Neural Kalman-filter

Gábor Szirtes [a] Barnabás Póczos [a] András Lőrincz [a,1]

[a] *Department of Information Systems*
Eötvös Loránd University
Pázmány Péter sétány 1/C., 1117 Budapest, Hungary

Abstract

Anticipating future events is a crucial function of the central nervous system and can be modelled by Kalman-filter like mechanisms which are optimal for predicting *linear* dynamical systems. *Connectionist* representation of such mechanisms with Hebbian learning rules has not yet been derived. We show that the recursive prediction error method offers a solution that can be mapped onto the entorhinal-hippocampal loop in a biologically plausible way. Model predictions are provided.

Key words: Kalman-filter, hippocampus, entorhinal cortex, neural prediction

1 Introduction

Linear dynamical systems (LDS) are widely applied tools in state estimation and control tasks. The so-called Kalman-filter recursion (KFR) makes the inference in LDS simple, the resulting estimations are unbiased and have minimized covariance. Here, an approximation of the KFR is provided (Section 2) by applying the recursive prediction error (RPE) method [1]. We show that the approximation (i) can be represented in neuronal form, (ii) is efficient and (iii) can be mapped (Section 3) onto the entorhinal-hippocampal (EC-HC) loop, the center of memory functions. Conclusions are drawn in Section 4.

2 Approximated Kalman-filter recursion

Consider the following LDS:

[1] Corresponding author. E-mail: andras.lorincz@elte.hu

$$\text{observation process}: \quad \mathbf{y}^t = \mathbf{H}\mathbf{x}^t + \mathbf{n}^t \tag{1}$$

$$\text{hidden process}: \quad \mathbf{x}^{t+1} = \mathbf{F}\mathbf{x}^t + \mathbf{m}^t \tag{2}$$

where variables $\mathbf{m}^t \propto \mathcal{N}(0, \Pi)$, $\mathbf{n}^t \propto \mathcal{N}(0, \Sigma)$ are independent Gaussian noise processes. The task is to estimate the hidden variables $\mathbf{x}^t \in \mathbf{R}^n$ given the series of observations $\mathbf{y}^\tau \in \mathbf{R}^p$, $\tau \leq t$. For squared norm in the cost, the optimal solution was derived in [2]. The *prediction equation* is used to estimate $\mathbf{x}$ *before* the $(t+1)^{th}$ measurement:

$$\hat{\mathbf{x}}^{(t+1|t)} = \mathbf{F}\hat{\mathbf{x}}^{(t|t-1)} + \mathbf{K}^t(\mathbf{y}^t - \mathbf{H}\hat{\mathbf{x}}^{(t|t-1)}) = \mathbf{F}\hat{\mathbf{x}}^{(t|t)}. \tag{3}$$

where $\mathbf{K}^t$ is the 'Kalman-gain', which can be computed by means of *a priori* and *posteriori* covariance matrices of $\hat{\mathbf{x}}^t$. Expression $\mathbf{e}^t = \mathbf{y}^t - \mathbf{H}\hat{\mathbf{x}}^{(t|t-1)}$ in Eq. 3 can be identified as the *reconstruction error*, because, for the noiseless case, $\mathbf{H}\hat{\mathbf{x}}^{(t|t-1)}$ should perfectly match the input. Kalman-gain balances error $\mathbf{e}^t$ and model based prediction $\mathbf{F}\hat{\mathbf{x}}^{(t|t-1)}$ to optimize the estimation.

The first problem of the classical solution is that covariance matrices of measurement and observation noises (Π and Σ) are generally assumed to be known. The second problem is that to ensure dynamic adaptation of the Kalman-gain, the algorithm requires the calculation of a matrix inversion, which is hard to interpret in neurobiological terms. In this section we derive an approximation of the Kalman-gain which eliminates these problems. Our approximation makes use of the RPE method. The resulting scheme is (i) local, (ii) well suited to 'track' the changing world and (iii) asymptotically optimal by construction under mild conditions [3]. Let $\mathbf{K}^t\mathbf{z} \approx \theta^t .* \mathbf{K}\mathbf{z}$ denote an arbitrary parametrization of $\mathbf{K}^t$, where $.*$ denotes element-wise multiplication. The RPE approximation of KFR using this arbitrary parametrization is as follows:

$$\hat{x}_i^{t+1} = \sum_j F_{ij}\hat{x}_j^t + \theta_i^t \sum_l K_{il} e_l^t \tag{4}$$

in which $\theta^t \in \mathbf{R}^p$, $\hat{\mathbf{x}}^{t+1} = \hat{\mathbf{x}}^{(t+1|t)}$. For simplicity, the notation of the error's explicit dependence on θ is dropped. Let us suppose that a suboptimal matrix $\mathbf{K}(\theta^0)$ is given at time $t = 0$. Our goal is to tune parameter θ^t in order to minimize $J_k(\theta_k) = \frac{1}{2}E[(\epsilon_k^t)^2]$ with respect to θ_k, where $E[.]$ is the expectation operator and $\epsilon_k^t = \sum_l K_{kl} e_l^t$ is the transformed error. Stochastic gradient approximation provides the following update equations:

$$\theta_k^{t+1} = \theta_k^t + \alpha \sum_{lj} K_{kl} H_{lj} W_{jk} \epsilon_k^t \tag{5}$$

$$W_{ik}^{t+1} = \sum_j F_{ij} W_{jk}^t \xi_k - \theta_i^t \sum_{lj} K_{il} H_{lj} W_{jk}^t \xi_k + \delta_{ik} \epsilon_k^t \tag{6}$$

where $W_{ik}^t = \frac{\partial \hat{x}_i^t}{\partial \theta_k}$ is an auxiliary matrix, δ is Kronecker's delta, $\xi = (\xi_1, \ldots \xi_n)^T$ has been introduced to provide a conventional neuronal equation, learning rate α may depend on time, $\mathbf{F}$ and $\mathbf{H}$ are equal to or approximate the quantities of Eqs. 1-2, ξ can be regarded as sparse, internally generated *noise*. The resulting model will be referred as 'O1' (1^{st} online KF model). To simplify the complexity of the iteration, we may suppose that the system is near optimal: $\mathbf{K} \approx \mathbf{H}^{-1}$. Now, the updates for $\mathbf{W}$ and θ (model 'O2') become:

$$W_{ik}^{t+1} \approx \sum_j F_{ij} W_{jk}^t \xi_k - \theta_i^t W_{ik}^t \xi_k + \delta_{ik} \epsilon_k^t \tag{7}$$

$$\theta_k^{t+1} \approx \theta_k^t + \alpha W_{kk} \epsilon_k^t \tag{8}$$

Noticing that only W_{kk}s play a role in tuning θ (Eq. 8), we may neglect the off-diagonal elements of matrix $\mathbf{W}$ in Eq. 7 (model 'O3'):

$$W_{ii}^{t+1} \approx F_{ii} W_{ii}^t \xi_i - \theta_i^t W_{ii}^t \xi_i + \epsilon_i^t \tag{9}$$

$$\theta_k^{t+1} \approx \theta_k^t + \alpha W_{kk} \epsilon_k^t \tag{10}$$

An important consequence is that $\mathbf{F}$ is involved in the tuning of $\mathbf{W}$ only through its diagonal elements. Further simplification neglects the self-excitatory contribution, $F_{ii} W_{ii}^t \xi_i$. The resulting approximation (Eq. 11-12, model 'O4') and its stabilized form (Eq. 13-14, model 'O5') are as follows:

$$W_{ii}^{t+1} \approx -\theta_i^t W_{ii}^t \xi_i + \epsilon_i^t \tag{11}$$

$$\theta_k^{t+1} \approx \theta_k^t + \alpha W_{kk} \epsilon_k^t \tag{12}$$

$$W_{ii}^{t+1} \approx W_{ii}^t + \gamma\{-\theta_i^t W_{ii}^t \xi_i + \epsilon_i^t\} \tag{13}$$

$$\theta_k^{t+1} \approx \theta_k^t + \alpha W_{kk} \epsilon_k^t \tag{14}$$

Model 'O5', that is Eqs. 13-14, resembles the more conventional forms used in artificial neural network literature.

The scheme also allows to introduce an additional matrix ($\mathbf{N}$) to change the graphical representation (resulting in a more 'convenient' mapping (see later):

$$\hat{x}_m^{t+1} = \sum_j F_{mj} \hat{x}_j^t + \sum_i N_{mi}\, \theta_i^t \sum_l K_{il} e_l^t \tag{15}$$

Approximations 'O1'-'O5' remain valid for this *extended form*.

Simulations. Figure 1 depicts the transients of the prediction and the reconstruction errors for a dynamical system that consists of two-dimensional rotational matrices ($\mathbf{F}, \mathbf{H}$) and $\mathbf{K} \approx \mathbf{H}^{-1}$. The left hand side of both subfigures show the transients when the system is turned on. The right hand side depicts

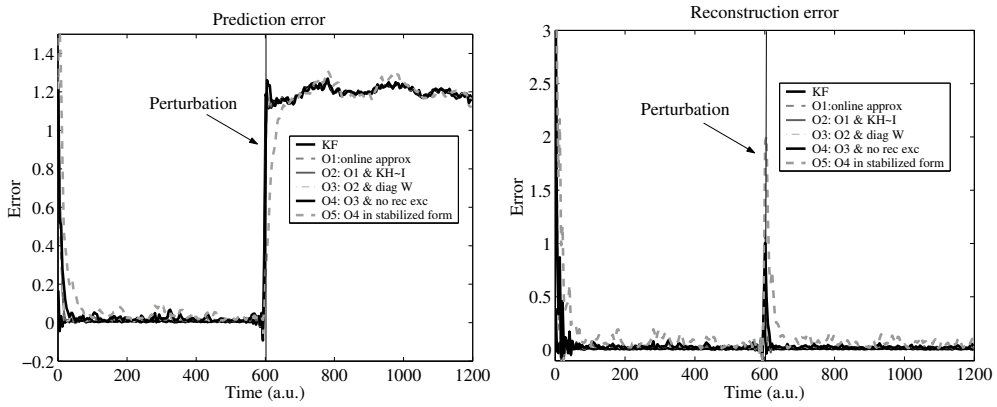

Figure 1. **Comparison of direct and approximated Kalman filters.** 'Prediction error': $\|\mathbf{x}-\hat{\mathbf{x}}\|$. 'Reconstruction error': $\|\mathbf{y}-\mathbf{H}\hat{\mathbf{x}}\|$. 'KF' and 'O1-O5' stand for the optimal direct KF method and for the approximations, respectively. At time $t = 600$ the observation matrix $\mathbf{H}$ was changed. Negative values around the sudden change are artifacts of the Chebyshev noise filter used to improve visualization. Dynamical system is made of rotational matrices (see text). The rotational angles: $\alpha_F = 10°$, $\alpha_H = 50°$ before, and $\alpha_{H2} = 20°$ after perturbation. Signal to noise ratios: hidden process: $59dB$, observation process: $51dB$. $\alpha = 0.01$ for all simulations.

the system's response to a sudden perturbation to the the observation matrix, $\mathbf{H}$. According to Fig. 1, online models converge to the optimal solution and are better suited for sudden changes than KFR, limited by Gaussian noise assumption. Convergence properties have been studied in the entire angle space (α_F and $(\alpha_H + \alpha_K)$). We found that for our toy problem, the simplest model ('O5') is almost as good as the full online KF approximation ('O1').

3 Mapping onto the hippocampal-entorhinal loop

Motivated by [4,5] we map the KF model onto the EC-HC loop (Fig. 2). The mapping is an extension of the memory model described in [5,6], which stated that the goal of internal representation is to reconstruct the input and to minimize the reconstruction error. The necessity of optimal information transfer then follows and is enabled by bottom-up signal separation (Independent Component Analysis, ICA [7] and references therein). The model assumed additional noise filtering and pattern completion mechanisms. [2] The signal separation step (ICA) processes the reconstruction error and consists of two connection systems, whitening and separation. According to the mapping, CA1 holds the separated error signal and the internal representation is maintained by persistent activity observed in EC V-VI [8] as predicted by [5,6].

[2] Note that the assumption of independent component transformation of $\hat{\mathbf{x}}$ is *exactly* what we made when dropping the off-diagonal elements of $\mathbf{W}$ in Eqs. 9-10.

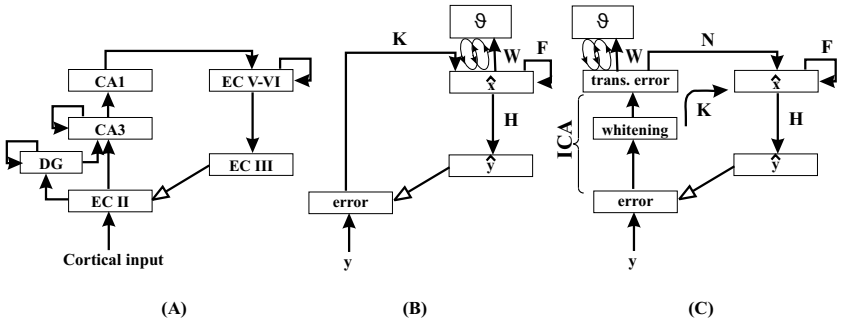

Figure 2. **(A)** Relevant anatomical connections of the HC and its environment (without the EC III→CA1 'direct pathway'). The dentate gyrus (DG) (not modelled here, but see [5]), the CA3 and CA1 subfields are three major regions of the hippocampus (HC). EC II and III and EC V-VI denote the superficial and deep layers of the entorhinal cortex, respectively. Solid (empty) arrows denote excitatory (inhibitory) connections. Local circuits can be found in every layer with different modulatory effects. **(B)** Minimal structure of the KFR model. KF is modulated at the internal representation, i.e., at EC V-VI. **(C)** Extended KFR model and its mapping onto the EC-HC loop. Transformations are realized in connection systems. Vectors represent neural activity at different layers. ICA is performed by CA3 (whitening) and CA1 (separation). KF is modulated at the output of CA1.

The θ-modulated output of CA1 corrects the internal representation, which, in turn, corrects the reconstruction of the input at EC layers II-III. The reconstructed input is compared to the original input and the iteration goes on.

It then follows that functional 'micro-circuits' representing θ may modify (i) the CA1 output *and/or* (ii) the synaptic inputs of EC V-VI. Case (i), which corresponds to Eq. 15, is detailed below. This model implies that *noise* is also required to originate at or flow through CA1 to adapt the diagonal components of auxiliary matrix $\mathbf{W}$. In turn, ICA transformed error and the intrinsic noise appear within the same subset of cells of CA1, but interference can be avoided only if these two signals work in *different* phases. During *error signal propagation* phase, the reconstruction error is transformed (at CA1) and the transformed signal modifies (a) the internal representation, (b) the θ-circuits, and (c) matrix $\mathbf{W}$. During the *noise phase*, error propagation stops and sparse random noise is generated in CA3 and/or in CA1 to optimize matrix $\mathbf{W}$.

4 Conclusions

We have extended the functional model of Lőrincz and Buzsáki [5,6] on the memory organization in the EC-HC loop. One of the verified predictions of this model is that specific sustaining mechanisms [8] should exist at the model layer – the deep layers of the EC – to ensure temporal integration of the

reconstruction error. However, for high noise, minimizing the reconstruction error itself is not sufficient to estimate the external world (i.e. to form optimal internal representation): the reconstruction error needs optimal attenuation with respect to the predictions of the internal model. We proposed a straightforward mechanism for such optimization. Interestingly, the constraints of the *original* model on signal separation have simplified the RPE equations considerably and provided a plausible mapping onto the EC-HC loop. In accord with [9], our model predicts the existence of functional micro-circuits in CA1 and/or EC deep layers, in which individual neurons may take part in several functions (e.g., 'θ-circuits' and noise propagation). In summary, we conjecture that (i) there exists a *predictive* modelling system in the deep layers of the EC, (ii) there should be specific local circuits either in CA1, or, possibly, in the EC deep layers with at least two functionally separable cell types and (iii) strict timing is needed to prevent interfering propagation of error and noise that have different functional roles. However, the possible frequency range of these processes is still undefined.

References

[1] L. Ljung, T. Soderstrom, Theory and practice of recursive identification, MIT Press, Cambridge, MA, 1993.

[2] R. E. Kalman, A new approach to linear filtering and prediction problems, Trans. ASME–Journal of Basic Engin. 82 (Ser. D) (1960) 35–45.

[3] P. Whittle, Optimization over time, Vol. I. of Dynamic Programming and stochastic control, Wiley, Chichester, 1982.

[4] R. P. N. Rao, D. H. Ballard, Dynamic model of visual recognition predicts neural response properties in the visual cortex, Neural Comp. 9 (1997) 721–763.

[5] A. Lőrincz, G. Buzsáki, Two–phase computational model training long–term memories in the entorhinal–hippocampal region, Vol. 911, NYAS, New York, 2000, pp. 83–111.

[6] A. Lőrincz, B. Szatmáry, G. Szirtes, Mystery of structure and function of sensory processing areas of the neocortex: A resolution, J. Comp. Neurosci. 13 (2002) 187–205.

[7] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley Interscience, 2001.

[8] A. V. Egorov, B. N. Hamam, E. Fransén, M. E. Hasselmo, A. A. Alonso, Graded persistent activity in entorhinal cortex neurons, Nature 420 (2002) 173–178.

[9] A. Gupta, Y. Wang, H. Markram, Organizing principles for a diversity of GABAergic interneurons and synapses in the neocortex, Science 287 (2000) 273–278.