Constructing Gaussian Processes for Probabilistic Graphical Models

Mattis Hartwig, Marisa Mohr and Ralf Möller
University of Lübeck
{hartwig,mohr,moeller}@ifis.uni-luebeck.de

Abstract

Probabilistic graphical models have been successfully applied in a lot of different fields, e.g., medical diagnosis and bio-statistics. Multiple specific extensions have been developed to handle, e.g., time-series data or Gaussian distributed random variables. In the case that handles both Gaussian variables and time-series data, downsides are that the models still have a discrete time-scale, evidence needs to be propagated through the graph and the conditional relationships between the variables are bound to be linear. This paper converts two probabilistic graphical models (the Markov chain and the hidden Markov model) into Gaussian processes by constructing covariance and mean functions, that encode the characteristics of the probabilistic graphical models. Our developed Gaussian process based formalism has the advantage of supporting a continuous time scale, direct inference from any time point to the other without propagation of evidence and flexibility to modify the covariance function if needed.

1 Introduction

A lot of applications from medical diagnosis, bio-statistics, ecology, maintenance, etc. have been represented by probabilistic graphical models (PGMs) (Weber et al. 2012; McCann, Marcot, and Ellis 2006). Markov chains, hidden Markov models (HMMs) and dynamic Bayesian networks (DBNs) are PGMs that model time series (Murphy 2002). All three models have been originally discrete models, but versions that allow for continuous Gaussian distributed variables have been developed and successfully applied (Grzegorczyk 2010). PGMs in general offer a sparse and interpretable representation for probabilistic distributions and allow to model (in)dependencies between its random variables (Koller, Friedman, and Bach 2009; McCann, Marcot, and Ellis 2006). The interpretability of the modeling language for a PGM also makes it possible to construct PGMs based on expert knowledge instead of or as an addition to learning them from data (Constantinou, Fenton, and Neil 2016; Flores et al. 2011). There are downsides of the Gaussian variants of PGMs for time-series. First, the time dimension is still discrete which brings up the problem of finding the right sampling rate. Second, evidence is usually propagated through the graphical structure which can be computational expensive. Third, the Gaussian variants are based on linear relationships between random variables which makes it difficult to model certain real-world phenomenon, e.g. periodic behaviors.

Gaussian Processes (GPs) are another approach applied for modeling time-series (Roberts et al. 2013; Frigola-Alcalde 2016) and have been rather recently brought into focus in the machine learning community (Rasmussen 2006). Both Gaussian PGMs and GPs have Gaussian distributions over their random variables at any point in time. In contrast to PGMs, GPs are continuous on the time dimension and allow direct inference without propagation of evidence through a network. Additionally, an existing GP that models a certain behavior can be easily extended or adapted by making changes to its covariance function. Drawbacks of GPs are that modeling multiple outputs at once is challenging (Alvarez et al. 2012) and that modeling a detailed interpretable (in)dependence structure as it is done in a PGM is currently not possible.

Since the GPs and Gaussian time-series PGMs are both based on Gaussian distributions along a time dimension, this paper aims to bring the two approaches together. More specifically, we convert two well known Gaussian PGMs for time-series - the Markov chain and the HMM - into a GP representation which unlocks the benefits mentioned above. The key is to build a covariance function that encodes the characteristics of the PGMs.

The remainder of the paper has following structure. We start by explaining the preliminaries about PGMs and GPs and their respective benefits. Afterwards we discuss related work that draw connections between relation based models and GPs and construct GPs for two PGMs - the Markov chain and the hidden Markov model. We conclude with a discussion of benefits and downsides of the created GPs and with an agenda for further research in that area.

2 Preliminaries

In this section we introduce PGMs, GPs and kernel functions for GPs. Afterwards, we discuss the advantages of the two models, which also motivates combining them.

2.1 Probabilistic Graphical Models

This section gives a brief overview about the three different types of PGMs used in this paper. For further details we re-

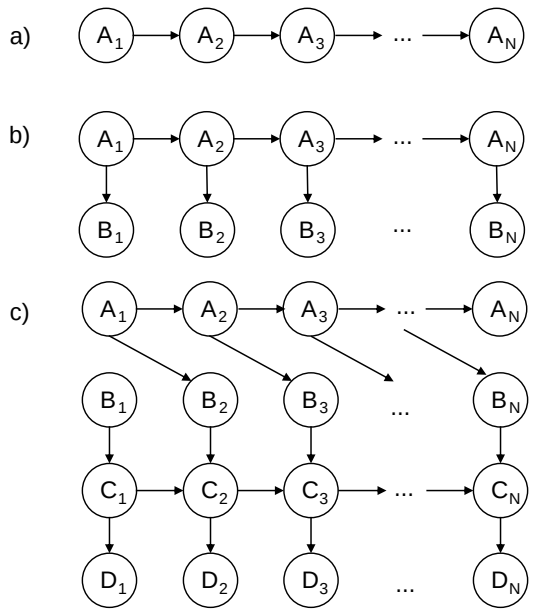

Figure 1: Three different types of PGMs: a) Markov chain b) hidden Markov model c) dynamic Bayesian network

fer to the work by Koller, Friedman, and Bach (2009), Pearl (1988) and Murphy (2002).

In general, a PGM is a network with nodes for the random variables and edges to describe relations between them. A Gaussian Markov chain describes the development of a single Gaussian distributed random variable over time. A Gaussian HMM has two variables that develop over time, one being a latent state and one being an observable state variable. A dynamic Gaussian Bayesian network is a general PGM, that allows arbitrary links between the random variables (Murphy 2002). Figure 1 contains illustrations of the three different types of PGMs. We denote the set of random variables as $\mathbb{X}$ and the set random variables that are influencing a specific random variable $A \in \mathbb{X}$ as its parents $Pa(A)$. Each random variable follows a conditional Gaussian probability distribution that is linearly dependent on its parent and is given by

$$P(A|Pa(A)) \sim N\left(\mu_A + \textstyle\sum_{\Pi \in Pa(A)} \beta_{A,\Pi}(\pi - \mu_\Pi), \sigma_A^2\right), \quad (1)$$

where μ_A and μ_Π are the unconditional means of A and Π respectively, π is the realization of Π, σ_A^2 is the variance of A and $\beta_{A,\Pi}$ represents the influence of the parent Π on its child A.

The joint probability distribution

$$P(\mathbb{X}) = \prod_{X \in \mathbb{X}} p(X|Pa(X)) \quad (2)$$

is the product of all conditional probability distributions.

In a PGM for time-series, the variables $\mathbb{X}$ develop over time which we denote by $\mathbb{X}_t$. It can be fully defined by starting distribution $P(\mathbb{X}_1)$ of $\mathbb{X}$ at time $t = 1$ and its transition over time which is described by the conditional probability distribution $P(\mathbb{X}_t|\mathbb{X}_{t-1})$ (Murphy 2002).

2.2 Gaussian Processes

A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution (Rasmussen 2006). A GP can be interpreted as a distribution over functions on a spatial dimension, which is in our case the time dimension t. It is completely specified by its mean $\mu = m(t)$ and its covariance function $k(t, t')$ and can be written as

$$f(t) \sim GP(m(t), k(t, t')). \quad (3)$$

The covariance function (also known as kernel function) describes the similarity of function values at different points in time (t and t') and influences the shape of the function space (Rasmussen 2006).

If we have a dataset that consists of an input vector $\mathbf{t}$ and an output vector $\mathbf{y}$, we can define any vector of time points $\mathbf{t}^*$ for which we would like to calculate the posterior distribution. The joint distribution over the observed and the unknown time points is given by

$$p\left(\begin{bmatrix} y \\ y* \end{bmatrix}\right) = N\left(\begin{bmatrix} \mu(t) \\ \mu(t*) \end{bmatrix}, \begin{bmatrix} K(\mathbf{t},\mathbf{t}) & K(\mathbf{t},\mathbf{t}^*) \\ K(\mathbf{t}^*,\mathbf{t}) & K(\mathbf{t}^*,\mathbf{t}^*) \end{bmatrix}\right), \quad (4)$$

where $K(\mathbf{t}, \mathbf{t}^*)$ is a covariance matrix produced by plugging all values from $(\mathbf{t}, \mathbf{t}^*)$ into the covariance function $k(t, t')$. By applying the conditional probability rules for multivariate Gaussians (Roberts et al. 2013) we obtain the posterior $P(\mathbf{y}^*)$ with mean $\mathbf{m}^*$ and covariance matrix C^*

$$P(\mathbf{y}^*) = N(\mathbf{m}^*, C^*), \quad (5)$$

where

$$\mathbf{m}^* = \mu(\mathbf{t}^*) + K(\mathbf{t}^*,\mathbf{t})K(\mathbf{t},\mathbf{t})^{-1}(\mathbf{y} - \mu(\mathbf{t})) \quad (6)$$

and

$$C^* = K(\mathbf{t}^*,\mathbf{t}^*) - K(\mathbf{t}^*,\mathbf{t})K(\mathbf{t},\mathbf{t})^{-1}K(\mathbf{t}^*,\mathbf{t})^T. \quad (7)$$

2.3 Kernel Functions

Rasmussen (2006) provides an overview of different possible kernels with the squared exponential kernel

$$k_{SE}(t, t') = \sigma^2 \exp\left(-\frac{(t - t')^2}{2l^2}\right), \quad (8)$$

where σ^2 and l are hyperparameters for the signal noise and the lengthscale respectively, being a commonly used one.

A valid kernel $k : \mathbb{T} \times \mathbb{T} \to \mathbb{R}$ for a GP needs to fulfill three characteristics (Rasmussen 2006):

- continuity, i.e., $\mathbb{T} \subset \mathbb{R}$,
- symmetry, i.e., $k(t, t') = k(t', t)$ for all t and t',
- being positive semidefinit, i.e., symmetry and $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(t_i, t_j)$ for $n \in \mathbb{N}$, $t_1, ..., t_n \in \mathbb{T}$, $c_1, ..., c_n \in \mathbb{R}$.

Valid Kernels can be constructed of other kernels. Bishop (2006) lists valid kernel operations from which we use the following subset in later sections. Given valid kernels $k_1(t,t')$, $k_2(t,t')$ and a constant c, the following kernels will also be valid:

$$k(t,t') = ck_1(t,t'), \tag{9}$$

$$k(t,t') = k_1(t,t') + k_2(t,t'), \tag{10}$$

$$k(t,t') = \exp(k_1(t,t')), \tag{11}$$

$$k(t,t') = k_1(t,t')k_2(t,t'). \tag{12}$$

2.4 Benefits of the Models

PGMs have several benefits that make them straightforward to use. One benefit is that they can capture (conditional) dependencies and independencies of the random variables very intuitively (Koller, Friedman, and Bach 2009). Another benefit is that PGMs can incorporate expert knowledge; it is possible to construct a network entirely by expert knowledge but it is also possible to use expert knowledge as a prior for the probability distribution (Flores et al. 2011). HMMs and DBNs can model the probability distribution over multiple random variables simultaneously. Last but not least, PGMs have already been used in many applications and therefore a wide range of inference and learning tactics have been developed (Koller, Friedman, and Bach 2009).

The usage of GPs has also benefits. GPs have a continuous sequential dimension which allows to model continuous changes directly and without the need of discretization. GPs are nonparametric and directly incorporate a quantification of uncertainty. Because of their joint Gaussian characteristics, calculating posterior distributions is straightforward and relatively efficient (Roberts et al. 2013).

Converting the PGMs to GPs while retaining the PGM characteristics is a promising approach that we pursue in this paper to exploit the benefits of both approaches.

3 Related Work

There have been three different streams to bring graphical or relational models together with GPs. One research stream known as relation learning uses multiple GPs to identify probabilistic relations or links within sets of entities (Xu, Kersting, and Tresp 2009; Yu et al. 2007). A second research stream uses GPs for transition functions in state space models. Frigola-Alcalde (2016) has researched different techniques for learning state space models that have GP priors over their transition functions and Turner (2012) has explored change point detection in state space models using GPs. A third research stream focuses on constructing covariance functions for GPs to mimic certain behaviors from other models. Reece and Roberts (2010b) have shown that they can convert a specific Kalman filter model for the near constant acceleration model into a kernel function for a GP and then Reece and Roberts (2010a) combine that kernel function with other known kernels to get better results temporal-spatial predictions. Rasmussen (2006) has introduced GP kernels for, e.g., a wiener process (the min kernel), that we will reuse in this paper as well. The results of this paper contribute to the third by providing a novel approach to converting two PGMs into a GP, which as far as we know has not yet been done.

4 Gaussian Processes for Markov Chains

In this section we construct a continuous GP for the Gaussian Markov chain model. Let A be a random variable with a linear relationship between the time points t and $t-1$. The Gaussian distributed Markov chain can be also interpreted as a dynamic Gaussian Bayesian network with one dimension. It is described by its initial distribution $P(A_1)$ with

$$P(A_1) = N(\mu_{A_1}; \sigma^2_{A_1}) \tag{13}$$

and a transition distribution $P(A_t|A_{t-1})$ with

$$P(A_t|A_{t-1}) = N(\mu_{A_t} + \beta_{A_t,A_{t-1}}(a_{t-1} - \mu_{A_{t-1}}); \sigma^2_{A_t}). \tag{14}$$

For simplicity, we assume that $\sigma^2_{A_1} = \sigma^2_{A_t} =: \sigma^2_A$ and $\mu_{A_1} = \mu_{A_t} =: \mu_A$.

4.1 Constructing the Kernel

Shachter and Kenley (1989) have developed an algorithm to convert a Gaussian Bayesian network into a multivariate Gaussian distribution. To prove correctness, they formulated the following Lemma that we will reuse.

Lemma 1. *For $N \in \mathbb{N}$ topological ordered random variables $X^i \in \mathbb{X}$, $i = 1,\dots,N$ in a Gaussian Bayesian network let σ_i^2 be the variance of the conditional distribution of X^i given its parents. Let $B \in \mathbb{R}^{N\times N}$ be a matrix, where the entries $\beta_{i,l}$, $l = 1,...,N$ describe the linear relationship between a child X^i and its parent X^l. If X^l is no parent of X^i the entry is zero. For a fixed $j \in \{1,\dots,N\}$ let Σ_{tt} be the covariance matrix between all random variables X^t, $t = 1,...,j$ and $B_{sj} \in \mathbb{R}^{j-1\times 1}$, $s = 1,\dots,j-1$ the corresponding part of B. We denote the matrices*

$$S_j := \begin{bmatrix} \Sigma_{tt} & 0 & \dots & 0 \\ 0 & \sigma^2_{j+1} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \sigma^2_N \end{bmatrix}, \tag{15}$$

$$U_j := \begin{bmatrix} I_{j-1} & B_{sj} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & I_{N-j} \end{bmatrix}. \tag{16}$$

Then it is

$$S_j = U^T_{j-1} S_{j-1} U_{j-1} \tag{17}$$

and

$$\Sigma = S_N = U^T_N ... U^T_1 S_0 U_1 ... U_N, \tag{18}$$

where $\Sigma \in \mathbb{R}^{N\times N}$ is the covariance matrix of the equivalent multivariate Gaussian distribution for the above defined Gaussian Bayesian network.

The $N \times N$ covariance matrix from Equation 18 is calculated by recursively multiplying the U-matrices. To define a GP, a kernel function must be constructed that maps arbitrary time points t and t' to a covariance value. Therefore we convert the recursive multiplication of the matrices in Equation 18 into a kernel function.

The matrix S_0 is diagonal with the individual variances for each individual node. In the case of the Markov chain, each value on the diagonal is σ_A^2. Additionally, the matrix B has entries $\beta_{A_t,A_{t-1}} := \beta_A$ at the positions $(s, s+1)$ for $s = 1, ..., N$, which describe the linear relationship along the time dimension of A, and zeroes everywhere else. Consequently, the matrix U_i is the identity matrix of the size $N \times N$ with a β_A at position $(i-1, i)$. By multiplying all U-matrices as indicated above, we obtain the diagonal matrix

$$\prod_{i=1}^{N} U_i = \begin{bmatrix} 1 & \beta_A & \beta_A^2 & \dots & \beta_A^N \\ 0 & 1 & \beta_A & \dots & \beta_A^{N-1} \\ 0 & 0 & 1 & \dots & \beta_A^{N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}. \tag{19}$$

The same applies for the left part of the Equation 18:

$$\prod_{i=1}^{N} U_i^T = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \beta_A & 1 & 0 & \dots & 0 \\ \beta_A^2 & \beta_A & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_A^N & \beta_A^{N-1} & \beta_A^{N-2} & \dots & 1 \end{bmatrix}. \tag{20}$$

Since all values on the diagonal-matrix S_0 have the same scalar value σ_A^2, we can multiply out the constant σ_A^2, resulting in

$$\Sigma = U_N^T ... U_1^T S_0 U_1 ... U_N = \sigma_A^2 \prod_{i=1}^{N} U_i^T \prod_{i=1}^{N} U_i. \tag{21}$$

We multiply row i from Equation 20 with column j from Equation 19. For $i = j$ (the diagonal of the resulting covariance matrix), we calculate the covariance with

$$\begin{bmatrix} \beta_A^{i-1} \\ \beta_A^{i-2} \\ \beta_A^{i-3} \\ \vdots \\ \beta_A^0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \cdot \begin{bmatrix} \beta_A^{i-1} \\ \beta_A^{i-2} \\ \beta_A^{i-3} \\ \vdots \\ \beta_A^0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T = \sum_{k=0}^{i} (\beta_A^k)^2. \tag{22}$$

For $i \neq j$, we denote the difference as $d = |i - j|$. Because of the symmetry of matrices in 19 and 20 there is no difference between the cases $j > i$ and $j < i$. Let $j > i$, then the j-th vector contains more elements and in addition elements with higher exponents. This changing values are only relevant in the first i entries because all other entries are multiplied with zero. Substituting j by $i + d$ results in:

$$\begin{bmatrix} \beta_A^{i-1} \\ \beta_A^{i-2} \\ \beta_A^{i-3} \\ \vdots \\ \beta_A^0 \end{bmatrix} \cdot \begin{bmatrix} \beta_A^{j-1} \\ \beta_A^{j-2} \\ \beta_A^{j-3} \\ \vdots \\ \beta_A^{j-i} \end{bmatrix}^T = \begin{bmatrix} \beta_A^{i-1} \\ \beta_A^{i-2} \\ \beta_A^{i-3} \\ \vdots \\ \beta_A^0 \end{bmatrix} \cdot \begin{bmatrix} \beta_A^{i+d-1} \\ \beta_A^{i+d-2} \\ \beta_A^{i+d-3} \\ \vdots \\ \beta_A^d \end{bmatrix}^T \tag{23}$$
$$= \sum_{k=0}^{i} \beta_A^k \beta_A^{k+d} = \beta_A^d \sum_{k=0}^{i} (\beta_A^k)^2.$$

To ensure symmetry of the kernel, we also need to include the case $i > j$, which leads to replacing i by $\min(i, j)$ and d by $|i-j|$ in Equation 23. We reformulate the whole equation using the formula for the partial sum of a geometric series and replace i and j with t and t' respectively:

$$k(t, t') = \sigma_A^2 \beta_A^{|t-t'|} \frac{1 - \beta_A^{\min(t,t')}}{1 - \beta_A}. \tag{24}$$

We have shown that Equation 24 can be used to construct a covariance matrix that encodes the characteristics of the Gaussian Markov chain.

4.2 Proving the Kernel

To prove the validity of the kernel for a GPs, we use the characteristics that kernels can be constructed of other kernels using the equations defined in Section 2.4. We show that the two factors

$$k_a(t, t') = \sigma_A^2 \beta_A^{|t-t'|}, \tag{25}$$

$$k_b(t, t') = \frac{1 - \beta_A^{\min(t,t')}}{1 - \beta_A}. \tag{26}$$

of the Equation 24 are valid kernels. Then, based on Equation 12, our constructed kernel being the product of the two factors is a valid kernel as well. The exponent $|t - t|'$ is the one dimensional case of the Euclidean distance kernel (Bishop 2006). With Equation 9 and the exponential rule from Equation 11 $k_a(t, t')$ is a valid kernel.

To show that the function $k_b(t, t')$ is a valid kernel, we use the summation in Equation 23. Equation 23 converts the fraction back into a sum of exponential functions that have a positive numbers $n = 1, ..., \min(t, t')$ as the exponents. Based on Equation 10, Equation 11 and the fact that $\min(t, t')$ is a valid kernel (Rasmussen 2006), the function $k_b(t, t')$ is a valid kernel. Consequently, Equation 24 is a valid kernel for a GP.

4.3 Defining the Gaussian Process

We just constructed the kernel function, so in order to complete the GP, the mean function needs to be defined. Equation 14 shows that the value of the random variable A at time t is defined by a linear function of the difference between the parents value and its mean denoted by $a_{t-1} - \mu_A$. Without any evidence fed into the GP, the difference is symmetrical distributed around zero, resulting in a constant mean function

$$m(t) = \mu_A. \tag{27}$$

With the covariance function from Equation 24 and the mean function from Equation 27 we have constructed the desired GP for a Gaussian Markov chain. The hyperparameters for the constructed GP are β_A, σ_A^2 and μ_A.

5 Gaussian Processes for Hidden Markov Models

In a HMM, there are two random variables. We denote the hidden variable as A and the observable variable as B. In notation of a dynamic Gaussian Bayesian network, we keep Equations 13 and 14 for the distributions over A_t and have a conditional distribution for B_t

$$P(B_t|A_t) = N(\mu_{B_t} + \beta_{B,A}(a_t - \mu_{A_t}); \sigma^2_{B_t}), \quad (28)$$

where the influence from the parent node A_t to its child node B_t is equal in every time step. Again we simplify with $\sigma^2_{B_t} =: \sigma^2_B$ and $\mu_{B_t} =: \mu_B$.

5.1 Constructing the Kernel

We generalize the kernel from the previous section to support the two random variables as outputs. We denote D as set of random variables with $D = \{A, B\}$ in specific HMM case. Alvarez et al. (2012) introduces multi-output kernels in the format of

$$k((D,t),(D',t')). \quad (29)$$

In the previous section the kernel described the similarity of the random variable A at two different time points t and t'. The new kernel from based on Equation 29 describes the similarity of a random variable D at time t to another random variable D' at time t'.

In Section 4.2, we used the Lemma from Shachter and Kenley (1989) to construct the covariance function. For the HMM case, we use following recursive covariance formula from Shachter and Kenley (1989):

$$\Sigma_{i,j} = \Sigma_{j,i} = \sum_{\pi \in Parents(Node_i)} \Sigma_{j,\pi}\beta_\pi, \quad (30)$$

$$\Sigma_{i,i} = \sum_{\pi \in Parents(Node_i)} \Sigma_{i,\pi}\beta_\pi + \sigma_i^2, \quad (31)$$

where $\Sigma_{i,j}$ is the covariance between two random variables X_i and X_j. We reformulate for the multidimensional case and replace Σ with the kernel function

$$\begin{aligned} &k((D,t),(D',t')) = k((D',t'),(D,t)) \\ &= \sum_{(\Pi,\pi)\in Pa(D,t)} k((D',t'),(\Pi,\pi))\beta_{D\Pi}. \end{aligned} \quad (32)$$

The kernel $k(t,t',A,A)$ is the same kernel as it is in the single variable case, because a B-node is never parent of an A-node:

$$k((A,t),(A',t')) = \sigma_A^2 \beta^{|t-t'|}\frac{1-\beta^{\min(t,t')}}{1-\beta}. \quad (33)$$

In the case $k(t,t',B,A)$ including the knowledge that $Pa(B,t) = (A,t)$, the recursive formula from Equation 32 is

$$\begin{aligned} k((B,t),(A,t')) &= k((A,t'),(B,t)) \\ &= k((A,t')(A,t))\beta_{B,A}. \end{aligned} \quad (34)$$

The last open case is the one where we calculate covariances between two B-nodes. We reuse the recursive formula from Equation 32 with the knowledge that the node B_t has node A_t as a parent resulting in following formula

$$\begin{aligned} k((B,t),(B,t')) &= k((B,t'),(B,t)) \\ &= k((B,t')(A,t))\beta_{B,A}. \end{aligned} \quad (35)$$

Reusing Equation 34 results in

$$\begin{aligned} k((B,t),(B,t')) &= k((B,t')(A,t))\beta_{B,A} \\ &= k((A,t)(A,t')\beta^2_{B,A}. \end{aligned} \quad (36)$$

In the case $t = t'$, we need to add the constant σ_B^2 resulting in the final kernel

$$k((D,t),(D',t')) = \begin{cases} k(t,t'), & \text{if } D,D' = A \\ k(t,t')\beta_{B,A}, & \text{if } D \neq D' \\ k(t,t')\beta^2_{B,A} + \delta_{tt'}\sigma_B^2, & \text{if } D,D' = B \end{cases}, \quad (37)$$

where $\delta_{tt'}$ is the Kronecker-Delta.

5.2 Proving the Kernel

A separable kernel is a multi-dimensional kernel, that can be reformulated into a product of single-dimensional kernels

$$k((D,t),(D',t')) = k(D,D')k(t,t'). \quad (38)$$

A sum of separable kernels is also a valid kernel (Rasmussen 2006; Alvarez et al. 2012).

We need to show that Equation 37 can be rewritten as a sum of separable kernels. Therefore we write the three different cases as separate kernels on D and D' which results in

$$k_1(D,D') = \epsilon_{D,D',A}, \quad (39)$$

$$k_2(D,D') = 1 - \delta_{D,D'}, \quad (40)$$

$$k_3(D,D') = \epsilon_{D,D',B}, \quad (41)$$

where $\epsilon_{D,D',D*}$ is equal to 1 if $D = D' = D^*$ and otherwise zero and $\delta_{D,D'}$ in the Kronecker-Delta. With Equations 39-41, Equation 37 can be rewritten to

$$\begin{aligned} k((D,t),(D',t')) &= k_1(D,D')k(t,t') \\ &+ k_2(D,D')(k(t,t')\beta_{B,A}) \\ &+ k_3(D,D')k(t,t')\beta^2_{B,A} + \delta_{t,t'}\sigma_B^2. \end{aligned} \quad (42)$$

That proves that the constructed kernel is a valid kernel as well.

5.3 Defining the Gaussian Process

As in Section 4.3, we also need to construct the mean function. Analogue to the single variable case, the mean functions for A and B are constant resulting in

$$m(D,t) = \begin{cases} \mu_A, & \text{if } D = A \\ \mu_B, & \text{if } D = B \end{cases}. \quad (43)$$

For the constructed GP we have a set of hyperparameters $\theta = \{\mu_A, \mu_B, \beta_A, \beta_{B,A}\}$. In HMMs, the mean of B is usually set to be equal to the mean of A which can be defined for the model by setting $\mu_B = \mu_A$.

6 Conclusion and Outlook

We have shown that we can convert two types of PGMs into Gaussian Processes. The transformation of the two PGMs into GPs brings significant benefits:

Continuity: The model supports a continuous time scale. This avoids setting a sample rate and rounding beforehand because evidence or queried variables can be at any real-valued point in time.

Kernel Combination: The created kernel can be combined with any other existing kernels. If the real-world phenomenon is relatively well described by the GDBN-kernel but also contains a slight periodic behavior, both kernels can easily combined by different operations, e.g. addition and multiplication (Rasmussen 2006).

Efficient Query Answering: In a dynamic PGM the evidence is usually propagated through the model along the time dimension. In the constructed GP the kernel allows us to explicitly define the effect of an observation to any other queried point in time which speeds up the querying answering process.

Markov Property: The defined kernels keep the Markov property and the transition behavior of the underlying model.

Further relaxing constraints on the underlying PGMs will be the focus of our further research in that area. More specifically, we would like to further generalize our approach to support dynamic Gaussian Bayesian networks, that allow arbitrary connections between random variables (as long as the acyclic-property is not violated). Additionally, we will add real-world evaluations. First, we would like to evaluate whether the newly defined kernels and their characteristics can enhance existing use cases of GPs for time series modeling in their prediction. Second, we would like evaluate if existing use cases of dynamic Gaussian PGMs can be enhanced by either having less run-time for inference, having a continuous time scale or by combining the new kernels with existing ones.

References

Alvarez, M. A.; Rosasco, L.; Lawrence, N. D.; and Others. 2012. Kernels for vector-valued functions: A review. *Foundations and Trends®in Machine Learning* 4(3):195–266.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. Springer.

Constantinou, A. C.; Fenton, N.; and Neil, M. 2016. Integrating expert knowledge with data in Bayesian networks: Preserving data-driven expectations when the expert variables remain unobserved. *Expert Systems with Applications* 56:197–208.

Flores, M. J.; Nicholson, A. E.; Brunskill, A.; Korb, K. B.; and Mascaro, S. 2011. Incorporating expert knowledge when learning Bayesian network structure: a medical case study. *Artificial intelligence in medicine* 53(3):181–204.

Frigola-Alcalde, R. 2016. *Bayesian time series learning with Gaussian processes*. Ph.D. Dissertation, University of Cambridge.

Grzegorczyk, M. 2010. An introduction to Gaussian Bayesian networks. In *Systems Biology in Drug Discovery and Development*. Springer. 121–147.

Koller, D.; Friedman, N.; and Bach, F. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

McCann, R. K.; Marcot, B. G.; and Ellis, R. 2006. Bayesian belief networks: applications in ecology and natural resource management. *Canadian Journal of Forest Research* 36(12):3053–3062.

Murphy, K. P. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. Dissertation.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rasmussen, C. E. 2006. Gaussian processes for machine learning. MIT Press.

Reece, S., and Roberts, S. 2010a. An introduction to Gaussian processes for the Kalman filter expert. In *2010 13th International Conference on Information Fusion*, 1–9. IEEE.

Reece, S., and Roberts, S. 2010b. The near constant acceleration Gaussian process kernel for tracking. *IEEE Signal Processing Letters* 17(8):707–710.

Roberts, S.; Osborne, M.; Ebden, M.; Reece, S.; Gibson, N.; and Aigrain, S. 2013. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371(1984):20110550.

Shachter, R. D., and Kenley, C. R. 1989. Gaussian Influence Diagrams. *Management Science* 35(5):527–550.

Turner, R. D. 2012. *Gaussian processes for state space models and change point detection*. Ph.D. Dissertation, University of Cambridge.

Weber, P.; Medina-Oliva, G.; Simon, C.; and Iung, B. 2012. Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Engineering Applications of Artificial Intelligence* 25(4):671–682.

Xu, Z.; Kersting, K.; and Tresp, V. 2009. Multi-relational learning with gaussian processes. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Yu, K.; Chu, W.; Yu, S.; Tresp, V.; and Xu, Z. 2007. Stochastic relational models for discriminative link prediction. In *Advances in neural information processing systems*, 1553–1560.