Gaussian Process Latent Variable Model (GPLVM)

James McMurray

PhD Student
Department of Empirical Inference

11/02/2014

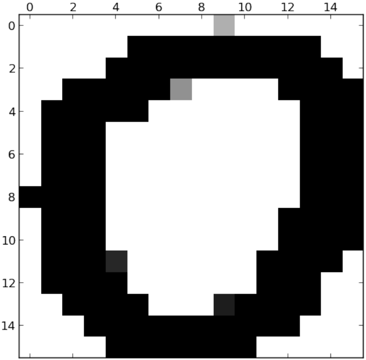
MAX-PLANCK-GESELLSCHAFT

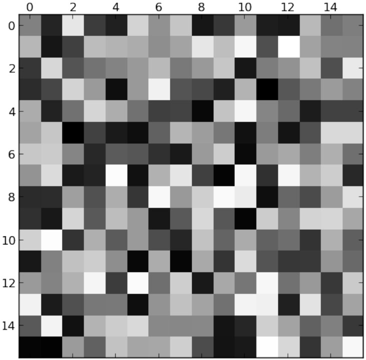
Outline of talk

MAX-PLANCK-GESELLSCHAFT

Why are latent variable models useful?

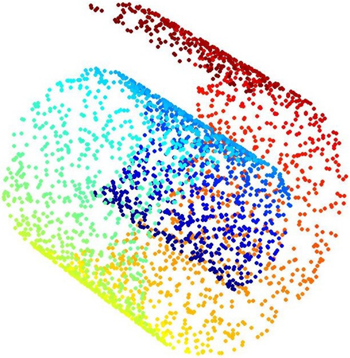
- Data has structure.





Why are latent variable models useful?

- Observed high-dimensional data often lies on a lower-dimensional manifold.



Example "Swiss Roll" dataset

MAX-PLANCK-GESELLSCHAFT

Why are latent variable models useful?

- The structure in the data means that we don't need such high dimensionality to describe it.
- The lower dimensional space is often easier to work with.
- Allows for interpolation between observed data points.

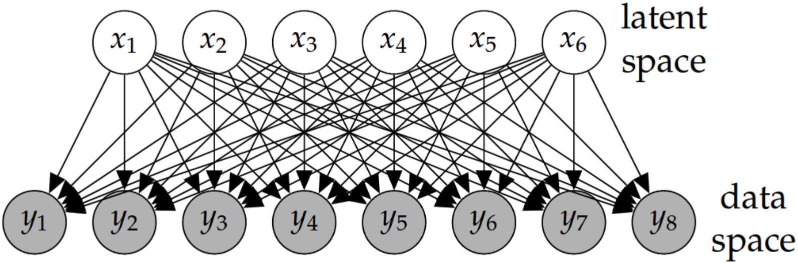
Definition of a latent variable model

- Assumptions:
  - Assume that the observed variables actually result from a smaller set of latent variables.
  - Assume that the observed variables are independent given the latent variables.
- Differs slightly from dimensionality reduction paradigm which wishes to find a lower-dimensional embedding in the high-dimensional space.
- With the latent variable model we specify the functional form of the mapping:

  $$y = g(x) + \varepsilon$$

  where x are the latent variables, y are the observed variables and ε is noise.
- Obtain different latent variable models for different assumptions on $g(x)$ and ε

Graphical model representation

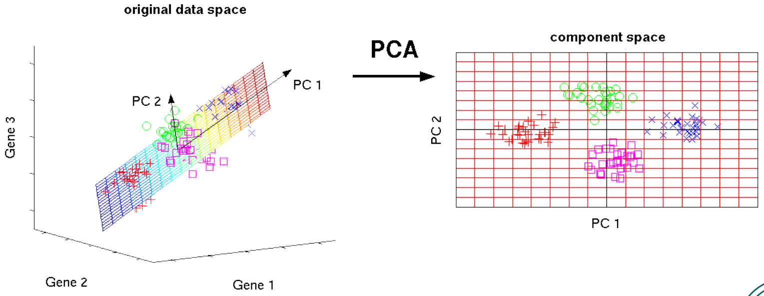


Graphical Model example of Latent Variable Model
Taken from Neil Lawrence:
http://ml.dcs.shef.ac.uk/gpss/gpws14/gp_gpws14_session3.pdf

Plan

Principal Components Analysis (PCA)

- Returns orthogonal dimensions of maximum variance.
- Works well if data lies on a plane in the higher dimensional space.
- Linear method (although variants allow non-linear application, e.g. kernel PCA).



Example application of PCA. Taken from
http://www.nlpca.org/pca_principal_component_analysis.html

MAX-PLANCK-GESELLSCHAFT

Plan

MAX-PLANCK-GESELLSCHAFT

Probabilistic PCA (PPCA)

- A probabilistic version of PCA.
- Probabilistic formulation is useful for many reasons:
  - Allows comparison with other techniques via likelihood measure.
  - Facilitates statistical testing.
  - Allows application of Bayesian methods.
  - Provides a principled way of handling missing values - via Expectation Maximization.

PPCA Definition

- Consider a set of centered data of n observations and d dimensions: $Y = [y_1, \ldots, y_n]^T$.
- We assume this data has a linear relationship with some embedded latent space data x_n. Where $Y \in \mathbb{R}^{N \times D}$ and $x \in \mathbb{R}^{N \times q}$.
- $y_n = \mathbf{W}x_n + n$, where x_n is the q-dimensional latent variable associated with each observation, and $\mathbf{W} \in \mathbb{R}^{D \times q}$ is the transformation matrix relating the observed and latent space.
- We assume a spherical Gaussian distribution for the noise with a mean of zero and a covariance of $\beta^{-1}\mathbf{I}$
- Likelihood for an observation y_n is:

$$p\left(y_n | x_n, \mathbf{W}, \beta\right) = \mathcal{N}\left(y_n | \mathbf{W}x_n, \beta^{-1}\mathbf{I}\right)$$

PPCA Derivation

- Marginalise latent variables x_n, put a Gaussian prior on $\mathbf{W}$ and solve using maximum likelihood.
- The prior used for x_n in the integration is a zero mean, unit covariance Gaussian distribution:

$$p(x_n) = \mathcal{N}(x_n|0, \mathbf{I})$$

$$p(y_n|\mathbf{W}, \beta) = \int p(y_n|x_n, \mathbf{W}, \beta)p(x_n)dx_n$$

$$p(y_n|\mathbf{W}, \beta) = \int \mathcal{N}\left(y_n|\mathbf{W}x_n, \beta^{-1}\mathbf{I}\right)\mathcal{N}(x_n|0, \mathbf{I})dx_n$$

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

- Assuming i.i.d. data, the likelihood of the full set is the product of the individual probabilities:

$$p(Y|\mathbf{W}, \beta) = \prod_{n=1}^{N} p(y_n|\mathbf{W}, \beta)$$

PPCA Derivation

- To calculate that marginalisation step we use the summation and scaling properties of Gaussians.
- Sum of Gaussian variables is Gaussian.

$$\sum_{i=1}^{n} \mathcal{N}(\mu_i, \sigma_i^2) \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

- Scaling a Gaussian leads to a Gaussian:

$$w\mathcal{N}(\mu, \sigma^2) \sim \mathcal{N}(w\mu, w^2\sigma^2)$$

- So:

$$y = \mathbf{W}x + \varepsilon \ , \ \ x \sim \mathcal{N}(0, \mathbf{I}) \ , \ \ \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

$$\mathbf{W}x \sim \mathcal{N}(0, \mathbf{W}\mathbf{W}^\mathsf{T})$$

$$\mathbf{W}x + \varepsilon \sim \mathcal{N}(0, \mathbf{W}\mathbf{W}^\mathsf{T} + \sigma^2\mathbf{I})$$

MAX-PLANCK-GESELLSCHAFT

PPCA Derivation

- Can find a solution for **W** by maximising the likelihood.
- Results in an eigenvalue problem.
- Turns out that the closed-form solution for **W** is achieved when **W** spans the principal sub-space of the data[1].
- Same solution as PCA: Probabilistic PCA
- Can it be extended to capture non-linear features?

[1] Michael E. Tipping and Christopher M. Bishop. "Probabilistic principal component analysis." (1997).

Dual PPCA

- Similar to previous derivation of PPCA.
- But marginalise $\mathbf{W}$ and optimise x_n.
- Same linear-Gaussian relationship between latent variables and data:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \beta) = \prod_{d=1}^{D} \mathcal{N}(y_{d,:}|\mathbf{W}x_{d,:}, \beta^{-1}\mathbf{I})$$

- Place a conjugate prior on $\mathbf{W}$:

$$P(\mathbf{W}) = \prod_{d=1}^{D} \mathcal{N}(w_{d,:}|\mathbf{0}, \mathbf{I})$$

- Resulting marginal likelihood is:

$$P(Y|X, \beta) = \prod_{d=1}^{D} \mathcal{N}(y_{:,d}|0, \mathbf{X}\mathbf{X}^{\mathbf{T}} + \beta^{-1}\mathbf{I})$$

MAX-PLANCK-GESELLSCHAFT

Dual PPCA

- Results in equivalent eigenvalue problem to PPCA.
- So what is the benefit?
- The eigendecomposition is now done on an $N \times q$ instead of a $d \times q$ matrix.
- Recall marginal likelihood:

$$P(Y|X,\beta) = \prod_{d=1}^{D} \mathcal{N}(y_{:,d}|0, \mathbf{X}\mathbf{X}^{\mathbf{T}} + \beta^{-1}\mathbf{I})$$

- The covariance matrix is a covariance function:

$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathbf{T}} + \beta^{-1}\mathbf{I}$$

- This linear kernel can be replaced by other covariance functions for non-linearity.
- This is the GPLVM.

MAX-PLANCK-GESELLSCHAFT

GPLVM

- Each dimension of the marginal distribution can be interpreted as an independent Gaussian Process[2].
- Dual PPCA is the special case where the output dimensions are assumed to be linear, independent and identically distributed.
- GPLVM removes assumption of linearity.
- Gaussian prior over the function space.
- Choice of covariance function changes family of functions considered.
- Popular kernels:
  - Exponentiated Quadratic (RBF) kernel
  - Matern kernels
  - Periodic kernels
  - Many more...

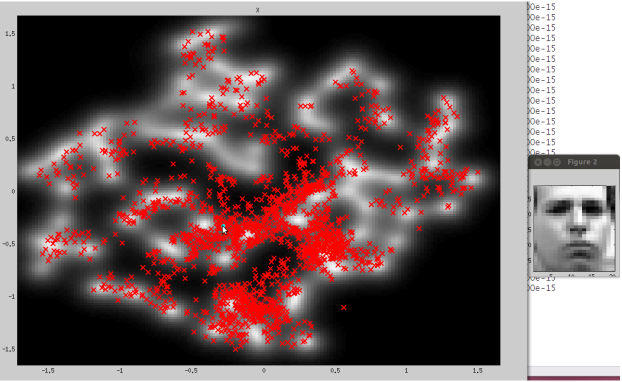
[2] Neil Lawrence: "Probabilistic non-linear principal component analysis with Gaussian process latent variable models." JMLR (2005)

MAX-PLANCK-GESELLSCHAFT

Plan

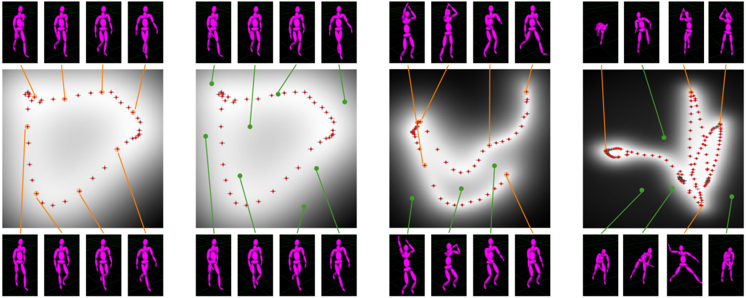
MAX-PLANCK-GESELLSCHAFT

Example: Frey Face data



Example in GPMat

Example: Motion Capture data



Taken from Keith Grochow, et al. "Style-based inverse kinematics." ACM Transactions on Graphics (TOG). Vol. 23. No. 3. ACM, 2004.

Plan

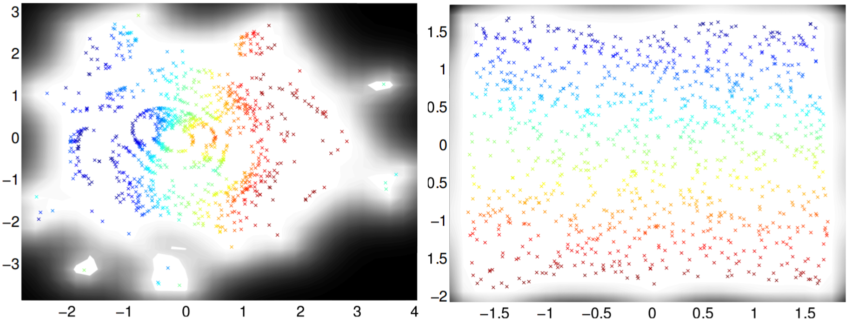
MAX-PLANCK-GESELLSCHAFT

Practical points

- Need to optimise over non-convex objective function.
- Achieved using gradient-based methods (scaled conjugate gradients).
- Several restarts to attempt to avoid local optima.
- Cannot guarantee global optimum.
- High computational cost for large datasets.
- May need to optimise over most-informative subset of data, the "active set" for sparsification.

Practical points

Initialisation can have a large effect on the final results.



(a) (b)

Effect of poor initialisation on Swiss Roll dataset. PCA left, Isomap right. Taken from "Probabilistic non-linear principal component analysis with Gaussian process latent variable models.", Neil Lawrence, JMLR (2005).

MAX-PLANCK-GESELLSCHAFT

Plan

MAX-PLANCK-GESELLSCHAFT

Variants

- There are a number of variants of the GPLVM.
- For example, the GPLVM uses the same covariance function for each output dimension.
- This can be changed, for example the Scaled GPLVM which introduces a scaling parameter for each output dimension[3].
- The Gaussian Process Dynamic Model (GPDM) adds another Gaussian process for dynamical mappings[4].
- The Bayesian GPLVM approximates integrating over both the latent variables and the mapping function[5].

---

[3] Keith Grochow, et al. "Style-based inverse kinematics." ACM Transactions on Graphics (TOG). Vol. 23. No. 3. ACM, 2004.

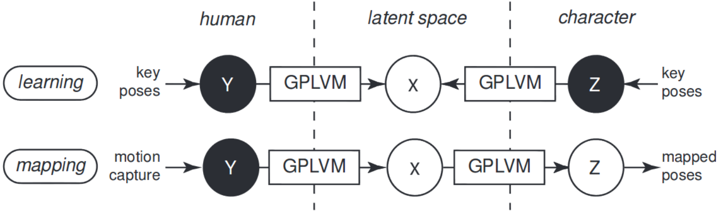
[4] Wang, Jack, Aaron Hertzmann, and David M. Blei. "Gaussian process dynamical models." Advances in neural information processing systems. 2005.

[5] Titsias, Michalis, and Neil Lawrence. "Bayesian Gaussian process latent variable model." (2010).

MAX-PLANCK-GESELLSCHAFT

Variants

- The Shared GPLVM learns mappings from a shared latent space to two separate observational spaces.
- Used by Disney Research in their paper "Animating Non-Humanoid Characters with Human Motion Data" for generating animations for non-human characters from human motion capture data.



Shared GPLVM mappings as used by Disney Research

- **Video**

Variants

- Can also put a Gaussian Process prior on X to produce Deep Gaussian Processes[6].
- Zhang *et al.* developed Invariant GPLVM[7] - permits interpretation of causal relations between observed variables, by allowing arbitrary noise correlations between the latent variables.
- Currently attempting to implement IGPLVM in GPy.

[6]Damianou, Andreas C., and Neil D. Lawrence. "Deep Gaussian Processes." arXiv preprint arXiv:1211.0358 (2012).

[7]Zhang, K., Schölkopf, B., and Janzing, D. (2010). "Invariant Gaussian Process Latent Variable Models and Application in Causal Discovery". UAI 2010.

MAX-PLANCK-GESELLSCHAFT

Plan

Conclusion

- Implemented in GPy (Python) and GPMat (MATLAB).
- Many practical applications - pose modelling, tweening
- Especially if smooth interpolation is desireable.
- Modeling of confounders.

Thanks for your time

Questions?

MAX-PLANCK-GESELLSCHAFT

References

- Neil Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data." Advances in neural information processing systems 16.329-336 (2004): 3.
- Neil Lawrence, "Probabilistic non-linear principal component analysis wit Gaussian process latent variable models." JMLR (2005)
- Gaussian Process Winter School, Sheffield 2013: `http://ml.dcs.shef.ac.uk/gpss/gpws14/`
- WikiCourseNote: `http://wikicoursenote.com/wiki/Probabilistic_PCA_with_GPLVM`