Dirichlet process–based component detection in state-space models

Botond A. Bócsi and Lehel Csató
Fac. of Mathematics and Informatics, Babeş-Bolyai University
Kogalniceanu str. 1, RO-400084 Cluj-Napoca, Romania
E-mail: *bocsiboti@gmail.com*, *lehel.csato@cs.ubbcluj.ro*

Abstract. An extension of the switching-state models (SSSM) that allows arbitrary number of components is presented. We introduce a Dirichlet process prior over the mixture components of the linear models. This prior allows the inference on the number of linear models to be put into the mixture. We develop a distance measure in the space of linear Kalman filters with the use of the Kullback-Leibler divergence over the conditional probabilities induced by the individual Kalman filters. The introduced distance measure allows to remove components that are no longer relevant, making the algorithm more effective. We test the proposed algorithm on both artificial and real-world data.

1 Introduction

The analysis of sequential data is an important research topic: this type of data is found in several domains like the analysis of medical data [1], the forecasting of economical fluctuations, or in robotics, where the information to be processed lies in sequential data obtained from the sensors [8]. In this paper we focus on analyzing sequential data obtained from robot manipulation.

We extend the linear Kalman filtering (KF) scheme [9] to a nonlinear framework that is better suited for the usually nonlinear real world data. Although nonlinear modeling is a better choice, it is very often computationally infeasible. A possible solution to the intractability is to use a nonlinear model built from *locally linear* models [7, 11], similar to the mixture models in clustering [2], called switching-state space models (SSSM). We extend the SSSM's with the possibility to determine the number of components in the data set. This is achieved by using a Dirichlet prior assumption on the structure of the model that allows the insertion of a new local KF and to fit it to the data we have. As the number of components can grow indefinitely, we need a mechanism that *trims* the model by *removing* filters that are not used. This is achieved with the introduction of a distance measure in the space of the local Kalman filters.

The paper is organized as follows: first an introduction is presented into the topic of SSSM's and the Dirichlet prior on the structure of the model is introduced. Section 3 contains the proposed parameter estimation algorithm for the SSSMs. The new Kalman filters measure is described in Section 4. The simulations we conducted to support our approach are presented in Section 5, with conclusions drawn in Section 6.

2 Switching state-space models

In this paper we build a generative model for the observed data assuming that the data are coming from several distinct sources. We further assume that these sources are "simple", although in general they can also be more complex than linear. Here we focus on simple individual models without knowing from which component a specific

data was obtained, known as mixture models in machine learning [7, 11]. For numerical tractability we assume that the sources are Kalman filters [2, 8, 9], therefore we keep the simple models and only mention that the nonlinear extension has been studied by Honkela [10] using neural networks. Nonlinearity is introduced via the locally linear switching state-space models (SSSM's), where the sources are Kalman filters. In this case the SSSM is called a switching state Kalman filter (SSKF). If we denote the observations with z_k and the latent vector as x_k, the state-space equations are as follow:

$$\begin{aligned} x_k^{(s)} &= \mathbf{F}^{(s)} x_{k-1}^{(s)} + w_k^{(s)} \\ z_k &= \mathbf{H}^{(s)} x_k^{(s)} + v_k^{(s)}, \end{aligned} \tag{1}$$

where $\mathbf{F}^{(s)}$ is the time transition matrix, $\mathbf{H}^{(s)}$ is the output observation matrix. We assume that the dimensions of the matrices are consistent such that all multiplications can be performed in eq. (1). The random variables $w_k^{(s)}$ and $v_k^{(s)}$ are the driving and observation noise processes, characterized by covariance matrices $\mathbf{R}^{(s)}$ and $\mathbf{Q}^{(s)}$ respectively. An important ingredient of the mixture model, the superscript $^{(s)}$ – with s the *switching variable* – defines which component KF produced z_k, the actual output. We assume N components: $s \in \{1, \ldots, N\}$. The dynamics of s is defined by a Hidden Markov Model (HMM) [2, 13]. Let Φ be the transition probability and π the initial state distribution of the HMM. To be able to estimate the parameters of the local filters, we have to assign the individual data points to a filter, i.e. we have to know the value of s for each z_k, done with the use of the HMM. The complete set of parameters of the SSKF is the following:

$$\begin{aligned} \theta &= \{\theta^{(s)}\}_{s=1}^{N} \cup \{\pi, \Phi\}, \\ \theta^{(s)} &= \{\mathbf{F}^{(s)}, \mathbf{H}^{(s)}, \mathbf{Q}^{(s)}, \mathbf{R}^{(s)}\}. \end{aligned} \tag{2}$$

We cannot assume we know N a-priori, thus we impose a prior distribution over the components and the number of components our model has. The choice of the Dirichlet prior looks convenient because of its useful properties (e.g. the measures drawn from a Dirichlet process (DP) are discrete with probability one). It is the multivariate generalization of the Beta distribution, it defines the belief that the probabilities of K rival events have given values x_i if each event has been observed $\alpha_i - 1$ times. The DP is the extension of the Dirichlet distribution to continuous spaces. We do not provide an exact definition of the DP, we point to the references by Ferguson [6] or Teh [16].

We model the parameters of a SSKF, and implicitly the number of components to be sampled from the following DP:

$$\theta^{(s)} \sim \mathrm{DP}(\alpha, \mathbf{G}),$$

where G is the base distribution, assumed to be uniform on the space of the parameters, this space is taken from eq. (2). The parameter α is a concentration parameter used to set the range of the number of components the model has. It is important to mention that with the proposed hierarchical approach we can deal with data build from a potentially infinite number of sources. This is possible since the structure of the model is not fixed and its complexity is dependent on the data. We next describe the inference algorithm.

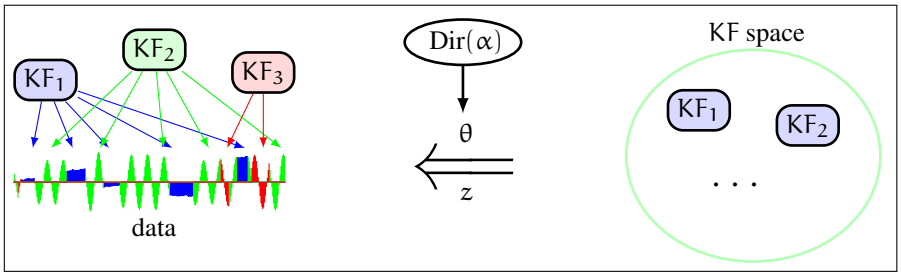

Fig. 1: Illustration of the component detection scheme.

3 Learning the SSKF parameters

The SSKF learning algorithm is introduced by Ghahramani and Hinton [7] and by Murphy [11]. The inference is based on a modified version of the Expectation-Maximization (EM) algorithm [7]. We extended this algorithm to infer not only the parameters of the local KF's and the global HMM, but also the number of Kalman filters required by the data. We summarize the learning algorithm for SSKF with Dirichlet prior as follows:
[E.] In the **expectation** step we calculate the observation probability – $p(z_k|S_k = m)$ – for every state-space model from the prediction error. Using this probability we obtain the responsibility assigned to every state-space model and every observation – $p(S_k = m)$ – using the forward-backward algorithm for the HMM. Lastly we run the Kalman smoother for every state-space model, using data weighted with the responsibility obtained.
[M.] In the **maximization** step we re-estimate the parameters for each state-space [2, 17], using data weighted by the responsibility from the E. step. Next we re-estimate the parameters of the HMM using the Baum-Welch algorithm [2, 13].
[Comp] We also introduce a third step that infers the **number of components**: adds or removes local KF's to the SSKF. First we define the removal procedure then the addition of new models. A component can be neglected in two cases: either when (1) its contribution drops below a threshold, as suggested by Bishop [2], or (2) when filters generate data very close to each other. The first case is detected by examining the responsibilities from step E. To detect the second case we developed a distance measure between two KF's – defined in Section 4.

The immediate approach that assumes a maximum number of components and during the learning process eliminates unnecessary SSKF-s is usually computationally unfeasible. Starting with a small initial component number and increasing its value based on data is a more convenient solution. This can be achieved by using the Dirichlet process extension, as introduced in Section 2. We want to get the posterior distribution of the states given by the Dirichlet process. The most convenient way is using Gibbs sampling [2, 14]. In the Gibbs sampling from a Dirichlet process mixture model [3, 12] we iterate for every KF model the following procedure: we sample a KF conditioned on all other switching variables except itself: s_{-n} and the whole observed data set z:

$$p(s_n^k = 1|z, s_{-n}) \quad \propto \quad p(z_n|z_{-n}, s_{-n}, s_n^k = 1)p(s_n^k = 1|s_{-n}), \tag{3}$$

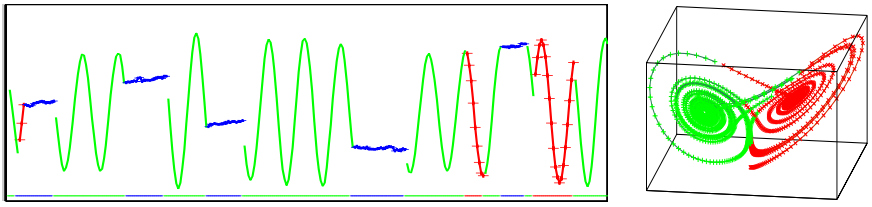

Fig. 2: Partitioning of (a) rotation data and (b) Lorenz attractor.

where the second term can be computed using eq. 4 below. Let k be the component we have chosen and assume that there are N components. We choose k with probabilities given by

$$p(s_n^k = 1|s_{-n}) = \frac{n_k}{\alpha + N - 1} \quad \forall k \leq N, \text{ and } p(s_n^{N+1} = 1|s_{-n}) = \frac{\alpha}{\alpha + N - 1}, \quad (4)$$

where n_k is the number of occurrences of the k-th filter. The right-side of eq. (4) is the probability that a new filter will be inserted into the SSKF.

The first term of eq. 3 can be obtained using the Kalman smoother for the new potential state and comparing its likelihood based on the prediction error of the filters. The Dirichlet process provides means to add new components to the SSKF and we now introduce a method that allows for simplification: we compute a "distance" in the space of Kalman filters presented next.

4 Distance measure of Kalman filters

The Kullback-Leibler (KL) divergence is widely used to measure distances between probability distributions. Additionally to its popularity, we know that predictions of the filter are Gaussian random variables and the computation of the KL divergence between Gaussians has analytical form [4]. We introduce this measure based on KL divergence:

$$d(KF^{(1)}, KF^{(2)}) \stackrel{o}{=} \int dp(z_0)\ KL\left(p(z_1^{(1)}|z_0) \| p(z_1^{(2)}|z_0)\right). \quad (5)$$

In the equation above we used $KF \stackrel{o}{=} p(z_1|z_0)$ with z_1 the predicted random variable conditioned on z_0. Since z_0 itself is unknown, we treat it again as a random variable $p(z_0) = \mathcal{N}(0, \mathbf{\Sigma}_{z_0})$ and average with respect to it, as shown above. $KF^{(1)}$ and $KF^{(2)}$ are the two Kalman filters and $KL(\cdot,\cdot)$ is the symmetric extension of the KL divergence [5] between the two conditional predictive distributions. Thus the evaluation of $p(z_1|z_0)$ is needed for each filter. For each KF this has the formula:

$$p(z_1|z_0) = \mathcal{N}(\mathbf{H}\mathbf{F}\mu_0, \mathbf{H}(\mathbf{F}\mathbf{\Sigma}_0\mathbf{F}^T + \mathbf{Q})\mathbf{H}^T + \mathbf{R}), \quad (6)$$

where $\mu_0^T = z_0^T\mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{P}_0^{-1})^{-1}$ and $\mathbf{\Sigma}_0 = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{P}_0^{-1})^{-1}$, where $\mathbf{P}_0$ satisfies $\mathbf{H}\mathbf{P}_0\mathbf{H}^T = (\mathbf{\Sigma}_{z_0} - \mathbf{R})$. Note that the KL distance between Gaussians has closed form, the evaluation of eq. (5) using eq. (6) is straightforward [4], therefore we have an easily computable measure between two KF's using minimalist assumptions about the

α	with DP	no DP
0.6	3.1250	2.0000
0.8	3.3000	2.4500
1.0	3.1500	2.7788
1.2	3.2500	2.9851
1.4	3.3250	3.1347

α	with DP	no DP
0.6	10.0125	3.0000
0.8	10.4875	17.0125
1.0	10.5125	21.4986
1.2	10.2750	23.2345
1.4	10.6375	24.5583

Table 1: (a) number of detected components, (b) steps taken to converge. The initial component number in no DP case was $\lfloor 4.3 * \alpha \rfloor$.

most probable value of z_0. Also interesting to note is that this distance does not depend on the dimension of the latent space of the filters, therefore the presented KL-based distance measure allows for direct comparison between filters of arbitrary latent spaces that produce output of the same dimension.

5 Simulations

We tested our method on artificial and real–world data, with our main interest being the inference on the number of components. We used three data–sets to examine the effectiveness of our proposal. The first was created using three filters, each with a two dimensional latent space and two dimensional output space. The first filter implemented rotation, the second one left the data unchanged, and the third one was also rotation, in the opposite direction to the first filter. The observations were corrupted with zero mean normal noise with variance 0.1. We plotted the first dimension on Figure 2.a, each detected component with a different style. We see that indeed the proposed algorithm identifies mostly correctly the components. In the second experiment we wanted to partition the three dimensional Lorenz attractor with parameters $\rho = 28$, $\sigma = 10$ and $\beta = 8/3$ [15]. There were just two components left, as it is shown in Figure 2.b. The third set is the KIN40 data–set[1], a realistic simulation of the forward dynamics of an 8 link all-revolute robot arm.

We run all three experiments with different hidden dimension of the internal states and our prior on the number of sources. Each parameter settings were tested twenty times, totally performed 3×600 experiments. Table 1 shows the experimental results. Table 1.a contains the number of active components after the SSKF converged as a function of the prior assumption about the number of it. We see with the use of a Dirichlet prior the component number is almost independent of the concentration parameter α. Table 1.b contains the number of EM steps needed to converge, one can again see the advantages of the use of the Dirichlet process together with a KL-distance based removal of components. Again, the convergence speed is almost constant, not depending on the prior values.

6 Discussion and further research

In this paper we presented a generalization of the SSKF framework by adding a Dirichlet prior over the structure of the model. This extension allows us to model SSKFs

[1] http://ida.first.fraunhofer.de/~anton/data/kin40k.mat

whose number of component is not fixed, theoretically can be very high, often infinitely many components and make the computations feasible. We also presented a new distance measure of the Kalman filters, this proved to be useful step in optimizing with respect to the number of components of the mixture. The presented simulations show that using our proposal a faster and a more efficient segmentation of the mixture model can be achieved, however, these results rely on generated data. We aim to test the method on real world data too, where the components have practical significance.

The presented component detection scheme is *unsupervised*: there is no "user" intervention required for the method to work. Furthermore, the result of this filtering is a set of "simple" models that can be used as building blocks for hierarchical motion planning systems, forming the basis of an ontology. It would be interesting to evaluate the component detection scheme in conjunction e.g. with a reinforcement learning algorithm that can borrow templates from the SSKF.

The authors acknowledge the support of the Romanian Ministry of Education, grant PN2 11-039/2007.

References

[1] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. The MIT Press, 1998.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] D. M. Blei and M. I. Jordan. Variational methods for the Dirichlet process. In *ICML 2004*, pages 1–12. ACM, 2004.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons, 1991.

[5] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[6] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[7] Z. Ghahramani and G. E. Hinton. Switching state-space models. Technical report, Universtiy of Toronto, 1996.

[8] M. S. Grewal and A. P. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB*. Wiley and Sons, second edition, 2001.

[9] S. Haykin. *Kalman Filtering and Neural Networks*. Wiley and Sons, 2001.

[10] A. Honkela. *Nonlinear Switching State-Space Models*. PhD thesis, Helsinki University of Technology, 2001.

[11] K. P. Murphy. Switching Kalman filters. Technical report, 1998.

[12] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[13] L. Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[14] C. P. Robert and G. Casella. *Monte Carlo Methods*. Springer, 2004.

[15] S. H. Strogatz. *Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry and Engineering*. Perseus Books Group, 2001.

[16] Y. W. Teh. Dirichlet processes. Submitted to Encyclopedia of Machine Learning, 2007.

[17] B. M. Yu, K. V. Shenoy, and M. Sahani. Derivation of Kalman filtering and smoothing equations. Technical report, Stanford University, 2004.