REPRESENTATION THEORY MEETS ANATOMY: FACTOR LEARNING IN THE HIPPOCAMPAL FORMATION[*]

ANDRÁS LŐRINCZ

Department of Information Systems, Eötvös Loránd University, Pázmány sétány 1/C Budapest 1117 Hungary

GÁBOR SZIRTES

Department of Cognitive Psychology, Eötvös Loránd University, Izabella u. 46 Budapest 1064 Hungary

In this paper we argue that computational issues like complexity, memory requirements and training time impose strong constraints on learning in any goal-oriented system. Along these constraints we derive a particular architecture that learns representations for optimizing plans e.g., trajectory planning. To comply with biological constraints as well, the resulting encoding mechanism is translated into a connectionist network. We argue that the goal-oriented framework implies distinct representations of place and direction in the hippocampal formation responsible for spatial navigation in mammals.

1. Introduction

Ample clinical evidence demonstrates the central role of the hippocampal region (HR) in forming long-term memories see, e.g. [1]. The complex structure of this brain region has inspired a plethora of models aiming to correlate its functioning with the peculiarities of the structure. One common view [2] in many models is that memory systems maintain representations (that is models) shaped to help executive control, because associative mapping of immediate stimuli to responses is insufficient to reflect the reward structure of the environment. Granted this assumption, computational constraints regarding speed and complexity on forming representations for control and desired goals should be applied on any model of memory. In this paper we focus on how these constraints may restrict potential models. In a general goal-oriented system

[*] This research has been supported by the EC NEST PERCEPT' Grant FP6-043261 and Air Force Grant FA8655-07-1-3077. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the European Commission, European Office of Aerospace Research and Development, Air Force Office of Scientific Research, Air Force Research Laboratory. G. Sz. has been supported by the Zoltán Magyary Fellowship. We are grateful for the anonym reviewer for his/her useful remarks.

learning has two stages; (i) the adaptation of the internal model, i.e., shaping of the *representations* and (ii) the optimization of sequential decision making, i.e., the state→desired state plans, to maximize the expectation value of discounted and cumulated long-term rewards. Sequential decision making may be treated within the framework of reinforcement learning (RL) and Markov decision process (MDP) can be used as its mathematical model motivated by psychology and neuroscience as well [3]. While the original MDP scales badly with the number of variables, its factored version (fMDP) [4] can avoid such complexity issues by assuming that dependencies among variables can be factored – under certain circumstances – to several components (see e.g. [5]). However, these factors should first be appropriately extracted and represented.

In what follows, we review the central notion of factors and how it implies the concept of independent process analysis (Section 2). The theoretical results are translated into a neural network architecture featuring Hebbian learning (Section 3). In Section 4 we show the close similarity between the derived architecture and the hippocampal region (HR). We shortly discuss how our model has evolved since the proposal of [17] in Section 5.

2. Factors and Independent Processes

We take navigation as an example to illustrate our ideas. The state of the animal is characterized by two independent factors, position and speed, according to Newton's second law. Acceleration depends on the net force g and the mass (m). Newton's equation $md^2s/dt^2 = g$ –where d^2s/dt^2 denotes the second temporal derivative of position s– can be discretized and for short time steps we get

$$s(t+1)=g(t)/m - 2s(t)+ s(t-1) \qquad (1)$$

This equation may be interpreted as an autoregressive process *driven by* $g(t)/m$. What makes this problem hard is that this simple form is typically hidden as observations are filtered through, for example, visual, olfactory or proprioceptive sensors thus providing a distorted function of the original process. Distortions may be caused by delays or by including differently transformed echoes (called moving averages, MA) of the same driving forces. The recovery of the hidden factors then includes the following steps: (i) translation of the problem into hidden autoregressive moving average (ARMA) processes, (ii) removal of both the AR and MA contributions, (iii) extraction of the hidden driving noises (or causes, e.g., the forces, or their immediate

consequences; the changes of the trajectory) and (iv) identification of their relation. Recent advances on hidden ARMA models enable the recovery of the hidden processes as well as the hidden causes from the observations up to certain forms of invariance provided that multi-dimensional components of the hidden driving noise can be assumed *independent* (see [6] and the cited references). Loosely speaking, in contrast to simple decorrelation, independence implies that second *and* higher order correlations between the multi-dimensional components can be neglected.

We restrict our considerations to first order AR processes that assume the following form

$$h(t+1) = Gh(t) + f(t+1) \quad (2)$$
$$x(t) = Ah(t). \quad (3)$$

where matrix $G \in R^{k \times k}$ represents the dynamics of the system, vector $h(t) \in R^k$ is the hidden state or hidden process at time t, $f(t) \in R^k$ is the hidden driving noise process, $A \in R^{k \times k}$ is the mixing matrix that prohibits direct observation, and $x(t) \in R^k$ is the observation at time t (from now on, capital letters denote matrices, lower case letters denote vectors). We also assume that matrix A is invertible. The goal is to compute the estimations, $\hat{h}(t)$ for $h(t)$, $e(t)$ for $f(t)$, and the so called separation matrix W for A^{-1}. Within this setup it also makes sense to distinguish *independent processes* (time series) $h_i(t)$ in the sense that they do not mix: F may assume block structure. ARMA process with independence assumption on the noise will be referred to as ARMA-IPA model, where IPA stands for Independent Process Analysis.

3. A connectionist network for identifying ARMA-IPA models

Let us remark first that the observation of Eq. (3) is also an AR process, as substituting (2) in (3) yields:

$$x(t+1) = \mathcal{M}\, x(t) + n(t+1) \quad (4)$$

According to the central limit theorem, $n(t+1)=Af(t+1)$ is approximately Gaussian and thus matrix $\mathcal{M} = AFA^{-1}$ and noise $n(t+1)$ can be estimated by least-mean square methods, i.e. the cost function can be read as:

$$J(M) = \tfrac{1}{2} \Sigma_t \, |x(t+1) - M\, x(t)|^2, \quad (5)$$

where M is the estimation of $\mathcal{M}$ yielding the following estimation of the observation noise: $\varepsilon(t) = x(t+1) - Mx(t)$. Because of the independence assumption on $f(t)$, term $\varepsilon(t)$ can now be analyzed by Independent Subspace Analysis (ISA, [7,8]) techniques to estimate W, which then leads to the estimations of the hidden causes $f(t)$, the hidden state $h(t)$, and hidden dynamics G. (On the peculiarities of the special combination of ISA and ARMA models, see [6, 9].) ISA is a generalization of the Independent Component Analysis (ICA) methods in that it assumes multi-dimensional hidden sources and its solution can be decomposed into two subtasks: 1, on an ICA *separation* step (yielding more or less independent *one-dimensional* components) and then 2, a *clustering* of these components into independent subspaces. ISA, alike other ICA derivatives, is speeded up considerably if data is decorrelated (whitened) before the actual separation. ISA also inherits ambiguities of the ICA methods: components are not ordered, and their sign and scale may also vary. Furthermore, the ISA solution is ambiguous up to orthogonal transformations within the recovered subspaces.

Having identified the most important algorithmic components (namely, identification of an AR process, whitening, ICA separation, and clustering) now we are in the position to derive a connectionist network featuring local (i.e. Hebbian) interactions only. Inter- and intralayer synaptic weights or connections are denoted by matrices, while activity of a given layer is defined by a vector. The derived architecture will be mapped onto the neural substrate in the next section, but fine details like temporal characteristics, spike generation sparse coding and others will be omitted. For simplicity, rate coding (manifesting analogue values) and mixed weights (that contradict Dale's Principle) are assumed throughout the derivations, but the proposed functioning can in principle be also realized by using either positive coding [10] or homogenous connection systems [11]. Inhibitions (or subtractions) are manifested by separate inhibitory populations within a layer using feedback or feed-forward inhibition.

3.1. *Identification of the AR predictive matrix*

The following online and local learning rule for matrix M can be derived from the cost function of Eq. (5) by computing its negative gradient according to matrix M:

$$\Delta M(t+1) \sim \varepsilon(t+1)x(t)^T \tag{6}$$

This rule already has an effect on the construction of the network.

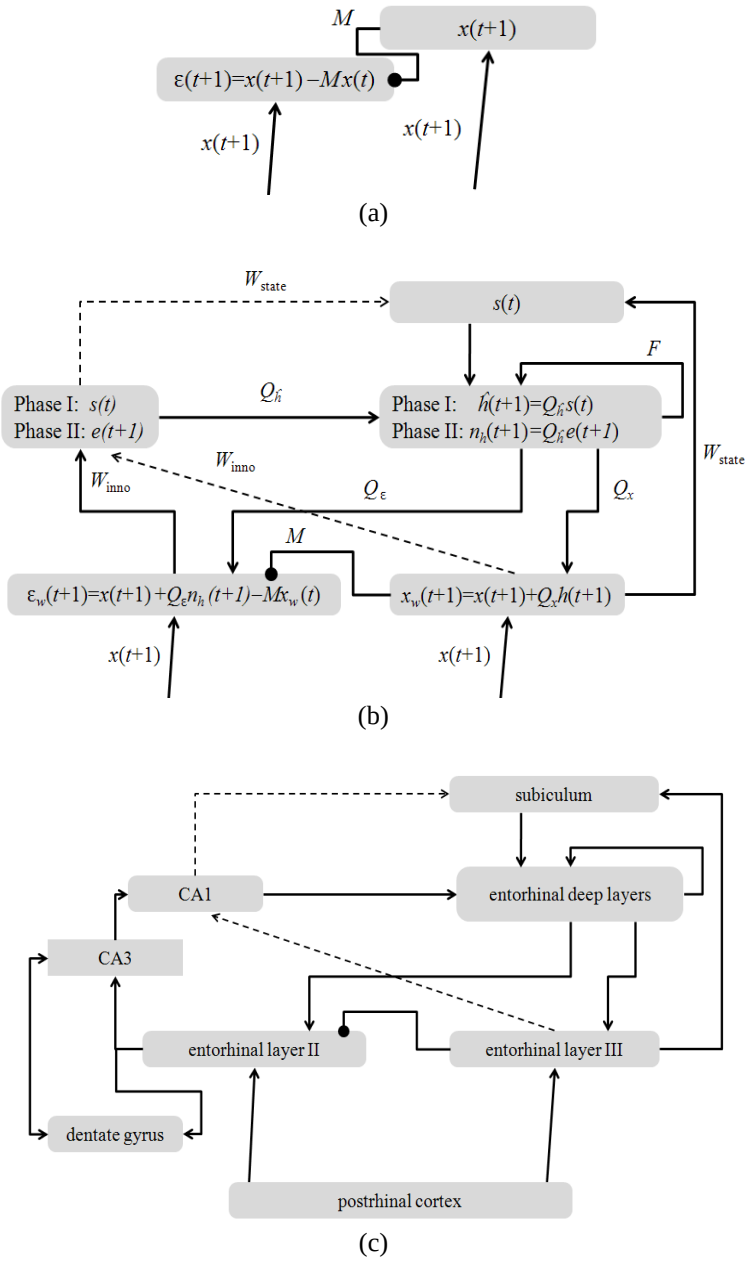

Figure 1. See page 6 for the Caption.

Caption for Figure 1. Hebbian network of the hippocampal region. (a): a network that can extract the noise of the observed input: *t* time index, *x* input, ε estimated noise, *M:* estimeted AR matrix. Open (circle) arrows: excitatory (inhibitory) connections. (b) Computational architecture. Q_x, Q_ε, $Q_{\hat{h}}$: whitening transformations targeting *x*, ε, $\hat{h}$, respectively. W_{state}, W_{inno}: separation matrices extracting state and innovation, respectively, *s* estimated independent components of the state, *e* estimated hidden independent components. ε_w and x_w whitened estimation of innovation and input, respectively. Estimated hidden dynamics: *F*, corresponding estimated state and noise $\hat{h}$ and n_h, respectively. Phase I and II refer to positive and negative phases of the theta oscillation, respetively. Straight lines: mainly proximal projections, dotted lines: mainly distal projections. (c): Most important connections in the hippocampal formation. CA3 subfield and dentate gyrus eliminate moving averages, or echoes, detailed in [16,17,18]. For more details, see text.

First, two distinct layers are required to sustain a term analogue to the estimated noise and the observed signal. The simplest construction depicted on Fig. 1(a) assumes two layers with a delayed feed-forward and mostly inhibitory connection system representing predictive matrix *M*. If the predictive connections are properly tuned, then activity of one layer represents the observed process, whereas activity of the other layer represents the observation noise.

3.2. *Whitening*

As we already noted, decorrelation is a necessary step prior to extracting the wanted independent components. Due to the special form (see below) of the remarkable online learning rule of [12], we need to introduce an additional layer whose activity is denoted by $\hat{h}(t)$ (and will be referred to as the 'internal model'), see Fig 1(b). Its role is quite central in our discussion, but for now it is enough to assume that it is already decorrelated and is projected through the linear connections Q_x and Q_ε onto layers holding $x(t)$ and $\varepsilon(t)$, respectively. Let us denote the modified observed signal by $x_w(t)=Q_x\hat{h}(t)+x(t)$, implying that $x_w(t)$ is still a linear function of $x(t)$. Decorrelation of $x_w(t)$ may be accomplished if tuning of Q_x is as follows:

$$\Delta Q_x(t+1) \sim Q_x(t) - x_w(t)x_w(t)^T Q_x(t) = Q_x(t) - x_w(t)\hat{h}(t)^T \qquad (7)$$

The second form is clearly Hebbian and it follows from the decorrelation assumption on $\hat{h}(t)$ and the fact that $Q_x=Q_x^T$. Whitening of the observation noise may follow similar principles in that the modified noise is $\varepsilon_w(t)= Q_\varepsilon n_h(t)+ x(t+1)$

$-Mx_w(t)$, where $n_h(t)$ is the noise of the internal 'model'. Tuning of Q_ε is also Hebbian:

$$\Delta Q_\varepsilon(t+1) \sim Q_\varepsilon(t) - \varepsilon_w(t) n_h(t)^T \quad (8)$$

In turn, transformation M extracts the deterministic part from ε, whereas Q_ε decorrelates it yielding ε_w.

3.3. *Separation*

Whitening allows for separation of the hidden, independent sources from the observation noise. An additional layer needs to be introduced to represent the emerging independent components. Along the logic presented so far, we are looking for an online learning rule that complies with Hebbian constraints and yields the proper components. Interestingly, as it was shown in [13], a rule very similar to those defined by Eqs. (7-8) can be used to tune the transformation matrix between the layers representing the decorrelated and the independent components:

$$\Delta W_{inno}(t+1) \sim W_{inno}(t) - f(e(t)) e(t)^T W_{inno}(t) = W_{inno}(t) - f(e(t)) \, (W_{inno}{}^T e_w(t))^T \quad (9)$$

Where $f(\cdot)$ denotes an almost arbitrary component-wise nonlinear transformation. Although W_{inno} is not symmetric, the second form of Eq. (9) is again Hebbian if signal backpropagation ($W_{inno}{}^T e(t)$) controls learning [14,15]. Note that whitening may also be promoted by signal backpropagation.

The correctly tuned W_{inno} (as it approximates the inverse of the hidden mixing matrix, A) yields the estimation of the hidden driving sources, $e(t)$. However, as Eqs. (4) and (3) show, the *very same* transformation should be carried on to yield the hidden states $s(t)$ from the decorrelated observation ($x_w(t)$). In turn, an additional layer targeted by the layer of $x_w(t)$ seems to be in need to be introduced to represent $s(t)$ (see Fig. 1(b)). While the mathematics requiring the double application of separation is simple, a truly connectionist or neuronal implementation is not straightforward and impose strong architectural and functioning constraints besides the Hebbian nature of the interactions. Some consequences and putative mechanisms of such 'transformation adjustments' for different pathways will be discussed later.

3.4. *Remark on ARMA processes of higher order*

Although we assumed first order dynamics in the derivations above (Eqs. (2-5)), many real problems are prevailed by higher order dynamics. Sensory stimuli, for

example, become higher order due to echoes (reflections) and delays (multi-modal integration, synaptic delays, etc.). As recent results show, even for hidden *integrated* [6] as well as *non-linear* hidden ARMA processes [16] can be recovered using ideas similar to what has been presented. Canceling the distortions should take place before attempting separation, so the network derived so far should be extended with a separate subsystem that handles delays prior to the actual separation process. As delays are ignored in the present study, we do not discuss the implementation nor the mapping of this subsystem. For more details on the implementation and the mapping, see our supplementary material [17] and [18], respectively.

4. Mapping the network onto the hippocampal region

Due to the space limit we only sketch a possible functional mapping onto the hippocampal region and highlight some relevant supporting physiological and anatomical properties (for terminology, see [19] and for a detailed enumeration of the supporting anatomical findings, see [18]). Despite some differences in the connectivity structure, the common view is that mammalian HR-s share the same gross anatomy as well as functionality, so we refer to findings on rats as well as monkeys. The resulting mapping is depicted on Fig. 1(b)-(c). First, it is known that the superficial layers of the entorhinal cortex (denoted by EC II and EC III) are the main recipients of the cortical inputs. Since the exact nature of these incoming signals is not known we assume they share the same input. While they both have excitatory recurrent connections, an important difference is that EC II has a widespread inhibitory circuitry, too. Excitatory projections from EC III to EC II in part form a unique feed-forward inhibitory system interpreted here as the operation of 'subtraction'. This special connection may allow for the comparison needed to represent the observation noise. The emerging activity in EC II is then assumed to represent the observation noise and is projected onto sub-region CA3 and the dentate gyrus. The latter part of HR is known to have tunable recurrent connections with extreme long time delays thus being able to diminish the distorting effects caused by larger order delays and reflections (see Subsection 3.4). CA3, on the other hand, is known to develop the so called 'place cells' (cells which are active only if the animal is at a given location ('place field' of the cell), like near the feeding place in the maze or at the center). It also has a very extensive recurrent collateral system, which, however, is mostly active only at one phase of the characteristic theta oscillation of the hippocampal formation [20, 21]. What it implies is that if approximate independence of the place cells is the end result of a separation process then separation can only take place when these recurrent connections are turned off so

interference can be avoided. CA3 projects onto CA1 which also features place cells showing even stronger independence. However, place field activity in CA1 can also be seen when only the direct connections from EC III are present, implying that these connections should also be able to form independent components. Subiculum shares many similarities with CA1, but instead of having place cells, it features the so called head direction cells which are active when the animal's head assumes a given relative direction. It is the major projection area of CA1, but it also receives input directly from EC III. An important finding [22] is that information flows from EC III and CA1 are specially crossed in the subiculum, that is, distinct loops are maintained carrying different bits of the flow of the same origin. Following the logic that requires separate representation of the factors, we conjecture that subicular activity represents the *ongoing* independent *processes* in synchrony with the independent driving sources represented by CA1. Intuitively, directional information would be such an independent process for navigational tasks. In order to be able to recombine the factors, an additional layer is needed, which is targeted by both components. The wiring diagram suggests that the wanted area should be the deep layers of EC which are also connected to the input layers of EC. As we noted earlier, separation can be greatly facilitated if decorrelation takes place first. To get a properly decorrelated estimated noise, activity at both EC II and EC III should be decorrelated. If the cortical input is not decorrelated, connections from the deep layers of EC should carry on this transformation. Furthermore the algorithmic derivations show that activity at EC V should also be decorrelated. To meet this requirement, we conjecture that projections from both CA1 and subiculum are tuned to help decorrelate the activity at the 'model' layer EC V. In this way the loop is closed and the information may go around till the right representations of the two factors, namely the position and the direction, are formed. For more information, see the supplementary material [17].

4.1. *Dynamics of the network*

Functional mapping would consist of description of the dynamics at two time scales: tuning (learning) of the synaptic weights and the working (information processing) of the whole system. Space limitations allow for discussing only some of the most intriguing characteristics but see [17].

4.1.1. *Co-learning in the parallel systems*

We have seen that the *very same transformation* (denoted by W_{inno}) is needed to recover the hidden driving sources, $e(t)$ and the hidden processes, $s(t)$. Due to

ambiguities of the ICA solution, like the undetermined order of components (see, e.g., [8]), the emerging components of the different factors should be put in register, otherwise the integrating model ($h(t)$) would result in distorted estimation of the incoming inputs. While it is easy to realize two identical matrices on a computer, neuronal implementation of such 'transformation adjustments' should be explicitly treated.

In our model, CA1 is the focal area that sub-serves the matching of indices in the whole network. This area is unique from the point of view of the anatomy as it has no excitatory recurrent collateral system and so intra-areal mixing is limited. Accordingly, CA1 is also central to our model; the 'transformation adjustments' of the matrices should be arranged by its activity. Although network level synchronizations unique to the HR and the parallel routes described above may allow for different schemes, experimental evidences about the exact nature of information flow through the two pathways are insufficient or controversial. In turn, we only propose some potential mechanism by which the required registration may be implemented.

It is likely [26] that the pace maker theta oscillation can differentiate the inputs projected onto CA1. We speculate that transformation along both pathways could be driven by the input statistics. In our scheme, however, ICA learning depends on output activity and the backpropagated signals it gives rise to. Learning may be coupled in turn, because signal backpropagation may effect both pathways. However, there is another condition for learning: synapses should be activated by the *proper* bottom-up signals that (could have) produce(d) the outputs. By assuming an interlaced signaling mechanism, it may be possible that at one pace the trisynaptic pathway is tuned in an unsupervised manner driven by the observation noise. At the next pace both channels may transmit signals proportional to the observed input and the resulting activity at CA1 constrained by the trisynaptic signals reshapes the direct pathway via signal backpropagation. An additional option is that the recurrent connection system of CA3 may in principle be able to integrate the observation noise represented by EC II and thus its output is proportional to the observation. This scenario implies that 1, the observation noise should be sustained long enough to be the input to CA1 at one pace and drive the activity in CA3 at the next one; 2, the recurrent connections of CA3 should be turned off when integration is not needed and 3, pair pulsed facilitation is required at CA1 so the corresponding components of $s(t)$ and $e(t)$ can be put in register. Interestingly, there are findings indicating that all these requirements are actually met [26, 27, 28].

4.1.2. *The subiculum*

The intriguing ‘cross-wired’ topographical projections from CA1 and EC to the subiculum, i.e., that proximal parts of CA1 project onto the distal side of subiculum [23], may allow for separate channelling of noise and state related information. On the ground of the similarities between the wiring diagrams of the CA1 and the subiculum [22] we may conjecture that subiculum also realizes a double encoding mechanism carrying representations of the moving averages of the *independent processes* and the related innovations, i.e., the changes between these processes, respectively. In navigational tasks, running in different directions may appear as an averaged process while a change of this process occurs when the animal stops, so positional information may be seen as the related innovation. In turn, CA1 is responsible to transfer the estimated independent driving sources while subiculum maintains the representations of the independent processes. These signals may then be integrated in the internal model maintained by the deep layers of EC. The mechanism of this integration remains to be answered.

5. Discussion

Lack of space only allows for a short account on how our model has evolved since an early proposal of [18]. Excellent reviews on other computational models can, for example, be found in [24,25] and [17] discusses how our models fares against some of the main ideas found in other models.

In the work of [18], (i) subiculum was not modeled, (ii) decorrelation or whitening was the putative role of CA3. Since 2000, considerable amount of information has emerged about the HR, including the time sharing mode of CA1 during the positive and the negative theta phases [23]. In addition, theoretical advances [6,18] extended independent component analysis to independent process analysis and it has also been shown that Hebbian learning is potentially feasible [9]. In the present work we have built on these results and derived a more detailed model of the hippocampal region.

There are three particular points of our work: (i) on the ground of computational considerations on goal-oriented systems we argued that forming representations is linked to the problem of independent process analysis (ii) we speculate that supervised aspects of signal backpropagation guided training ‘adjusting the transformations’ of two parallel pathways may indeed be realized in the hippocampal region through the activity of CA1 (iii) our construction may shed light on the role of the subiculum in forming representations for navigation. Supporting numerical simulations can be found in [17].

References

1. L. R. Squire and S. Zola-Morgan, *Science* **253**, 1380 (1991).
2. G. L. Chadderdon, *J. of Cogn. Neurosci.* **18**, 242 (2006).
3. W. Schultz, *Neuron* **36**, 241 (2002).
4. C. Boutilier, R. Dearden and M. Goldszmidt, *Artif. Intell*. **121**, 1104 (2000).
5. I. Szita and A. Lőrincz, *Acta Cybern*, (in press).
6. B. Póczos, Z. Szabó, M. Kiszlinger and A. Lőrincz, *Lect. Notes in Comp. Sci*. **4666**, 252 (2007).
7. J. F. Cardoso. *In Proc. of ICASSP*, **4**, 1941(1998).
8. B. Póczos and A. Lőrincz. *In Proc. of ICML*, 673 (2005).
9. A. Lőrincz and Z. Szabó. *Neurocomputing*, **70**,1569 (2007).
10. M. Plumbey, *IEEE Signal Proc. Lett.* **9**, 177 (2002).
11. C. Parisien, C. H. Anderson and C. Eliasmith. *Neural Comp.*, **20,** 1473 (2008)
12. J.-F. Cardoso and B. H. Laheld. *IEEE Tr. Signal Proc.*, **40,** 3017 (1996).
13. S. Amari, A. Cichocki, and H. H. Yang, *Adv. in NIPS*,**8**, 757 (1996).
14. Gy. Buzsáki, M. Penttonen, Z. Nádasdy, A. Bragin, *PNAS.*,**90,** 9921 (1996).
15. G. Chechik. *Neural Computation*, **15**,1481 (2003).
16. Z. Szabó, B. Póczos, G. Szirtes, and A. Lőrincz. *Lect. Notes in Comp. Sci.* **4668**, 677 (2007).
17. A. Lőrincz, M. Kiszlinger, and G. Szirtes. http://arxiv.org/abs/0804.3176, (2008).
18. A. Lőrincz and Gy. Buzsáki. *Annals New York Acad Sci.* **911**, 83 (2000).
19. M. P. Witter and D. G. Amaral. In: *The Rat Nervous System* 635 (2004).
20. M. E. Hasselmo and C. Bodelon and B. Wyble (2002) *Neural Comp.*, **14**, 793 (2002).
21. Gy. Buzsáki, Rythms of the Brain, *Oxford Univ. Press* (2006).
22. J. Gigg. *Behavioural Brain Research*, **174**, 265 (2006).
23. G. Dragoi and Gy. Buzsáki. *Neuron,* **50,** 145 (2006).
24. L. R. Squire, *Psych. Rev.* **99**, 195 (1992).
25. P. Andersen, R. Morris, D. Amaral T. Bliss, and J. O'Keefe (Eds.) *Oxford Univ. Press* (2007).
26. J. R. Manns, E. A. Zilli, K. C. Ong, M. E. Hasselmo and H. Eichenbaum, *Neurobiol. of Learn. and Mem. ,* **87,** 9 (2007).
27. D. M. Villarreal, A. L. Gross, and B. E. Derrick, *J. of Neurosci.,* **27,** 49 (2007).
28. T. Klausberger and P. Somogyi, *Science,* **321,** 53 (2008).