Identifiability of Gaussian Bayesian bandit models

Maarten Speekenbrink (m.speekenbrink@ucl.ac.uk)
UCL Experimental Psychology, 26 Bedford Way
London WC1H 0AP, England UK

Abstract

The Kalman filter, combined with heuristic choice rules such as softmax, UCB, and Thompson sampling, has been a popular model to identify the role of uncertainty in exploration in human reinforcement learning. Here we show that the Kalman filter combined with a softmax or UCB choice rule is not fully identifiable. By this structural identifiability, we mean that with unlimited data, the true parameter values are determinable. Perhaps surprisingly, the Kalman filter with Thompson sampling is fully identifiable.

Keywords: Identifiability; Kalman filter; Softmax; UCB; Thompson sampling; Multi-Armed Bandits

Introduction

There has been much interest in identifying the role of uncertainty in exploration in human reinforcement learning (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Gershman, 2018; Knox, Otto, Stone, & Love, 2012; Speekenbrink & Konstantinidis, 2015; Wilson, Geana, White, Ludvig, & Cohen, 2014; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). Restless multi-armed bandit tasks are a useful paradigm to empirically investigate this (Daw et al., 2006; Speekenbrink & Konstantinidis, 2015), as they require continued exploration long after all options have been tried initially. A prominent learning model for human behaviour in such tasks is the Kalman filter (Daw et al., 2006; Gershman, 2015; Speekenbrink & Konstantinidis, 2015). The Kalman filter provides a principled and computationally efficient way to track both estimated value and the uncertainty in these estimates. Combining the Kalman filter with heuristic choice rules which aim to balance exploration and exploitation, such as the softmax, upper-confidence bound (UCB), or Thompson sampling, offers a powerful and flexible computational framework to assess the role of uncertainty in human reinforcement learning. When the Kalman filter is an adequate descriptive model of how agents learn (expected) rewards, estimating the relevant parameters of the Kalman filter provides a window into their inductive biases, such as how variable they believe rewards are, and how changeable the environment is over time.

This paper addresses to what extent the parameters of models combining the Kalman filter with heuristic choice rules are structurally identifiable, in the sense that with unlimited data, we would be able to determine the true value of their parameters. For identifiable models, parameter estimates can be appropriately compared between or within people and related to neural functioning. While the Kalman filter combined with a softmax or UCB rule is not fully identifiable, perhaps surprisingly, the Kalman filter with Thompson sampling is.

Gaussian restless bandits

We will focus on a simple and canonical version of a restless bandit, where we assume that rewards $R_{t,i}$ at time t for bandit i are continuous and normally distributed around the average reward $\mu_{t,i}$ for that option at that time, while the average rewards vary over time according to a simple random walk:

$$R_{t,i} = \mu_{t,i} + \varepsilon_{t,i} \qquad \varepsilon_{t,i} \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{1}$$

$$\mu_{t,i} = \mu_{t-1,i} + \xi_{t,i} \qquad \xi_{t,i} \sim \mathcal{N}(0, \sigma_\xi^2) \tag{2}$$

where $\mathcal{N}(m, v)$ denotes a Gaussian (normal) distribution with mean m and variance v. We refer to σ_ξ^2 as the *innovation variance* and σ_ε^2 as the *noise variance*.

Although we focus on an environment like above, for which the Kalman filter is optimal, the results hold for any task in which a Kalman filter is assumed to be an (approximate) learning model for average rewards.

Bayesian learning and decision models

Kalman filter

The Kalman filter (Kalman, 1960; Kalman & Bucy, 1961) is an efficient and algorithm to compute the posterior distributions for $\mu_{t,i}$ for linear Gaussian dynamical systems such as that defined by Equations 1 and 2. Assuming that the innovation and error variances are known, and assuming a Gaussian prior for the initial mean: $\mu_{0,i} \sim \mathcal{N}(m_0, v_0)$, the posterior distributions are all Gaussian:

$$p(\mu_{t,i}|C_{0:t}, R_{0:t}) = \mathcal{N}(m_{t,i}, v_{t,i}) \tag{3}$$

where $C_{0:t} = (C_1, \ldots, C_t)$ and we have taken the liberty to define $C_0 = \varnothing$ and apply the same notation for $R_{0:t}$.

The Kalman filter provides an efficient way to sequentially calculate the mean $m_{t,j}$ and variance $v_{t,j}$ of these posterior distributions. The Kalman filter update equations are:

$$m_{t,i} = m_{t-1,i} + k_{t,i}(R_t - m_{t-1,i}) \tag{4}$$

and

$$v_{t,i} = (1 - k_{t,i})(v_{t-1,i} + \sigma_\xi^2) \tag{5}$$

with the *Kalman gain*:

$$k_{t,i} = \begin{cases} \frac{v_{t-1,i} + \sigma_\xi^2}{v_{t-1,i} + \sigma_\xi^2 + \sigma_\varepsilon^2} & \text{if } C_t = i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

At time t, before making a choice, the Bayesian value of each bandit (its average reward) can be derived from the prior predictive distribution

$$p(\mu_{t+1,i}|C_{0:t},R_{0:t}) = \mathcal{N}(m_{t-1,i}, v_{t-1,i}+\sigma_\xi^2)$$

Softmax

The softmax rule can be viewed as a stochastic choice rule in which the probability of choosing a bandit depends solely on the estimated mean rewards $m_{t,i}$. It can be stated as:

$$P(C_t = i|C_{0:t-1},R_{0:(t-1)}) = \frac{\exp(\gamma m_{t,i})}{\sum_{j=1}^N \exp(\gamma m_{t,j})} \tag{7}$$

where the inverse temperature parameter γ allows choices to vary from uniformly random ($\gamma = 0$) to deterministically always choosing the option with the highest estimated mean reward ($\gamma \to \infty$).

Upper confidence bound (UCB)

The upper confidence bound (UCB) strategy can be defined as follows:

$$P(C_t = i|C_{0:(t-1)},R_{0:(t-1)}) = \begin{cases} 1 & \text{if } i = \arg\max_j u_{t,j} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where the upper confidence bound is defined as

$$u_{t,i} = m_{t-1,i} + \beta(\sqrt{v_{t-1,i}+\sigma_\xi^2}), \tag{9}$$

and the parameter β defines the width of the confidence interval, e.g. setting $\beta \approx 1.96$ results in always choosing the bandit with the highest 95% upper confidence interval.

As the UCB rule is deterministic, it is generally not applied in the manner above to human choices. One way to allow for deviations is to use an "epsilon-greedy" style implementation, such that the bandit with the highest UCB is chosen with probability $1-\varepsilon$, while with probability ε, a bandit is chosen uniformly at random. Another – more popular – stochastic version of the UCB rule is to use a softmax version (e.g. Daw et al., 2006; Speekenbrink & Konstantinidis, 2015; Wu et al., 2018):

$$P(C_t = i|C_{0:(t-1)},R_{0:(t-1)}) = \frac{\exp \gamma u_{t,i}}{\sum_{j=1}^N \exp \gamma u_{t,j}} \tag{10}$$

Thompson sampling

Thompson sampling (Thompson, 1933; May, Korda, Lee, & Leslie, 2012), like the UCB rule, depends on both estimated value and the uncertainty in those estimates. In words, it matches the probability of choosing a bandit with the probability that it has the highest expected reward. A Bayesian decision rule, this is based on the prior predictive distributions $p(\mu_{t,j}|C_{0:(t-1)},R_{0:(t-1)})$:

$$P(C_t = i|C_{0:(t-1)},R_{1:(t-1)}) = P(\forall j \neq i : \tilde{m}_{t,i} > \tilde{m}_{t,j}) \tag{11}$$

where

$$\tilde{m}_{t,i} \sim \mathcal{N}(m_{t-1,i}, v_{t-1,i}+\sigma_\xi^2)$$

is a sample from the prior predictive distribution of the mean $\mu_{t,i}$. In contrast to the softmax and UCB rule, there are no further adjustable parameters, it only needs (sensible) values for the environmental parameters m_0, σ_ξ^2, and σ_ε^2.

Model identifiability

Identifiability of a statistical model roughly means that any change in model parameters implies a change in the likelihood. More formally, a model with parameters $\theta \in \Theta$, where Θ is the parameter space, is identifiable when, for (almost) all possible observations $c \in \mathcal{C}$,

$$P(c|\theta) = P(c|\theta') \leftrightarrow \theta = \theta' \tag{12}$$

Identifiability of the KF-SM model

The Kalman filter softmax (KF-SM) model, with $\theta_{\text{sm}} = (\gamma, m_0, v_0, \sigma_\xi^2, \sigma_\varepsilon^2)$, is not identifiable. The problem here is that we can rescale the variance parameters v_0, σ_ξ^2, and σ_ε^2 by a common scaling factor α, such that $v_0' = \alpha v_0$, $\sigma_\xi^{2'} = \alpha\sigma_\xi^2$, and $\sigma_\varepsilon^{2'} = \alpha\sigma_\varepsilon^2$, and get the same likelihood for θ_{sm} and $\theta_{\text{sm}}' = (\gamma, m_0, v_0', \sigma_\xi^{2'}, \sigma_\varepsilon^{2'})$. Firstly, at $t = 1$, it is clear that

$$\frac{v_{0,i}' + \sigma_\xi^{2'}}{v_{0,i}' + \sigma_\xi^{2'} + \sigma_\varepsilon^{2'}} = \frac{\alpha v_{0,i} + \alpha\sigma_\xi^2}{\alpha v_{0,i} + \alpha\sigma_\xi^2 + \alpha\sigma_\varepsilon^2} = \frac{v_{0,i} + \sigma_\xi^2}{v_{0,i} + \sigma_\xi^2 + \sigma_\varepsilon^2} \tag{13}$$

Hence, θ_{sm} and θ_{sm}' lead to identical Kalman gain $k_{1,i}$ for all bandits i. In fact, the Kalman gain is identical at all $t > 1$. For θ_{sm}', the posterior variance is

$$v_{1,i}' = (1-k_{1,i})(v_{0,i}' + \sigma_\xi^{2'}) = (1-k_{1,i})(\alpha v_{0,i} + \alpha\sigma_\xi^{2'}) = \alpha v_{1,i},$$

from which it follows that $v_{t,i}' = \alpha v_{t,i}$ for all $t > 0$. Hence, we can replace $v_{0,i}'$ in Eq 13 by $v_{t,i}'$, which shows that $k_{t,i}$ is identical for θ_{sm} and θ_{sm}' for all $t > 0$.

This means that only the relative values of v_0, σ_ξ^2, and σ_ε^2, are identifiable in the KF-SM model. By fixing one of the variance parameters to an arbitrary value (not equal to 0), the remaining parameters are identifiable.

Identifiability of the KF-UCB model

The Kalman filter UCB model (KF-UCB), with $\theta_{\text{ucb}} = (\gamma, m_0, v_0, \sigma_\xi^2, \sigma_\varepsilon^2)$, is not identifiable. Although the likelihood of this model depends both on the means $m_{t,i}$ and variances $v_{t,i}$, rescaling v_0, σ_ξ^2, and σ_ε^2 by a common factor α, as above, will again provide identical likelihood values. As shown above, the prior predictive variance then becomes $v_{t,i}' + \sigma_\xi^{2'} = \alpha(v_{t,i} + \sigma_\xi^2)$ and setting by $\beta' = \beta/\sqrt{\alpha}$, the likelihood is identical for θ_{ucb} and $\theta_{\text{ucb}}' = (\beta', m_0, v_0', \sigma_\xi^{2'}, \sigma_\varepsilon^{2'})$. The same will hold for the stochastic versions of the KF-UCB model. Again, one of the variance parameters can be fixed to an arbitrary value $\neq 0$, which will result in the other parameters being identifiable.

Identifiability of the KF-TS model

The Kalman filter Thompson sampling (KF-TS) model with $\theta_{\mathsf{ts}} = (m_0, v_0, \sigma^2_\xi, \sigma^2_\varepsilon)$ is identifiable. While scaling the variances as above will provide the same prior predictive means, the prior predictive variances will be affected uniquely, and with that the choice probabilities.

Conclusions

We have shown that only the Kalman filter with Thompson sampling is fully identifiable, while the Kalman filter with softmax or UCB is not. For the latter models, one of the variance parameters needs to be fixed to an arbitrary value. While this often has been done (e.g. Speekenbrink & Konstantinidis, 2015), the reason has not been laid out clearly. While identifying all the parameters may not be a primary concern, in many cases researchers are interested in comparing parameter estimates between people, in correlating these with neural signals. In these cases, it is important to realise what the consequences are of only having access to e.g. relative variances. When the Kalman filter with Thompson sampling is an adequate descriptive model, as all parameters are identifiable, it is possible to compare people according to e.g. the level of their prior uncertainty.

The results about identifiability presented here generalize immediately to the "mean-stable" version of the Kalman filter (where σ^2_ξ is fixed to 0). In this case, the other variances (v_0 and σ^2_ε) are still not identifiable, so one of these needs to be fixed to an arbitrary value $\neq 0$. Generalization to other Gaussian learning models, such as Gaussian Process regression (e.g. Wu et al., 2018; Schulz, Speekenbrink, & Krause, 2018), will also be relatively straightforward.

Parameter identifiability is an important but often overlooked aspect of computational modelling of empirical data. We have focused here on the structural identifiability of relatively simple models, and we could show that particular models were not fully identifiable. For more complex models, structural identifiability may not be as straightforward to determine analytically. In those cases, one may attempt to address the identifiability of a model by more "empirical" methods, such as assessing the "flatness" of the profile likelihood (Raue et al., 2009). Such methods may also be able to detect practical non-identifiability (in the sense that a particular data set is insufficient to estimate the parameters with any precision). As the models become more complex, assessing the structural and practical identifiability of models will become an increasingly difficult but important aspect of computational cognitive neuroscience.

References

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS computational biology*, *11*(11), e1004567.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineers, Series D, Journal of Basic Engineering*, *82*, 35–45.

Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Transactions of the American Society of Mechanical Engineers, Series D, Journal of Basic Engineering*, *83*, 95–108.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology*, *2:398*, 1–12.

May, B. C., Korda, N., Lee, A., & Leslie, D. S. (2012). Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, *13*(Jun), 2069–2106.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, *25*(15), 1923–1929.

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*, 351–367. doi: 10.1111/tops.12145

Thompson, W. R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, *25*, 285–294. doi: 10.2307/2332286

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General*, *143*, 2074–2081. doi: 10.1037/a0038199

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behavior*, *2*(12), 915.