Enformation Theory — Part II:
—
From *Elements* to *Structures*[⋆]

Heikki Hyötyniemi

Aalto University School of Electrical Engineering
Department of Automation and Systems Technology
P.O. Box 15500, 00076 Aalto, Finland

Abstract. The local *enformation maximization principle* of Part I is studied closer and global-level results are found: it turns out that the system implements *principal subspace analysis* of the experienced data, and it can be extended towards *sparsity pursuit*. The resulting model of the environmental enformation is based on linear *features* that together span the observed patterns. As a fundamental guideline in the derivations one has the fact: *There exist no pure information flows.*

1 Introduction

In Part I, it was observed that for an m dimensional vector $\bar{u}$ of input data and an n dimensional vector $\bar{x}$ of system activations, where typically $m \gg n$, and for some diagonal *coupling matrix* Q, one can write the enformation-maximizing mapping applying the “emergence operator” $\mathcal{E}$ as

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

However, one knows that this kind of Hebbian-style learning structure is unstable, because the adaptation law is based on an internal positive feedback. To keep the system stable, there are different kinds of ways to attenuate the signals, like adding nonlinearities in the structure (for example, applying *Oja's rule,* the weight vectors are explicitly normalized). Here, on the other hand, *negative feedback* is employed; the additional explicit loop structure becomes seemingly complex, but, however, it is structurally simple — the overall mapping remains *linear.* The motivation for the negative feedback is that *if a signal is exploited, it simultaneously becomes exhausted.*

Previously, the focus was on an *individual*; now, the emphasis will be extended to the *system*. Here, a brief excursion through the resulting *neocybernetic system structure* is shown.

[⋆] Submitted to *The Eleventh International Symposium on Intelligent Data Analysis* (IDA 2012), 25–27 October, 2012, in Helsinki, Finland; Part 2 of 5

2 Interplay among levels

One can find many expressions governing the signal covariances (or the system enformations). When multiplying (1) from the right by $\bar{x}^{\mathrm{T}}$ and applying the emergence operator, one has the following expression:

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\} = Q\,\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}. \tag{2}$$

The transpose of this gives yet another expression (remember that $Q^{\mathrm{T}} = Q$):

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}} Q. \tag{3}$$

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

$$Q\,\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} Q, \tag{4}$$

so that also

$$f\left(Q\right)\, g\left(\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right) = g\left(\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)\, f\left(Q\right), \tag{5}$$

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.

Further, assuming invertibility of $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$, and noting (5), from (2) or (3) one has

$$I_n = Q^{1/2}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}Q^{1/2}. \tag{6}$$

When defining

$$\theta^{\mathrm{T}} = Q^{1/2}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}, \tag{7}$$

one has

$$I_n = \theta^{\mathrm{T}}\,\theta. \tag{8}$$

The columns in this new matrix θ are thus orthonormal. Further, by multiplying (1) from the right this time by $\bar{u}^{\mathrm{T}}$ and applying the emergence operator, one has

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

Substituting this in (3), there holds

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

Noting (5), this can be changed to read

$$Q^{-1} = Q^{1/2}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}Q^{1/2}, \tag{11}$$

so that one gets

$$Q^{-1} = \theta^{\mathrm{T}}\,\mathcal{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}\theta. \tag{12}$$

This means that if ever the basic assumption (1) is fulfilled, the statistical properties of the *effective*, observed input $\bar{u}$ are fixed to the selected Q. This modification of the environment can be seen as a manifestation of a more general *observer effect.* As will be shown later, the coupling parameters q_i in Q can be seen as determining the "stiffnesses" of the coupled system elements. Even though it is the environment that is "in charge", the system makes it become somehow organized.

The expression (12), together with (8), reveals that the columns in θ span a space determined by *eigenvectors* of the data covariance matrix $\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$. It can further be shown that after convergence the employed eigenvectors correspond to the most significant eigenvalues (see Sec. 4); thus, the vectors span the subspace capturing the maximum of data covariation (enformation). The original *local behaviors* have become *global*.

What comes to the actual axis vectors θ_i, there are the following two essentially opposite possibilities of interest:

1. If all q_i in Q are distinct, according to (12), the original data eigenvalues λ_j change to $\bar{\lambda}_j = 1/q_i$ (assuming that the unit i has become coupled to mode j); further, for (4) to hold, $\mathcal{E}\{\bar{x}\bar{x}^{\mathrm{T}}\}$ must become diagonal, and from (7) it is evident that there is no shuffling of basis vectors — the system implements *principal component analysis.*
2. If all q_i in Q are equal, on the other hand, so that $Q = q\,I_n$, all eigenvalues are *equalized,* all $\bar{\lambda}_j$ of the closed loop system 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 (4), so that the system implements only *principal subspace analysis* with rotatable basis vectors.

The latter case is the more interesting one, 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, and "higher order" properties in data have become visible.

3 Feedback through the environment

The above analyses apply if such a mapping matrix really exists as proposed in (1). 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"? In the cybernetic spirit, of course, dynamics and feedback is here proposed.

When the system element 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 connection strength $q_i\mathcal{E}\{\bar{x}_i\bar{u}_j\}$ (possible scaling needs are included in x; see details in [1]). This means that the change in the whole set of inputs can be written in matrix form as

$$\Delta\bar{u} = \mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}} Q\,\bar{x}. \tag{13}$$

If such feedback is not enough to implement stabilization of the loop, the adaptation increases the signals until the "balance of tensions" is reached for some $\bar{x}_i' = c_i \bar{x}_i$, or $\bar{x}' = C\bar{x}$ for some diagonal C. Because

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

one can freely scale the state variables; in what follows, it is assumed that the signals $\bar{x}$ and $\bar{u}$ have already been scaled to match each other.

This far, symbols like $\bar{u}$ and $\bar{x}$ have been used all the time; they are the final effective 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*). For the disturbed input, or *residual,* there holds

$$\tilde{u}(t) = u - \Delta u(t), \tag{14}$$

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

$$\bar{u} = \lim_{t\to\infty} \left\{\tilde{u}(t)\right\} \tag{15}$$

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

$$\bar{x} = \lim_{t\to\infty} \left\{x(t)\right\} = \lim_{t\to\infty} \left\{Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\tilde{u}(t)\right\}. \tag{16}$$

In the asymptotic case, when the balance has been found, the situation looks like that shown in (1). 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: 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 $\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}$.

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

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

or, when solved,

$$\bar{x} = \left(I_n + Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}^{\mathrm{T}}Q\right)^{-1} Q\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u. \tag{18}$$

Using (3) and simplifying, one has an expression for $\bar{x}$ directly in terms of u:

$$\bar{x} = \left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1}\mathcal{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u. \tag{19}$$

The variables $\bar{x}_i$ present the functionally different approaches to surviving in the environment, that is, they represent some kinds of *ecological lockers* or *niches*. Typically in nature, these lockers are inhabited by *populations* of individuals, so that the numerical value of the variable reveals the (scaled) *abundance*.

4 Maximum of enformation captured

In the formula (19) there is a discrepancy: the input is u but the covariances are given in terms of $\bar{u}$. The other of the variables can be eliminated by manipulating the expression:

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

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

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

Combining this with (19)

$$\bar{x} = \underbrace{\left(Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-2} Q^{-1}}_{M_1} \underbrace{\mathcal{E}\left\{\bar{x}u^{\mathrm{T}}\right\}}_{M_2} u. \qquad (20)$$

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 Section 5), 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 (20), proceeds as

$$\begin{aligned}
\bar{x}^{(0)} &= M_1^{(0)} M_2^{(0)}\, u \\
\\
\bar{x}^{(1)} &= M_1^{(1)}\mathcal{E}\left\{\bar{x}^{(0)}u^{\mathrm{T}}\right\} u = M_1^{(1)}\mathcal{E}\left\{M_1^{(0)}M_2^{(0)}uu^{\mathrm{T}}\right\} u \\
&= M_1^{(1)}M_1^{(0)}M_2^{(0)}\mathcal{E}\left\{uu^{\mathrm{T}}\right\} u \\
\bar{x}^{(2)} &= M_1^{(2)}\mathcal{E}\left\{\bar{x}^{(1)}u^{\mathrm{T}}\right\} u = M_1^{(2)}\mathcal{E}\left\{M_1^{(1)}M_1^{(0)}M_2^{(0)}\mathcal{E}\left\{uu^{\mathrm{T}}\right\}uu^{\mathrm{T}}\right\} u \\
&= M_1^{(2)}M_1^{(1)}M_1^{(0)}M_2^{(0)}\mathcal{E}\left\{uu^{\mathrm{T}}\right\}^2 u \\
&\vdots \\
\bar{x}^{(k)} &= M_1^{(k)}M_2^{(k)}\, u = \left(\prod_{i=0}^{k} M_1^{(k-i)}\right) M_2^{(0)}\mathcal{E}\left\{uu^{\mathrm{T}}\right\}^k u.
\end{aligned} \qquad (21)$$

Assume that the eigenvalue decomposition of the data covariance is written as

$$\mathcal{E}\left\{uu^{\mathrm{T}}\right\} = \Theta\, \Lambda\, \Theta^{\mathrm{T}}. \tag{22}$$

The final mapping matrix in (21) becomes

$$M_1^{(k)} M_2^{(k)} = M_1^{(k)} M_2^{(0)} \mathcal{E}\left\{uu^{\mathrm{T}}\right\}^k = \underbrace{M_1^{(k)} M_2^{(0)} \Theta}_{n\times n}\ \underbrace{\Lambda^k \Theta^{\mathrm{T}}}_{n\times m}. \tag{23}$$

The former part is a scaling matrix of dimension $n \times n$ and it does not affect the subspace being spanned by the mapping; from the latter part one can see 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). □

5 Emergence of new structures

Despite the analyses above, there are *two* classes of solutions to (1). In addition to the case that was discussed in previous sections, 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 entity i, whether it fades away or stays "alive", depends on the corresponding *coupling* to the environment.

From (19) one can write yet another expression for the covariance by multiplying the expression by its transpose, and applying the emergence operator, giving

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

that can be written

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

Observing the commutativity of the matrices:

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

Further, because of the orthogonality of θ,

$$Q^{-1} + \mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q^{-1/4}\, \theta^{\mathrm{T}}\, \mathcal{E}\left\{uu^{\mathrm{T}}\right\}^{1/2} \theta\, Q^{-1/4}, \tag{24}$$

or

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q^{-1/4}\,\theta^{\mathrm{T}}\,\mathcal{E}\left\{uu^{\mathrm{T}}\right\}^{1/2}\theta\,Q^{-1/4} - Q^{-1}. \tag{25}$$

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

$$\mathcal{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q^{-1/4}\,P^{\mathrm{T}}\Lambda_{[n]}^{1/2}P\;Q^{-1/4} - Q^{-1}, \tag{26}$$

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

$$\mathcal{E}\left\{\bar{x}_i^2\right\} = \sqrt{\frac{\lambda_j}{q_i}} - \frac{1}{q_i}. \tag{27}$$

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 (26) becomes fully compensated. To keep the entity functional, there must hold

$$q_i > \frac{1}{\lambda_j}. \tag{28}$$

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 (assuming the the input data is rich enough).

Where is this activation lost, where does the "static friction" come from, introducing nonlinearity in the linear structures as seen from above? This threshold 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.

6 Getting rid of free parameters

To minimize the number of unknown parameters in a large system, it has to be assumed that there is some local mechanism for adjusting the coupling factors q_i. A practical formulation is found when the determination of $\bar{x}$ for given u is seen as a stochastic estimation task; intuition can then be gained from *recursive least-squares identification*. It is reasonable to scale down the variable by the *inverse of its variance*, so that one can select

$$q_i = b\,\frac{1}{\mathcal{E}\left\{\bar{x}_i^2\right\}}, \tag{29}$$

with $b > 0$ being some scaling factor. There are various technical motivations for selecting q_i in such a way: first, variables 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. It also assures that the variable remains "alive" (or "cybernetic"), increasing the coupling if the activation seems to vanish; indeed, the selection (29) assures *maximum system activation.* — What is more, also natural neurons turn out to implement similar activity-based controls.

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 (12), 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.

As all variances $\mathcal{E}\{\bar{x}_i^2\}$ become equal with the selection (29), one can easily apply the matrix trace to (25), and one has for all i and j a formula for the variances:

$$\mathcal{E}\left\{\bar{x}_i^2\right\} = \frac{b}{q_i} = b\,\bar{\lambda}_j = \left(\frac{\sum_{\iota=1}^{n}\sqrt{\lambda_\iota}}{n\left(\sqrt{b}+\frac{1}{\sqrt{b}}\right)}\right)^2 . \tag{30}$$

When one selects $b = 1$, there is an intuitively appealing balance between the internal and external variances, so that the value of $\mathcal{E}\{\bar{x}_i^2\} = \bar{\lambda}_j$ 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.

One can even propose a system size optimization scheme based on the formula (30): 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 (for large n, and for $b = 1$) hold

$$\sqrt{\lambda_n} > \frac{1}{2}\,\frac{\sum_{\iota=1}^{n-1}\sqrt{\lambda_\iota}}{n-1}, \tag{31}$$

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. Because of the regularization term Q^{-1} in the formulas, 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" model structures.

Examples of the neocybernetic algorithm being applied to real data are available in [1]. The operation of the algorithm seems "very interesting and promising". It is easy to propose extensions to the basic algorithm, too, without jeopardizing the basic stabilizing nature of the approach.

Assume that the data consists of sums of positively weighted (orthogonal) *sparse components*, and one would like to extract these physically motivated basis vectors out from the principal subspace. By adding an explicit nonlinearity in the internal loop, so that all negative values of x_i are cut off, it turns out that the model implements sparsity pursuit, because then the maximum overall variance becomes captured. After convergence, the overall behavior of the system can still be linear. — One can also explicitly only select the most significant of the variables (in the spirit of "magical number 7 ± 2").

In practice, a more relevant case is faced when the structure is extended into an *input-output system*. It turns out that if there are two sets of signals, u and y, *canonical correlation regression* between these spaces can be approximated when the two individual signal models are combined by calculating the averages of the corresponding x vectors; when the average is recirculated and used for learning in both subsystems, after converges it can be seen as the *latent variable* representing the lower-dimensional intermediate space for implementing the mapping from the input to the output.

References

1. *Neocybernetics — Pragmatic Semiosis by Complex Adaptive Systems.* Research pages accessible in Internet through http://neocybernetics.com.