

Coupling Gaussian Process Dynamical Models with Product-of-Experts Kernels

Dmytro Velychko, Dominik Endres*, Nick Taubert, Martin A. Giese*

Section Computational Sensomotrics, Department of Cognitive Neurology,
University Clinic Tübingen, CIN, HHH and University of Tübingen,
Otfried-Müller-Str. 25, 72076 Tübingen, Germany
dmytro.velychko@gmail.com
{nick.taubert,dominik.endres,martin.giese}@uni-tuebingen.de

Abstract. We describe a new probabilistic model for learning of coupled dynamical systems in latent state spaces. The coupling is achieved by combining predictions from several Gaussian process dynamical models in a product-of-experts fashion. Our approach facilitates modulation of coupling strengths without the need for computationally expensive re-learning of the dynamical models. We demonstrate the effectiveness of the new coupling model on synthetic toy examples and on high-dimensional human walking motion capture data.

Keywords: Gaussian Process, Products of Experts, Computer Graphics

1 Introduction

Mathematical models of dynamical systems are used in many fields of science. For example, coordinated motor patterns have been accounted for by networks of coupled dynamic movement primitives or 'central pattern generators' [7]. We are primarily concerned with the modeling of human motion data for experiments in psychophysics and neuroscience, but our approach lends itself naturally to applications in computer graphics and robotics.

While whole-body human motion data is high-dimensional, the intrinsic dimensionality is usually much smaller. Applying dynamical models to such data directly often results in poor generalization abilities, e.g. when one wants to vary parameters affecting the dynamical coupling strength between body parts. Therefore, a dimensionality reduction component is usually part of such a model. Lawrence [11] introduced a new probabilistic, non-linear dimensionality reduction method, the Gaussian process latent variable model (GPLVM), which is based on Gaussian processes (GP). A GP can be obtained from a neural network with a particular prior on the weights and biases in the limit of infinitely many hidden units [16]. The GPLVM was extended by a latent dynamics in [23] resulting in the Gaussian process dynamical model (GPDm). Due to its

* equal contribution

2 D. Velychko, D. Endres, N. Taubert, M.A. Giese

probabilistic nature, it is well equipped to handle the variability of natural motion data. It is possible to model full-body human motion with just one GPDM [21, 23, 20]. However, such 'monolithic' motion models do not allow for modulation of parts of the dynamics (e.g. different motion styles of body parts for movement design in computer graphics) or for recombining previously learned component dynamics for complex coordinated movements. Such a recombination would allow us to construct a rich repertoire of full-body movements from a much smaller set of dynamical primitives. We therefore present an approach to coupling GPDMs based on a product-of-experts (PoE) [9] construction. PoE results in less uncertain overall predictions than any of its parts, which is conducive to stability. Furthermore, we can then modulate the coupling strengths after learning, *without* costly re-training of the components.

We briefly review related work in section 2, and introduce the model's building blocks in section 3. The main theoretical development of this paper, the product-of-experts kernel is derived in section 4. Section 5 presents results on illustrative toy examples, and on human locomotion data.

2 Related work

There are, broadly speaking, two approaches for learning of dynamical systems: as a deterministic system of differential equations (see e.g. [10, 5]), and statistical approaches, where the evolution of a system is described in terms of a probabilistic mapping from the previous state to the next, for example [3, 22, 8, 4]. Both approaches may be augmented with deep hierarchies.

Deterministic systems based on differential equations have to be carefully designed and tuned. Even though the theory of learning of complex dynamical systems for motion synthesis is in active development, and some sophisticated applications of it for robotics, computer graphics and neuroscience exist [12, 10, ?], the nature of nonlinear dynamical systems makes it hard to design and train such models. One approach for their design is contraction theory [14], which allows for the construction of dynamically stable systems from stable components [1, 15].

On the other hand, probabilistic approaches promise to capture the variability of human motion and its styles [22, 8, 20]. Ease of learning and manipulating of the parameters are crucial advantages for such applications as psychophysical experiments in emotions perceptions [20], computer graphics [13] and human locomotion modeling [21]. However, a stability analysis of these models is non-trivial, and has not been accomplished to date.

3 Model components

Gaussian Process Latent Variable Model (GPLVM). A GPLVM comprises a prior on mappings $f_Y(X)$ from a (possibly vector-valued) latent variable X onto observable variables Y [11]. The $f_Y(X)$ is drawn from a Gaussian

Process (GP) prior, parametrized by a mean function (constant zero in this paper) and kernel (covariance function) $k(X, X')$. Furthermore Y may be corrupted by additive Gaussian noise with standard deviation β , see fig. 1, left. The prior on X is typically an isotropic Gaussian, too, but may be replaced by predictions from a higher level model in a hierarchical architecture [20].

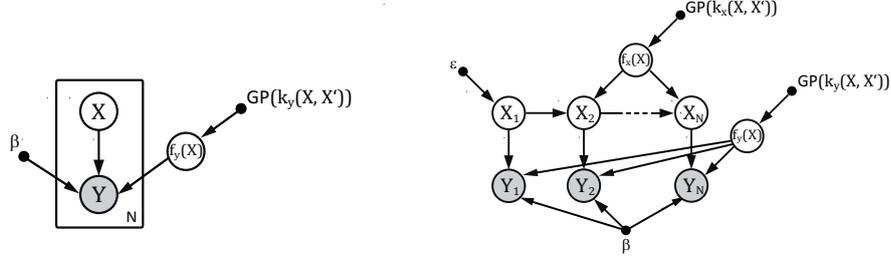


Fig. 1. Left: Graphical model representation of a GPLVM, which is a prior on functions from a (low-dimensional) latent space X to a (high-dimensional, observed) space Y [11]. **Right:** in a GPDM, a Markov chain in the latent space X models the dynamics of the observed data Y [23]. For details, see text.

Assume $D = \dim(Y)$, $q = \dim(X)$ and that we had observed N instances \mathbf{y}_i of Y . We index component d of instance i as $y_{d,i}$. We also use slice notation for arrays, e.g. $\tilde{\mathbf{y}} = y_{d,:}$ denotes the vector comprised of all instances of component d , whereas $\mathbf{y}_i = y_{:,i}$. We write $p(\mathbf{y})$ as a short-hand for $p(Y = \mathbf{y})$. Like in [11], we learn the GPLVM by maximizing the joint posterior density of the corresponding $\mathbf{x}_:$. Since all finite-dimensional marginals of a GP are multivariate Gaussian with density $\mathcal{N}(\mathbf{y}|\mu, \Sigma)$, and the components of a vector-valued GP are independent, it follows that the likelihood of $y_{:,d}$ is given by

$$p(y_{:,d}|k_Y(X, X'), \beta, \mathbf{x}_:) = \mathcal{N}(y_{:,d}|\mathbf{0}_N, \mathbf{K}_Y + \beta^2 \mathbf{1}_{N,N}) \quad (1)$$

where \mathbf{K}_Y is the kernel matrix, with $(\mathbf{K}_Y)_{i,j} = k_Y(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{1}_{N,N}$ is the N -dimensional identity matrix. Thus, the total posterior of the latent variables is proportional to

$$p(\mathbf{x}_:|\mathbf{y}_:, k(X, X')) \propto \prod_d p(y_{:,d}|k(X, X'), \beta, \mathbf{x}_:) \prod_i p(\mathbf{x}_i) \quad (2)$$

which can be optimized by standard non-linear methods; we use [19].

Gaussian Process Dynamical Model (GPDM). Wang [23] extended the GPLVM with a dynamical auto-regressive prior on the latent X_i , where the data point index i now denotes discrete time. The evolution function $f_X(X)$ of this dynamics is drawn from a GP with kernel $k_X(X, X')$ and $\mathbf{x}_{i+1} = f_X(\mathbf{x}_i) + \eta$; $\eta \sim \mathcal{N}(\mathbf{0}_q, \xi^2 \mathbf{1}_{q \times q})$. This approach leads to a non-linear, continuous generalization of

4 D. Velychko, D. Endres, N. Taubert, M.A. Giese

a hidden Markov model. For a first-order dynamics, one obtains (cf. fig. 1, right):

$$p(\mathbf{x}_i | k_X(X, X'), \xi, \epsilon) = p(\mathbf{x}_1 | \epsilon) \prod_{i=2}^N p(\mathbf{x}_i | \mathbf{x}_{i-1}, f_X(X), \xi) \quad (3)$$

$$p(\mathbf{x}_i | \mathbf{x}_{i-1}, f(X), \xi) = \mathcal{N}(\mathbf{x}_i | f(\mathbf{x}_{i-1}), \xi^2 \mathbf{1}_{q \times q}) \quad (4)$$

$$p(\mathbf{x}_1 | \epsilon) = \mathcal{N}(\mathbf{x}_1 | \mathbf{0}_q, \epsilon^2 \mathbf{1}_{q \times q}), \quad f(X) \sim GP(k_X(X, X')) \quad (5)$$

The GPDM is easily extensible to higher-order dynamics. The mapping onto observable variables Y is done in the same fashion as for the GPLVM.

4 Coupling GPDMs with a Product-of-Experts Kernel

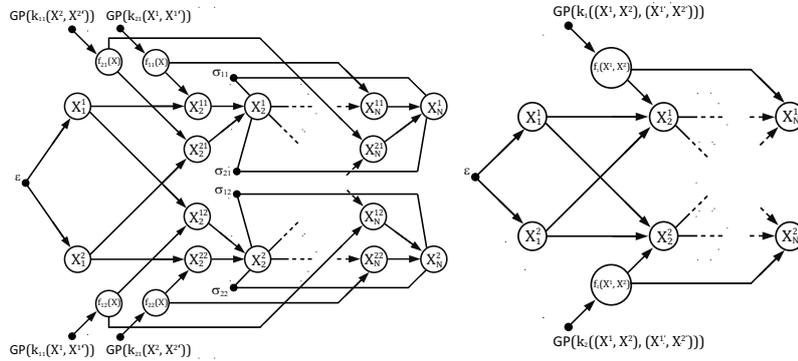


Fig. 2. Coupled latent space dynamics with two parts, observables Y^m omitted for clarity. **Left:** full model. Every part m of the latent X_{i-1}^m at time step $i-1$ generates a prediction $X_i^{m,r}$ about every part r at time step i via an evolution function $f_{m,r}(X^m)$. These individual predictions are combined through a Product-of-Experts approach. **Right:** after the (closed-form) marginalization of the $X_i^{m,r}$ and the $f_{m,r}(X^m)$, the model can be equivalently written as having only one evolution function $f_m(X^1, \dots, X^M)$ per part m . For details, see text.

We now derive a way of coupling the dynamics of a latent space partitioned into M parts, $X^{1:M}$. In a nutshell, we introduce $M \times M$ dynamics, each of which makes a prediction about a part of the latent space from the previous state of some (other) part, and these predictions are combined via product-of-experts (PoE) [9]. Observables Y^m are generated from the corresponding X^m as in the GPLVM.

More specifically, let X_{i-1}^m be the latent state of part m at (discrete) time $i-1$. We introduce $M \times M$ evolution functions $f_{m,r}(X^m)$ generating the prediction $X_i^{m,r}$ which the latent state of part m makes about the latent state of part r at time i , see fig. 2, left, for an example with $M=2$. The $f_{m,r}(X^m)$ are drawn from GPs with kernels $k_{m,r}(X^m, X^{m'})$. Observables Y^m have been omitted to keep graphical model less cluttered. The $X_i^{m,r}$ are the means of Gaussian 'experts' with isotropic coupling variances $\sigma_{m,r}^2$. Denoting the total predictive PoE variance of part r by

$$\sigma_r^2 = \left(\sum_m \sigma_{m,r}^{-2} \right)^{-1} \quad (6)$$

we find, by multiplying the parts' densities together and renormalizing:

$$\begin{aligned} p(\mathbf{x}_i^r | \mathbf{x}_i^{:,r}, \sigma_{:,r}) &\propto \prod_m \mathcal{N}(\mathbf{x}_i^r | \mathbf{x}_i^{m,r}, \sigma_{m,r}^2) \\ \Rightarrow p(\mathbf{x}_i^r | \mathbf{x}_i^{:,r}, \sigma_r) &= \frac{\exp \left[-\frac{1}{2\sigma_r^2} \left(\mathbf{x}_i^r - \sigma_r^2 \sum_m \frac{\mathbf{x}_i^{m,r}}{\sigma_{m,r}^2} \right)^2 \right]}{(2\pi\sigma_r^2)^{\frac{\dim(X^r)}{2}}} \end{aligned} \quad (7)$$

Note that the total PoE variance (eqn. 6) is smaller than any of the individual coupling variances.

Next, we marginalize the part predictions $X_i^{m,r}$ and the evolution functions $f_{m,r}(X^m)$, to obtain the joint density of the X_i^m . To this end, we make use of conditional independence properties of the model, which can be read off the graphical model (fig. 2, left) using the D-separation rules [17]. In the following, assume the X_i^j were fixed. Then, the tail-to-tail paths from $X_i^{m,r}$ to any $X_i^{m,q}$ with $r \neq q$ are blocked. Also, the head-to-tail paths from $X_i^{m,r}$ to any $X_j^{q,p}$ for $j \neq i$ are blocked, because there is (at least one) fixed node between them. There are no other open paths from $X_i^{m,r}$ to any $X_j^{q,p}$ for $r \neq p$, hence the part predictions about part r are independent from those about p across all time steps. On the other hand, fixing X_i^r opens head-to-head paths between the $X_i^{:,r}$. Finally, observe that there is an open tail-to-tail path from $X_i^{m,r}$ and $X_j^{m,r}$ for all i, j through the function node $f_{r,m}(X^m)$, which induces a dependency between predictions about the same part across time steps. Hence, we can marginalize the $X_i^{m,r}$ separately for each r , but we need to do so jointly across all m and i .

The dependency between the $X_i^{:,r}$ through fixed X_i^r is multivariate Gaussian for every i , see eqn. 7. The dependency induced by the unobserved evolution functions is multivariate Gaussian in time, because these functions are drawn from GPs. Since the priors on X_1^m are Gaussian, the joint density of the X_i^m must be a multivariate Gaussian as well. Hence, the marginalization boils down to a multivariate Gaussian integral, which we carry out using the following

Lemma 1. *Let \mathbf{v} and \mathbf{w} be multivariate Gaussian random variates. Assume $p(\mathbf{v}|\mathbf{w}) = \mathcal{N}(\mathbf{v}|\mathbf{P}\mathbf{w}, \mathbf{\Sigma})$ and $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{K})$, where \mathbf{P} is a $\dim(\mathbf{v}) \times \dim(\mathbf{w})$ projection matrix, and both $\mathbf{\Sigma}$ and \mathbf{K} are positive definite. Then $p(\mathbf{v}) = \mathcal{N}(\mathbf{P}\boldsymbol{\mu}, \mathbf{\Sigma} + \mathbf{P}\mathbf{K}\mathbf{P}^T)$.*

Proof. Marginalize \mathbf{w} using standard matrix algebra results [18].

To use this lemma, let $\mathbf{v} = x_{d,2:N}^r$ for some part r and component d . Construct the $M(N-1)$ -dimensional vector \mathbf{w} by stacking the $x_{d,2:N}^{m,r}$ for all m . Then, by virtue of eqn. 7, $\mathbf{\Sigma} = \sigma_r^2 \mathbf{1}_{(N-1) \times (N-1)}$ and $\mathbf{P} = \sigma_r^2 (\sigma_{1,r}^{-2}, \dots, \sigma_{M,r}^{-2}) \otimes \mathbf{1}_{(N-1), (N-1)}$ (\otimes denotes the Kronecker product). \mathbf{K} is a block-diagonal matrix, with M blocks

6 D. Velychko, D. Endres, N. Taubert, M.A. Giese

K^m , where the entries of these kernel matrices are computed from the kernel functions $k_{m,r}(X^m, X^{m'})$ (cf. fig. 2) as $\mathbf{K}_{i,j}^m = k_{m,r}(x_{:,i}^m, x_{:,j}^m)$ with $i, j = 2, \dots, N$. Since all GPs in our model have zero mean, $\boldsymbol{\mu} = \mathbf{0}_{M(N-1)}$. Thus, the mean of $\mathbf{v} = \mathbf{P}\boldsymbol{\mu} = \mathbf{0}_{N-1}$. The covariance matrix of \mathbf{v} is given by

$$\boldsymbol{\Sigma} + \mathbf{P}\mathbf{K}\mathbf{P}^T = \sigma_r^2 \mathbf{1}_{(N-1) \times (N-1)} + \sigma_r^4 \sum_{m=1}^M \frac{\mathbf{K}^m}{\sigma_{m,r}^4} \quad (8)$$

Since this holds for any choice of $x_{:,i}^m$, the Kolmogorov extension theorem guarantees the existence of a GP with constant zero mean function and a kernel function k_r generating these covariance matrices:

$$k_r(X^:, X'^:) = \sigma_r^2 \delta(X^:, X'^:) + \sigma_r^4 \sum_{m=1}^M \frac{k_{m,r}(X^m, X^{m'})}{\sigma_{m,r}^4} \quad (9)$$

where $X^:$ denotes the tuple (X^1, \dots, X^M) and $\delta(X, Y)$ is the Dirac delta function. We can therefore rewrite the graphical model of the coupled dynamical systems as depicted in fig. 2, right: for every part r , there is one evolution function that generates the current X_i^r from all previous X_{i-1}^m . This function is drawn from a GP prior with zero mean and a kernel as in eqn. 9. Note that we could in principle choose different kernel parts $k_{m,r}(X^m, X^{m'})$ for every m, r . That eqn. 9 is a valid kernel also follows from standard 'kernel engineering' rules [2].

5 Results

We tested the model on simple synthetic data sets first, see fig. 3. The data (blue lines) were created by sampling sine waves at 50 time steps with amplitudes 1, 2, and 3, and adding isotropic Gaussian noise with standard deviation 0.1. All three curves in one panel have the same frequency but different phase. We learned a coupled 2nd order GPDM with three parts and an RBF+linear dynamics kernel for the parts. The latent space had 2 dimensions, latent points formed a circle after learning. The coupling matrices in fig. 3 show the learned values of the relative coupling variances $\sigma_{r,r}^2/\sigma_{m,r}^2$, which are the higher the stronger the coupling is. After learning, we generated the red data by running the GPDMs in generative mode to show that they can reproduce and continue the training data. In the data on the left side of fig. 3 (panels A-C), all generating frequencies were different between panels. Consequently, the CGPDM learns that there should be no strong coupling between the parts (small off-diagonal values). In contrast, on the right side of fig. 3, the sine waves of panels E and F have the same frequencies, but shifted phases. Here, the coupling from part E to parts E and F is strong. Hence, part E can be used to drive both parts E and F.

Human walking data. To illustrate the power of the CGPDM on real-world data, we learned a two-part model on human walking data recorded with a Vicon motion capture system. The latent space of each part was three-dimensional, we used a RBF+linear+isotropic noise kernel for each coupling, and second order dynamics. The raw data were converted into exponential map format [6], which

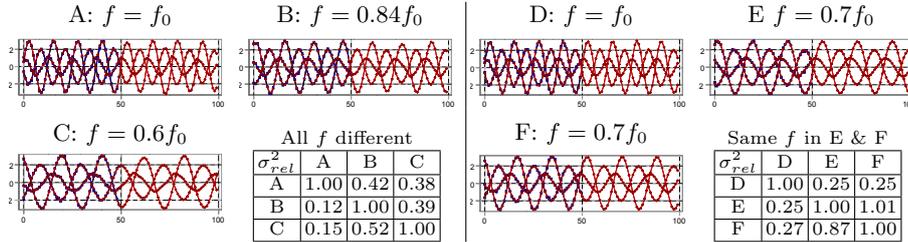


Fig. 3. Toy example. Observed data for three parts (panels A-C and D-F) were synthesized from sine waves with different or same frequencies f and phases, Gaussian noise was added (blue lines). The learned model could reproduce and continue the data (red lines). Matrices show relative coupling variances $\sigma_{rel}^2 = \sigma_{r,r}^2 / \sigma_{m,r}^2$ from part m (row) to r (column), large values indicate strong coupling (cf. eqn. 7). For details, see text.

is suitable for learning with GPs since it represents a joint rotation as a 3 dimensional real vector with unconstrained component values. The data were divided into upper body (thorax, arms and head) and lower body (legs and pelvis). After learning, we found that the relative coupling variances were ≈ 1 for both upper-to-lower and lower-to-upper coupling. We then synthesized walking motions by running the CGPDM generatively. **Panel a)** of the supplemental movie available at <http://www.compsens.uni-tuebingen.de/icannCoupledDynamics.html> shows a generated walk with the learned coupling variances. It looks quite natural, including variability between steps, but a rigorous psychophysical test of this observation has yet to be conducted. In **panels b) & c)**, the upper and lower body were started with a phase-shift of 20 frames, and coupled strongly (b) or weakly (c, small relative coupling variances). The difference in synchronization speed is clearly visible. **Panels d) & e)** show the result of driving one body part completely by the other: using the lower body as the driver leads to a smooth walking motion, whereas unnatural variability in the legs appears when the coupling is reversed. Finally, **panel f)** demonstrates that the body parts will not synchronize when completely decoupled.

6 Conclusion

We have derived a coupled GPDM from a product-of-experts principle, and demonstrated its ability to learn complex full-body human motion. Natural variability in the data is preserved. The coupling strengths can be modulated without needing to re-train the individual dynamics. Future work will focus on establishing dynamical stability conditions for the coupling kernel, possibly employing contraction theory [14]. Furthermore, we would like to extend our approach towards learning causal relations, e.g. for EEG and human motion data analysis. **Acknowledgments:** This work was supported by EU projects TANGO FP7-249858-TP3, AMARSi- EC FP7-ICT-248311; DFG GI 305/4-1, DFG GZ: KA 1258/15-1; BMBF, FKZ: 01GQ1002A, FP7-PEOPLE-2011-ITN(Marie Curie): ABC PITN-GA-011-290011, HBP FP7-ICT-2013-FET-F/ 604102; Koroibot FP7-ICT-2013-10/ 611909.

References

1. Ajallooeian, M., van den Kieboom, J., Mukovskiy, A., Giese, M.A., Ijspeert, A.: A general family of morphed nonlinear phase oscillators with arbitrary limit cycle shape. *Physica D: Nonlinear Phenomena* 263, 41–56 (2013),
2. Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer (2006)
3. Brand, M., Hertzmann, A.: Style machines. In: *Proc. SIGGRAPH'00*. pp. 183–192 (2000)
4. Chai, J., Hodgins, J.K.: Performance animation from low-dimensional control signals. *ACM Trans. Graph.* 24(3), 686–696 (2005)
5. Giese, M.A., Mukovskiy, A., Park, A., Omlor, L., Slotine, J.J.E.: Real-Time Synthesis of Body Movements Based on Learned Primitives. In: *Statistical and Geometrical Approaches to Visual Motion Analysis, LNCS*, vol. 5604, pp. 107–127. Springer Verlag (2009)
6. Grassia, F.S.: Practical parameterization of rotations using the exponential map. *J. Graph. Tools* 3(3), 29–48 (Mar 1998),
7. Grillner, S., Wallen, P.: Central pattern generators for locomotion, with special reference to vertebrates. *Ann. Rev. Neurosci.* 8(1), 233–261 (1985)
8. Grochow, K., Martin, S.L., Hertzmann, A., Popovic, Z.: Style-based inverse kinematics. *ACM Trans. Graph.* 23(3), 522–531 (2004)
9. Hinton, G.E.: Products of experts. In: *Proc. ICANN'99*. vol. 1, pp. 1–6 (1999)
10. Ijspeert, A.J., Nakanishi, J., Hoffmann, H., Pastor, P., Schaal, S.: Dynamical movement primitives: Learning attractor models for motor behaviors. *Neu. Comp.* 25(2), 328–373 (2013)
11. Lawrence, N.D.: Gaussian process latent variable models for visualisation of high dimensional data. In: *NIPS'03* (2003)
12. Lee, S.H., Sifakis, E., Terzopoulos, D.: Comprehensive biomechanical modeling and simulation of the upper body. *ACM Trans. Graph.* 28(4), 99 (2009)
13. Levine, S., Wang, J.M., Haraux, A., Popović, Z., Koltun, V.: Continuous character control with low-dimensional embeddings. *ACM Trans. Graph.* 31(4), 28 (2012)
14. Lohmiller, W., Slotine, J.J.E.: On contraction analysis for non-linear systems. *Automatica* 34(6), 683–696 (1998)
15. Mukovskiy, A., Slotine, J.J., Giese, M.: Design of the dynamic stability properties of the collective behavior of articulated bipeds. In: *10th IEEE-RAS Intl. Conf. Humanoid Robots*. pp. 66–73 (2010)
16. Neal, R.: *Bayesian Learning for Neural Networks*. Ph.D. thesis, Dept. of Computer Science, University of Toronto (1994)
17. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann (1997)
18. Petersen, K.B., Pedersen, M.S.: *The matrix cookbook* (2012), version 20121115
19. Rasmussen, C.E.: minimize.m. <http://learning.eng.cam.ac.uk/car1/code/minimize/> (2006)
20. Taubert, N., Endres, D., Christensen, A., Giese, M.A.: Shaking hands in latent space. In: *Adv. in AI, LNCS*. vol. 7006, pp. 330–334 (2011)
21. Urtasun, R., Fleet, D.J., Lawrence, N.D.: Modeling human locomotion with topologically constrained latent variable models. In: *Workshop on Human Motion, LNCS*. vol. 4814, pp. 104–118 (2007)
22. Wang, J.M., Fleet, D.J., Hertzmann, A.: Multifactor gaussian process models for style-content separation. In: *ICML*. pp. 975–982 (2007)
23. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Anal. Mach. Intell.* 30(2), 283–298 (2008)