Learning Conditional Structures in Graphical Models from a Large Set of Observation Streams through efficient Discretisation

Carl Henrik Ek Dan Song Danica Kragic

***Abstract*— The introduction of probabilistic graphical models to robotics research has been one of the success stories in the field over the last couple of years. Application of principles from statistical learning have allowed researchers to create systems which are aware and capable to reason about uncertainty in its observations. This has led to more robust and reliable systems.**

Many robotic applications are modeled from observations which are related by an underlying and unobserved conditional or casual relationship. One example of this is the study of affordances. Learning conditional structures from data has proved to be a tremendous challenge in the general case. Some progress have been made in special cases where the observations are discrete and low-dimensional. However, most interesting scenarios in robotics are characterized by high dimensional and continuous data. This means that for all but the simplest scenarios conditional structures had to be assumed in an ad-hoc manner and could not be learned from observations.

In previous work [1] we have presented a method which allows for principled discretisation of continuous variables. In specific we applied this method in order to model the task of robotic grasping. In this paper we extends the work by introducing an additional prior into the model. The aim of this prior is to encourage an additional degree of sparseness in order to reduce the complexity of the discrete representation. Further, we also extend the learning domain by applying the model to a more challenging data-set, which further shows the benefits of the suggested approach.

I. Introduction

The scenario we are interested in is characterized by a set of m different corresponding observation spaces $\mathbf{Y}_i = [\mathbf{y}_1^i, \dots, \mathbf{y}_N^i]$ where $\mathbf{y}_j^i \in \Re^{d_i}$. From this set of observations we want to build a model through which we can infer information about unobserved variables. We do not want to introduce any additional requirements on the data except it being vectorial. Therefore the model needs to be able to handle both discrete variables describing properties such as class and continuous variables describing for example visual information or motion features.

The goal of the model is to, given an observed subspace, infer the corresponding distribution over the unobserved subset of the model. Such questions can be answered given the joint distribution of the data $p(\mathbf{Y}_1, \dots, \mathbf{Y}_M)$ together with Bayes Rule. Modeling this distribution is potentially very complicated as the total dimensionality $\sum_{i=1}^{M} d_i$ for many real world scenarios is very big requiring significant amounts of data in order to be well represented. However, we expect the observations to exhibit a conditional relationships, which is something we can exploit to factorize the joint distribution. More specifically this implies a factorization,

$$p(\mathbf{Y}_1, \dots, \mathbf{Y}_M) = \prod p(\mathbf{Y}_i | \pi_i), \tag{1}$$

where π_i indicates the set of variables upon which $\mathbf{Y}_i$ is conditionally dependent. In terms of a graphical models this can be visualized by a network of nodes where π_i takes the role of the parents to $\mathbf{Y}_i$ connected by a directed edges from π_i to $\mathbf{Y}_i$, referred to as a Bayesian Network (BN) [2].

A BN is defined by the conditional distributions and the connectivity of the vertices. We will refer to the conditional distributions as the parameters and the connectivity as the structure of the network. Ideally we would like to learn both the parameters and the structure of the network from data. This poses a significant challenge and remains to be solved for the general case. In order to proceed a large range of heuristic methods have been developed which have shown good results in practice. In specific, significant advances have been made for modeling with discrete variables [3].

The requirement that each variable should be described by a set of discrete states is a significant limitation in many scenarios. In order to proceed the structure of the network can be defined a-priori using expert knowledge. However, the conditional structure of many data-sets is not observable and non-trivial to uncover.

In this paper we present a model which learns an intermediate discrete representation of each continuous variable opening up the possibility for heuristic structure learning of networks with continuous variables. The next section we will describe the mathematical foundation of our model. We will then proceed to present the proposed approach applied in a robotic grasping scenario.

II. Models

The Machine Learning task of dimensionality reduction is concerned with recovering the intrinsic or latent representation $\mathbf{X} \in \Re^{N \times q}$ from a set of observed data $\mathbf{Y} \in \Re^{N \times D}$ where $q < D$. In general we will assume that the observed data has been generated from the latent through a generative mapping,

$$\mathbf{y}_i = f(\mathbf{x}_i). \tag{2}$$

The problem of dimensionality reduction is severely ill-constrained as an infinite combination of mappings f and latent representations $\mathbf{X}$ could have generated the data $\mathbf{Y}$. In order to proceed assumptions have to be made. In specific these assumptions can be divided into two sets of approaches, spectral methods which assumes that the generative mapping is invertible, and generative models which directly model f.

C.H Ek, D.Song and D. Kragic are with KTH – Royal Institute of Technology, Stockholm, Sweden, as members of the Computer Vision & Active Perception Lab., Center for Autonomous Systems, www: http://www.csc.kth.se/cvap, e-mail addresses: {chek,dsong,danik}@csc.kth.se.

However, the scenario we are interested in is characterized by several different high-dimensional observation spaces $\mathbf{Y}^1, \ldots, \mathbf{Y}^M$ from which we are interested in learning their conditional structure using a directed graphical model. As described in Section I this requires the data to be discrete.

Our approach in this paper is to take inspiration from dimensionality reduction and learn a new representation of each observation space independently. The new parametrisation of the data is both low-dimensional and structured in such a manner that is "good" for clustering. The model is probabilistic, which we exploit in order to recover a continuous estimate from the conditional predictions of the BN. In specific we make use of Gaussian Process priors and a sparse formulation which is learned using a variational approach.

A. Gaussian Process

A Gaussian Process ($\mathcal{GP}$) [4] is defined as a collection of random variables of which any subset has a joint Gaussian distribution. The $\mathcal{GP}$ is defined completely by its mean function $\mu(\mathbf{x}_i)$ and its covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ defined by a set of parameters θ. Being defined over infinite index sets, a $\mathcal{GP}$ can be used to specify distributions over infinite objects. This has meant that one of the typical uses of $\mathcal{GP}$s is to define distributions over functions.

Given two sets of corresponding data $\mathbf{X} \in \Re^{N \times q}$ and $\mathbf{Y} \in \Re^{N \times D}$ and assuming the data to be related by a functional relationship with additive Gaussian noise, $\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon$, we can formulate a Gaussian likelihood of the data. Using a $\mathcal{GP}$ to specify a prior over the mapping we can marginalise out f and reach the marginal likelihood of the data,

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \int p(\mathbf{Y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \theta) \mathrm{d}f. \tag{3}$$

Learning in the $\mathcal{GP}$ framework can then be performed by finding the maximum likelihood solution of the parameters θ of the $\mathcal{GP}$ referred to as the hyper-parameters of the model. In specific we can often translate the data and assume the mean function to be constant which leaves only the parameters of the covariance function to learn.

Finding the maximum likelihood solution has a computational complexity of $\mathcal{O}(N^3)$ which severely limits the size of the data-sets which can be treated using this approach. In order to apply $\mathcal{GP}$s to larger problems, a significant amount of work has been applied to formulating efficient methods which are computationally more beneficial.

In specific, [5] formulated an augmented likelihood function in terms of a set of variables $\mathbf{U} \in \Re^{m \times q}$, $m \ll N$, referred to as the *inducing variables* of the model. The underlying assumption is that the input locations are conditionally independent given the inducing variables which leads to the following likelihood,

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{U}, \theta) = \int p(\mathbf{Y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \theta) p(\mathbf{X}|\mathbf{U}, \theta) \mathrm{d}f \tag{4}$$

In [6] a variational framework is developed which can efficiently learn this sparse model.

B. GP-LVM

The Gaussian Process Latent Variable Model (GP-LVM) is a generative dimensionality reduction approach which builds on $\mathcal{GP}$s. By placing a $\mathcal{GP}$-prior over the generative mapping f in Eq. 2 and a prior distribution over the latent locations $\mathbf{X}$, both the hyper-parameters and the latent locations can be found through the marginal likelihood,

$$\{\hat{\mathbf{X}}, \hat{\theta}\} = \mathrm{argmin}_{\mathbf{X}, \theta} \int p(\mathbf{Y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \theta) p(\mathbf{X}) \mathrm{d}f. \tag{5}$$

The regularizing effect of the $\mathcal{GP}$ and the latent prior together with fixing the dimensionality of the latent space means that a solution can be found using this approach. Learning in the GP-LVM framework is done by finding the MAP solution to the latent representation $\mathbf{X}$ and the ML solution to the hyper-parameters θ. This is done using gradient based methods. In order to make learning more efficient the augmented model using the variational approach of [6] can also be applied to the GP-LVM.

C. GP-LVM for Clustering

When introducing the GP-LVM model we deliberately avoided explaining the form of the latent prior $p(\mathbf{X})$. As the task of dimensionality reduction is severely ill-constrained we need to introduce a preference for the latent configuration into the system. In the original formulation of the model a non-informative Gaussian prior was used in order to remove the invariance to the scale of the latent representation [7]. Other works have exploited this ambiguity and formulated different priors in order to learn latent structures which respect the class constraints [8], sequential information [9] or "topological" structures [10]. There is nothing right or wrong about these different priors, as stated above, an infinite number of valid solutions exists, different priors just works as a mean of encoding a preference among the possible solutions making the problem better conditioned. It could be argued that from a learning perspective the original uninformative prior is the least intrusive of the ones proposed. However, there is no one single valid latent representation but rather an infinite set to chose from, if the generative mapping is capable of modeling the data from a specific latent structure then this representation is valid. Given that we want to use the representation in a specific manner we can introduce a prior which directs us to a more desirable solution from an application perspective.

In this paper we are interested in learning a representation of the data which can be represented in a discrete form with a good accuracy, i.e. well-separated clusters. Translating this into the notion of $\mathcal{GP}$'s this would mean that the covariance function evaluated on a specific latent point $\mathbf{x}_i$ would have a high degree of covariance with the points within the same cluster but small for the remaining points. Specifying such a prior is non-trivial as it requires dealing with equivalent classes of associations, i.e. we need to implicitly associate a point with a cluster center.

However, by using the augmented model from the sparse approximation we can proceed. The augmented model ap-

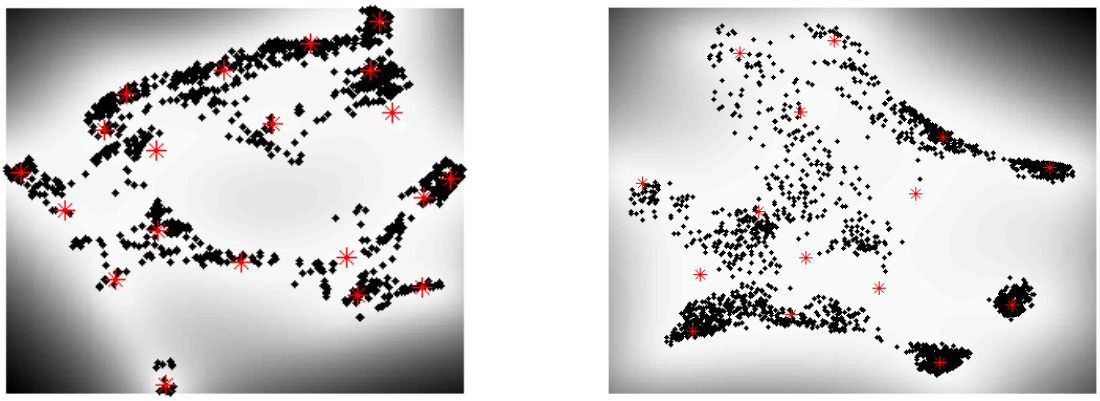

Fig. 1. *The plots above shows the latent representation learned by applying the sparse variational GP-LVM approximation. The left image shows the results for the standard approach while the right plot shows the result when the proposed prior is applied. The training data-points are represented in black and the inducing points by red crosses, the grey-scale background shows the variance associated with a prediction from the corresponding latent location. As can be seen from above the prior makes the inducing points better separated on the latent space.*

pends the likelihood function with a set of inducing points. The inducing points makes the latent locations conditionally independent. This means that we can interpret the inducing points as cluster centers as they are assumed to be capable to describe the full latent space. Using this analogy we can specify a prior over the inducing points such that they represents "good" cluster centers. This means that we do not need to deal with associating data points to the clusters as this will automatically be taken care of when learning the generative mapping.

We would like to encourage a representation where each of the cluster centers are well separated on the latent space while at the same time capable of accurately reconstructing the data. The latter is handled by the reconstruction of the generative mapping. What remains is to formulate a prior which encourages the inducing points to be well separated.

Such a prior can be formulated by penalizing the L_1 norm of the off-diagonal elements of the inner-product matrix computed between the inducing points,

$$p(\mathbf{U}|\theta_U, \beta_U) = \mathcal{N}(\sqrt{D(\mathbf{U}, \theta_U)}|0, \beta_U^{-1}), \tag{6}$$

$$D(\mathbf{U}, \theta_U) = \sum_{ij}^{M} \lambda_{ij} k_u(\mathbf{u}_i, \mathbf{u}_j, \theta_U), \lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}.$$

If $k_u(\mathbf{u}_i, \mathbf{u}_j)$ is a smooth monotonically decreasing function with respect to $||\mathbf{u}_i - \mathbf{u}_j||$ the distribution will encourage a representation with well separated clusters. The parameters β_U and θ_U control the strength of the prior and the smoothness of $k_u(\mathbf{u}_i, \mathbf{u}_j, \theta_U)$ respectively. In specific we will use a RBF function where θ_U controls the width of the function which relates to the strength of the prior.

Once a model has been trained, discretisation can be done by the posterior distribution over each inducing point $p(\mathbf{y}_i|\mathbf{u}_j)$. This means that we can acquire an uncertainty associated with the discretisation which can be exploited when learning the network, i.e. using the soft-evidence. The results presented in this paper, however, used a point estimate for the discretisation (hard-evidence).

III. Inference

A trained network defines an efficient factorization of the joint distribution of the observations in terms of a network of conditional dependencies. An acyclic graph can be converted into a tree through the junction tree algorithm [11] allowing us to formulate any conditional distribution over any subset of the observations. This allows us to ask any set of questions about the state of any node in the network. The output of the network is in terms of a multinomial distribution over each of the discrete states of the network,

$$\mu_{ijk} = p(\mathbf{x}_i = \mathbf{u}_k | \pi_i = \mathbf{U}_j). \tag{7}$$

We will now describe how we can, in an efficient manner, recover a continuous estimate from this distribution.

Each point on the latent space $\mathbf{X}$ defines a distribution over the observed data space $\mathbf{Y}$ through the $\mathcal{GP}$ modeling the generative mapping. Therefore in order to acquire a continuous estimate we need to determine a distribution over the latent space associated with the multi-nominal distribution from the tree. In order to achieve this we learn a parametric mixture model with the location of the inducing points as the mixture centers. In specific we use full-covariance Gaussian basis functions,

$$p(\mathbf{x}) \propto \prod_{i=1}^{M} \lambda_i \mathcal{N}(\mathbf{x}|, \mathbf{u}_i, \Sigma_i^{-1}) \tag{8}$$

The parameters of the mixture model are learned using the standard EM approach. The multinomial output distribution from the network defines a distribution over the inducing points. We use this distribution to specify the mixture components to create the following conditional mixture model,

$$p(\mathbf{x}_j^i|\mathbf{Y}^V) \propto \prod_{k=1}^{M} \mu_{ijk} N(\mathbf{x}|, \mathbf{u}_k, \Sigma_i^{-1}). \tag{9}$$

We can then sample from the above distribution in order to find locations over the latent space.

In the next section we show experimental results for the task of robot grasping when applying the proposed model.

IV. Experiments

We apply the proposed approach to model grasping tasks. Grasping is a very challenging problem as it is specified by several different observation spaces from which we acquire

noisy and uncertain data. We wish to model a complete joint set of the observations in order to be able to make predictions on any set of variables when observing other sub-sets. As we expect there to exist significant conditional (in)dependencies between different variables. However, the observation spaces are often high-dimensional, consisting of both discrete and continuous variables. Further, the correlations in the data are often not obvious. This has eluded a direct application of the BN formalism and applications have to resort to pre-defined structures using only a subset of the available observation streams such as in [12].

In our current full paper submission [1], we applied the proposed discretisation method in order to learn both structure and parameters of a directed graphical model. In this paper we have extended the model with the prior on the inducing points as explained above. We also introduced one additional task and more objects into the data-set making the problem even more challenging from a modeling perspective. We will now give a brief explanation of the extensions to the data-set.

A. Data Generation

We extracted synthetic data describing each object-grasp configuration using the grasp-planner BADGr [13] in a simulation environment provided by GraspIt! [14]. The hand model we used was the human 20 degrees-of-freedom hand provided in GraspIt. The features describing each grasp were divided into three sub-sets: *object features* (O) extracted from the object representation, *action features* (A) extracted from the planned grasps, and *constraint features* (C) resulting from the complementation of both, i.e. the hand-object configuration. Each grasp was visualized in GraspIt! to a human tutor who associated it with a task label (T). The extraction process was the same as in [1] where we refer the reader for additional detail. However, compared to [1] we extended the task space to include four tasks, *Hand-Over, Moving, Tool-Use* and *dish-washing*. The new task *dish-washing* was defined as grasping an object to put it into a dish-washing machine. In addition, we also added more object models within each of the 6 object classes (knife, hammer, screwdriver, bottle, glass and mug), leading totally to 48 object models in total (approximately doubling the object numbers in [1]).

V. Results

In order to evaluate the effects of adding the additional sparseness prior to the model we first compare the resulting latent representation with and without the prior. To do so we use the original data-set first presented in [1]. In Figure 1 we show the resulting latent spaces for the two models for a specifically challenging observation sensory stream, the final hand configuration.

The motivation behind the prior is to encourage the model to cluster data-points together and explain them with as few of the inducing points as possible. The result shown on the left plot of Figure 1 clearly indicates that this works well. Especially for the regions of the right part of the latent space, we see several distinct clusters that are formed associated with a single inducing point. This is in contrast to the model without the prior whose results are shown in the left plot in Figure 1: there are less clearly defined clusters. But more importantly the location of the inducing points seems to be more correlated with the location of the data-point and not with their shape. This means that a rather tight cluster like the one in the far right of the latent space in the left plot of Figure 1 are associated with two inducing points which makes clustering challenging. This exemplifies the competition between the sparseness prior and the reconstruction error in the model.

TABLE I

Confusion matrix for task classification when observing all the action and the object features.

hand-over	35	26	31	8
pouring	1	81	0	18
tool-use	0	0		0
dish-washing	8	6	6	80

However, we can also see that additional problems occur with the introduction of the prior. In regions in the middle of the space, the latent representation has been stretched out as the priors are forcing the inducing points away from each other. This is an effect of using incorrect number of inducing points which in the model has to be chosen a prior. This introduces a competition over the points that could have been associated with the same cluster, to follow several inducing points. As the prior wants to spread them apart this will stretch the representation and result in a less clustered representation. Therefore choosing the right number of inducing points is a key aspect of the algorithm. At the moment this needs to be chosen by inspection.

Given a trained Bayesian network, we can also evaluate its performance in discriminating between different tasks. In order to do so we observe the action and the object features and aim from those to determine the task. Classification is done by selecting the task associated with the highest probability of the conditional distribution $p(\text{task}|\text{action}, \text{object})$. In Table I the confusion matrix computed on a test data-set is shown. As can be seen the model is capable of classifying three of the four classes with good accuracy. The performance for classifying the *hand-over* task shows a large confusion over *pouring* and *tool-use*. This is expected as grasps that are good for *hand-over* are in many cases also likely to work well for the other two. This indicates that our classification of task might need a hierarchical structure rather than the flat class association we use here. We would like to point out that the results here are the point estimates, while the BNs provides probabilities for each task; making use of this would benefit the classification.

Once given a trained model we are interested in the capabilities of inferring unobserved data spaces when we have observed only a subset of the sensory streams. In Figure 2 the results of inferring the grasp position from

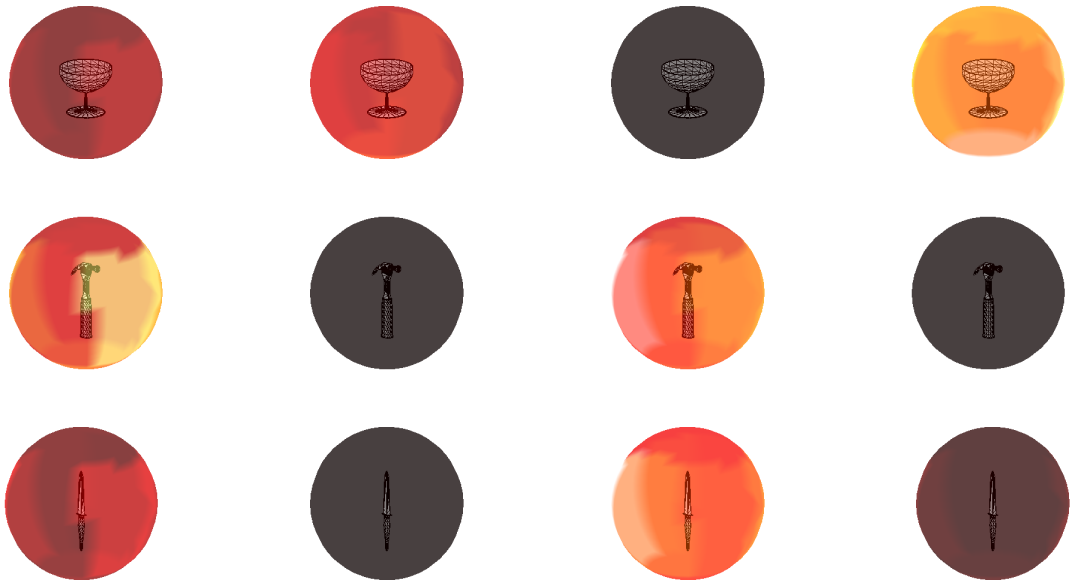

Fig. 2. *Each row shows a specific object where the colour of the surrounding sphere encodes the most likely position for the object to be grasped when performing the associated task. Each column defines a different task; from left to right:* **hand-over, pouring, tool-use** *and* **dish-washing**. *The probability mass is normalised over the different tasks in order to indicate affordances, i.e. what task does a specific object affords. As we can clearly see this is apparent where the glass does not afford the task tool-use while the hammer and the knife does. On the other hand, glass affords dish-washing which has a very low likelihood to be associated with the other objects.*

object features and task are shown. The heat map shows the likelihoods of the the grasping positions, and the likelihoods are normalized across columns on the four tasks. The darker color indicate the grasping position is not likely given the object features and task assignment. We see from this figure, the models successfully ruled out the glass for *tool-use* task and the hammer and screwdriver for *pouring* and *dish-washing*. And for the tasks the object can afford, for example when hammer is used for *tool-use*, then heat map indicate the grasps should not be applied from the hammer head direction (indicated by darker map). And the preferred position of grasp is on the two side of the handles.

VI. Conclusion

In this paper we have presented a principled approach for learning low-dimensional clustered representations of high-dimensional continuous data. We extend previous work on sparse probabilistic dimensionality reduction with a prior encouraging learning a model suitable for clustering of data. The proposed model allows us to learn a factorised conditional structure of a large set of continuous variables through efficient discretisation. We show how by application of the proposed method we can learn accurate models of the task of robotic grasping from a large range of sensory streams.

The addition of the prior to our model results in spaces which are more compact and beneficial for discretisation. However, this addition also introduces new challenges to the model. In specific, choosing the right number of inducing points becomes more important. We are currently working towards extending the model in such a manner that it will be capable of learning the number of states from data.

Acknowledgments

This work was supported by EU IST-FP7-IP GRASP, EU IST-FP6-IP-027657 PACO-PLUS, and Swedish Foundation for Strategic Research. The authors would like to thank Dr. Kai Huebner for his support for data generation and experiment design.

References

[1] D. Song, C. H. Ek, K. Hubner, and D. Kragic, "Multivariate discretization for bayesian network structure learning in robot grasping," in *International conference on robotics and automation (ICRA)*, 2011.

[2] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, September 1988.

[3] L. D. Fu and I. Tsamardinos, "A Comparison of Bayesian Network Learning Algorithms from Continuous Data," in *Proceedings of Annual Symposium of American Medical Informatics Association (AMIA)*, 2005.

[4] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[5] Q. Candela and C. Rasmussen, "A unifying view of sparse approximate gaussian process regression," *JMLR*, vol. 6, pp. 1935–1959, 2005.

[6] M. Titsias, "Variational Learning of Inducing Variables in Sparse Gaussian Processes," in *Artificial Intelligence and Statistics*, 2009.

[7] N. D. Lawrence, "Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, November 2005.

[8] R. Urtasun and T. Darrell, "Discriminative Gaussian process latent variable model for classification," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 927–934.

[9] J. Wang, D. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 283–298, 2008.

[10] R. Urtasun, D. Fleet, A. Geiger, J. Popović, T. Darrell, and N. Lawrence, "Topologically-constrained latent variable models," in *Proceedings of the 25th international conference on Machine learning*. ACM New York, NY, USA, 2008, pp. 1080–1087.

[11] S. Lauritzen and D. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society. Series B*, pp. 157–224, 1988.

[12] D. Song, K. Hubner, and D. Kragic, "Learning task constraints for robot grasping using graphical models," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[13] K. Huebner, "BADGr - A Toolbox for Box-based Approximation, Decomposition and GRasping," in *IROS*, 2010.

[14] A. T. Miller and P. K. Allen, "GraspIt! A Versatile Simulator for Robotic Grasping," *IEEE Robotics and Automation Magazine*, 2004.