A Neural Implementation of the Kalman Filter

Robert C. Wilson
Department of Bioengineering
University of Pennsylvania
Philadelphia, PA 19103
rcwilson@seas.upenn.edu

Leif H. Finkel
Department of Bioengineering
University of Pennsylvania
Philadelphia, PA 19103

Abstract

There is a growing body of experimental evidence to suggest that the brain is capable of approximating optimal Bayesian inference in the face of noisy input stimuli. Despite this progress, the neural underpinnings of this computation are still poorly understood. In this paper we focus on the problem of Bayesian filtering of stochastic time series. In particular we introduce a novel neural network, derived from a line attractor architecture, whose dynamics map directly onto those of the Kalman Filter in the limit where the prediction error is small. When the prediction error is large we show that the network responds robustly to change-points in a way that is qualitatively compatible with the optimal Bayesian model. The model suggests ways in which probability distributions are encoded in the brain and makes a number of testable experimental predictions.

1 Introduction

There is a growing body of experimental evidence consistent with the idea that animals are somehow able to represent, manipulate and, ultimately, make decisions based on, probability distributions. While still unproven, this idea has obvious appeal to theorists as a principled way in which to understand neural computation. A key question, then, is how such Bayesian computations could be performed by neural networks. Several authors have proposed models addressing aspects of this issue [17, 11, 10, 21, 2, 3, 18, 4, 13, 20, 19, 8, 7, 9], but as yet, conclusive experimental evidence in favour of any one of them has been fairly thin on the ground and the question remains open.

In this paper we focus on the problem of tracking a randomly moving, one dimensional stimulus in a noisy environment. In particular, we develop a neural network whose dynamics can be shown to approximate those of a one-dimensional Kalman filter - which is the Bayesian model when all of the distributions are Gaussian. Where the approximation breaks down, for large prediction errors, the network performs something akin to outlier and / or change detection and this 'failure' suggests ways in which the network can be extended to deal with more complex, non-Gaussian distributions over multiple dimensions.

Our approach rests on the modification of the line attractor network of Zhang [28]. In particular, we make three changes to Zhang's network, modifying the activation rule, the weights and the inputs in such a way that the network's dynamics map exactly onto those of the Kalman filter when the prediction error is small. Crucially, these modifications result in a network that is no longer a line attractor and thus no longer suffers from the limitations of these networks.

2 The model

Our network is a modification of Zhang's line attractor network for head direction cells [28]. In particular we update the membrane potential, $\mathbf{u}(t)$, of the network according to the update equation

$$\mathbf{u}(t+1) = w\mathbf{J}\mathbf{f}\left[\mathbf{u}(t)\right] + \mathbf{I}(t) \tag{1}$$

where w is a scalar that scales the strength of the weights, $\mathbf{J}$ is the connectivity matrix, $\mathbf{f}[\cdot]$ is the activation rule and $\mathbf{I}(t)$ is the input at time t.

Like Zhang we choose $\mathbf{J}$ such that the the connection matrix is made up of a mixture of symmetric and asymmetric components, i.e.

$$\mathbf{J} = \mathbf{J}^{sym} + \gamma\mathbf{J}^{asym} \tag{2}$$

for some constant γ. We use the following form for $\mathbf{J}^{sym}$ and $\mathbf{J}^{asym}$ (which is just the derivative of $\mathbf{J}^{sym}$ with respect to i) although it is important to note that the results presented here do not depend strongly on the exact form of $\mathbf{J}^{sym}$ and $\mathbf{J}^{asym}$.

$$J^{sym}_{ij} = K_w \exp\left[\frac{\cos\left(\frac{2\pi(i-j)}{N}\right) - 1}{\sigma_w^2}\right] - c \quad ; \quad J^{asym}_{ij} = -\frac{2\pi}{N\sigma_w^2}\sin\left(\frac{2\pi(i-j)}{N}\right) J^{sym}_{ij} \tag{3}$$

For constant σ_w, K_w and c that determine the width and excitatory and inhibitory connection strengths respectively. N is the number of neurons in the network.

The form of the activation function $\mathbf{f}[\cdot]$ turns out to be crucial if the network is to implement a line attractor. In particular we will show that the following form allows the network to implement a Kalman filter.

$$\mathbf{f}[\mathbf{u}] = \frac{[\mathbf{u}]_+}{S + \mu\sum_i [u_i]_+} \tag{4}$$

where $[\mathbf{u}]_+$ denotes recitification of $\mathbf{u}$; and S and μ are parameters. Note that this activation rule is a functional, depending on the entire profile of u and implements divisive normalization [15, 14].

When $w = 1$, $\mathbf{I}(t) = 0$ and $\gamma = 0$, over a wide range of K_w, σ_w and c, this network is a line attractor network having (in the limit as $N \to \infty$) a continuum of fixed points. Each fixed point takes the form of a bump of activity, $\mathbf{U}(x)$, centered at location x in the network. In general, we are not concerned with the exact form of the fixed point activity profile, but only that it is a smooth bump centered at some location and that it satisfies

$$\mathbf{U}(x) = \mathbf{J}\mathbf{f}\left[\mathbf{U}(x)\right] \tag{5}$$

3 Relation to Kalman filter - small prediction error case

To understand the effect of external input on the network and, ultimately, relate the network dynamics to those of the Kalman filter, we limit our discussion to inputs of the form

$$\mathbf{I}(t) = A(t)\mathbf{U}(z(t)) \tag{6}$$

i.e. inputs which are scaled versions of the fixed point current profile centered at $z(t)$. We then propose an *ansatz* of the form

$$\mathbf{u}(t) = \alpha(t)\mathbf{U}(\hat{x}(t)) \tag{7}$$

Substituting this into the right hand side of equation 1 gives

$$\begin{aligned} RHS &= w\left[\mathbf{J}^{sym} + \gamma\mathbf{J}^{asym}\right]\mathbf{f}\left[\alpha(t)\mathbf{U}\left(\hat{x}(t)\right)\right] + A(t+1)\mathbf{U}\left(z(t+1)\right) \\ &= w\frac{(S_0 + \mu_0\mathcal{I})\alpha(t)}{S + \mu\mathcal{I}\alpha(t)}\left[\mathbf{U}\left(\hat{x}(t)\right) + \gamma\mathbf{U}'\left(\hat{x}(t)\right)\right] + A(t+1)\mathbf{U}\left(z(t+1)\right) \end{aligned} \tag{8}$$

where S_0 and μ_0 are the settings of S and μ used to get the fixed point activity profile, $\mathbf{U}(z)$. $\mathbf{U}'(x)$ is the spatial derivative of the fixed point profile centered at x and we have introduced the variable $\mathcal{I}$ which is given by

$$\mathcal{I} = \sum_i \left[U_i(\hat{x}(t))\right]_+ \tag{9}$$

If we also introduce the variable C as

$$C = \frac{w(S_0 + \mu_0\mathcal{I})\alpha(t)}{S + \mu\mathcal{I}\alpha(t)} \tag{10}$$

and assume that γ is small then we can write the right hand side as

$$RHS \approx C\mathbf{U}\left(\hat{x}(t) + \gamma\right) + A(t+1)\mathbf{U}\left(z(t+1)\right) \tag{11}$$

If we write

$$\bar{x}(t+1) = \hat{x}(t) + \gamma \tag{12}$$

and suppose that the difference $\delta = z(t+1) - \bar{x}(t+1)$ is small, then we can write

$$RHS \approx \left[A(t+1) + C\right]\mathbf{U}\left(\bar{x}(t+1) + \frac{A(t+1)}{A(t+1)+C}\left[z(t+1) - \bar{x}(t+1)\right]\right) \tag{13}$$

Equating the two sides of 1 and rearranging into a more revealing form gives

$$\alpha(t+1) \approx \frac{1}{\frac{S}{w(S_0+\mu_0\mathcal{I})}\frac{1}{\alpha(t)} + \frac{\mu\mathcal{I}}{w(S_0+\mu_0\mathcal{I})}} + A(t+1) \tag{14}$$

and

$$\hat{x}(t+1) = \bar{x}(t) + \frac{A(t+1)}{\alpha(t+1)}\left[z(t+1) - x(t+1)\right] \tag{15}$$

Now, the update equations for the Kalman filter take the form

$$\bar{x}(t+1) = \hat{x}(t) + v(t) \tag{16}$$

$$\hat{x}(t+1) = \bar{x}(t+1) + \frac{\hat{\sigma}_x(t+1)^2}{\sigma_z(t+1)^2}\left[z(t+1) - \bar{x}(t+1)\right] \tag{17}$$

$$\frac{1}{\hat{\sigma}_x(t+1)^2} = \frac{1}{\hat{\sigma}_x(t)^2 + \sigma_v(t)^2} + \frac{1}{\sigma_z(t+1)^2} \tag{18}$$

where $\hat{\sigma}_x(t)$ is the model's uncertainty in the estimate of the mean at time t, $\sigma_z(t)$ is the uncertainty in the observed position, $v(t)$ is the velocity signal and $\sigma_v(t)$ is the uncertainty in the velocity signal.

Comparing these expressions with equations 12, 14 and 15 suggests that if we require

$$\gamma = v(t) \tag{19}$$

and define w such that

$$w = \frac{S}{S_0 + \mu_0\mathcal{I}} \tag{20}$$

then we arrive at the following update equation for $\alpha(t)$

$$\alpha(t+1) = \frac{1}{\frac{1}{\alpha(t)} + \frac{\mu\mathcal{I}}{S}} + A(t+1) \tag{21}$$

and we can make the identifications

$$\alpha(t) \propto \frac{1}{\hat{\sigma}_x(t)^2} \tag{22}$$

$$A(t) \propto \frac{1}{\sigma_z(t)^2} \tag{23}$$

and

$$\frac{\mu\mathcal{I}}{S} \propto \sigma_v(t)^2 \tag{24}$$

which maps the network dynamics, when the prediction error is small, directly onto the Kalman filter equations. This is our main result.

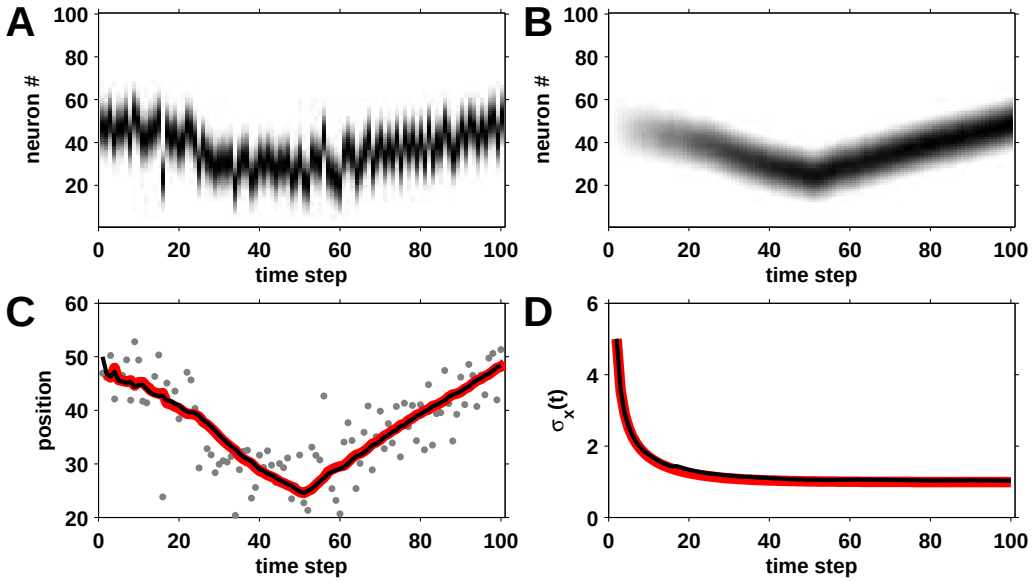

Figure 1: Comparison of noiseless network dynamics with dynamics of the Kalman Filter for small prediction errors.

3.1 Implications

3.1.1 Reciprocal code for uncertainty in input and estimate

First of all, equation 23 suggests that the uncertainty in the data, $\sigma_z(t)$, should be reflected in the strength of the input, with decreasing uncertainty associated with a stronger input. Interestingly, such a scaling of input (*without* a corresponding narrowing of the tuning curves) *is* seen in the brain [22, 5, 2].

Next, equation 22 provides a link between the strength of activity in the network and the overall uncertainty in the optimal estimate of the Kalman filter, $\hat{\sigma}_x(t)$. As the uncertainty decreases the activity increases.

3.1.2 Code for velocity signal

Equation 19 shows how the mean of the velocity signal is encoded into the weights of the network - with the degree of asymmetry in the weights, γ, proportional to the speed. Such hard coding of the velocity signal represents a potential limitation of the model, as we would like to be able to deal with arbitrary, time varying speeds. However, this kind of change could be implemented by pre-synaptic inhibition [26] or by using a 'double-ring'network similar to [28, 27]. Equation 24 suggests that the variance of the velocity signal could be varied by changing μ the amount of divisive inhibition present in the model. Recent modeling work has suggested that such rapid changing of the strength of divisive inhibition is feasible [6].

3.1.3 The network is no longer a line attractor

Finally, the weight scaling factor, w, defined in equation 20 is such that the resulting network is no longer a line attractor. This can be seen by considering the fixed points of equation 1 when the input current, $I(x,t) = 0$ as a function of w.

Let α^* be the fixed value of the scale factor. From equation 21 it is clear that this satisfies

$$\alpha^* = \frac{w\left(S_0 + \mu_0 \mathcal{I}\right)\alpha^*}{S + \mu \mathcal{I} \alpha^*} \tag{25}$$

which has the solutions $\alpha^* = 0$ and

$$\alpha^* = \left(\frac{S_0 + \mu_0 \mathcal{I}}{\mu \mathcal{I}}\right) w - \frac{S}{\mu \mathcal{I}} \tag{26}$$

This second solution turns out to be exactly zero when w satisfies equation 20, hence the network only has one fixed point corresponding to the all zero state and is not a line attractor. This is a key result as it removes all of the constraints required to achieve line attractor dynamics such as infinite precision in the weights and lack of noise in the network and thus the network is much more biologically plausible.

3.2 An example

In figure 1 we demonstrate the ability of the network to approximate the dynamics of the one-dimensional Kalman filter. The input, shown in figure 1A, is a noiseless bump of current centered at position random position $z(t)$, which is given by

$$z(t) = x_{actual}(t) + n_z(t) \tag{27}$$

where $n_z(t)$ Gaussian random noise with standard deviation, $\sigma_z(t) = 5$, and $x_{actual}(t)$ is the 'true' stimulus position determined by a random walk with drift speed $v = 0.5$ for $1 \leq t < 50$ and $v = -0.5$ for $50 \leq t < 100$ and standard deviation $\sigma_v(t) = 0.2$. In accordance with equation 23 the height of each bump is scaled by $1/\sigma_z(t)$.

In figure 1B we plot the output activity of the network over time. Darker shades correspond to higher firing rates. We assume that the network gets the correct velocity signal, i.e. $\gamma = v$ and μ is set such that $\frac{\mu \mathcal{I}}{S} = \sigma_v(t)^2$. The other parameters are set to $K_w = 1$, $\sigma_w = 0.2$, $c = 0.05$ and $S = 1$ which gives $\mathcal{I} = 5.47$. As can be seen from the plot, the amount of activity in the network steadily grows from zero over time to an asymptotic value, corresponding to the network's increasing certainty in its predictions. The position of the bump of activity in the network is also much less jittery than the input bump, reflecting a certain amount of smoothing.

In figure 1C we compare the positions of the input bumps (grey dots) with the position of the network bump (black line) and the output of the equivalent Kalman filter (red line). The network clearly tracks the Kalman filter estimate extremely well. In figure 1D we compare the uncertainty in the estimate of the mean between the equivalent Kalman filter (red line) and the network, computed as $1/\sqrt{\alpha(t)}$ and, again, the two estimates are in close agreement.

4 The effect of input noise

We now consider the effect of noise on the ability of the network to implement Kalman filter dynamics. In particular we consider noise in the input signal, which for this simple one layer network is equivalent to having noise in the update equation. For brevity, in this section we only present the main results along with the results of simulations, leaving the more detailed analysis to the supplementary information.

Specifically, we consider input signals where the only source of noise is in the input current i.e. there is no additional jitter in the position of the bump as there was in the noiseless case, thus we write

$$\mathbf{I}(t) = A(t)\mathbf{U}(x(t)) + \epsilon(t) \tag{28}$$

where $\epsilon(t)$ is some noise vector. The main effect of the noise is that it perturbs the effective position of the input bump this can be modeled by extracting the maximum likelihood estimate of the input position given the noisy input and then using. This position as the input to the equivalent Kalman filter. Because of the noise, this extracted position is not, in general, the same as the noiseless input position and for zero mean Gaussian noise with covariance Σ, the variance of the perturbation, σ_z is

$$\sigma_z = \frac{1}{A}\sqrt{\frac{2}{\mathbf{U}'^T \Sigma^{-1} \mathbf{U}'}} \tag{29}$$

Now, for the network to approximate a Kalman filter, equation 23 must hold which means that we require the magnitude of the covariance matrix to scale in proportion to the strength of the input

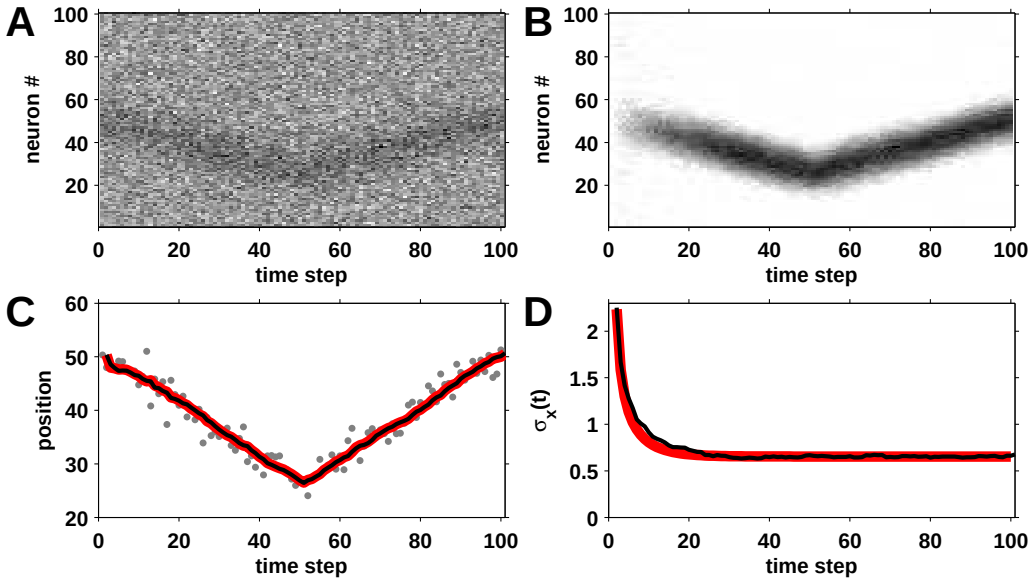

Figure 2: Response of the network when presented with a noisy moving bump input.

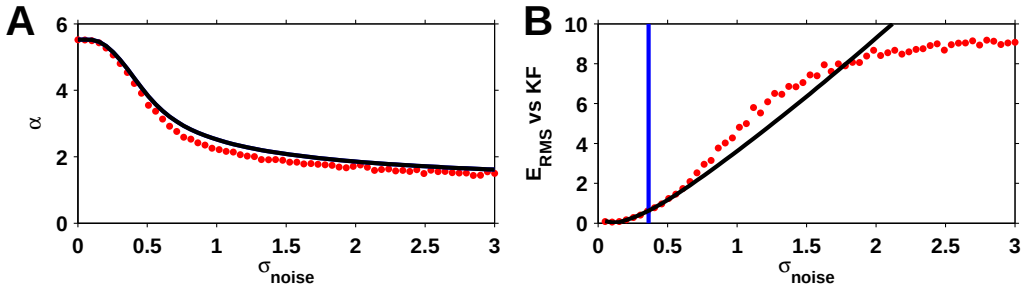

Figure 3: Effect of noise magnitude on performance of network.

signal, A, i.e. $\Sigma \propto A$. Interestingly this relation is true for Poisson noise, the type of noise that is found all over the brain.

In figure 2 we demonstrate the ability of the network to approximate the dynamics of a Kalman filter. In panel A we show the input current which is a moving bump of activity corrupted by independent Gaussian noise of standard deviation $\sigma_{noise} = 0.23$, or about two thirds of the maximum bump height of the noise free bump. Clearly this is a high noise setting and it is hard to see the bump location by eye.

In figure 2B we show the output activity of the network. Clearly it has cleaned up the signal dramatically and the bump of activity (although still noisy) reflects the position of the underlying stimulus much more faithfully than the input. (Note that the colour scales in A and B are different).

In panel C we compare the position of the output bump in the network (black line) with that of the equivalent Kalman filter. To do this we first fit the noisy input bump at each time to obtain input positions $z(t)$ shown as grey dots. Then using $\sigma_z = 2.23$ computed using equation 29 we make predictions of the equivalent Kalman filter (thick red line). As before, the network closely approximates the dynamics of the equivalent Kalman filter. Finally in panel D we plot the Kalman filter's (red line) and network's (black line) estimate of the error in the mean, $\sigma_x(t)$ and as before these are in good agreement.

4.1 Performance of the network as a function of noise magnitude

The noise not only effects the position of the input bump but also, in a slightly more subtle manner, causes a gradual decline in the ability of the network to emulate a Kalman filter. The reason for this (outlined in more detail in the supplementary material) is that the output bump scale factor, α, decreases as a function of the noise variance, σ_{noise}. This effect is illustrated in figure 3A where we plot the steady state value of α (for constant input strength, $A(t)$) as a function of σ_{noise}. The average results of simulations on 100 neurons are shown as the red dots, while the black line represents the results of the theory from the supplementary material.

The reason for the decline in α as σ_{noise} goes up is that, because of the rectifying non-linearity in the activation rule, increasing σ_{noise} increases the amount of noisy activity in the network. Because of inhibition (both divisive and subtractive) in the network, this 'noisy activity' competes with the bump activity and decreases it - thus reducing α.

This decrease in α, by equation 15, results in a change in the Kalman gain of the network, making it different from that of the equivalent Kalman filter thus degrading the network's performance. We quantify this difference in figure 3B where we plot the root mean squared error (in units of neural position) between the network and the equivalent Kalman filter as a function of σ_{noise}. As before, the results of simulations are shown as red dots and the theory (outlined in the supplementary material) is the black line. To give some sense for the scale on this plot, the horizontal blue line corresponds to the maximum height of the (noise free) input bump. Thus we may conclude that the performance of the network and the theory are robust up to fairly large values of σ_{noise}.

5 Response to change-points (and outliers) - large prediction error case

We now consider the dynamics of the network when the prediction error is large. By large we mean that the difference between the prediction $\bar{x}(t)$ of the network and the input $z(t)$ is greater than the width of the bump of activity in the network. In terms of tracking, such an input could be caused by an outlier or a change-point - i.e. sustained large and abrupt changes in the input position at random times. In the interests of space we focus only on the case of change-points. Such an input is shown in figure 4A where the position of the input bump undergoes a change-point at $t = 50$.

In figure 4B we show the output of the network when presented with this stimulus. As before, prior to the change-point, the network produces a single bump of activity whose position approximates that of a Kalman filter. However, after the change-point, the network maintains two bumps of activity for a significant period of time. One, that decreases in magnitude over time, at the original position - essentially predicting where the input would be if the change had not occurred. And a second bump, that grows over time, at the location of the input after the change-point. Thus in the period immediately after the change-point, the network can be thought of as maintaining two separate and competing hypotheses about the position of the stimulus - one corresponding to the case where no change has occurred, and the other, the case where a change occurred at $t = 50$.

In figure 4C we compare the position of the bump(s) of activity in the network (black dots whose size reflects the size of each bump) to the output from the equivalent Kalman filter (thick red line). Before the change-point, the two agree well, but after the change, the Kalman filter becomes suboptimal, taking a long time to move to the new position. The network, however, by maintaining two hypotheses reacts much better. Finally, in figure 4D we plot the scale factor, $\alpha_i(t)$, of each bump as computed from the simulations (black dots) and from the approximate analytic solution described in the supplementary material (red line for scale factor of bump at 30, blue line for that of bump at 80). As can be seen, there is good agreement between theory and simulation - with the largest discrepancy occurring for small values of the scale factor.

Thus the network no longer approximates a Kalman filter when confronted with a change-point and instead maintains two competing hypotheses in a way that is qualitatively similar to that of the run-length distribution in [1]. This is an extremely interesting result and hints at ways in which more complex distributions may be encoded in these type of networks.

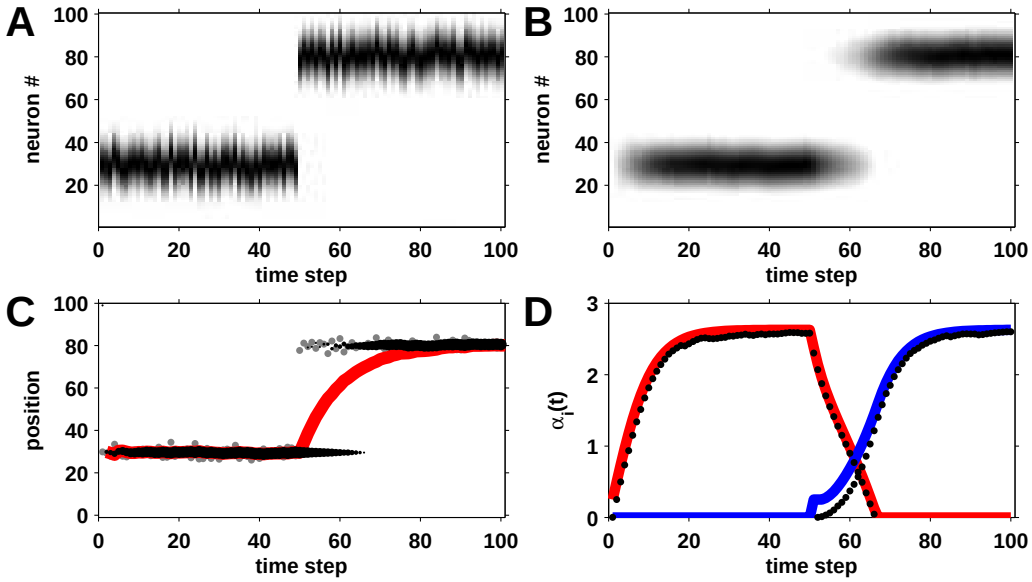

Figure 4: Response of the network to change-points.

6 Discussion

6.1 Relation to previous work

Of the papers mentioned in the introduction, two are of particular relevance to the current work. In the first, [9], the authors considered a neural implementation of the Kalman filter using line attractors. Although this work, at first glance, seems similar to what is presented here, there are several major differences, the main one being that our network is not a line attractor. Also, in [9], the input gain is changed manually and the form of non-linearity is also different.

Probabilistic population coding [18, 4] is a more closely related model to the one presented here. Combined with divisive normalization, these networks can implement a Kalman filter exactly. The main difference is that in the current work the network 'only' computes an approximation to the Kalman filter. While this may seem like a limitation of our network, we see it as an advantage as the breakdown of the approximation leads to a more robust response to outliers and change-points than a pure Kalman filter.

6.2 Extension beyond one-dimensional Gaussians

A major limitation of the current model is that it only applies to one dimensional tracking - clearly an unreasonable restriction for the brain. One possible way around these problems is hinted at by the response of the network in the change-point case where we saw two, largely independent bumps of activity in the network. This facility to encode multiple 'particles' in the network may allow networks of this kind to approximate the dynamics of the ensemble Kalman filter [16], an extension of the Kalman filter that uses several, coupled Kalman filters to approximate the inference process for non-linear and non-Gaussian systems. Such a possibility is an intriguing idea for future work.

6.3 Experimental predictions

The model makes at least two easily testable predictions about the response of head direction cells [23, 24, 25] in rats. The first comes by considering the response of the neurons in the 'dark'. Assuming that all bearing cues can indeed be eliminated, by setting $A(t) = 0$ in equation 14, we expect the activity of the neurons to fall off as $1/t$ and that the shape of the tuning curves will remain approximately constant. Note that this prediction is vastly different from the behaviour of a line attractor, where we would not expect the level of activity to fall off at all in the dark.

Another, slightly more ambitious experiment would involve perturbing the reliability of one of the landmark cues. In particular, one could imagine a training phase, where the position of one landmark is jittered over time, such that each time the rat encounters it it is at a slightly different heading. In the test case, all other, reliable, landmark cues would be removed and the response of head direction cells measured in response to presentation of the unreliable cue alone. The prediction of the model is that this would reduce the strength of the input, A, which in turn reduces the level of activity in the head direction cells, α. In particular, if σ_z is the jitter of the unreliable landmark, then we expect α to scale as $1/\sigma_z^2$. This prediction is very different from that if a line attractor which would predict a constant level of activity regardless of the reliability of the landmark cues.

7 Conclusions

In this paper we have introduced a novel neural network model whose dynamics map directly onto those of a one-dimensional Kalman filter when the prediction error is small. This property is robust to noise and when the prediction error is large, such as for change-points, the output of the network diverges from that of the Kalman filter, but in a way that is both interesting and useful. Finally, the model makes two easily testable experimental predictions about head direction cells.

References

[1] R.P. Adams and D.J.C. MacKay. Bayesian online changepoint detection. Technical report, University of Cambridge, Cambridge, UK, 2007.

[2] J. S. Anderson, I. Lampl, D. C. Gillespie, and D. Ferster. The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. *Science*, 290:1968–1972, 2000.

[3] M. J. Barber, J. W. Clark, and C. H. Anderson. Neural representation of probabilistic information. *Neural Computation*, 15:1843–1864, 2003.

[4] J. Beck, W. J. Ma, P. E. Latham, and A. Pouget. Probabilistic population codes and the exponential family of distributions. *Progress in Brain Research*, 165:509–519, 2007.

[5] K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon. Response of neurons in macaque mt to stochastic motion signals. *Visual Neuroscience*, 10(1157-1169), 1993.

[6] F. S. Chance, L. F. Abbott, and A. D. Reyes. Gain modulation from background synaptic input. *Neuron*, 35:773–782, 2002.

[7] S. Deneve. Bayesian spiking neurons i: Inference. *Neural Computation*, 20:91–117, 2008.

[8] S. Deneve. Bayesian spiking neurons ii: Learning. *Neural Computation*, 20:118–145, 2008.

[9] S. Deneve, J.-R. Duhammel, and A. Pouget. Optimal sensorimotor integration in recurrent cortical networks: a neural implementation of kalman filters. *Journal of Neuroscience*, 27(21):5744–5756, 2007.

[10] S. Deneve, P. E. Latham, and A. Pouget. Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2(8):740–745, 1999.

[11] S. Deneve, P. E. Latham, and A. Pouget. Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, 4(8):826–831, 2001.

[12] G. Evensen. The ensemble kalman filter: theoretical formulation and practical implementation.

[13] J. I. Gold and M. N. Shadlen. Representation of a perceptual decision in developing oculomotor commands. *Nature*, 404(390-394), 2000.

[14] D. J. Heeger. Modeling simple cell direction selectivity with normalized half-squared, linear operators. *Journal of Neurophysiology*, 70:1885–1897, 1993.

[15] D. J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–198, 1993.

[16] S. J. Julier and J. K. Uhlmann. A new extension of the kalman filter to nonlinear systems. 1997.

[17] P. E. Latham, S. Deneve, and A. Pouget. Optimal computation with attractor networks. *Journal of Physiology Paris*, 97(683-694), 2003.

[18] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, 2006.

[19] R. P. N. Rao. Bayesian computation in recurrent neural circuits. *Neural Computation*, 16:1–38, 2004.

[20] R. P. N. Rao. Hierarchical bayesian inference in networks of spiking neurons. In *Advances in Neural Information Processing Systems*, volume 17, 2005.

[21] M. Sahani and P. Dayan. Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Computation*, 15:2255–2279, 2003.

[22] G. Sclar and R. D. Freeman. Orientation selectivity in the cat's striate cortex is invariant with stimulus contrast. *Experimental Brain Research*, 46:457–461, 1982.

[23] J. S. Taube, R. U. Muller, and J. B. Ranck. Head-direction cells recorded from postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435, 1990.

[24] J. S. Taube, R. U. Muller, and J. B. Ranck. Head-direction cells recorded from postsubiculum in freely moving rats. ii. effects of environmental manipulations. *Journal of Neuroscience*, 10(2):436–447, 1990.

[25] S. I. Wiener and J. S. Taube. *Head direction cells and the neural mechanisms of spatial orientation*. MIT Press, 2005.

[26] L.-G. Wu and P. Saggau. Presynaptic inhibition of elicited neurotransmitter release. *Trends in Neuroscience*, 20:204–212, 1997.

[27] X. Xie, R. H. Hahnloser, and H. S. Seung. Double-ring network model of the head-direction system. *Physical Review E*, 66:0419021–0419029, 2002.

[28] K. Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.