Sparse Summarization of Robotic Grasp Data

Martin Hjelm Carl Henrik Ek Renaud Detry Hedvig Kjellström Danica Kragic

***Abstract*— In this paper we propose a new approach for learning a summarized representation of high dimensional continuous data. We apply the model to learn efficient representations of grasp data for two robotic scenarios that facilitates a compact summarization. Our technique consists of a Bayesian non-parametric model capable of encoding high-dimensional data from complex distributions using a sparse summarization. In specific the method marries techniques from probabilistic dimensionality reduction and clustering to reach a solution. We show how the summarization provided by the model significantly benefits interpretations of data for a robotics grasping task.**

I. Introduction

The data-driven paradigm has had a profound effect on computer science in general and robotics in specific. With increasing amounts of data we have been able to approach diverse topics such as, grasp transfer [1], gestures [2], imitation learning [3] and action recognition [4] to name just a few, from a data-driven perspective. As new richer sensors and more efficient computational units are developed the hope is that the same approaches can be applied to new areas. Between hope and realization stands the ever-present foe accompanying data-driven learning, *the curse of dimensionality* [5]. This important notion can simply be summarized as, when the dimensionality increases linearly, the amount of data necessary for making statistically sound conclusions increases exponentially. This is also true for many algorithms where the computational cost scales badly with the dimension. In order to alleviate these effects it is important to represent data in as simple and efficient form as possible. However, this notion goes beyond just considering the dimensionality of the data. More generally it is important to try to understand the characteristics of each algorithm and prepare the data such that it has the right form. As an example, most classification methods will project the data into a high-dimensional space and produce simple linear boundaries compared to a low-dimensional one that exhibits a more complicated class separation. However, in many situations even with solid domain knowledge and intuition understanding the data can be hampered, and low-dimensionality can be assumed to be a desirable characteristic for finding the right form of the representation. To that end it is common, as a pre-processing step, to reduce the dimensionality of data using one of the many dimensionality reduction techniques that has been suggested in the literature [6]–[10].

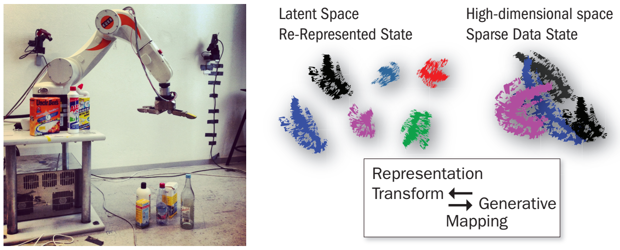


Fig. 1: In robotics, data is typically embedded in a high dimensional space and structured in such a way that it represents the sensory data but does not reflect any deeper insights. By forcing a re-representation in a favorable form inference becomes a much simpler task. Making generalizations which in our case mean such disparate things as relating actions objects and tasks or finding grip similarities both results in the need for re-representation into clustered form but also the need for a generative approach that lets one move from a ”theory of forms” to the possibilities offered in the high-dimensional space.

Another important class of problems is those that prefers or requires clustered or discrete data. As an example, most features in computer vision such as [11]–[13] rely on an efficient discretization and representation of a very high-dimensional feature space. As another example, in machine learning, the structures of a graphical model have a profound effect on its descriptive power. Little general progress has been made for learning structure from continuous data while for discrete data a range of methods exists (for a review see [14]). As a tool for visualizing and summarizing data, clustering and discretization can be very effective. In [15] clustering was used to extract a prototypical grasp, which allowed generalization of grasps to novel objects. However, discretization is a ”hard-choice” where once a data-point is associated to a state its relationship to the original feature representation is lost.

Our approach to these two problems is a method that simultaneous performs dimensionality reduction and clustering. We learn a latent space via the augmentation of the observed data together with a sparse generative mapping. The generative mapping is coupled with the latent space representation of the augmented data in such a way as to enforce the latent representation of the observed data to be explainable through the augmented data, which is constrained to be uncorrelated. We perform a thorough analysis, testing the model on both real and synthetic data.

II. Related Work

There exist an infinite number of possible parameterizations of any type of data. This means that in order to proceed, one needs to make assumptions and specify a preference

M. Hjelm, C. H. Ek, R. Detry, H. Kjellström and D. Kragic are with the Centre for Autonomous Systems and the Computer Vision and Active Perception Lab, CSC, KTH Royal Institute of Technology, Stockholm, Sweden. Email: {martinhjelm,chek,detryr,hedvig,danik}@csc.kth.se

This work was supported by the Swedish Foundation for Strategic Research, the Belgian National Fund for Scientific Research (FNRS), the Swedish Research Council, and the EU project TOMSY (IST-FP7-Collaborative Project-270436).

of the characteristics of the representations that we seek. Of specific importance are methods that aim to find low-dimensional representations of data, commonly referred to as dimensionality reduction methods. Formally such methods make the assumption that the observed representation of the data has been generated from an underlying representation through a generative mapping. Such methods can be divided into two different approaches, (i) spectral methods that aim to find a mapping from the observed data to the new parameterization and (ii) generative models, which aim to find a representation that can generate the observed data. Spectral methods aim to model the inverse of the generative mapping and are therefore much more restricted and only considers the set of solutions where the generative mapping takes the form of a bijection [16]. Generative methods on the other hand are much more flexible but in order to reach a solution additional information needs to be provided to limit the space of solutions. In this paper we will exploit the flexibility of a generative approach.

A generative method that has seen wide success is the Gaussian Process Latent Variable Model (GP-LVM) [6]. In the GP-LVM framework the generative mapping is modeled using a flexible Gaussian Process ($\mathcal{GP}$) prior [17] and the resulting representation is referred to as the latent representation. One of the main benefits of the model is that it is straightforward to incorporate priors on the latent representation. In the original presentation of the model an uninformative prior was used to regularize the solution space [6]. However, many different priors have been suggested encoding different preferences on the representations. In [18] a model that enforces the latent locations to respect the local distance in the observed space was presented. In [19] the authors propose a prior based on class information learning a representation that reflects the class division, similarly [20] constrains the latent space to respect a certain topology. More similar to the approach we will present here is [21] which uses a prior that encourages the latent space to reflect the dynamics of the data by using an auto-regressive prior. The representations learned using such priors have had a big impact on modeling of high-dimensional dynamic data such as human pose [22] where it has provided representations suitable for dynamical models. In robotics this have been further developed with the proposal of additional regularization which encourage temporally regular shaped latent spaces for data over multiple repetitions such that each matches a simple template sequence [23].

The wide variety of priors that have been discussed above all aims at the same goal: *finding a representation that will suit a specific task or model*. In this paper we will present work on a prior that generates latent representations aimed at summarizing high dimensional data using a sparse clustered representation. Our approach is an extension of the work presented in [1], [24] but provides a much more principled formulation that significantly increases the strength and applicability of the model.

III. Model

In this section we will formalize the notion underpinning our view of data-representation. Given a set of observed data $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ represented in a space or parametrization Y we wish to find a new parametrization X that summarizes and represents the observed data. Formally there is a mapping f that relates elements $\mathbf{x}_i$ in the latent representation X to its corresponding observed parametrization $\mathbf{y}_i \in Y$,

$$f : X \rightarrow Y. \tag{1}$$

For a set of observed data $\mathbf{Y}$ there exists an infinite number of possible representations that respects the above relationship. In this paper we will adopt a latent variable approach where the mapping f will take functional form. The scenario we are interested in is finding a well separated and grouped representation such that it can easily be summarized in terms of a clustering. We think about the cluster centers in terms of being explanatory for the data, i.e. we should be able to explain any given point in the cluster using the point that defines the cluster.

More formally, for the latent representation we adopt a GP-LVM model, assuming the observed data to be generated from the latent representation through a functional mapping with additive Gaussian noise. This leads to the likelihood of the model $p(\mathbf{Y}|\mathbf{f})$, where $\mathbf{f}$ is the instantiation of the function. The GP-LVM proceeds by assuming that each dimension of the observed data is independent, given the latent locations, and by placing a $\mathcal{GP}$ prior over the latent locations.

A $\mathcal{GP}$ is a finite collection of random variables that have a joint normal distribution, where every random value is considered a function value, $f(x)$ at the point x. A $\mathcal{GP}$ is uniquely determined by its mean and covariance function, where the covariance function relates the influence of the other random variables in the collection that is: $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, where μ and $k(\cdot,\cdot)$ are the mean and the covariance function respectively.

The benefit of the GP-LVM compared to other latent variable models is that the mapping f can be integrated out which leads to a marginal likelihood which averages over every possible mapping f.

$$p(Y|X;\theta) = \prod_{i=1}^{D} \int p(Y^i|F)p(F|X;\theta)dF = \prod_{i=1}^{D} \mathcal{N}(Y^i|0, K;\theta) \tag{2}$$

Here $\mathbf{Y}^i$ corresponds to the i:th dimension of the observed data, K the specific covariance and θ is the hyper-parameters of the covariance function. The $\mathcal{GP}$ prior in the model is extremely flexible and can, with an appropriately chosen kernel function, be made to have a non-zero probability for a very large range of functions allowing for a very representative model. Learning implies seeking the latent location $\mathbf{X}$ and the hyper-parameters θ and can be found by maximizing the posterior of the model

$$\arg\max_{\mathbf{X},\theta} p(\mathbf{X},\theta|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X},\theta)\, p(\mathbf{X}) \tag{3}$$

In the original model presented in [6] an uninformative prior was used for the latent space to get a solution that was as unrestricted as possible. However, when additional information is available, such as in [19], or when a specific structure of the space is desired such can easily be accommodated within the framework by formulation of a more specific prior $p(\mathbf{X})$ [23]. We will now proceed to explain how a prior that facilitates a clustered representation can be formulated.

A. *Latent space priors*

As previously stated the prior on the latent representation encodes a preference that allows us to reach a solution as it regularizes the infinite solution space. It is illustrative to think about the problem from another direction, imagining already clustered, well-represented lower dimensional input that are related to some higher dimensional target values via some function. This is essentially a regression problem and to solve it we want to use $\mathcal{GP}$ regression. However the regression must meet the condition that some points in the data, the cluster centers represents the same information as the data in the clusters e.g. if we were to remove some data points the solution should still be the same. Therefore we augment our data set with additional input, target value pairs, $(\mathbf{U}, \mathbf{f}_u)$ that have this explanatory capability. We refer to the pairs as *inducing inputs* and *inducing points*. The complete probability is now written as

$$p(\mathbf{Y}, \mathbf{f}, \mathbf{f}_u|\mathbf{X}, \mathbf{U}) = p(\mathbf{Y}|\mathbf{f}, \mathbf{f}_u)\, p(\mathbf{f}|\mathbf{f}_u)\, p(\mathbf{f}_u) \qquad (4)$$

where $\mathbf{f}_u$ are the inducing points in the observed space and $\mathbf{U}$ the corresponding latent inputs. This is the same formulation as in sparse $\mathcal{GP}$ regression [25]. However there the augmentation data, the inducing points and inputs are considered as additional variables for approximating the true posterior but here they are cluster centers with the power to represent the data points that belong to the clusters.

If we go back to our GP-LVM and use the augmentation idea to model our representation of the data to a latent clustering we realize that if the cluster centers should be responsible for explaining the points belonging to the cluster then the inducing points in the observed space, should be as uncorrelated as possible. By reducing the correlation between the inducing points we are forcing them to choose which points in the data set to explain. The GP-LVM will thus be forced to find a balance between a good latent representation and one that clusters the latent variables. A good latent representation will mean a mapping to the observed data that is as probable as possible. The decorrelation property and the explanatory capacity of the augmentation data will mean a latent representation that is as separated as possible and where the cluster centers explain the cluster points as good as possible.

In practice such behavior will manifest itself when the covariance function evaluated on the inducing variables $\mathbf{U}$ is diagonal. For example in a grasp representation this would mean recovering a set of independent postures that are capable of *inducing* the full range of possible postures in the data. To that end, we are motivated by the inducing prior that was defined in [1], [24] which penalizes the $\mathcal{L}_1$ norm of the off-diagonal elements of a kernel matrix evaluated on the inducing variables,

$$p(\mathbf{U}|\theta, \beta) \sim \mathcal{N}(\sqrt{D(\mathbf{U}, \theta)}|0, \beta_U^{-1}) \qquad (5)$$

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

where $k(u_i, u_j, \theta_u)$ is the covariance function of the $\mathcal{GP}$ prior on the inducing points $\mathbf{f}_u$, θ the support and β_U^{-1} the precision parameter or what we will later refer to as the constraint parameter. This will force the covariance matrix of the inducing points, in the $\mathcal{GP}$ prior into a more diagonal form since any non-zero off-diagonal values will be penalized.

Introducing the augmented clustering data into the GP-LVM we can marginalize out f_u in the same manner as f was marginalized out, leading to the following posterior

$$p(\mathbf{X}, \mathbf{U}, \theta|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \mathbf{U}, \theta)\, p(\mathbf{U}|\theta)\, p(\mathbf{X}) \qquad (6)$$

where the optimization now is also over the cluster centers.

In the original model proposed in [24], a discretization of the latent space i.e. the inducing inputs of the posterior above were sought. However, the prior over them, $P(\mathbf{U})$ and the generative mapping was decoupled meaning that we did not fully exploit and model the relation between U and f_u. In the optimization process they were separate from each other in the sense that the kernel support parameter was not shared. This means, that in the optimization process the parameter on $\mathbf{f}_u$ would focus on explaining the observed data $\mathbf{Y}$, and would not be affected by the prior on $\mathbf{U}$ to the same extent, if a certain form of the latent representation part would be highly likely.

By coupling the support parameter in the cluster prior and the inducing $\mathcal{GP}$ prior we enforce consistency, since the support parameter for the inducing $\mathcal{GP}$ prior now will be less likely to move in a direction that would increase the representative probability at the expense of the cluster prior probability. Hence the explanatory capacity of the inducing inputs for the latent variables in terms of clusters becomes much more effective. In the experimental section we will show results comparing the original uncoupled approach with this approach where the parameters of the kernel function are shared.

Since our formulation of the augmentation of the data is mathematically equivalent to sparse $\mathcal{GP}$ regression we can utilize the same methods, we therefore use the sparse variational approximation of Titsias et al. [26] for the $\mathcal{GP}$ prior. The variational approach has several good qualities, it avoids over fitting, is quite fast compared to other approximations etc.

The maximization of the posterior i.e. the learning is done via conjugate gradient descent. The latent variables are assigned to the cluster centers / inducing inputs with which they have the largest covariance. In this case this is equivalent to choosing the minimum Euclidian distance since we are using a RBF kernel. This is an attractive aspect of our method since it opens up for other kernels or other distance measures for assigning points to clusters. It also means that our assignments are not hard assignments unless we specify so, since the covariance gives a degree of association. In our case this is useful since it allows us to be more coarse or fine-tuned when selecting grasp generalizations for example. In a more general setting it allows for agglomerative clustering depending on the cut-off value of the covariance one set to rule the association.

IV. Experiments

In order to fully appreciate the flexibility of our method we use three different data sets for evaluation, a synthetic data set allowing us to have full control over the characteristics of the data and two recently presented data sets related to grasping in robotics. We begin by providing an intuition to how the initialization, parameters and number of clusters affect the latent output and then proceed to the experiments.

A. Initialization, Parameters, Number of clusters

The initialization process can be thought of as a pre-step to help the clustering. We can think of the log posterior we are trying to optimize as an energy function being the sum of two parts

$$E = E_{\text{Representation}} + E_{\text{Clustering}} \tag{7}$$

where $E_{\text{Representation}}$ is our GP-LVM posterior, the representative part and $E_{\text{Clustering}}$ is the clustering part, our prior on the latent variables and cluster centers. This energy function has several local minima due to both the large representation space, the latent variables and inducing inputs configuration in the latent space. Initializating using standard probabilistic PCA (PPCA) [8] as is the default case for the GP-LVM might not be the most beneficial, since it will create strong local minima for the representative part while not taking into account the clustering part. The conjugate gradient descent will be highly likely to get stuck in local minima close to the initialization. To give the optimization process a carte blanche if the PPCA is to biased we utilize a random initialization for the latent variables while the inducing variables are chosen using K-means to ensure an even distribution of the cluster centers.

The precision parameter of the inducing prior can be thought of as a parameter to tighten or relax our constraints on the correlation between the inducing points. It affects the separation of the clusters where a tight (large) constraint will force the latent points closer to the cluster centers and a more relaxed (small) constraint will have a lesser clustering effect.

That the number of cluster centers affects the solution is a natural and desirable property of the model. However that the ratio of data points to clusters has an effect on the solution is more subtle, it can be seen by thinking about the assumption of the inducing inputs and points as explanatory for the observed data. If the ratio is too big and the data is noisy then the explanatory capacity and the power of the cluster center prior will be reduced, since the covariance will have a natural diagonal and representative part will take over in the optimization.

B. Synthetic dataset

As a good litmus test for our approach we generated two test datasets consisting of 400 points sampled from 4 (fig2) and 10 (fig3) different two-dimensional Gaussians with random means and identical co-variances, on the unit square. The points were then linearly projected into 10 dimensions where Gaussian noise was added. During generation we also saw to it that the sampled points would overlap, to not make it too simple for a PPCA-K-means solution. We choose the following scenarios to test our model:

1) True amount of clusters: The true amount of clusters is of course an ambiguous statement. It is dependent on the definition criteria for the clusters as well as how hard the lines are drawn between categories. If one tries to group apple and pears the line is quite clear but if one chooses damaged fruit there is all of a sudden a sliding scale. The notion of the true amount remains in the questions we ask and the way we structure the data. Thus without encoding a preference on the latent variables via the prior, the representation and clustering is going to be the one most likely to have generated the observed data but not necessarily the one we deem to be the true. Of course some representations are going to be more likely than others. With this in mind we see in figure 2b that if we ask for 4 clusters corresponding to the clusters underlying the observed data our method delivers 4 Gaussian looking clusters.

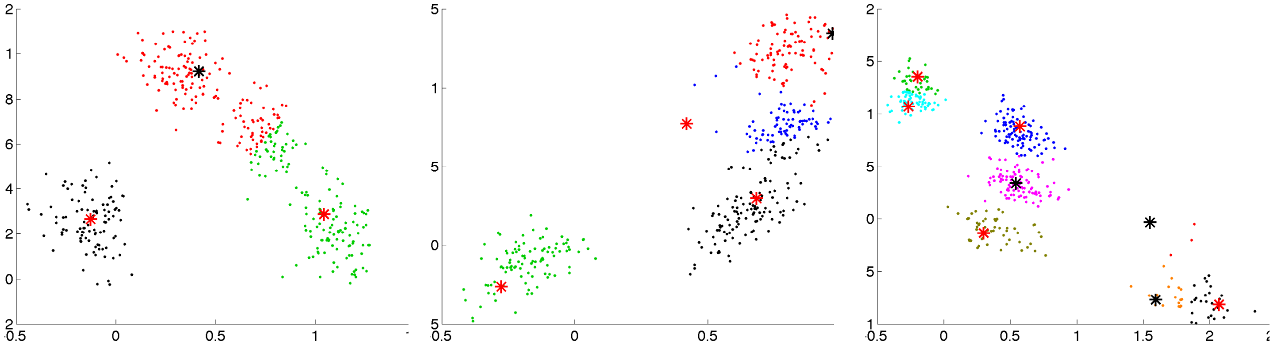
2) Too few clusters: Using too few cluster centers has the effect that the model tries to push together the data even when there is an inherent and more probable separation. Basically the inducing inputs are forced to represent more of the data, and have to stretch and compress the latent variables. This can be seen in both figure 2a and 3a.

3) Too many clusters: Enforcing a less likely clustering onto the data means less separation and divisions of natural clusters between two or more centers. Again we can see the trade-off between the representation and clustering - the representation being some larger chunks of latent points crammed together, while the cluster prior tries to divide the data, which can be seen in figure 2c and 3c.

C. Grasping data

In this experiment we apply the proposed model to the same data set as in the originally proposed method using the de-coupled prior. The data set consists of a 20 dimensional representation of a hand configuration performing different grasps. Different grasps are related to objects depending on how the objects are used.In the paper the authors use the discretization approach where the covariance of the representation and the clustering are uncoupled to facilitate learning of structure in a Bayesian Network.

By being able to discretize the data according to the learned clustered representation the authors could learn an efficient factorization of the data. In the experiments below we wish to clarify the benefits of the coupled prior by comparing the results of our method with the original one from [1], [24]

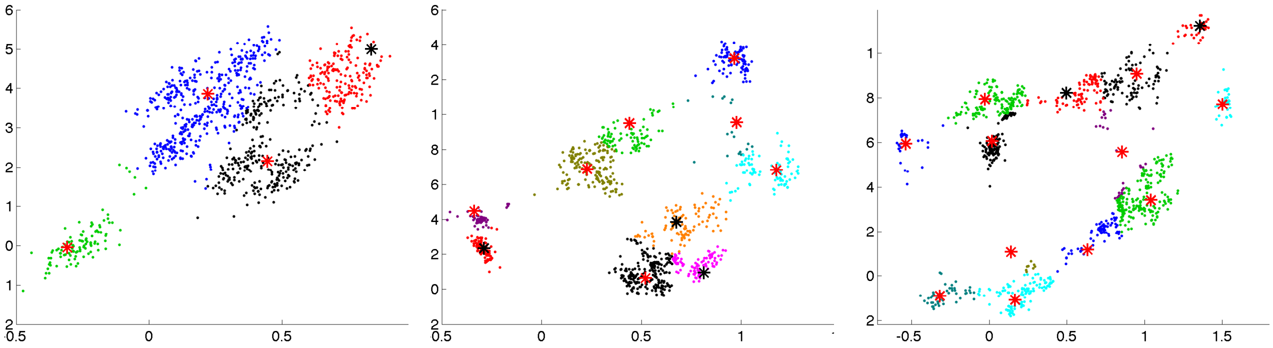


(a) 3 clusters

(b) 4 clusters

(c) 8 clusters

Fig. 2: Synthetic dataset generated by sampling from 4 Gaussians and projected into to 10 dimensions with added noise. In (a) we use 3 cluster points, which results in a pulling a part of the data as two points split the explaining of the 4 underlying clusters. (b) Using 4 cluster centers we find 4 Gaussian looking clusters. The blue center is forced left due to the constraint parameter. In (c) we choose 8 clusters this divides the data such that some bigger clusters of latent variables are explained by two cluster centers.



(a) 4 clusters

(b) 10 clusters

(c) 15 clusters

Fig. 3: Synthetic dataset generated in the same way as in figure 2 but with 10 Gaussian clusters. The behavior is consistent with the analysis in figure 2.

We first run the GPLVM algorithm without a prior using 10 inducing points on the data. Comparing the results, 4b, with the initialization, 4a, we can see that the difference is mostly in spreading the data. Introducing the un-coupled prior, 4c,d, the data points get even more drawn out and clusters starts to form. The uncoupled prior is still to weak compared to the representation part to start separating the data from the initialization and the solution ends up being close to the no-prior. The large amount of data compared to the number of centers makes the solution sensitive to the constraint parameter, the random initialization and the initial placement of the cluster centers. To get good results we have to rely on PPCA-K-means initialization, implying that the uncoupled prior in this case is not strong enough to move past uninteresting minima.

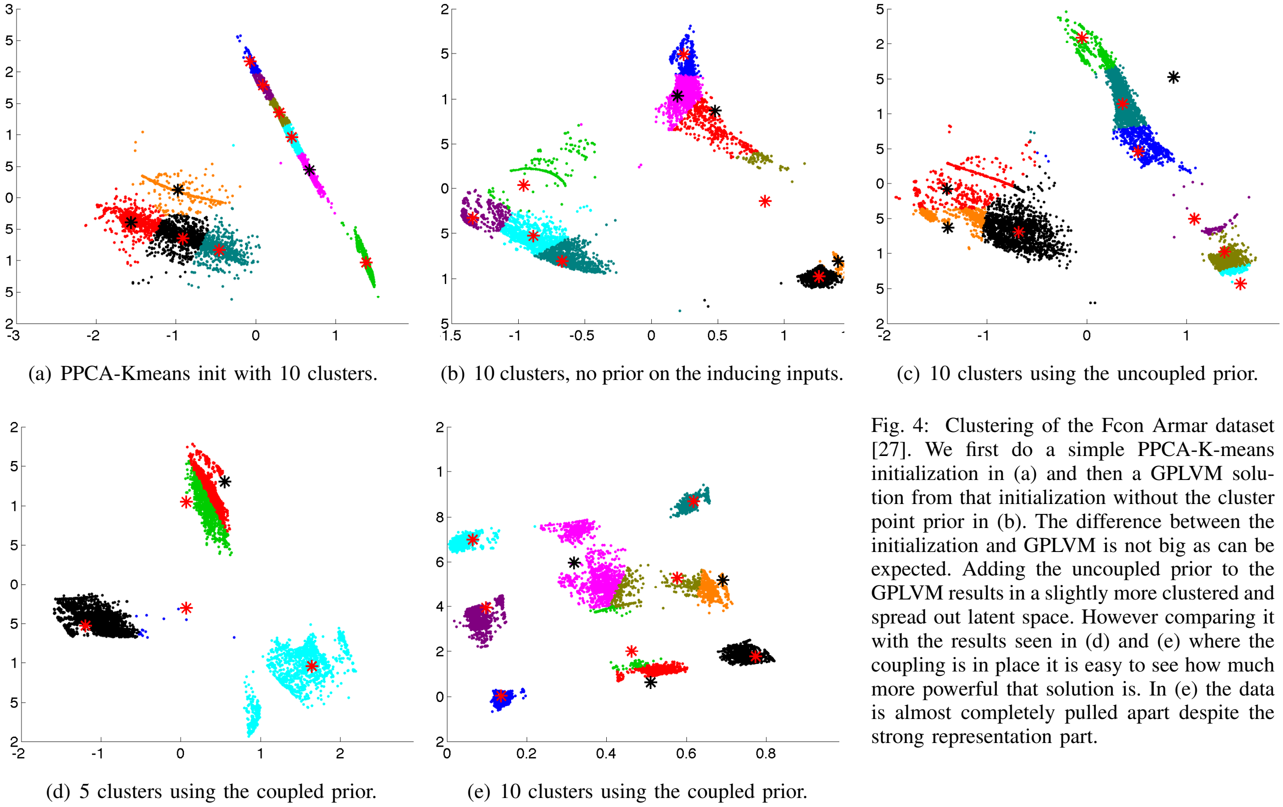
When we introduce the coupling something interesting happens, the data starts getting more pulled apart as can be seen in 4d and 4e. In 4d one cluster center is responsible only for a few data points, some observed data having high variance could explain this. Increasing the number of cluster centers, 4e, gives a more even distribution of the data points but some clusters still shares two cluster centers. This suggest that there could be a number of clusters that is more probable with respect to the representative part or that the intra-cluster variance makes the centers need more inducing inputs than one to be explained.

D. Grasp Shape Experience

In this section, we present an instance of such a problem in the context of robot grasp learning. In order to efficiently grasp new objects, it has been argued that robots should learn from experience and transfer acquired grasping skills to new objects as they come [28]–[31]. We have recently argued that it can be done by learning the shape of parts by which objects are often grasped [32]. The core of our approach is to compare the shapes of surface segments extracted around a robot's hand while it is trained to grasp different objects. Our rationale is that shapes that are observed across multiple grasps should be helpful for grasping new objects. We suggested a two-step solution to this problem: (1) measure the similarity in shape between all pairs of grasps in the dataset, and (2) cluster the grasps in the space induced by the similarity measure. As it is costly to teach grasps to a robot, the dataset is sparse, which makes clustering challenging.

In this paper we illustrate our work on the data and similarity measure presented in ICRA 2013 submission 923. The data was collected by tele-operating a robot to grasp 8 objects in several different ways[1]. Then, surface segments of

[1] http://www.csc.kth.se/~detryr/research/raster/grasp_transfer.mp4



(a) PPCA-Kmeans init with 10 clusters.

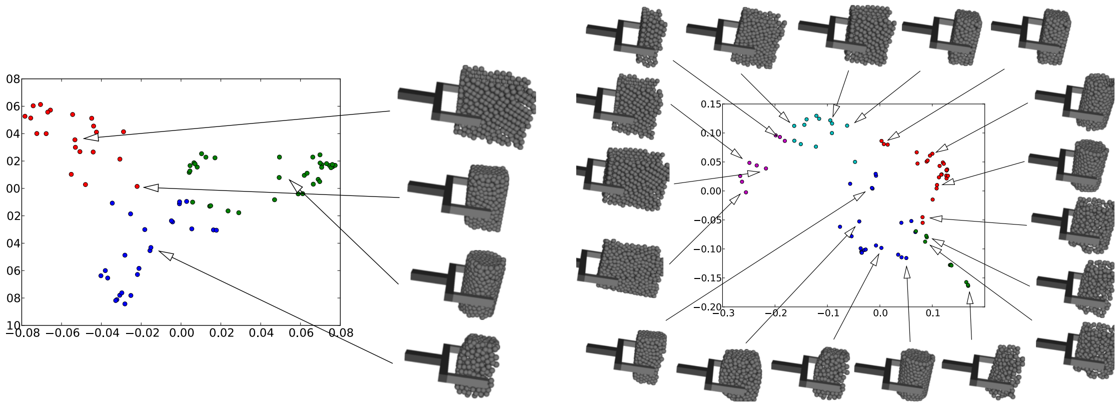
(b) 10 clusters, no prior on the inducing inputs.

(c) 10 clusters using the uncoupled prior.

(d) 5 clusters using the coupled prior.

(e) 10 clusters using the coupled prior.

Fig. 4: Clustering of the Fcon Armar dataset [27]. We first do a simple PPCA-K-means initialization in (a) and then a GPLVM solution from that initialization without the cluster point prior in (b). The difference between the initialization and GPLVM is not big as can be expected. Adding the uncoupled prior to the GPLVM results in a slightly more clustered and spread out latent space. However comparing it with the results seen in (d) and (e) where the coupling is in place it is easy to see how much more powerful that solution is. In (e) the data is almost completely pulled apart despite the strong representation part.



(a) Random initialization with 3 cluster centers

(b) Random initialization 5 clusters using prior

Fig. 5: GSE data clustering solution using 3 and 5 clusters. Using 3 cluster centers groups the data similar to a PPCA-K-means solution meaning that the representation part has a possible strong minima. The 5 cluster solution in (b) infuses a finer granularity. The representation part now has to take much more of the constraints of the clustering part into consideration and thereby forces a more grouped and different latent configuration.

various extents were segmented out of the objects around the grasping points, and a shape similarity measure was applied to all pairs of such segments.

We start out by applying PCA to the GSE similarity matrix and project the data down into 20 dimensions from where we proceed to run our model using 3 cluster centers that can be seen in figure (5). The 3 cluster centers corresponds to our prior belief that the data will generalize well across the division into three categories. The clustering turns out to be similar to the PPCA-K-means solution. If the natural division is 3, then the data could be expected to have a high variance implying strong minima for the representation part, which in practice would be close to a PPCA solution. However, when we move on to 5 clusters (5) we can see a big change from the PPCA-K-means solution. With more cluster centers, the prior now splits the data into a finer granularity forcing the representation to move away from the 3-cluster representation and find the tradeoff between the two.

We also make the observation that the clustering is consistent in that the same points roughly gets assigned to the same clusters when using a random initialization of the latent space. This means that there is a clear underlying natural grouping, which we find. Since the clustering is confident in the optimal solution it implies that the posterior has few local minima.

V. Conclusion

In this paper we have presented an approach to learn a sparse low-dimensional representation of high-dimensional robotic grasp data. The model is general and provides a summarizing of any continuous data. We have shown how the method extends and outperforms our previous method and provide a thorough analysis of the model. In future work we are interested in evaluating the possibilities of integrating the coupled inducing point prior model with the full Bayesian variational approach to the GP-LVM [33]. Further, we are interested in integrating the proposed prior with a dynamic model which should facilitate learning of key-frames for dynamic data. In terms of application domains we believe that our approach has the potential to provide beneficial representations for many tasks. Initially we aim to target high-dimensional data for grasping and motion modeling.

References

[1] D. Song, Carl Henrik Ek, K. Huebner, and D Kragic. Multivariate discretization for bayesian network structure learning in robot grasping. In *IEEE International Conference on Robotics and Automation*, pages 1944–1950. IEEE, 2011.

[2] S Calinon, F D'halluin, E L Sauser, D G Caldwell, and A G Billard. Learning and reproduction of gestures by imitation: An approach based on Hidden Markov Model and Gaussian Mixture Regression. *IEEE Robotics and Automation Magazine*, 17(2):44–54, 2010.

[3] Stefan Schaal, AJ Ijspeert, and A Billard. Computational approaches to motor learning by imitation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):537, 2003.

[4] Guoliang Luo, N. Bergström, Carl Henrik Ek, and D Kragic. Representing actions with Kernels. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2028–2035, 2011.

[5] Richard Bellman. On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, August 1952.

[6] Neil D Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[7] M Cox and T Cox. Multidimensional scaling. *Handbook of data visualization*, January 2008.

[8] Michael E Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[9] Neil D Lawrence. A Unifying Probabilistic Perspective for Spectral Dimensionality Reduction: Insights and New Models. *The Journal of Machine Learning Research*, 98888, June 2012.

[10] Neil D Lawrence. Spectral dimensionality reduction via maximum entropy. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 51–59, 2011.

[11] N Dalal and B Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[12] G Mori, S Belongie, and J Malik. Efficient shape matching using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1832–1837, 2005.

[13] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[14] P Leray and O Francois. BNT structure learning package: Documentation and experiments. Technical report, 2004.

[15] Renaud Detry, Carl Henrik Ek, Marianna Pronobis, Justus Piater, and Danica Kragic. Generalizing Grasps Across Partly Similar Objects. In *IEEE International conference on robotics and automation*, May 2012.

[16] Carl Henrik Ek. Shared Gaussian Process Latent Variable Models. *PhD Thesis*, 2009.

[17] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[18] Neil D Lawrence and Joaquin Quiñonero-Candela. Local distance preservation in the GP-LVM through back constraints. In *Proceedings of the 23rd international conference on Machine learning*, pages 513–520, New York, NY, USA, 2006. ACM.

[19] Raquel Urtasun and Trevor Darrell. Discriminative Gaussian process latent variable model for classification. *International Conference on Machine Learning*, page 934, 2007.

[20] Raquel Urtasun, David J Fleet, Andreas Geiger, Jovan Popovic, Trevor Darrell, and Neil D Lawrence. Topologically-Constrained Latent Variable Models. *International Conference on Machine Learning*, 2008.

[21] J.M Wang, David J Fleet, and A Hertzmann. Gaussian Process Dynamical Models for Human Motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):283–298, 2008.

[22] Raquel Urtasun, David J Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:238–245, 2006.

[23] S Bitzer and Sethu Vijayakumar. Latent Spaces for Dynamic Movement Primitives. *International Conference on Humanoid Robots*, January 2009.

[24] Carl Henrik Ek, Dan Song, and Danica Kragic. Learning Conditional Structures in Graphical Models from a Large Set of Observation Streams through efficient Discretisation. In *IEEE International Conference on Robotics and Automation, Workshop on Manipulation under Uncertainty*, pages 1–5. Royal Institute of Technology, 2011.

[25] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

[26] Michalis K Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. *Journal of Machine Learning Research - Proceedings Track*, 5:567–574, 2009.

[27] Dan Song, Carl Henrik Ek, Kai Huebner, and Danica Kragic. Multivariate discretization for Bayesian Network structure learning in robot grasping. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1944–1950, 2011.

[28] Jefferson Coelho, Justus Piater, and Roderic Grupen. Developing haptic and visual perceptual categories for reaching and grasping with a humanoid robot. In *Robotics and Autonomous Systems*, volume 37, pages 7–8, 2000.

[29] I. Kamon, T. Flash, and S. Edelman. Learning to grasp using visual information. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 2470–2476, 1996.

[30] A. Morales, E. Chinellato, A. H. Fagg, and A. P. del Pobil. Using experience for assessing grasp reliability. *International Journal of Humanoid Robotics*, 1(4):671–691, 2004.

[31] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic Grasping of Novel Objects using Vision. *International Journal of Robotics Research*, 27(2):157, 2008.

[32] Renaud Detry, Carl Henrik Ek, Marianna Madry, Justus Piater, and Danica Kragic. Generalizing grasps across partly similar objects. In *IEEE International Conference on Robotics and Automation*, 2012.

[33] Michalis Titsias and Neil D Lawrence. Bayesian Gaussian Process Latent Variable Model. In *International Conference on Artificial Inteligence and Statistical Learning*, 2010.