Autoregressive model of the hippocampal representation of events

András Lőrincz and Gábor Szirtes

***Abstract*—The hippocampal formation is believed to play a central role in forming long lasting representation of events. However, in contrast to the continuous nature of sensory signal flow, events are spatially and temporally bounded processes. In this paper we are interested in the kind of representation that allows for detecting and/or predicting events. Based on new results on the identification problem of linear hidden processes, we propose a general signal encoding model that can represent causal relationships used to define events. We translate the model into a connectionist structure in which parameter learning follows biologically plausible rules. We also speculate on the resemblance of the resulting structure to the connection system of the hippocampal formation. When our signal encoding model is applied on *spatially* anchored inputs, its different parts feature spatially localized *and* periodic neural activity similar to those found in the hippocampus and in the entorhinal cortex, respectively. These emergent forms of spatial activity differentiates our model from other computational models of (spatial) memory as the model has not been explicitly designed to deal with spatial information. We speculate that our model may describe the core function of the hippocampal region in forming episodic memory and supporting spatial navigation.**

I. Introduction

ONE of the spectacular functions of the brain is to detect and predict events which may intuitively be defined as spatially and temporally limited change of a given state or condition of the observed environment (which may also include internal state variables of the observer). While events may or may not be in a causal relation (since end state of one event may be the start state of another one), it makes sense to learn causality of signals that trigger a switch between states and the resulting events. For example, an animal may be still or moving fast and a switch between these states (the event) can be triggered by different causes (detecting a predator or a potential mating partner). By learning this causality, it becomes possible to predict succeeding states from the detection of a particular trigger. However, in contrast to the widely used coding mechanisms where codewords can easily be distinguished, we receive a continuous flow of sensory signals from which we should first decode the triggers. These triggers then can be used to segment the sensory signals into finite chunks that allow for efficient representation of events (where efficiency may, for example, regard memory capacity). In Section II we formalize our assumptions on the sensory signals and show how the resulting model can be used for segmentation in which the notion of statistical independence plays a key role. In Section III we translate the algorithms into a connectionist network in which parameter tuning can only be realized via biologically plausible local interactions. In Section IV we parallel some interesting properties of the architecture with the features unique to the hippocampal region (HR), the brain area centrally involved in forming the so called episodic memory [1], [2], [3]. We also show some simulation results on the dynamics of the model. Finally, in Section V we shortly discuss a few open questions and formalize predictions on the functioning and dynamics of HR.

András Lőrincz and Gábor Szirtes are with the Department of Information Systems, Eötvös Loránd University, Pázmány P. sétány 1/C, Budapest H-1117, (email: andras.lorincz@elte.hu, szirtes@inf.elte.hu).

II. Identification of the sensory input

It is natural to interpret the observations ($x(t) \in \mathbf{R}^d$) as mixed signals emitted by a hidden (not directly observable) state variable ($s(t) \in \mathbf{R}^d$) that evolves in time thus forming a process. The simplest case is if linearity is assumed both for the mixing and the dynamics of the process. Accordingly, the observations and the hidden process may be written as:

$$x(t) = As(t), \quad (1)$$

$$s(t+1) = \sum_{i=1}^{I} F_i s(t-i) + \sum_{j=0}^{J} H_j e(t+1-j), \quad (2)$$

that is the observations are instantaneous mixtures of the state components whose evolution follows an autoregressive moving average process (ARMA) of order (I, J) with driving noise (or innovation process) $e(t) \in \mathbf{R}^d$. In general the driving noise components are assumed to be *temporally* independent and identically distributed (i.i.d.) stochastic variables. However, in accord with our causal definition of events we also assume that noise components (or at least their subgroups) are *spatially* (i.e., index-wise) independent. We assume that matrix $H_0 \in \mathbf{R}^{d\times d}$ is the identity matrix $I_d \in \mathbf{R}^{d\times d}$. The ARMA$(I, J)$ model comprises the contributions of previous states transferred by the predictive matrices F_i $(i = 1, \ldots, I)$ and the different echoes of the driving noise transferred by matrices H_j $(j = 1, \ldots, J)$. The goal is to find the $\hat{F}_i \in \mathbf{R}^{d\times d}$ $(i = 1, \ldots, I)$ and $\hat{H}_j \in \mathbf{R}^{d\times d}$ $(j = 1, \ldots, J)$ estimations of the hidden dynamics and the echo structure, respectively, and to learn to separately represent the estimated hidden state, $\hat{s}(t)$, and the independent subgroups of the estimated driving noise $\hat{e}(t)$.

Regarding the structure of F_i, $i = (1, \ldots, I)$, a special case seems relevant: *joint block-diagonal structure* implies dynamical sub-processes that do not *mix*. These hidden processes are independent in a dynamical sense so their identification could reduce the representational complexity.

Below, for the sake of conceptual simplicity, we assume a hidden ARMA$(1, 0)$=AR(1) process and the issue of delays will be discussed later. Theoretical details of higher order

978-1-4244-3553-1/09/$25.00 ©2009 IEEE

hidden ARMA processes and post-nonlinear extensions can be found in [4], [5], [6] and in [7], respectively.

The key step to solve the identification problem is to recognize that the observation process is also an AR(1) process, if matrix A can be inverted [6]:

$$x(t+1) = Mx(t) + n(t), \tag{3}$$

where the observation noise is $n(t+1) = Ae(t+1)$. According to the d-central limit theorem [8] $n(t+1)$ is approximately Gaussian so the predictive matrix ($M = AFA^{-1}$) may be estimated by least-mean square approximations and then the wanted independent driving noise components can be extracted by applying Independent Component/Subspace Analysis (ICA/ISA) [9], [10], [11] on the observation noise. An important result [12], [13] is that – for a large class of source distributions – separation of independent subspaces (i.e., the ISA problem) can be solved in two steps. First, traditional ICA methods yield one dimensional components and second, the resulting components should be grouped to form independent subspaces. In turn, ISA of the estimated noise can simultaneously recover the estimated mixing matrix $\hat{A}$, the hidden state $\hat{s}(t) = \hat{A}^{-1}x(t)$ as well as the assumed independent subspaces of the multidimensional driving source components $\hat{e}(t) = \hat{A}^{-1}\hat{n}(t)$. If $\hat{A}$ is recovered, then the hidden predictive matrix can also be approximated ($\hat{F} = \hat{A}^{-1}\hat{M}\hat{A}$).

III. A connectionist network implementation

In this section we provide local learning rules for parameter estimation of the identification task described above. The resulting algorithms can be translated into a neural network in which 'activity' of a 'neuronal layer' is represented by a vector, connection weights between layers or within layer components (i.e., recurrent connections) are represented by matrices and neurons may realize non-linear transformations of their inputs.

A. Assumptions and simplifications

For simplicity, rate coding (manifesting analogue values) and mixed, i.e., positive and negative weights (which realization contradicts Dale's Principle) are assumed throughout the derivations, but the proposed functioning can in principle be also realized by using either positive coding [14] or homogenous connection systems [15]. Inhibitions (or subtractions) are manifested by separate inhibitory populations within a layer using feedback or feed-forward inhibition.

We also assume that after each transformation the resulting entities (represented by a given layer) get decorrelated and normalized (i.e. they are subject to *whitening*). Whitening helps compressed encoding [16] and it speeds up ICA [17], [18] as a preprocessing step. Whitening may take place within a layer with the help of inhibitory recurrent connections for which biologically feasible learning rules can also be given [19]. Interestingly, whitening can also be realized by utilizing inter-layer feed-back connections. The so called forward-inverse model [20] or reconstruction network [21] assumes a *loopy* structure with a *bottom-up* (BU) and a *top-down* (TD) transformations in which representation of the input is used to regenerate an estimate of the input and different constraints can be put on the transformations. It can be shown that in a reconstruction network the two branches can learn to invert each other – at least in the pseudo-inverse sense – applying simple Hebbian rules that we introduce later. Both whitening and ICA can be realized in a reconstruction network for which the mathematical details and a possible implementation scheme will soon be provided.

The different algorithmic tasks and the related learning rules will be described in separate subsections below.

B. Innovation

The first algorithmic step is the estimation of the innovation: $\hat{n}(t+1) = x(t+1) - \hat{M}x(t)$. The corresponding cost function $J(\hat{M}) = \frac{1}{2}\sum_t |x(t+1) - \hat{M}\,x(t)|^2$ leads to the following Hebbian learning rule: $\Delta\hat{M}(t+1) \propto \hat{n}(t+1)x(t)'$, where $'$ denotes transpose. Since this rule requires simultaneous access to the error and the observation, the realization of this rule assumes two separate layers representing the observation and the innovation, respectively. As we see, while the learning rule is based on positive correlations, but the transformation itself is a subtraction, $\hat{M}$ is supposed to act through feed-forward inhibition. It also implies that observation is projected onto both layers. To fulfill the whitening requirement on the represented entities, both the architecture and the learning rule need to be modified. Let $z(t)$ and $\hat{m}(t)$ denote the whitened observation and the whitened estimated innovation, respectively. Distinct whitening procedures are needed for the two signals. Observation can be made white by intra-layer inhibitory collaterals. Innovation, however, may *also* get whitened by integrating the two layers into a loop structure. The effect of the loop is represented by an auxiliary term in the cost function:

$$J(\hat{M}) = \frac{1}{2}\sum_t |x(t+1) - \hat{M}\,z(t) + Q\,\nu(t+1)|^2 \tag{4}$$

The exact role of $Q \in \mathbf{R}^{d\times d}$ and $\nu(t+1) \in \mathbf{R}^d$ will be given later. For now it is enough to conceive $\nu(t+1)$ as an internal representation of the innovation created by a reconstruction network (see, Fig. 1). The advantage of the auxiliary connection system Q is that learning rule for Q

$$\Delta Q \propto [Q']^{-1} - \hat{m}(t+1)\nu(t+1)' \tag{5}$$

can make the activity at the innovation layer white [17]. This learning rule can be strictly Hebbian by means of *effective coincidences* [22] (see later). In addition, the online learning rule of $\hat{M}$ still assumes a pure Hebbian form:

$$\Delta\hat{M}(t+1) \propto \hat{m}(t+1)z'(t), \tag{6}$$

C. Separation

The next module extracts the independent noise components from the innovation: $\hat{e}(t) = W\hat{m}(t)$. Among several solutions that can estimate *demixing matrix* W, the so called Infomax online rule [23] suits our architecture:

$$\Delta W(t+1) \propto [W(t)']^{-1} - f\big(\hat{e}(t)\big)\hat{m}(t)', \tag{7}$$

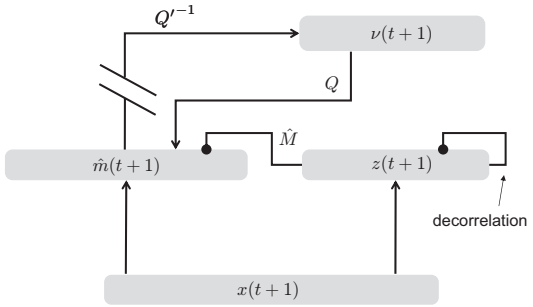

Figure 1. Connectionist architecture that learns AR dynamics by Hebbian learning and computes whitened innovation.

Here, $f(.)$ denotes tangent hyperbolic component-wise nonlinearity. This rule is actually a nonlinear version of Eq. (5) and locality can also be ensured by the loopy structure. Let us remark that pre-whitening significantly accelerates Eq. (7) [17], [18]. In order to couple the hidden states with the corresponding driving noise, a transformation that matches indices and compensates for differences of preprocessing steps should be learnt on the white observation: $\hat{s}(t) = \tilde{W}z(t)$. The trivial solution is if $\hat{m}$ and an estimation of z are channeled through the same connection system. Another solution is if the learnt transformation is 'mirrored' onto a parallel BU channel thus transmitting the two signals separately. The second choice fits the innovation module that yielded a *separate* representation of the white innovation and the state. We will argue in Subsection IV-E that both solutions may be realized in a neurally plausible way and that HR may indeed use both types of processing, offering a *robust solution* to this delicate problem.

D. Learning to predict

The *excitatory* predictive matrix of the hidden process ($\hat{F}$) can be learnt with the help of the recovered independent driving noise components and the estimated hidden states. To comply with the whitening requirement, the internal predictive model is supposed to work on white signals. The resulting model could be used to predict future hidden states even if detection of the sensory stimuli is perturbed (for example, by occlusion or noise). We propose a separate layer with a recurrent network that receives two inputs; the estimated driving noise ($\hat{e}$) and the estimated hidden state ($\hat{s}$) channeled through the same connection system, but in a temporally multiplexed *two-phase* manner. Let h denote the (intra-layer) whitened hidden states. The prediction of future hidden states assumes the following dynamics:

$$h(t+1) = \hat{F}h(t) + \nu(t+1). \tag{8}$$

Since the transformed driving source, $\nu(t+1)$, acts as an error signal, a supervised Hebbian learning rule can be given:

$$\Delta\hat{F}(t+1) \propto \nu(t+1)h(t)' \tag{9}$$

which minimizes the cost function

$$J(\hat{F}) = \frac{1}{2}\sum_t |h(t+1) - \hat{F}\,h(t)|^2. \tag{10}$$

Although computationally such rule is feasible, the neuronal implementation is not straightforward as we will discuss.

IV. Functional mapping

As an interesting motivation to our work, the gross connection system of the mammalian hippocampal region (comprising the entorhinal cortex (EC), subfields CA3 and CA1, dentate gyrus (DG), para- and presubiculum and the subiculum [24]) is depicted in Fig. 2(a).

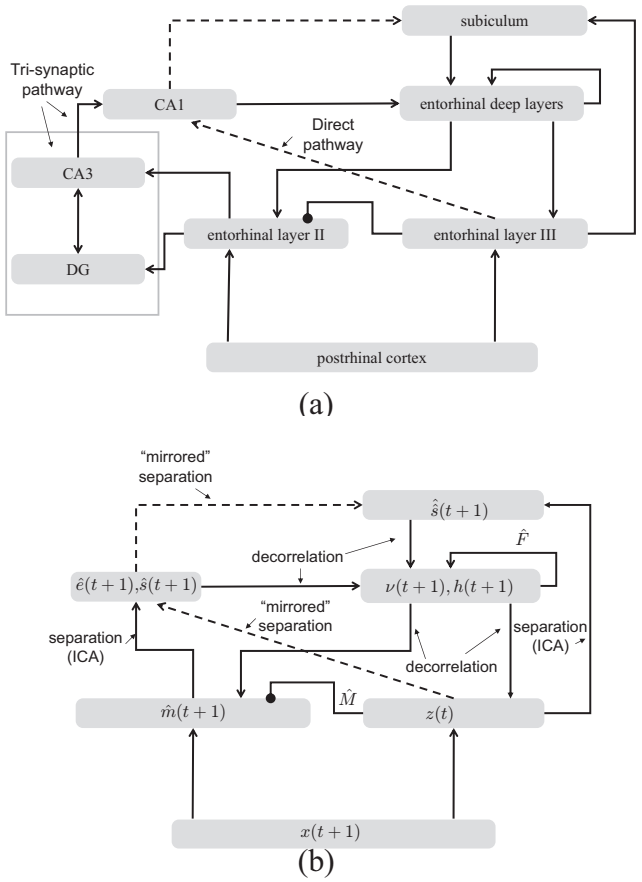

Figure 2. (a) The main wiring diagram of the hippocampal region. For the sake of completeness, loops between the dentate gyrus (DG) and CA3 are also depicted. In our model, the putative role of these loops is to remove large delays (moving averages). (b) Connectionist architecture for representing the independent noise and state recovered from the observation signals. Dashed lined denote connections tuned in a supervised manner. Arrowhead lines denote mostly excitatory, circle head denotes inhibitory connections. Note the two ICAs giving rise to $\hat{s}$ and $\hat{\hat{s}}$, respectively.

Here we argue that our proposed learning model may describe an important aspect of episodic memory learning in the HR. First we highlight some relevant and unique features of the neural substrate, which may also support other models, but their absence would probably *falsify* ours. Experimental data mainly come from rat studies, but believed to be similar for other mammalian brains, too. Also we combine data from anatomical, episodic memory and spatial navigation studies, including the particular place and grid like activity in CA1 (CA3) and in the EC, respectively.

A. Direction of information flow

First, there is a dominantly unidirectional ([25], but see [26]), and parallel connection system among all parts: superficial layers of EC receive input from adjacent cortical regions and transmit the signals toward CA1 and also toward the subiculum. Signal transmission to CA1 takes place in two parallel routes: the so-called *tri-synaptic* connection system (EC II–DG–CA3–CA1) and the *direct* route from EC III to CA1. They presumably transmit different information, yet they can control each other in driving CA1 [27]. Another interaction is the so called *cooperative learning* of the proximal and the distal synapses on CA1 [28] when local dendritic spikes can induce learning in other synapses even without postsynaptic activity. As the exact nature of the input received by EC II and EC III is not known, so we assume that the superficial layers share the same cortical input. We also assume that differences in the activity of these layers stem from their differing intrinsic physiology (e.g. the ratio of interneurons that enables strong feedforward inhibition in EC II), anatomy (role of recurrent collaterals) and the received feedback (EC layers V/VI project back to both layers).

CA1 projects to the subiculum, which receives information directly from EC III, too. CA1 as well as the subiculum in turn project back to the deep layers of EC. The parallel systems in part preserve topographical arrangement [29] but there exists a separation along the lateral to medial direction. The lateral and medial parts of the entorhinal cortex (LEC and MEC, respectively) receive input from different cortical areas and, in turn, project to non-overlapping portions of CA1 and the subiculum. In contrast, DG and CA3 receive convergent input from both LEC and MEC. The functional consequence is that the fusion of spatial and non-spatial information may be strictly controlled within the HR [30], [31].

EC deep layers, which presumably also receive modulatory signals from different cortical areas, close the loop: they send mostly excitatory [32] feedback to the superficial layers.

B. Unique intra-regional interactions

Although place cells (that is cells with spatially localized unimodal activity distribution, [33]) can be found in DG, CA3 and CA1, their coding mechanism may be quite different, as the underlying connection systems have significantly distinct features. DG is unique for its temporally tunable connections [34]. CA3 has a dense collateral system which has a particular role in memory replay [35], [36], [37], [38], [39]. CA1, as a single exception in the whole circuitry, has no recurrent collaterals and the activity of the principal cells seems to be independent[40].

C. Temporal synchrony

In addition to the intricate anatomy, the physiology of the separate modules is also striking. The prominent synchronized membrane potential oscillations (theta: 4–10Hz, gamma: 40–100 Hz) can be generated within (EC II or EC III) and outside the HR (septum) and have differential effects on the different modules. The oscillations have been proposed to control synchrony throughout the circuitry [41] or to provide an internal reference clock [42], [43].

Interestingly, while EC III is phase locked to the main theta and can maintain persistent activity [44], EC II is very close to EC III and can also initiate theta activity yet it shows phase precession, similarly to the place cells in the hippocampus [45].

Deep layers of the EC show peculiar functioning as well. In contrast to the superficial layers, EC V can generate input specific graded persistent activity in individual neurons [46] which is generally considered the underlying neural mechanism of working memory [47]. Furthermore, the relative homogeneity of the CA1 response to changing inputs as compared to that seen in the deep EC may suggest [48] that active CA1 neurons are engaged in representing one environment, while deep EC may contain multiple sub-populations, some tied to CA1 output while others are more independent of CA1. Interestingly, separate modules or ‘cell islands’ can be found in EC II as well [31]. As a consequence, if deep layers can represent several likely models concerning the world, there should be a switching mechanism that can help select the one that best serves the correct predictive coding. It is intriguing that layer III of the EC has been found to receive such switching signals [44].

D. Space related activity

During spatial navigation tasks, signals carrying different aspects of spatial information, such as position (DG, CA3 and CA1), head-direction (mainly subiculum) or speed (all subfields), seem to interfere at several stages. While activity in most CA3 and CA1 cells does not show correlation with directional information, postsubicular head-direction cells directly innervate the deep layers of EC, which in turn send this information to the superficial layers. According to this scenario, grid cells in EC III show clear conjunctive correlation representing mixed information[49]. However, the activity of neurons in EC layer II is free of directional modulation.

E. Mapping of the architecture

The proposed connectionist architecture is shown in Fig. 2(b) to help the comparison with the gross anatomy of HR, depicted on Figs. 2(a). The superficial entorhinal layers are supposed to receive inputs (highly processed sensory stimuli and motion related activity) from the cortex. Due to the proximity of EC layers II and III, they may share the same input thus providing the parallel pathways required by the innovation step. As EC II has an extensive inhibitory network, we speculate that connections from EC III mainly act through feed-forward inhibition and EC II represents innovation. The recurrent connections in both layers may take part in whitening, because the diversity of inhibitory populations may enable concurrent functions without interference. In addition, projections from EC deep layers may also contribute to whitening of EC II activity. Due to lack of space connections between EC II and CA3

through DG are collapsed and the tri-synaptic pathway is treated as one transformation (but see [21], [50], where compensation of long delays was assigned to this pathway). In turn, the superficial layers communicate with CA1 through two parallel channels that we identify as the separation and mirrored separation that transmit the estimated independent noise and the transformed states, respectively. The supervised training of the hidden model requires the matching between the noise and state components. In turn, separation matrices between EC II to CA1 and EC III to CA1 should be tightly coupled in spite of the potential differences of the whitening transformations acting on EC II and EC III and the permutation invariance of ICA. We speculate that this *index matching* of the two learning processes can be realized by 'cooperative learning' between the distal and proximal synapses of CA1 [28], [51]. We also speculate that CA3 is able to transmit different signals in different phases of theta. When its recurrent connections are suppressed, it transmits the innovation and tuning may be governed by Eq. (7) yielding the separation transformation. However, when the collaterals are active, CA3 can integrate the previous and current innovations to signal an estimate of the observation. In this phase, CA3 to CA1 transformation yields the estimated hidden state which serves as a *supervisory* output of the *other* channel thus substituting the role of the postsynaptic activity in Hebbian learning. Nevertheless the direct route shares common properties with the tri-synaptic route and may also learn to separate via learning rule (7) unless supervisory CA3 signals override it. This proposal is in accord with the findings [52] that approximate place fields emerge first in CA1, but place fields of CA3 cells get stabilized faster. When the parallel channels are correctly tuned, then their transformed inputs are adjusted to each other. We also conjecture that in one phase of the theta wave CA1 transmits the estimated hidden driving sources and in the other phase it transmits the estimated hidden states. Since the components of the latter term may not be independent it implies that CA1 output collected during the two phases should show different statistics.

Another critical issue about the separation is that the necessary synaptic enhancement of separation (regarding the term $[W']^{-1}$ in Eq. (7)) cannot be realized by a recurrent connection system as it is absent in CA1. Our proposal is built on the reconstruction network idea utilizing loopy connections. As Eq. (7) shows, the second Hebbian term directly depends on the pre- and postsynaptic activity. However, the first term is proportional to the inverse transpose of the transformation matrix ('top-down' (TD) connections) which is basically equals the rest of the loop. Since the reconstruction network is aiming to regenerate the input, the reconstruction error decreases as parameter tuning advances. In turn, this TD term will transmit mostly noise for which the so called *effective coincidences* [22] can maintain locality. This proposal is in accord with the anatomy of HR, because CA1 projects onto the deep layers of EC, which in turn project onto the input (superficial) layers of EC.

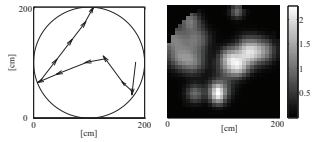

Figure 3. (a): Circular maze, diameter: 2m, with a short sample trajectory. Step size varies between 11 and 55 cm. (b): Sample input to the loop in the form of an activity map within the maze. Activity map is shown in arbitrary units.

EC deep layer to superficial layer transformations may also take part in the whitening of the activity at the superficial layers. In turn, the auxiliary term $Q\nu(t+1)$ of Eq. (4) may augment intra-layer whitening through recurrent collaterals at the superficial layers, provided that noise of the model network is transmitted top-down in one of the theta phases.

The subiculum as one of the chief projection areas of CA1 is not modeled now, although we hypothesize that the special cross connections found in CA1 and subiculum [30] are used to create a second separation loop in order to have a distinct place for representing hidden states $\hat{s}$ denoted by $\hat{\hat{s}}$ in Fig. 2(b), because – in the absence of another index matching process – they may differ. Information from CA1 and subiculum are optimized for grouping them into independent subspaces for the hidden causes and into independent deterministic processes for the hidden states. Such grouping may occur at the deep layers of EC as well as at other parts of the brain including the frontal lobe.

F. Simulation results

In our simulations[1], the 'rat' 'runs' in a 2m wide, open-field circular arena on a linear path at a constant speed and makes a random turn at each step with a given chance. It makes a random turn also, if possible collision with the wall is detected. Input sampling has been fixed to 55cm (Fig. 3). The most restrictive approximation in our simulations is that the inputs to the HR do not contain information about distant cues. It implies that distortions of spatial activity distributions as a function of maze distortions [53] cannot be modeled this way. On the other hand, this simplification enabled us to avoid the arbitrariness in modeling low-level sensory processing, because we could simply mimic postrhinal [54] inputs.

Information was cumulated on the trajectories and was analyzed by Auto-Regressive Independent Process Analysis methods [12], [13]. The following two types of inputs have been used: 1, 1000 mixed local cues similar to the one shown in the right hand side of Fig. 3 and 2, conjunctive inputs, where the same patches have directional tuning with randomly chosen directions.

As opposed to real spiking data, linear transformations may give rise to negative signals. In turn, the correspondence between the unit activity values after each transformation and the neurons's response is not straightforward. For generating the activity maps of the input units, first we discretized the

[1]For a detailed description of the simulations see the online supplementary material [50].

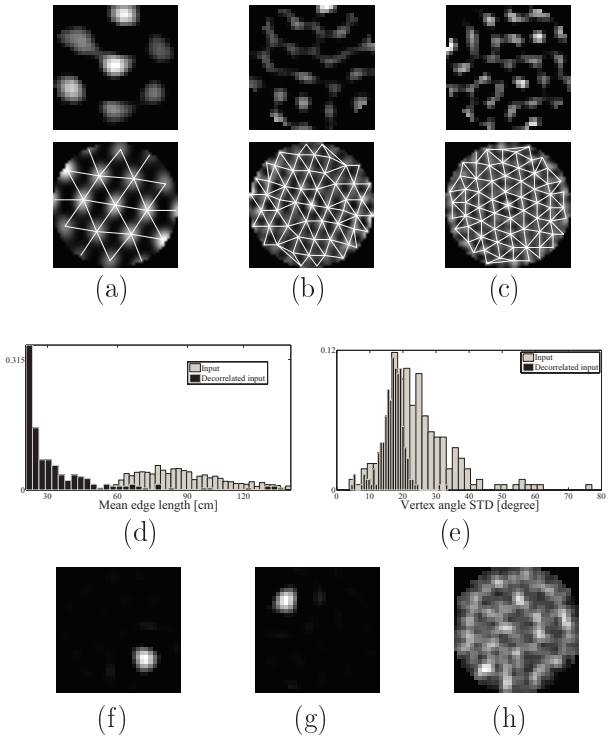

Figure 4. Position dependent input. (a-d): each column shows the output of different *decorrelating* units (whitening). First row: half-wave rectified and scaled activity maps (0: black, 1: white). Second row: 2D autocorrelation function of the activity maps and the fitted grids. Third row: vertex angle histogram for the fitted grids. (e-f): cumulative statistics over all grids. (e): histogram of the mean edge length of the grids for the input set and the whitening units. (f): histogram of the standard deviation of the vertex angles for the input set and the whitening units. (g-i): sign corrected activity maps of three *separating* (ICA) units. Response is localized. (j): superimposed map of all ICA units demonstrating that the localized units cover the full maze. For more details, see, [50]

space (the resolution was 30 so a bin is $6.67\,cm \times 6.67\,cm$, which is comparable of [55], and for each bin we summed up the activity measured in those steps that ended in the given bin. This spatial averaging smoothes out the artifacts caused by unattended spots. The activity after, e.g., decorrelation may assume negative values so the data were clipped below zero and was scaled to the $[0, 1]$ range. Separation is invariant for the change of sign [10]. In turn, we chose to set the sign as follows: we collected the 10 highest absolute value bins, took the sign of the majority of these 10 values, multiplied the activity map with this sign, clipped all negative bins afterwards, and scaled the data to the $[0, 1]$ range. We also computed the 2 dimensional normalized autocorrelation for each activity map. For the spatial analysis of the peak activity regions of the autocorrelation image we fitted a grid on the locally maximal points using Delaunay-triangulation [56], [57].

For the case of local cues we found reasonable hexagonal grid structure upon whitening (Fig. 4(a-d)). Mean edge lengths of all grids covered a large domain upon decorrelation, because individual grids had different characteristic length (Fig. 4(e)). Angles between the vertices of the hexagonal grids were close to 60 degrees (Fig. 4(a-d)) and their standard deviation was less than 20 degrees, much sharper than for random grids (Fig. 4(f)). ICA provided place cells (Fig. 4(g-i)). The size of the cells depended on the number of ICA components; but for all cases, they covered the arena uniformly (Fig. 4(j)).

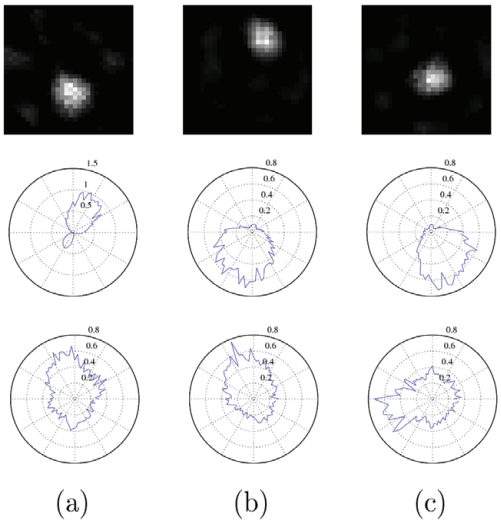

Figure 5. ICA of *conjunctive inputs* and ICA of the *innovation of conjunctive inputs*. Upper and middle rows: ICA on conjunctive inputs. Bottom row: ICA on innovation of the conjunctive inputs. Each column corresponds to 3 sample output units of the *separation* (a-c). First row: clipped and scaled activity maps of conjunctive inputs. Maximal activity values are: $2.9, 4.4, 4.3$. Middle row: overall direction selectivity depicted on a polar plot. Bottom row: overall direction selectivity depicted on a polar plot. Learning the innovations considerably decreases direction sensitivity and yet the responses remain localized (not shown).

For the case of conjunctive inputs, whitening and ICA were both direction dependent (we show the results for ICA in Fig. 5). We computed the innovation along the straight lines of the trajectories that decreased the directional dependence considerably. (Compare the two sets of polar plots in Fig. 5). At the same time we have experienced a slight degradation of the place fields [50]. In another set of experiments (not shown, but see [50]) directional information was mixed with local cues giving rise to apparent conjunctive representation upon whitening. Separation led to rotationally invariant place cells and global directional cells in this case.

V. Discussion

Episodic memory has been identified as one of the key components of the memory system, yet the term 'episode' or 'event' is usually defined by arbitrary examples. In this paper we argued that events should be identified as spatially and temporally bounded changes in the hidden state behind the observations about the environment. This interpretation of events, however, poses a problem as the wanted events should be detected from the continuous flow of sensory signals. Our solution to this issue is based on modeling the sensory signals as a hidden ARMA process driven by independent sources. For continuous signals, independence usually is coupled with sparse activity, so the sources can

naturally act as time stamps for events. In addition, extraction of *multidimensional* independent sources requires extraction and grouping of the one dimensional components (cf. ISA). We hypothesize that *CA1's unique function* is to recover the simple sources which then are grouped (second step of Independent Subspace Analysis, [6], [13]) and integrated by EC as well as the neocortex. This grouping is then an integral part of memory consolidation (on systems consolidation, see [58]). Furthermore, once this integration is fully learnt, it may function even without HR's separating ability. The proposed signal encoding model is able to extract the hidden driving sources (triggers or causes), represent separately the hidden states and learn a transformed form of the hidden dynamics that allows for prediction of relevant changes (events) in the environment even if full access to the observation cannot be granted. In addition, the resulting internal model for predicting the dynamics of the hidden states is a natural place for path-integration which is thought to be a fundamental part of any biologically realistic spatial navigation model (see, e.g. [59]).

Description of events, however, is not complete without defining a corresponding time scale. Changes at a given time scale may trigger other changes at different time scales, thus defining a *hierarchy* of events. In our simulations signals interpreted as triggers have shown place cell like activity. Due to the coarse sampling and the lack of distal cues, place fields are always traversed in one time step. In real experiments, however, place cells are active for several theta periods. In turn, our model corresponds to dynamics with slow triggers present for a time span compatible with the animal's speed instead of the sampling rate.

Next, we translated the proposed algorithmic model into a connectionist framework in which biologically appealing (Hebbian or local) interactions are allowed for parameter learning. Based on the apparent similarity between the resulting network and the gross anatomy of the hippocampal region we mapped each functional module onto a given part of HR guided by computational and biological constraints as well. When our signal encoding model is applied on spatially defined inputs, the emergent (that is not designed) spatial behavior of the different modules resemble the corresponding parts of the HR thus giving additional support to our functional mapping. This resemblance may give rise to a different interpretation of the place cells. They may not describe space per se, but they represent a set of observations that can trigger given changes of the hidden state described by the learnt internal model, so sequences of events without spatial information should activate similarly these cells.

Although our computational assumptions may seem overly simplified, they resulted in a consistent computational model of event representation in HR. In spite of its mathematical simplicity, *Hebbian constraint on learning* gives rise to a sophisticated architecture. The model provides several testable predictions regarding properties at synapse and network level of HR. The first prediction is about the importance of the loopy structure generally overlooked by other models. We predict if no *output* of CA1 is allowed (but the cells' functioning is left intact) than formation of place fields in new environments will be impaired. The second prediction is about the particular coupling of the tuning mechanism by CA1 (index matching of the separation transformation). We predict if this system is perturbed then the representation of head directions in the subiculum will be impaired. The last prediction is that if the tight control over the two phase processing is perturbed then learning of the predictive internal model represented by the deep layers of EC will be impaired.

The current model does not provide a full description of HR yet as the integration with the tri-synaptic pathway is missing, but see [21]. Also, the issue of multiple environments and goals/contexts (learning and switching of separate representations)[60] need to be treated. In addition, several questions remained open. One of the most important problems is the learning of invariance of the input space. In our simulation we used local cues only thus avoiding this problem. Some models have already been proposed to overcome this difficulty [61], [62], [63].

References

[1] J. O'Keefe and L. Nadel, *The Hippocampus as a Cognitive Map*. Clarendon, Oxford, 1978.

[2] H. A. Eichenbaum, "cortical-hippocampal system for declarative memory," *Nature Rev. Neurosci.*, vol. 1, pp. 41–50, 2000.

[3] F. Vargha-Khadem, D. G. Gadian, K. E. Watkins, A. Connelly, W. Van Paesschen, and M. Mishkin, "Differential Effects of Early Hippocampal Pathology on Episodic and Semantic Memory," *Science*, vol. 277, no. 5324, pp. 376–380, 1997.

[4] A. Lőrincz and Z. Szabó, "Neurally plausible, non-combinatorial iterative independent process analysis," *Neurocomputing*, vol. 70, pp. 1569–1573, 2007.

[5] Z. Szabó, B. Póczos, and A. Lőrincz, "Undercomplete blind subspace deconvolution via linear prediction," *Lecture Notes in Artificial Intelligence*, vol. 4701, pp. 740–747, 2007.

[6] ——, "Undercomplete blind subspace deconvolution," *Journal of Machine Learning Research*, vol. 8, pp. 1063–1095, 2007.

[7] Z. Szabó, B. Póczos, G. Szirtes, and A. Lőrincz, "Post nonlinear independent subspace analysis," *Lecture Notes in Computer Science*, vol. 4668, pp. 677–686, 2007.

[8] V. V. Petrov, "Central limit theorem for m-dependent variables," in *Proceedings of the All-Union Conference on Probability Theory and Mathematical Statistics*, 1958, pp. 38–44.

[9] P. Comon, "Independent Component Analysis, a new concept?" *Signal Processing, Elsevier*, vol. 36, no. 3, pp. 287–314, Apr. 1994, special issue on Higher-Order Statistics.

[10] C. Jutten and J. Hérault, "Blind separation of sources. Part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[11] J. Cardoso, "Multidimensional independent component analysis," in *Proc. of ICASSP*, vol. 4, 1998, pp. 1941–1944.

[12] B. Póczos and A. Lőrincz, "Non-combinatorial estimation of independent autoregressive sources," *Neurocomputing*, vol. 69, pp. 2416–2419, 2006.

[13] B. Póczos, Z. Szabó, M. Kiszlinger, and A. Lőrincz, "Independent process analysis without a priori dimensional information," *Lect. Notes in Comp. Sci.*, vol. 4666, pp. 252–259, 2007.

[14] M. D. Plumbley, "Conditions for non-negative independent component analysis," *IEEE Signal Processing Letters*, vol. 9, no. 6, pp. 177–180, 2002.

[15] C. Parisien, C. H. Anderson, and C. Eliasmith, "Solving the problem of negative synaptic weights in cortical models," *Neural Computation*, vol. 20, pp. 1473–1500, 2008.

[16] M. Brand, "Fast low-rank modifications of the thin singular value decomposition," *Linear Algebra and its Applications,*, vol. 415, no. 1, pp. 20–30, 2006.

[17] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, vol. 44, pp. 3017–3030, 1996.

[18] S. I. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems*. San Mateo, CA: Morgan Kaufmann, 1996, pp. 757–763.

[19] P. Foldiak, "Adaptive network for optimal linear feature extraction," in *IEEE/INNS International Conference on Neural Networks*, 1989, pp. 401–405.

[20] M. Kawato, H. Hayakawa, and T. Inui, "A forward-inverse model of reciprocal connections between visual neocortical areas," *Network*, vol. 4, pp. 415–422, 1993.

[21] A. Lőrincz and G. Buzsáki, *The parahippocampal region: Implications for neurological and psychiatric diseases*, ser. Annals of the New York Academy of Sciences, 2000, no. 911, ch. Two-phase computational model of the entorhinal-hippocampal region, pp. 83–111.

[22] A. Lőrincz and G. Buzsáki, "Two–phase computational model training long–term memories in the entorhinal–hippocampal region," *Ann. New York Acad. Sci.*, vol. 911, pp. 83–111, 2000.

[23] A. Bell and T. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[24] M. P. Witter and D. G. Amaral, *The Rat Nervous System*, 3rd ed. Academic Press, San Diego, CA, 2004, ch. Hippocampal Formation, pp. 635–704.

[25] P. A. Naber, F. H. Lopes da Silva, and M. P. Witter, "Reciprocal connections between the entorhinal cortex and hippocampal fields CA1 and the subiculum are in register with the projections from CA1 to the subiculum," *Hippocampus*, vol. 11, pp. 99–104, 2001.

[26] L.-R. Shao and F. E. Dudek, "Electrophysiological evidence using focal flash photolysis of caged glutamate that CA1 pyramidal cells receive excitatory synaptic input from the subiculum," *Journal of Neurophysiology*, vol. 93, pp. 3007–3011, 2005.

[27] G. Buzsáki, *Rhythms of the Brain*. Oxford: Oxford University Press.

[28] N. L. Golding, N. P. Staff, and N. Spruston, "Dendritic spikes as a mechanism for cooperative long-term potentiation," *Nature*, no. 418, pp. 326–331, 2002.

[29] M. P. Witter, "Connections of the subiculum of the rat: Topography in relation to columnar and laminar organization," *Behavioural Brain Research*, vol. 174, pp. 251–264, 2006.

[30] J. Gigg, "Constraints on hippocampal processing imposed by the connectivity between CA1, subiculum and subicular targets," *Behavioural Brain Research*, vol. 174, pp. 265–271, 2006.

[31] M. P. Witter and E. I. Moser, "Spatial representation and the architecture of the entorhinal cortex," *Trends in Neurosciences*, vol. 29, pp. 671–678, 2006.

[32] T. van Haeften, L. B. te Bulte, P. H. Goede, F. G. Wouterlood, and M. P. Witter, "Morphological and numerical analysis of synaptic interactions between neurons in deep and superficial layers of the entorhinal cortex of the rat," *Hippocampus*, vol. 13, pp. 943–952, 2003.

[33] J. O'Keefe and J. Dostrovsky, "The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat," *Brain Research*, vol. 34, pp. 171–175, 1971.

[34] D. A. Henze, L. Wittner, and G. Buzsáki, "Single granule cells reliably discharge targets in the hippocampal CA3 network in vivo," *Nature Neuroscience*, vol. 5, pp. 790–795, 2002.

[35] K. Louie and M. A. Wilson, "Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep," *Neuron*, vol. 29, pp. 145–156, 2001.

[36] D. J. Foster and M. A. Wilson, "Reverse replay of behavioural sequences in hippocampal place cells during the awake state," *Nature*, vol. 440, pp. 680–683, 2006.

[37] K. Diba and G. Buzsáki, "Forward and reverse hippocampal place-cell sequences during ripples," *Nature Neuroscience*, vol. 10, pp. 1241–1242, 2007.

[38] J. Csicsvari, J. O'Neill, K. Allen, and T. Senior, "Place-selective firing contributes to the reverse-order reactivation of ca1 pyramidal cells during sharp waves in open-field exploration," *European Journal of Neuroscience*, vol. 26, pp. 704–716, 2007.

[39] J. O'Neill, T. J. Senior, K. Allen, J. R. Huxter, and J. Csicsvari, "Reactivation of experience-dependent cell assembly patterns in the hippocampus," *Nature Neuroscience*, vol. 11, pp. 209–215, 2008.

[40] A. D. Redish, F. P. Battaglia, M. K. Chawla, A. D. Ekstrom, J. L. Gerrard, P. Lipa, E. S. Rosenzweig, P. F. Worley, J. F. Guzowski, B. L. McNaughton, and C. A. Barnes, "Independence of firing correlates of anatomically proximate hippocampal pyramidal cells," *Journal of Neuroscience*, vol. 21, pp. 1–6, 2001.

[41] M. J. Denham and R. M. Borisyuk, "A model of theta rhythm production in the septal-hippocampal system and its modulation by ascending brain stem pathways," *Hippocampus*, vol. 10, pp. 698–716, 2000.

[42] J. Jefferys, R. Traub, and M. Whittington, "Neuronal networks for induced '40 Hz' rhythms," *Trends in Neuroscience*, vol. 19, pp. 202–208, 1996.

[43] O. Jensen, M. Idiart, and J. Lisman, "Physiologically realistic formation of autoassociative memory in networks with theta/gamma oscillations: role of fast NMDA channels," *Learning and Memory*, vol. 3, pp. 243–256, 1996.

[44] B. Tahvildari, E. Fransen, A. A. Alonso, and M. E. Hasselmo, "Switching between "on" and "off" states of persistent activity in lateral entorhinal layer iii neurons," *Hippocampus*, vol. 17, pp. 257–263, 2007.

[45] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Topographic organization of a spatial map in the entorhinal cortex," ser. Neuroscience 2005. SfN, 2005, p. 198.3.

[46] A. V. Egorov, B. N. Hamam, E. Fransen, M. E. Hasselmo, and A. A. Alonso, "Graded persistent activity in entorhinal cortex neurons," *Nature*, vol. 420, pp. 173–178, 2002.

[47] P. S. Goldman-Rakic, "Cellular basis of working memory," *Neuron*, vol. 14, pp. 477–485, 1995.

[48] L. M. Frank, E. N. Brown, and G. B. Stanley, "Hippocampal and cortical place cell plasticity: Implications for episodic memory," *Hippocampus*, vol. 16, pp. 775–784, 2006.

[49] F. Sargolini, M. Fyhn, T. Hafting, B. L. McNaughton, M. P. Witter, M.-B. Moser, and E. I. Moser, "Conjunctive representation of position, direction, and velocity in entorhinal cortex," *Science*, vol. 312, pp. 758–762, 2006.

[50] A. Lőrincz, M. Kiszlinger, and G. Szirtes, "Model of the hippocampal formation explains the coexistence of grid cells and place cells," http://arxiv.org/abs/0804.3176, 2008.

[51] N. Mehta, S. Natarajan, P. Tadepalli, and A. Fern, "Transfer in variable reward hierarchical reinforcement learning," in *NIPS 2005 Worksop on Inductive Transfer*, 2005.

[52] S. Leutgeb, J. K. Leutgeb, A. Treves, M. Moser, and E. I. Moser, "Distinct ensemble codes in hippocampal areas ca3 and ca1," *Science*, vol. 305, no. 5688, pp. 1295–1298, 2004.

[53] C. Barry, R. Hayman, N. Burgess, and K. J. Jeffery, "Experience-dependent rescaling of entorhinal grids," *Nature Neuroscience*, vol. 10, no. 6, pp. 682–684, 2007.

[54] R. D. Burwell and D. M. Hafeman, "Positional firing properties of postrhinal cortex neurons," *Neuroscience*, vol. 119, no. 2, pp. 577–588, 2003.

[55] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, pp. 801–806, 2005.

[56] E. J. Markus, Y.-L. Qin, B. Leonard, W. E. Skaggs, B. L. McNaughton, and C. A. Barnes, "Interactions between location and task affect the spatial and directional firing of hippocampal neurons," *Journal of Neuroscience*, vol. 15, pp. 7079–7094, 1995.

[57] B. Takács and A. Lőrincz, "Simple conditions for forming triangular grids," *Neurocomputing*, vol. 70, pp. 1741–1747, 2007.

[58] L. Nadel, G. Winocur, L. Ryan, and M. Moscovitch, "Systems consolidation and hippocampus: two views," *Debates in Neuroscience*, vol. 1, no. 2–4, pp. 55–66, 2007.

[59] B. L. McNaughton, F. P. Battaglia, O. Jensen, E. I. Moser, and M. Moser, "Path integration and the neural basis of the cognitive map," *Nature Rev. Neuro.*, vol. 7, pp. 663–678, 2006.

[60] V. Hok, P.-P. Lenck-Santini, S. Roux, E. Save, R. U. Muller, and B. Poucet, "Goal-related activity in hippocampal place cells," *Journal of Neuroscience,*, vol. 27, no. 3, pp. 472–482, 2007.

[61] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances." *Neural Computation*, vol. 14, no. 4, pp. 715–770, 2002.

[62] M. Franzius, H. Sprekeler, and L. Wiskott, "Slowness and sparseness lead to place, head-direction, and spatial-view cells," http://cogprints.org/5711/, 2007.

[63] P. Byrne and S. Becker, "A principle for learning egocentric-allocentric transformations." *Neural Computation*, vol. 20, no. 3, pp. 709–737, 2008.