Cake Talk

James McMurray

PhD Student
Department of Empirical Inference

26/03/2014

Outline of talk

MAX-PLANCK-GESELLSCHAFT

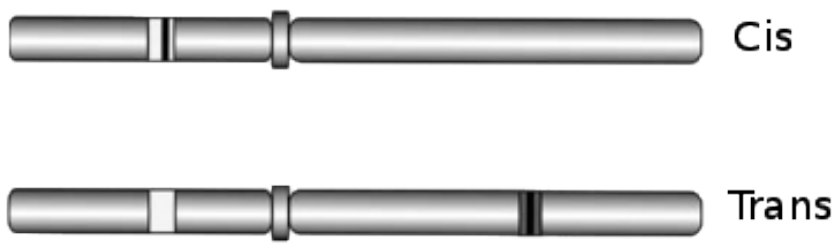
MLPM ITN

- I am part of the Machine Learning for Personalised Medicine (MLPM) Marie Curie Initial Training Network (ITN), which involves 14 PhD students and 14 network nodes (institutions and companies) throughout Europe.
- My project title: "Predicting Phenotype through Interaction of Genotype, Epigenotype and Environment with Probabilistic Models".

MAX-PLANCK-GESELLSCHAFT

eQTL studies: Molecular phenotype

- Expression levels of gene transcripts can be regulated by genetic loci (bases in the genome) that are local to (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, so this can also be useful for establishing prior information for GWASs.



The white bar represents the coding gene, the black line is the eQTL. Adapted from http://openi.nlm.nih.gov/detailedresult.php?img=2817885_CG-10-540_F1&req=4.

MAX-PLANCK-GESELLSCHAFT

Example data

- Table of SNP alleles against samples (i.e. majority allele or minority):

SNP	Sample 1	Sample 2	...
YAL069W	1	1	...
NAL013C	1	0	...

- Table of expression of genes against samples:

Gene exp.	Sample 1	Sample 2	...
YOL161C	0.037	0.187	...
YJR107W	0.078	0.081	...

- For each vector of SNPs (sample), we have a vector of gene expression levels.

MAX-PLANCK-GESELLSCHAFT

Outline of talk

MAX-PLANCK-GESELLSCHAFT

PANAMA model

- Fusi, Stegle *et al.* developed PANAMA[1] 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.

[1] 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

PANAMA model

- Based on a linear model:

$$\mathbf{y_g} = \mu_g + \sum_{k=1}^{K} v_{k,g}\mathbf{S_k} + \sum_{q=1}^{Q} w_{g,q}\mathbf{X_q} + \varepsilon_{\mathbf{g}}$$

  Where $\mathbf{y_g}$ is the gene expression of gene g across N individuals, given K SNPs and Q confounders (latent variables), μ_g is a gene specific mean term (does not vary with individuals), and ε_g is a noise term.

- But the weights $\mathbf{v}$, $\mathbf{w}$ and the latent variables $\mathbf{X}$ are unknown a priori.

- Integrate out the weights using Gaussian priors:

$$p(\mathbf{W}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{w}_q|0, \alpha_q^2\mathbf{I}) \ , \ p(\mathbf{V}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{v}_k|0, \beta_k^2\mathbf{I})$$

MAX-PLANCK-GESELLSCHAFT

GPLVM

- This means one only needs to find the prior parameters and the latent variables.
- This is achieved by putting everything in a GPLVM (Gaussian Process Latent Variable Model):

$$\mathbf{y_d} = g_d(\mathbf{X}; \mathbf{\Theta}) + \mathbf{e_d}$$

- Assume that the observed variables result from a mapping, g, from set of latent variables: $\mathbf{X}$.
- Note this is not the same as the set of confounders $\mathbf{X}$ in the PANAMA model, since in the PANAMA model all of the variables are put in to the GPLVM.
- Assume that the observed variables are independent given the latent variables.

MAX-PLANCK-GESELLSCHAFT

PANAMA GPLVM

- To implement the PANAMA model, use a combination of linear kernels for each part.
- The latent variables in the GPLVM are then the confounders, SNPs, and possible covariates combined.
- Fix the values of the SNPs and covariates, so it only optimises over the parameters and confounders.
- After fitting the model one can calculate the likeilhood ratio between the gene expression given the confounders, with a certain SNP and without it (the null model).

$$LOD_{g,k} = \log\left(\frac{\mathcal{N}(\mathbf{y_g}|\theta\mathbf{s_k}, \sigma_k^2\mathbf{K} + \sigma_e^2\mathbf{I})}{\mathcal{N}(\mathbf{y_g}|\mathbf{0}, \sigma_k^2\mathbf{K} + \sigma_e^2\mathbf{I})}\right)$$

PANAMA GPLVM

- Demonstrates use of GPLVM in accounting for confounders.
- Can add additional covariance matrices for known covariates/population structure e.g. a matrix of genetic relatedness.
- LIMMI[2] model is an extension which adds a covariance matrix to account for SNP-confounder interactions.

[2] Fusi *et al.*, "Detecting regulatory gene-environment interactions with unmeasured environmental factors", Bioinformatics (2013)

MAX-PLANCK-GESELLSCHAFT

Outline of talk

MAX-PLANCK-GESELLSCHAFT

IGPLVM

- Zhang *et al.* developed IGPLVM[3] - permits interpretation of causal relations between observed variables.
- Allows arbitrary noise correlations (whereas the GPLVM assumed independent Gaussian noise in each dimension).
- Invariant to linear non-singular transformations of the data.
- Allows application of causal inference methods (LiNGAM) to estimate causal relations between observed variables.

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

LiNGAM

- Linear Non-Gaussian Acyclic Model[4] (LiNGAM) for causal discovery.
- Assume the underlying causal model is linear, non-gaussian and acyclic.
- Model is:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

  Where $\mathbf{B}$ is a matrix that could be permuted to be lower triangular, given the causal ordering of the variables (acyclic). Diagonal of $\mathbf{B}$ constrained to zeroes.
- Can be written as:

$$\mathbf{x} = \mathbf{A}\mathbf{e}$$

  Where $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$.
- This can be solved by Independent Component Analysis (ICA).

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

LiNGAM

The LiNGAM discovery algorithm:

1. Given an $(m \times n)$ data matrix $\mathbf{X}$, where each column contains one sample vector $\mathbf{x}$ and each row has its mean subtracted.
2. Apply an ICA algorithm to obtain a decomposition $\mathbf{X} = \mathbf{AS}$ where $\mathbf{S}$ has the same size as $\mathbf{X}$ and contains in its rows the independent components. Let $\mathbf{W} = \mathbf{A}^{-1}$.
3. Find the one and only permutation of rows of $\mathbf{W}$ which yields a matrix $\tilde{\mathbf{W}}$ without any zeros on the main diagonal.
4. Divide each row of $\tilde{\mathbf{W}}$ by its corresponding diagonal element, to yield a new matrix $\tilde{\mathbf{W}}'$ with all ones on the diagonal.
5. Compute an estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$ using $\hat{\mathbf{B}} = \mathbf{I} - \tilde{\mathbf{W}}'$.
6. To find a causal order, find the permutation matrix $\mathbf{P}$ which yields a matrix $\tilde{\mathbf{B}} = \mathbf{P}\hat{\mathbf{B}}\mathbf{P}^T$ which is as close as possible to strictly lower triangular.

MAX-PLANCK-GESELLSCHAFT

LiNGAM

- Allows estimation of causal network given we meet the assumptions:
  - Acyclic
  - Linear
  - Non-Gaussian
- We shall see that the IGPLVM model can meet these assumptions in some circumstances.

IGPLVM

- Generating process of $\mathbf{y_t}$ is:

$$\mathbf{y_t} = \mathbf{B}\mathbf{y_t} + \tilde{\mathbf{g}}(\mathbf{x_t}; \theta) + \tilde{\mathbf{e}}_\mathbf{t}$$

  Where $\mathbf{B}$ is a matrix of coefficients of the linear instantaneous influences (the diagonal of $\mathbf{B}$ is constrained to zeros).

- Therefore we can write:

$$(\mathbf{I} - \mathbf{B})\mathbf{y_t} = \tilde{\mathbf{g}}(\mathbf{x_t}; \theta) + \tilde{\mathbf{e}}_\mathbf{t}$$

$$\mathbf{y_t} = (\mathbf{I} - \mathbf{B})^{-1}\tilde{\mathbf{g}}(\mathbf{x_t}; \theta) + (\mathbf{I} - \mathbf{B})^{-1}\tilde{\mathbf{e}}_\mathbf{t}$$

- Linear causal relations are implied in the structure of the noise.

IGPLVM Causal Inference

1. Obtain estimates of noise terms $\hat{e}_{it}$ by fitting IGPLVM.
2. Can recover the Gaussian-Markov graph implied by the precision matrix of $\hat{e}_{it}$.
3. Test if $\hat{e}_{it}$ are Gaussian.
   - If yes, then we can only recover Markov-equivalence class by conditional independence testing.
   - If not, we can apply LiNGAM to $\hat{\mathbf{e}}_{\mathbf{t}}$.
4. Assume causal relations are acyclic, although there are methods for discovering cyclic causal models too[5]

[5] Lacerda, *et al.*, "Discovering Cyclic Causal Models by Independent Components Analysis", arXiv:1206.3273, 2008

IGPLVM Implementation

- If we naïvely added new parameters to the GPLVM for arbitrary noise, it would result in a $DN \times DN$ kernel matrix - this is very large, and storage would be a problem.
- Can achieve a more efficient parameterization through minor approximations.
- Use Cholesky factorisation to decompose the noise covariance:

$$\text{cov}(\mathbf{e_t}) = \mathbf{L}\mathbf{L}^{\mathbf{T}}$$

  Where $\mathbf{L}$ is lower-triangular with positive diagonals.
- Can write the noise as:

$$\mathbf{e_t} = \mathbf{L}\mathbf{e}_\mathbf{t}^*$$

IGPLVM Implementation

- Using this to reformulate the model:

$$\mathbf{y_t} = \mathbf{g}(\mathbf{x_t}; \theta) + \mathbf{L}\mathbf{e}_\mathbf{t}^*$$

Can be written as:

$$\mathbf{y_t} = \mathbf{L}\mathbf{y}_\mathbf{t}^\mathbf{I}$$

Where:

$$\mathbf{y}_\mathbf{t}^\mathbf{I} \triangleq \mathbf{L}^{-1}\mathbf{g}(\mathbf{x_t}; \theta) - \mathbf{e}_\mathbf{t}^*$$

- $\mathbf{y}_\mathbf{t}^\mathbf{I}$ can then be modelled by the original GPLVM.
- The resulting (positive) data log likelihood is:

$$\mathcal{L}_y = \log(\mathbf{Y}|\mathbf{X}, \mathbf{L}, r, \gamma)$$

$$= -N \log |\mathbf{L}| - \frac{1}{2}\sum_{i=1}^{D} \mathbf{Y}_\mathbf{i}^{\mathbf{I}\mathbf{T}} \mathbf{K}_{\mathbf{y}^\mathbf{I}}^{-1} \mathbf{Y}_\mathbf{i}^\mathbf{I} - \frac{D}{2}\log |\mathbf{K}_\mathbf{y}^\mathbf{I}| - \frac{DN}{2}\log(2\pi)$$

MAX-PLANCK-GESELLSCHAFT

IGPLVM Implementation

- Due to the linear transformation $\mathbf{L}$, the components of $\mathbf{y_t^i}$ are independent.
- The normal GPLVM can be applied to $\mathbf{y_t^l}$ because the dependence is captured in $\mathbf{L}$.
- Results that we only need $\frac{D(D-1)}{2}$ more parameters due to $\mathbf{L}$, but these can be obtained in closed form.
- $\mathbf{K_{y^l}}$ is $N \times N$, not $DN \times DN$.
- Paper uses following kernel:

$$\mathbf{K_{y^l}} = r \exp\left(\frac{-\gamma}{2}||\mathbf{x_t} - \mathbf{x_{t'}}||^2\right) + \delta_{\mathbf{x_t},\mathbf{x_{t'}}}$$

IGPLVM Implementation

- $\mathbf{L}^{-1}$ can be optimally set each iteration in closed form:

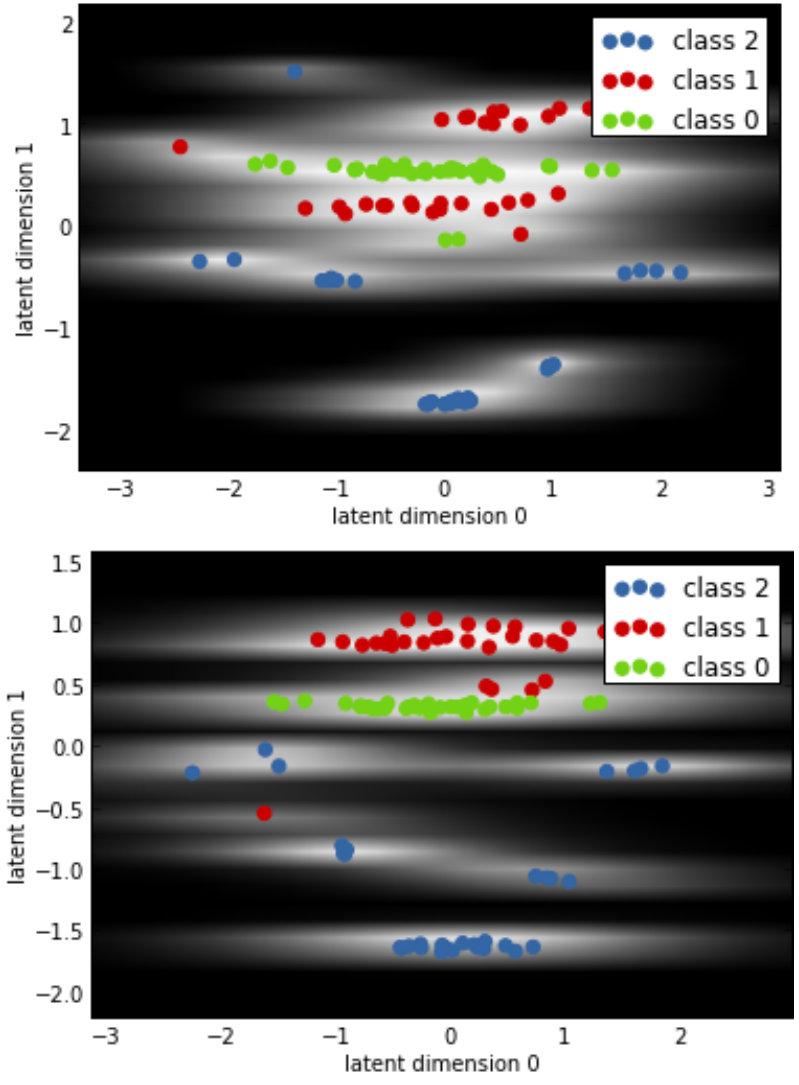
$$\mathbf{L}^{-1} = \text{inv}\left(\text{chol}_{\text{LT}}\left(\frac{1}{N}\mathbf{y}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y}^{\mathbf{T}}\right)\right)$$

  Obtained by differentiating the log likelihood with respect to $\mathbf{L}^{-1}$

- The kernel parameters are learnt by minimizing the negative log likelihood as in the normal GPLVM. The optimisation is done by the Scaled Conjugate Gradient method.
- The latent variables are initialised by PCA.

Example plots: Oil flow dataset

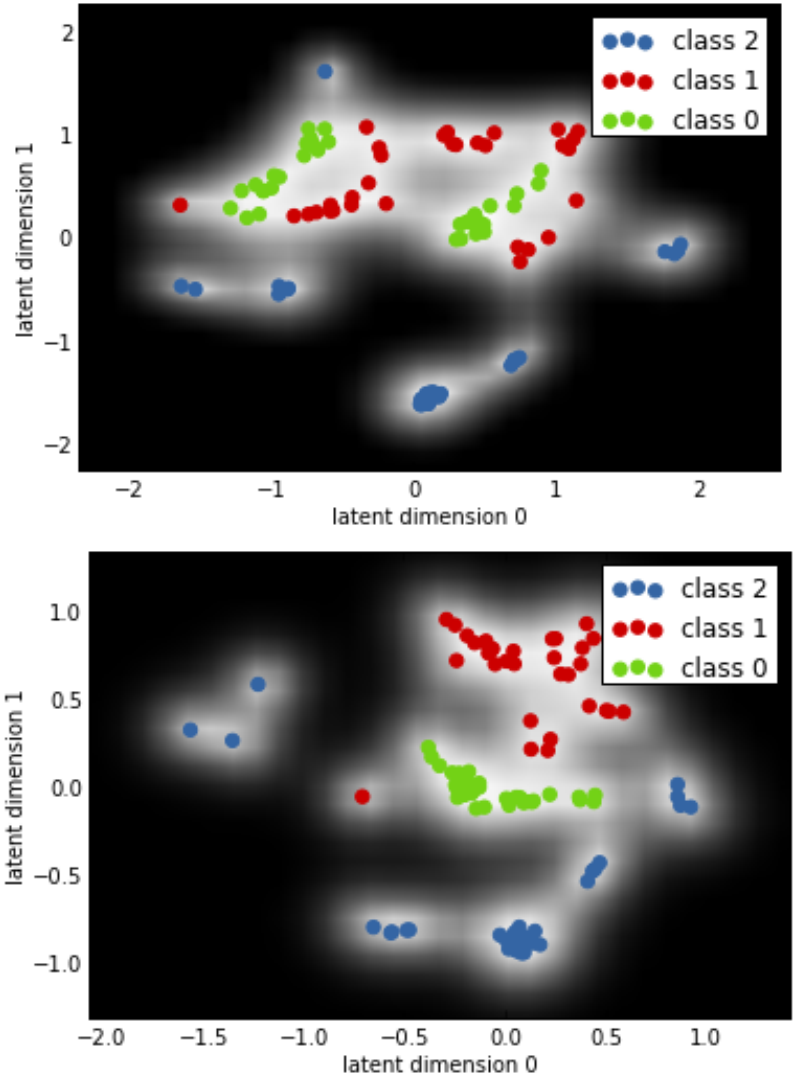
- RBF kernel, ARD enabled (one length-scale per X dimension), 2 latent dimensions. GPVLM top, IGPLVM bottom:



MAX-PLANCK-GESELLSCHAFT

Example plots: Oil flow dataset

- RBF kernel, ARD disabled, 2 latent dimensions. GPVLM top, IGPLVM bottom:



MAX-PLANCK-GESELLSCHAFT

Outline of talk

MAX-PLANCK-GESELLSCHAFT

GPDM

- The Gaussian Process Dynamical Model (GPDM) adds a dynamic mapping across the latent variables.
- Can incorporate temporal ordering of data.
- Very useful if the confounders are expected to change with time.
- Add additional Gaussian Process from $\mathbf{X}_{:\mathbf{N}-\mathbf{1}}$ to $\mathbf{X}_{\mathbf{1}:\mathbf{N}}$
- Need to add new terms to log-likelihood:

$$\mathcal{L}_x = \frac{-q}{2}\log|\mathbf{K}_\mathbf{x}| + \frac{1}{2}\mathrm{tr}\left(\mathbf{K}_\mathbf{x}^{-\mathbf{1}}\mathbf{X}_{\mathbf{out}}\mathbf{X}_{\mathbf{out}}^{\mathbf{T}}\right)$$

- Need to add new gradients for the log likelihood.
- Can optimise parameters for both kernels simultaneously.

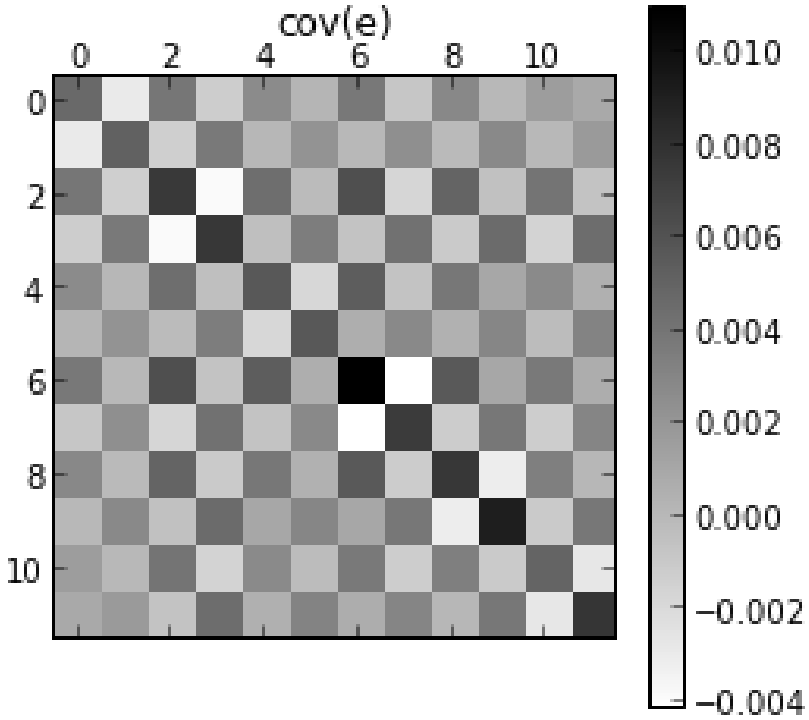
MAX-PLANCK-GESELLSCHAFT

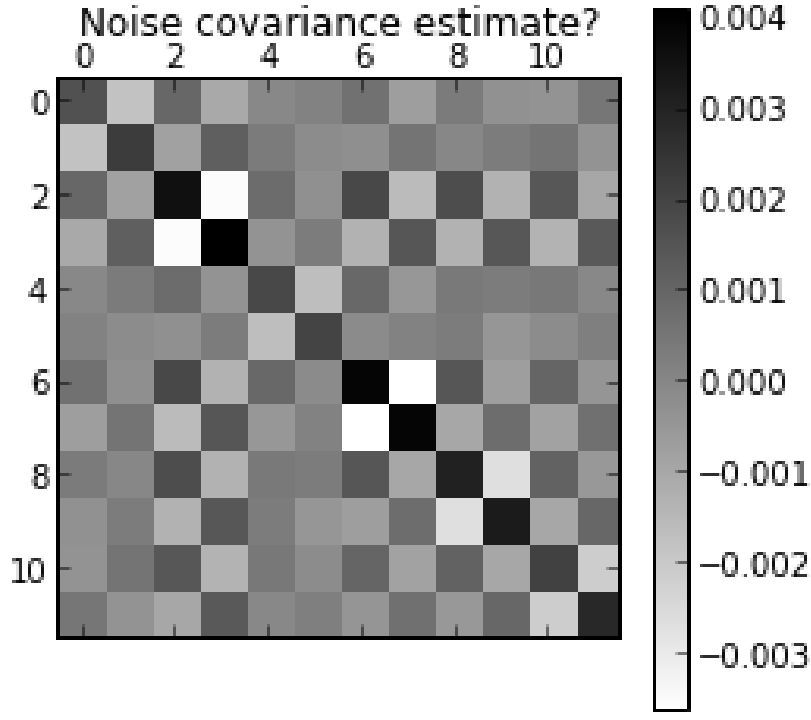
Outline of talk

MAX-PLANCK-GESELLSCHAFT

Current Problems: Noise covariance reconstruction from IGPLVM

- The reconstructed noise from the estimated error, does not exactly match that from the estimate of **L** itself - not clear that scaling problems are solved.





MAX-PLANCK-GESELLSCHAFT

Current Problems: GPDM

- Enabling the GPDM currently has no effect on latent variables.
- But kernel parameters are definitely being learnt.
- Likely a problem in the gradients of the likelihood - probably $\frac{d\mathcal{L}}{d\mathbf{X}}$ is not correctly including the GPDM terms.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = \frac{\partial \mathcal{L}}{\partial \mathbf{K_y}} \odot \frac{\partial \mathbf{K_y}}{\partial \mathbf{X}} + \frac{\partial \mathcal{L}}{\partial \mathbf{K_x}} \odot \frac{\partial \mathbf{K_x}}{\partial \mathbf{X}} - \mathbf{K_x^{-1}} \mathbf{X}_{\text{out}} \odot \frac{\partial \mathbf{X}_{\text{out}}}{\partial \mathbf{X}}$$

- But $\mathbf{K_y}$ is $N \times N$, while $\mathbf{K_x}$ is $(N-1) \times (N-1)$.
- Cannot simply add the gradients, MATLAB code must do something more sophisticated.

MAX-PLANCK-GESELLSCHAFT

Outline of talk

MAX-PLANCK-GESELLSCHAFT

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.
- May need more sophisticated ways to deal with computational cost for large datasets. There has been recent work on Gaussian Processes for Big Data[6].

[6] J. Hensman, N. Fusi, and N. Lawrence. "Gaussian processes for big data." arXiv:1309.6835 (2013)

Outline of talk

MAX-PLANCK-GESELLSCHAFT

Conclusion

- IGPLVM allows application of causal inference methods to scenarios where the GPLVM is used (i.e. latent confounders).
- Appear to be useful possible applications in the real world - such as gene regulatory networks.
- Hopefully application to PANAMA model and eQTL data will be useful, although this does not use temporal data. But data and model (with Python code) are freely available.
- Still need to solve some remaining implementation problems in Python.

Thanks for your time

Questions?