Metric for Probabilistic Manifold Embeddings

Max Zwiessele[a,b]

The University Of Sheffield.

Problem

High throughput sequencing technologies [Shendure and Ji, 2008] reveal vast amounts genomic traits. We approach a cap in the ability to associate these genomic traits to measurable phenotypes (such as clinical measurements, visible phenotypes and proteomic measurements). It has been shown that high dimensional data can be analysed and visualized in lower dimension representations using the so called Gaussian Process Latent Variable Model [Fusi et al., 2012, Damianou et al., 2012, 2011, Fusi et al., 2011, Titsias and Lawrence, 2010, Stegle et al., 2010]

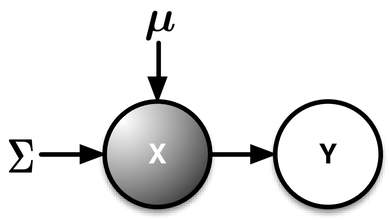
We propose to use a recent development of deeper analysis of the probabilistic embedding using Riemanian methods [Tosi et al., 2014] in order to reveal more about the latent structure imposed by the underlying processes in biology.

Metric for Embedding

In order to learn such latent structure we need to apply latent variable models, such as Bayesian GPLVM [Titsias and Lawrence, 2010], which tries to learn latent input variables $\mathbf{X} \in \mathbb{R}^{n\times q}$ explaining observed output variables $\mathbf{Y} \in \mathbb{R}^{n\times d}$ by variational bayesian treatment of the input space of a Gaussian process [Rasmussen, 2006], leading to a non-linear learning of the latent embedding emposed by the Gaussian process (GP).

$$\begin{aligned}\ln p(\mathbf{Y}) &= \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})d\mathbf{X}\\ &\geq \int q(\mathbf{X}) \ln p(\mathbf{Y}|\mathbf{X})d\mathbf{X} - \mathrm{KL}(q\|p)\end{aligned}$$

$$q(\mathbf{X}) = \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



As the Bayesian GPLVM is a probabilistic model, the method of Tosi et al. [2014] can be easily applied. We find the Wishart embedding of the Bayesian GPLVM by defining a joint Gaussian process on the observed data and the derivative of the latent function:

$$\begin{pmatrix}\mathbf{Y}\\ \mathbf{J}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{Y}\\ \frac{\partial \mathbf{F}}{\partial \mathbf{X}}\end{pmatrix}\middle| \mathbf{0}, \begin{bmatrix}\mathbf{K}+\sigma^2\mathbf{I} & \frac{\partial \mathbf{K}}{\partial \mathbf{X}}\\ \frac{\partial \mathbf{K}^\top}{\partial \mathbf{X}} & \frac{\partial^2 \mathbf{K}}{\partial \mathbf{X}\partial \mathbf{X}}\end{bmatrix}\right)$$

With that, we can use standard GP properties to get the posterior distribution over the derivative of the latent function values $\frac{\partial \mathbf{F}}{\partial \mathbf{X}}$ having seen the data: [a]

$$\begin{aligned}p(\mathbf{J}|\mathbf{Y},\mathbf{X}) &= \mathcal{N}(\mathbf{M},\boldsymbol{\Sigma})\\ \mathbf{M} &= \frac{\partial \mathbf{K}}{\partial \mathbf{X}}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{Y}\\ \boldsymbol{\Sigma} &= \frac{\partial^2 \mathbf{K}}{\partial \mathbf{X}\partial \mathbf{X}} - \frac{\partial \mathbf{K}}{\partial \mathbf{X}}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\frac{\partial \mathbf{K}^\top}{\partial \mathbf{X}}\end{aligned}$$

With this we define the Wishart distribution over over the metric tensor $\mathbf{G}$ as explained by Tosi et al. [2014]

$$\begin{aligned}\mathbf{G}(\mathbf{X}) &\sim \mathcal{W}(d, \boldsymbol{\Sigma}, \left\langle \mathbf{J}^\top \right\rangle \langle \mathbf{J} \rangle)\\ \Rightarrow \langle \mathbf{G}(\mathbf{X}) \rangle &= \left\langle \mathbf{J}^\top \mathbf{J} \right\rangle = \left\langle \mathbf{J}^\top \right\rangle \langle \mathbf{J} \rangle + d\boldsymbol{\Sigma} \ .\end{aligned}$$

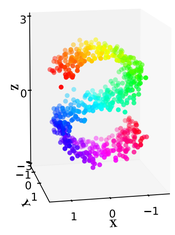
This metric tensor can be computed at every point of the latent space and is a proxy for the density (and in the squared exponential case non-linearity) of the latent space.

For visualization purposes we reduce the expected value over the metric tensor into a scalar, by its footprint

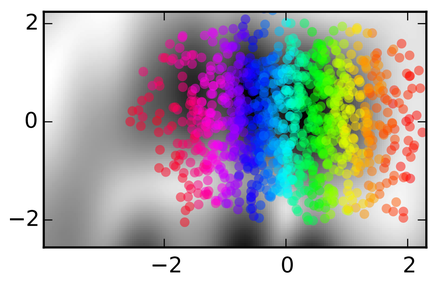
$$\mathbf{M} = \sqrt{\det \langle \mathbf{J}^\top \mathbf{J} \rangle} \ .$$

[a] This directly translates into the Bayesian case (using e.g. [Titsias and Lawrence, 2010] posterior) and prediction at unseen inputs $p\left(\frac{\partial \mathbf{F}^\star}{\partial \mathbf{X}^\star}\middle| \mathbf{Y}, \mathbf{X}, \mathbf{X}^\star\right)$

Magnification Factor



In order to show the magnification factor, we will learn a lower dimensional representation of the S-curve dataset, created by the sklearn package [Pedregosa et al., 2011]. The latent embedding is usually done in two dimensions, so that the manifold has to curve along the data (but not fold into itself). This curvature should be visible in the magnification factor, as along the curvature, the manifold is highly non-linear, and therefore more dense then elsewhere.



Latent space embedding as learnt by a Bayesian GPLVM of the S-curve data. The grey background shows the magnification factor. You can clearly see how the manifold needs to bend (and how it bends) along the x-axis in order to capture data embedding.

Distances in Manifold

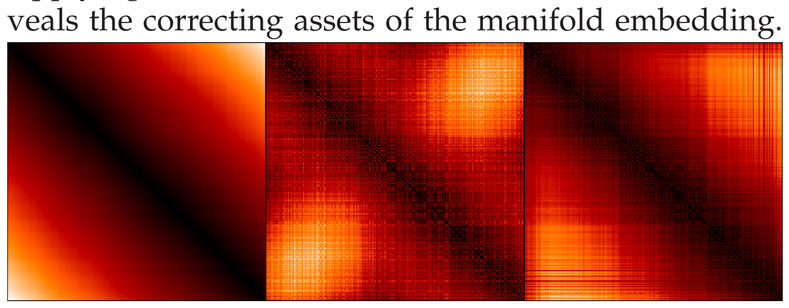
As the measure of manifold learning, we will compare corrected distances in the manifold

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^\top \mathbf{A} (\mathbf{x}-\mathbf{y})}$$

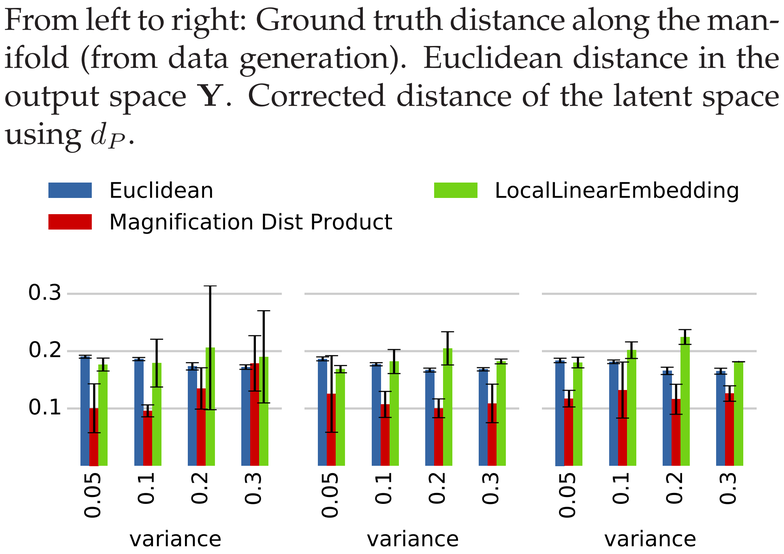
where $\mathbf{A}$ is the correction for the manifold distortion. We define a correction method[a] for the distance in the manifold:

$$d_P(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^\top \mathbf{L}_\mathbf{x}\mathbf{L}_\mathbf{y}^\top (\mathbf{x}-\mathbf{y})} \ ,$$

where $\mathbf{G}(\mathbf{X}) = \mathbf{L}_\mathbf{X}\mathbf{L}_\mathbf{X}^\top$ the cholesky decomposition of the manifold embedding for the given datapoint. Applying this distance to the S-curve dataset reveals the correcting assets of the manifold embedding.



From left to right: Ground truth distance along the manifold (from data generation). Euclidean distance in the output space $\mathbf{Y}$. Corrected distance of the latent space using d_P.

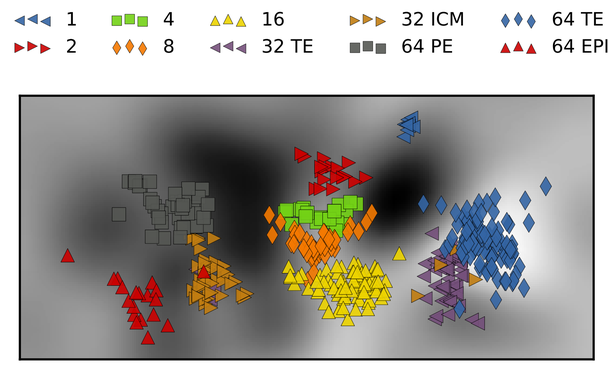


From left to right. with increasing variance in the simulated data, we plot error between the true distances in the latent space and the corrected distances for $200, 400$ and 600 datapoints, respectively. 'Euclidean': Euclidean distance in original space, 'Magnification Dist Product': d_P and 'LocalLinearEmbedding': Local Linear Embedding for comparison.

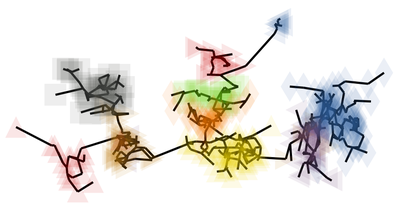
[a] Here d_P is not fully a distance anymore (it is not necessarily symmetric). This is because we only correct for local distortion, not for the full paths. See Tosi et al. [2014] for full integration over the geodesics.

Single Cell experiment

Applying the described method to the singlecell experiment from Guo et al. [2010] we can see that the corrected distances unravel the time structure of the embedding.



We used the corrected distances to plot a minimal spanning tree into the latent space, which follows the time structure of the data. Thus, using the corrected distances we were able to unravel the timestructure embedded in the data.



Future Directions

In order to further deepen the knowledge about the manifold embedding we will have a look into using the Fisher information of the latent embedding

$$\mathcal{F}(\mathbf{X}) = -\int \left(\frac{\partial^2}{\partial \mathbf{X}\partial \mathbf{X}} \log p(\mathbf{Y}|\mathbf{X})\right) p(\mathbf{Y}|\mathbf{X})\, \mathrm{d}\mathbf{X}$$

This is the amount of information the probability distribution carries about the latent space $\mathbf{X}$, and therefore a direct proxy for the information embedded in the manifold.

Acknowledgements

I am grateful for financial support from the European Union 7th Framework Programme through the Marie Curie Initial Training Network "Machine Learning for Personalized Medicine" MLPM2012, Grant No. 316861.

References

A. Damianou, M. Titsias, and N. Lawrence. Variational Gaussian Process Dynamical Systems. *Neural Information Processing System (NIPS)*, 2011.

A. C. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence. Manifold Relevance Determination. *ICML*, 2012.

N. Fusi, O. Stegle, and N. D. Lawrence. Accurate modeling of confounding variation in eQTL studies leads to a great increase in power to detect trans-regulatory effects. *Nature Precedings*, 2011.

N. Fusi, O. Stegle, and N. D. Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput Biol*, 8(1): e1002330, Jan 2012. doi: 10.1371/journal.pcbi.1002330.

G. Guo, M. Huss, G. Q. Tong, C. Wang, L. Li Sun, N. D. Clarke, and P. Robson. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental cell*, 18(4):675–685, 2010.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

C. E. Rasmussen. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

J. Shendure and H. Ji. Next-generation DNA sequencing. *Nat Biotechnol*, 26(10):1135–45, Oct 2008. doi: 10.1038/nbt1486.

O. Stegle, L. Parts, R. Durbin, and J. Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol*, 6(5): e1000770, May 2010. doi: 10.1371/journal.pcbi.1000770.

M. K. Titsias and N. D. Lawrence. Bayesian Gaussian Process Latent Variable Model. *Artificial Intelligence and Statistics*, 2010.

A. Tosi, S. Hauberg, A. Vellido, and N. Lawrence. Metrics for Probabilistic Geometries. In *Proceedings of the Thirtieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-14)*, pages 800–808, Corvallis, Oregon, 2014. AUAI Press.

[a] University of Sheffield, Department of Computer Science, UK
Corresponding author: m.zwiessele@sheffield.ac.uk

[b] Sheffield Institute for Translational Neuroscience (SITraN), UK