Visual Analytics
Lecture week 11: GP-LVM

Ian T. Nabney

University of BRISTOL

Overview

- Understand dual view of probabilistic PCA
- Understand fundamentals of Gaussian Processes Latent Variable Models (GP-LVM)
- Able to apply GP-LVM to datasets

Further reading:

- Lawrence, Neil D. "Gaussian process latent variable models for visualisation of high dimensional data." NIPS. Vol. 2. 2003.
- Lawrence, Neil. "Probabilistic non-linear principal component analysis with Gaussian process latent variable models." Journal of machine learning research 6.11 (2005).
- Li, Ping, and Chen, Songcan. "A review on Gaussian Process Latent Variable Models." CAAI Transactions on Intelligence Technology 1 (2016), 366–376.

Probabilistic PCA (again!)

- The relationship between the latent variable and the data is linear: $\mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \eta_n$, where the η_n are an independent sample from a spherical Gaussian distribution with mean zero and covariance $\beta^{-1}\mathbf{I}$.
- The likelihood of a data point is

$$p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{W}, \beta) = \mathcal{N}(\mathbf{y}_n|\mathbf{W}\mathbf{x}_n, \beta^{-1}\mathbf{I}) \tag{1}$$

- To obtain the marginal likelihood we integrate over the latent variables, with a unit covariance, zero mean Gaussian prior over $\mathbf{x}_n$.

$$p(\mathbf{y}_n|\mathbf{W}, \beta) = \int p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{W}, \beta)p(\mathbf{x}_n)d\mathbf{x}_n \tag{2}$$

which gives

$$p(\mathbf{y}_n|\mathbf{W}, \beta) = \mathcal{N}(\mathbf{y}_n|\mathbf{0}, \mathbf{W}\mathbf{W}^T + \beta^{-1}\mathbf{I}) \tag{3}$$

The parameters $\mathbf{W}$ can then be found analytically which gives the usual eigendecomposition.

Dual approach to PPCA

- Marginalising the latent variables and optimising the parameters via maximum likelihood is a standard approach for fitting latent variable models. We can try the dual approach of marginalising parameters, $\mathbf{W}$, and optimising with respect to latent variables, $\mathbf{X}$.
- The prior for $\mathbf{W}$ is $\prod_{i=1}^{D} \mathcal{N}(\mathbf{w}_i|\mathbf{0}, \mathbf{I})$ where $\mathbf{w}_i$ is the ith row of $\mathbf{W}$. We then marginalise over $\mathbf{W}$

$$p(\mathbf{Y}|\mathbf{X}, \beta) = \prod_{d=1}^{D} p(\mathbf{y}_{:,d}|\mathbf{X}, \beta), \tag{4}$$

where $\mathbf{y}_{:,d}$ represents the dth column of $\mathbf{Y}$ and

$$p(\mathbf{y}_{:,d}|\mathbf{X}, \beta) = \mathcal{N}(\mathbf{y}_{:,d}|\mathbf{0}, \mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}) \tag{5}$$

- Optimising this with respect to $\mathbf{X}$ (i.e. finding the optimal points in latent space to map to the data) gives $\mathbf{X} = \mathbf{U}\mathbf{L}\mathbf{V}^T$ where $\mathbf{U}$ is an $N \times q$ matrix whose columns are the first q eigenvectors of $\mathbf{Y}\mathbf{Y}^T$, $\mathbf{L}$ is a $q \times q$ diagonal matrix with entries $l_j = (\lambda_j - \beta^{-1})^{-1/2}$ where λ_j is the jth eigenvalue of $D^{-1}\mathbf{Y}\mathbf{Y}^T$ and $\mathbf{V}$ is an arbitrary rotation.

Principles of GP-LVM

- Dual PPCA is special case where the output dimensions are assumed to be linear, independent, and identically distributed.
- If we write $\mathbf{K} = \mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}$, then the log likelihood is

$$L = -\frac{DN}{2}\ln 2\pi - \frac{D}{2}\ln|\mathbf{K}| - \frac{1}{2}\text{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T) \tag{6}$$

- We replace the inner product kernel with a covariance function that allows for non-linear functions and obtain a non-linear latent variable model.
- This can be viewed as a non-linear probabilistic version of PCA.

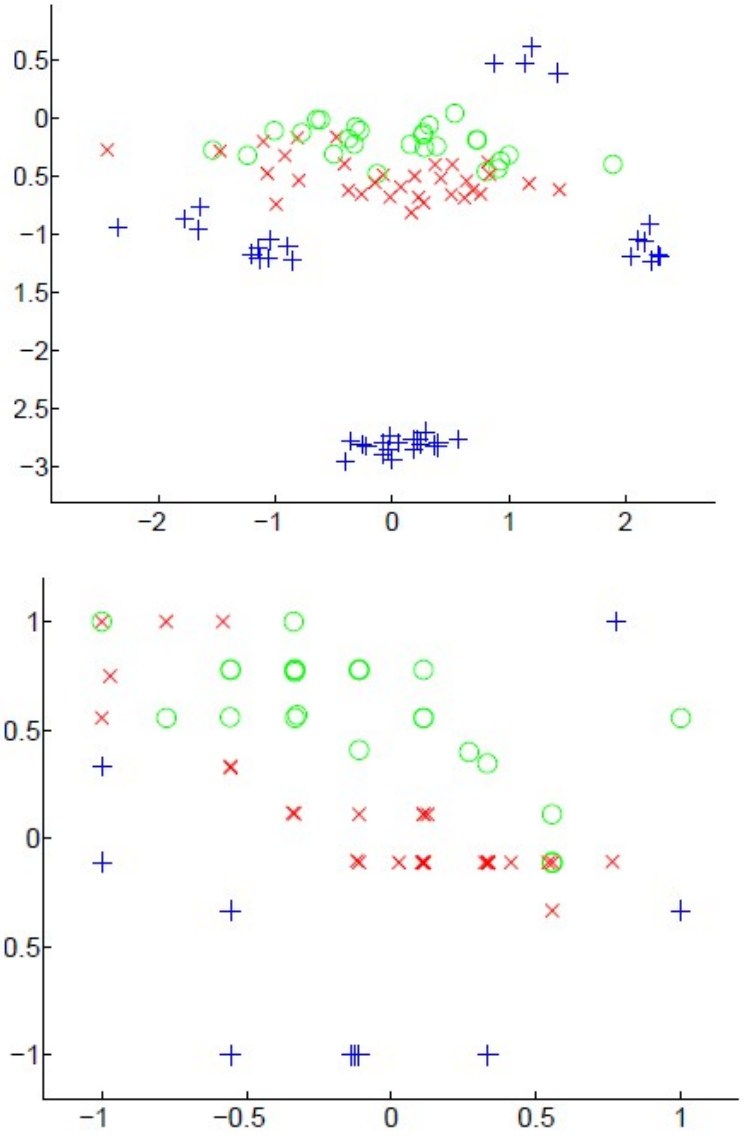
Fitting a non-linear GP-LVM

- For a linear kernel a closed form solution can be found (up to a rotation matrix).
- For non-linear kernels there will be no closed form solution and there are likely to be multiple local optima.
- Gradients of (6) with respect to the latent points can be found through first taking the gradient with respect to the kernel

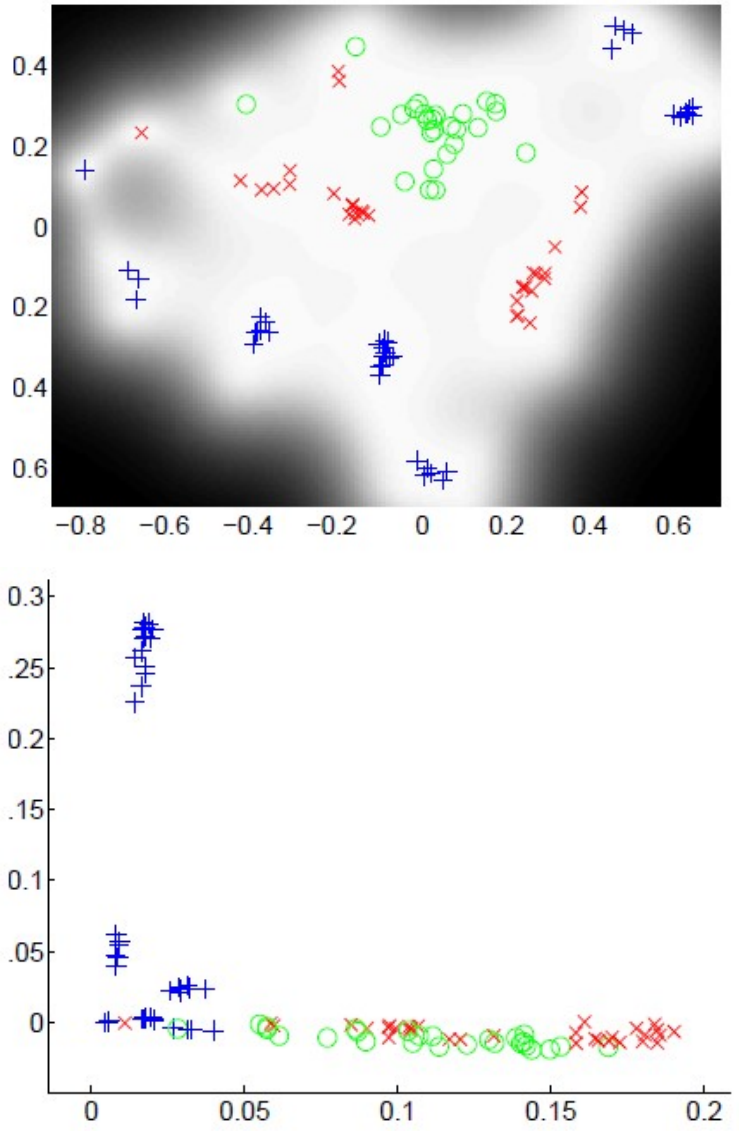
$$\frac{\partial L}{\partial \mathbf{K}} = \mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{T}\mathbf{K}^{-1} - D\mathbf{K}^{-1} \tag{7}$$

  and then combining it with $\partial \mathbf{K}/\partial \mathbf{x}_{n,j}$ through the chain rule.
- These gradients may then be used in combination with (6) in a non-linear optimiser to obtain a latent variable representation of the data.
- Gradients with respect to the parameters of the kernel matrix may be computed and used to jointly optimise $\mathbf{X}$ and the kernel's parameters.

GP-LVM example: oil dataset



PCA (top), GTM (bottom)

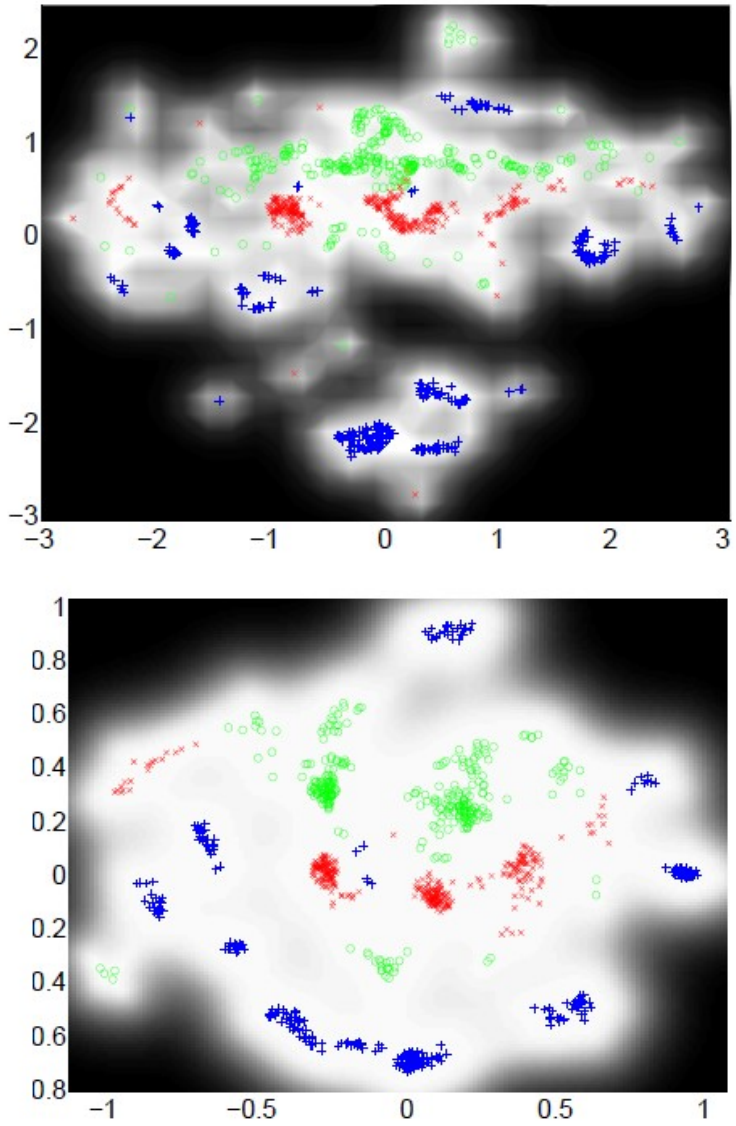


GP-LVM (top), kernel PCA (bottom)

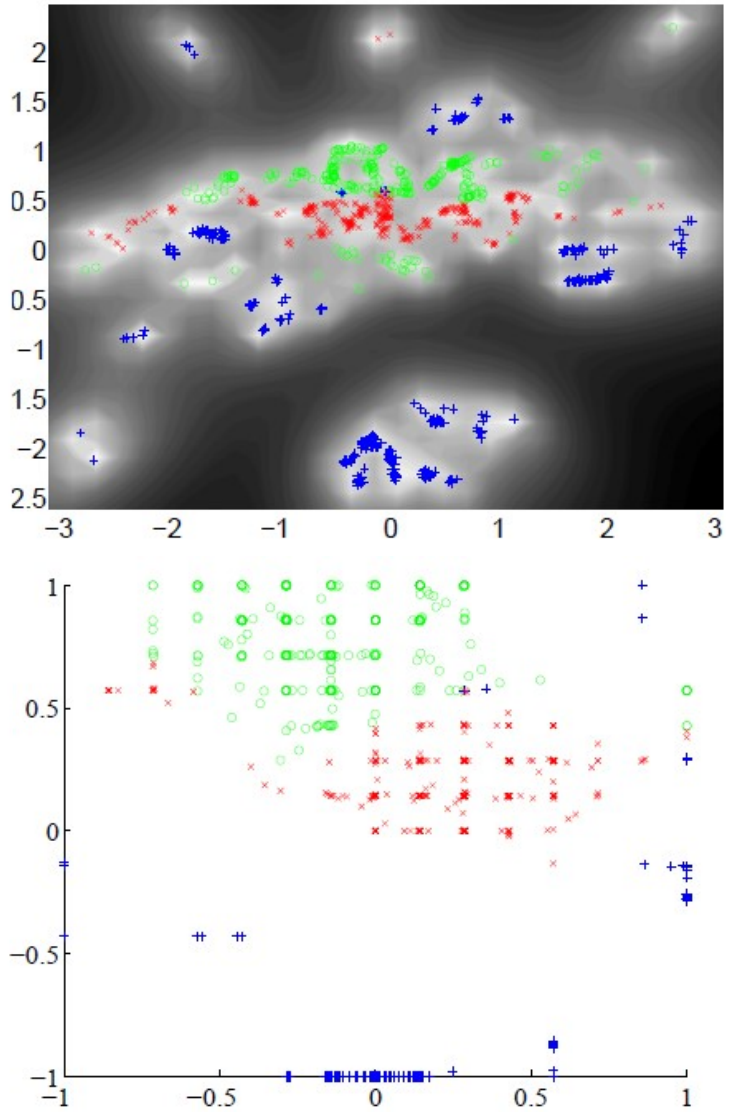
Practical issues

- The optimisation problem both non-linear and high dimensional (Nq interdependent parameters/latent variables before we consider the parameters of the kernel).
- Kernel methods may be sped up through sparsification, i.e. representing the data set by a subset, I, of d points known as the active set. We make use of the informative vector machine (IVM) which selects points sequentially according to the reduction in the posterior process's entropy that they induce.
- Optimisation with respect to the kernel's parameters and $\mathbf{X}_I$ with gradient evaluations costs $O(d^3)$ rather than the prohibitive $O(N^3)$ from the full model. The dominant cost (asymptotically) becomes that of the active selection which is $O(d^2 N)$.
- We can optimise the inactive points independently.

GP-LVM example: full oil dataset



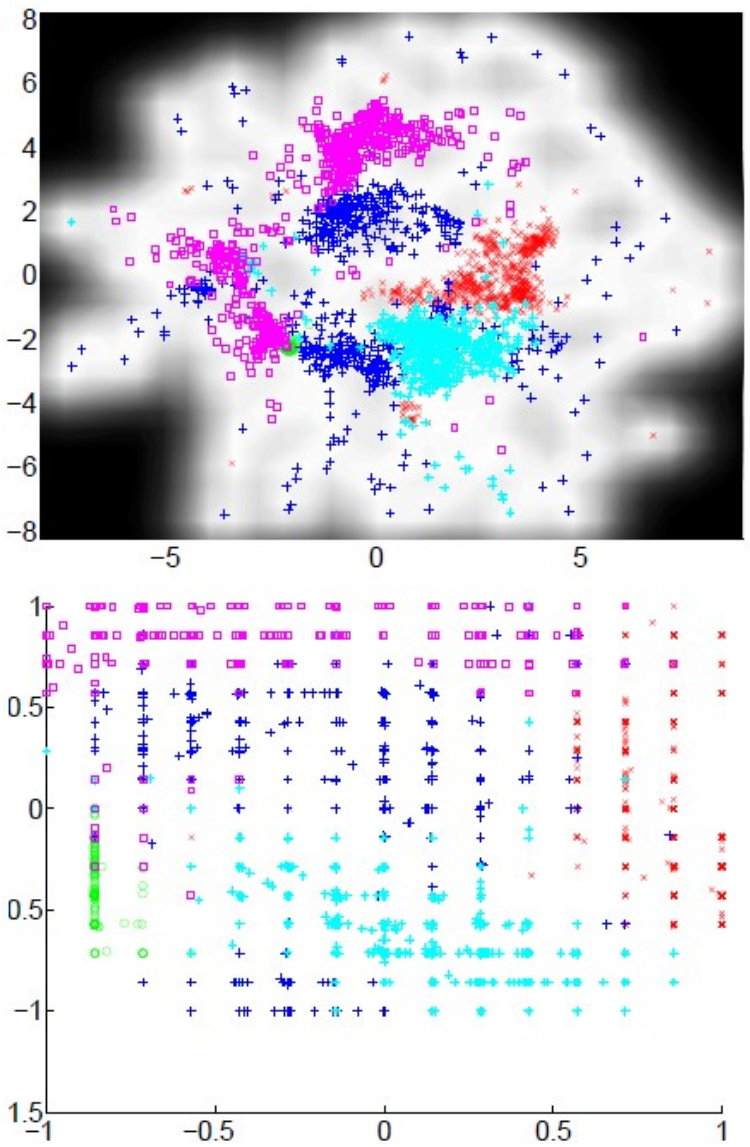
Sparse GP-LVM RBF kernel (top), GP-LVM RBF kernel (bottom)



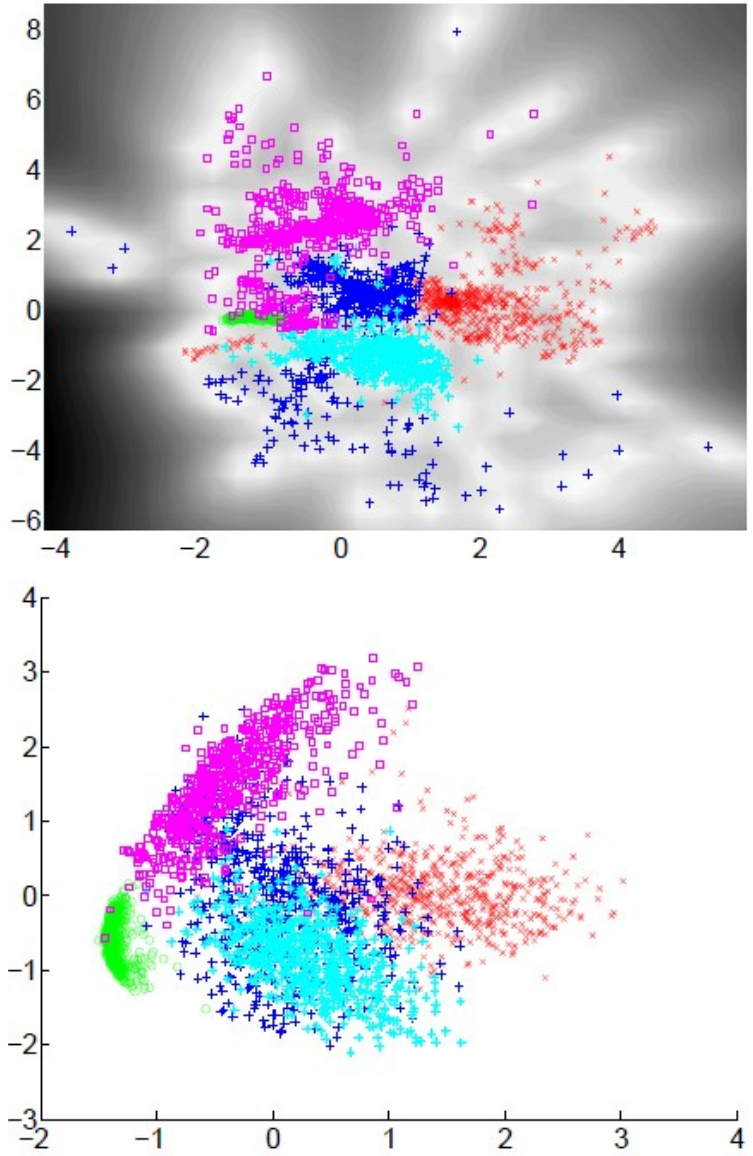
Sparse GP-LVM MLP kernel (top), GTM (bottom)

University of BRISTOL

GP-LVM example: digit dataset



Sparse GP-LVM RBF kernel (top), GTM (bottom)

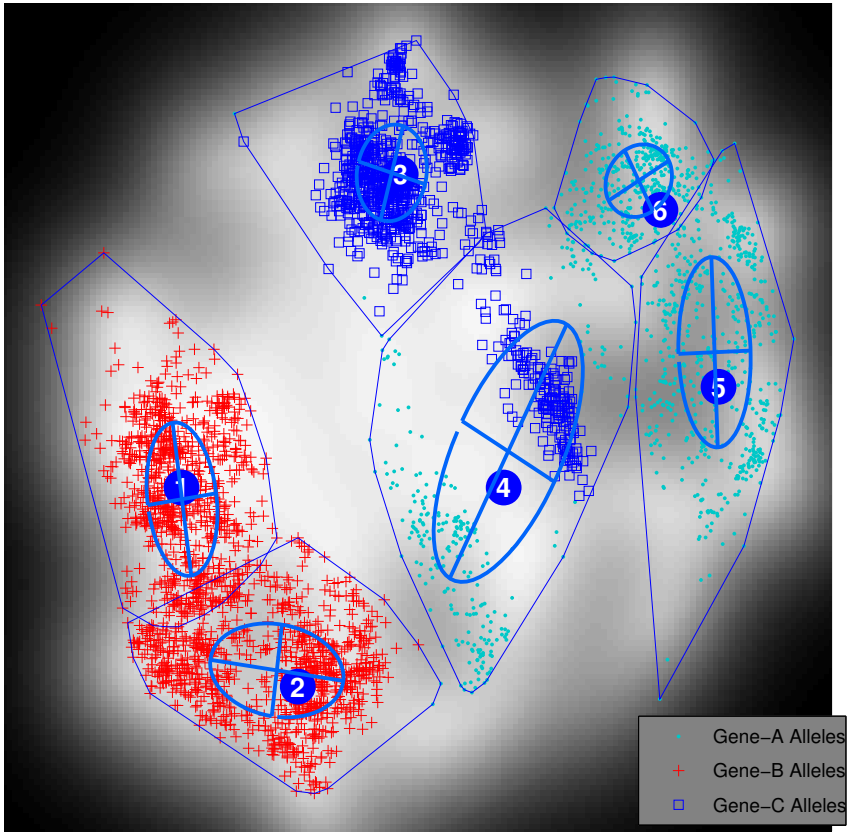


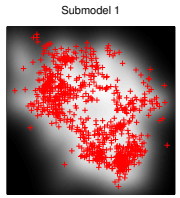
Sparse GP-LVM MLP kernel (top), PCA (bottom)

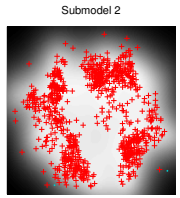
Generalisations of GP-LVM

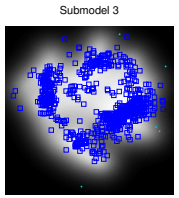
- Non-Gaussian noise models: using a Bernoulli noise model for binary data.
- Bayesian GP-LVM: variational inference that can automatically select the dimensionality of the non-linear latent space and also handle missing data.
- Hierarchical GPLVM (cf. hierarchical GTM).

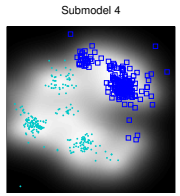
Hierarchical GP-LVM

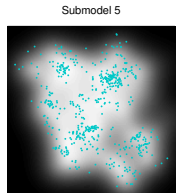


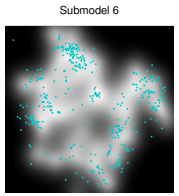












University of BRISTOL

Overview

- Understand dual view of probabilistic PCA
- Understand fundamentals of Gaussian Processes Latent Variable Models (GP-LVM)
- Able to apply GP-LVM to datasets

This topic has brought together all elements of the second half of the unit: PCA, numerical linear algebra, optimisation, Bayesian methods, latent variable models. The unifying thread is the principled application of probability theory to uncertainty.