Multi-level Visualisation using Gaussian Process Latent Variable Models

Shahzad Mumtaz[1], Darren R. Flower[2] and Ian T. Nabney[1]
[1]*Non-Linearity and Complexity Research Group, Aston University, Birmingham B4 7ET, U.K.*
[2]*School of Life and Health Sciences, Aston University, Birmingham B4 7ET, U.K.*

Keywords: Multi-level Gaussian Process Latent Variable Model, *k*-means, Gaussian Mixture Model, Trustworthiness, Continuity, Negative Log-likelihood, Visualisation Distance Distortion, Mean Relative Rank Errors, Major Histocompatibility Complex.

Abstract: Projection of a high-dimensional dataset onto a two-dimensional space is a useful tool to visualise structures and relationships in the dataset. However, a single two-dimensional visualisation may not display all the intrinsic structure. Therefore, hierarchical/multi-level visualisation methods have been used to extract more detailed understanding of the data. Here we propose a multi-level Gaussian process latent variable model (MLGPLVM). MLGPLVM works by segmenting data (with e.g. *K*-means, Gaussian mixture model or interactive clustering) in the visualisation space and then fitting a visualisation model to each subset. To measure the quality of multi-level visualisation (with respect to parent and child models), metrics such as trustworthiness, continuity, mean relative rank errors, visualisation distance distortion and the negative log-likelihood per point are used. We evaluate the MLGPLVM approach on the 'Oil Flow' dataset and a dataset of protein electrostatic potentials for the 'Major Histocompatibility Complex (MHC) class I' of humans. In both cases, visual observation and the quantitative quality measures have shown better visualisation at lower levels.

1 INTRODUCTION

Recent advances in sciences as diverse as astronomy, biology, weather forecasting and economics have led to the generation and storage of large high-dimensional datasets. Such datasets have not only presented new challenges for researchers but also created new openings for theoretical developments (Donoho, 2000).

In the machine-learning domain, projection of large and high-dimensional datasets onto a single two-dimensional plot is a useful and popular way of extracting intrinsic structures. A single two-dimensional visualisation plot (using either linear or non-linear visualisation methods) seldom captures all the intrinsic structure in complex datasets. Consider, for example, situations where a single two-dimensional visualisation plot of a complex and large high-dimensional dataset can only show major clusters whereas a hierarchical or multi-level visualisation approach could show more interesting detailed structures. Such a tree-like visualisation model uses a root-level visualisation plot to give a high-level overview of a dataset and child visualisation plots for more detailed views.

In the last two decades, unsupervised hierarchical visualisation or clustering models have been introduced and are reviewd as in (Vicente and Vellido, 2004; Murtagh and Contreras, 2011). Two well known visualisation approaches are: probabilistic hierarchical models and multi-level models. Both categories are based on a top-down divisive approach to build a tree-like visualisation structure. The term 'hierarchical' is used here to indicate a probabilistic way of assigning data points (also known as soft memberships) to child visualisations whereas the term 'multi-level' indicates that we partition the dataset into subsets (using hard clustering) for building the child visualisations.

Probabilistic hierarchical models use density estimation to build a complete and consistent hierarchical model of the data. For example, a hierarchical mixture of latent variables model (HMLVM) uses a single linear latent variable model to obtain a top-level visualisation plot and uses a probabilistic mixture of latent variable models to represent the lower-level models (Bishop and Tipping, 1998). This process can be continued recursively where child models at level $N+1$ represent a mixture decomposition of the parent model at level N. A non-linear extension of the HM-

Mumtaz S., Flower D. and Nabney I..
Multi-level Visualisation using Gaussian Process Latent Variable Models.
DOI: 10.5220/0004686801220129
In *Proceedings of the 5th International Conference on Information Visualization Theory and Applications* (IVAPP-2014), pages 122-129
ISBN: 978-989-758-005-5

LVM was proposed in (Tino and Nabney, 2002). This non-linear variant uses the Generative Topographic Mapping (GTM) as a building block to represent visualisation of high-dimensional datasets in a tree structure of multiple two-dimensional plots and is known as hierarchical GTM (HGTM). A multiple manifold learning framework based on HGTM was proposed in (Wang et al., 2008): this uses an approximation method to initialize a hierarchical model to represent each sub-model as a single manifold. These probabilistic hierarchical visualisation methods are based on soft clustering.

Multi-level approaches based on the Self Organizing Map (SOM) use hard clustering where each of the low-dimensional code vectors represents a single cluster and the corresponding points in the high-dimensional space are projected onto a separate lower-level two-dimensional plot to show subclusters (Miikkulainen, 1990; Versino and Gambardella, 1996; Lampinen and Oja, 1992).

In this paper we propose a multi-level visualisation with Gaussian Process Latent Variable Models (MLGPLVM). In MLGPLVM we visualise a complete dataset at the root level which gives a high-level view. We apply clustering on this root-level visualisation plot in order to create a hard partition into subsets. Each subset is then used to build a child-level visualisation model. These child-level or subset-level models may help us to visualise detailed local structures or clusters in the dataset. We take this approach since there is no simple way of modifying the GPLVM to take account of 'soft' cluster membership, as would be needed for a probabilistic hierarchy. In our experiments we used K-means, Gaussian mixture models and interactive clustering for partitioning the data. An interactive clustering approach permits the user to draw polygons on the visualisation plot to identify clusters. The benefit of clustering in the visualisation space is that the user can see the nature of the segmentation, and correct it manually if necessary. To measure the effectiveness of our proposed multi-level visualisation approach we apply it to two datasets and compute quantitative visualisation quality measures.

The outline of the paper is as follows. In Section 2, we briefly discuss the theory of GPLVM with sparse and back-constrained extensions. In Section 3 we explain our proposed MLGPLVM approach with brief details about the clustering approaches we used in our analysis. Section 4 defines the quantitative quality matrics used in our analysis to show the effectiveness of our approach. In Section 5 we describe briefly the 'Oil Flow' and the 'MHC class-I' datasets. In section 6 we explain the visualisation experiments and discuss the results for both datasets. Section 7 concludes our paper with potential advantages and disadvantages of this new method and proposes future work.

2 GAUSSIAN PROCESS LATENT VARIABLE MODEL (GPLVM)

2.1 Probabilistic Dimensionality Reduction Process

A latent-variable model is used to represent a dataset $Y \in \mathbb{R}^{N\times D}$ with N data points in D dimensions by mapping from the low-dimensional $X \in \mathbb{R}^{N\times Q}$ with N data points in Q dimensions (usually $Q = 2$). The mapping between a low-dimensional data point x_n and a high-dimensional data point y_n is defined by

$$y_{ni} = f_i(x_n) + \eta_{ni}, \tag{1}$$

where η_{ni} represents noise for the ith feature of the nth data point. The noise model we assume is a Gaussian with zero mean and inverse variance β. So the conditional distribution of a data point y_n given a data point x_n is

$$p(y_n|x_n) = \prod_{i=1}^{D} N(y_{ni}|f_i(x), \beta^{-1}). \tag{2}$$

If the mapping is assumed to be linear $f_i(x_n) = w_i^T x_n$, and the latent variable x is drawn from a Gaussian prior (with zero mean and unit variance) then the maximum likelihood solution of the model represents the principal subspaces of the data (Tipping and Bishop, 1999): this is Probabilistic Principal Component Analysis (PPCA). Integrating out the latent variable gives the marginal likelihood,

$$p(y_n) = \int p(y_n|x_n)p(x_n)dx_n. \tag{3}$$

So the marginal distribution for the complete dataset is given as

$$p(Y) = \prod_{n=1}^{N} p(y_n). \tag{4}$$

In a standard latent-variable model we use maximum likelihood to optimize weights and marginalize out latent variables.

2.2 Standard GPLVM

A non-linear extension of PPCA is the Gaussian process latent variable model (GPLVM), which uses a smooth mapping from the latent space to the data space. In the GPLVM instead of optimizing weights

they are marginalized out and instead of marginalizing over the latent space it is optimized (i.e. the position of each point in the latent space is optimized). A conjugate prior over the weights is chosen, taking the form of a spherical Gaussian distribution for each dimension

$$p(w) = \prod_{i=1}^{D} N(w_i|0,I). \quad (5)$$

The likelihood after marginalizing the weights is

$$p(Y|X) = \prod_{i=1}^{D} N(y_{(:,i)}|f_i(x),\beta^{-1}), \quad (6)$$

where $p(y_{(:,i)}) = N(y_{(:,i)}|0,XX^T+\beta^{-1}I)$ represents a distribution over a single feature in the data space. GPLVM uses the following likelihood function to optimize the latent variables (similar to the likelihood used in (Tipping and Bishop, 1999))

$$L = -\frac{DN}{2}\log(2\pi) - \frac{D}{2}\log(\det K) - \frac{1}{2}\mathrm{tr}(K^{-1}YY^T). \quad (7)$$

If $K = XX^T + \beta^{-1}I$ is a linear kernel, then it is similar to PPCA. But for the GPLVM a non-linear RBF kernel is used as explained in (Lawrence and Hyvrinen, 2005; Lawrence, 2004) and then the optimization of the latent variable can be achieved using the non-linear optimization algorithm using the gradient of the likelihood with respect to the kernel. The kernel parameters are optimized by combining this gradient with the derivative of the kernel parameters using the chain rule. The gradient calculation uses the inverse of the kernel matrix; it has $O(N^3)$ complexity thereby making it less practical for large datasets. Due to this complexity, a GPLVM is usually trained using sparse approximations where a small subset of data points of size $k << N$ known as 'inducing points' or the 'active set' is used to reduce the complexity from $O(N^3)$ to $O(k^2N)$. For sparse approximation the standard GPLVM uses informative vector machine (IVM) (where data points are chosen sequentially based on the reduction of the posterior process's entropy (Lawrence et al., 2003)). But we used the GPLVM with the improved sparse approximation approach compared to IVM based sparse approach as proposed in (Lawrence, 2008). This new improved approximation process was originally proposed for Gaussian process regression and is based on the unified view process as explained in (Quionero-candela et al., 2005).

2.3 Preserving Local Distances

The standard GPLVM uses mapping from the latent space to the data space for the training data only which constraints distant points in the data space to be distant in the latent space at the expense of the local distance preservation. When users visualize data, it is the local structure that is most relevant to their analysis (for example, when they identify clusters). Therefore we use the variant of GPLVM where constrained smooth mapping as in Neuroscale (Lowe and Tipping, 1996) is employed to overcome the problem of local distance preservation because the data points x are no longer freely optimized. Instead they are the image of points y in the data space under the non-linear function like a Radial Basis function (RBF) kernel or multi-layer kernel perceptron (MLP). This constrained mapping (also known as back-constraint) ensures that the data points which are close in the visualisation space are also close in the data space. We use an MLP kernal as a back-constraint. Further details on preserving local distances with GPLVM can be found in (Lawrence, 2006).

3 MULTI-LEVEL GPLVM

We propose here a multi-level GPLVM (MLGPLVM) visualisation method for analysing complex datasets. The fundamental building block of our proposed visualisation model is GPLVM with back constraint. The steps involved to generate a MLGPLVM visualisation are:

1. Generate a root visualisation plot using standard GPLVM with back constraint which represents the mapping from data space Y to latent visualisation space X.
2. Cluster the data in the two-dimensional visualisation space.
3. Partition the high-dimensional dataset Y into subsets based on the clustering information at step 2.
4. Build a separate visualisation model for each subset to generate local visualisation sub-models.
5. Repeat steps 2 to 4 on local visualisation sub-models if required to add more levels in the multi-level visualisation.

The structure of the multi-level GPLVM visualisation approach is shown in Figure 1 where the top-level visualisation shows only three major clusters whereas second-level visualisations show more detailed local structures (e.g. cluster-1 at level-1 shows four clear sub-clusters at level-2). The similar high-level and detailed views have been found in real datasets (as shown in Section 6). For the purpose of comparison and findnig the effectiveness of our proposed ML-GPLVM approach, we use three different clustering

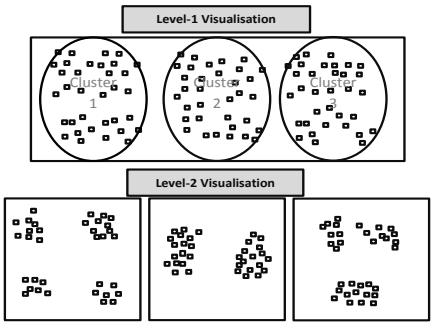


Figure 1: Structure of multi-level GPLVM visualisation.

methods to partition a dataset to generate lower-level visualisations: $K-$means, Gaussian mixture models and interactive clustering by drawing polygons.

3.1 Cluster Identification

Traditional well known clustering approaches such as $K-$means and Gaussian mixture models can be used for generating subsets of a dataset required to generate level-2 visualisations.

K-means is an unsupervised clustering method where K cluster centres are initialised as randomly selected data points (MacQueen, 1967) and is trained iteratively in a two step process: in the first step the cluster centres are kept fixed to compute cluster memberships and in the second step cluster centres are updated to be the mean of the assigned data points. This continues until no cluster memberships change. This algorithm faces certain limitations:there is no principled way of determining the true number of clusters, it is sensitive to outliers and it is difficult to identify true boundaries of clusters due to the unknown optimal number of clusters.

As argued in (Iwata et al., 2012; Lawrence and Hyvrinen, 2005), the standard GPLVM considers a single Gaussian in the latent space to represent a more complex dataset. Therefore applying Gaussian mixture model (GMM) could not be a useful way to identify clusters in the latent space. We also observed the same by applying variational mixture modeling (Corduneanu and Bishop, 2001) approach on the latent visualisation space in order to determine the true number of Gaussians. As GPLVM gives good clustering results in the visualisation latent space (Lawrence, 2006) therefore only for the purpose of comparison with our proposed interactive region selection approach we applied finite Gaussian mixture models (McLachlan and Basford, 1988) for the purpose of segmenting the dataset based on the clustering in the visualisation latent space. Visualisation results using $K-$means and GMM under MLGPLVM framework are available in a technical report (Mumtaz et al., 2013).

As K-means and GMM have limitations in defining the true number of clusters and identification of cluster boundaries onto visualisation space, we therefore propose involving user in identifying clusters by drawing polygon regions interactively on visualisation space can give us better clusters by defining true boundaries (Larkin and Simon, 1987). This interactive approach of defining clusters using human perception requires no mathematical or statistical modeling but enables the user to control the drill-down directly (Shneiderman, 2002). Clusters were identified by using a polygon region-selection approach proposed in (Hormann and Agathos, 2001). We compute mapping precision for the GPLVM (see Figures 2 and 4 where it is represented as grey background). This can be helpful to identify clusters interactively.

4 MLGPLVM VISUALISATION QUALITY MEASURES

Evaluating visualisation performance quantitatively is necessary but difficult because there is no true target output. The log likelihood is a global model fit measure. Because visual interpretation is often focused on clusters of points, we need to use metrics that capture local neighbourhood preservation. To compare the mapping at different levels of the hierarchy we use local quality measures such as visualisation distance distortion, trustworthiness, continuity and mean relative rank errors. We briefly explained them in the following sub-sections whereas detailed description of each of these is avaialable in a technical report (Mumtaz et al., 2013).

4.1 Visualisation Distance Distortion

The visualisation distance distortion (VDD) measure is used to compare the distances between the points in the data space Y and the projection space X for each data point and its k nearest neighbours. VDD is calculated as the norm of the difference vectors between the scaled distances in the data space and the visualisation latent space. The scaled distances are used to make the distance comparable between the data space and the latent visualisation space. The idea of VDD is similar to the projection precision score (PPS) as discussed in (Schreck et al., 2010) where it was used to observe projection precision quality on the visualisation plot. We compute the sum of the VDD values of all the data points in a subset to compare the subset

visualisation quality in the level-1 and level-2 visualisation plots.

4.2 Rank Based Neighbourhood Visualisation Quality Measures

In the information visualisation domain, two well known visualisation quality measures based on comparing neighbourhoods in the data space Y and projection space X are trustworthiness and continuity (Venna and Kaski, 2001). Mapping is assumed to be trustworthy if k-neighbourhood in the visualised space matches in the data space but if the k-neighbourhood in the data space matches the visualized space it maintains continuity. Two another quality measures mean relative rank errors (MRREs) with respect to data and latent space are also used (Lee et al., 2007). MRREs are computed using the exact rank position differences within the k-neighbourhood of the data space and visualisation space.

The higher the value of trustworthiness and continuity (ranges from 0 to 1) the better the proximity preservation is whereas for mean relative rank errors the lower the measure is the better the proximity is preserved.

5 DATASETS

For evaluating the MLGPLVM visualisation, we consider two datasets: 'oil flow' and 'MHC class I'.

5.1 Oil Flow Dataset

The 'oil flow' dataset is a twelve-dimensional dataset collected from a simulation of a non-invasive monitoring system (Bishop and James, 1993) and used previously to demonstrate the Generative Topographic Mapping (Bishop and Svensen, 1998) and hierarchical visualisation algorithms (Bishop and Tipping, 1998; Tino and Nabney, 2002). The dataset comprises 1000 data points and it was generated artificially in a multiphase flow configuration of three liquids (oil, water and gas) by defining three known classes: homogeneous, annular and laminar. From knowledge of the generation process, the data is expected to lie on low-dimensional manifolds.

5.2 Major Histocompatibility Complex class-I

The second dataset is related to MHC class-I that we used previously to demonstrate variants of generative topographic mapping (GTM) and GTM with simultaneous feature saliency (Mumtaz et al., 2012).

Here we briefly explain the process of generating a Poisson-Boltzmann electrostatic potential data for the MHC class-I genes. MHC genes in humans are known as Human Leukocyte Antigen (HLA). We modelled $3,944$ three-dimensional protein structures of HLA class I (1236 for HLA-A, $1,779$ for HLA-B and 929 for HLA-C) using homology modelling (as in (Doytchinova et al., 2004)). We then computed a Poisson Boltzmann electrostatic potential for all the modelled proteins by placing a three dimensional grid box (with $17^3 = 2,601$ grid points) around the top surface (covering $\alpha 1$ and $\alpha 2$ region). We are interested in analysing interaction with other molecules therefore ignored all those grid points that were inside the van der Waals surface of the target protein. We end up with $2,418$ grid points which were definitely outside the van der Waals surface of all the target proteins. Each grid point worked as a variable in our dataset and each protein is represented as a row in our dataset. Further details for the data generation process can be found in (Mumtaz et al., 2012; Mumtaz et al., 2013).

6 EXPERIMENTS

We used full GPLVM for the 'oil flow' dataset and sparse GPLVM for the 'HLA class I' dataset for visualisation under the MLGPLVM framework. The 'oil flow' dataset has fewer data points and fewer variables and therefore, applying full GPLVM, it was possible to generate results in a matter of a few hours whereas applying full GPLVM on the MHC class-I dataset could take months as this has thousands of data points with more than a couple of thousand variables. Therefore, we used sparse GPLVM (as briefly explained in Section 2.2) for the 'HLA class-I' dataset to create visualisations under the MLGPLVM framework in a matter of few hours. Each visualisation model under the MLGPLVM framework is trained (with 1500 iterations for the 'oil flow' dataset and 2000 iterations for 'HLA-Class I' dataset) using the scaled conjugate gradient optimisation method.

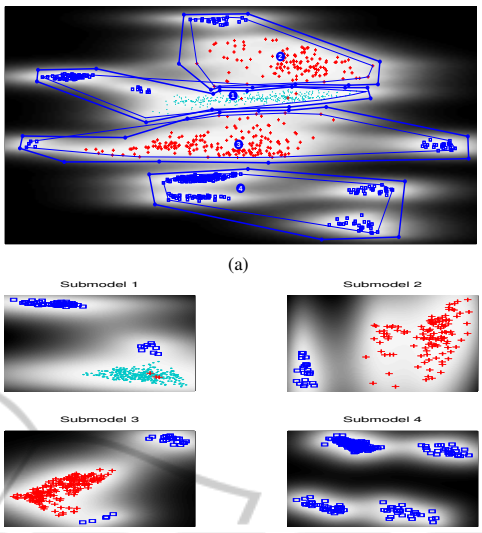
To evaluate the visualisation quality of MLGPLVM, we compute the quality measures defined in Section 4 for a range of number of neighbours $k = 5, 10, \ldots, 50$ for each cluster as indicated at level-1 and its corresponing subset visualisation at level-2. We intially computed the mean of these measures (over k) (see in (Mumtaz et al., 2013)) and then summarised them by taking mean across all the subsets to compare the performance across levels (see Table 1).

6.1 Results

Root level visualisation plot for both the datasets (the 'oil flow' and 'HLA class I') have shown that the three classes in each dataset case are well separated with a number of clusters observed for each class (see Figure 2(a) for 'oil flow' dataset and Figure 4(a) for 'HLA class I' dataset whereas applying linear visualisation such as Principal Component Analysis (PCA) has not shown clear separation of the alleles of each HLA gene but instead the alleles of all three genes overlap (as shown in Figure 3)). We present in this paper second level visualisation results (see Figures 2(b) and 4(b)) generated by applying interactive clustering at the root level visualisation plots (See (Mumtaz et al., 2013) for second level visualisation results generated by applying K-means and GMM clustering on the root level visualisation plots). Experiments were performed with a different numbers of clusters at root level but here we present only the results with 4 clusters used to generate the second-level visualisation.

Visual inspection of all these local second-level visualisation models show that they provide a more detailed clustering/visualisation structure compared to the root visualisation. Tabel 1 show the mean of the quality measure over clusters at level-1 and over submodels at level-2 using three clustering approaches under MLGPLVM framework: K-means, GMM and interactive clustering. We compute per point negative log-likelihood for each cluster at level-1 and submodel at level-2. The mean negative log-likelihood (per point) is then computed over clusters at level-1 and sub-models at level-2 and presented in terms of ratio increase or decrease of level-2 with respect to level-1. For the 'oil flow' dataset we observe that the mean quantitative quality matrics (over clusters) to compare visualisation quality across levels appeared better for level-2 compared to level-1 visualisations (see Table 1). For the 'HLA class I' dataset trustworthiness, continuity and TVDD are observed better for all the second level visulisation models using each of the clustering algorithm applied to generate second visualisations (as shown in Table 1). The other measures such as MRREs and negative log-likelihoods were slightly better for level-1.

By rigorous state-of-the-art analysis of projected properties, we have identified clusters corresponding to the three class I human MHC loci, and sub groups therein. It is notable that the analysis recovers the HLA-A; HLA-B, and HLA-C alleles without prior knowledge of such a division at the root level visualisation and hence such grouping is refined by adding the lower level visualisations. This gives confidence to any assertion we might make regarding the division



(a)

(b)

Figure 2: MLGPLVM visualisation of 'oil flow' dataset with K-means clustering. (a) Root visualisation plot: blue circles with numbers indicate cluster centres and blue lines represent cluster boundaries, cyan dots ('.') for 'Homogeneous', red plus signs ('+') for 'Annular', blue squares ('□') for 'Laminar', and the grey background shows mapping precision (lighter regions correspond to better precision in mapping). (b) Level-2 visualisation.

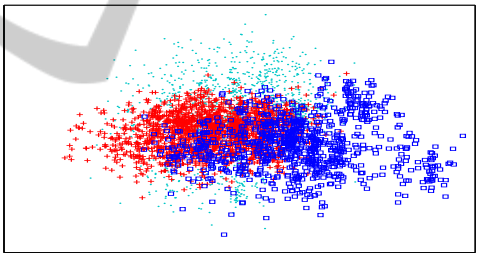
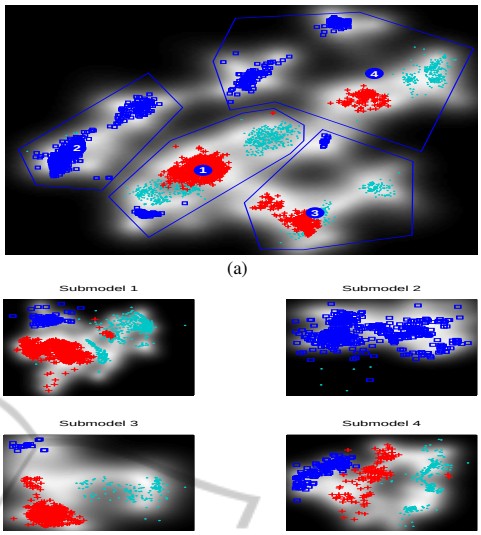


Figure 3: PCA visualisation (cyan dots ('.') for HLA-A, red plus signs ('+') for HLA-B and blue squares ('□') for HLA-C).

of the allele population into structurally and functionally similar sub-groups. The results of our analysis are fully consistent with both the choice of Poisson-Boltzmann electrostatic potential as a meaningful indicator of molecular spatial interactions and with the sophisticated methods of data reduction used to derive the final clustering. It is also consistent with the evolutionary argument, since it suggests that with the exception of a handful of genes, the three class-I loci exhibit quite distinct specificities for peptides and TCRs, since redundant specificities shared between loci would be not favourable since it would reduce the diversity of peptides that a host could recognize and respond to and thus the diversity of pathogens it could effectively combat. It will be interesting to extend our analysis to investigate the structural basis for this phenomenon.

Table 1: Visualisation Quality Matrics of MLGPLVM evaluation for two datasets (i.e. Oil flow and HLA class 1 dataset). ('Trust' for trustworthiness, 'Cont' for continuity, 'MRREd' and 'MRREl' for mean relative rank errors with respect to data space and latent space respectively, 'TVDD' for total visualsiation distance distortion) and 'NLL' for negative log likelihood. Measures such as 'Trust' and 'Cont' the higher the better visualisation whereas measures such as 'MRREd', 'MRREl' 'TVDD' and 'NLL' the lower the better the visualisation.

Clustering		**Oil Flow**		**HLA class 1**	
		Level-1	**Level-2**	**Level-1**	**Level-2**
K-means	Trust	0.9484	**0.9705**	0.7895	**0.8191**
	Cont	0.9409	**0.9749**	0.8022	**0.8406**
	MRREd	0.1974	**0.1352**	**0.0440**	0.0445
	MRREl	0.1956	**0.1346**	**0.0407**	0.4130
	TVDD	0.5807	**0.4752**	0.8121	**0.7978**
	NLL	1.0000	**0.9911**	**1.0000**	1.0295
GMM	Trust	0.9501	**0.9729**	0.7896	**0.8285**
	Cont	0.9428	**0.9743**	0.8022	**0.8406**
	MRREd	0.1661	**0.1103**	**0.0440**	0.0445
	MRREl	0.1658	**0.1103**	**0.0407**	0.0413
	TVDD	0.5806	**0.4853**	0.8121	**0.7978**
	NLL	1.0000	**0.9931**	**1.0000**	1.0303
Interactive	Trust	0.9469	**0.9788**	0.7932	**0.8243**
	Cont	0.9353	**0.9817**	0.8055	**0.8341**
	MRREd	0.1682	**0.1037**	0.0439	**0.0433**
	MRREl	0.1700	**0.1043**	**0.0405**	0.0408
	TVDD	0.5990	**0.4360**	0.8112	**0.7917**
	NLL	1.0000	**0.9918**	**1.0000**	1.1661



(a) (b)

Figure 4: MLGPLVM visualisation of 'MHC class-I' dataset with interactive clustering. (a) Root visualisation plot (numbered blue circles with numbers indicate cluster centres and blue lines represent cluster boundaries, cyan dots ('.') for HLA-A, red plus sign ('+') for HLA-B and blue squares ('□') for HLA-C, and the grey background show mapping precision (lighter regions show better precision in mapping)). (b) Level-2 visualisation.

7 CONCLUSIONS AND FUTURE WORK

In this paper we propose a multi-level visualisation using Gaussian process latent variable models where the root-level visualisation gives an overview of the complete dataset and the second-level view gives refined visualisation results for the clustered data. We experiment the generation of second level visualisations using three different clustering algorithms: *K*-means, GMM and interactive. Both the datasets we used for the demonstration of MLGPLVM have shown promising improvements on the root-level visualisation by giving refined lower level visualisations. We briefly conclude here the results of 'MHC class-I' dataset by saying that the present approach, which combines the established protocol of chemical landscape profiling with calculated properties and state-of-the-art data visualization and clustering, is promising. We will seek to extend this to approach and apply it to the classification of MHC alleles in terms of peptide specificity, TCR specificity, and antibody interaction and use it to investigate practical problems in epitope prediction, solid organ and bone marrow transplantation, mate-choice, and MHC-mediated adverse drug reactions.

We have also incorporated the code of MLGPLVM in our recently developed visualisation tool called Data Visualisation and Modelling System (DVMS). This tool is freely accessible from our website[1]. We plan to extend this work with a probabilistic hierarchical visualisation framework (based on soft assignments to child models).

REFERENCES

Bishop, C. and James, G. (1993). Analysis of Multiphase Flows Using Dual-Energy Gamma Densitometry and Neural Networks. *Nuclear Instruments and Methods in Physics Research*, 327(2-3):580–593.

Bishop, C. M. and Svensen, M. (1998). GTM: The generative topographic mapping. *Neural Compuatation*, 10(1):215–234.

Bishop, C. M. and Tipping, M. E. (1998). A hierarchical latent variable model for data visualization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3):281–293.

Corduneanu, A. and Bishop, C. M. (2001). Variational bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA.

Donoho, D. L. (2000). High-dimensional data analysis: the curses and blessings of dimensionality. In *American*

[1] http://www.aston.ac.uk/ncrg

Mathematical Society Conf. Math Challenges of the 21st Century.

Doytchinova, I. A., Guan, P., and Flower, D. R. (2004). Identifying human MHC supertypes uisng bioinformatics methods. *The Journal of Immunology*, 172:4314–4323.

Hormann, K. and Agathos, A. (2001). The point in polygon problem for arbitrary polygons. *Comput. Geom. Theory Appl.*, 20(3):131–144.

Iwata, T., Duvenaud, D., and Ghahramani, Z. (2012). Warped mixtures for nonparametric cluster shapes. *arXiv preprint arXiv:1206.1846*.

Lampinen, J. and Oja, E. (1992). Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, 2:261–272.

Larkin, J. H. and Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1):65–100.

Lawrence, N. and Hyvrinen, A. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816.

Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, pages 329–336, Cambridge, MA. MIT Press.

Lawrence, N. D. (2006). Local distance preservation in the GPLVM through back constraints. In *In ICML*, pages 513–520. ACM Press.

Lawrence, N. D. (2008). Large scale learning with the gaussian process latent variable model. Technical report, university of sheffield, United Kingdom.

Lawrence, N. D., Seeger, M., and Herbrich, R. (2003). Fast sparse gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems 15*, pages 609–616. MIT Press.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8:995–1005.

Lowe, D. and Tipping, M. E. (1996). Neuroscale: Novel topographic feature extraction using RBF networks. In *NIPS*, pages 543–549.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. M. and Neyman, J., editors, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - Vol. 1*, pages 281–297. University of California Press, Berkeley, CA, USA.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

Miikkulainen, R. (1990). Script recognition with hierarchical feature maps. *Connection Science*, 2:83–101.

Mumtaz, S., Flower, D. R., and Nabney, I. T. (2013). Multi-level visualisation. Technical report, NCRG, Aston University, Birmingham, UK. http://eprints.aston.ac.uk/.

Mumtaz, S., Nabney, I. T., and Flower, D. (2012). Novel visualization methods for protein data. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*, pages 198 –205.

Murtagh, F. and Contreras, P. (2011). Methods of hierarchical clustering. *CoRR*, abs/1105.0121.

Quionero-candela, J., Rasmussen, C. E., and Herbrich, R. (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959.

Schreck, T., von Landesberger, T., and Bremm, S. (2010). Techniques for precision-based visual analysis of projected data. *Information Visualization*, 9(3):181–193.

Shneiderman, B. (2002). Inventing discovery tools: combining information visualization with data mining. *Information Visualization*, 1(1):5–12.

Tino, P. and Nabney, I. T. (2002). Hierarchical GTM: Constructing localized nonlinear projection manifolds in a principled way. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 24, pages 639–656.

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

Venna, J. and Kaski, S. (2001). Neighborhood preservation in nonlinear projection methods: An experimental study. In *Proceedings of the International Conference on Artificial Neural Networks*, ICANN '01, pages 485–491, London, UK, UK. Springer-Verlag.

Versino, C. and Gambardella, L. M. (1996). Learning fine motion by using the hierarchical extended kohonen map. In *MAP, Proceedings Proc. ICANN96, International Conference on Artificial Neural Networks*, pages 221–226. Springer-Verlag.

Vicente, D. and Vellido, A. (2004). Review of hierarchical models for data clustering and visualization. *Tendencias de la Minera de Datos en Espaa, Espaola de Minera de Datos*.

Wang, X., Tiňo, P., and Fardal, M. A. (2008). Multiple manifolds learning framework based on hierarchical mixture density model. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, ECML PKDD '08, pages 566–581, Berlin, Heidelberg. Springer-Verlag.