The IGPLVM and Causal Inference

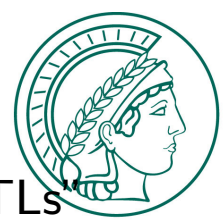
James McMurray

PhD Student
Department of Empirical Inference

19/03/2014

Background: eQTL studies

- Expression levels of proteins and mRNA can be regulated by genetic loci (bases in the genome) that are within, (cis) or far (trans) from the coding gene itself. Expression Quantitative Trait Loci (eQTL) mapping studies attempt to discover which loci regulate the expression of which products (and therefore, the associated genes).
- Trait-associated SNPs are more likely to be eQTLs[1], so this can also be useful for establishing prior information for GWASs.



[1] D. Nicolae, *et al.* "Trait-Associated SNPs Are More Likely to Be eQTLs" PLoS Genet 6(4): e1000888. (2010)

PANAMA model

- Fusi, Stegle *et al.* developed PANAMA[2] model (Probabilistic ANAlysis of genoMic dAta) for eQTL studies, which uses the Gaussian Process Latent Variable Model (GPLVM) to model confounding factors.
- Assuming the smaller set of confounding factors have a broad influence on the gene expression levels.

[2] 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 Computational Biology, 2012

The IGPLVM

- Zhang *et al.* developed IGPLVM[3] - permits interpretation of causal relations between observed variables, by allowing an arbitrary noise structure amongst the latent variables.
- This means that the mapping can be expressed in a LiNGAM[4] (Linear Non-Gaussian Acyclic Model) if the noise turns out to be non-Gaussian, and the LiNGAM causal inference method can be applied. If the noise is Gaussian, one is restricted to a Markov equivalence class.
- Currently attempting to implement IGPLVM using GPy. Produces reasonable latent variable mapping, but there appear to be bugs with the matrix reconstruction for causal inference.

---

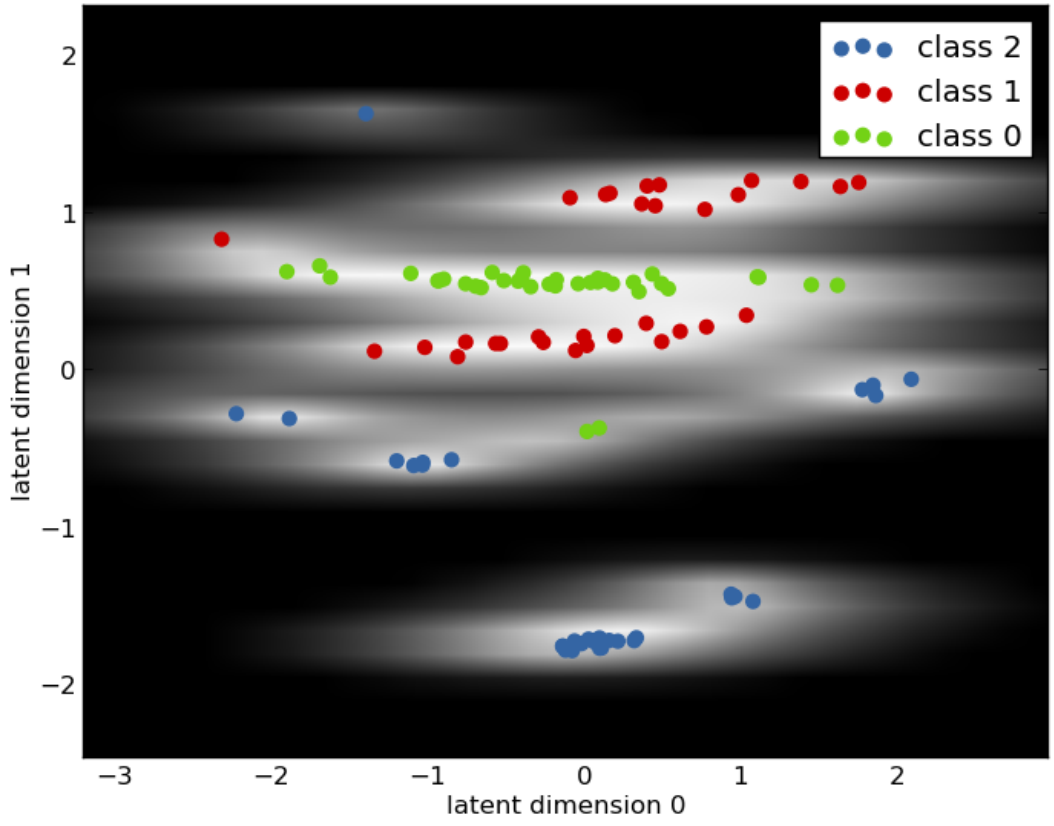
[3] Zhang, K., Schölkopf, B., and Janzing, D. (2010). "Invariant Gaussian Process Latent Variable Models and Application in Causal Discovery". UAI 2010.

[4] S. Shimizu, et al. "A linear, non-gaussian acyclic model for causal discovery." JMLR 2006.

MAX-PLANCK-GESELLSCHAFT

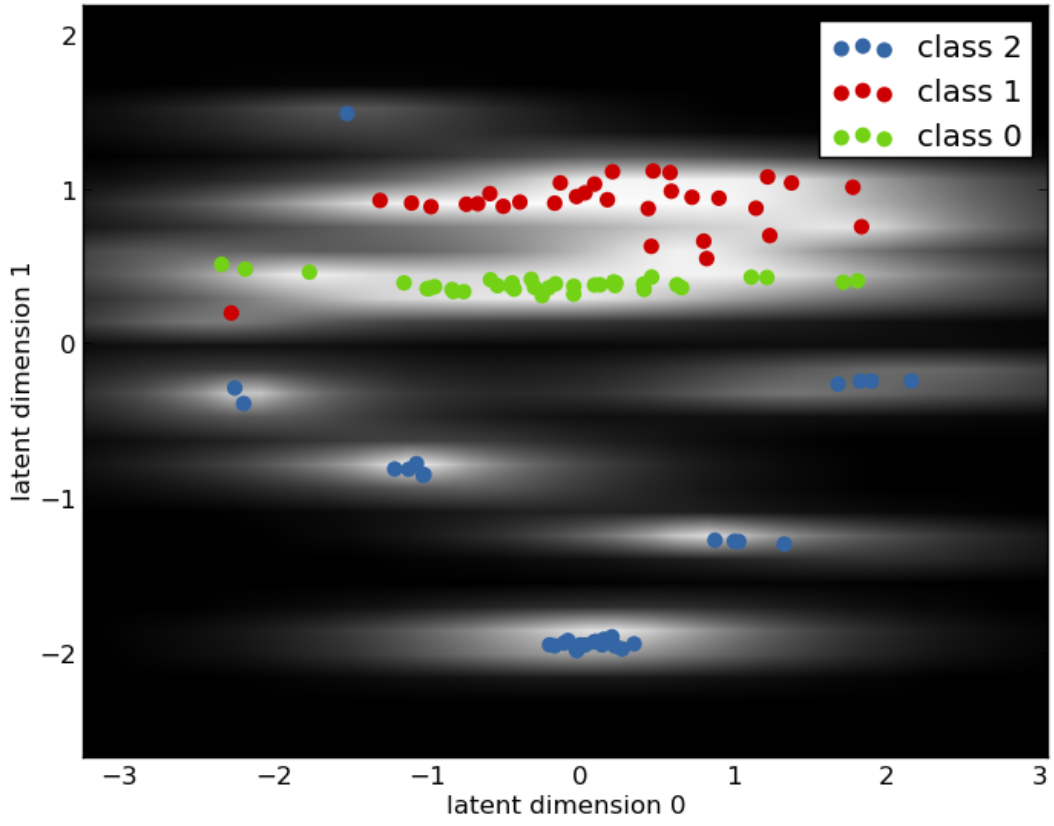
Example plots: Oil flow dataset

- GPLVM, 2 latent dimensions:



Example plots: Oil flow dataset

- IGPLVM, 2 latent dimensions:



Applications

- First idea was to simply replace GPLVM with IGPLVM in the PANAMA model, to see if one can obtain better results.
- However a better idea, may be to apply it to the learning of regulatory networks, where one has time series data. Since it can be used with the GPDM (Gaussian Process Dynamic Model) which learns a mapping in time of the latent variables.
- This could then be applied to account for confounders in gene expression time series data, and perhaps give a better interpretation of the confounders.

Thanks for your time

Questions?

MAX-PLANCK-GESELLSCHAFT