The Variational Coupled Gaussian Process Dynamical Model

Dmytro Velychko DMYTRO.VELYCHKO@STAFF.UNI-MARBURG.DE
Benjamin Knopp KNOPPBE@STAFF.UNI-MARBURG.DE
Dominik Endres DOMINIK.ENDRES@UNI-MARBURG.DE
Theoretical Neuroscience Group, Dept. Psychology
Philipps-University of Marburg
Gutenbergstr. 18, 35032 Marburg, Germany

Abstract

We present a full variational treatment of the Coupled Gaussian Process Dynamical Model (CGPDM) with non-marginalized coupling mappings. A CGPDM is comprised of several latent dynamical models, each of which predicts a part of the observations with a mapping drawn from a Gaussian process. Furthermore, the dynamical models are coupled in a product-of-experts fashion, allowing for a flexible adjustment of the coupling strength at run time (Velychko et al., 2014). However, this previous treatments of the CGPDM integrated out the coupling mappings, making it virtually impossible to exchange the dynamics of the parts after learning. This exchange possibility would be crucial for the construction of modular movement primitive models, which is our primary application goal. Furthermore, such models need a compact representation of the individual dynamical models, to allow for the storage of a large library of movements. We tackle the first problem by a non-marginalized treatment of the CGPDM, and the second by representing each mapping by a small collection of inducing points (Titsias and Lawrence, 2010). Our work builds on similar developments in Gaussian state-space models (Frigola et al., 2014), but we obviate the need for sampling, which results in a fast deterministic approximation for the posterior of latent states. We test the model against human perception and illustrate the modularity on human movement data.

Keywords: Gaussian Process, Dynamical Systems, Coupling, Motor Primitives, Bayesian Statistics

1. Introduction

Planning and execution of full-body movements is a formidable control problem. Modular movement primitives (MP) have been suggested as a means to simplify this control problem while retaining a sufficient degree of control flexibility for a wide range of task, see Bizzi et al. (2008) for a recent review. 'Modular' in this context usually refers to the existence of an operation which allows for the combination of (simple) primitives into (complex) movements.

While evidence for the existence of modular MPs has been accumulated in biological motor control for quite some time, starting with Sherrington (1906), technical applications have also been devised. For example in computer graphics, especially combined with dynamics models (Giese et al., 2009; Lee et al., 2009) and robotics, e.g. the dynamical MP (DMP) (Ijspeert et al., 2013; Rückert and d'Avella, 2013). Each DMP is encoded by

a canonical second order differential equation with guaranteeable stability properties and learnable parameters.

To lift the restriction of canonical dynamics, we describe a model that learns MPs comprised of coupled dynamical systems and associated kinematics mappings, where *both* components are learned. We build on the Coupled Gaussian Process Dynamical Model (CGPDM) by Velychko et al. (2014), which combines the advantages of modularity and flexibility in the dynamics, at least theoretically. In a CGPDM, the temporal evolution functions for latent dynamical systems are drawn out of a Gaussian process (GP) prior Rasmussen and Williams (2005). These dynamical systems are then coupled probabilistically, the result is mapped onto observations by functions drawn from another GP. One drawback of the CGPDM is its fully non-parametric nature, which leads to a prohibitive run-time scaling for large data sets. We improve this scaling by employing sparse variational approximations Titsias and Lawrence (2010); Frigola et al. (2014) and obviating the need for sampling. In our model, which we call 'variational CGPDM' (vCGPDM), each MP is effectively parametrized by a small set of inducing points (IPs), leading to a compact representation. This compactness is important for real-world applicability of the model, since it results in a small memory footprint and significantly reduced learning time. While our target application here is human movement modeling, the vCGPDM could be easily applied to other systems where modularized control is beneficial, e.g. humanoid robotics. It was recently demonstrated Koch et al. (2015); Clever et al. (2016) that such modular MPs can indeed substantially simplify the control problem of a humanoid robot while yielding solutions that are not only feasible, but close to optimal.

The paper is organized as follows: in section 2, we briefly discuss related work before introducing the vCGPDM in section 3. An overview of the calculation of the variational lower bound on the model evidence (ELBO, evidence lower bound) is provided in section 4, the full derivation is contained in the supplementary material. In section 5, we present perceptual validation results with the objective of determining the degree of human-tolerable sparseness. We also show that perceptual performance can be predicted from the ELBO. Second, we demonstrate the modularity of the model by re-using learned MPs for the generation of novel movements.

2. Related work

The Gaussian process (GP) is a machine learning staple for classification and regression tasks Rasmussen and Williams (2005). Its advantages include theoretical elegance, tractability and closed-form solutions for posterior densities. Its main disadvantage are cubic run time scaling with the number of data points. A number of solutions have been proposed for this problem. Many of these involve a sparse representation of the posterior process via a small set of IPs Snelson and Ghahramani (2006). If the input space is unobserved, one obtains a GP latent variable model (GPLVM), for which sparse approximations have also been devised Lawrence (2007). One problem with sparse GP approximations is their tendency to overfit Lawrence (2007), leading to incorrect variance predictions which can be alleviated by a *variational* approximation Titsias and Lawrence (2010). This prompted us to develop a similar approach for the CGPDM.

The latent space of the GPLVM can be endowed with dynamics in discrete time. If the temporal evolution function is also drawn from a GP, the resulting model is called Gaussian Process Dynamical Model (GPDM) Wang et al. (2008). The GPDM can model the variability of human movements and has been used for computer animation with style control Urtasun et al. (2007); Taubert et al. (2012); Levine et al. (2012). It has also been used with an additional switching prior on the dynamics for motion tracking and recognition Chen et al. (2009). Inspired by the successes of the variational posterior approximation for the GPLVM, such approximations have also been developed for GPDM-like architectures Frigola et al. (2014) and even deep extensions thereof Mattos et al. (2016). However, with the exception of the coupled GPDM (CGPDM) Velychko et al. (2014), all these approaches have a 'monolithic' latent space(s) and thus lack the modularity of MPs.

In the following, we therefore introduce a variational approximation to CGPDM learning and inference based on an approach similar to Frigola et al. (2014), but we aim to obviate the need for sampling altogether to allow for fast, repeatable trajectory generation. While deriving a variational approximation is not trivial, we expect it to avoid overfitting and yield a good bound on the marginal likelihood Bauer et al. (2016).

3. The model

A vCGPDM is basically a number of GPDMs (the 'parts') run in parallel, with coupling between the latent space dynamics. The model operates in discrete time $t = 0, \ldots, T$. For every part $i = 1, \ldots, M$ there is a Q_i-dimensional latent space with second-order autoregressive dynamics and inputs from the latent spaces of the other parts. Let $\vec{x}_t^i \in \mathbb{R}^{Q_i}$ be the state of latent space i at time t. Then

$$\vec{x}_t^i = \vec{f}^i(\vec{x}_{t-2}^1, \vec{x}_{t-1}^1, \ldots, \vec{x}_{t-1}, ^M, \vec{x}_{t-1}^M). \tag{1}$$

We chose a second-order model, because our target application is human movement modeling, and the literature indicates (e.g. Taubert et al. (2012)) that this is a good choice for this task. However, we note that this can be easily changed in the model. The latent states $\vec{x}_t^i$ give rise to D_i-dimensional observations $\vec{y}_t^i \in \mathbb{R}^{D_i}$ via functions $\vec{g}^i(.)$ plus isotropic Gaussian noise η_t^i

$$\vec{y}_t^i = \vec{g}^i(\vec{x}_t^i) + \eta_t^i \tag{2}$$

The functions $\vec{g}^i(.)$ are drawn from a GP prior with zero mean function and a suitable kernel. In a vCGPDM, the functions $\vec{f}^i(\ldots)$ are also drawn from a GP prior with zero mean function, and and a kernel that is derived with product-of-experts (PoE, Hinton (1999)) coupling between the latent spaces of the different parts, as described by Velychko et al. (2014): each part generates a Gaussian prediction about every part (i.e. including itself). Let $\vec{x}_t^{i,j} = f^{i,j}(\vec{x}_{t-2}^i, \vec{x}_{t-1}^i)$ be the mean of the prediction of part i about part j at time index t, and $\alpha^{i,j}$ its variance. Following the standard PoE construction of multiplying the densities of the individual predictions and re-normalizing, one finds

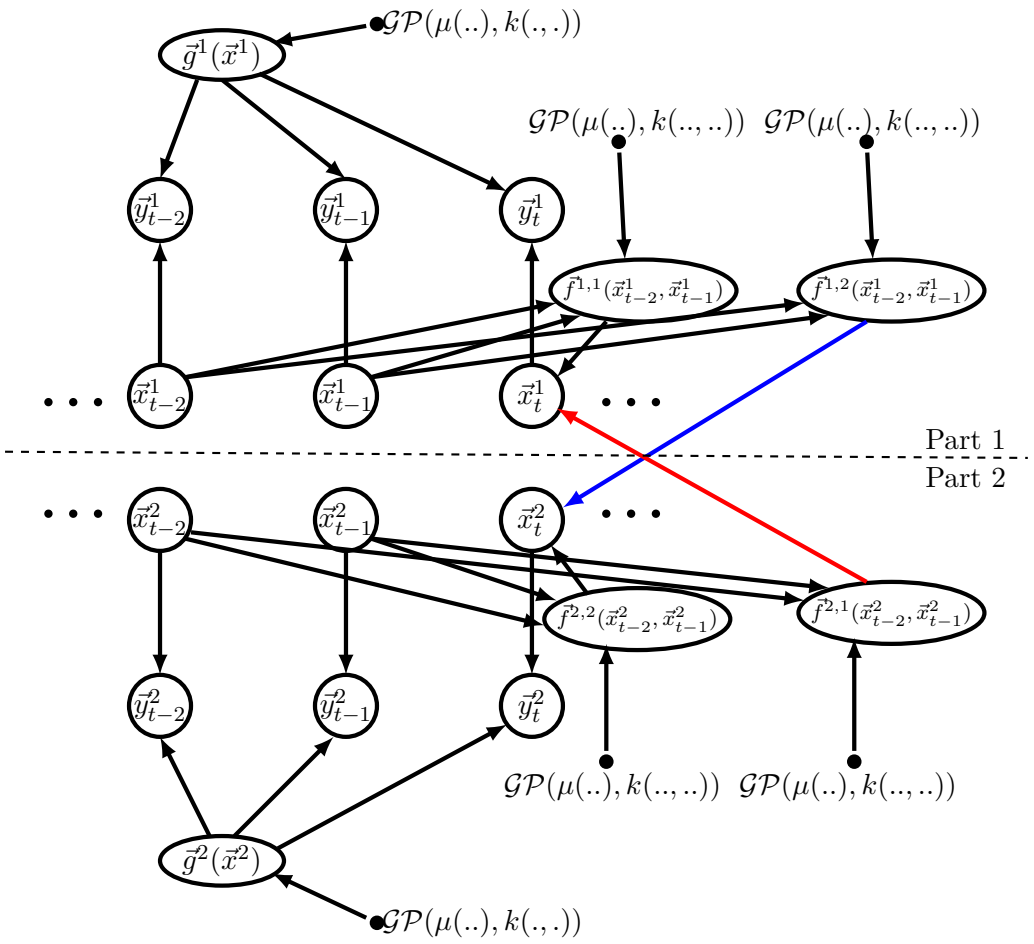

Figure 1: The graphical model representation of the variational Coupled Gaussian Process Dynamical Model (vCGPDM). Shown is a model with 2 parts, indicated by the superscripts $i, j \in \{1, 2\}$. Latent, vector-valued dynamical random variables $\vec{x}^i_t$ generate (vector-valued) observations $\vec{y}^i_t$ via functions $\vec{g}^i(\vec{x}^1_t)$ drawn from a $\mathcal{GP}$ Titsias and Lawrence (2010). Gaussian noise η is added to the observations, i.e. $\vec{y}^i_t = \vec{g}^i(\vec{x}^i_t) + \eta^i_t$ (not shown). The second-order mean dynamics functions $\vec{f}^{i,j}(\vec{x}^i_{t-2}, \vec{x}^i_{t-1})$, whose product-of-expert combination Hinton (1999) govern the temporal evolution of $\vec{x}^j_t$ are also drawn from $\mathcal{GP}$s. The blue and red connections indicate the coupling between the parts.

$$p(\vec{x}_t^j|\vec{x}_t^{\cdot,j},\alpha^{\cdot,j}) = \frac{\exp\left[-\frac{1}{2\alpha^j}\left(\vec{x}_t^j - \alpha^j \sum_i \frac{\vec{x}_t^{i,j}}{\alpha^{i,j}}\right)^2\right]}{(2\pi\alpha^j)^{\frac{\dim(X^j)}{2}}}$$
$$\propto \prod_i \mathcal{N}\left(\vec{x}_t^j|\vec{x}_t^{i,j},\alpha^{i,j}\right) \quad (3)$$

where $\dim(X^j)$ is the dimensionality of the latent space of part j and $\alpha^j = \left(\sum_i \alpha_{i,j}^{-1}\right)^{-1}$. It was shown in Velychko et al. (2014) that the individual predictions $\vec{x}_t^{i,j}$ can be marginalized out in closed form, if each of them is generated by a function drawn from a GP with mean zero and kernel $k^{i,j}(.,.)$. In contrast to that work, we will keep the individual predictions, because this allows us to couple a previously learned dynamics model for a part (including its predictions about the other parts) to any other dynamics model for the other parts, thus obtaining a modular MP model.

The form of eqn. 3 indicates the function of the coupling variances: the smaller a given variance, the more important the prediction of the generating part. When the $\alpha^{i,j}$ are optimized during learning, the model should be able to discover which couplings are important for predicting the data, and which ones are not (see section 5). Put differently, if an $\alpha^{i,j}$ is small compared to α^j, then part i is able to make a prediction about part j with high certainty, and vice versa. Furthermore, as demonstrated in Velychko et al. (2014), the $\alpha^{i,j}$ can be modulated after learning to generate novel movements which were not in the training data. Fig. 1 shows a graphical model representation with $M = 2$ parts (top and bottom half of figure).

The basic CGPDM exhibits the usual cubic run time scaling with the number of data points, which prohibits learning from large data sets. We therefore developed a sparse variational approximation, following the treatment in Titsias and Lawrence (2010); Mattos et al. (2016). We augment the model with IPs $\vec{r}^i$ and associated values $\vec{v}^i$ such that $g^i(\vec{r}^i) = \vec{v}^i$ for the latent-to-observed mappings $g^i(X^i)$, and condition the probability density of the function values of $g^i(X^i)$ on these points/values, which we assume to be a sufficient statistic. We apply the same augmentation strategy to reduce the computational effort for learning the dynamics mappings, which are induced by $\vec{z}^{i,j}$ and $\vec{u}^{i,j}$.

Key assumption: to obtain a tractable variational posterior distribution q over the latent states $\vec{x}_t^i = (x_{t,1}^i, \ldots, x_{t,Q^i}^i)$, we choose a distribution that factorizes across time steps $0, \ldots, T$, parts $1, \ldots, M$ and dimensions $1, \ldots, Q^i$ within parts, and assume that the individual distributions are Gaussian:

$$q(\vec{x}_0^1, \ldots, \vec{x}_T^M) = \prod_{t=0}^{T}\prod_{i=1}^{M}\prod_{q=1}^{Q^i} q(\vec{x}_{t,q}^i) \quad (4)$$
$$q(\vec{x}_{t,q}^i) = \mathcal{N}(\mu_{t,q}^i, \sigma_{t,q}^{2,i}). \quad (5)$$

This approximation assumption is clearly a gross simplification of the correct latent state posterior. However, it allows us to make analytical progress: an ELBO will be a sum of

two main contributions. First, the latent-to-observed component of the model is now a variational GPLVM as described by Titsias and Lawrence (2010). Hence, the results of that paper can be reused without alteration. Second, the dynamics component. We can derive closed-form expressions for the required integrals for certain kernels $k^i\left(\vec{X}, \vec{X}'\right)$. Specifically, we use an ARD (automatic relevance detection) squared exponential kernel Bishop (2006) for every part-i-to-j prediction GP:

$$k^{i,j}(\vec{X}, \vec{X}') = \exp\left(-\frac{1}{2}\sum_{q}^{Q_i}\frac{(\vec{X}_q - \vec{X}'_q)^2}{\lambda_q^{i,j}}\right). \tag{6}$$

The computations required for this are lengthy (and error-prone) but straightforward, and resemble those for the variational GPLVM of Titsias and Lawrence (2010). We outline the computation in the next section, details can be found in section 2 of the supplementary material. Whether our simplistic approximation assumption (eqn. 4) is useful depends on the data, but at least for human movement it seems appropriate (see section 5).

4. Computing a variational lower bound for the dynamics model

Since we work with a second order dynamics, we denote the last two time steps $\vec{x}_{-t} = \{\vec{x}_{t-2}, \vec{x}_{t-1}\}$. Superscripts refer to parts, subscripts to time indexes $t = 0, \ldots T$ and dimensions d. Note that an extension to higher order dynamics is relatively easy, by redefining $\vec{x}_{-t}$ accordingly. As mentioned above, we construct a sparse variational approximation by augmenting each of the $M \times M$ dynamics mappings $f^{j,i}$ IPs $\vec{z}^j$ and values $\vec{u}^{j,i}$. Likewise, the observation function of each part g^i is augmented with inputs $\vec{r}^i$ and outputs $\vec{v}^i$.

The variational distribution for $q(\vec{x}_t^i)$ factorizes (eqn. 4), whereas $q(\vec{u}^i)$ and $q(\vec{v}^i)$ are unconstrained distributions. The joint density of the augmented model factorizes as

$$p(\vec{x}, \vec{u}, \vec{f}, \vec{v}, \vec{g}, \vec{y}) = p(\vec{y}|\vec{g})p(\vec{g}|\vec{x}, \vec{v})p(\vec{v})p(\vec{x}, \vec{f}|\vec{u})p(\vec{u}) \tag{7}$$

where omission of indexes or slice notation signifies collections of variables, e.g. $\vec{x} = \{\vec{x}_0^1, \ldots, \vec{x}_T^M\}$; $\vec{x}_:^i = \{\vec{x}_0^i, \ldots, \vec{x}_T^i\}$. In eqn. 7,

$$p(\vec{y}|\vec{g}) = \prod_{i=1}^{M}\prod_{d=1}^{D_i} p(\vec{y}_d^i|\vec{g}_d^i) \tag{8}$$

$$p(\vec{g}|\vec{x}, \vec{v}) = \prod_{i=1}^{M}\prod_{d=1}^{D_i} p(\vec{g}_d^i|\vec{x}^i, \vec{v}^i) \tag{9}$$

$$p(\vec{v}) = \prod_{i=1}^{M} p(\vec{v}^i); \;\; p(\vec{u}) = \prod_{i=1}^{M}\prod_{j=1}^{M} p(\vec{u}^{j,i}). \tag{10}$$

The density of the latent variables and the individual parts' predictions thereof is:

$$p(\vec{x}, \vec{f}|\vec{u}) = p(\vec{x}|\vec{f}, \vec{u})p(\vec{f}|\vec{u}) \tag{11}$$

with

$$p(\vec{x}|\vec{f},\vec{u}) = \prod_{t=1}^{T}\prod_{i=1}^{M} p(\vec{x}_t^i|\{\vec{f}_t^{:,i},\alpha^{:,i}\}) \tag{12}$$

$$= \prod_{t=1}^{T} p(\vec{x}_t|\{\vec{f}_t,\alpha^{:,:}\}) \tag{13}$$

$$p(\vec{f}|\vec{u}) = \prod_{t=1}^{T}\prod_{i=1}^{M}\prod_{j=1}^{M} p(\vec{f}_t^{j,i}|\vec{f}_{1:t-1}^{j,i},\vec{x}_{0:t-1}^{j},\vec{u}^{j,i}) \tag{14}$$

$$= \prod_{t=1}^{T} p(\vec{f}_t|\vec{f}_{1:t-1},\vec{x}_{0:t-1},\vec{u}) \tag{15}$$

where an empty slice ($t < 2$) implies no conditioning. The first two latent states at $t = 0, 1$ are drawn from independent Gaussians, $\prod_{i=1}^{M} p(\vec{x}_0^i)p(\vec{x}_1^i)$. Eqn. 12 is given by the PoE construction (eqn. 3).

We choose the following proposal variational posterior:

$$q(\vec{x},\vec{u},\vec{f},\vec{v},\vec{g}) = p(\vec{g}|\vec{x},\vec{v})q(\vec{v})p(\vec{f}|\vec{x},\vec{u})q(\vec{x})q(\vec{u}) \tag{16}$$

with $p(\vec{g}|\vec{x},\vec{v})$ given by eqn. 9 and $p(\vec{f}|\vec{x},\vec{u})$ by eqn. 11. The densities $q(\vec{v})$ and $q(\vec{u})$ are assumed to be multivariate Gaussian.

With these distributions, we derive the standard ELBO Bishop (2006), denoting $\vec{\theta} = (\vec{x},\vec{u},\vec{f},\vec{v},\vec{g})$:

$$\log p(\vec{y}) \geq \mathcal{L}(\vec{\theta}) = \int d\vec{\theta} q(\vec{\theta}) \log\left(\frac{p(\vec{y},\vec{\theta})}{q(\vec{\theta})}\right) \tag{17}$$

using the assumption that the IPs and values $\vec{r}^i, \vec{v}^i$ are a sufficient statistic for the function values $\vec{g}^i$ for every part i and exploiting the fact that the variational posterior (eqn. 16) and the joint model density (eqn. 7) contains common factors (cf. Titsias and Lawrence (2010)). Thus

$$\mathcal{L}(\vec{\theta}) = \sum_{i=1}^{M} \mathcal{L}_{kin}^i + \mathcal{L}_{dyn} \tag{18}$$

where

$$\mathcal{L}_{kin}^i = \sum_{d=1}^{D} \int_{\vec{x}^i,\vec{v}^i,\vec{g}_d^i} p(\vec{g}_d^i|\vec{x}^i,\vec{v}^i)q(\vec{x}^i)q(\vec{v}^i)\log\frac{p(\vec{y}_d^i|\vec{g}_d^i)}{q(\vec{v}^i)}. \tag{19}$$

is similar to the Bayesian GPLVM bound of Titsias and Lawrence (2010). The remaining integral

$$
\begin{aligned}
&\mathcal{L}_{dyn} = \\
&+ \int q(\vec{u}) \left[\sum_{t=1}^{T} \int q(\vec{x}_{1:t}) \left(\int p(\vec{f}_t | \vec{f}_{1:t-1}, \vec{x}_{0:t-1}, \vec{u}) \right. \right. \\
&\left. \left. \times \log p(\vec{x}_t | \{\vec{f}_t, \alpha^{\cdot,\cdot}\}) d\vec{f}_t \right) d\vec{x}_{1:t} \right] + q(\vec{u}) \log \frac{p(\vec{u})}{q(\vec{u})} d\vec{u} \\
&+ \int q(\vec{x}_0) \log p(\vec{x}_0) d\vec{x}_0 + H(q(\vec{x}))
\end{aligned} \tag{20}
$$

is derived in the supplementary material. Briefly, we again use the assumption that the IPs and values $\vec{z}^{j,i}$ and $\vec{u}^{j,i}$ are a sufficient statistic for the function values $\vec{f}_t^{j,i}$. We write $\vec{K}^{j,i}_{\vec{z}^{j,i},\vec{z}^{j,i}}$ for the kernel matrix computed from the IPs of the dynamics/coupling GPs. Optimizing with respect to $q(\vec{u}^i)$ can be carried out in closed form using variational calculus and yields

$$\mathcal{L}_{dyn}(\vec{\theta}) \geq \log \int p(\vec{u}) \exp(\mathcal{C}) d\vec{u} + H(q(\vec{x})) \tag{21}$$

$$q(\vec{u}^i) = \mathcal{N}(\vec{u}^i | \vec{\mu}^i, \vec{\Sigma}^i) \tag{22}$$

$$\vec{\mu}^i = (\vec{K}^{-1}_{\vec{z}^{\cdot,i},\vec{z}^{\cdot,i}} + \mathcal{F}^i)^{-1} \mathcal{G}^i \tag{23}$$

$$\vec{\Sigma}^i = (\vec{K}^{-1}_{\vec{z}^{\cdot,i},\vec{z}^{\cdot,i}} + \mathcal{F}^i)^{-1} \tag{24}$$

The inequality on the first line is due to the sufficient statistics assumption, which introduces another approximation step that lower-bounds $\mathcal{L}_{dyn}$. The matrices $\mathcal{F}^i$ and $\mathcal{G}^i$ contain kernel statistics, which are derived in the appendix, likewise the expressions for $\mathcal{C}$. We now have all the ingredients to compute the ELBO for the whole model.

5. Results

We implemented the model in `Python 2.7` using the machine-learning framework `Theano` Bastien et al. (2012) for automatic differentiation to enable gradient-based maximization of the ELBO with the `scipy.optimize.fmin_l_bfgs_b` routine Jones et al. (2001–). Latent space trajectories were initialized with PCA.

While the sparse approximations in the vCGPDM greatly reduce the memory consumption of the model, they might also introduce errors. Also, our fully factorized latent posterior approximation (eqn. 4) might be too simple. We tried to quantify these errors in a cross-validatory model comparison, and in a human perception experiment.

5.1 Human movement data

Comparisons were carried out on human movement data. We recorded these data with a 10-camera PhaseSpace Impulse motion capture system, mapped them onto a skeleton with 19 joints and computed joint angles in angle-axis representation, yielding a total of 60 degrees of freedom. The actors were instructed to walk straight with a natural arm swing, and to walk while waving both arms. Six walking-only and walking+waving sequences each were used to train the models.

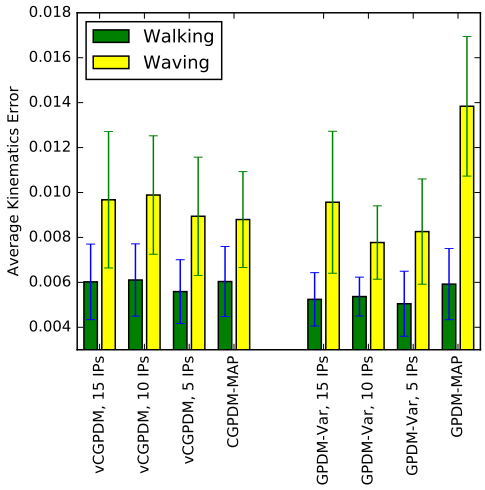

Figure 2: Model comparison results. Shown is the average squared kinematics error after dynamic time warping. The vCGPDM and variational GPDM are at least as good or even better than the full capacity CGPDM and GPDM models with maximum-a-posterior (MAP) estimation of latent space trajectories. IPs: numbers of inducing points for both dynamics and kinematics mapping. Error bars are standard errors of the mean. For details, see text.

5.2 MAP is worse than variational approximation

To check how predictive quality for both dynamics and kinematics (pose) is affected by our sparse variational approximation, we conducted a comparison by five/four-fold cross-validation of the following models for walking/waving: 1.) a GPDM with maximum-a-posteriori (MAP) estimation of the latent variables Wang et al. (2008), called GPDM-MAP in fig. 5.2. 2.) a fully marginalized two-part (upper/lower body) CGPDM with MAP estimation of the latent variables Velychko et al. (2014), called CGPDM-MAP. 3.) Their variational counterparts, two-part vCGPDM and GPDM-Var, for five, ten and 15 inducing points (IPs) for both the dynamics (latent-to-latent) and the kinematics (latent-to-observed) mapping. The variational GPDM is implemented as a vCGPDM with one part. All latent spaces were three-dimensional. Cross-validation scores were dynamics and kinematics error, computed in the following way: we generated trajectories by initializing the model with the first two states of a held-out trial. Then, the model was run forward, predicting the mean trajectory. This generated trajectory was mapped onto the actual held-out trajectory using dynamic time warping Sakoe and Chiba (1978). The kinematics error in 5.2 is the remaining mean squared error. Generally, all models perform better on walking only dataset, than on walking+waving. This might be due to the arms sometimes being out of sync with the legs during walking and waving, as can be seen in the movie `modular_primitives.mov` in the supplementary data. Comparing the full-capacity (no IPs) model, we find that the CGPDM-MAP gives better predictions than the GPDM-MAP on both data sets. Somewhat

surprisingly, all variational models are at least as good (within the error bars) as the CGPDM-MAP, which indicates that a relatively small number of IPs is enough to model these data – $\mathcal{O}(10)$ inducing points compared to $\mathcal{O}(10^4)$ data points which have to be stored for full capacity models. This also shows that our fully factorized latent state posterior approximation, eqn. 4 is not too drastic an approximation for human movement data, a conclusion which is furthermore supported by the fact that the two-part vCGPDM does not overfit the data compared to the GPDM-Var, even though the former has two latent spaces.

5.3 A small number of IPs is enough to fool human observers

Next, we wanted to investigate the number of inducing points needed for perceptually plausible movements. Six walking sequences were used to train vCGPDMs with 2 parts (upper/lower body), between 4-16 IPs for the kinematics mapping and 2-16 IPs for the dynamics and latent-to-latent mappings. We generated walking sequences from the vCGPDMs by initializing them with starting conditions taken from the training data. Furthermore, we recorded another 9 walking sequences for catch trials during the perception experiment, to rule out memorization effects. Generated and recorded sequences were rendered on a neutral avatar. Examples of stimuli, for different numbers of IPs, can be found in the movie `example_stimuli.mov` in the supplementary material.

Experiment: in a 2-AFC task, human observers were presented simultaneously with videos of natural and generated movements. The duration of an individual stimulus was 1.8 seconds, with a total of 1170 presentations. After the stimulus presentation participants were asked choose which movement they perceived as more natural. 31 participants (10 male, mean age: 23.8 ± 3.5a) participated in this experiment. To test whether participants simply memorized the six natural stimuli during the experiment, we added 10 catch trials in the last quarter of the experiment where previously unknown natural movements were tested against the known natural stimuli. The trial sequence was randomized for every subject. All experimental procedures were approved by the local ethics commission.

Results: we computed the frequency f_{gen} of choosing the CGPDM-generated movement across all subjects as a function of the number of IPs for the dynamics/latent-to-latent mappings and the number of IPs for the latent-to-observed mapping ('GPLVM'), see fig. 3, A. At best, we might expect f_{gen} to approach $\frac{1}{2}$ when the generated movements are indistinguishable from the natural ones. We also fitted those data with a logistic sigmoid $\frac{1}{1+\exp(a \cdot r(.)+c)}$ and a Bernoulli observation model, using three different regressor functions $r(.)$: a soft-minimum between the number of IPs, the ELBO (eqn. 18) and a linear combination of the kinematics $\mathcal{L}_{kin}$ and the dynamics ELBO part $\mathcal{L}_{kin}$, called sELBO. Panels C & D depict the sELBO regressors, panel B the fit. Panel E shows 107-fold crossvalidation results for the three regressors, using the average negative log-probability on the held-out data as score. Error bars are standard deviations. 'Constant' is the constant regressor, any other regressor should predict better. 'Data' uses the data mean of the individual #IP combinations as a predictor, and constitutes a lower bound on the cross validation score.

Clearly, f_{gen} increases with the number of IPs, approaching (but not quite reaching) 0.5 for a sufficiently large number of IPs, this is true for the sELBO regression, too. Thus, a rather small number of IPs is sufficient for modeling this data. This allows for compactly parametrized MPs. The cross-validation data indicate that prediction of subjects'

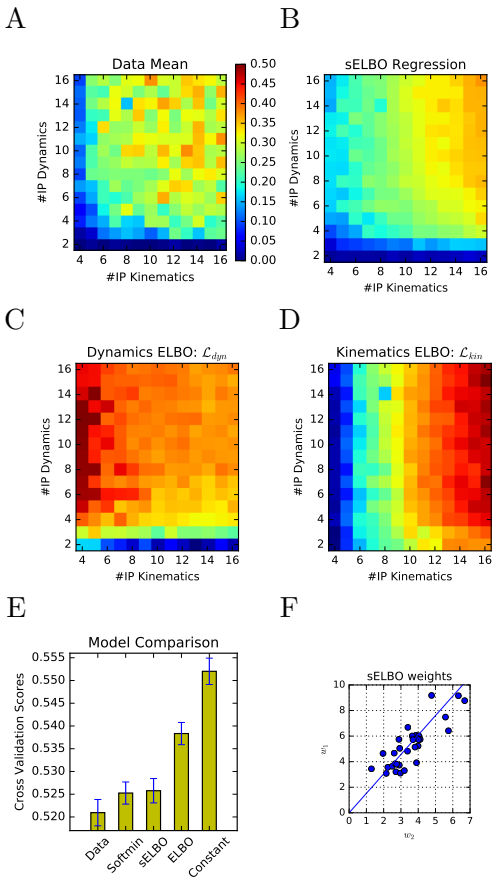

Figure 3: Perceived naturalness of the model, as a function of the number of inducing points (#IP) **A**: Rate of perceiving vCGPDM-generated stimulus as more natural than natural stimulus, averaged across all participants. **B**: Regression of data in panel A, with $\mathcal{L}_{dyn}$ and $\mathcal{L}_{kin}$ as regressors and logistic sigmoid as psychometric function **C & D**: $\mathcal{L}_{dyn}$ and $\mathcal{L}_{kin}$ regressors. Color scale from cold (low) to hot (high), but not the same as in panels A,B. **E**: Regression model comparison with 107-fold cross-validation. Softmin and sELBO perform comparably well, ELBO is significantly worse. **F**: Weights of dynamics (w_1) and kinematics (w_2) in the sELBO regression. The line indicates the ratio w_1/w_2 averaged across subjects.

performance is better, and almost optimal, when using the sELBO as regressors, rather than ELBO. The average ratio between dynamic/kinematic regression weights of $w_1/w_2 = 1.52$ indicates a stronger influence of dynamics for human perceptual performance. We did not find evidence for stimulus memorization effects during the catch trials.

5.4 Modularity test

Next, we examined if the intended modularization of our model can be used to compose novel movements from previously learned parts. To this end, we recorded motion capture

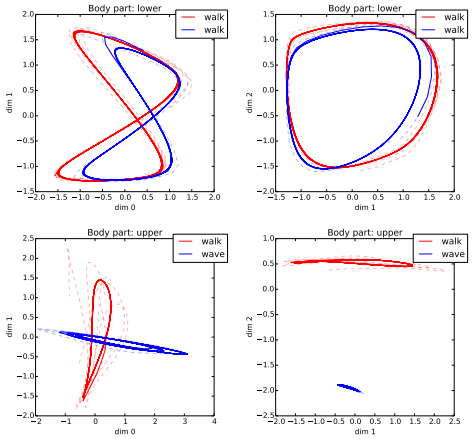

Figure 4: Modularity example. Shown are the first two dimensions of generated 3-dimensional latent space trajectories (solid) and training data (dashed). Blue: walk + wave movements, red: walk + normal arm swing. Dynamics IPs re-used across movements for lower body. For details, see text.

data from one human participant who executed a walk with normal arm swing, and a walk with two-handed waving. We trained a vCGPDM consisting one part for the lower body (below and including pelvis), and a second part for the upper body. 25 IPs for the latent-to-observed mapping of each part were shared across all movements. The walking MP, parametrized by 16 IPs for the lower-body dynamics and the the lower-to-upper mappings, was also shared. We used a different set of 16 IPs for the upper body MPs between arm-swing and waving. Furthermore, the coupling $\alpha^{m,n}$s, were also learned anew for each combination of upper/lower MPs. The resulting latent space trajectories are plotted in fig. 5.4. All generated trajectories (solid lines) are on average close to the training data (dashed lines). While the walking trajectories for the lower body are very similar for the two movements, the upper body trajectories clearly differ. Movements generated from this model are very natural (see movie `modular_primitives.mov` in the suppl.mat.). This is a first demonstration that the vCGPDM with non-marginalized couplings can be used to learn a library of compactly parametrized MPs, from which novel movements can be produced with minimal re-training effort (i.e. learning the couplings only).

6. Conclusion

We developed a full variational approximation of the CGPDM, the vCGPDM, which obviates the need for sampling the latent space trajectories. We demonstrated that the vCGPDM with a small number of IPs performs equally well as the full-capacity CGPDM. Next, we showed that it produces perceptually believable full-body movements. While perceptual evaluations of full and sparse GPDM-like models Taubert et al. (2012, 2013) have been done before, we are the first to investigate systematically the number of IPs of all model components required for perceptual plausibility. Furthermore, we showed that the parts

of the ELBO (eqn. 18) and the number of IPs can be used to predict average human classification performance almost optimally. This indicates that the model selection process on large databases of training movements for the model could possibly be automated. Within the range of IPs which we tested, the ELBO was still increasing with their number. We chose that range because we wanted to see how few IPs would still lead to perceptually indistinguishable movements. Due to experimental time constraints, we did not investigate perceptual performance at the point where the ELBO begins to decrease with increasing IPs (i.e. the approximately optimal model), but we plan to do that in the future.

Second, we showed that the model can be employed in a modular fashion, using one lower-body dynamics model, and coupling it to two different models for the upper body. Note that the lower-to-upper coupling function was the same for the two upper-body models. Each of these models, including the coupling functions to the other model parts, may therefore be viewed as a modular MP that is parametrized compactly by a small number of IPs and values. While our modularity demonstration can only be regarded as a very first step, we are now in a position to learn a large library of movements with a CGPDM, and study its compositionality. This is possible due to the compact representation of each MP. To generate complex movement sequences, we will put a switching prior on top of the dynamical models, like Chen et al. (2009). We then plan to evaluate if the extra flexibility afforded by the GP based dynamical models is necessary by comparing our approach directly to DMPs on a large dataset of natural (human) movements. Instead of direct connections between parts in the vCGPDM, it is also conceivable to embed the parts into a hierarchical architecture, like Lawrence and Moore (2007); Taubert et al. (2012). While the vCGPDM is suitable when the number of parts is relatively small (computational complexity $\mathcal{O}(T * M * (M * \#IP)^3)$ per optimization iteration), a hierarchical architecture might enable more computational savings for many parts.

A further direction of future research are *sensorimotor* primitives, i.e. MPs that can be conditioned on sensory input. This conditioning can take place on at least two timescales: a short one (while the MP is running), thus effectively turning the MPs into flexible control policies, like the probabilistic MPs described by Paraschos et al. (2013) or Mattos et al. (2016). And a long timescale, i.e. the planning of the movement. This could be implemented by learning a mapping from goals and affordances onto the coupling weights, comparable to the DMPs with associative skill memories Pastor et al. (2012).

Acknowledgments

We would like to acknowledge funding from DFG under IRTG 1901 'The Brain in Action', SFB-TRR 135 project C06, and the European Union Seventh Framework Program (FP7/2007 - 2013) under grant agreement no 611909 (KoroiBot). We thank Olaf Haag for help with rendering the movies, and Björn Büdenbender for assistance with MoCap.

References

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

M.S. Bauer, M. van der Wilk, and C.E. Rasmussen. Understanding probabilistic sparse gaussian process approximations. Technical report, University of Cambridge, UK, 2016. arXiv:1606.04820.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

E. Bizzi, V.C.K. Cheung, A. d'Avella, P. Saltiel, and M. Tresch. Combining modules for movement. *Brain Research Reviews*, 57(1):125 – 133, 2008. ISSN 0165-0173. doi: http://dx.doi.org/10.1016/j.brainresrev.2007.08.004. URL `http://www.sciencedirect.com/science/article/pii/S0165017307001774`. Networks in Motion.

Jixu Chen, Minyoung Kim, Yu Wang, and Qiang Ji. Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2655–2662, June 2009. doi: 10.1109/CVPR.2009.5206580.

D. Clever, M. Harant, K. H. Koch, K. Mombaur, and D. M. Endres. A novel approach for the generation of complex humanoid walking sequences based on a combination of optimal control and learning of movement primitives. *Robotics and Autonomous Systems*, 2016. doi: 10.1016/j.robot.2016.06.001.

Roger Frigola, Yutian Chen, and Carl Rasmussen. Variational gaussian process state-space models. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3680–3688. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5375-variational-gaussian-process-state-space-models.pdf`.

M. A. Giese, A. Mukovskiy, A.-N. Park, L. Omlor, and J.-J. E. Slotine. Real-Time Synthesis of Body Movements Based on Learned Primitives. *In Cremers D, Rosenhahn B, Yuille A L (eds): Statistical and Geometrical Approaches to Visual Motion Analysis, Lecture Notes in Computer Science*, 5604:107–127, 01 2009. URL `pub/pdf/LNCS5604proc2009.pdf`.

Geoffrey E. Hinton. Products of experts. In *Proc. ICANN'99*, volume 1, pages 1–6, 1999.

Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. Dynamical movement primitives: Learning attractor models for motor behaviors. *Neu. Comp.*, 25(2):328–373, 2013.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL `http://www.scipy.org/`. [Online; accessed 2015-10-09].

K. H. Koch, D. Clever, K. Mombaur, and D. M. Endres. Learning movement primitives from optimal and dynamically feasible trajectories for humanoid walking. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids 2015)*, pages 866–873, 2015.

Neil D. Lawrence. Learning for larger datasets with the gaussian process latent variable model. In Marina Meila and Xiaotong Shen, editors, *AISTATS*, volume 2 of *JMLR Proceedings*, pages 243–250. JMLR.org, 2007.

Neil D. Lawrence and Andrew J. Moore. Hierarchical gaussian process latent variable models. In Zoubin Ghahramani, editor, *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 481–488. ACM, 2007. ISBN 978-1-59593-793-3.

S.-H. Lee, E. Sifakis, and D. Terzopoulos. Comprehensive biomechanical modeling and simulation of the upper body. *ACM Trans. Graph.*, 28(4):99, 2009.

Sergey Levine, Jack M. Wang, Alexis Haraux, Zoran Popović, and Vladlen Koltun. Continuous character control with low-dimensional embeddings. *ACM Trans. Graph.*, 31(4):28, 2012.

César Lincoln C. Mattos, Zhenwen Dai, Andreas Damianou, Jeremy Forth, Guilherme A. Barreto, and Neil D. Lawrence. Recurrent gaussian processes. Technical report, University of Sheffield, 2016. arXiv:1511.06644.

Alexandros Paraschos, Christian Daniel, Jan Peters, and Gerhard Neumann. Probabilistic movement primitives. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2616–2624. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/5177-probabilistic-movement-primitives.pdf`.

P. Pastor, M. Kalakrishnan, L. Righetti, and S. Schaal. Towards associative skill memories. In *IEEE-RAS International Conference on Humanoid Robots*, page 7, 2012.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

E. Rückert and A. d'Avella. Learned parametrized dynamic movement primitives with shared synergies for controlling robotic and musculoskeletal systems. *Frontiers in Computational Neuroscience*, 7(138), 2013.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, Feb 1978. ISSN 0096-3518. doi: 10.1109/TASSP.1978.1163055.

Charles S. Sherrington. *The Integrative Action of the Nervous System*. Yale University Press, 1906.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT press, 2006.

N. Taubert, A. Christensen, D. Endres, and M.A. Giese. Online simulation of emotional interactive behaviors with hierarchical gaussian process dynamical models. In *Proceedings of the ACM Symposium on Applied Perception*, pages 25–32. ACM, 2012. doi: 10.1145/2338676.2338682.

N. Taubert, M. Löffler, N. Ludolph, A. Christensen, D. Endres, and M.A. Giese. A virtual reality setup for controllable, stylized real-time interactions between humans and avatars with sparse Gaussian process dynamical models. *Proceedings of the ACM Symposium on Applied Perception*, pages 41–44, 2013. doi: 10.1145/2492494.2492515. URL `http://doi.acm.org/10.1145/2492494.2492515`.

Michalis K. Titsias and Neil D. Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 844–851, 2010. URL `http://www.jmlr.org/proceedings/papers/v9/titsias10a.html`.

Raquel Urtasun, David J. Fleet, and Neil D. Lawrence. Modeling human locomotion with topologically constrained latent variable models. In *Workshop on Human Motion, LNCS*, volume 4814, pages 104–118, 2007.

D. Velychko, D. Endres, N. Taubert, and M. A. Giese. Coupling Gaussian process dynamical models with product-of-experts kernels. In *Proceedings of the 24th International Conference on Artificial Neural Networks, LNCS 8681*, pages 603–610. Springer, 2014. doi: 10.1007/978-3-319-11179-7_76.

Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):283–298, 2008.