James McMurray

Project Title: Predicting Phenotype through Interaction of Genotype, Epigenotype and Environment with Probabilistic Models

Location: Max Planck Institute for Intelligent Systems, Tübingen, Germany

Supervisor: Bernhard Schölkopf

Background:

- Completed MPhys degree in **Physics** at the University of Exeter in July 2013.
- Completed two programming internships in Germany during university (at Uni Konstanz and Uni Tübingen).
- Also completed research project on Social Network Analysis - awaiting publication.
- Became interested in Machine Learning after completing "MOOCs" on **Coursera** by Andrew Ng and Daphne Koller, and on Udacity by Peter Norvig.

Research Interests:

- Models which combine different sources and forms of information, i.e. combining genetic, epigenetic and environmental data.
- Causal inference from observational data.
- Consumer personalized medicine.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861

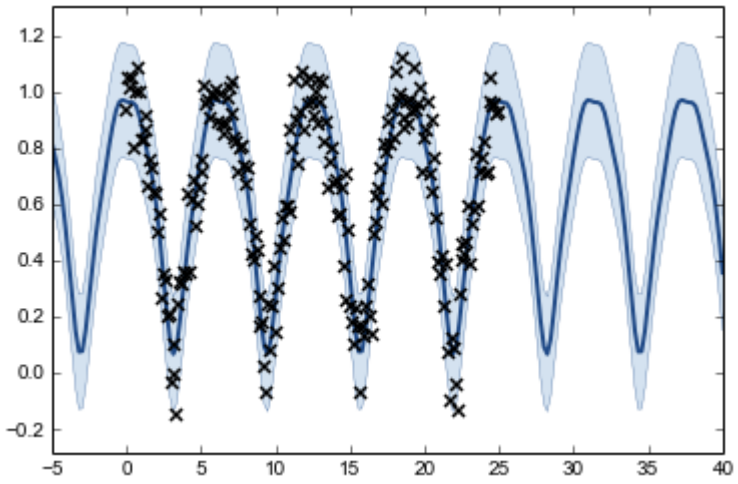
Overview:

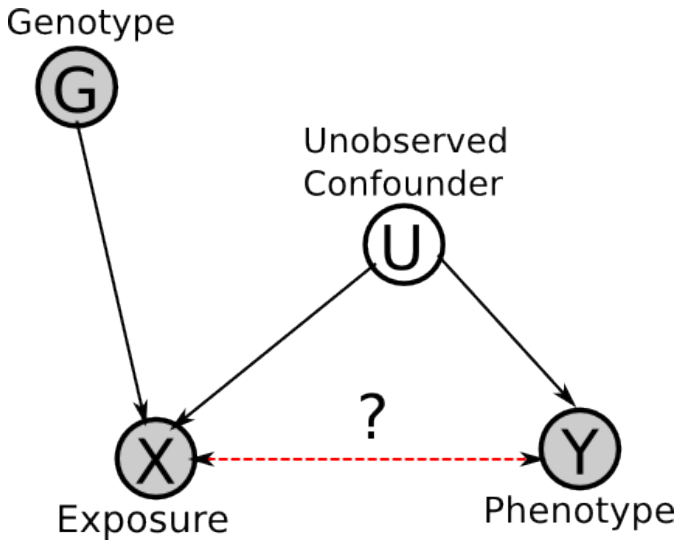
- Work has largely focussed on **causal inference** and **unsupervised learning**.
- Throughout the work I have used the **GPy** software package developed by Neil Lawrence's group in Sheffield.
- Including some of Max Zwiessele's work on the Bayesian GPLVM.
- Focussed on work with possible clinical applications (cancer diagnosis) - with the **DREAM9 challenge** and **TCGA** data.
- Have also worked on some more theoretical areas such as causal inference, however I do not think it is currently mature enough to be applicable scientifically.
- Interested in working on more **practical applications** - hope to co-operate with other nodes to this end.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861

Start of project:

- Started by learning background to **causal inference** methods and **Gaussian Processes**.
- Interested in methods to deduce **causal direction** of interactions from **observational data**.
- Gaussian Processes are a non-parametric, Bayesian non-linear **regression** method.
- Provide a convenient way of incorporating **prior information**.

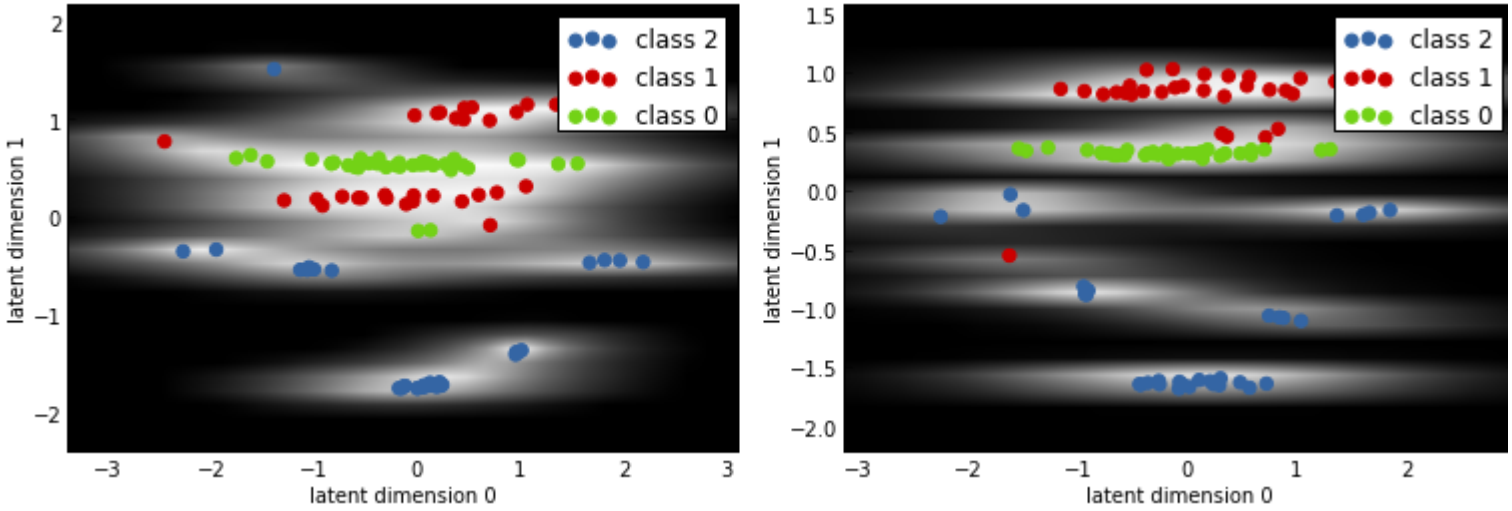




This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861

Invariant GPLVM:

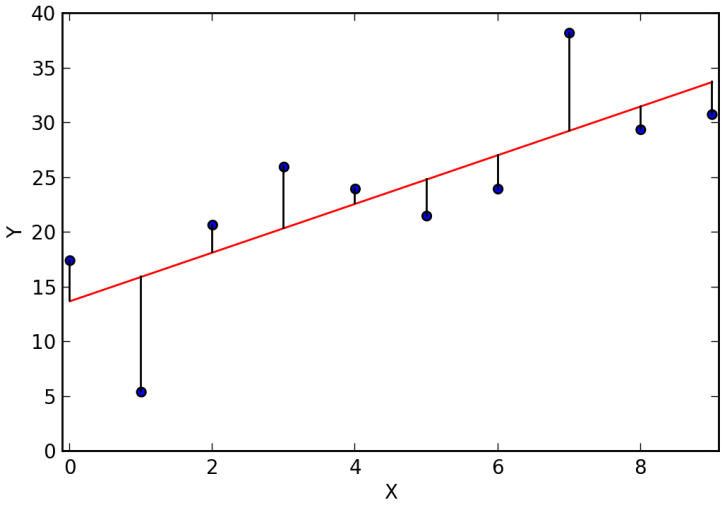
- Started implementation of the IGPLVM, an extension of the Gaussian Process Latent Variable Model, using the **GPy** Python module from Neil Lawrence's group in **Sheffield** (other ITN node).
- IGPLVM was originally developed by Kun Zhang, a post-doc at Tübingen.
- Allows one to infer instantaneous causal relations amongst observed variables.
- However, in biological applications it is often difficult to interpret instantaneous relations.
- Also introduces storage dependency on the number of dimensions of the original data.



This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861

eQTL discovery and independence-based regression:

- Did some work on **eQTL** (expression Quantitative Trait Loci) discovery with the assistance of Oliver Stegle (previously at Tübingen, now a group leader at the European Bioinformatics Institute).
- Aim is to find which SNPs are significantly correlated with changes in gene expression.
- Wanted to see if **HSIC**-based regression methods (HSIC is the Hilbert-Schmidt Independence Criterion), work better than the Maximum Likelihood approach when the model assumptions do not hold (i.e. there are causal SNPs which we do not consider, etc.)
- However, could not outperform the Maximum Likelihood approach in simulations.

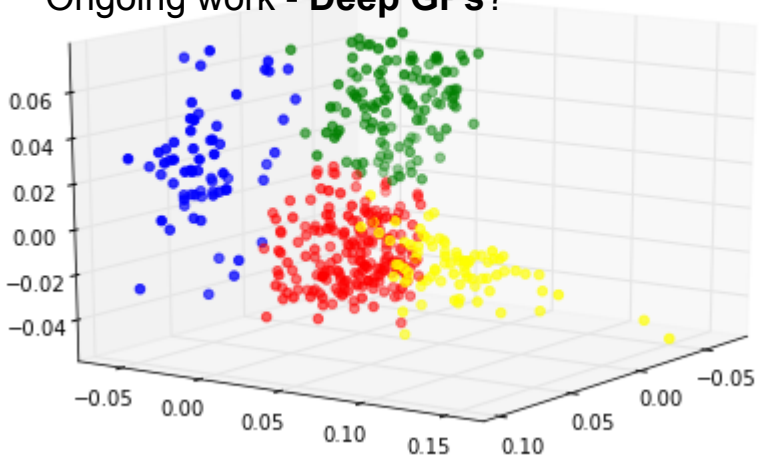


- HSIC-based regression tries to maximise the independence between the residuals and the regression variable (x).
- Has applications in causal inference.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861

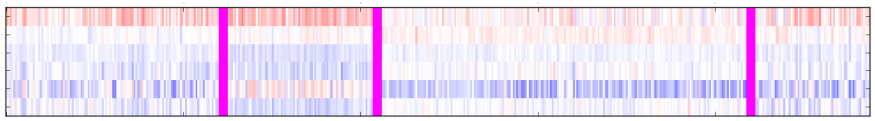
Investigating cancer subtypes in TCGA data:

- The Cancer Genome Atlas (TCGA) provides a lot of public data of various types (RNASeq, DNA Methylation, MicroRNA, SNPs (restricted access), expression arrays, etc.) for many different types of cancer.
- Main aim is to discover links between the different types of cancer.
- Verhaak, R.G., et al. (2010) **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1**. Cancer Cell. **17** (1):98-110 - used Factor Analysis and Consensus Clustering.
- Attempt to repeat using **GPLVM** and **K-means**.
- Find different significant genes.
- Ongoing work - **Deep GPs**?



- Clinical differences:

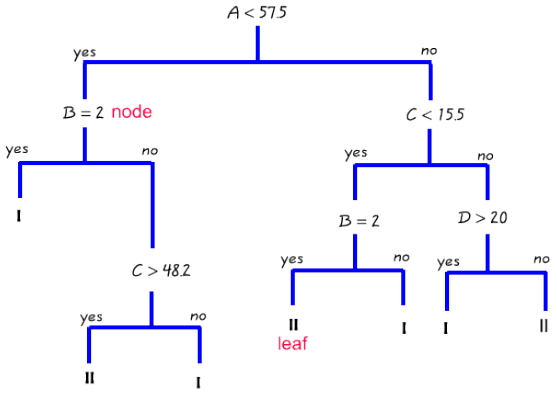
Cluster 0:
Dead: 91/120(75.8333333333%)
Mean Survival time of dead: 40.20
Cluster 1:
Dead: 65/82(79.2682926829%)
Mean Survival time of dead: 41.82
Cluster 2:
Dead: 139/206(67.4757281553%)
Mean Survival time of dead: 37.68
Cluster 3:
Dead: 47/65(72.3076923077%)
Mean Survival time of dead: 528.38



This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861

DREAM9 AML Prediction challenge:

- DREAM programme provides annual challenges in Bioinformatics.
- Chose AML (**Acute Myeloid Leukemia**) Outcome Prediction challenge.
- AML is a very lethal form of leukemia.
- Only approximately a quarter of the patients diagnosed with AML survive beyond 5 years.
- Provided with medical covariates and proteomics data.
- Consists of three sub-challenges:
  1. Predicting whether a patient will go in to remission or not (classification).
  2. Predicting the length of remission (regression).
  3. Predicting the overall survival time (regression).
- Have tried many approaches so far, most successful have been **Random Forests**, sometimes combined with the **GPLVM** for dimensionality reduction.

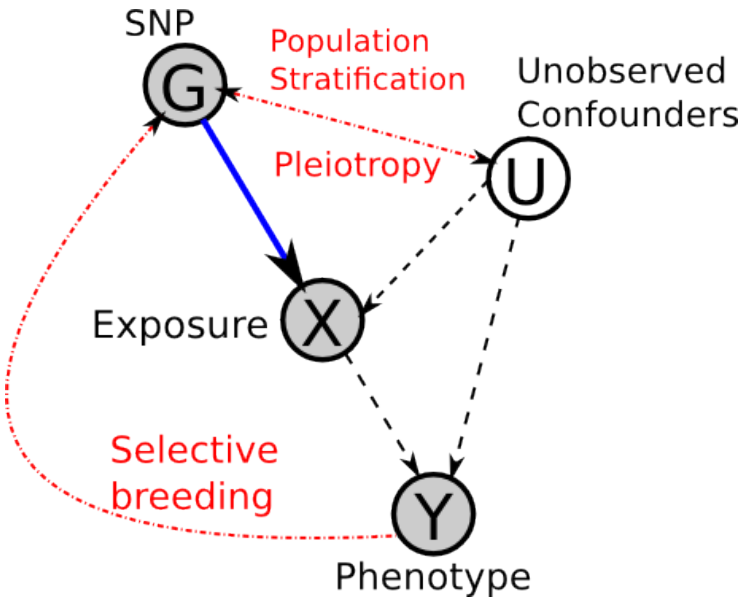


- The Random Forests algorithm uses an ensemble of Decision Trees to make its predictions.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861

Causal inference in epidemiology?

- I believe there are many possible applications of causal inference in epidemiology.
- A classic example is **Mendelian Randomisation** - whereby the genotype is used as an instrumental variable for causal inference, to determine if the environmental exposure (X) has an effect on the disease/phenotype (Y).
- Can we improve the robustness using intermediate epigenetic measurements, etc.?
- Can we determine the necessary covariates to observe.?



- I would appreciate any data for this problem!

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861

Conclusion

- The importance of Machine Learning and Personalized Medicine will only increase as more data becomes available, and data access restrictions are relaxed.
- To take advantage of this, we need viable methods to extract actionable information from the data.
- The **MLPM ITN** provides us with an excellent opportunity to work on cutting-edge methods with world-class institutions and scientists.
- In addition to the opportunities to receive great training from summer schools and workshops.
- I hope to develop practically applicable methods for Personalized Medicine, such as the problems of cancer diagnosis.
- I hope I can co-operate with other nodes in the network, and make the most of the opportunities we have.

Thanks for your time!

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861