal dimension* in the input space. Assuming that at any specific time instant, the "synaptic weight vector" is fixed, so that $a_{i1}^2 + \cdots + a_{im}^2$ has some constant value; then one can write

$$\begin{array}{ll} \text{Maximize} & \sum_{j=1}^{m} a_{ij}\, \mathrm{E}\left\{\bar{x}_i \bar{u}_j\right\} \\ \text{when} & \sum_{j=1}^{m} a_{ij}^2 = \text{constant}. \end{array} \tag{3}$$

This constrained optimization problem can be solved applying the method of Lagrange multipliers, giving

$$\left\{ \begin{array}{rcl} a_{i1} & = & q_i\, \mathrm{E}\left\{\bar{x}_i \bar{u}_1\right\} \\ & \vdots & \\ a_{im} & = & q_i\, \mathrm{E}\left\{\bar{x}_i \bar{u}_m\right\}, \end{array} \right. \tag{4}$$

where a new constant parameter q_i is introduced. Thus, from (1) it can be seen that if the neuron is optimally coupled to its environment, for some *coupling factor* q_i, there holds

$$\bar{x}_i = q_i \sum_{j=1}^{m} \mathrm{E}\left\{\bar{x}_i \bar{u}_j\right\}\, \bar{u}_j. \tag{5}$$

The same reasoning applies to all neurons in the system, so that the set of n similar equations (5) can be expressed in a compact matrix form as

$$\bar{x} = Q\, \mathrm{E}\left\{\bar{x} \bar{u}^{\mathrm{T}}\right\}\, \bar{u}, \tag{6}$$

where the vectors $\bar{x}$ and $\bar{u}$ contain the variables $\bar{x}_i$ and $\bar{u}_j$, respectively, and the diagonal matrix Q contains the individual q_i's on its diagonal. The matrix $\mathrm{E}\{\bar{x}\bar{u}^{\mathrm{T}}\}$ is the *covariance matrix,* but without the traditional mean-centering or normalization. One needs to remember that even though the matrix representation is employed, all operations in the system are completely local.

Essentially, (6) is a compact formulation for the *Hebbian learning principle* that has traditionally been studied in neural networks research: synaptic connection between the input and the neuron increases if they correlate. However, this matrix formulation seems to offer fresh views to the functionalities of the neuron population. What is more, now this principle can be extended: note that, in 7.2, the expression $\mathrm{E}\{\bar{x}_i{}^{\mathrm{H}} \bar{x}_i\}$ still reveals the *real* neuronal benefit.

2.2 Engineering of emergence

After all, modeling of neuron systems is such an ambitious task that one cannot only discuss the technical details — a wider perspective is needed. During the first reading, these issues can be skipped: learning, too, is a *constructivistic process* consisting of cybernetic convergent loops!

As an example of the challenges, study the distinction between *data* and *information.* Information is a *higher level* concept; intuitively, there must be some kind of *emergence* taking place between them. Indeed, in this context, intuition is seen as a resource that can make it easier to understand complex issues, and the mathematical constructs here just happen to carry the appropriate connotations, as shown below; to maintain the intuitive plausibility of the following discussions, a lengthy motivation is justified.

Here, we are defining *weak emergence* in strictly mathematical terms. This is accomplished through the expectation operation that captures the *memory of the prior behaviors*, abstracting individual sample details away. Now this operation determines how information cumulates as given by $\mathrm{E}\{\bar{x}\bar{u}^{\mathrm{T}}\}$, thus emerging from the lower-level data $\bar{x}$ and $\bar{u}$. The formulation (6) couples the two levels, *filtering* new data; indeed, one could even speak of *knowledge* in a very narrow sense, as this formula determines the *operational structure for applying information for data manipulation*. These observations can only contain the kernels of the complex ideas, at best, but if they manage to carry their essence, less trivial results can perhaps emerge when the essence cumulates.

Emergence is a *holistic* concept, whereas engineering is, by definition, purely *reductionistic*, and, *intuitively*, there is an *infinity* between them. So, the shortest route from the other to the other is through the infinity, in the spirit of Greek *apeiron*. Here it is simply assumed that when there is an infinite number of elementary operations applied as the expectation is calculated, in the limit, when the details cannot any more be detected, the *quantitative* changes to *qualitative*.

As compared to the infinite philosophical discussions concerning emergence, the nice thing about the current simple definition of weak emergence is that there are efficient mathematical tools to operate on the infinities.

Still, as the concepts that are being exploited here are so semantically loaded, there are more objections against the simplistic definitions: for example, there needs to be some *direction* in the underlying aspirations to capture the everyday intuition about emergence, otherwise the summations within the underlying chaos are somehow *meaningless* and nothing meaningful can come out of them. The key challenge is, really, how to capture *meaning* in the formal expressions, or how to capture the domain area *semantics*. — In our situation, however, things become easy: what is relevant in the environment is determined simply through the *resources* it can supply — and, in this case, the resource that everybody assumedly agrees with is the available information that is being competed for. Information is now the Aristotelian *energeia*, or the new idea of *emergy*, beyond the neuronal survival strategies. Understanding the "desires" of underlying subsystems makes it possible to master the emergence in the large even though the individual neuronal solutions remain out of supervision.

Of course, as modeling of complex systems is so important, the problems of emergence have been attacked before, and there are many approaches to "emergence engineering". There is little consensus about the general principles, but everybody seems to emphasize the role of *self-organization* in such systems, meaning that new structures have to come out from computations. Now, when applying the cybernetic approach, it is more like *self-reorganization* that takes place: first, there is *self-simplification* of prior structures, and then there is *self-complexification*,

or construction of new ones. The former process, or the compression of data into its inner structure, is studied in Section 3, and the latter process, or reshuffling the data kernels towards new structures, is studied in Section 4.

From the practical point of view, it is reasonable to make a concrete distinction between the expectation operator E and a special *emergence operator* $\mathcal{E}$ (an operator that makes things emerge from a tiny epsilon ϵ!). Whereas expectation is a mathematical abstraction, never available for real observation data, the new operator that is based on practical measurements, *pragmatic* or *sample expectation,* or *experience,* really, can be defined, for example, as

$$\frac{d\,\mathcal{E}_\tau\{\bar{x}\bar{u}^{\mathrm{T}}\}}{dt/\tau}(t) = \bar{x}(t)\bar{u}(t)^{\mathrm{T}} - \mathcal{E}_\tau\{\bar{x}\bar{u}^{\mathrm{T}}\}(t), \tag{7}$$

assuming that at the τ time scale the signals are always available, that is, the convergence of the internal signals is fast enough. The above formula defines a low-pass filter, the "visibility horizon" exponentially receding towards the past. — Later on, explicit time variables and scale subscripts will be omitted in formulas, and, at the appropriate time scale, the emergence operator $\mathcal{E}$ works virtually like the expectation operator E. For example, the formula (6) becomes

$$\bar{x} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\bar{u}. \tag{8}$$

Again, facing the reality and its non-idealities (here, admitting that the mathematical expectation is never available) makes it possible to see the beauty in the fractal nuances of the time scales (see also 3.1 and 7.2).

2.3 Interplay among emergent levels

To exploit the mathematical machinery, one has to connect the level of signals and their emergent counterparts. This can be accomplished by concentrating on their common statistical properties: in practice, a still longer time scale τ can be selected so that the data and the information are *together* seen in that wider perspective. Now, the statistical properties are captured in the covariance matrices.

One can find many expressions governing the covariances. When multiplying (8) from the right by $\bar{x}^{\mathrm{T}}$ and applying the emergence operator, one has the following expression:

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}. \tag{9}$$

This comes from the fact that the operator is linear and it traverses through constant expressions just as the expectation operator does. The transpose of this gives yet another expression (remember that $Q^{\mathrm{T}} = Q$):

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}\,Q. \tag{10}$$

Multiplying the former expression by Q from the right and the latter from the left, it becomes evident that there must hold

$$Q\,\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}Q, \tag{11}$$

so that also

$$f\left(Q\right)\, g\left(\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right) = g\left(\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)\, f\left(Q\right), \tag{12}$$

where f and g are any functions that can be defined in terms of matrix power series. This commutativity property means that many mathematical manipulations of the matrix data structures become very much like scalar algebra in later analyses. In principle, there are two classes of solutions fulfilling (11): if the factors q_i are distinct, $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$ must become diagonal, but if there holds $q_i = q$ for all i, then there are no constraints for the covariance. These cases are studied closer in 3.3.

Further, assuming invertibility of $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$, and noting (12), from (9) or (10) one has

$$I_n = Q^{1/2}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}Q^{1/2}. \tag{13}$$

When defining

$$\theta^{\mathrm{T}} = Q^{1/2}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}, \tag{14}$$

one has

$$I_n = \theta^{\mathrm{T}}\,\theta. \tag{15}$$

The columns in this new matrix θ are thus orthonormal. In 3.3 the role of the matrices like θ are studied closer; here, let us just show some alternative formulations for it:

$$\theta^{\mathrm{T}} = \left(Q\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1/2}Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} = \left(\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}\right)^{-1/2}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}. \tag{16}$$

Further, by multiplying (8) from the right this time by $\bar{u}^{\mathrm{T}}$ and applying the emergence operator, one has

$$\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}. \tag{17}$$

Substituting this in (10), there holds

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}Q. \tag{18}$$

Again assuming invertibility of $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$, and noting (12), this can be changed to read

$$Q^{-1} = Q^{1/2}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}Q^{1/2}, \tag{19}$$

so that we get

$$Q^{-1} = \theta^{\mathrm{T}}\,\mathcal{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}\theta. \tag{20}$$

This means that if ever the basic assumption (8) is fulfilled, the statistical properties of the input $\bar{u}$ are fixed to the selected Q. As will be shown later, the coupling parameters q_i in Q can be seen

as determining the "stiffnesses" of the coupled neurons. Especially, as it turns out in 3.3, the visible data $\bar{u}$ becomes *diagonalized* by the system, and if all q_i are equal, data gets *whitened.* This modification of the environment can be seen as a manifestation of a more general *observer effect.* How can the system dictate the properties of its environment — this is studied next.

The above analyses apply if such a mapping matrix really exists as proposed in (8). How to make signals stationary and the formulas meaningful? How to avoid the excessive growth (explosion) of $\bar{x}$ and the resulting instability of adaptation? Indeed, this instability problem is the traditional curse of all Hebbian-based approaches. How to supply the "integrated intelligence" to assure the balance on the "edge between order and chaos", and, specially, how to reach that in linear terms? In the cybernetic spirit, of course, dynamics and feedback is here proposed.

3 Facing the reality and exploiting it

There are no pure information flows in nature: exploitation of signals also means exhaustion of them. When this implicit effect is included in the signal flow diagrams, it turns out that there is *negative feedback* from $\bar{x}$ back to $\bar{u}$: activation consumed by a neuron is not available to others. This feedback *stabilizes the closed loop.*

3.1 Feedback through environment

When the neuron i has been put running, and it has activity $\bar{x}_i$, it sucks from resource j such an amount of resource that is proportional to the synaptic strength a_{ij} or $q_i\mathcal{E}\{\bar{x}_i\bar{u}_j\}$. This means that the change in the input activation j because of the neurons can be written as

$$\Delta\bar{u}_j = a_{1j}\bar{x}_1 + \cdots + a_{nj}\bar{x}_n = \sum_{i=1}^{n} q_i\mathcal{E}\left\{\bar{x}_i\bar{u}_j\right\}\bar{x}_i. \tag{21}$$

For the whole population of inputs one can write in matrix form

$$\Delta\bar{u} = \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}} Q\,\bar{x}. \tag{22}$$

The key point in (22) is that the neuron is thought to be an active entity: after being launched, it ruthlessly pulls the activity it needs to itself. The above formula, together with (8), determines the connection between the internal and external realms, and there can be some scaling effects taking place in the "conversions" between $\bar{u}$ and $\bar{x}$. If $\bar{x}$ alone is not strong enough to "make a difference that makes a difference" in the environment. or implement the feedback, one has to scale up its elements. So, assume that it is some $\bar{x}' = C\bar{x}$ for some diagonal C that would only implement the necessary balancing effect. Study what happens if the variable $\bar{x}'$ is used instead:

$$\bar{x}' = C\bar{x} = C\,Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\bar{u} = Q\,C\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\bar{u} = Q\mathcal{E}\left\{C\bar{x}\bar{u}^{\mathrm{T}}\right\}\bar{u} = Q\mathcal{E}\left\{\bar{x}'\bar{u}^{\mathrm{T}}\right\}\bar{u}. \tag{23}$$

This means that all references to the old variable have vanished. One can also employ the new variable $\bar{x}'$ and the discussions above are still valid; however, from now on, assume that such

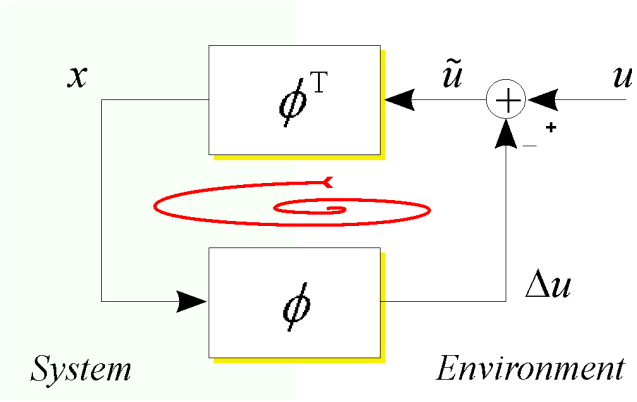


Figure 1: Cybernetic loop structure and signals therein

necessary data scalings have been carried out for the variable $\bar{x}$ itself (that is, instead of using $\bar{x}'$ in the subsequent discussions, it is the familiar $\bar{x}$ that is used). How to scale the elements of $\bar{x}$ to make it somehow compatible with the environment — this is discussed in 4.2.

Above, symbols like $\bar{u}$ and $\bar{x}$ have been used all the time; they are the final, effective, visible variables, dynamic balance values that result after underlying interactions have converged in the equilibrium of tensions. The original undisturbed resource vector u is *invisible* for the local actors, because in reality it is disturbed by the systems (this can be called the *observer effect*). The actual signal structure is shown in Fig. 1. For the disturbed input, or *residual,* there holds

$$\tilde{u}(t) = u - \Delta u(t), \tag{24}$$

and the asymptotic values are defined (in a somewhat sloppy way) as

$$\bar{u} = \lim_{t\to\infty} \{\tilde{u}(t)\} \tag{25}$$

and, correspondingly, $\bar{x}$ can be found only after convergence:

$$\bar{x} = \lim_{t\to\infty} \{x(t)\} = \lim_{t\to\infty} \left\{Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\tilde{u}(t)\right\}. \tag{26}$$

In the asymptotic case, when the balance has been found, the situation looks like that shown in (8). Here it is assumed that one only studies some kind of "local infinities" at the local time scale that is relevant to the dynamics of x. Indeed, to capture the "momentary nature" of behaviors in the system, one has to concentrate on the following scales separately (when concentrating on a specific time scale, signals from other scales look like constants):

- Fastest, the internal time scale in the neurons: relevant to momentary signals like x
- Moderate, environmental time scale: applies to signals like u, $\bar{u}$ and $\bar{x}$
- Slowest, "system scale": models of (co)variation, for example $\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$, and $\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}$.

When the above expressions concerning the feedback are combined, one has

$$\bar{u} = u - \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}} Q\,\bar{x}, \tag{27}$$

and, further, for $\bar{x}$

$$\bar{x} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\bar{u} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u - Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}} Q\,\bar{x}, \tag{28}$$

or, when solved,

$$\bar{x} = \left(I_n + Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}Q\right)^{-1} Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u. \tag{29}$$

Using (10), one has

$$\bar{x} = \left(I_n + Q\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u, \tag{30}$$

and, simplifying further, one has an expression for $\bar{x}$ directly in terms of u:

$$\bar{x} = \left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u. \tag{31}$$

Using corresponding manipulations, by setting (27) in the other location (within the emergence operator) in (8), one can also derive, for example,

$$\bar{x} = \left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} \mathcal{E}\left\{\bar{x}u^{\mathrm{T}}\right\}\bar{u}. \tag{32}$$

3.2 Maximum of variance inherited!

In the formula (31) there is a discrepancy: the input is u but the covariances are given in terms of $\bar{u}$. This can be resolved by manipulating the expression:

$$\begin{aligned} \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} &= \mathcal{E}\left\{\bar{x}\left(u - \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}Q\bar{x}\right)^{\mathrm{T}}\right\} \\ &= \mathcal{E}\left\{\bar{x}u^{\mathrm{T}}\right\} - \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}. \end{aligned} \tag{33}$$

Solving this for $\mathcal{E}\{\bar{x}\bar{u}^{\mathrm{T}}\}$, one has

$$\begin{aligned} \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} &= \left(I_n + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}Q\right)^{-1} \mathcal{E}\left\{\bar{x}u^{\mathrm{T}}\right\} \\ &= \left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} Q^{-1}\mathcal{E}\left\{\bar{x}u^{\mathrm{T}}\right\}. \end{aligned} \tag{34}$$

Combining (31) and (34):

$$\bar{x} = \underbrace{\left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-2} Q^{-1}}_{M_1}\ \underbrace{\mathcal{E}\left\{\bar{x}u^{\mathrm{T}}\right\}}_{M_2}\ u. \tag{35}$$

Using this expression, one can study the connection between the undisturbed u and $\bar{x}$. If the statistical properties of the input data u are assumed to remain intact, one has

Theorem.
If data is rich enough (non-zero variation dimensions in data $d \geq n$), and if each mode remains cybernetic *or* alive *(see 4.1), after convergence the neuronal mapping from u to $\bar{x}$ spans the principal subspace of data variation in u, corresponding to the n most significant eigenvector directions of the data covariance matrix $\mathcal{E}\{uu^{\mathrm{T}}\}$.*

Proof.
Rather than studying the adaptation process as a continuous process, the time axis is here assumed to be divided in long enough subparts; these subparts are indexed below using superscript numbers in parentheses. The expectations, when calculated as sample averages within each interval, are already assumed to be accurate enough. If one starts from some arbitrary mapping matrices $M_1^{(0)}$ and $M_2^{(0)}$, the step-by-step covariance adaptation, iterating (35), proceeds as

$$\begin{aligned}
\bar{x}^{(0)} &= M_1^{(0)} M_2^{(0)}\, u \\
\bar{x}^{(1)} &= M_1^{(1)} \mathcal{E}\left\{\bar{x}^{(0)} u^{\mathrm{T}}\right\} u = M_1^{(1)} \mathcal{E}\left\{M_1^{(0)} M_2^{(0)} u u^{\mathrm{T}}\right\} u \\
&= M_1^{(1)} M_1^{(0)} M_2^{(0)} \mathcal{E}\left\{u u^{\mathrm{T}}\right\} u \\
\bar{x}^{(2)} &= M_1^{(2)} \mathcal{E}\left\{\bar{x}^{(1)} u^{\mathrm{T}}\right\} u = M_1^{(2)} \mathcal{E}\left\{M_1^{(1)} M_1^{(0)} M_2^{(0)} \mathcal{E}\left\{u u^{\mathrm{T}}\right\} u u^{\mathrm{T}}\right\} u \\
&= M_1^{(2)} M_1^{(1)} M_1^{(0)} M_2^{(0)} \mathcal{E}\left\{u u^{\mathrm{T}}\right\}^2 u \\
&\vdots \\
\bar{x}^{(k)} &= M_1^{(k)} M_2^{(k)}\, u = \left(\prod_{i=0}^{k} M_1^{(k-i)}\right) M_2^{(0)} \mathcal{E}\left\{u u^{\mathrm{T}}\right\}^k u.
\end{aligned} \tag{36}$$

The former part $M_1^{(k)} = \prod_{i=0}^{k} M_1^{(k-i)}$ is a scaling matrix of dimension $n \times n$ and it does not affect the subspace being spanned by the mapping. On the other hand, $M_2^{(k)}$ deserves more attention. Assume that the eigenvalue decomposition of the data covariance (see 3.3) is written as

$$\mathcal{E}\left\{u u^{\mathrm{T}}\right\} = \Theta\, \Lambda\, \Theta^{\mathrm{T}}. \tag{37}$$

The resulting mapping matrix $M_2^{(k)}$ becomes

$$M_2^{(k)} = M_2^{(0)} \mathcal{E}\left\{u u^{\mathrm{T}}\right\}^k = \left(M_2^{(0)} \Theta\right)\, \Lambda^k \Theta^{\mathrm{T}}. \tag{38}$$

This means that in the mapping matrix the relevance of the principal component direction j is weighted by λ_j^k. At each iteration, the eigenvectors become better aligned with the most significant eigenvectors. Because the variables $\bar{x}_i$ are linearly independent, it is the n most significant covariance matrix eigenvectors that determine the mapping after adaptation (assuming that in an ordered list of decreasing eigenvalues, there holds $\lambda_n > \lambda_{n+1}$). These eigenvectors define the same subspace as in the case of $\bar{x}$ vs. $\bar{u}$ (but the eigenvalues differ; see below). □

3.3 Relation between principal subspaces

To understand the properties of the Hebbian neuron populations, the structure of input data needs to be studied closer. For stationary input data, one can always write the *eigenvalue decomposition* for the covariance matrix $\mathcal{E}\{uu^{\mathrm{T}}\}$ as

$$\mathcal{E}\left\{uu^{\mathrm{T}}\right\} = \Theta\,\Lambda\,\Theta^{-1}, \tag{39}$$

where the $m \times m$ matrix Θ contains the *eigenvectors* of the covariance matrix of u as its columns, and the diagonal matrix Λ contains the corresponding *eigenvalues* on its diagonal. Because of the structure of the covariance matrix, all of its eigenvalues are real and non-negative, and they can be ordered in the order of descending significance, revealing the proportion of variation that is distributed in that eigenvector direction. Because of the symmetricity of the covariance matrix, all eigenvectors are normal to each other, so that when they are normalized, there holds $\Theta^{\mathrm{T}}\Theta = I_m$, or $\Theta^{-1} = \Theta^{\mathrm{T}}$. When data is projected onto the basis determined by the covariance matrix eigenvectors, so that $z = \Theta^{\mathrm{T}}\,u$, the new latent variables z are known as *principal components.*

The same kind of eigenvalue decomposition as in (39) can be carried out also for the modified data $\bar{u}$, and one has now

$$\mathcal{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\} = \bar{\Theta}\,\bar{\Lambda}\,\bar{\Theta}^{-1}. \tag{40}$$

According to (15) and (20), n of the m eigenvectors in $\bar{\Theta}$ (collected as columns) are present in the matrix θ:

$$\theta = \bar{\Theta}_{[n]}D. \tag{41}$$

Here, notation $\bar{\Theta}_{[n]}$ means that only n of the constructs are selected; D is some orthogonal $n \times n$ matrix shuffling these vectors. It is the matrix $\mathcal{E}\{\bar{x}\bar{u}^{\mathrm{T}}\}$ that spans the n dimensional subspace of θ, so that this mapping is characterized by the eigenvectors of $\bar{\Theta}$; but, according to (31), the same $\mathcal{E}\{\bar{x}\bar{u}^{\mathrm{T}}\}$ spans a subspace in Θ, too, and because of the Theorem in 3.2, it must even be the most relevant of the subspaces for data u. This all means that Θ and $\bar{\Theta}$ are closely related. However, there is an essential difference between the above eigenvalue decompositions: whereas eigenvectors are the same, the eigenvalues are not, or $\Lambda \neq \bar{\Lambda}$. The relevance ordering of eigenvectors can change. It may even be so that the n most significant eigenvectors are not the most significant ones after the system coupling; this is because the most visible data dimensions are suppressed by the special structure of the feedback control (this case of excessive coupling could be called "hyper-cybernetic"). The modification of the data variance structure caused by the cybernetic coupling between the system and its environment is illustrated in Fig. 2.

What comes to the coupling between the environment and the system, the above discussion is not yet the whole story. When looking at the final latent variables, or the vector $\bar{x}$, there are the following two essentially opposite possibilities of interest:

1. If all q_i in Q are distinct, according to (20), the original data eigenvalues λ_j change to $\bar{\lambda}_j = 1/q_i$ (assuming that neuron i has become coupled to mode j); further, for (11) to hold, $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$ must become diagonal, and from (14) it is evident that there is no shuffling of basis vectors — the system implements *principal component analysis.*

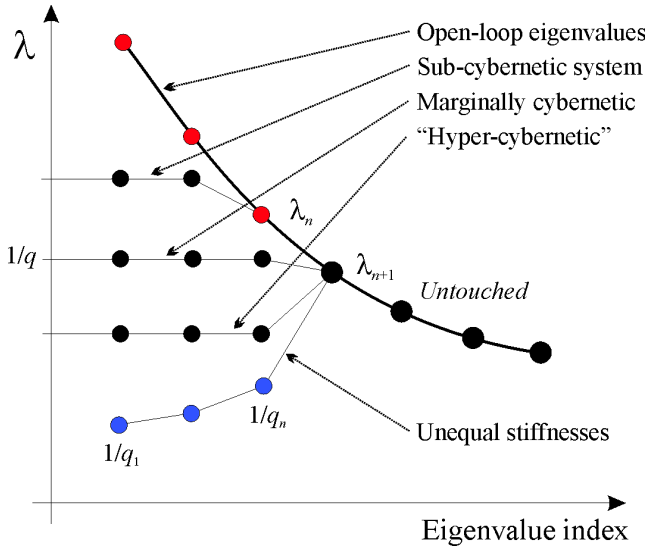


Figure 2: Eigenvalues in the coupled system

2. If all q_i in Q are equal, on the other hand, so that $Q = q\,I_n$, all eigenvalues are *equalized,* all $\bar{\lambda}_j$ equalling $1/q$, no matter what the original $\lambda_j > 1/q$ are; now there are no limitations for $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$ because of (11), so that the system implements only *principal subspace analysis* with rotatable basis vectors.

In practice, in the case of distinct q_i there must hold $D = P$, P being a permutation matrix, otherwise there are no formal limitations for D in addition to orthogonality. The latter case is the more interesting, and it is reasonable to study how the internal feedback structures rotate the basis axes. — It deserves to be recognized that "whitening" of the *effective data* in $\bar{u}$ has been automatically accomplished without any preprocessing (centering or scaling) of the original data in the coupling process (compare to *independent component analysis,* etc.). Now the latent variables are not orthogonal, or $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$ is not diagonal, and ϕ_i are not orthogonal. There is still a tendency towards *independence* of variables (at least if there are nonlinearities in the model, see 5.3): remember that signals are shuffled in the loops, and the system tries to eliminate all correlations between the generated functions.

4 Emergence of new structures

Despite the analyses above, there are *two* classes of solutions to (8). In addition to the case that was discussed in Sec. 3, the trivial solution $\bar{x} \equiv 0$ for all inputs, or $\bar{x}_i \equiv 0$ for a subset of them, also satisfies the assumed constraint, the corresponding mappings vanishing, $\mathcal{E}\{\bar{x}_i\bar{u}\} \equiv 0$. To understand the faith of the neuron i, whether it fades away or stays "alive", depends on the corresponding coupling to the environment.

4.1 Details of the neuronal mapping

From (31) one can write yet another expression for the covariance by multiplying the expression by its transpose, and applying the emergence operator:

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \mathcal{E}\left\{uu^{\mathrm{T}}\right\} \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}} \left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1}.$$

Eliminate the matrix inverses by multiplication, so that

$$\left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right) \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} \left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right) = \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \mathcal{E}\left\{uu^{\mathrm{T}}\right\} \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}},$$

and observe the commutativity of the matrices:

$$\begin{aligned}
&\left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^2 \\
&\quad = Q^{-1/2}\, Q^{1/2} \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2} \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \mathcal{E}\left\{uu^{\mathrm{T}}\right\} \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}} \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2} Q^{1/2}\, Q^{-1/2} \\
&\quad = Q^{-1/2}\, \theta^{\mathrm{T}}\, \mathcal{E}\left\{uu^{\mathrm{T}}\right\} \theta\, Q^{-1/2}.
\end{aligned}$$

Further, because of the orthogonality of θ,

$$Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q^{-1/4}\, \theta^{\mathrm{T}}\, \mathcal{E}\left\{uu^{\mathrm{T}}\right\}^{1/2} \theta\, Q^{-1/4}, \tag{42}$$

or

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q^{-1/4}\, \theta^{\mathrm{T}}\, \mathcal{E}\left\{uu^{\mathrm{T}}\right\}^{1/2} \theta\, Q^{-1/4} - Q^{-1}. \tag{43}$$

If the coupling factors q_i are distinct for all i, the θ mapping has a diagonalizing property, and

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q^{-1/4}\, P^{\mathrm{T}} \Lambda_{[n]}^{1/2} P\; Q^{-1/4} - Q^{-1}, \tag{44}$$

where $\Lambda_{[n]}$ is a diagonal $n \times n$ matrix containing the most significant eigenvalues of the original data u, and P is a permutation matrix. Assuming that the eigenvalue λ_j in the data has become coupled with variable x_i, one can write

$$\mathcal{E}\left\{\bar{x}_i^2\right\} = \sqrt{\frac{\lambda_j}{q_i}} - \frac{1}{q_i}. \tag{45}$$

The behavior of this as a function of q_i is shown in Fig. 3. Incidentally, the square root form of activity inheritance is also motivated by the *Penrose's voting rule* that gives equal weight to all "individuals" beyond the emergent-level activities. Additionally, there is now the *threshold* term $-1/q_i$. Because the variances always must be non-negative, meaning that variations in each direction must have real values, one can see that the non-trivial solutions are only possible if the variation level in the input data is high enough, so that the additional factor $-Q^{-1}$ in (44) becomes fully compensated. To keep the neuron functional, there must hold

$$q_i > \frac{1}{\lambda_j}. \tag{46}$$

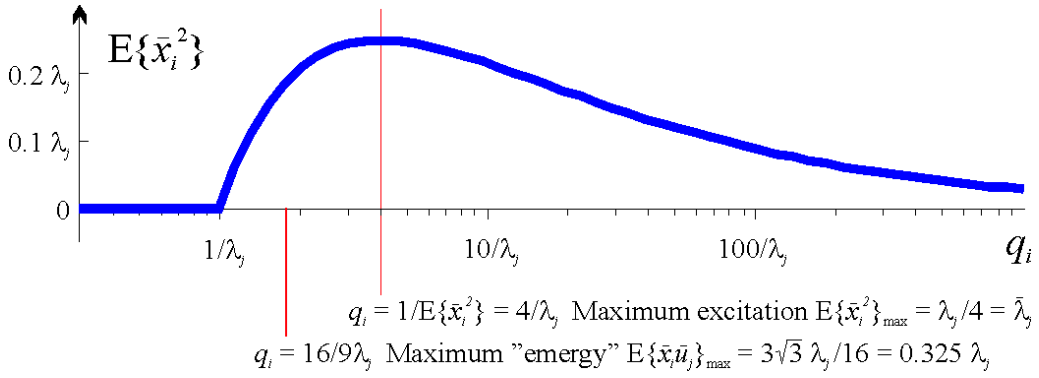


Figure 3: System activation as a function of the coupling parameter

This assures that the studies in the previous sections are relevant; this also assures that the matrix $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$ remains invertible. — Strange structures emerging in the strictly linear model!

Where is this activation lost, where does the "static friction" come from? This loss can perhaps be seen as some kind of minimum dissipation that is needed to keep the mills rolling. It is the loop-based iteration that essentially solves a set of linear equations when finding the equilibrium in the algebraic loop, providing data whitening, and only using enough pressure (strong enough coupling q_i), this can be accomplished.

On the other hand, if the incoming activation flow is strong and if there are limitations for individual neurons, so that a single neuron cannot exhaust all available activation as revealed by (45), additional neurons can start sharing the modal load; this means that the variable $\mathcal{E}\{\bar{x}_i^2\}$ represents their total energy.

4.2 Adding "inverse noise"

Thus, not to introduce the burden of adjustable parameters in the system, it has to be assumed that there is some local mechanism assuring that the neuron i remains "alive" (or "cybernetic") by increasing the value of q_i if the activity in the neuron i seems to be vanishing altogether. A clever choice to reach such adaptive sensitivity seems to be to define

$$q_i = b\,\frac{1}{\mathcal{E}\,\{\bar{x}_i^2\}}, \tag{47}$$

with $b > 0$ being some scaling factor, so that, assuming that there is similar local compensation in all neurons,

$$Q = \begin{pmatrix} \frac{b}{\mathcal{E}\{\bar{x}_1^2\}} & & \\ & \ddots & \\ & & \frac{b}{\mathcal{E}\{\bar{x}_n^2\}} \end{pmatrix} = b\ \mathrm{Var}\,\{\bar{x}\}^{-1}, \tag{48}$$

where $\mathrm{Var}\,\{\bar{x}\}$ is a diagonal matrix containing the (uncentered) variances of variables $\bar{x}_i$ on its diagonal, variances being defined in the familiar τ scale.

There are various technical motivations for selecting Q in such a way. First, neurons with such compensation are always stable, so that the system as a whole remains stable even if the negative feedback through the environment would fail. Concerning the exact structure of (48), the best motivation is perhaps given by the convergence considerations: the adaptation process of the mapping matrix ϕ (see 5.1) now becomes a data-based *identification* routine when applying a "robustified" *stochastic Newton method.* — From the plausibility point of view, such additional activity control is not as disturbing as it seems, because it is strictly local; what is more, also natural neurons turn out to implement similar activity-based controls.

Variance compensation keeps the activity in the neuron constant even if there were some additional activity losses in neurons; for example, the neurons can be exploited by further neurons as inputs, so that one can have a sequence of neuron populations without changing their theoretical properties.

The "exaggerated variance compensation" against the growth of variables means that activity is aggressively pushed to other neurons; as the total variance still remains to be shared, the neurons finally end in having the same variance load. This means that, as the variances are then equal, also q_i are, and, according to (20), eigenvalues $\bar{\lambda}_i$ get equalized and variance structure in $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$ gets blurred, becoming non-diagonal. Rotations can then be introduced, but only so that this equality among the latent variable variances remains intact: the D matrix has to supply for such weighted combinations of eigenvalues that the sums are equal. With this special selection of Q, equation (43) can be studied closer:

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \frac{1}{\sqrt{b}}\,\mathrm{Var}\,\{\bar{x}\}^{1/4}\,\theta^{\mathrm{T}}\,\mathcal{E}\left\{uu^{\mathrm{T}}\right\}^{1/2}\,\theta\,\mathrm{Var}\,\{\bar{x}\}^{1/4} - \frac{1}{b}\,\mathrm{Var}\,\{\bar{x}\}, \tag{49}$$

and multiplying this by $\mathrm{Var}\,\{\bar{x}\}^{-1/2}$ from the left and by $\mathrm{Var}\,\{\bar{x}\}^{-1/2}$ from the right one has

$$\begin{aligned}\overbrace{\mathcal{E}\left\{\mathrm{Var}\,\{\bar{x}\}^{-1/2}\bar{x}\,\bar{x}^{\mathrm{T}}\mathrm{Var}\,\{\bar{x}\}^{-1/2}\right\}}^{\mathrm{Corr}\{\bar{x}\}} \\ = \tfrac{1}{\sqrt{b}}\,\mathrm{Var}\,\{\bar{x}\}^{-1/4}\,D^{\mathrm{T}}\,\Lambda_{[n]}^{1/2}\,D\,\mathrm{Var}\,\{\bar{x}\}^{-1/4} - \tfrac{1}{b}\,I_n.\end{aligned} \tag{50}$$

The left-hand side of this expression is the (uncentered) correlation matrix, all diagonal elements being 1. The right hand side is some rotation and scaling of the principal subspace data covariance matrix; what is more interesting, however, is the additional term, or $-\frac{1}{b}\,I_n$, in the end of the expression. To understand its role, remember that normally adding (or subtracting) noise can only increase the variation level, and, specially, adding white uncorrelated noise only increases the diagonal elements in the covariance matrix (or suppresses the non-diagonal ones in the correlation matrix). Now, on the other hand, the diagonal elements are being artificially *reduced*; an intuitively appropriate name for such effect is *black noise.*

Applying such active noise suppression, variation in the n most significant data directions becomes attenuated, reducing uncorrelated information (or noise) that is only visible on the

diagonal of the covariance matrix. The qualitative net effect is perhaps *not* sparse coding with maximally distinct codes in the traditional sense; rather, one could speak of some kind of "companion coding" that tries to find groupings, emphasizing non-diagonal correlations. Perhaps this could be characterized as resulting in a *mixture model* among sparse subspaces.

Another interpretation concerning the threshold term in the formula (49) is also possible. In theory, variation among the activations in the system can be freely distributed, and the sum of variances still remains the same. However, when there is now the threshold, all variance contributions below the threshold are zeroed; to reach maximum of the effective variation, it is better to concentrate the activity in a few neurons, while letting the others (those remaining under the threshold anyway) "voluntarily" have little activity. This kind of activity redistribution based on "variance difference maximization" is carried out also in *factor analysis*, and, in its extreme, this interpretation results in sparsity pursuit.

4.3 Analysis of the inherited variance

As all variances $\mathcal{E}\{\bar{x}_i^2\}$ become equal with the selection (48), one can easily apply the matrix trace to (50), and one has for all i and j a formula for the variances:

$$\mathcal{E}\left\{\bar{x}_i^2\right\} = \frac{b}{q_i} = b\,\bar{\lambda}_j = \left(\frac{\sum_{\iota=1}^{n}\sqrt{\lambda}_\iota}{n\left(\sqrt{b}+\frac{1}{\sqrt{b}}\right)}\right)^2. \tag{51}$$

When one selects $b = 1$, there is an intuitively appealing balance between the internal and external variances, so that the value of $\mathcal{E}\{\bar{x}_i^2\} = \bar{\lambda}_j = \left(\sum_{\iota=1}^{n}\sqrt{\lambda}_\iota\right)^2/4n^2$ is the same for all i and j. Following the terminology of Geoffrey Hinton, variables become "equivariant capsules".

It is interesting to note that the square roots of the data covariance matrix, or the numbers $\sqrt{\lambda}_j$, are directly the *singular values* of the data matrix; and the expression $\sum_{j=1}^{n}\sqrt{\lambda}_j$ for ordered λ_j is called the *Ky Fan n-norm* of the data matrix.

Now we can return to the discussion in (3.1): the above formula applies only for such $\bar{x}$ that make the loops in the system balanced so that the signals are compatible in the system and in its environment (so that the formula (21) really has its intended effect). It is reasonable to scale the elements of $\bar{x}$ explicitly by selecting the diagonal elements of the scaling matrix C in (23) as

$$c_i = \frac{\sum_{\iota=1}^{n}\sqrt{\lambda}_\iota}{n\left(\sqrt{b}+\frac{1}{\sqrt{b}}\right)} \cdot \frac{1}{\sqrt{\mathcal{E}\left\{\bar{x}_i^2\right\}}}, \tag{52}$$

where $\mathcal{E}\{\bar{x}_i^2\}$ is the current variance level. When implementing the algorithms outside the brain, this scaling can be applied after each state update (until, hopefully, this correction is no more needed).

One can even propose a system size optimization scheme based on the formula (51): for the coupling to take place, there must hold $\bar{\lambda}_j < \lambda_j$ for each $j \leq n$ within the system; now, then, select n so that the maximum number of modes gets captured without violating this eigenvalue

criterion. Assuming that the eigenvalues λ_j are ordered in descending order, for the last $j = n$ to be included there should still hold

$$\lambda_n > \left(\frac{\sum_{\iota=1}^{n-1} \sqrt{\lambda_\iota}}{n(b+1)-1} \right)^2 , \tag{53}$$

or, approximately for large n, and for $b = 1$,

$$\sqrt{\lambda_n} > \frac{1}{2} \frac{\sum_{\iota=1}^{n-1} \sqrt{\lambda_\iota}}{n-1}, \tag{54}$$

so that *the new singular value to be included must be at least half of the average of the previous ones.* This test can be used for all $n \geq 2$ (there are never coupling problems for the model size $n = 1$). The maximum n is dictated by the properties of the original data, or by the outlook of the λ_j eigenvalue envelope. The criterion can be relaxed using data preprocessing, that is, by making the distribution range of the eigenvalues narrower, and, in the extreme case, if eigenvalues are made equal, there are no theoretical limitations for the system size. Such a formal criterion, model size being determined without closer data analysis, suggests that the feature representations cannot be unique.

If n is selected below the maximum, the system can become "hyper-cybernetic" with twisted eigenvalue structure: in the visible residual data, it *seems* that the most significant of the eigenvalues are left outside the model, the modes that are included in the model being over-compensated. Values of n beyond the optimum result in redundancy, neurons sharing each others' activity patterns, meaning that $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$ becomes singular. No matter how high n is, variation beyond the "visibility horizon" cannot be seen by the system; depending on the parameter b, or the level of added virtual noise, this information will automatically be regarded as somehow suspicious. Because of the regularization term Q^{-1} in the regularized regression formula (see 5.1), the extra variables do not collapse the numerical behavior of the system, however, and the abundance of nodes (even beyond m) makes it possible to emulate special "lossy" neural network structures (see 5.3).

In principle, application of the criterion (53) makes the *final* free parameter n in the cybernetic model fixed. In practice, when implementing neuromimetic algorithms, such fact would be very nice, as the traditionally decisive role of parameter tunings in neural network algorithms would be thus eliminated.

5 Neuron systems as models

So, it seems that the neuron population implements a special kind of representation of the properties of the incoming data — but why does it do that? It turns out that the constructed *model of the environment* makes it possible to implement *model-based control* to *maximally exhaust information from the environment*, and, as seen from the neuronal point of view, this makes it possible to *maximally exploit the available resources*.

5.1 Controlling of information

Even though everything in neural populations is based on elementary operations, the systemic properties can best be understood in terms of multivariate linear theory and as *mappings between spaces*. Interactions between a system and its environment are mappings between the space of inputs and the space of neuron activities. When the dynamic equilibrium is found not only on the signal level but also on the statistical level, the relation between the inputs and the linear neurons is captured by the *explicit mapping*

$$\phi^{\mathrm{T}} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} = \left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1}\mathcal{E}\left\{\bar{x}u^{\mathrm{T}}\right\}, \tag{55}$$

according to formulas (8) and (32). This means that the feedforward mapping can be expressed as $\bar{x} = \phi^{\mathrm{T}}\bar{u}$ and the feedback as $\Delta\bar{u} = \phi\bar{x}$. Further, when the effective mapping from the original, undisturbed u to the system state $\bar{x}$ is solved, so that $\bar{x} = \varphi^{\mathrm{T}}u$, one has the following formulation for this *implicit mapping*, according to (31),

$$\varphi^{\mathrm{T}} = \left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}. \tag{56}$$

Using these notations, one can find new formulations; for example, as in (16), one can express the eigenvector matrix as an "orthogonalization" of the mappings:

$$\theta^{\mathrm{T}} = \left(\phi^{\mathrm{T}}\phi\right)^{-1/2}\phi^{\mathrm{T}} = \left(\varphi^{\mathrm{T}}\varphi\right)^{-1/2}\varphi^{\mathrm{T}}. \tag{57}$$

However, to truly understand what takes place in the cybernetic loop of neurons, one needs to take a wider perspective.

Assume that there is some data $\xi(k)$ of dimension n, and there is some other related data $\zeta(k)$ of higher dimension m, with $1 \leq k \leq K$. One would like to find the best possible (approximate) mapping from the space of ξ to the space of ζ so that the average of the squared *reconstruction error*, or $\|\zeta(k) - \hat{\zeta}(k)\|_2^2$, would be minimized (note that now one would like to find the optimal mapping from the lower to the higher dimension, whereas in principal component analysis the direction is opposite). The standard solution to this problem is provided by the least-squares method, giving the *multilinear regression estimate*

$$\hat{\zeta}(k) = \left(\mathrm{E}\left\{\xi\xi^{\mathrm{T}}\right\}^{-1}\mathrm{E}\left\{\xi\zeta^{\mathrm{T}}\right\}\right)^{\mathrm{T}}\xi(k). \tag{58}$$

However, this estimate is typically not *robust* for high-dimensional data, as *colinearities* can cause the covariance matrix $\mathrm{E}\{\xi\xi^{\mathrm{T}}\}$ to become practically non-invertible. A simple fix to this problem is to add uncorrelated white noise to the data ξ; then the eigenvalues of the covariance matrix get farther from zero (this is closely related to *regularization* in the neural network algorithms). Thus, if the added white noise

has covariance C, diagonal matrix with all positive entries, having always full rank, one has the (somewhat conservative) *ridge regression* formula

$$\hat{\zeta}(k) = \left(\left(C + \mathrm{E}\left\{ \xi \xi^{\mathrm{T}} \right\} \right)^{-1} \mathrm{E}\left\{ \xi \zeta^{\mathrm{T}} \right\} \right)^{\mathrm{T}} \xi(k). \tag{59}$$

When one selects $C = Q^{-1}$, $\xi = \bar{x}$, and $\zeta = u$ (or $\zeta = \bar{u}$) in (59), and when E is identified with $\mathcal{E}$, one can see the connection to formulas (55) and (56). Indeed, one can summarize the mappings in the following form with intriguing *dual symmetry*:

$$\left\{ \begin{array}{rcl} \bar{x} & = & \phi^{\mathrm{T}} \, \bar{u} \\ \bar{x} & = & \varphi^{\mathrm{T}} \, u \\ \hat{u} & = & \phi \, \bar{x} \\ \hat{\bar{u}} & = & \varphi \, \bar{x}, \end{array} \right. \tag{60}$$

where the residual error is

$$\bar{u} = u - \hat{u}. \tag{61}$$

This all means that local level maximizations result in global level modeling. In the sense of information capture, the cybernetic model is the *best possible*:

- The feedforward section implements optimal (robust) modeling of the input data in terms of variance (information) preservation.
- The feedback implements optimal (robust) estimation (or "generative modeling") of the input data in terms of variance preservation.
- Thus, the closed loop with negative feedback implements optimal (robust) "statistical level control" of the input, or elimination of excitation from the environment.

Here, *optimality* in estimation is to be interpreted in the linear regression framework, and in modeling it means principal component (subspace) analysis perspective, in both cases meaning optimality in the statistical second moment sense. On the other hand, *robustness* in regression means reducing sensitivity to colinearity of variables; in the modeling part this robustness means pre-matching against candidate constructs, thus filtering noise. Briefly, as regression is enhanced through introduction of white noise, modeling is facilitated by introduction of black noise.

5.2 Interpretation of mathematical patterns

To summarize: the end result of running the cybernetic neuron system is also a model of the input data, where the constant *features,* or columns ϕ_i, together explain the changing *patterns* in u. The features are weighted by the variables $\bar{x}_i$ so that their sum maximally reconstructs each

individual input pattern; each variable $\bar{x}_i$ defines a *degree of freedom* of its own along its feature axis. The matching process between the patterns and the weightings of features is an iteration where a "balance" between the pattern and its reconstruction is searched for. As an extension of principal component analysis, this approach could perhaps be called *emergent component analysis* (ECA, or $\mathcal{E}$CA).

The learning in the structures can be fast, as it is the *reconstruction error* $\bar{u}$ only that is used for training, meaning that it is the difficulties in matching that are especially concentrated on. And, as Geoffrey Hinton has observed, many filter layers make the learning easier; now there are virtually an infinite number of layers, but, because of the signal recirculation, the filter is always the same! Regarding the simultaneous generative nature of the cybernetic network, another of his sayings is that *to recognize shapes, first learn to generate images.*

In other words, now this generation of "mental images" is based on models of sparse subspaces in the space of observations. The basis vectors spanning the subspaces are the features, summable prototypes, being the rotated eigenvectors of the data covariance matrix. With these, the original observation pattern is decomposed into a (low-level) perception. Indeed, this view matches well with the studies on *eigenbehaviors* and *eigenfaces*, etc., that have successfully been used for compressing complex phenomena. And, as experimentally demonstrated by Tom Mitchell, low-dimensional linear basis is enough to distinguish between linguistic concepts.

Patterns are assumed to be linear sums of features, and everything is linear. This is not only a pragmatic simplification: it turns out that the *linear models are optimal* in the cybernetic framework. When it is assumed that everything of interest is based on information, and information content is based on correlations, the theoretically best models for capturing this information are linear. Even though the natural computing elements may be non-ideal, evolution tries to make the models linear!

One more issue deserves to be mentioned about the modeling property of the cybernetic neuron population: as seen from above, the behaviors in the network can be explained and analyzed also in terms of an *energy function*

$$J(x) = \frac{1}{2} x^{\mathrm{T}} \left(Q^{-1} + \mathcal{E} \left\{ \bar{x}\bar{x}^{\mathrm{T}} \right\} \right) x - x^{\mathrm{T}} \mathcal{E} \left\{ \bar{x}\bar{u}^{\mathrm{T}} \right\} u, \tag{62}$$

where there are terms for the *internal energy* and for the *external energy*; following the intuition from mechanics, this expression could be called the *deformation energy* of the system, measuring how appropriately the *stiffness axes* match the directions of external pressures.

The criterion (62) also connects the time scales: it can be used for determining $\bar{x}$ (when minimizing $J(x)$), and, on the higher level, for determining the model itself (when minimizing $\mathcal{E}\{J(\bar{x})\}$). Indeed, also the model structure (the model size n) can be included in the criterion: it is $\mathcal{E}\{J(\bar{x}, n)\}$ that is to be minimized with respect to all variables to find the best model in the cybernetic setting. The first claim is easy to see as (31) characterizes the fixed point of the gradient descent implemented for the given criterion, giving $J(\bar{x})$ as its minimum; the latter claims deserve closer study. Again, assume that one selects $Q^{-1} = \mathrm{Var}\{\bar{x}\}$; it turns out that the minimum is reached when the average system activation is maximum ($\bar{\mu}_i$ being the eigenvalues

of $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$):

$$\begin{aligned}
\mathrm{Tr}\left\{\mathcal{E}\left\{J(\bar{x})\right\}\right\} &= \mathrm{Tr}\left\{\mathcal{E}\left\{ \tfrac{1}{2}\ \bar{x}^{\mathrm{T}}\left(\mathrm{Var}\left\{\bar{x}\right\}+\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)\bar{x}-\bar{x}^{\mathrm{T}}\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u\ \right\}\right\} \\
&= \mathcal{E}\left\{\mathrm{Tr}\left\{\ -\tfrac{1}{2}\,\bar{x}^{\mathrm{T}}\left(\mathrm{Var}\left\{\bar{x}\right\}+\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)\bar{x}\ \right\}\right\} \\
&= \mathrm{Tr}\left\{\mathcal{E}\left\{\ -\tfrac{1}{2}\ \left(\mathrm{Var}\left\{\bar{x}\right\}+\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)\bar{x}\bar{x}^{\mathrm{T}}\ \right\}\right\} \\
&= -\tfrac{1}{2}\ \mathrm{Tr}\left\{\ \left(\mathrm{Var}\left\{\bar{x}\right\}+\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\ \right\} \\
&= -\tfrac{1}{2}\ \mathrm{Tr}\left\{\ \mathrm{Var}\left\{\bar{x}\right\}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}+\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{2}\right\} \\
&= -\tfrac{1}{2}\ \mathrm{Tr}\left\{\ \mathrm{Var}\left\{\bar{x}\right\}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\ \right\}-\tfrac{1}{2}\ \mathrm{Tr}\left\{\ \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{2}\ \right\} \\
&= -\mathcal{E}\left\{\bar{x}_1^2\right\}^2-\cdots-\mathcal{E}\left\{\bar{x}_n^2\right\}^2=-\textstyle\sum_{i=1}^{n}\bar{\mu}_i^2 \\
&= -n\left(\frac{\sum_{\iota=1}^{n}\sqrt{\lambda}_\iota}{2n}\right)^4 .
\end{aligned} \tag{63}$$

This minimum can be reached only if the appropriate subspace with the maximal singular values is spanned by the model; what comes to the model size determination as given by (53), it seems that the above expression is (approximately, for large n) minimized for the optimum n.

The expression (62) evidently gives "negative energies", and perhaps it is better to apply the following "shifted" energy formulation

$$J'(x)=J(x)+n\left(\frac{\sum_{\iota=1}^{n}\sqrt{\lambda}_\iota}{2n}\right)^4 . \tag{64}$$

However, note that this criterion only measures how well the model matches the principal subspace, not how good rotations have been found; this means that simpler criteria for evaluating this aspect of model convergence can be used in practice.

The energy function defines a "landscape" in the data space, characterizing the surveyed properties of the environment; the model is a "map" of that terrain, and $\bar{x}_i$ are the "coordinates". As the matching process is iterative, and as there are many local minima (in the nonlinear case), typically the models remain suboptimal. Indeed, the cybernetic approach represents a model of a multitude of local minima rather than a single model of the global minimum; characterization of the landscape is more interesting than knowing the single optimum point. A pool of moderate non-unique solutions better characterizes the complex system, as nature, too, only finds sets of suboptimal solutions — and, if those evolutionary processes would be repeated, the results would never be the same, only their *nature* remains (remember the Heraclitean metaphors: "you cannot step in the same river twice").

5.3 "Metamodels" of neuron systems?

A wider view and fresh associations can help to get rid of outdated intuitions concerning neural networks. For example, what is the role of *data inflation*, *central control*, *preprogrammed structure*, and *nonlinearity* in neural network models? Normally, such ideas are never questioned. Let us study them here.

First, take the expansion of data, or introduction of a multitude of more or less hidden variables. As the goal of traditional modeling is compression, or reduction of data, it seems that the modeling principles are not applicable here. However, in the case of sparse coding, augmentation of the variable basis is well-founded, and optimization among some kind of kernel functions can be based on traditional minimization schemes. Second, central control of operations (or some kind of *positive feedback*) is seen as necessary to reach emergence of structures; however, as shown by the cybernetic approach, what if self-regulation and purely negative feedback makes it possible to reach non-trivial results in a truly distributed setting? Each neuron is independent, other neurons being visible only through their effects in the environment.

As an example of the above two intuitions, study a distributed version of the *self-organizing map*, also known as SOM. There one has a high number of candidate nodes representing the input pattern, and one has to implement the selection of the *winner* node; the winner and its *neighbors* are dragged towards that input pattern, resulting in a *map* getting formed. The key point is the definition of *topology* among the nodes $\bar{x}_i$, that is, determination of the *neighborhoods* among them. In the cybernetic model, the matrix Q can be made initially non-diagonal, thus facilitating crosstalk among neighboring neurons: this has the role of the *neighborhood matrix*. Because of the distribution of activity, now there are various winners that together represent the input pattern. This approach is applied in 6.2, so that similar-looking features are ordered near each other.

The two latter intuitions, the need for structural complexity and functional complexity, are discussed next — and, again, it turns out that there is something more that can be said about them in the cybernetic setting.

Because it is the reconstruction error $\tilde{u}$ that is driving the signal adaptation, and the asymptotic error $\bar{u}$ that is driving the model adaptation, smooth and (strictly) monotonous nonlinearities do not essentially change the big picture: in the steady state, as the fluctuations in error cancel each other, the feedback mapping constructs an estimate of the input, and, as a whole, the system constitutes a lower-dimensional model of the environment (the nonlinearity $f(x)$ is included also in the model as $\mathcal{E}\{f(x)u^{\mathrm{T}}\}$). When the nonlinearity is added in such a late phase, the basic functionality of the neuron system is not jeopardized, the neuron population still constructing a model maximally trying to capture the input variation, even though the variables are "crippled". For example, the sparse nature of coding can further be emphasized by adding a nonlinearity in the loop. Cutting negative x element values to zero, so that $f(x) = \mathrm{cut}(x)$, one can implement "symmetry breaking" to reach *non-negative sparse coding*. Search for positive features can be further emphasized by selecting $f(x) = |x|$; here, if strictly positive features are found, the model still behaves in a linear way. Even more complicated nonlinearities can be proposed: for example, introducing the *sigmoid function* in the loop, one gets nearer to binary representation, that is, towards the traditional on/off style sparsity.

But there is more: the cybernetic version of a sigmoid neuron population can be seen as a "collapsed" *multi-layer perceptron net*. Rather than employing various independent layers, now one has a single iterated layer, all hidden neuron activities being collected in the vector x. The "subpatterns" reside in the grid side by side, and further "layers" select from these their inputs; because of the nonlinearity, n can exceed m, or the signal space can get inflated. The forward

signal in the structure, or $\bar{u} = u - \Delta u$ can be seen as an error signal (see 5.1) — compare this to the backpropagation algorithm: now it is the *same* signal that traverses forward and backwards! There is no need for separate training phases or inverted flows, as all information that is needed is present all the time. And there are no such limitations for the number of variables now: as there are no hidden layers, all effects being directly visible, training the parameters in the net is much more straightforward. The *hierarchy of functional structures emerges if it is justified by the data properties*, inputs for each layer being selected among the already available kernel functions defined by the "prior layers". What is more, the role of the "hidden nodes" can also be studied in the input space now, and they can even be "preprogrammed". — If the nonlinearity is applied in each neuron, the internal mappings are all nonlinear, but the output mapping is linear — just as one usually selects also in the standard perceptron nets.

There exists plenty of literature for understanding the behaviors in multilayer perceptron networks, and this pool of knowledge can thus be applied also for understanding the behaviors in nonlinear cybernetic networks, and for enhancing the convergence of the parameters. But there is contribution also in the inverse direction: there are now new fruitful interpretations available. What is more, most of the objections against the physical plausibility of the backpropagation algorithm seem to vanish in the cybernetic setting.

As the presented cybernetic model structure makes it possible to support and adapt distributed kernel functions in a self-organized manner, *many different kinds of neural network approaches can be emulated within the cybernetic neuron system framework in a compact fashion*.

6 Step aside: an example application

The above discussions are next illustrated using a simple case example. During the tutorial presentation, *more experiments will be carried out and presented by Mr. Petri Lievonen*.

6.1 Implementation of the algorithm

In Fig. 4, following the mathematics above, the cybernetics-inspired regression algorithm is presented in its basic form. In the `Matlab` style pseudocode, `U` is the $\dim(u) \times k$ matrix of k input vectors u, `Y` is the $\dim(y) \times k$ matrix of k output vectors y, and `Xbar` is the $n \times k$ matrix of neuronal activities. The matrices representing the covariances are denoted `Exx`, `Exu`, and `Exy`. In addition to `n` and `q`, there are additional parameters for affecting the adaptation: `taux` is the time constant for the state adaptation, and `tau` is the time constant for the model adaptation (note that the discrete-time integrator does not exactly match the continuous one; furthermore, within one step the whole data is employed). The model matrices are initialized to random values, and the algorithm is iterated for the data until convergence is reached.

The algorithm is for batch data, assuming that all data is immediately available, so that matrix operations can be applied, whole data material being operated on in one step. This also means that the statistical properties of u (like singular values) are available.

```
% Initializations
Exu = eps*randn(n,dimU);
Exy = eps*randn(n,dimY);
q = (cumsum(svd(U))./(2*[1:dimU]')).^-2;

ITERATE until models in Exu and Exy converge

   Xbar = zeros(n,k);
   ITERATE until states in Xbar converge

      % Residual of the environmental signals
      Ubar = U - Exu'*q(n)*Xbar;

      % Balance of latent variables
      Xbar = (1-1/taux)*Xbar + (1/taux)*q(n)*Exu*Ubar;

      % Enhance model by nonlinearity?
      if nonlinear
         Xbar = Xbar.*(Xbar>0);      % Simple "cut"

      % Explicit scaling to the "natural scale"?
      Xbar = scaletovariance(Xbar,1/q(n));

   % Estimate of the output
   Yhat = Exy'*q(n)*Xbar;

   % Model adaptations
   Exu = (1-1/tau)*Exu + (1/tau)*Xbar*Ubar'/k;
   Exy = (1-1/tau)*Exu + (1/tau)*Xbar*(Y-Yhat)'/k;

END
```

Figure 4: **Algorithm**. Pseudocode for cybernetic feature extraction

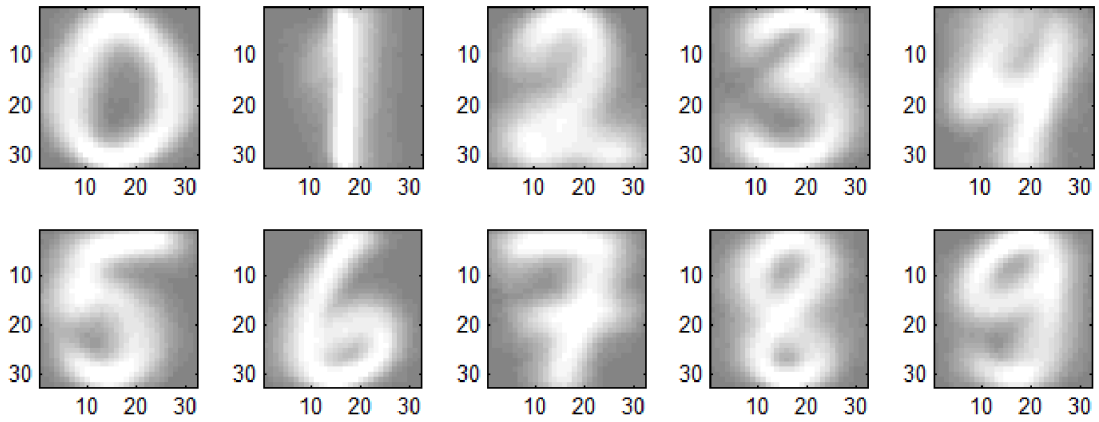


Figure 5: Averages of number classes

Because of the outlook of the regression formula (59), after the de-correlation of the $\bar{x}_i$ variables, the final regression can be accomplished not only back to u but to *any* variable by substituting the $\mathrm{E}\{\bar{x}\bar{u}^{\mathrm{T}}\}$ by some $\mathrm{E}\{\bar{x}\bar{y}^{\mathrm{T}}\}$. Then one can implement regression from u to y through the latent variable $\bar{x}$. Projecting the data through the intermediate latent variables can filter out noise from the data, if the selection of the lower-dimensional latent basis has been carried out in a clever way. Thus, data has been divided here in two parts, in input (containing the observation data) and output (containing the classification information). The reason for this is that only the input data may affect the determination of the internal representation (the system state), because only this data is available during model application. For the output data, zero error is forwarded to the system (even though the appropriate reconstruction error is applied for model adaptation). Because the output is not used for determination of $\bar{x}$ and its model, the output mapping could be constructed separately afterwards.

To minimize free parameters, it is assumed that $b = 1$; and to have one less adaptation processes to manage, q_i are kept constant rather than updating them online. Their final value is calculated directly using the known singular values (note that `q` in the algorithm is a vector of optimal values for different values of n), and the latent variables, or the rows in `Xbar` are explicitly scaled to each have the variance $1/q$. In practice, one should *not* apply such shortcut.

6.2 Data and its model

As an example, a case of coding hand-written digits is presented. As data material, there were 8940 samples of digits written in a 32×32 grid of binary intensity values (courtesy of Jorma Laaksonen, Dr.Tech; see `http://lib.tkk.fi/Diss/199X/isbn9512254794/`). In Fig. 5, the number class averages are shown, grey color denoting intensity value 0 and white denoting 1. The statistical properties of this data are shown in Fig. 6, where the envelope of the eigenvalues of the data covariance matrix are plotted together with the criterion (53); it seems that the optimum

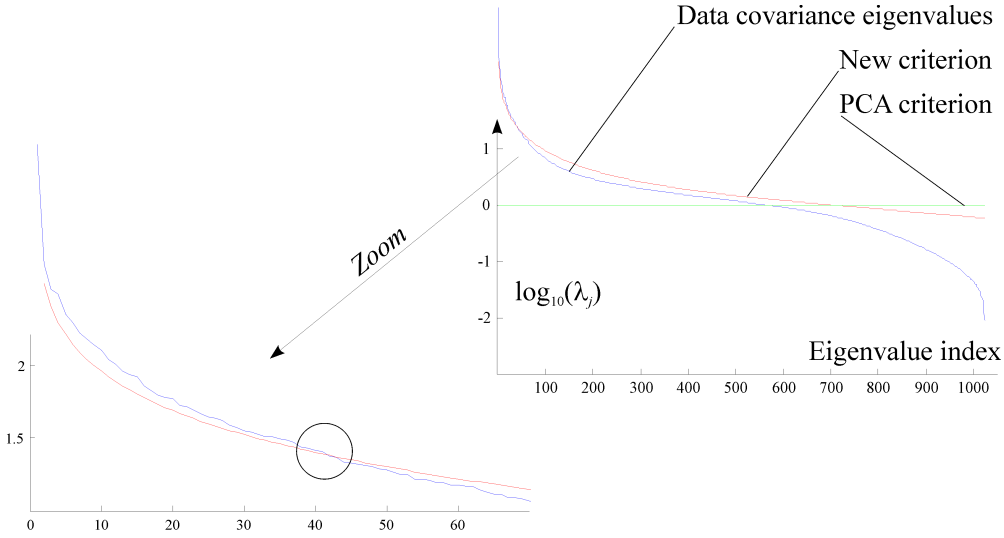


Figure 6: Selection of the model size

model size is in the vicinity of 40 (even though the behaviors of the curves is not very radical). Here, $n = 36$ was selected.

The intensity vector values were normalized, so that $\mathrm{E}\{u_j^2\} = 1$, but not mean-centered, and they were collected in the matrix `U`, with $k = 8940$ and $\dim(u) = 1024$. Correspondingly, the class matrix `Y`, with $k = 8940$ and $\dim(y) = 10$, was constructed of the labels: there is a single "1" in each column, corresponding to the correct classification of that input pattern.

The adaptation parameters were selected so that `tau` was 50 ($\tau = 50$) and `taux` was 2 ($\tau_x = 2$). In the spirit of the *self-organizing map*, the coupling factor q was now *not* scalar but a matrix Q, and this matrix was *not* originally diagonal: it started from a *neighborhood matrix*, and it was adapted towards a diagonal matrix as

$$Q(\kappa) = \left(1 - \frac{1}{\tau}\right) Q(\kappa - 1) + \left(\frac{1}{\tau}\right) q\, I_n, \tag{65}$$

with κ being the epoch index, and $Q(0) = q\, N_\sigma$. Here, this notation means a square grid topology with Gaussian neighborhood; that is, the neighborhood effect decays as a Gaussian function, with standard deviation σ. In this experiment, the distance between the nearest neighbors in the 6×6 grid is selected to equal the standard deviation.

As the adaptation process is based on gradient descent, and as the "fitness landscape" is very complicated, final adaptation takes a long time, and more sophisticated algorithms could be proposed. For example, the problem of local minima could be circumvented to some degree using the *momentum method*: the steps taken in adaptation are not steps in location but in *velocity*. Then, adaptation can speed up along long, flat valleys of the energy function.

When the algorithm is run with this data, results differ from a run to another; in Fig. 7, typical outcome is shown. Typically, representations become more and more complicated over

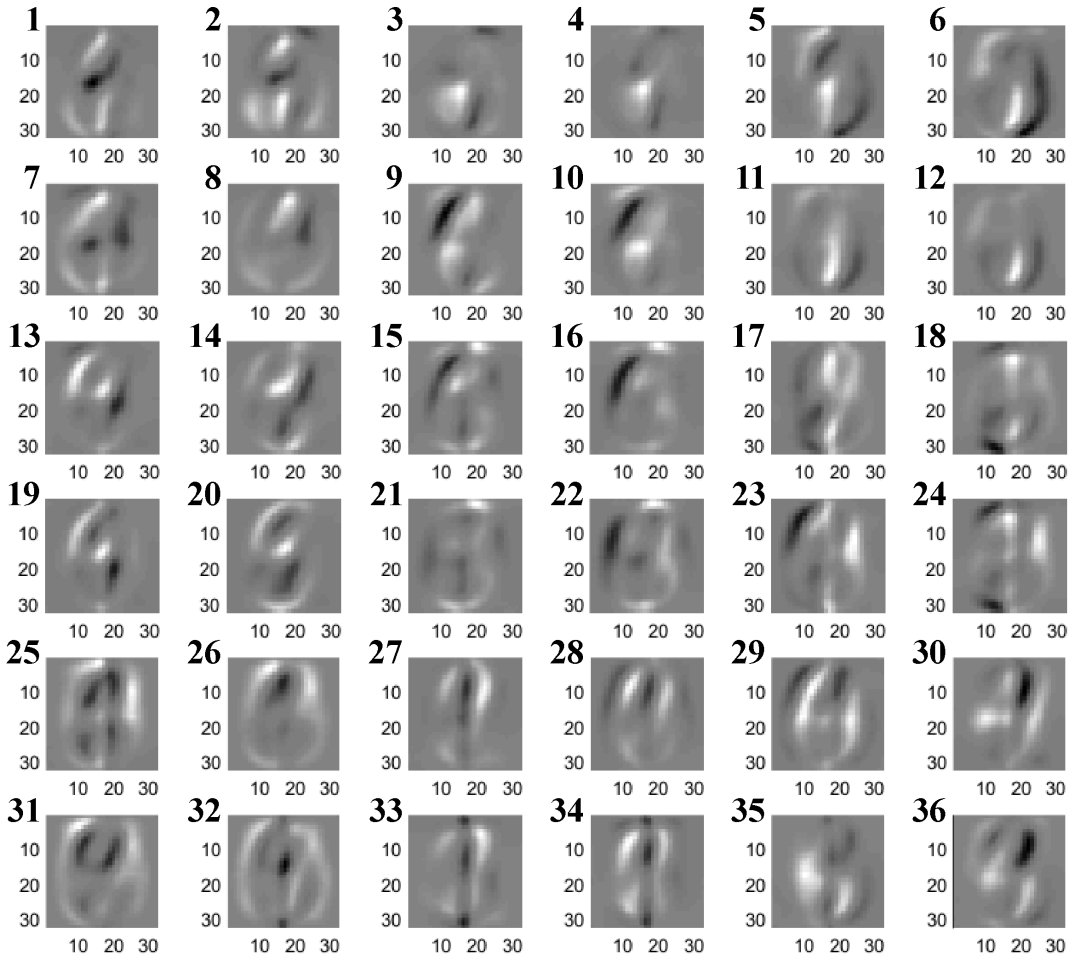


Figure 7: The resulting model of data

time: first, one has some kind of category prototypes; after that, there are "strokes"; and finally, there are some kind of *spatial gradients*, as in the figure. It turns out, however, that extreme decomposition does not help when constructing efficient mappings between input and output.

6.3 Classification results

In Fig. 8, the resulting regression model from $\bar{x}$ to y is visualized (compare this to Fig. 7). Whereas in Fig. 7 it is $\phi_{\text{in},i}$ that is shown, now it is $\phi_{\text{out},i}$. When used in classification, the estimates $\hat{y}$ are recorded, and the index of the maximum of these is the class estimate.

When the classifier performance was evaluated, 1000 fresh test cases were used, 100 for each class. The results were not good: only 70.0% of the validation samples were correctly classified (see Fig. 9). Different kinds of preprocessing methods could enhance the results; and, specially,

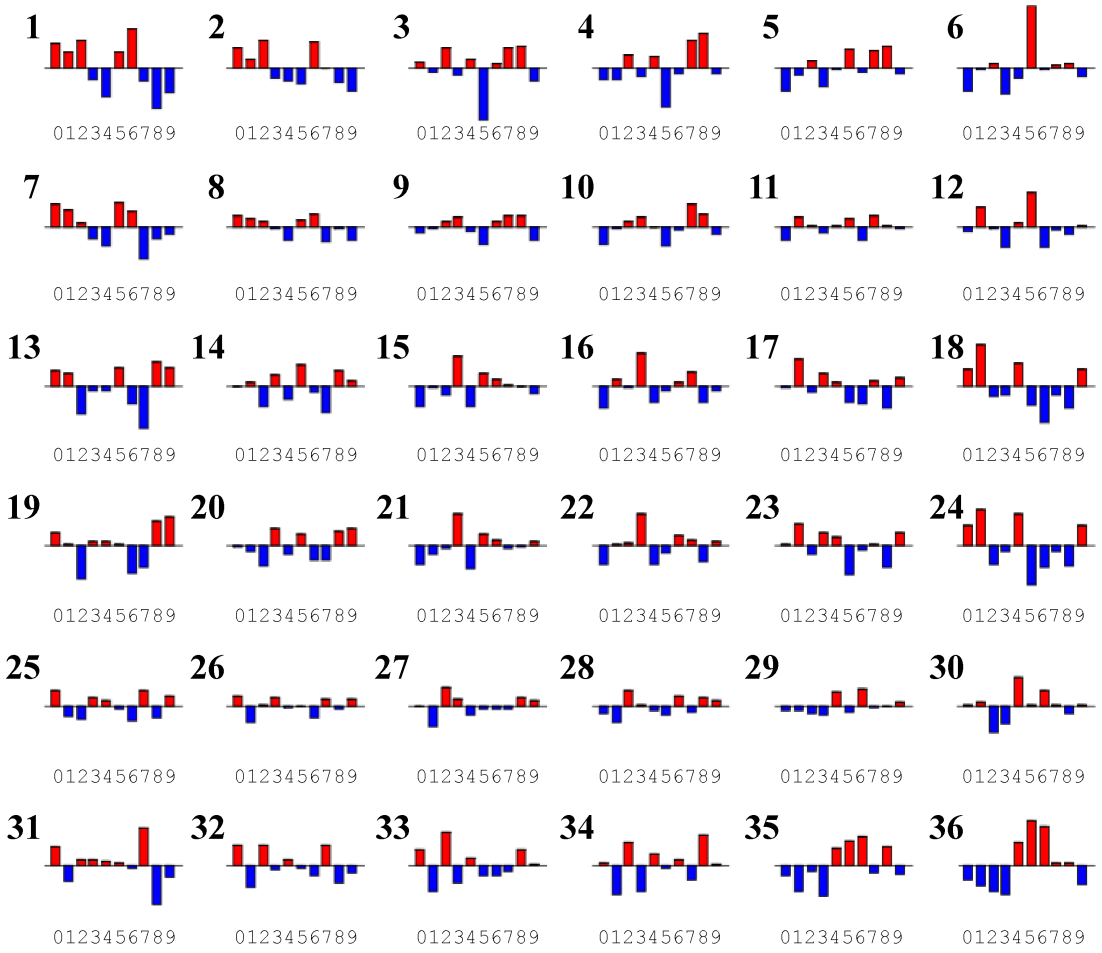


Figure 8: Relevance of neurons when explaining the classes

the following possibilities could be studied:

- Now, the model is strictly linear. It is known that linear classifiers cannot perform very well, and a sigmoid activation function, for example, could be added in the neurons.
- The chosen regularization level with $b = 1$ in $q_i = b/\mathcal{E}\{x_i^2\}$ is cautious and features become overlapping. With larger b there would be better separation, and the features would be more orthogonal and specific.
- The algorithm searches for common features within all training data. The inter-category similarities among the classes are not well-suited for distinguishing between them; perhaps there should be a separate model for each class, and the model with minimum reconstruction error would be selected to represent a sample?

		Classification by the model									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Correct class	**0**	**71**						29			
	1		**99**								1
	2	4	4	**50**	7	2	4	12	11	5	1
	3	8		1	**87**				3		1
	4	2	4			**63**		12	12	6	1
	5	7	1	2	6		**69**	14	1		
	6	8	2			1	2	**87**			
	7		17	1		7		2	**70**	3	
	8		2	5	3		5	7	6	**72**	
	9	14	5		17	16	2		13	1	**32**

Figure 9: Classification results using the basic model are not good

Not all variation carries information that helps in classification; on the other hand, some essential separating nuance can be rather delicate. It seems that the classifier could be enhanced by following the intuition about the structure of natural neural networks again (see Fig. 10).

If the model is extended to have *three* layers, so that the middle one with x exhausts the input layer with u, but there is also the output layer with y exhausting x, one can augment the original minimization of the input variance by the following (robust) minimization of the state variance:

$$\begin{array}{ll} \text{Minimize} & \mathcal{E}\left\{\| \bar{x} \|_2^2\right\} = \mathcal{E}\left\{\| {\phi_u}^{\mathrm{T}} \bar{u} - {\phi_y}^{\mathrm{T}} \bar{y} \|_2^2\right\} \\ \text{when} & {\phi_u}^{\mathrm{T}} \mathcal{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\} \phi_u = {\phi_y}^{\mathrm{T}} \mathcal{E}\left\{\bar{y}\bar{y}^{\mathrm{T}}\right\} \phi_y = const \cdot I_n. \end{array} \tag{66}$$

The constraints are known to hold because ϕ_u and ϕ_y differ from orthogonal eigenvector matrices through the same expressions containing only $\bar{x}$ (strictly speaking, as $\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$ is not diagonal, there are simultaneous rotations being applied to the basis vectors). The reason to write the expressions in the above form is that now one can employ one's intuition about mathematical patterns: the formulation can be shown to lead to a *generalized eigenvalue problem*, and its solution, or the converged matrices ϕ_u and ϕ_y determine the matrices of *canonical correlation analysis* between the spaces of u and y. There are still more statistical concepts available: the system also implements *linear discriminant analysis* to data, based on the known classifications in y for data u. Yet another concept related to matching of two data sets is *time warping* that can now be implemented by appropriate pre-ordering of the data. — Note that however the upper-level variables are disturbed, they are still interpreted as a model for the lower-level ones.

It needs to be recognized that the three-layer model can be emulated using only two layers by putting both u and y vectors on the input side, but adding one extra minus sign in front of ϕ_y in the formulas. The key intuition here is: *what gets squeezed, becomes modeled*.

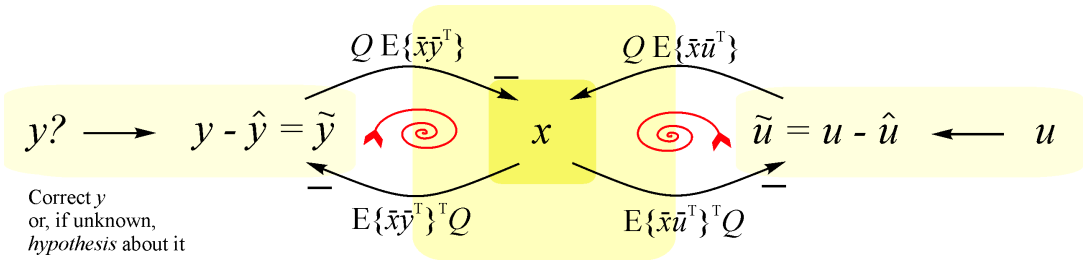


Figure 10: A clever (natural) implementation of an input/output mapping

7 Towards full-scale neuron systems

This far, only rigid grids of neurons have been studied. But this cannot be the whole truth — there is not only data filtering taking place in the brain. There must be some kind of *interaction and coordination among neuronal subsystems*. How to broaden the views? Again, applying the cybernetic intuition, abandon the static views, and actively face the dynamic realm as such approaches only can reveal the underlying hidden patterns.

7.1 Capturing dynamics

Above, the analyses were carried out applying a view from above, looking neuronal behaviors from a higher emergent level, so that the variables like $\bar{x}$ were assumed constant for given u. The actual signal behaviors were assumed to be irrelevant, and it was assumed that the steady state had been reached, signals having reached their asymptotic values. In practice, however, such situation seldom exists — there are transients taking place all the time.

The feedforward/feedback loop in the heart of the model is a dynamic self-referential structure, and this fact alone makes it necessary to concentrate on its properties to some extent. It turns out that *beauty hides itself in details.*

One can also benefit from the complex-looking situation. The loop structure itself implements *iteration*; by exploiting dynamicity, special kinds of seemingly complex tasks can be implemented in a straightforward manner. Plausible neuron models must include dynamics as *dynamicity is the only computational tool nature has available.* For example, the mathematical operation (matrix inverse) in formulas like (59) for finding the solution to a group of linear constraints can be motivated without having to abandon locality when one employs dynamics. Indeed, in the algorithm 6.1, dynamics enters the loop in a form of a (discrete-time) *integrator*. To get more intuition concerning integrators, study the following extended model:

$$\frac{d\,x}{dt/\tau}(t) = \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\bar{u} - Q^{-1}x(t). \tag{67}$$

Now, there is an additional local negative feedback loop within each computational unit, each neuron having an inhibitory link to itself. And, assuming (48), the additional synapses again

adapt in the Hebbian fashion (but now there is the minus sign). Because the system matrix $-Q^{-1}$ is negative definite, this model is always asymptotically stable; however, at the same time, this internal dissipation structure makes the system lossy. — When setting the derivative to zero (meaning that equilibrium is found), and studying what kind of balance signal $\bar{x}$ matches other variables (or causes appropriate tensions exactly compensating opposing ones), one can see that it is exactly the expression (8) that is being implemented by this "naural computation" (or "netural computation"!).

However, one would like to avoid *any* structural complications; the internal stabilization is not necessary because there already is the balancing feedback outside, being reflected in the experienced signal $\tilde{u}(t) = u - \phi x(t)$. The key forward is to apply intuition again: there are similar-looking mathematical patterns when one extends views to *probability distributions*, allowing us to discuss *optimal* dynamics.

> The *Ensemble Kalman Filter* is an iterative implementation of the probability density update problem: given an estimate of the pdf, called the *prior*, and the *likelihood* of some new data, find the new enhanced estimate, or the *posterior*. The Kalman filter is known to be the optimal update strategy for Gaussian data; the *ensemble* formulation means that the distribution is stored implicitly in the form of compressed "virtual data" in the *state vectors*. Using the adopted notation, the goal now is to find the model vectors x so that the conditional Gaussian probability for data u
>
> $$p\left(u|x\right) \;\propto\; \exp\left(-\frac{1}{2}\left(u-\phi x\right)^{\mathrm{T}} R^{-1}\left(u-\phi x\right)\right) \tag{68}$$
>
> would be maximized for all data in the *maximum likelihood* sense; there is *a priori* uncertainty in the data that is revealed in terms of the covariance R. The best result is reached when one updates the model in x iteratively as
>
> $$x^{\text{posterior}} = x^{\text{prior}} + C\phi^{\mathrm{T}}\left(\phi C\phi^{\mathrm{T}} + R\right)^{-1}\left(u - \phi x^{\text{prior}}\right). \tag{69}$$

In (69), the model uncertainty (the sample covariance $C = \mathrm{Cov}\{x^{\text{prior}}\}$) is projected into the space of data u, employing the covariance of the reconstruction ϕx^{prior}. However, for practical reasons one would like to implement robust matrix inverses in the lower dimension, even with the cost of less "optimal" estimates. The projection mapping ϕ^{T} can be moved to the other side of the inversion:

$$x^{\text{posterior}} \approx x^{\text{prior}} + C\left(C + R'\right)^{-1}\phi^{\mathrm{T}}\left(u - \phi x^{\text{prior}}\right). \tag{70}$$

Here, R' is now the *a priori* model covariance. If one assumes that covariances are very small, so that $R' \ll \mathrm{Cov}\{x^{\text{prior}}\}$, the expression can be simplified. Further, one can observe that x^{prior} is simply $x(t)$, and $x^{\text{posterior}}$ can be denoted $x(t+dt)$, or the state after a time dt; letting the originally

discrete-time update process become faster and faster, the difference between the posterior and the prior becomes the *derivative* with some time constant τ_x, and there approximately holds

$$\frac{dx}{dt/\tau_x}(t) = \phi^{\mathrm{T}}\,\tilde{u}(t). \tag{71}$$

This expression resembles the cybernetic model formula in (8). Indeed, now one can interpret the static model to have been implemented as a dynamic Kalman filter that also models its environment. Because of the optimality of the Kalman update scheme, it can be assumed that also natural neurons have adopted similar dynamic state update strategy during evolution. The model is lossless, there is no dissipation, with neurons acting as pure integrators.

The above assumption of R' being very small, or assuming that the *a priori* covariance would be negligible as compared to the observed sample covariance, is clearly incorrect, and in such case the current assumption of zero mean would be even less appropriate. A more plausible down-shifted formula for state adaptation is found when one simply selects $R' = \gamma C$ for some scalar $\gamma > 0$; then one has an additional scalar factor $\frac{1}{1+\gamma}$ in front of $\phi^{\mathrm{T}}\tilde{u}(t)$ in (71). Or, indeed, should one not select $R' = Q^{-1}$ for the state uncertainty! The extra factors (together with those caused by the time constants) become anyway dissolved in the asymptotic model structures, as studied in 3.1. — One more simplification can be made in the model formulas, substituting

$$\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \approx \mathcal{E}\left\{x\tilde{u}^{\mathrm{T}}\right\}, \qquad \text{and} \qquad \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} \approx \mathcal{E}\left\{xx^{\mathrm{T}}\right\}, \tag{72}$$

if it can be assumed that the internal dynamics is much faster than the environmental dynamics: then there is no need to wait for the convergence of signals, and learning can be continuous.

More intuition about the properties of the cybernetic model has thus been gained. Seemingly, one has made additional limiting assumptions: here, the data was assumed Gaussian. However, there is no increase of assumptions, because our model has been linear all the way — and it is well known that linear models and Gaussianity of data are in one-to-one correspondence with each other!

Something more needs to be said about this new probabilistic way to look at the data. It is the *central limit theorem* that assures that a sum of a large number of independent random variables will be approximately normally distributed — and this is a good assumption in our case, too, so that this way the adopted linearity assumption can be naturally motivated. Similarly, Gaussianity also helps to motivate the definition of information in 2.1: it is known that for normally distributed data all statistical cumulants beyond the second one (variance or covariance) vanish. To capture the statistical properties of the data, to be better prepared to its behavior and to have the best benefit of it, *there is no need to employ more complicated definitions of information*. All that is valuable in data must be already present in covariances.

7.2 Entering the frequency domain

When differential equations are integrated as a part in the neuronal model, it seems that the possibility of simple static calculations is lost. Dynamic signal structures are more difficult

to analyze and grasp. However, this increase in complexity does not take place, when one employs the *frequency domain,* where it is assumed that individual signals are irrelevant, and it is the resultant group behaviors or wave fronts that are of importance; in steady state, then, it is *frequencies* and their *phases* that count.

Now the original ambition, or sticking to linearity, is nicely rewarded: there are strong tools available for analysing signals in the frequency domain. The mathematical tool to manipulate and analyze systems with linear differential equations is the *Laplace transform.* Applying this transformation, differential equations change back to static algebraic equations, but the signal-domain variables become substituted with frequency-domain ones. In a way, a separate model is constructed for each frequency, and signals are thought to be superpositions of those frequencies. After the system has been solved in frequency domain, the dynamic trajectories in time-domain can be solved (if this is needed) applying the inverse transform. But *vibration patterns* can best be studied directly in Laplace domain (or applying the related *Fourier analysis*).

There is one catch, though: frequency domain signals are *complex,* as the amplitudes and phases both count. But this is not a problem now, as complex numbers can readily be used in the cybernetic models; it even seems that convergence properties of neural algorithms typically become *faster* and *more robust* in complex domain. However, there is one essential change: *all transposed expressions are substituted with Hermitean ones,* that is, formulas like $\mathrm{E}\{xu^{\mathrm{T}}\}$ change to $\mathrm{E}\{xu^{\mathrm{H}}\}$, etc. In matrices that are Hermited, in addition to transposition, all complex values of the form $x + y\mathrm{i}$ (or $r\mathrm{e}^{\mathrm{i}\psi}$) are changed to complex conjugates $x - y\mathrm{i}$ (or $r\mathrm{e}^{-\mathrm{i}\psi}$). This change in formulas can be motivated so that the symbol ψ in $r\mathrm{e}^{\mathrm{i}\psi}$ represents the *phase difference*; if the mapping ϕ conveys some phase lag, it is only natural that in the matching balance the backward mapping ϕ^{H} conveys the corresponding phase lead.

So, presenting the model (71) in complex domain, one has a model that can be interpreted as defining a *multivariate electric circuit*, where the driving force $\tilde{u}$ is the difference between the two neuronal vectors of potentials (variables representing deviations from some nominal levels):

$$\frac{dx}{dt/\tau_x}(t) = \phi^{\mathrm{H}}\,\tilde{u}(t), \tag{73}$$

corresponding to $\tau_x sX(s) = \phi^H \tilde{U}(s)$ when Laplace transformed. Here $s = \mathrm{i}\,2\pi f$ is the Laplace domain variable with f being the frequency (strictly speaking, the time constant is now not τ_x but the matrix $\tau_x(\phi^{\mathrm{H}}\phi)^{-1}$, meaning that the stiffnesses affect the dynamics). Transformed variables are typically written in capital letters.

In frequency domain the signal activities change to signal amplitudes. In both cases, averages of their squares are related to "information energy" (remember the *Parseval's theorem*, etc.), and motivations presented earlier concerning the "neuronal hunger" still hold: the adaptation of synapses can be assumedly carried out directly in Laplace domain. The synaptic weights, too, become complex-valued then. — The sensitivity to phases opens up new possibilities, as in the succession of input patterns, for example, anticipation of change (a kind of "pattern derivative") can become coded in the phase when data is appropriately preprocessed.

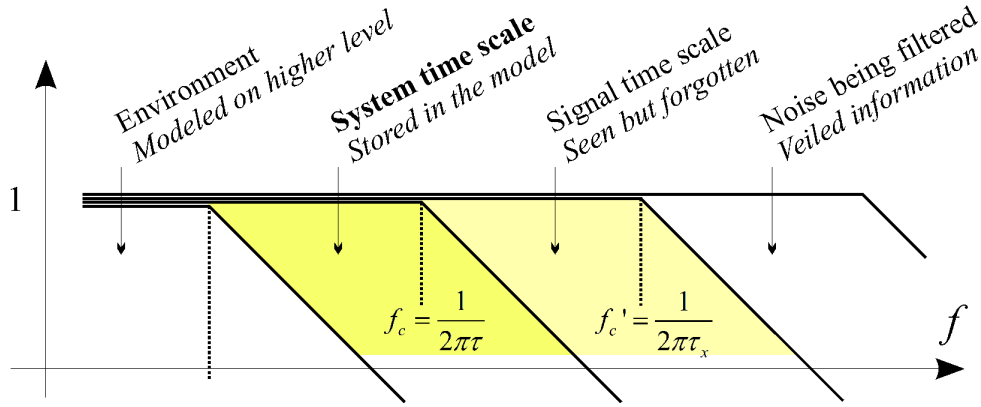


Figure 11: Hierarchy of scales (cut-off frequencies determined by the time constants)

Note that the "weakly emergent" model updates that are based on the formula (7) are also linear, and they can also be studied in frequency domain; however, if combined with signal-level dynamics, behaviors become *bilinear*. To preserve model linearity, one has to study one frequency scale at a time: too fast changes at any selected scale are just noise, and they should be low-pass filtered, whereas too slow changes get captured by some higher-level lower-frequency models. In Fig. 11, a schematic illustration (log/log scale) shows how the system recognizes the energy spectrum, or *information spectrum* around it, and how signal variation gets filtered. As time passes, information about long-term cycles can be detected so that models can be constructed on ever lower frequency scales.

Exploitation of frequency and phase information between neural subsystems can be the key to get onto the next-level neural models, not only quantitatively, as in the figure, but also qualitatively. Indeed, the cybernetic model inspires new hypotheses.

7.3 Scenario: higher-level models?

When extending our view to larger-scale neural systems, one thing that becomes clear is that the number of neurons increases very much as compared to the number of available sense inputs. This means that it is *other neurons* that have to serve as inputs to each other. As the neurons operate on the same emergent level, on the same time scale, the dynamic considerations become necessary when trying to capture their interactions. This observation of neurons acting as inputs raises a question: the inputs themselves are also dynamic now, being governed by similar differential equations. The increase in the activity of x is sucked from neurons u, and this loss can be expressed as

$$\frac{d\,\tilde{u}}{dt/\tau_u}(t) = -\phi\, x(t). \tag{74}$$

To get rid of the other variable, apply further differentiation to (73):

$$\frac{d^2x}{dt^2/\tau_x\tau_u}(t) = \phi^{\mathrm{H}}\,\frac{d\,\tilde{u}}{dt/\tau_u}(t) = -\phi^{\mathrm{H}}\phi\;x(t). \tag{75}$$

This expression now characterizes the lossless coupling between the neuron groups. By experimenting, one can recognize that there is a class of signal forms that fulfills this expression:

$$x(t) = A\sin\left(\sqrt{\frac{1}{\tau_x\tau_u}\,\phi^{\mathrm{H}}\phi}\;\;t + \psi\right). \tag{76}$$

There is also an unattenuated harmonic oscillation taking place between the neurons, defining a set of *resonators,* at least if assuming diagonality of $\phi^{\mathrm{H}}\phi$ (eigenvalues are real in any case). Frequencies of the resonators are determined by the coupling strengths in ϕ, tighter coupling resulting in higher frequency. The system dynamics is autonomous, and the external inputs can only affect the initial values, determining the free parameters, or the amplitudes in A and the phases in ψ. In steady state, with constant frequency patterns, Fourier transform only is needed.

How can such frequency coupling take place in practice, how can frequency domain be addressed in real time? Perhaps it is now time to get back to the basic principles, and beyond the abstraction of "neural activities". Neuron activity is implemented in terms of pulses; if there is a pulse train of constant intervals, this defines a frequency. Then, if there is a frequency, and if a neuron starts following the corresponding pulse train, one can have a *phase-locked loop* with synchronization of pulses. One can now extend the neuronal learning principle from static activities to frequencies: *try to make pulses match*! If there already exists a consistent frequency to be experienced, adapt towards it: increase the synaptic coupling if own pulses are delayed, so that the signal path becomes faster, and decrease it in the opposite case. It seems that such "frequency locking" can be presented in time domain when the time constant τ_x is allowed to be imaginary-valued? — Otherwise, in the purely stochastic case of no detectable nearby frequencies, things reduce to original case of Hebbian learning based on average activity levels.

Indeed, as indicated by EEG recordings, it has long been known that there are brainwaves that reflect the overall cognitive status, like alertness. What is more, it has been observed that there really exist some kind of rhythmic interactions between brain regions, and in some situations their synchronizations take place.

Such visions of oscillating neurons can help to attack some of the age-old dilemmas of mental functioning: for example, there can exist *dynamic coupling* of models and data being determined by their mutual resonances. Perception can be like a "blackboard system", where competing resonators (or the *complex cells* and other kinds of cell complexes) bootstrap characteristic frequencies, thus implementing *feature augmentation* "in place". A neuron can sustain various frequencies by implementing appropriate pulse firings. The processing hierarchy is collapsed, as all resonators operate side by side on the same data but on statistically orthogonal frequencies, constructing frequency-domain "fingerprints" to mental contents, facilitating far-reaching associations, neuron groups actively reacting to their characteristic chords. — As examples for need of this kind of dynamic construction of information structures, one can think of the

two-dimensional case of analysing visual views, or the one-dimensional case of natural language comprehension. The *deep structure* of thought is assumedly found when the vibration fields have converged.

It needs to be remembered that the frequency representation based on complex numbers is essentially richer than any familiar representation that is based only on real numbers. The possibilities of the additional phase information can be understood when one studies the operation of *holographic memories*.

In short, the new frequencies-based view makes it possible to see *mental states as being characterized in terms of standing waves.* It makes it also possible to assume that some kind of *delocalization* and *fast coupling* in information processing can take place. Combined with cybernetic self-organization, the pulse-coded neuron system can *make sense*.

8 Cybernetic minds

The above discussions have intuitive appeal and it is easy to make even more brave hypotheses. It seems that there are, for example, connections not only to cognitive theories, but also to the *philosophy of mind,* and there are connections to complex systems in general; below, some examples along these lines of thought are presented.

8.1 Further interpretations

There exists a large body of research on cognition, but the cybernetic approach does not fit very well with that tradition. However, *embodied embedded cognition* is a philosophically oriented position in cognitive science, closely related to *situated cognition* and *embodied cognition*, etc., that appropriately matches our intuitions. This theory states that intelligent behaviour emerges out of the interplay between brain, body, and world: all these are equally important factors in the explanation of how particular intelligent behaviours come about in practice.

The cybernetic metaphor is a good basis for *constructivism*: neuronal structures and mental constructs are being built gradually from non-existence. There is natural *complexification* driven by the hunger for information. The predicted functionalities — sparse coding, or detection of "strokes", etc. — have been observed in natural vision systems, but how about the higher levels, can the intuitions be expanded beyond the lowest level?

As neurons extract activation from their environment, it is only *their* activation that can be seen as a resource by others. This results in chains of cells. From the modeling point of view, what is the added value when there are multiple layers in the neural net structure? To understand this, study a case where input variables are logarithmic, that is, high sensory values are heavily attenuated (this is what senses typically do, at least what comes to visual and auditory signals). Summations of modified signals correspond to multiplications of the original. If the original signals have *probability interpretation* (or unscaled *relevance interpretation* truly), adding variables can be seen as a logical AND operation; sparse subsets approximately correspond to

OR operations among the sets of variables. Altogether a chain of cybernetic models can be seen as an AND/OR graph, thus making ever more accurate categorizations among patterns possible. The succession of elementary models facilitates better modeling and exploitation of information.

When simple manipulations cumulate, at some point the purely physical information changes to something non-physical: information granules change to symbols and concepts.

It seems that there are various cognitivistic concepts that can be interpreted in terms of the cybernetic model. The computer metaphor can be relaxed as the constructs are truly distributed: for example, the *long-term memory* gets implemented through the vectors ϕ_i in all cybernetic subsystems, and the *short-term memory* consists of the references to them — that is, it is the activations in variables $\bar{x}_i$ that stand for temporary storages. There is no need for computer-like transfer of data between memory registers. Simultaneous fuzziness or continuity of concepts and their crispness can be explained: normally, the locations of minima are smooth functions of the inputs, but in appropriate conditions stability of an attractor can suddenly get lost. As seen from another point of view: attributes (or *style*) of an object are its features, or degrees of freedom, and the category prototype is the most significant of them; again, all this can be seen in the cybernetic perspective. — One can even become arrogant: if some ideas concerning mental functioning do not have a correspondence in a cybernetic model, one can claim that *such intuitions are incorrect*. Many of the today's ideas are based on the computer metaphor; computer has been the only concrete example of information processing outside the brain, and its role is over-emphasized, even though the shortcomings of such interpretation can easily be seen.

It would seem that one cannot reach higher levels of mental processing applying the cybernetic ideas, but, at least to some extent, this is an illusion caused by a bias: subsymbolic processing cannot be captured through introspection as it cannot be explicated. It is evident that only novice-level knowledge is declarative, whereas expert knowledge is based on pattern matching. Expertise is based on balance models, matching of observations against a model, filtering irrelevant details (or "noise") away. And it is typically not the most visible features that make the difference.

The cybernetic model is based on balance structures — how about true transients, then? Study a process of detecting sequences of correlated inputs entering a realm of "free" neurons searching for activation. Assume that the "hungriest" neuron builds a momentary connection to the simultaneously active units; if this coupling is relevant, that is, if the neuron can get its livelihood from this combination later, too, this neuronal substructure can remain there. Later activations start connecting such neurons together, and, finally, the directed graphs of neurons become part of the pancausal "cybernetic medium". This way, the process of a one-time-only (declarative) event becoming a statistically relevant (associative) structure can, in principle, be explained; however, *generation* of time-domain representations is a bigger problem. How can the activity in neurons be discharged into an ordered sequence of bursts, how to explain attention control, what about the seemingly controlled construction of linguistic structures, and *thinking*?

The building blocks for a complete neurophilosophy are not yet there. But even though the details are not understood, some hypotheses about the big picture can be made. In the cybernetic spirit, one can perhaps trust that some kind of *higher-level feedbacks*, or infinite number of iterations through some kind of *conscious mind*, can again do the difference.

8.2 What is consciousness?

The convergent processes determine attractors in data space, being somehow *relevant* in their environment, defining a *grounding* for higher-level constructs to build further. In the neuronal domain, such higher-level "atoms" can be seen as *chunks*, or (more or less subsymbolic) *concepts* or *categories.* Based on such concepts, next emergent levels are needed as new modalities, etc., get involved, and finally in the succession of levels, assumedly there will be *consciousness* as the highest-level attractor.

This is a nice thought with all-embracing dynamics; but why should all mental levels follow the cybernetic ideals? — The higher-level systems always try to capture the lower level activation patterns in a compressed way; and to accomplish this, they need a model. According to the above studies, the cybernetic one is the best of all models what comes to pumping of activation to higher levels, so that this strategy has *evolutionary advantage.*

What does the cybernetic "model of models" then look like? The low-level model of own action is a blur of observations representing the whole closed loop between the inner and outer worlds; after that, however, the sparsity goal in modeling means that there will be a separate model of *self,* with more or less realistic attributes, being in interaction with environment. Is this not what today's experts on human psyche say about the nature of the *homunculus*, too?

The above model-oriented technical view can satisfy some, but simultaneously others will ask: do you really think that this is all there is about consciousness to say? Indeed, it is interesting to note that the cybernetic approach offers yet *another* intuition of what consciousness could consist of. The vibration fields of 7.3 assumedly permeate in all mental structures, causing an unexplainable experience of wholeness ... the process philosophical *feeling of what happens,* and *qualia* (as studied in 8.3), is the essence of feeling *alive,* and the experience of being part of it all. The feeling of pain, or anguish, is not only a category of thought.

Thus, consciousness can be deterministic without being algorithmic. One does not need to employ the ideas of Roger Penrose to attack the mechanistic interpretations of *strong AI*. And one does not need to resort to quantum phenomena, etc., to explain *free will* and creativity: it is about finding new freedoms among constraints determined by the controls, finding escape from the balance tensions, instantiating new variables as couplings become too intense. There are no random *butterfly effects* in cybernetic systems. Structures emerge only if they are *relevant*; but as soon as a new stable attractor becomes instantiated, minor effects explode "saltationistically".

Such "fields-based consciousness" can perhaps someday be studied when the methodology of frequency domain pattern recognition matures: rather than studying individual EEG signals, the frequency-domain "multivariate scenes" should be compared to the pattern analyses of visual scenes. The Kantian basic assumptions about mental functioning being anchored in space and time can be escaped: frequencies are not bound to locations or directions, and, in a way, frequencies address both the past and the future in current time. Different mental constructs define characteristic chords, together hopefully constituting a (Pythagorean) harmony of spheres!

It is evident that if it is the cybernetic principles that make the minds emerge, there is some level of consciousness in animals, too. On the other hand, infants are *not yet* truly conscious, as

their world is "holistic", the *inner* still being mixed with the *outer*.

One objection against reductionistic theories about consciousness is that mental models are *causal*, so that the structure of action and reaction is somehow integrated in them. Causality cannot be observed in data, only correlations can — how could one address this intuitive feel of causality in observations-based mental models, then? — Indeed, now there is an easy answer: because of the all-embracing observer effect in the cybernetic models, it is *models of one's actions on the environment* as being induced by the environment that are constructed.

In 2.2, knowledge was formally defined — similarly, in a very narrow sense, one can go still further: *wisdom* means that one knows one's self, what are one's capabilities, and how these can be used to change the world, and how the world (or others) will respond. The models are there ready to be used: whereas clever ones can find their ways out of troubles, the wise ones never end in problems in the first place.

Such high-level models that reside in ideasphere are not limited to exist within only a single brain, and the world models need not be subjective. There can exist *intelligent societies*, distributed models among groups of people, if the communication among the atomic minds is complete enough. The *systems thinking* in one mind can change to a *thinking system* as seen from outside — and this need not be only metaphorical speaking. In still wider scales, the Hegelian *self-consciousness of nature* can perhaps be implemented as the models of cybernetic systems in human's and nature's history are constructed by humans, detecting the cycles and their frequencies in one's world. The human implements the distributed consciousness of nature.

8.3 Extensions beyond neuronal realms

This far it has been assumed that the inputs were sensory signals, or neuronal ones; however, one can extend the view. It is evident that information or energy capture in general means evolutionary benefit, and the formulation $\mathcal{E}\{\bar{u}_j^2\}$ generally has the interpretation of energy, or some kind of *capacity* (or *emergy*) for many different kinds of signals $\bar{u}_j$. The framework of domain semantics changes to finding relevant *signs* in the environment, or determination of the *system semiosis*: one has to recognize where there are resources available in the environment.

Assume that there exist some receptors of the internal state of the body, for example, some hormone level indicators. Including such measurements in the input vector u delivers valuable information about the environment of the neuron system, and becomes most probably included in the models, together with the purely neuronal signals. Together with the sense signals, these measurements can be seen to deliver the *lowest-level semantics* to the otherwise hermeneutic neuronal models. One could assume that *feelings,* for example, inherit their contents at least partly from the prevailing adrenaline and testosterone levels, etc. And, more generally, the artificial intelligence problem of *qualia* can perhaps be explained through the connection to the bodily state. The dichotomy between the mind and the body need not be sharp.

One can generalize further. There are feedbacks also outside the brain: the motivation for the existence of the whole neuronal system is to implement clever cybernetic feedbacks between sensors and actuators so that the perceived environment would become better controlled. The

observations can be affected through real changes in the environment, or through one's own actions (either through some kind of adaptation, or through *moving* to another location). There are different levels of models involved here, too, starting from the lowest-level *reflexes* to *goals* of high-level cognition. The *limbic system* can be seen as a "chemical controller" of its own.

And, further, it is not only the neuronal system that can adapt; automatic model construction and model-based control can take place also without the brain involved at all in different kinds of agent societies reaching for resources. Such agents can be humans or animals, but, more generally, they can be cells or even mere molecules that experience the *selection of the fittest*. Assuming that the exploitation of scarce resources is the common goal in nature in general, and there is competition for them (implementing negative feedback), it is the above principles that still rule: the *domain area semantics* in its most basic form can be reduced to these resources, and, after all, the cybernetic approaches promise the best possible utilization of them. The similar adaptation strategies should then have evolutionary advantage, and, using the presented methodology, one can perhaps proceed from explaining intelligence to understanding *life.*

In the preceding discussions, optimality in information processing was emphasized. It may be that the "tools" available to nature do not make it easy to reach the structures that would assure such optimality, but it is interesting to see what the cybernetic balance in the slowest time scale dynamic system is: what does the hypothetical *evolutionary equilibrium* look like? Given enough time, where would nature aim at?

Indeed, in the cybernetic perspective, something can be said about nature's "goals" in general. It has been claimed that the extended view of entropy, or *maximum energy dispersal* would be the principle also beyond complex systems; however, the "whirls" in the entropy flow cannot still be explained. Such phenomena can be understood on the higher scale: energy must be seen as a form of *cybernetic emergy*, and what takes place is *maximum emergy dispersal*. Information coagulation on the higher level (seemingly against the entropy flow) can be explained in terms of control, in terms of still more effective emergy elimination on the lower level of more abundance. The neuronal system, and also the cognitive system, are nature's ways to reach "heat death" still more effectively, escaping the qualitative barriers that are chaining the flow of *emtropy*.

There are still deeper philosophical questions lying beyond such considerations. For example, assume that the modeling methodologies are the same in different domains; then, given the same input data, the resulting models should be effectively identical. This should apply also if the modeling takes place in the mind instead of taking place in the nature. The mind should construct mental models that capture the *essence* of the systems being studied — this view can be called *interobjectivity.* And this modeling can also take place in a computer using the cybernetic algorithms: it is not only artificial intelligence, mimicking humans, but it can be *universal intelligence*, explaining survival and prosperity in an environment. Omitting proof here: on the highest level, it is *pattern recognition of vibration patterns* that is the key issue. What is more, it can be claimed that the "incomprehensibility of comprehensibility", or (as interpreted here) the separability of individual problems, can be caused by the sparse coding in the cybernetic processes taking place deep in the nature; and the fine tuning of the natural parameters can perhaps be explained so that they *are* tuned by the ongoing natural optimization!

9 Discussion: what lies ahead?

Above, the theory of *neocybernetic systems* was discussed as applied to modeling of neural networks, and beyond. Not very many conclusions can yet be drawn; instead, let us see some prospects. Finding a model of a mind, as claimed above, should have wide-ranging consequences what comes to one's own mind, too! — Are there any practical lesssons to be learned?

Here, towards the end, let us get back to the research on artificial neural networks. This field is very incoherent, being a collection of diverse methods, the only integrative factor being the intuitive ambition: one is not only searching for some technical tools for data manipulation, but, after all, one tries to understand the functioning of the brain. Today, the field of neural networks research seems to have "converged", and conclusions can perhaps be drawn. Even though there are new algorithms being developed and minor breakthroughs being found every now and then, the overall picture seems to be missing. The field is a *fiddler's paradise* with no consistent paradigm, consisting of a multitude of distinct ideas; indeed, approaches are *collections of elaborated intuitions*, visions of what is relevant when capturing the essence of brain functions.

There is nothing bad in intuitions *per se*. To reach back towards a unified vision, one has to look at the research in a wider perspective, and it has to be admitted that research cannot be merely empiristic: one needs some guidelines to do experiments. Consistent research is *theory-driven* — or, more fundamentally, good research is *intuition-driven*, even though this fact is shamefully ignored afterwards. Perhaps speaking of such facts can someday be approved, as creativity and workings of the mind are so important components of scientific work.

To reach forward in the research on neural networks, goals need to be discussed, and the means to reach towards them. Concerning one's own work, one should answer *why* that approach is good, what it can say about higher brain functions, and *where* that vision would take us. Again, intuition is the key issue: what do *you* think is important? Indeed, one needs to discuss one's "values". The claim here is that modeling in ideasphere makes it possible to refine intuitions.

It needs to be emphasized once again: intuition has the key role in research, and in creativity in general. Intuitions are the sparks of new thinking, escaping the constraints of the current world, making it possible to find a new freedom, igniting the fire. Intuitions are kernels, changing to ideas when they have appropriately emerged and when they have been mentally elaborated on with the help of imagination. What one needs is a *tool for analysis and modeling of not only information but ideas*, a delicate machinery that would not embrace the ideas to death. And it seems that the *intuition of a cybernetic model is that it can be also a model of intuitions*.

> In ideasphere, intuitions can be seen as the "data", and competing visions find their model in the balance of tensions, such equilibrium revealing the structure among elementary intuitions: what is relevant when studying the domain field.
>
> One can define the *formalized Delphi method* for refining the intuitions and making them compatible. Raw intuitions as "data" can be collected from expert opinions: what features does each of the researchers think are the most important what comes

to the domain field? Entries of the "intuition vectors" are weighted by all experts. When the weight vectors are compressed in the cybernetic adaptation, one can perhaps see a structured view of the domain.

The key point in the Delphi method is iteration: as the experts see the other's opinions in a coherent framework, interpretations compete with each other, and the experts can rethink and refine their intuitions. The cybernetic principle can be a tool for *modeling expertise* in general — this can be the key to *reusability* of expertise.

10 About bibliographies

A traditional-style list of references is *not* included here. Rather, for reaching further material, a "cybernetic" mechanism is proposed: to have a (more or less) complete, up-to-date, balanced view of what lies beyond the ideas, *put keywords in an internet search engine* and follow links.

Internet is a distributed memory, but together with the search engines it is becoming the tool for implementing *distributed cognition,* ever better matching the users' intentions. It is dynamic and changing as the network contents are continuously updated — and, indeed, it is not the fixed *truth* you find there, but it is the *relevance-directed* view of the real world. The resource for the nodes to compete for is the users' interest and trust, and, thus, the evolution in the net seems to follow the cybernetic principles. The vision of distributed expertise works today already, offering glimpses to the "net-scape", and, regardless of the stochastic nature of the more or less random developments, this network of knowledge is getting better all the time.

In principle, following the cybernetic ideas, structure (knowledge and understanding) emerges automatically from the body of information available in the net — but this can take a very long time and a lot of shuffling of data (that is, thinking). This all is, of course, much easier, if there is some structure readily available to put the pieces of information in. Not to just get lost in the limitless networld, the key problem is to be capable of characterizing the intuitions and intentions in an appropriate order to make the *mental monads* start when they are needed. The role of presentations like this text is to deliver an organized list of structured "sparks of thought" to constitute a platform to build on, to set on the fire in the mind. Hopefully this new node in the network of distributed cognition becomes "alive" in your node of localized personal brains!

However, there cannot be found very much material along the presented lines of thought in the net, as this research is rather original. — The general cybernetic theory of complex systems (or *neocybernetics*) is available in internet at the pages of `http://neocybernetics.com`.

Acknowledgement

I am grateful to Mr. Petri Lievonen for his insightful comments, and for his never-ending encouragement.

Appendix: additional emulations

Here, some additional simulations of structure emergence applying the cybernetic approach are presented. The experiments were carried out by Mr. Petri Lievonen.

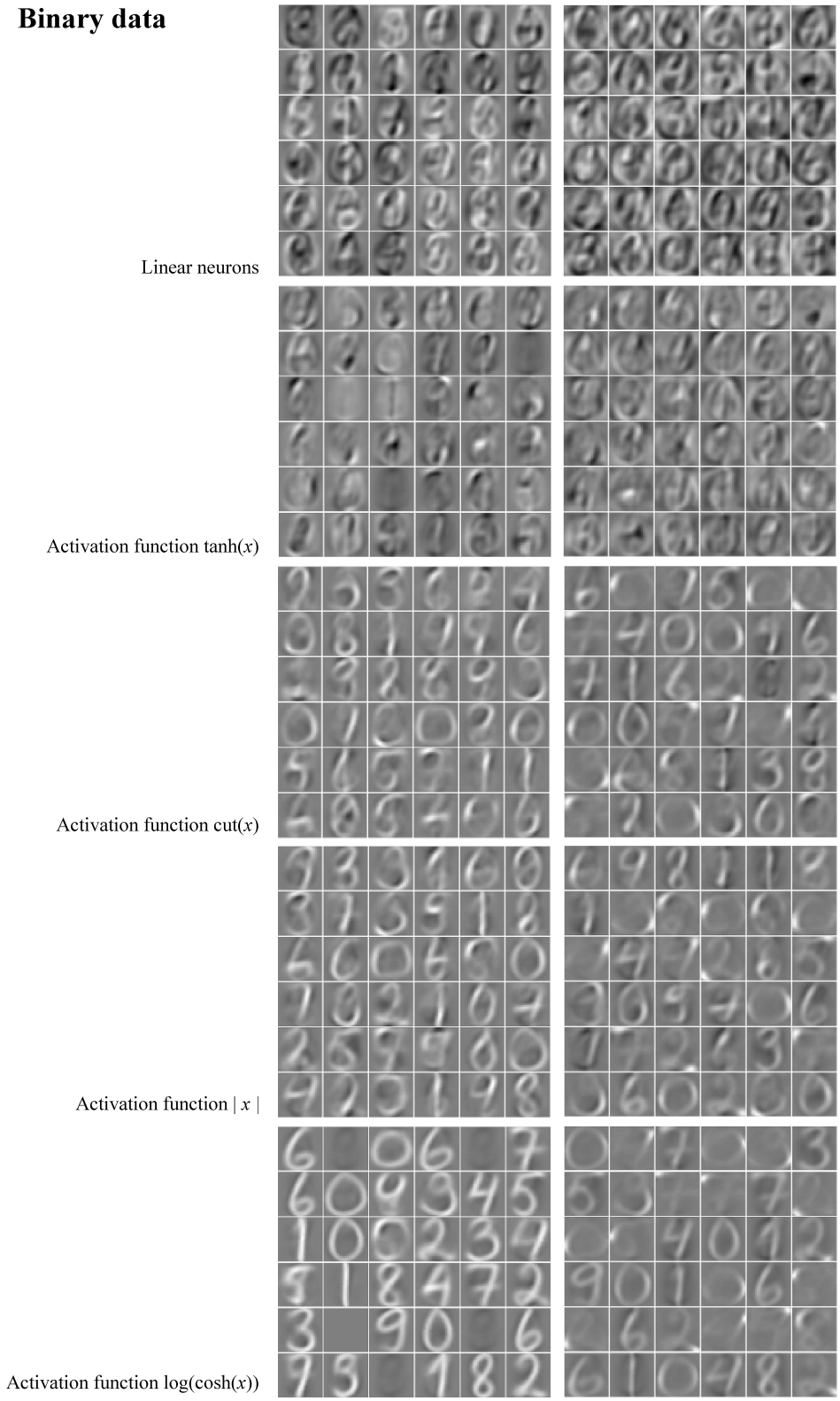
1. Examples of models for the digit data

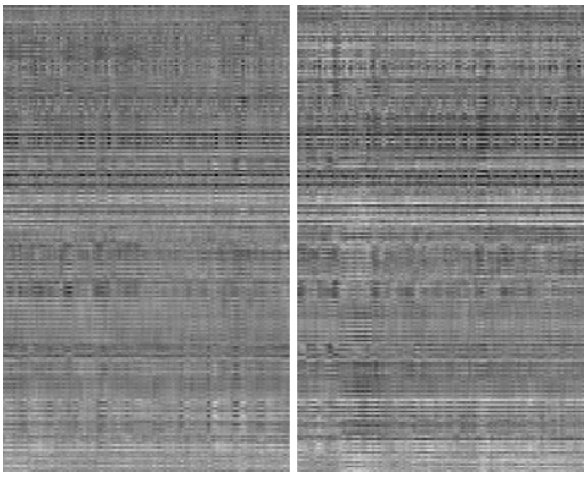
On page 46, additional modeling processes were applied to the data presented in 6.2. Again, only input data was employed; this time no SOM-like self-organization of neurons is applied. These experiments illustrate the robustness of the cybernetic adaptation: not even data normalization is necessary; and it seems that there is much freedom when selecting the nonlinearity (the activation function can even be nonmonotonous).

2. Frequency-domain model of MEG data

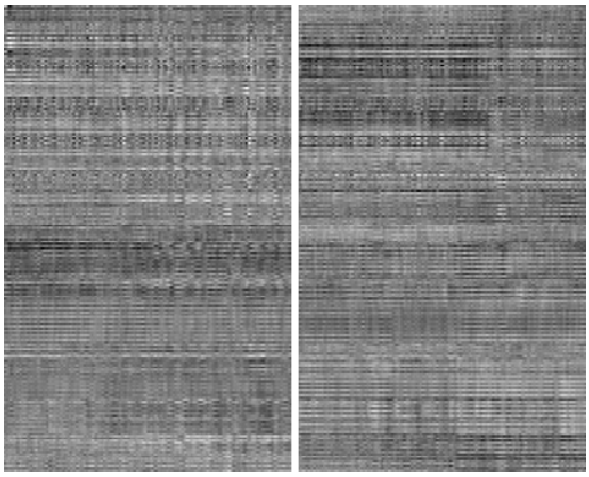
Pages from 47 onwards demonstrate the cybernetic frequency-level model hypothesis. As data, there are spatially distributed spectra, and the goal is to reconstruct the underlying structure. This task is rather ambitious, as what is used is the *ICANN 2011 benchmark data*, real measurements of brain functioning, as explained in `http://www.cis.hut.fi/icann11/mindreading.php`:

> *The challenge combines two recent trends in neuroscience: Analysis of naturalistic stimulation and mind reading. The task in the challenge is to decode the stimulus identity based on magnetoencephalography (MEG) recording done during naturalistic stimulation. In more detail, the subject is viewing video stimuli of different kinds (football match, feature film, recording of natural scenery etc), and the goal is to classify unlabeled test examples into these categories based on the MEG signal alone. . . .*
>
> *The data consists of MEG recordings of a single subject, made during two separate measurement sessions (consecutive days). In each session the subject was watching visual stimuli consisting of five different movie categories. The stimuli were presented without audio.*
>
> *The data contains measurements of 204 planar gradiometer channels at 200Hz rate, segmented into samples of one second length. The samples are given in random order, to enforce prediction based on the 1s window alone. The data is provided after standard preprocessing (removal of external inference, motion correction, low-pass filtering), and the raw signal measurements are complemented with outputs of five bandpass filters. . . .*
>
> *Task: Design and implement a classifier that takes as an input the MEG signals of the test samples (one second of time) and produces as an output the predicted class label (the type of the video stimulus). . . .*

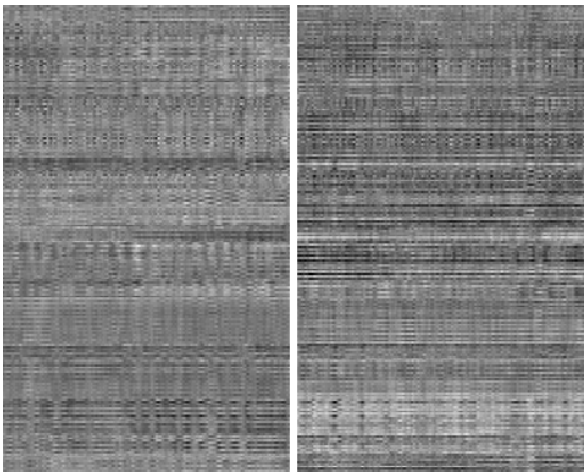

Binary data
Centered and normalized data
Linear neurons
Activation function tanh(x)
Activation function cut(x)
Activation function | x |
Activation function log(cosh(x))



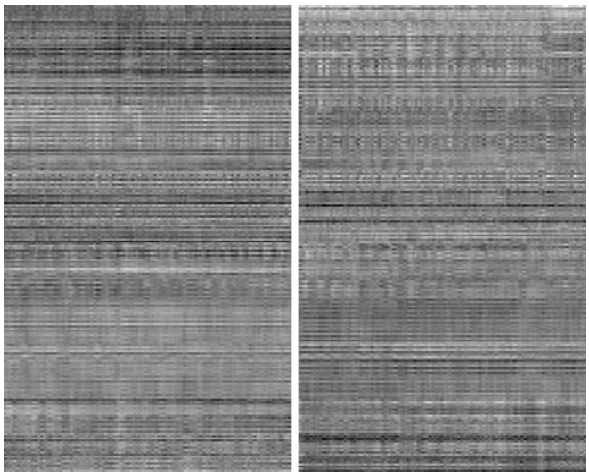
(a) Two seconds of MEG recording of a subject watching animated shapes or text.



(b) Two seconds of MEG recording of a subject watching sequences from a football match.

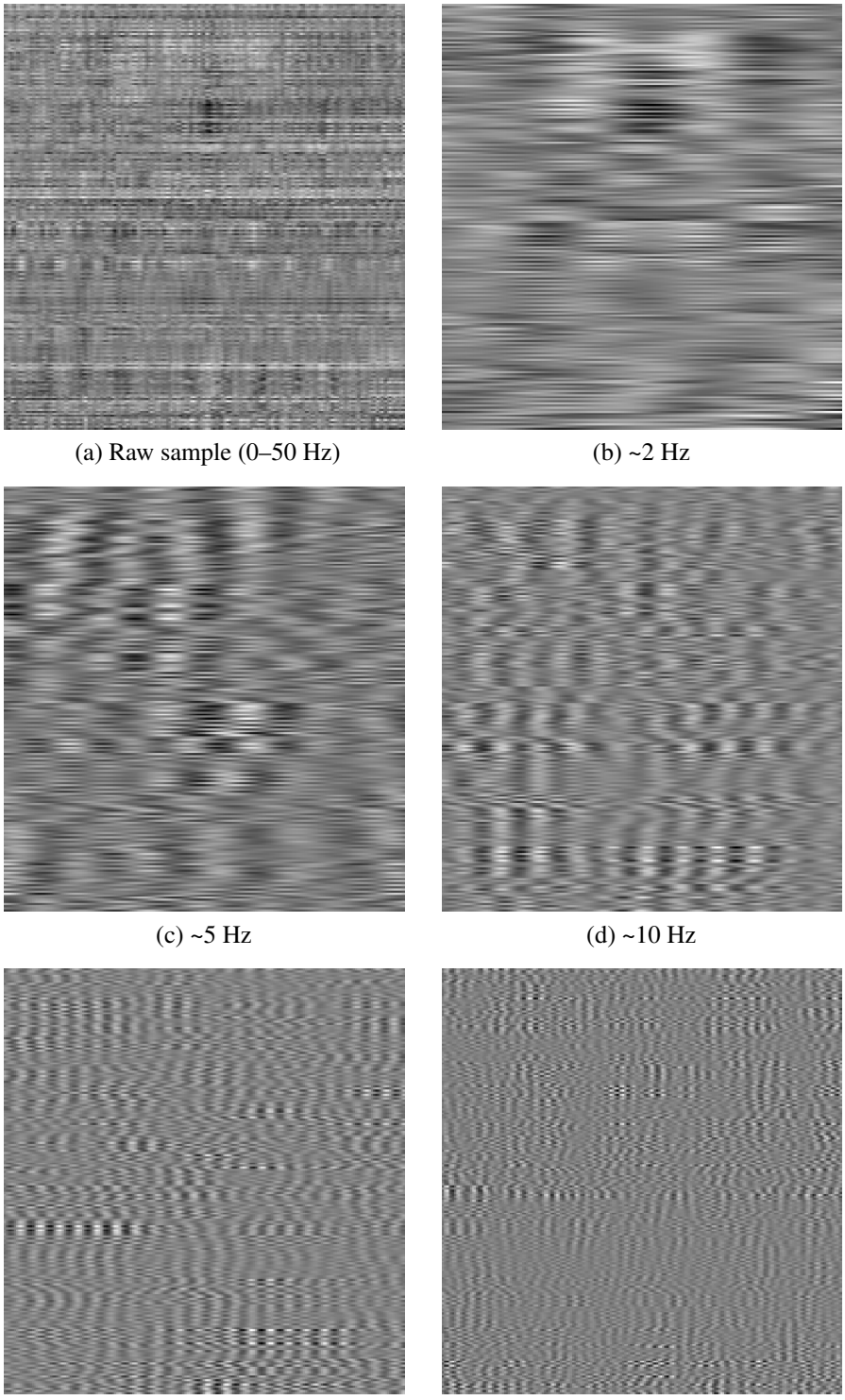


(c) Two seconds of MEG recording of a subject watching a nature documentary.



(d) Two seconds of MEG recording of a subject watching a Charlie Chaplin feature film.

Figure A.2: A few samples of the MEG recordings. The 204 MEG channels around the head are in the rows, and the signals flow from left to right for a total duration of one second in each sample. Synchronized oscillations are clearly visible.



(a) Raw sample (0–50 Hz)

(b) ~2 Hz

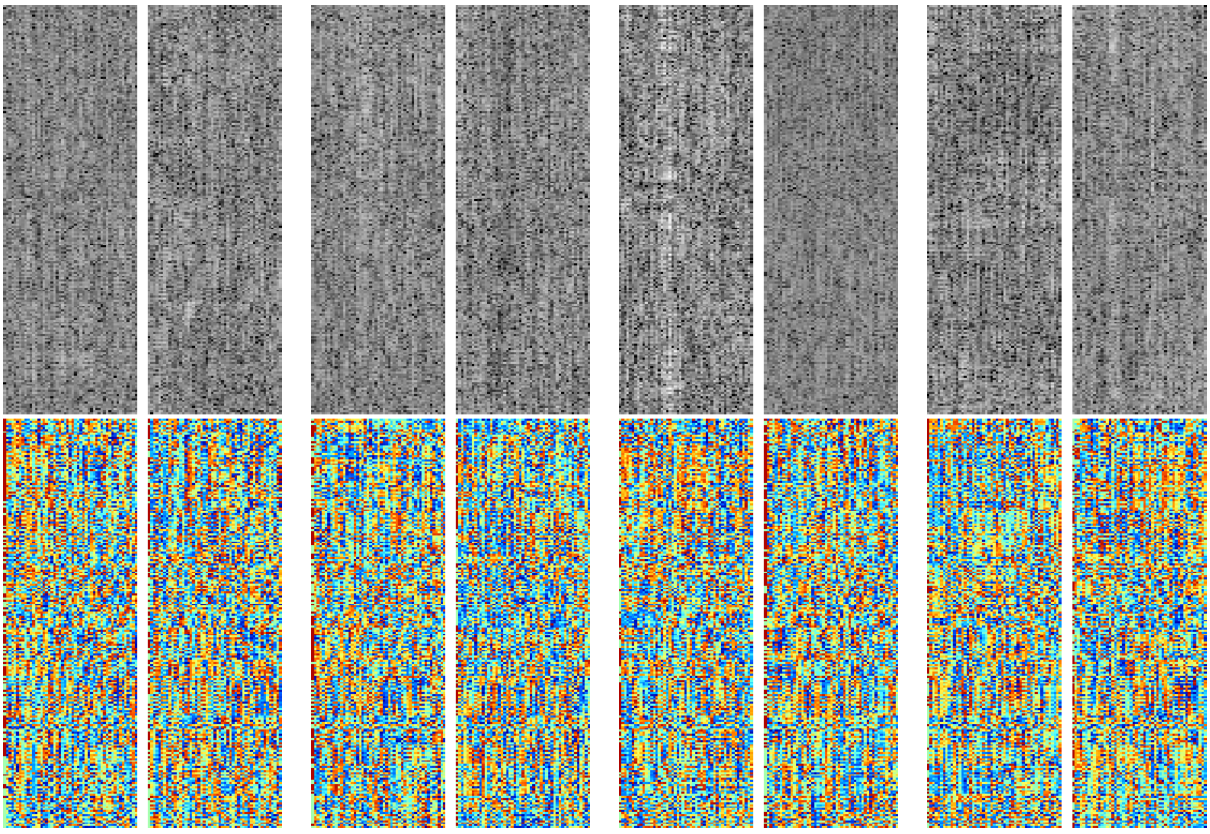
(c) ~5 Hz

(d) ~10 Hz

(e) ~20 Hz

(f) ~35 Hz

Figure A.3: Different oscillatory activity manifest on different scales. The dimensions are the same as in Figure A.2.



(a) Two seconds of MEG recording of a subject watching animated shapes or text.

(b) Two seconds of MEG recording of a subject watching sequences from a football match.

(c) Two seconds of MEG recording of a subject watching a nature documentary.

(d) Two seconds of MEG recording of a subject watching a Charlie Chaplin feature film.

Figure A.4: The same samples as in Figure A.2, but now Fourier-transformed into the frequency domain. The 204 MEG channels around the head are again in the rows, but now the frequencies 0–45 Hz are depicted from left to right in each sample. The upper part is the logarithm of the magnitude, and the lower part is the phase. Different frequency bands manifest in data as areas of larger magnitude, but there are also other interesting areas with coherent phases.

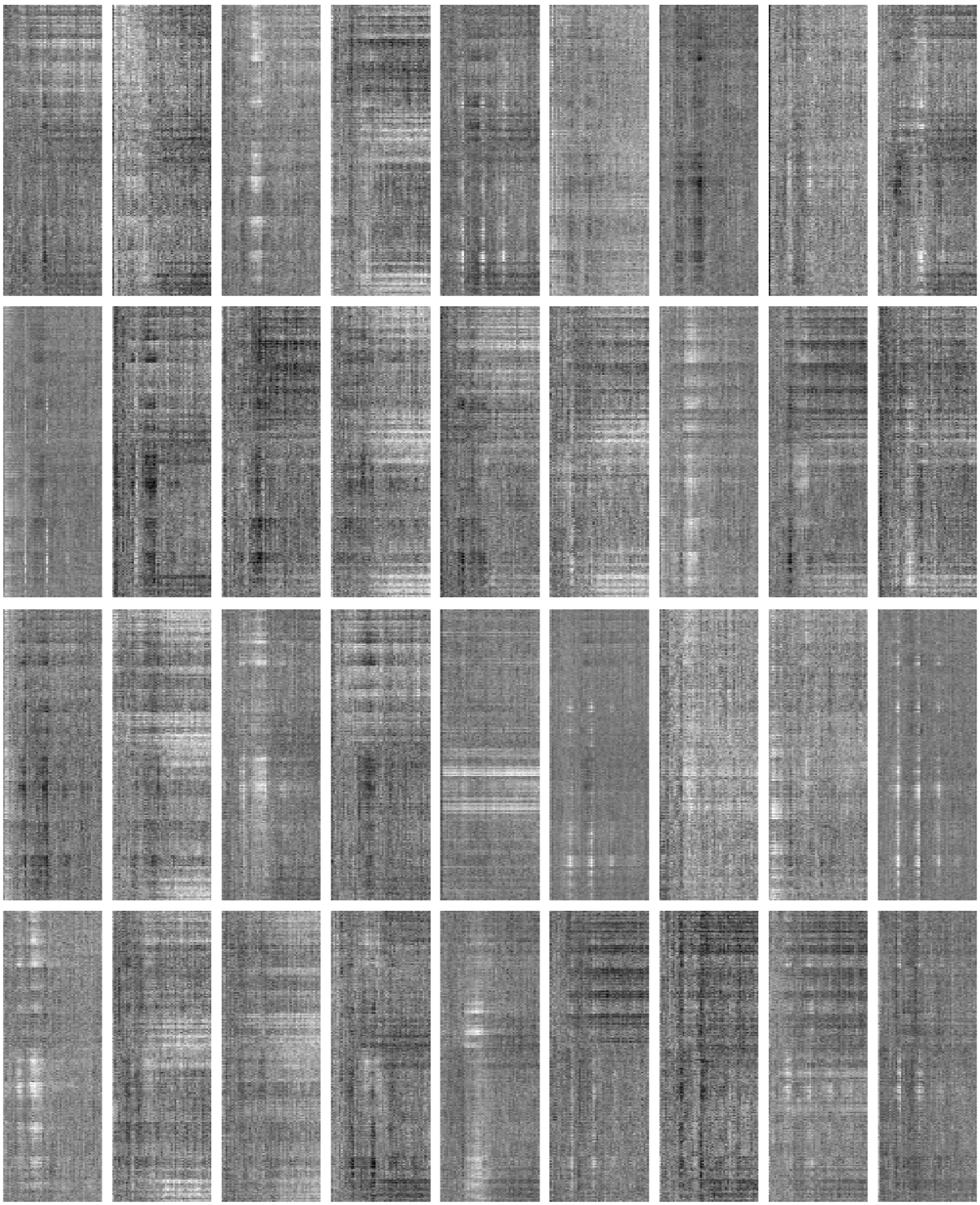


Figure A.5: Profiles of 36 non-negative emergent components of MEG data (677 samples as in Figure A.4) in random order. The 204 MEG channels around the head are in the rows, and the logarithm of frequency magnitudes 0–45 Hz are depicted from left to right in each profile. The profiles span most of the energy in data.

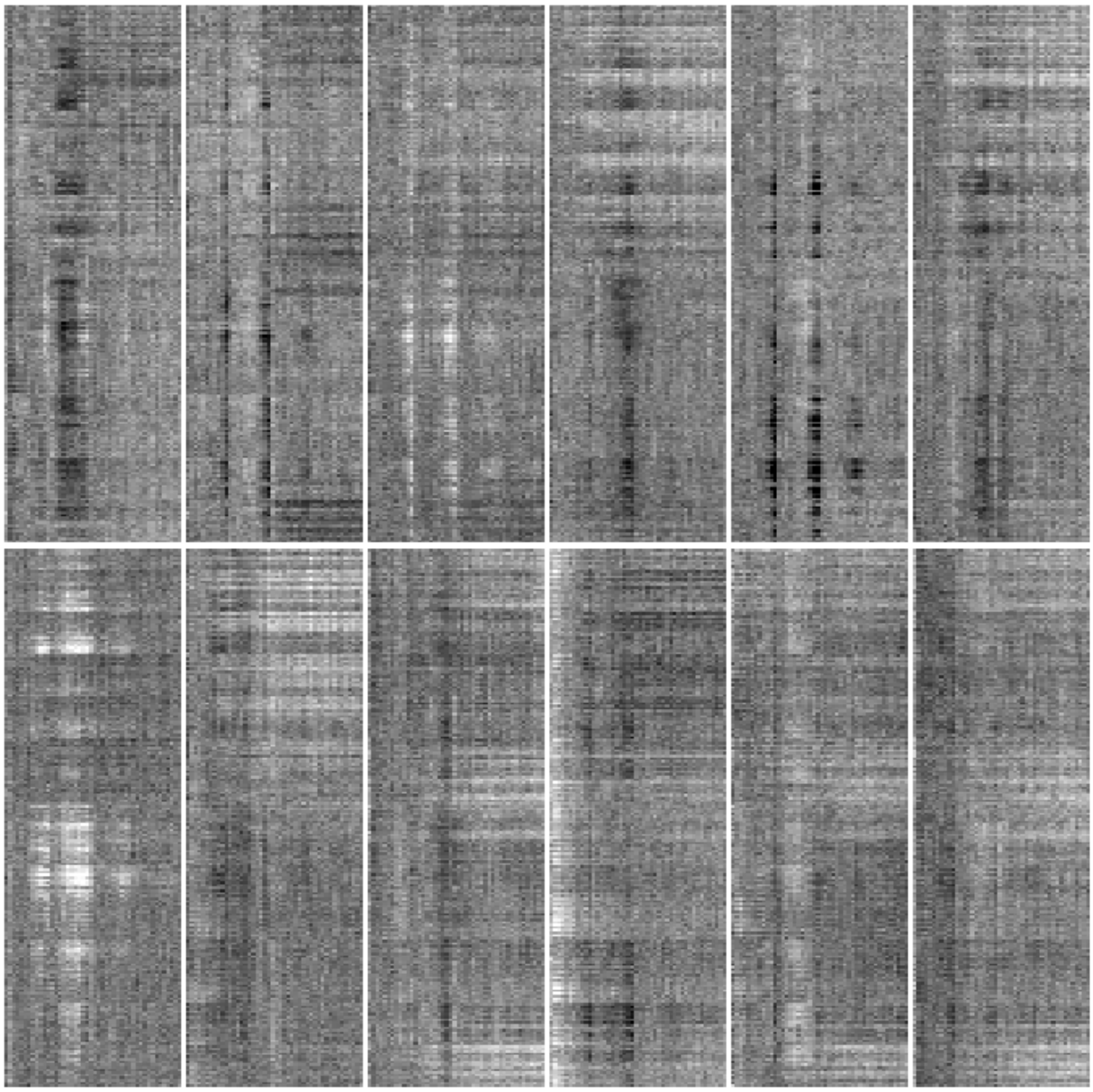


Figure A.6: Profiles of 12 emergent components of MEG data, now augmented with class information in extra dimensions. This external information rotates and mixes the profiles to facilitate classification. With these linear components one can predict with 60 % accurary which of the five classes the subject is watching.

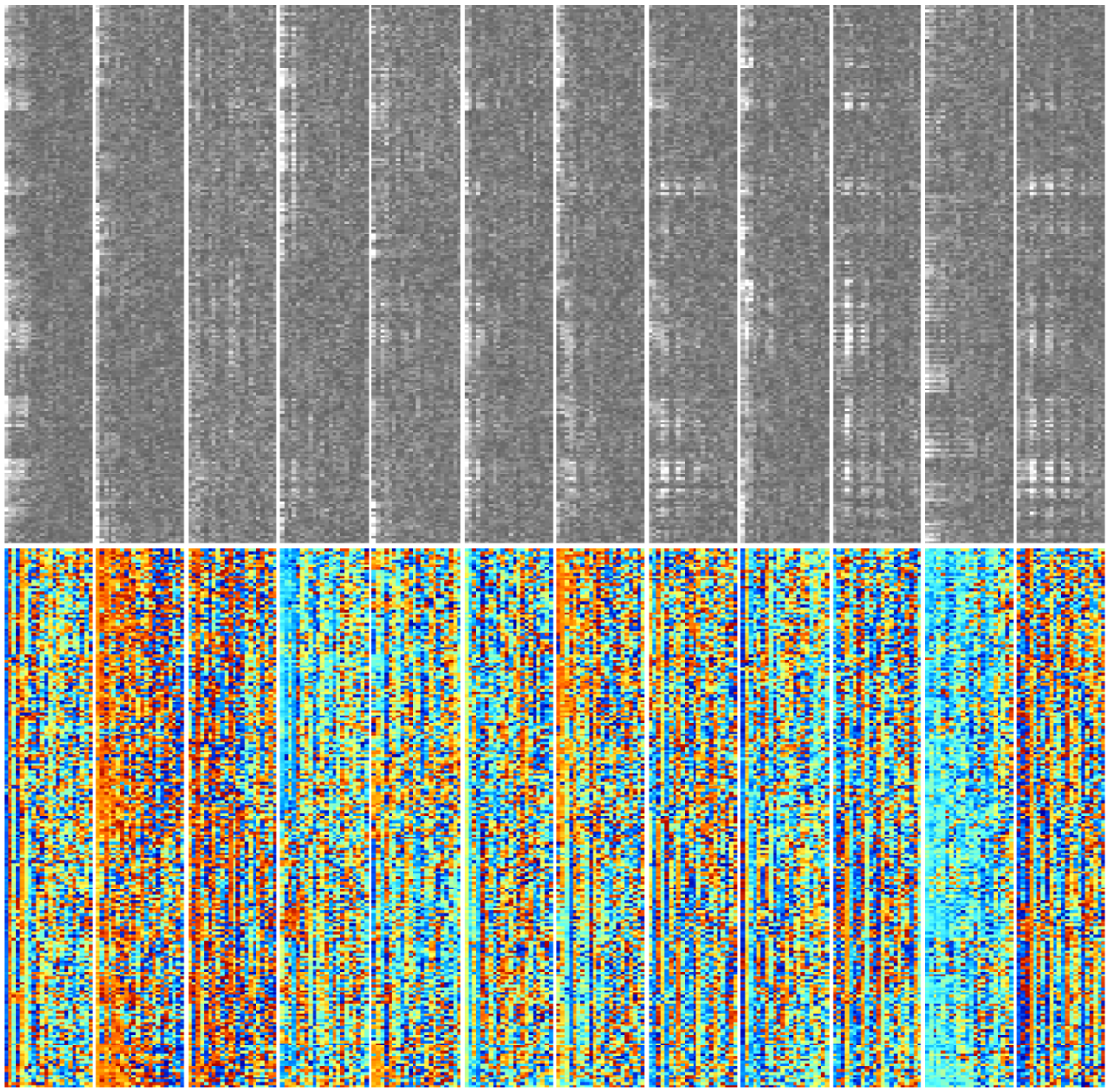


Figure A.7: ECA can also be applied on complex values. Here are 12 profiles of absolute-value components of cepstrum (spectrum of the logarithm of the power spectrum). The dimensions are the same as in Figure A.4.