Computational Model of the Entorhinal-Hippocampal Region Derived from a Single Principle

András Lőrincz[1] and György Buzsáki[2]

[1]Department of Information Systems, Eötvös Loránd University, Pázmány Péter sétány 1/D Budapest, Hungary H-1117, email: lorincz@valerie.inf.elte.hu

[2]Center for Molecular and Behavioral Neuroscience, Rutgers, The State University of New Jersey, 197 University Ave, Newark, NJ 07102, email: buzsaki@axon.rutgers.edu

Abstract

We show that several properties of the highly elaborate structure of the EC-HC loop can be explained using the single principle that to recall past and to foresee future events a predictive structure is necessary. Prediction based on information emerging from a high dimensional sensory system becomes less demanding if the processed sensory information can be separated into components that evolve independently. Networks that develop independent components (ICs) in an efficient manner can be built from two stages. We identify these stages with the CA3 and CA1 layers of the hippocampus (HC). The forming of ICs requires non-linear operation, whereas IC outputs arise under linear operation and thus two-phase operation follows. Concurrent occurrences of past and present events are required by Hebbian learning and can be achieved by delaying structures, e.g., by loops. The loop structure requires a third layer that we identify with the entorhinal cortex (EC). The output is not modified if the loop forms a dynamic reconstructing (generative) network (DRN), i.e., if the EC encodes a representation of the ICs. Proper encoding into the EC is possible during linear operation in a supervised manner. The DRN can be seen as an error compensating control architecture. The HC part of the DRN is inputted by the error, the mismatch between the primary input to the EC-HC loop and the reconstructed input conveyed by the hippocampus. Errors between primary input and reconstructed input can arise when the information is "novel". Delays can be compensated by tuned predictive structures. The two-phase operation requires two predictive structures. We assume that those correspond to the EC to CA1 connections that operate during theta phase and the recurrent collateral system of the CA3 field that operates during sharp wave phase. Learning modifies the memory system that may temporarily spoil the predictive network. Thus the novel information processed by the DRN may be temporally convolved. Convolution should be counteracted before IC analysis. We assume that blind source deconvolution is executed by the dentate gyrus and show that the dentate gyrus can satisfy the requirements.

I. Introduction

In this paper, we put forward a top-down computational model in an attempt to understand how different parts of the entorhinal-hippocampal region of the brain function as a complex, cooperative network. A main goal is to illustrate how the various sub-networks of the entorhinal-hippocampal system can perform different operations depending on the "state" of the brain. A traditional modeling approach is to regard distinct anatomical structures as functional subsystems. Each subsystem is assigned to a particular operation and eventually the subsystems are linked together (cf. [1]). Whereas modeling subsystems has intrinsic merits, oftentimes it is the interactions between structurally distinct brain regions that correspond to a measurable function. Accordingly, we attempt to describe the entorhinal-hippocampal system as a collection of structure-function relationships in which the particular cooperation of the various fields is determined by external factors, such as the availability of subcortical neuromodulators. Using the available anatomical and physiological knowledge, we propose various roles for the different regions (fields). When experimental data do not exist, the lack of empirical observation is "filled in" by operations, which provide a plausible fit for the overall function of the entorhinal-hippocampal region. Our basic assumption is that the major function subserved by the hippocampus is to develop a representation that can replay information and can make predictions. From this single principle we derive

the main components of the hippocampal-entorhinal loop. We also derive that an important function of the entorhinal-hippocampal system is to modify synaptic connections in the structures which provide inputs to the hippocampal formation. We refer to this function as the formation of memory traces or as the formation of long-term memory. Ample clinical and experimental evidence is available to support the view that the hippocampal formation is essential for forming certain types of long-term memories [1,2]. The exact definition of the types of memories involving the entorhinal-hippocampal system is still debated [3-6]. In the present paper, long-term memory formation is simply defined as alteration of synaptic connections in those cortical networks whose activity gave rise to hippocampal input.

Some of the model's features share similarities with previous models of the hippocampus and other cortical structures. The architecture of the reconstruction networks has some resemblance to the Adaptive Resonance Theory (ART) network developed for categorization [7]. ART has a "hidden layer" and develops "resonating" representations. ART selects the best representation based on resonance properties. Our network differs from ART, because it is not utilizing resonances and ART is not designed to optimize information transfer. Our network operates more like a comparator[8,9]: competition between units arises by the correcting dynamics that improves the representation based on the comparator's output, the reconstruction error. Our model is closely related to generative networks [10,11] including the dynamical sparse representation network introduced by Olshausen and Field [12]. Another part of our model concerns predictive structures that have been proposed for the modeling of the visual cortex by Rao and Ballard [13] in their Kalman-filter architecture. Roweis and Ghahramani [14] give a unifying view of this network family. Economical predictive structures can be formed between independent components (ICs) that minimize mutual information. Lörincz and coworkers have suggested that the EC-HC loop develops ICs [15-17]. IC analysis is an unsupervised blind method, which is intimately related to information maximizing principles [18-24].

II. The model

The activity of a computational unit is proportional to the sum of the filtered inputs received through its input connections:

$$\mathbf{a} = P^T\mathbf{x} \qquad (1)$$

where superscript T denotes transposition, $\mathbf{x}$ is the input vector that has m components: $\mathbf{x}=(x_1,...,x_m)$, there are n computational units, the output vector of the computational units is given by $\mathbf{a}=(a_1,...,a_n)$, and the and the ij^{th} element of the connection strength matrix P connects j^{th} input component to the i^{th} computational unit c_i and its strength is equal to p_{ij}. In most cases it is assumed that the number of computational units is smaller or equal to the number of inputs, i.e., $n \le m$. Activity vector $\mathbf{a}$ is sometimes called the internal representation of input $\mathbf{x}$ computed by the layer.

Relative to the number of pyramidal cells the number of feedback inhibitory neurons is small. Some interneuronal types exert a local control on the principal cell populations whereas others form "interstitium-like" networks with axonal connections different from, or opposite to, the excitatory circuits of the principal neurons (cf. [25]). We consider those interneuron classes whose axon collaterals have a similar spatial extent to the excitatory inputs. For example the HIPP cells of the dentate gyrus and the O-LM cells of the CA3 field belong to this class [26-28]. We assume that the inhibitory feedback through these interneurons is fast compared to the characteristic averaging time window of rate coding. Then, from the viewpoint rate coding, the effect of inhibitory neurons on each pyramidal cell represents an average inhibition. This average inhibition can be balanced by tunable excitatory synapses. We assume that the effects of excitatory and inhibitory synapses can be summed. No distinction will be made between inhibitory and excitatory synapses, instead, we simplify the mathematical formalism by using synapses that can have both signs. For example, if the synaptic strength of the inhibitory feedback axonal terminals overrides the synaptic strength of the matching excitatory terminals then we shall talk about negative synaptic strength, or negative connection strength.

The generic dynamic equation for each layer can be written as

$$\partial_t\mathbf{a} = -I(\mathbf{a}) + P_1^T\mathbf{x} - P_2^TQ\mathbf{a} \qquad (2)$$

where ∂_t denotes temporal derivation, P_1 denotes synapses on pyramidal cells dendrites, and Q denotes synapses that excite those inhibitory feedback neurons which target the somata of pyramidal cells via synapses given by matrix P_2. The membrane potential may be subject to different leakages (losses) that may assume non-linear form and is represented by function $I(.)$ identical for all of the components of its argument. Eq. (2) suits the model in [12] provided that $I(.)$ represents a saturating function. Eq. (2) can be extended by additive predictive structures, such as $vT\mathbf{a}$ (where v is a constant and T denotes an n by n predictive matrix) at the level of the input representation in the case of the CA3 recurrent collaterals, or acting between the input and the output layer in the

form of $\nu F\mathbf{x}$ (where F is an n by m matrix) for the case of the EC-CA1 connections, as in [13]. Equation (2) also describes the EC-HC loop: $P_1 = P_2$ represents the HC, $\mathbf{a}$ is the output of the HC, $Q\mathbf{a}$ represents the effect of the HC output on the deep layers and then on the superficial layers of the EC, and $(\mathbf{x}–Q\mathbf{a})$ is the reconstruction error.

Synaptic modification requires a large postsynaptic membrane depolarization of sufficient duration to open Ca^{2+}-permeable N-methyl-d-aspartate (NMDA) channels or voltage-gated Ca^{2+} channels [29]. This may be brought about by bursting presynaptic neurons or synchronous firing of several presynaptic fibers. We assume that "effective coincidences" are responsible for another type of learning that gives rise to the development of ICs. We assume that this part of learning depends on the synaptic strength in a linear fashion. In other words, during the bursting phase a stronger synapse will lead to stronger postsynaptic depolarization. The increasing postsynaptic activity is also dampened by the increased discharge of recurrently connected interneurons. Consequently, the magnitude of potentiation is inversely related to the activity of inhibitory interneurons impinging upon the postsynaptic principal cell. Interneuronal activity limits potentiation. The learning equation is written in the form:

$$\Delta P_1 = \eta_1 \mathbf{x}\mathbf{a}^T + \eta_2(P_1 - P_1\mathbf{a}\mathbf{a}^T) \qquad (3)$$

where matrix P_1 denotes the excitatory memory matrix, η_1, η_2 and γ are positive numbers. We assume that parameter η_2 depends on the firing rate and becomes zero at low firing rates. We assume that low firing rate is the result of strong inhibitions that can be modeled by a lossy mode. It is also assumed that in lossy mode the output activity vector can be approximated as $\mathbf{a}=g(P_1^T\mathbf{x})\approx I^{-1}(P_1^T\mathbf{x}-P_2^TQ\mathbf{a})$ where function $I(.)$ is monotone increasing and function $g(.)$ can be taken, e.g., as function $\tanh$. During linear operation – that we identify with the sharp wave phase (SWP) – losses are absent and mostly the second term of Eq. (3), i.e., $(P_1 - P_1\mathbf{a}\mathbf{a}^T)$ controls learning. The first term in this parenthesis, P_1, has been explained; the second term has a negative sign, it represents a limiting effect on synaptic learning. The second component of the second term is the synaptic memory matrix P_1. It describes that the stronger the synapses that transmit a given input, the larger the probability that the postsynaptic neuron discharges. Because the number of inhibitory neurons is much smaller than the number of pyramidal cells it allows us to approximate the cumulated inhibition at synapse P_{ik} as $\Sigma_j P_{ij}a_j$. The second term is also proportional to the postsynaptic activity of the neuron because in case of no postsynaptic activity no limiting effect is assumed for the synapse. Taken together, the expression $P_1\mathbf{a}\mathbf{a}^T$ follows. The term in the parentheses of Eq. (1) will be referred to as the statistical part of the learning rule. It can be considered as a normalizing term. The linear (strengthening) term dominates for small synaptic strength values. Thus, the synaptic vectors cannot diminish provided that the learning rate η_2 is sufficiently large. On the other hand, the limiting term may override the strengthening effect when the synaptic weight is large and synaptic vectors cannot grow without limit. Note that the firing rate-dependence of the statistical part of the learning rule makes the traditional Hebbian part relatively less effective at higher discharge rates.

We assume that supervised training is influenced by the membrane potential: the sign of learning is controlled by the sign of $(\mathbf{x}–Q\mathbf{a})$ and the learning rule can be given as

$$\Delta Q = (\mathbf{x}–Q\mathbf{a})\mathbf{a}^T \qquad (4)$$

where $\mathbf{a}$ denotes the output vector to be learnt.

The learning rule for predictive matrices can be written as $\Delta T=\partial_t \mathbf{a}\mathbf{a}^T$ for the CA3 recurrent collaterals and $\Delta T=\partial_t \mathbf{a}\mathbf{x}^T$ for the EC-CA1 connections. We assume that $\partial_t \mathbf{a}$, i.e., the instantaneous balance of excitatory and inhibitory components at the neuron that the synapse belongs to plays a role here.

The CA3 layer is assumed to operate in lossless mode during both phases. Then learning in the CA3 layer corresponds to whitening [23]. The recurrent collaterals learn to predict during theta phase and replay the learnt sequences during SWP. The CA1 layer operates in lossless mode during SWP and in lossy mode during the theta phase when it is relatively quiet. It follows [24] that the CA1 layer learns to separate. The CA1 layer produces separated outputs during linear operation (i.e., during SWP). We assume that synaptic integration occurs at the HC-to-EC synapses and that the reconstruction error approximates the reconstruction error of linear operation in spite of the saturating non-linearities. Thus the reconstruction error approximates $(\mathbf{x}-Q\mathbf{a})$. Learning rule (4) represents latent changes in the deep layer-to-superficial layer synapses of the EC during the theta phase and consolidating neural activities at the deep layer of the EC during SWP. Temporal integration gives rise to a multiplicative factor during SWP. The learning procedure encodes the ICs into the LTM of the EC [24]. Encoding ensures that the loop can output the ICs and thus upon encoding predictive structures can learn to predict ICs.

The relaxing DRN convolves the input that spoils statistical analysis unless inputs to the CA3 stage are deconvolved. It has been argued [15,16] that the EC-HC loop can be seen as an error compensating control architecture with a proportional and an integrating arm, the latter being the dentate gyrus (DG). Integration can

approximately linearize the non-linear losses and we assume that the EC-HC loop has linear losses. A DRN with linear losses has special convolving properties: a separating matrix can unmixes different convolutions. In short, our model requires that (i) the DG should be the substance for deconvolution, (ii) learning in the DG should modify its integrating role to a deconvolving role, and allows that (iii) the DG may be built from a separation stage and a component-wise deconvolution stage.

The DG is ideally suited for this task. Recurrent excitatory feedback loops (via the mossy cells) exist in the DG that can temporally integrate. Excitatory feedback from CA3 pyramidal cells also impinges onto the granule cells. These excitatory feedback loops are matched by inhibitory feedback loops via the HICAP cells (feedback from granule cells) and via the HIPP cells (feedback from CA3 cells). Mossy cell synapses can be tuned and thus the net feedback could be strictly positive (temporal integration) or of either polarities for blind source deconvolution.

The assumptions of our model for the dentate gyrus are as follows:

i. Granule cells and inhibitory cells form a lossy structure
ii. Losses of granule cells are supra-linear for small output activities and become approximately constant in the high output activity region
iii. Losses can be considered as contrast enhancing means that promote the forming of a sparse representation
iv. Sparse representation decomposes different configurations into different subspaces
v. Mossy cell – granule cell loops form delay lines with different (including zero) delays
vi. Granule cell – CA3 pyramidal cell – mossy cell – granule cell loops form delay lines with different delays
vii. The excitatory loops (formed by mossy cells) and the inhibitory loops (formed by HIPP and HICAP cells) can give rise to either positive or negative net feedback effects and the recurrent feedback loops can play an integrating as well as a deconvolving role
viii. Active granule cell synapses that are far from the somata (entorhinal afferents) are subject to the statistical learning rule and perform whitening upon tuning
ix. Active granule cell synapses that are close to the somata (mossy afferents) undergo non-linear tuning and the non-delaying loops learn to separate, whereas the delaying loops learn to deconvolve
x. The neocortex is assumed to have a “switching property” [30] and thus active dentate gyrus neuronal subsets represent amplitude distributions determined by temporal convolution alone
xi. The existence of recurrent feedback loops from the CA3 layer allows to deconvolve the direct EC – to – CA3 inputs
xii. Upon tuning inputs to the CA3 layer becomes deconvolved and thus the CA3 layer can perform proper whitening.

III. Computational results on predictive structures

Simulations were performed on a one-dimensional ring [17]. The ring was constructed by considering a finite input interval ([0.0, 1.0]) and by setting points 0.0 and 1.0 identical. It was assumed that a "previous processing stage" provided local excitations in the form of Gaussian functions. The ring was discretized by 20 input units, and input units sampled the Gaussian excitations at equidistant discrete points around the ring. The sampled value of the Gaussian function at a discretization point and at time t is considered as the excitatory input of the corresponding component of the input vector at time t. The Gaussian function had a variance of 0.1. The network consisted of 20 computational units. The connection strength vector for each unit was “prewired”. Connection strengths to each unit were sampled values of Gaussian functions of variance of 0.1 centered at different grid points. The centers of connection strength vectors were placed equidistantly on the grid. The excitatory input was moved around the ring with unit speed, i.e., the Gaussian excitation circled once per unit time. Figure 1 depicts the evolution of the predictive matrix T and its effect on the reconstruction error. The time of the simulation was 20 time units that corresponds to the time of 20 round trips. Therefore, the second and the third rows of the third column illustrate that reconstruction error can be improved considerably in a "single session", i.e., by making a single presentation of the input along the circle.

The effect of the predictive contribution $\mathbf{Ta}$ can be adapted by modifying the constant v in the term $v\mathbf{Ta}$ using the following adaptation rule:

$$\partial_t v = - \beta \mathbf{a}^T \mathbf{T} \mathbf{a} \qquad (5)$$

This adaptation rule can be derived for constant motion speeds and for locally linear representations. Simulations on the one dimensional ring and fast switching of the motion speed exhibited fast relaxing transients and, in turn, proper predictions after these transients have passed.

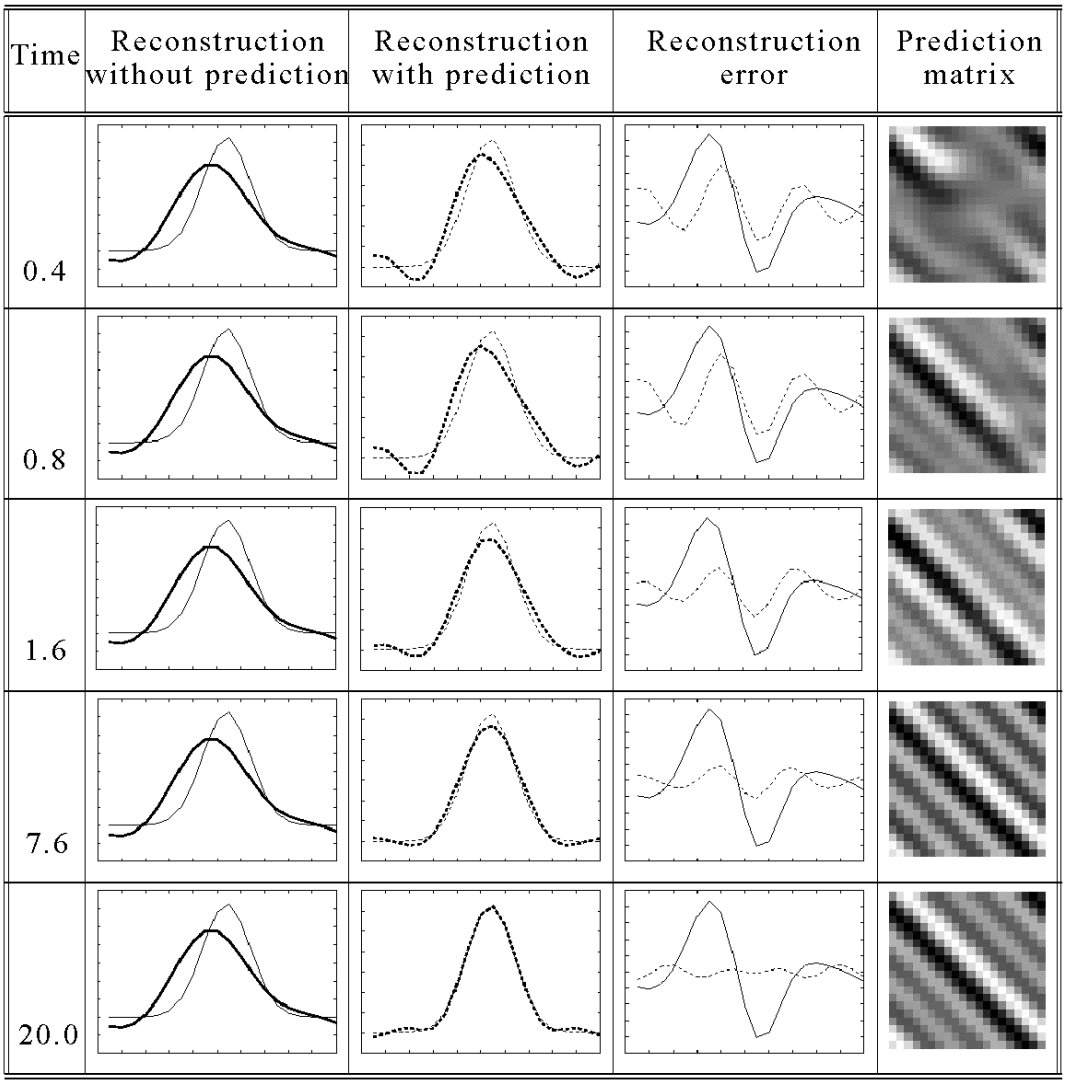

Figure 1. Training of predictive matrix T and the evolution of reconstruction. The first (TIME) column gives the actual time for Gaussian excitation to move around a circle. 20 time units correspond to one round trip time. Bold lines: reconstructed inputs. Thin solid lines: inputs and reconstruction errors without prediction. Thin dashed lines: inputs and reconstructed error with prediction. Note that training the prediction matrix in a single run decreases the reconstruction error. Predictive matrix elements (depicted in gray scale) are shown in the right-hand column at different stages of the training. Matrix elements were initialized by zeros and are depicted in matrix form. The strength of an element is illustrated in gray scale in the right-hand column at different stages of training. The farther a matrix element is from the diagonal the further apart are the receptive fields of the corresponding computational units and the larger is the time span of prediction. Gray level 0.5 corresponds to zero connection strength, darker (lighter) gray levels represent positive (negative) connection strengths.

IV. Conclusions

We have presented a model of the EC-HC loop. The model starts from the assumption that representation serves prediction. Considering the "cocktail-party problem", one can argue that predictive structures are less complex if we have a representation that "separates the speakers" and if temporal learning can be restricted (i) to the prediction of individual "speakers" and (ii) to the learning of temporal relations between the "speakers". Independent components are, however, not orthogonal and we have utilized a dynamic reconstruction network to optimize information transfer. This DRN reverberates information allowing to use Hebbian means for the tuning of the predictive structures. Learning of the independent components could be, however, severely limited by the convolving (reverberating) properties of this DRN itself, and thus an architecture performing blind source deconvolution is

required in order to counteract the undesirable convolving properties. We argued that the dentate gyrus can satisfy the strict requirements imposed on the deconvolving structure.

Acknowledgments. Thanks are due to Lehel Csató and Zoltán Gábor who performed the numerical simulations. This work was supported by OTKA (T14566, T17100), NIH (NS34994, MH54671), the Human Frontier Science Program and the US-Hungarian Joint Fund (No. 519).

References

1. Gluck MA (guest editor) (1996) Computational models of hippocampal function in memory. Hippocampus 6:565-762.
2. Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. J Neurol Neurosurg Psychiat 20:11-21.
3. O'Keefe J, Nadel L. (1978) The hippocampus as a cognitive map. Oxford: Clarendon Press.
4. Squire, LR (1992) Declarative and nondeclarative memory: multiple brain systems supporting learning and memory. J Cog Neurosci 4:232-243.
5. Eichenbaum H, Otto T, Cohen NJ (1994) Two functional components of the hippocampal memory system. Behav Brain Sci. 17:449:472.
6. Burgess N, O'Keefe, J (1996) Neuronal computation underlying the firing of place cells and their role in navigation. Hippocampus 7: 1-15.
7. Carpenter GA, Grossberg S (1993) Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. TINS 16:131-137.
8. Grastyan E, Lissak K, Madarasz I, Donhoffer H (1959) The hippocampal electrical activity during the development of conditioned reflexes. Electroencephal Clin Neurophysiol 11:409-430.
9. Sokolov EN (1963) Perception and the conditioned reflex. London: Pergamon Press.
10. Hinton GE, Sejnowski TJ (1983) Optimal perceptual inference. Proc. of the IEEE Computer Society Conf. on Vision and Pattern Recognition, pp. 448-453.
11. Hinton GE, Ghahramani Z (1997) Generative models for discovering sparse distributed representations. Proc Trans Roy Soc B 352: 1177-1190
12. Olshausen BA, Field DJ (1996) Emergence of Simple-Cell Receptive field properties by learning a sparse code for natural images. Nature 381: 607-609.
13. Rao RPN, Ballard DH (1997) Dynamic Model of Visual Recognition Predicts Neural Response Properties in the Visual Cortex. Neural Computation 9:721-763.
14. Roweis A, Ghahramani Z (1999) A unifying review of linear Gaussian models. Neural Computation (In press).
15. Lörincz A (1997) Hippocampal formation trains independent components via forcing input reconstruction In: Proceedings of ICANN'97 (Gerstner W, Germond A, Hasler M, and Nicoud, JD, eds) Berlin: Springer-Verlag, pp. 163-168.
16. Lörincz A (1998) Forming independent components via temporal locking of reconstruction architectures: a functional model of the hippocampus. Biological Cybernetics 79: 263-275.
17. Lörincz A, Csató L, Gábor Z, M. Molnár, Gy. Buzsáki (1998) A two stage computational model training long-term memories in the entorhinal-hippocampal region Collection of Abstracts, 28th Annual Meeting of the Neuroscience Society, Los Angeles, CA, p. 921.
18. Jutten C, Herault J (1991) Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing 24:1-10.
19. Comon P (1994) Independent component analysis, a new concept? Signal Processing 36:287-314.
20. Bell AJ, Sejnowski TJ (1995) An information maximization approach to blind separation and blind deconvolution. Neural Computation 7:1129-1159.
21. Wang L, Karhunen J, Oja E (1995) A bigradient optimization approach for robust PCA, MCA, and source separation. Proceedings of the IEEE ICNN, Perth, Australia, USA: IEEE Publishing, pp. 1684-1689.
22. Amari SL, Cichocki A, Yang HH (1996) A new learning algorithm for blind signal separation. In: Advances. In: Neural information processing systems 8, (Touretzky D, Mozer M, and Hasselmo M, eds.) pp. 757-763. Cambridge MA: MIT Press.
23. Cardoso JF, Laheld B (1996) Equivalent adaptive source separation. IEEE Trans on Signal Proc 44:3017-3030.
24. Karhunen J, Oja E, Wang L, Vigario R, Joutsensalo J (1997) A class of neural networks for independent component analysis. IEEE Trans. Neural Networks 8:486-504.
25. Freund TF, Buzsáki G (1996) Interneurons of the hippocampus. Hippocampus 6:345-470.
26. Han ZS, Buhl EH, Lörinczi Z, Somogyi P (1993) A high degree of spatial selectivity in the axonal and dendritic domains of physiologically identified local-circuit neurons in the dentate gyrus of the rat hippocampus. Eur J Neurosci 5:395-410.
27. Sik A, Penttonen M, Ylinen A, Buzsáki G (1995) Hippocampal CA1 interneurons: an in vivo intracellular labeling study. J Neurosci 15:6651-6665.
28. Sik A, Penttonen M, Buzsáki G (1997) Interneurons in the hippocampal dentate gyrus: an in vivo itnracellular study. Eur J Neurosci 9:573-588.
29. Markram H, Lubke J, Frotscher M, Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275:213-215.
30. Ghahramani Z, Hinton GE (1999) Switching space models. (Submitted).