

# Latent factor models and Gaussian processes with applications to scRNA-seq data



Magdalena Strauß

Wellcome Sanger Institute

22nd September 2020

# Overview

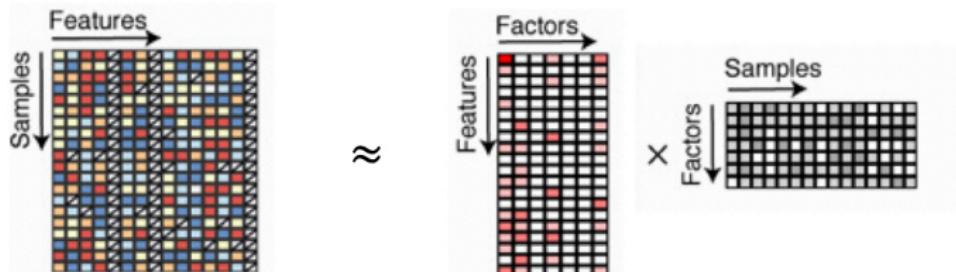
- 1 Latent factor models
- 2 Gaussian processes
- 3 Applications of GPs to pseudotime ordering
- 4 Gaussian processes for clustering
- 5 GP clustering for pseudotemporal data

# Introduction: latent variable methods

- Latent: not directly observable
- Assumption that high-dimensional data can be explained by a lower-dimensional representation of it
- Helps with datasets with large numbers of observed variables that reflect a smaller number of underlying variables
- Difference to general dimensionality reduction: interpretability

# Latent factors I

## Linear matrix decomposition



Adapted from Argelaguet et al. (2018)

$$\mathbf{X} \approx \mathbf{W}\mathbf{z}$$

$$\mathbf{X} \sim N(\mathbf{W}\mathbf{z}, \Psi)$$

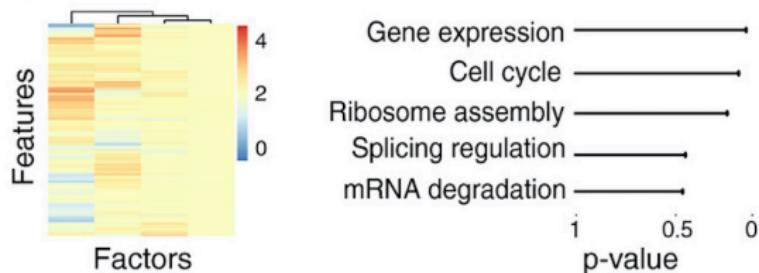
**W** weight matrix, **z** factors

# Interpretation of factors and weights I

- Factor model separates all the sources of variation in the data into factors
- Then we look at which genes, mutations etc. comprise each factor - large entries in  $\mathbf{W}$ 
  - top-weighted features
  - enrichment analysis to find relevant features

## Annotation of factors

Inspection of loadings Feature set enrichment analysis



Extract from Fig. 1 in Argelaguet et al. 2018

$$\mathbf{X} \approx \mathbf{Wz}$$

$$\mathbf{X} \sim N(\mathbf{Wz}, \Psi)$$

$\Psi$  is a diagonal matrix  $\rightarrow \mathbf{z}$  explains the correlation between the samples

If  $\mathbf{X}$  is not centred (has mean  $\mathbf{0}$ ), then

$$\mathbf{X} \sim N(\mathbf{Wz} + \mu, \Psi)$$

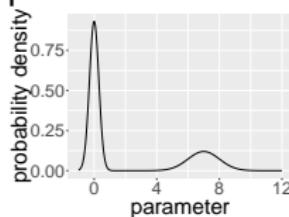
With  $\Psi = \sigma^2 \mathbf{I}$ , *probabilistic PCA*

The latent factors approximate the covariance matrix of  $\mathbf{X}$ :

$$\text{cov}(\mathbf{X}) = \mathbf{W}\mathbf{W}^T + \Psi$$

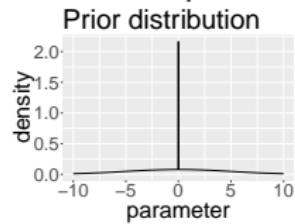
# Introduction to Bayesian methods I

- Parameters as stochastic quantities with a probability distribution → posterior distribution



# Introduction to Bayesian methods II

- There may be prior information concerning the parameter → prior distribution; Example: sparsity inducing prior: each individual variable has a low probability of being relevant



# Introduction to Bayesian methods III

- Bayes' theorem:  $\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- That is if  $H(\theta)$  is the prior distribution of a parameter  $\theta$ , and  $p(x | \theta)$  is the distribution of the data, then for the posterior distribution  $q(\theta | x)$ :  $q(\theta | x) = \frac{p(x|\theta)H(\theta)}{\int p(x|\theta)d\theta}$

Model: For each factor  $i$

$$\begin{aligned}\mathbf{z}_i &\sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ \mathbf{x} &\sim N(\mathbf{Wz}, \boldsymbol{\Psi})\end{aligned}$$

Inference:

$$\mathbf{z}_i \sim N(\mathbf{m}_i, \boldsymbol{\Sigma}_i)$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_i &= (\boldsymbol{\Sigma}_0^{-1} + \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W})^{-1} \\ \mathbf{m}_i &= \boldsymbol{\Sigma}_i (\mathbf{W}^T \boldsymbol{\Psi}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \boldsymbol{\Sigma}_0 \boldsymbol{\mu}_0)\end{aligned}$$

Murphy 2012, chapter 12

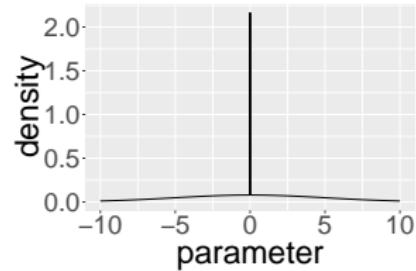
# Unidentifiability of latent factors

- Multiplying  $\mathbf{W}$  by a rotation matrix (an orthogonal matrix) will be compensated by a rotation in  $\mathbf{z}$
- To obtain unique solution, need to remove  $L(L - 1)/2$  degrees of freedom, where  $L$  is the number of factors - as there are  $L(L - 1)/2$  rotation matrices of size  $L \times L$
- Ways to address this
  - Force  $\mathbf{W}$  to be orthogonal, order columns by decreasing variance of factor, this is probabilistic PCA, similar to standard PCA, problem: makes factors identifiable, but not more interpretable
  - $\mathbf{W}$  lower triangular, also interpretability problems
  - Sparsity: methods that set entries of  $\mathbf{W}$  to 0, without prespecifying which one - widely used for applications in genomics,

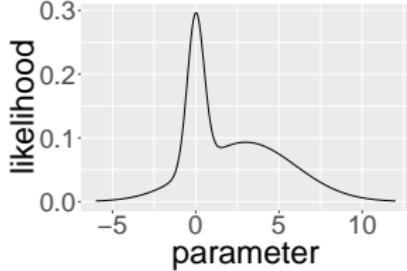
# Spike-and-slab priors I

Example 1:

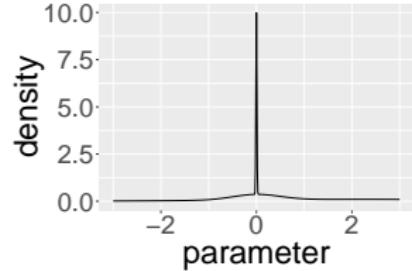
Prior distribution



Likelihood



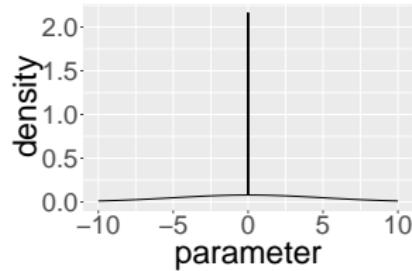
Posterior



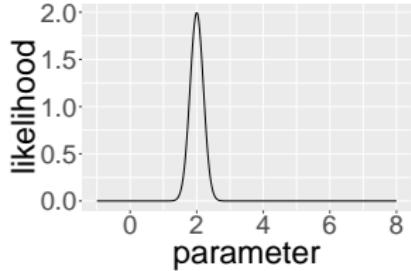
# Spike-and-slab priors II

Example 2:

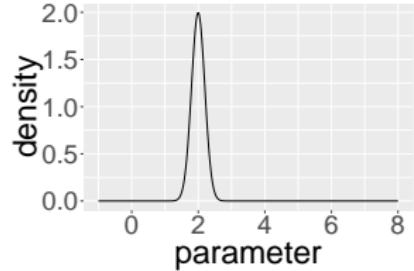
Prior distribution



Likelihood



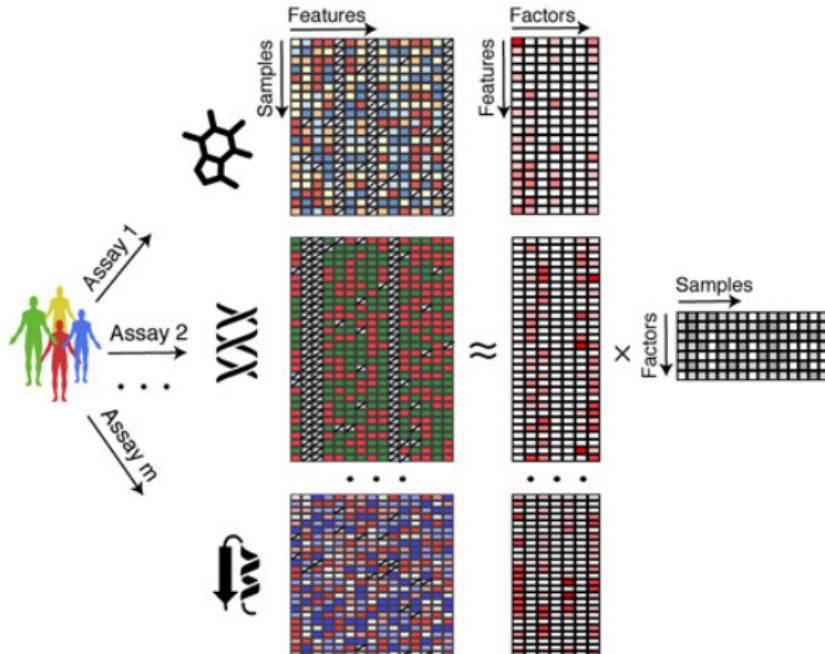
Posterior



- Based on group factor analysis: Virtanen et al. 2012:
- dependencies between variable groups, groups could be e.g. gene expression, methylation, somatic mutations , etc.; or each group could represent a different tissue
- groups may correspond to multi-omics data (multiple views of the same data), alternative measurements ...

- group-wise sparse factors, present in a subset of the groups
- factors explain set-specific and allow understanding variation across sets
- Bioconductor implementation and framework for multi-omics data:  
Argelaguet et al. 2018

# multi-omics factor analysis



Argelaguet et al. 2018

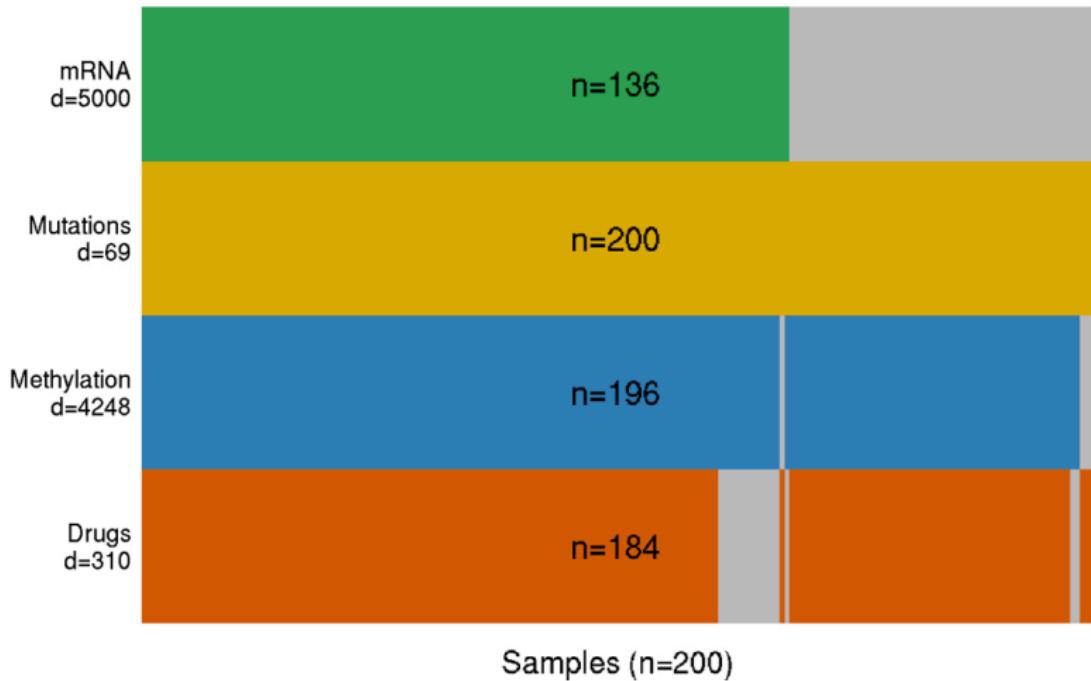
# MOFA-example I

Taken from Argelaguet et al. 2018 and Velten and Argelaguet 2020

Data:

- Dietrich et al. 2018
- chronic lymphocytic leukaemia
- ex vivo drug response measurements
- somatic mutations
- RNA-seq
- DNA methylation

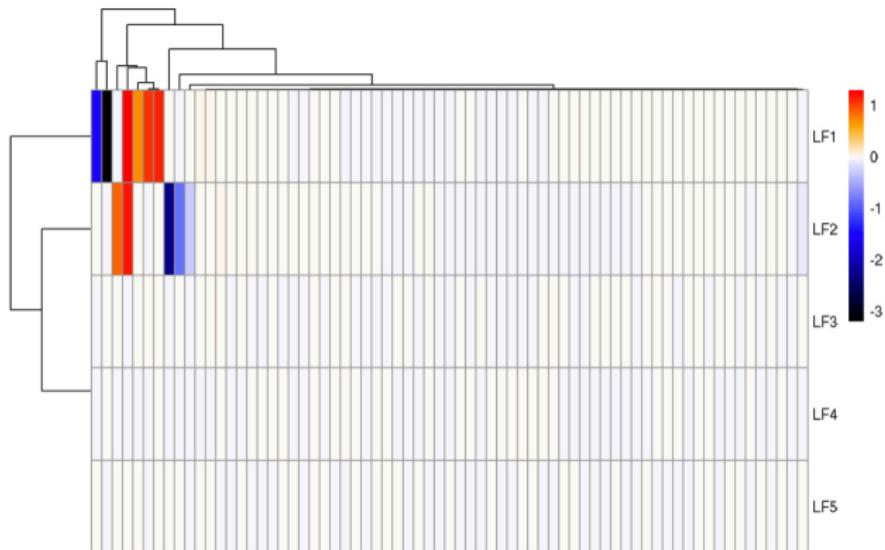
# MOFA-example II



From Velten and Argelaguet 2020

# MOFA-example III

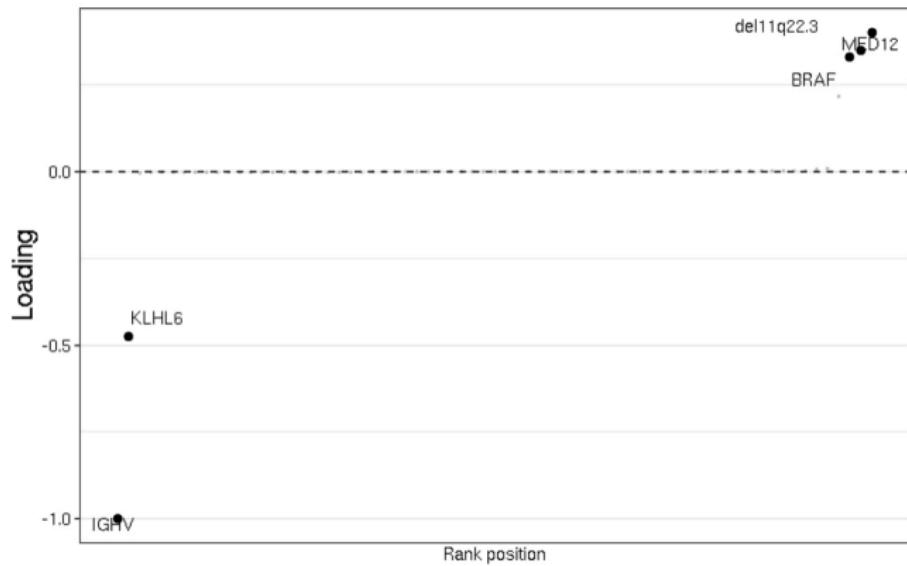
Heatmap of weights matrix  $\mathbf{W}$  for mutation data



Velten and Argelaguet 2020

## MOFA-example IV

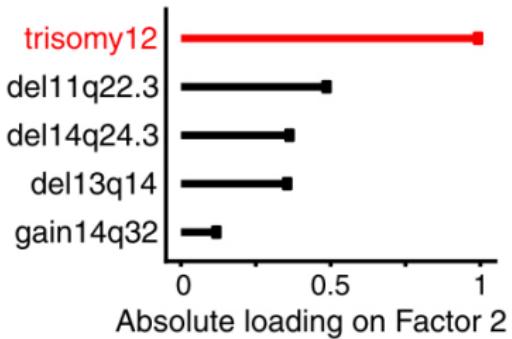
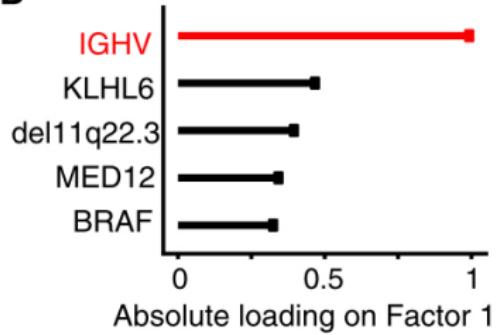
For factor 1, there are two features with highly negative, and three with highly positive weights.



Velten and Argelaguet 2020

# MOFA-example V

D



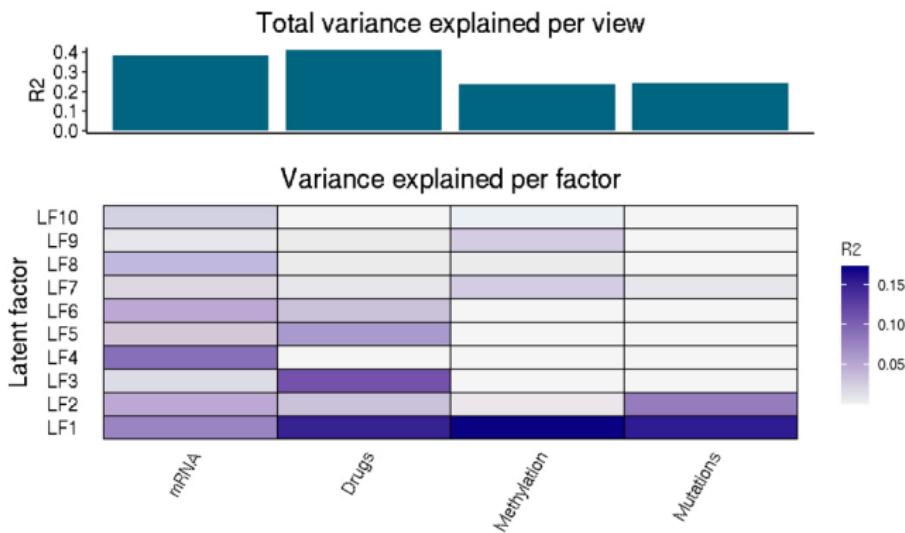
Argelaguet et al. 2018

IGHV is a major clinical marker

# MOFA-example VI

Remember that  $\mathbf{W}^T \mathbf{W}$  model approximates the covariance matrix of  $\mathbf{X}$ :

$$\text{cov}(\mathbf{X}) = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}$$



# Combining observed and latent factors

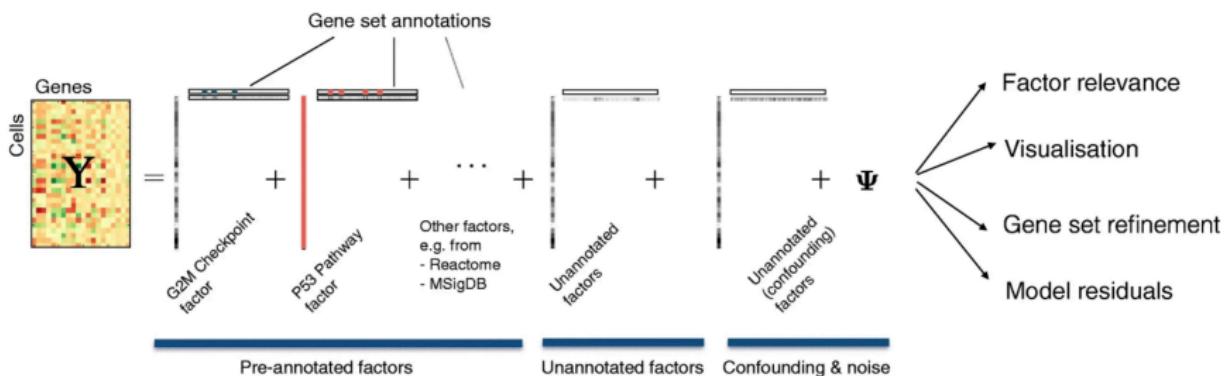
## - f-scLVM

Buettner et al. 2017

Combines observed and latent factors

**a**

### Factor decomposition



**b**

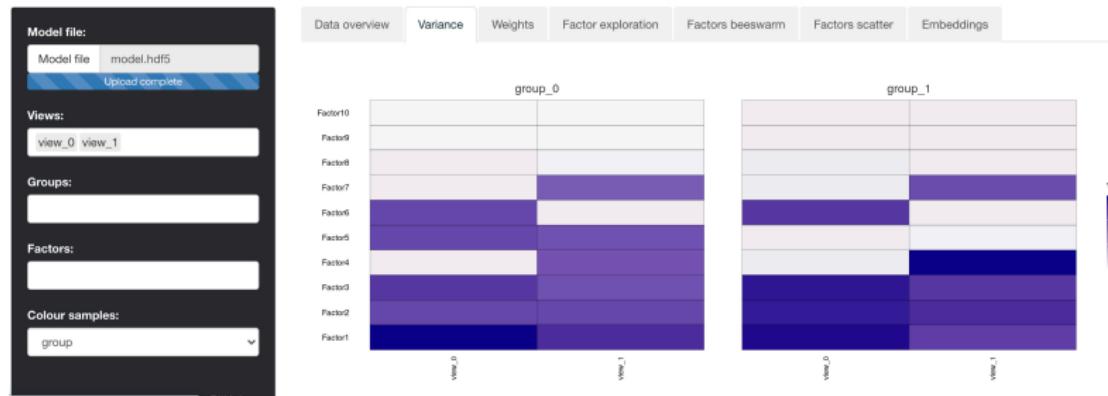
### Downstream analyses

Figure 1 from Buettner et al. 2017

- MOFA (Argelaguet et al. 2018):  
<https://www.bioconductor.org/packages/release/bioc/html/MOFA.html>
- f-scLVM(Buettner et al. 2017: ht-  
[tps://www.bioconductor.org/packages/release/bioc/html/slalom.html](https://www.bioconductor.org/packages/release/bioc/html/slalom.html)
- In addition to multi-omics, sample groups: e.g. batch, donors, or experimental conditions, MOFA+ (Argelaguet et al. 2020): Python or R, with downstream analysis in R:  
<https://github.com/bioFAM/MOFA2>
- Web interface for downstream analysis for small datasets:  
<http://www.ebi.ac.uk/shiny/mofa/>



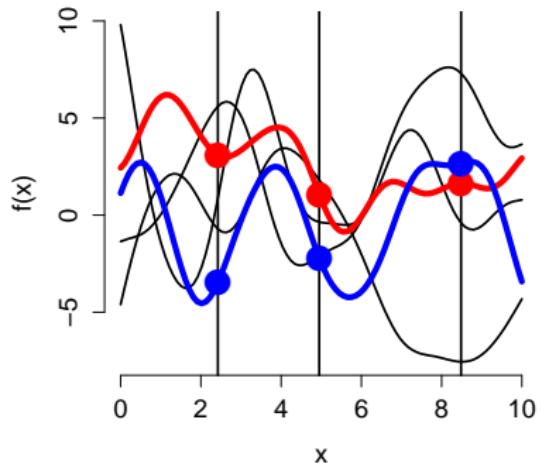
## MOFA+ model exploration



# Summary

- Latent factors find the linear directions of largest variation in the data
- Sparse latent factors allow identifiability and interpretability for high dimensional data
- Extensions for multi-omics data and multi-sample, multi-batch data
- We learn which genes and groups of genes are linked to the major directions of variability in the data

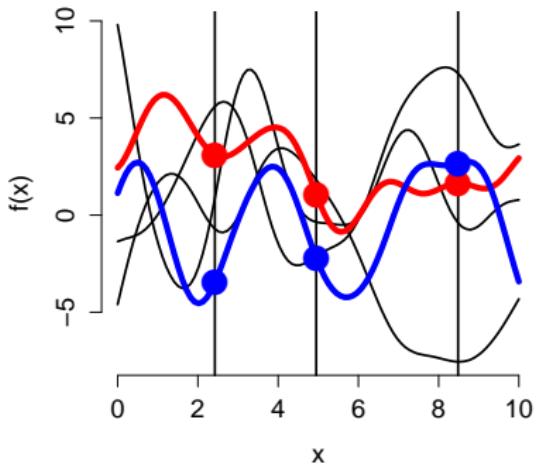
# Introduction to Gaussian processes



Family of functions via covariance  $K$   
on input points  $x$   $y \sim N(m_x, K_{xx})$

Figure by Lorenz Wernisch

# Introduction to Gaussian processes



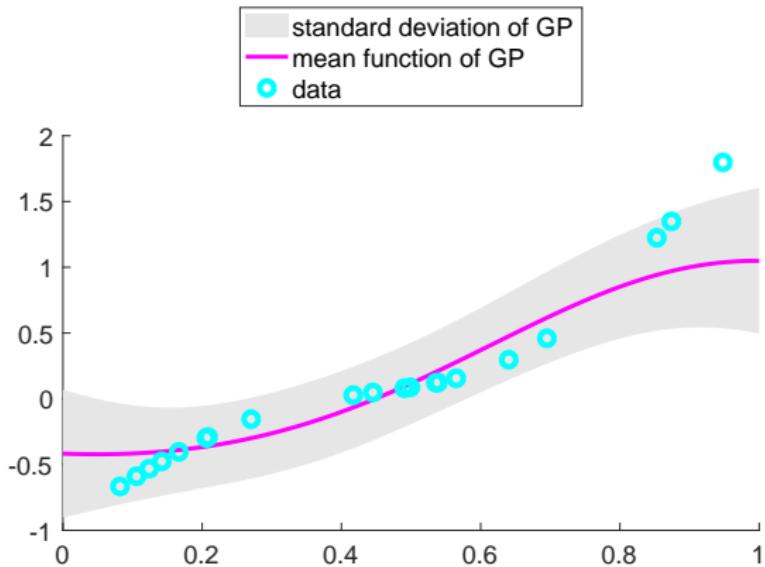
Family of functions via covariance  $K$   
on input points  $x \sim N(m_x, K_{xx})$

Figure by Lorenz Wernisch

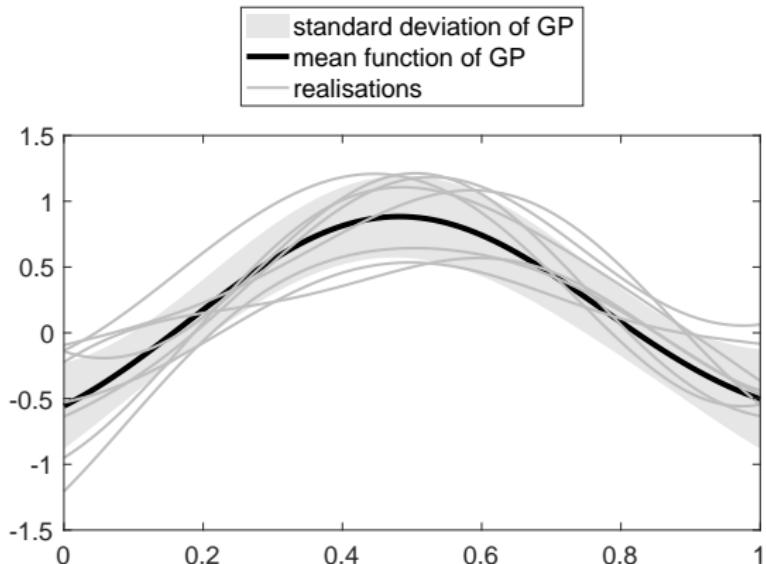
$$\text{Gaussian } \text{cov}(x, x^*) = \sigma_W^2 \exp\left(-\frac{1}{2L^2}(x - x^*)^2\right)$$

$$\text{Matérn}(3/2) \text{ cov}(x, x^*) = \sigma_W^2 \left(1 + \frac{\sqrt{3}}{L}|x - x^*|\right) \exp\left(-\frac{\sqrt{3}}{L}|x - x^*|\right)$$

# Noise-free Gaussian processes-mean function and standard deviation I

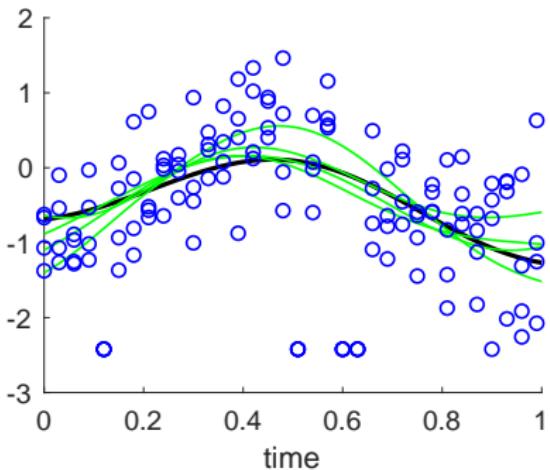


# Noise-free Gaussian processes-mean function and standard deviation II



# Noisy observations

— mean function of GP  
 — realisations  
 ○ noisy observations



$$\text{cov}(x, x^*) = \sigma_W^2 \exp\left(-\frac{1}{2L^2}(x - x^*)^2\right) + \sigma^2 I_{xx}$$

# Noisy observations

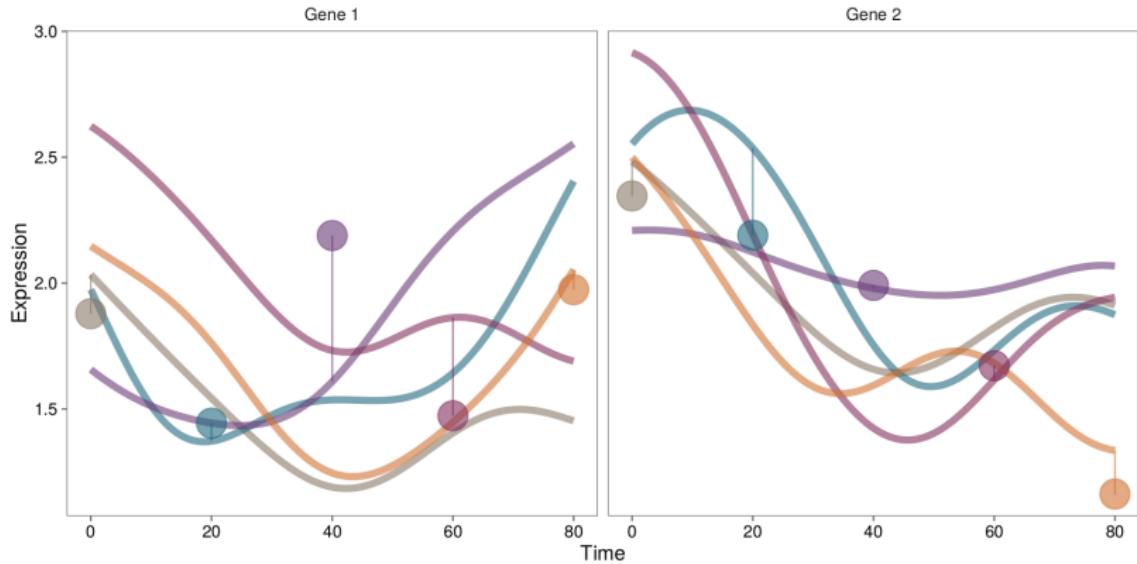
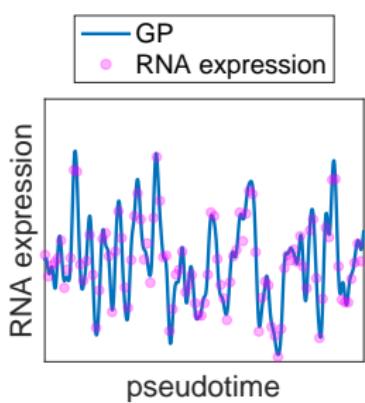


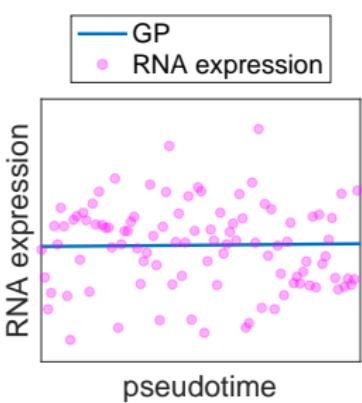
Figure by John Reid

# Parameters $L$ and $w$

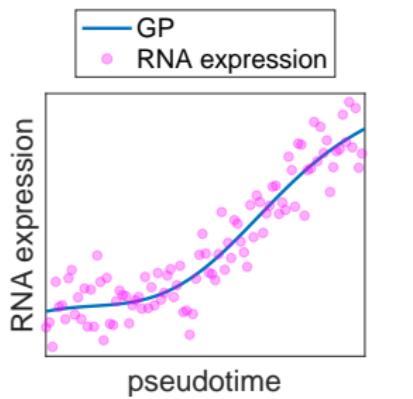
## Length and scale parameters



Short  $L$ , large  $w$

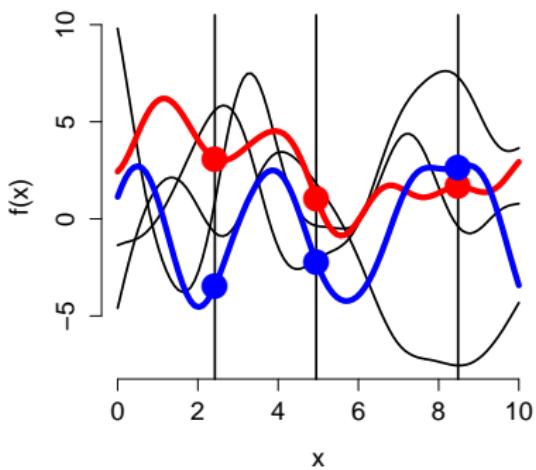


Large  $L$ , small  $w$



good fit of hyper-  
parameters

# GP regression: mean function and covariance given the data



Family of functions via covariance  $K$   
on input points  $x$

$$y \sim N(0, K_{xx})$$

Prediction for  $x^*$  from  $(x, y)$

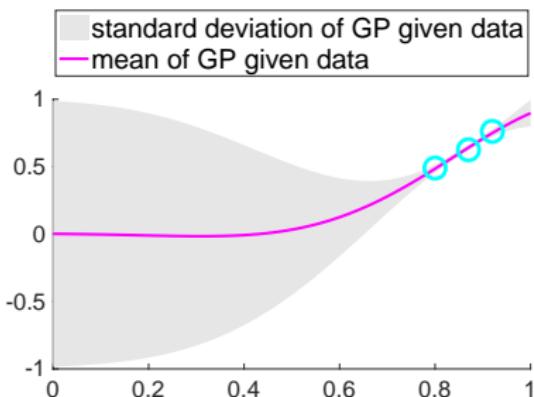
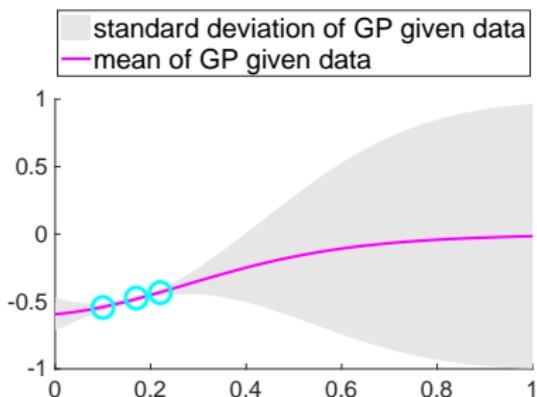
$$y^* \sim N(K_{x^*x}(K_{xx} + \sigma^2 I_{xx})^{-1}y, \Sigma)$$

$$\Sigma = K_{x^*x^*} - K_{x^*x} K_{xx}^{-1} K_{x^*x}$$

Figure by Lorenz Wernisch

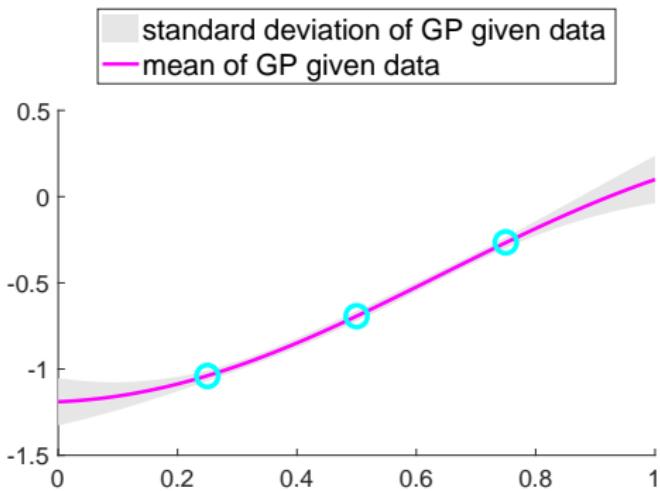
$$\text{Gaussian cov}(x, x^*) = \sigma_W^2 \exp\left(-\frac{1}{2L^2}(x - x^*)^2\right)$$

# GP regression: mean function and covariance given the data I



## GP regression:

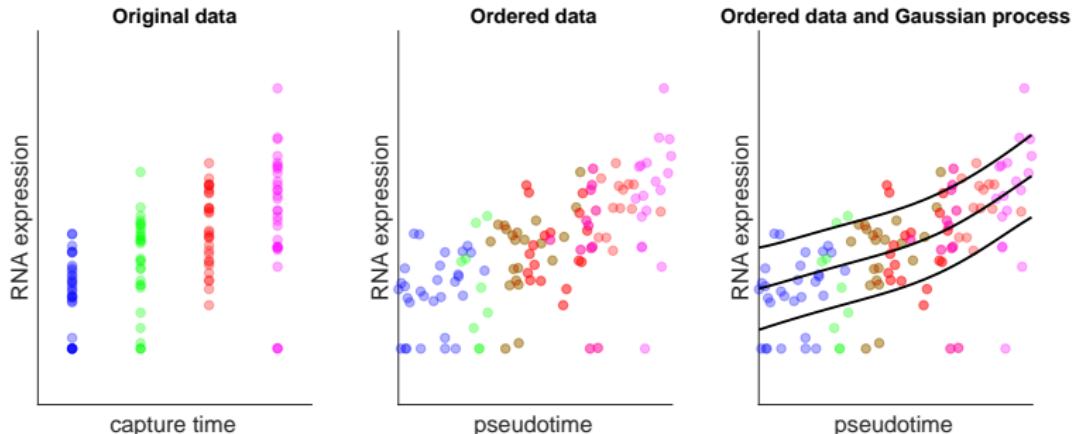
## mean function and covariance given the data II



# Why pseudotime ordering

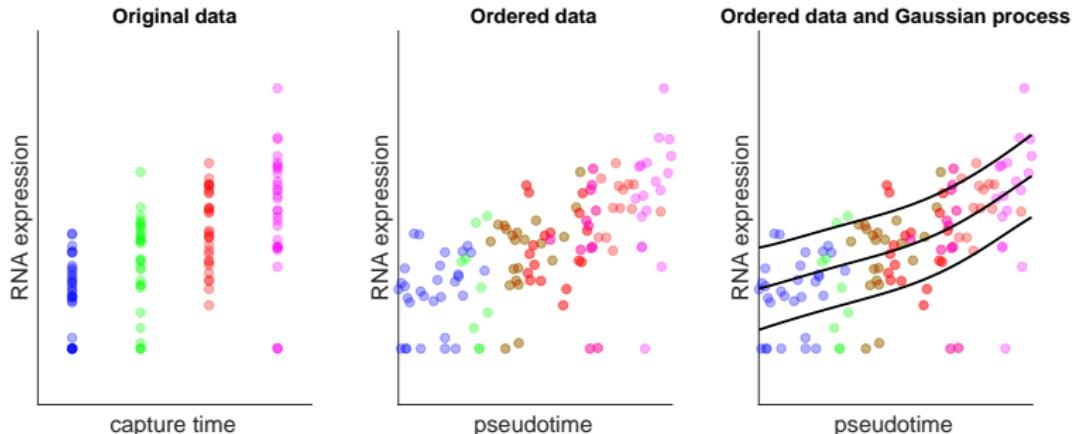
- Better resolution of response or developmental trajectories than we get from only a few capture times, e.g. identify potential master regulators whose expression changes before that of other genes
- Can be used to infer gene **regulatory networks** thanks to the extra dynamic information (e.g. Hamey et al. (2017))
- Understand which genes are **differentially expressed** (e.g. Campbell and Yau (2016))
- Identify precocious cells ahead in the development compared to other cells

# Pseudotime ordering of scRNA-seq/qPCR data



- Data at capture times (sometimes only one), only a single measurement per cell
- Individual cells are at slightly different developmental stages at the same capture time.

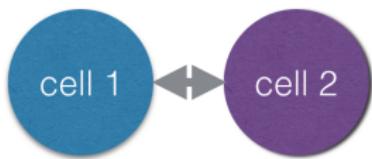
# Pseudotime ordering of scRNA-seq/qPCR data



- Pseudotime: measures progress of cells through a transition
- Using 100-200 such RNA expression series of genes selected by differences of mean expression for different capture times or high mean and variance genes

# Model: the likelihood

- Gene expression trajectories  $y_g = (y_g(o_1), \dots, y_g(o_T))$  for each gene  $g$  modelled by Gaussian processes (GPs), conditional on an ordering  $\mathbf{o}$  of the cells
- $y_g \sim N(\mathbf{0}, \sigma^2 \cdot \mathbf{K}) + \text{noise}$ , multivariate Normal
- Variance-covariance matrix  $\mathbf{K}$  determined by a length parameter
- Covariance between two cells depends on their pseudotime distance and on the length parameter



stronger correlation

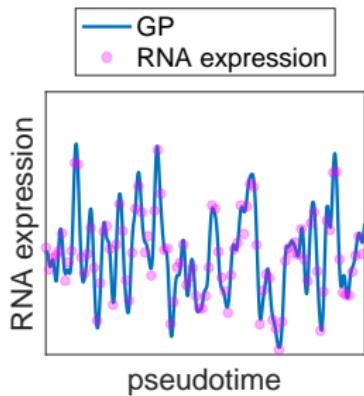


weaker correlation

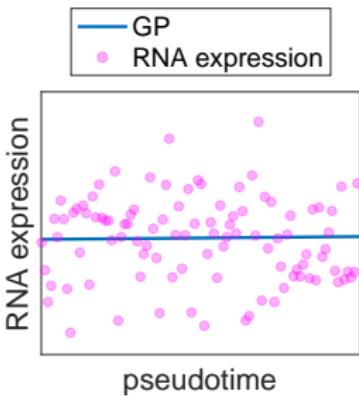
# Model

- Correct ordering  $\bullet$  of cells:

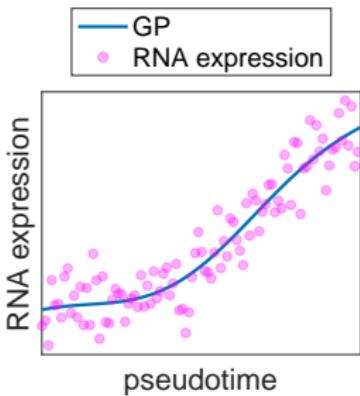
- Relatively low noise
- More of the variation captured by the profile
- Informative priors for noise and length parameters of the GP



Low posterior probability

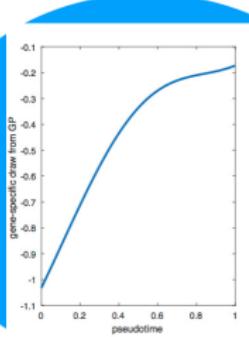
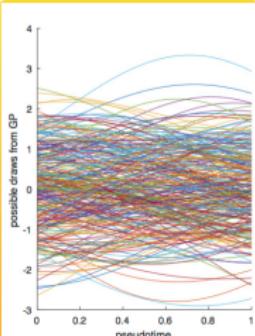
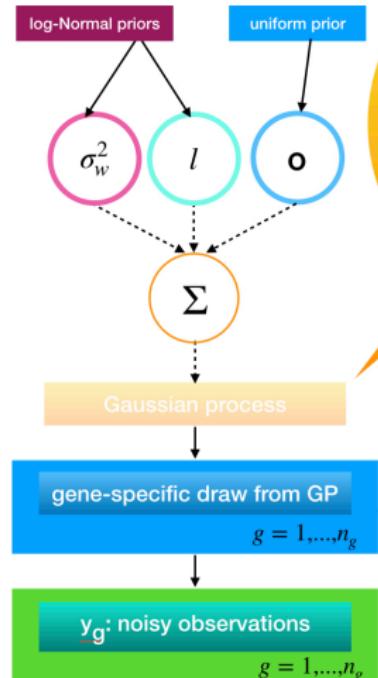


Low posterior probability

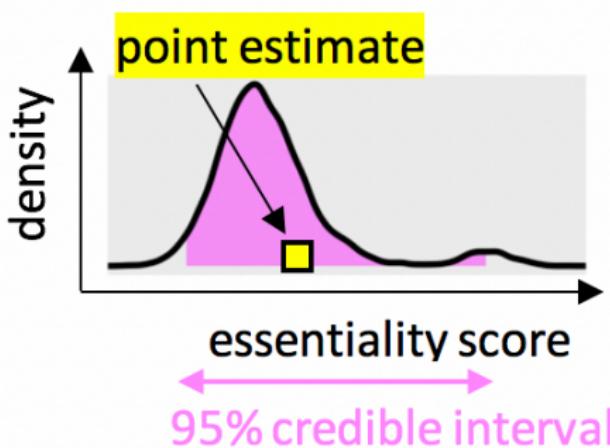


High posterior probability

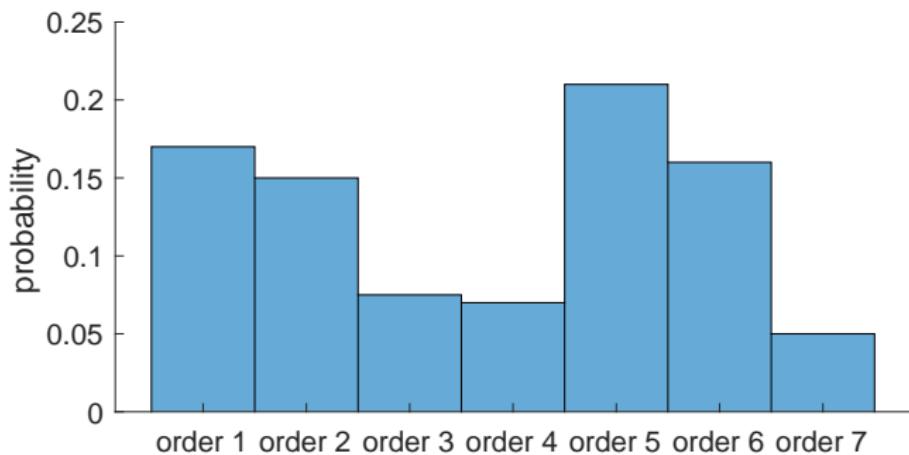
# Summary of pseudotime model



- Recall that the posterior distribution of a parameter is given by  $q(\theta | x)$ :  $q(\theta | x) = \frac{p(x|\theta)H(\theta)}{\int p(x|\theta)d\theta}$
- Sampling from the posterior distribution or
- Approximating the posterior distribution - variational inference, fast, not able to infer uncertainty accurately



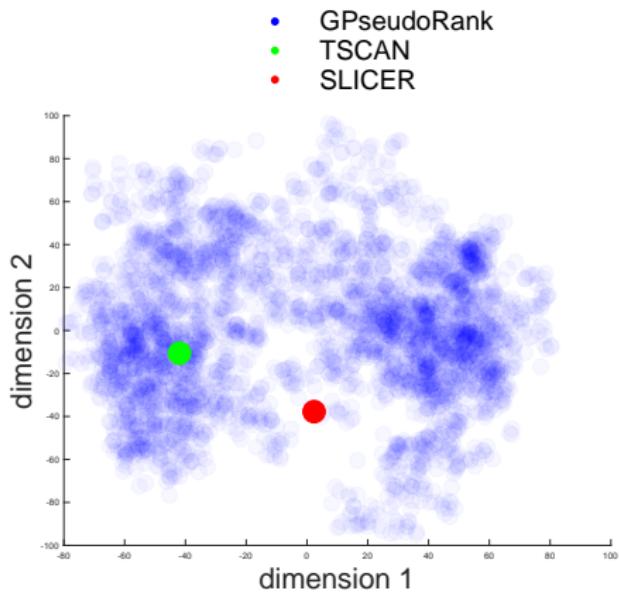
# Inference methods for Bayesian models II



A posterior distribution for cell orders

- Huge number of theoretically possible orders:
  - For 150 cells:  $5.7e+262$  orders
  - But for most the posterior probability is close to 0
  - Sample orders from all neighbourhoods around areas of higher posterior density
  - Convergence
- Additional approximation step for large data sets (more than 550-600 cells)

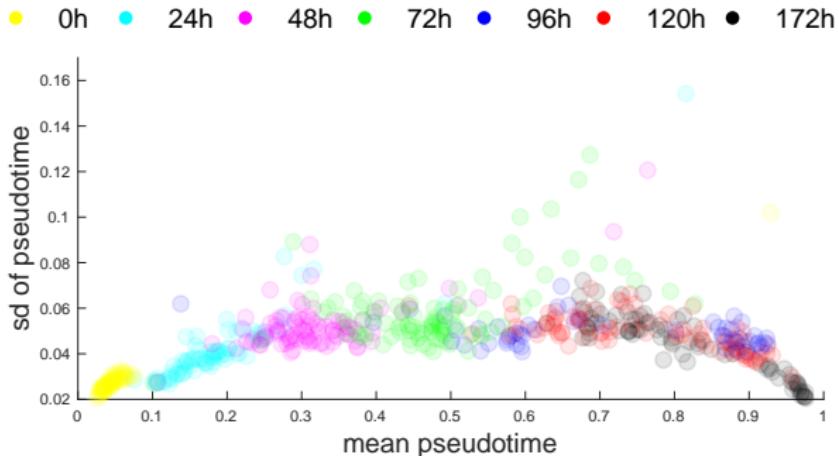
# Distribution of orders



- GPseudoRank
- TSCAN
- SLICER
- Mouse dendritic cell response to LPS stimulation, scRNA-seq, data: Shalek et al. (2014).
- 2-dimensional representation (MDS) of cell position vectors → modes of distribution of orders
- Methods without uncertainty Ji and Ji 2016; Welch et al. 2016

Strauß, Reid and Wernisch 2019

# Varying pseudotime uncertainty



**Figure:** Progression of mouse embryonic stem cells along the neuronal lineage  
 (data: Stumpf et al. (2017))

Strauß, Reid and Wernisch 2019

# Varying pseudotime uncertainty

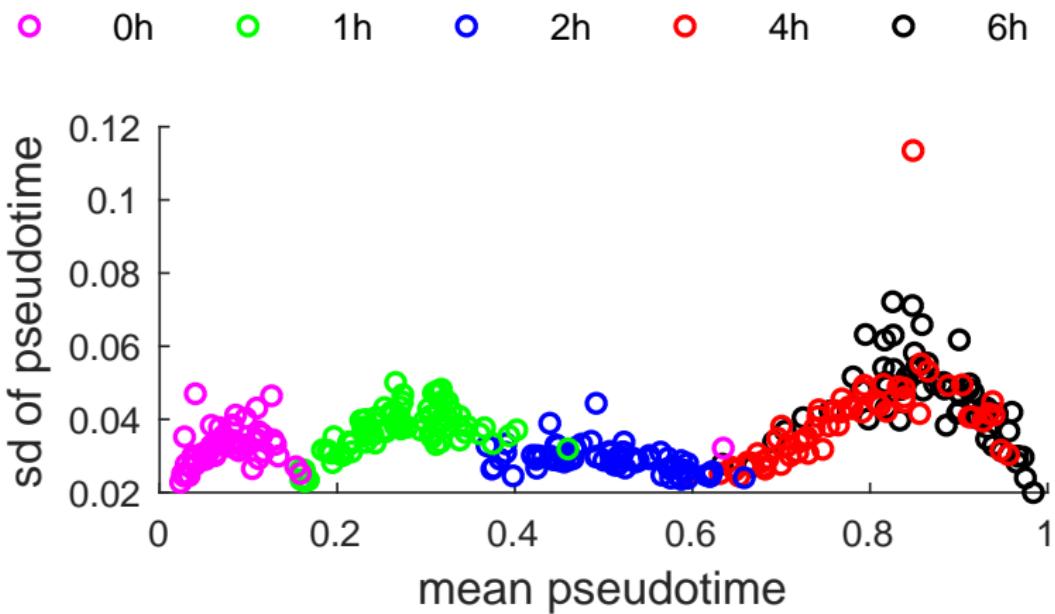


Figure: Data: Shalek et al. (2014). Capture times not required for initialisation.

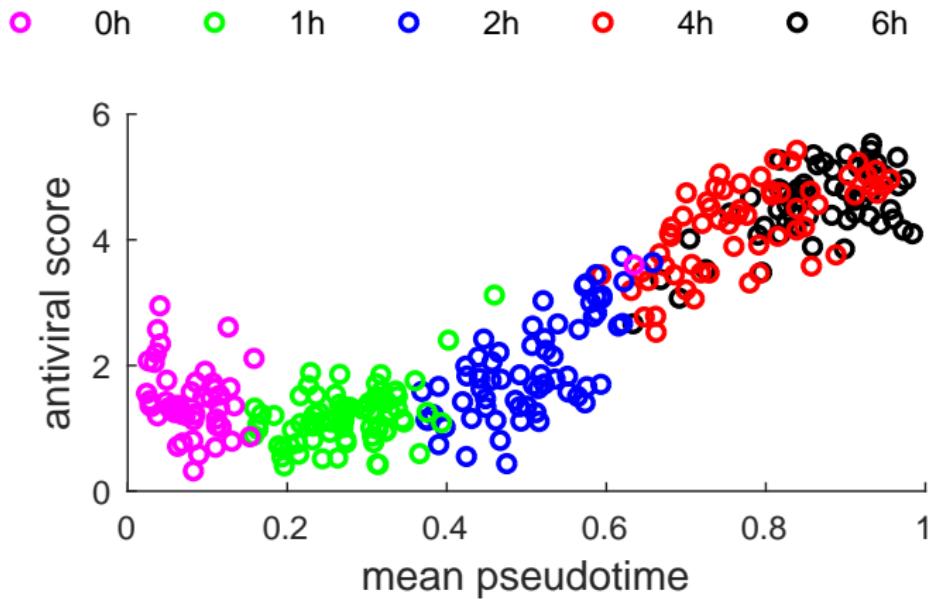
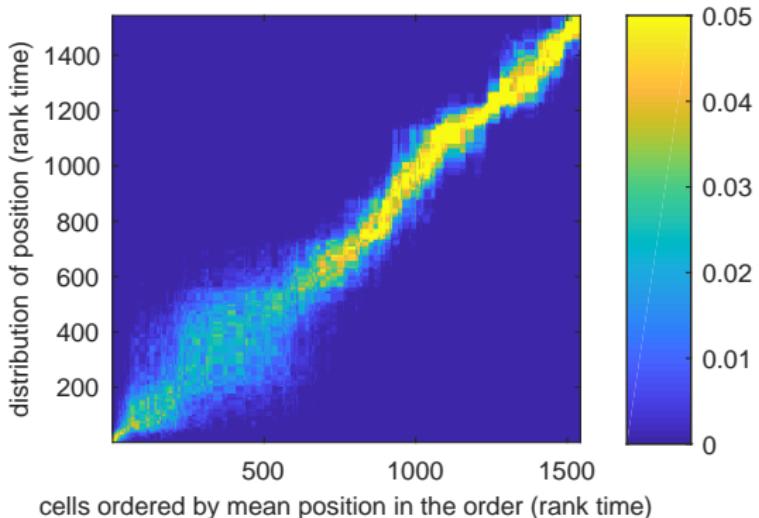


Figure: Data: Shalek et al. (2014)

# Metastable state

Group of cells that are essentially in the same developmental stage

Data: Klein et al. (2015),  
Mouse embryonic stem  
cells after leukemia  
inhibition factor (inhibits  
differentiation) withdrawal,  
droplet-based scRNA-seq  
data set, main branch



Strauß, Reid and Wernisch 2019

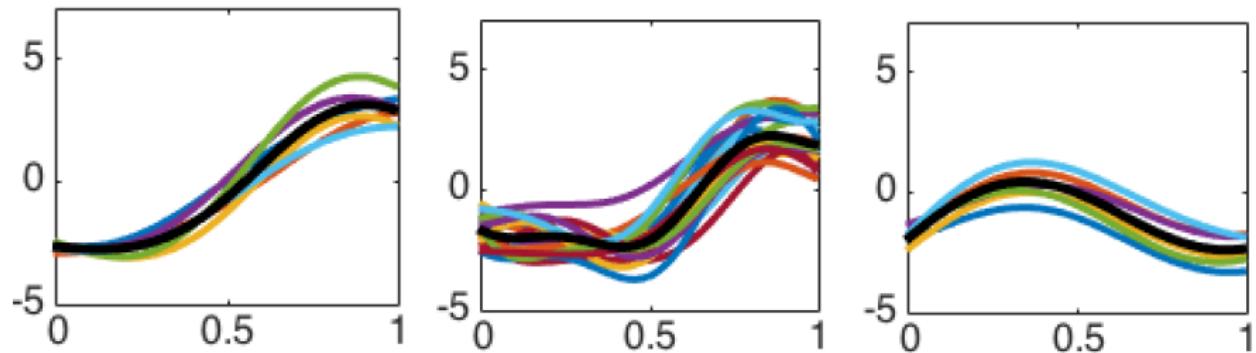
# Why GPs for pseudotime ordering I

- Have been well-tested to recapitulate known results (Ahmed, Rattray and Boukouvalas 2018; Reid and Wernisch 2016)
- Can be scaled up using approximation methods (sparse GPs as in Ahmed, Rattray and Boukouvalas 2018, small clusters of cells as in Strauß, Reid and Wernisch 2019)
- Bayesian methods can be used to understand measurement uncertainty and the consequences for how we interpret the data (Strauß, Reid and Wernisch 2019)

# Why GPs for pseudotime ordering II

- If uncertainty estimates are less important, efficient Variational methods allow efficient inference on huge datasets (Ahmed, Rattray and Boukouvalas 2018)
- Software:  
GrandPrix:<https://github.com/ManchesterBioinference/GrandPrix> for Ahmed, Rattray and Boukouvalas 2018, DeLorean R package for Reid and Wernisch 2016, and <https://github.com/magStra/GPseudoRank> for Strauß, Reid and Wernisch 2019

# GPs for clustering longitudinal data



# Deconvolution of shared profiles

Assume for a moment we knew the order of the cells



Figure: Several genes

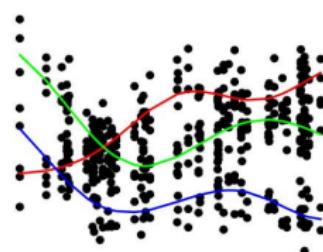


Figure: Shared trajectories

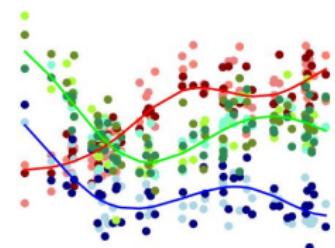
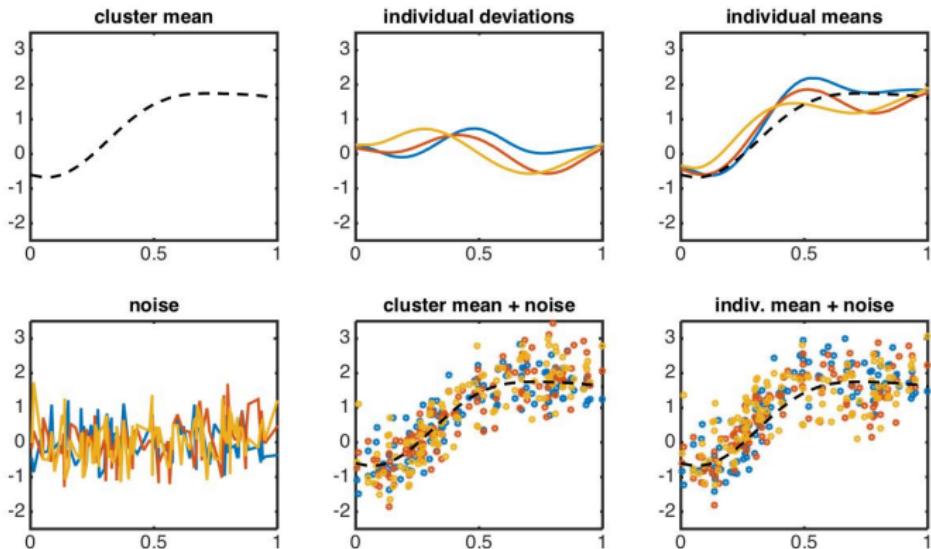


Figure: Allocation to trajectories

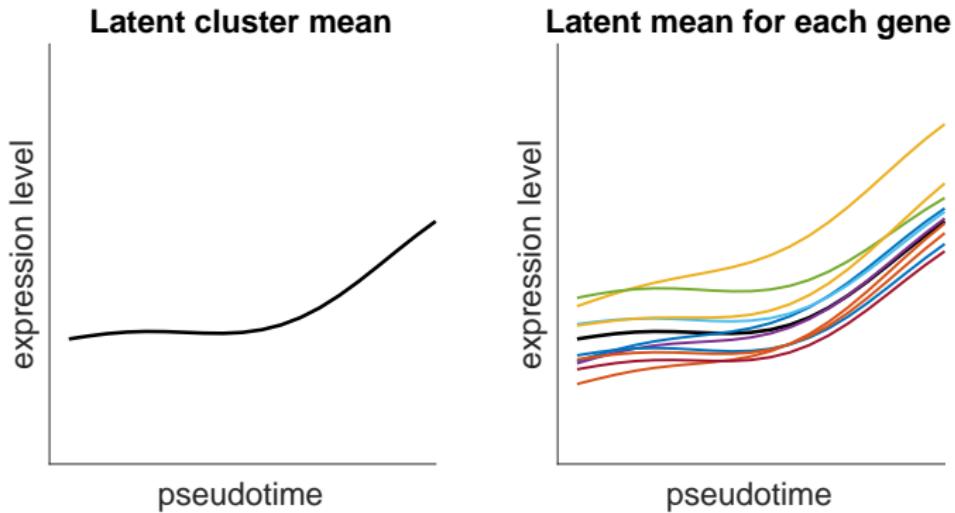
Figure: Mixture model for genomic time course. Cells are ordered horizontally in pseudotime, gene expressions vertically. Genes are allocated to clusters in terms of similarity of pseudotime profile.

# Structure of one cluster with given cell ordering



**Figure:** Large GP with block-matrix structure, developed efficient method for matrix decompositions

# Structure of one cluster with given cell ordering



# Gaussian process mixture models

## Mixtures of hierarchical Gaussian processes

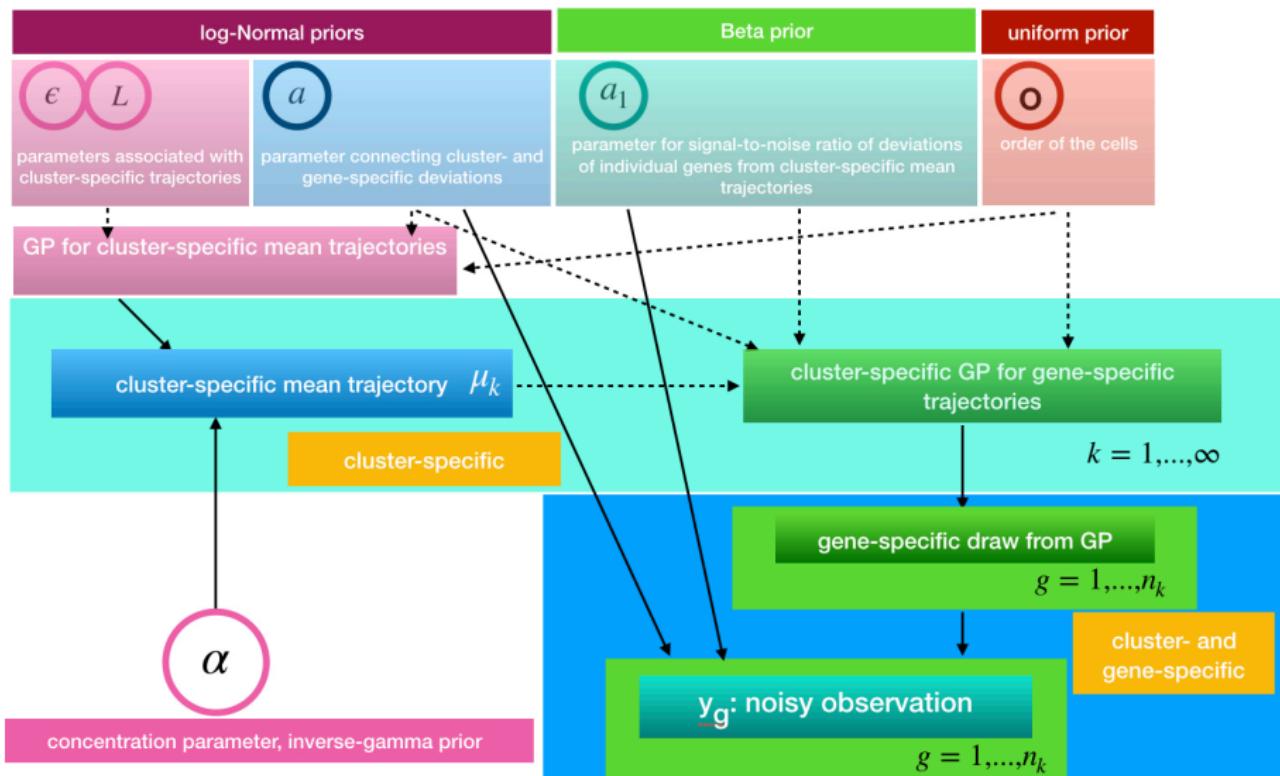
We use a mixture model:

$$p(\mathbf{y}) = \sum_{c=1}^C \pi_c f(\mathbf{y} | \mathbf{z}_c, \Sigma_c). \quad (1)$$

- $C$  is the number of mixture components
- $\pi_c$  are the mixture proportions
- $f$  is a Gaussian process
- $\mathbf{z}_c \sim GP(\mathbf{0}_T, \mathbf{K}_c)$
- $\mathbf{K}_c(t_i, t_j) = \sigma_{W,c}^2 \exp(-\frac{1}{2L_c^2}(t_i - t_j)^2)$
- $\Sigma_c(t_i, t_j) = \sigma_{W,c}^2 \exp(-\frac{1}{2L_c^2}(t_i - t_j)^2) + \delta_{ij}\sigma_{\epsilon,c}^2$

- Clusters genes in terms of their response or developmental **trajectories**: methods not looking at the trajectories cannot do this
- **Dirichlet process**: automatic determination of number of clusters
- **Bayesian** method with **MCMC** → uncertainty

# Dirichlet process mixture model of hierarchical GPs with pseudotime



# Dirichlet process mixture model of hierarchical GPs with pseudotime

## DP mixture model of GPs with pseudotime

$$\sigma_W^2 \sim \log\mathcal{N}(\cdot, \cdot) \quad L \sim \log\mathcal{N}(\cdot, \cdot)$$

$$H = GP(\cdot, \mathbf{K}(\sigma_W^2, L))$$

$$S_w = \log\mathcal{N}(\cdot, \cdot) \quad S_l = \log\mathcal{N}(\cdot, \cdot) \quad S_\epsilon = \log\mathcal{N}(\cdot, \cdot)$$

$$G_0 = H \otimes S_w \otimes S_l \otimes S_\epsilon$$

$$G|G_0 \sim DP(\alpha, G_0) \quad \mu_j, \sigma_{w,j}^2, l_j, \sigma_{\epsilon,j}^2 | G \sim G, \quad j = 1, \dots, n$$

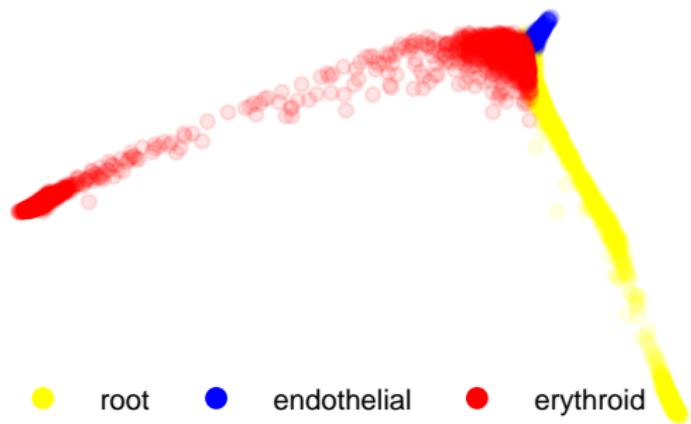
$$\alpha \sim \text{Gamma}(2, 4)$$

$$\mathbf{o} \sim \text{Uniform}$$

$$\mathbf{y}_j | \mathbf{o}, \mu_j, \sigma_{w,j}^2, l_j, \sigma_{\epsilon,j}^2 \sim GP(\mu_j, \Sigma_j(\sigma_{w,j}^2, l_j, \sigma_{\epsilon,j}^2))$$

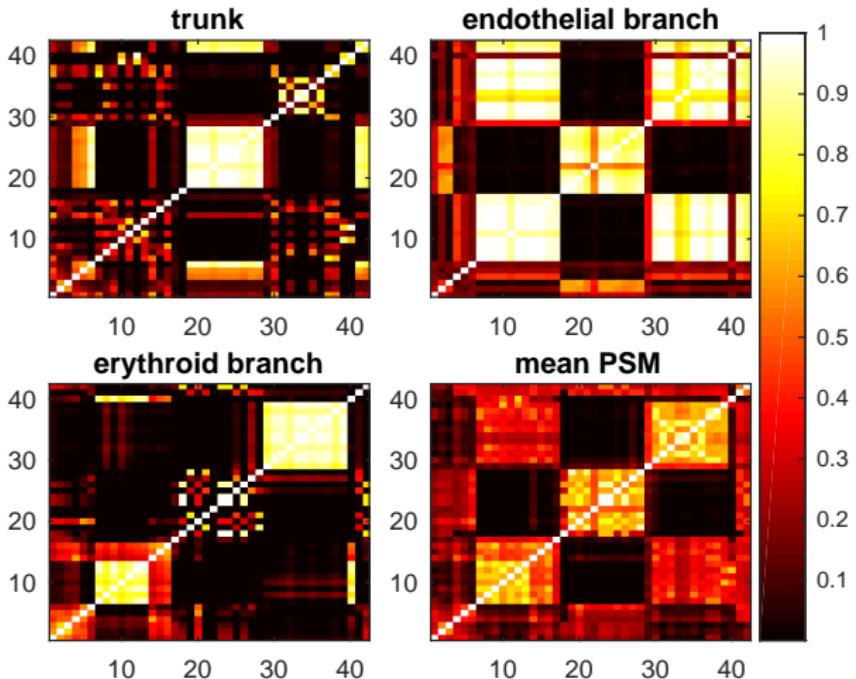
- Moignard et al. 2015 applied single-cell RT-qPCR to 3,934 mouse early hematopoietic cells in-vivo
- four time points between embryonic day 7.0 and 8.5.
- Moignard et al. 2015 and Haghverdi et al. 2016: diffusion maps are used to identify two branches, a blood and an endothelial branch
- We now identify and compare different clustering structures for genes for the different branches

# Different cluster structures in branching data II



Moignard data, branches inferred using diffusion maps

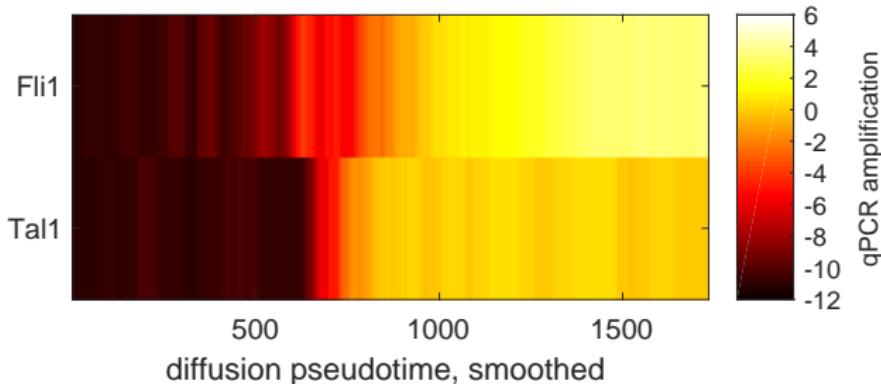
# Different cluster structures in branching data III



Strauss et al. 2020

Four groups of genes with a pairwise posterior co-clustering probability of more than 80% for the trunk:

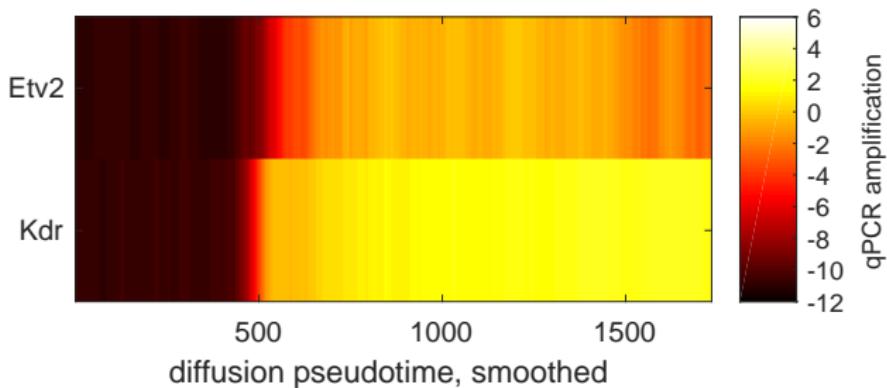
**Group 1:** *Fli1*, *Tal1*: These two transcription factors are switched on at a very similar pseudotime. The main role of *Tal1* is to induce a blood programme Scaldone. *Fli1* is also essential for maintaining hematopoiesis Badwe.



**Figure:** Moignard data, group 1 in trunk. Diffusion pseudotime was used for the ordering of the cells. Gene expression is smoothed by averaging over 50 cells.

**Group 2:** *Etv2*, *Kdr*: These two genes (see Figure 10) are switched on only slightly earlier in pseudotime compared to *Fli1* and *Tal1*. The posterior pairwise co-clustering probabilities with the genes in group 1 are therefore also relatively high: *Etv2-Fli1*: 0.65, *Etv2-Tal1*: 0.72, *Kdr-Fli1*: 0.62, *Kdr-Tal1*: 0.69. *Etv2* expression is required for the initiation of the hematopoietic program Wareing. *Kdr* is a well-known endothelial marker.

# Different cluster structures in branching data VI

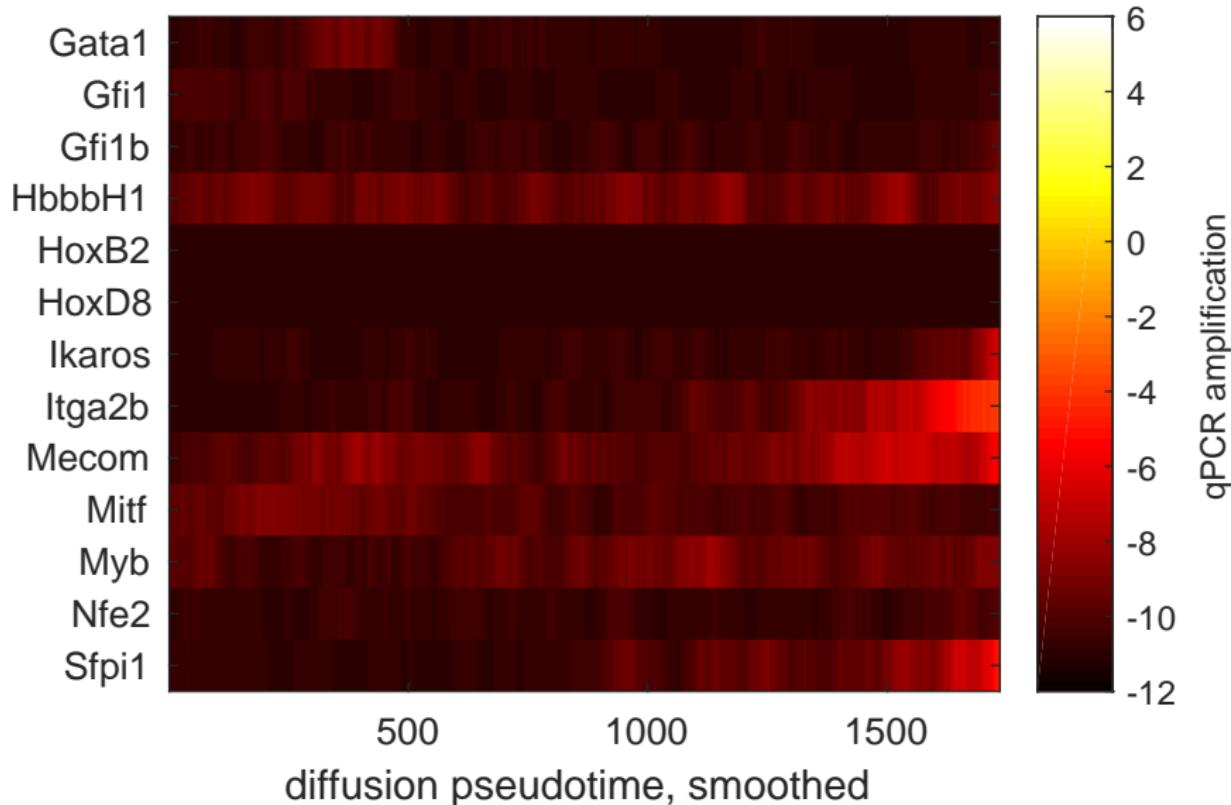


**Figure:** Moignard data, group 2 in trunk. Diffusion pseudotime was used for the ordering of the cells. Gene expression is smoothed by averaging over 50 cells.

This underlines the advantage of a method like GPseudoClust that incorporates uncertainty of pseudotime ordering into the clustering.

**Group 3:** *Gata1*, *Gfi1*, *Gfi1b*, *Hbbbh1*, *HoxB2*, *HoxD8*, *Ikaros*, *Itga2b*, *Mecom*, *Mitf*, *Myb*, *Nfe2*, *Sfpi1*: All of these genes have very low expression levels or are not observed.

# Different cluster structures in branching data VII

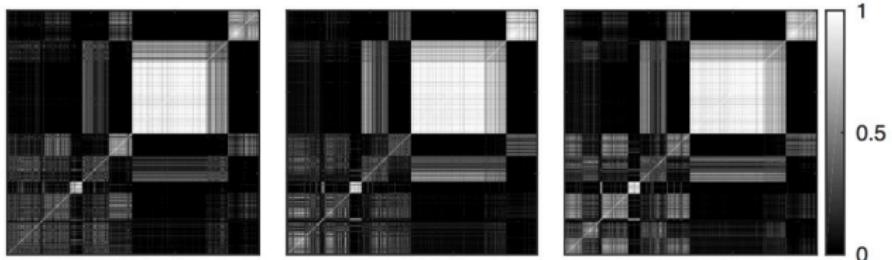


# Summary

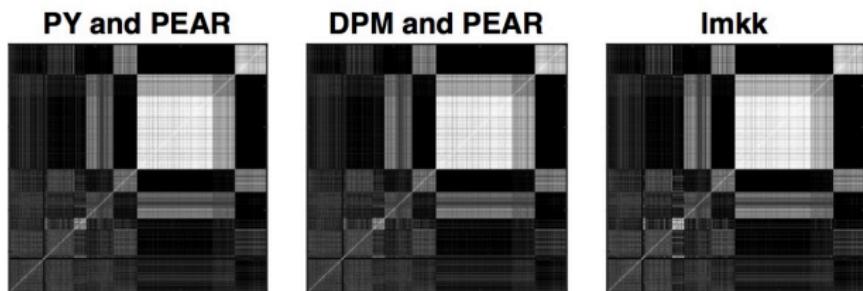
- Latent factor models and Gaussian process latent variable methods allow us to understand underlying factors driving variation in the data such as sets of important regulatory genes or pseudotime
- Gaussian processes also allow the clustering of data in terms of similarity of response or developmental profile
- Bayesian inference for Gaussian processes allows more understanding of the data by modelling the noise-related uncertainty of the inference
- Efficient methods have been developed to apply GPs to large single-cell datasets

- We take much fewer cells and repeat the method many times in parallel using the restricted number of cells
- From a certain number of points onwards the GP is well identified and we do not need to add points
- The method uses Gaussian covariance matrices, but is able to infer clusters simulated using Mat'ern and linear covariance matrices
- The clusters are fairly robust to considerable levels of dropout

# Efficient inference - Robustness of GPs II



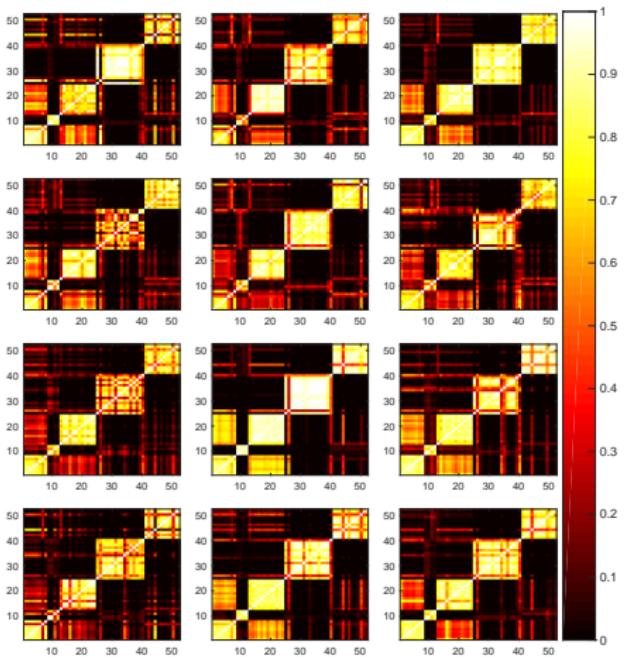
(a) GPpseudoClust without subsampling: PSMs for four chains.



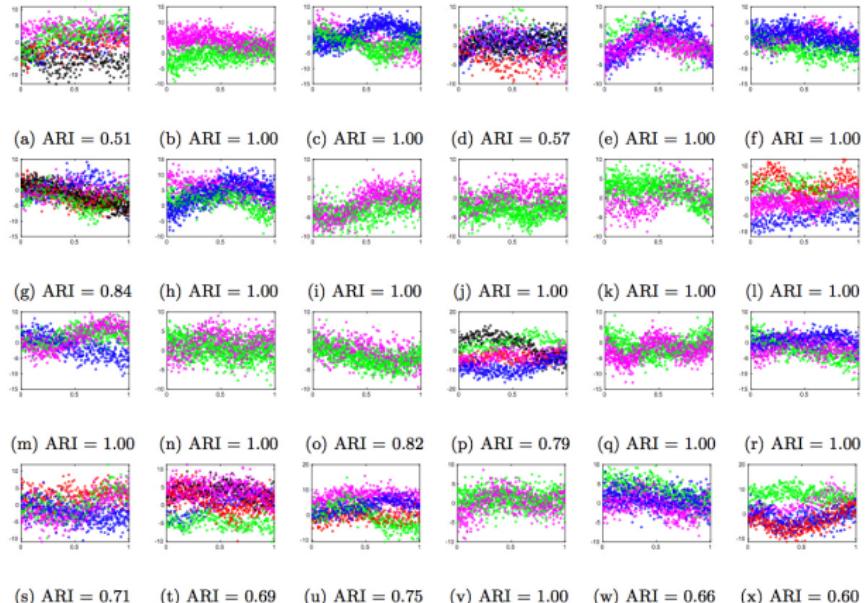
(b) GPpseudoClust with subsampling: 36 subsampled chains with 15 cells each.

Strauss et al. 2020

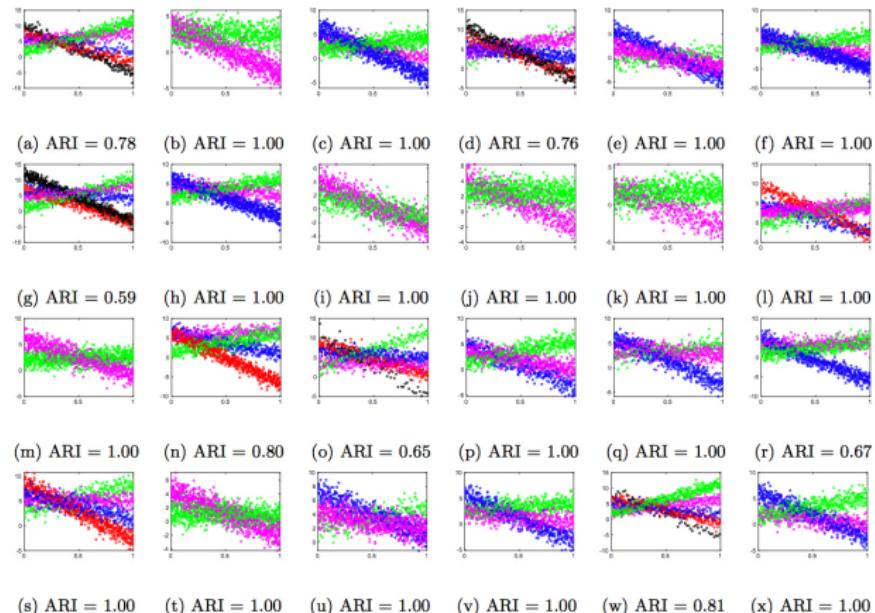
# Efficient inference - Robustness of GPs III



# Efficient inference - Robustness of GPs IV



# Efficient inference - Robustness of GPs V



# Efficient inference - Robustness of GPs VI

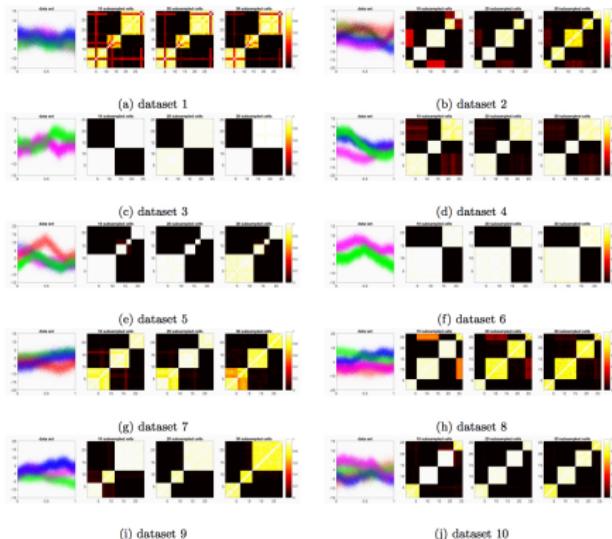


Figure S16: Simulated datasets with 9000 cells. The data were simulated using hierarchical GPs with Matérn-3/2 covariance matrices. Each subfigure illustrates one simulated dataset and summary PSMs computed using the 'PY+PEAR' method for 10, 20, and 30 subsampled cells per capture time.

Strauss et al. 2020, supplementary materials

# References I

-  Ahmed, S, M Rattray and A Boukouvalas (2018). "GrandPrix: Scaling up the Bayesian GPLVM for single-cell data". In: *Bioinformatics* 35.1, pp. 47–54. DOI: [10.1093/bioinformatics/bty533](https://doi.org/10.1093/bioinformatics/bty533).
-  Argelaguet, Ricard et al. (2018). "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets". In: *Molecular Systems Biology* 14.6, e8124.
-  Argelaguet, Ricard et al. (2020). "MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome Biology* 21.1, p. 111.
-  Buettner, Florian et al. (2017). "f-scLVM: Scalable and versatile factor analysis for single-cell RNA-seq". In: *Genome Biology* 18.1, p. 212.

## References II

-  Campbell, KR and C Yau (2016). "Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference". In: *PLOS Comput Biol* 12.11, e1005212.
-  Dietrich, Sascha et al. (2018). "Drug-perturbation-based stratification of blood cancer". In: *Journal of Clinical Investigation* 128.1, pp. 427–445.
-  Haghverdi, L et al. (2016). "Diffusion pseudotime robustly reconstructs lineage branching". In: *Nat Meth* 13.10, pp. 845–848. DOI: [10.1038/nmeth.3971](https://doi.org/10.1038/nmeth.3971).
-  Hamey, Fiona K. et al. (2017). "Reconstructing blood stem cell regulatory network models from single-cell molecular profiles". In: *PNAS* 114 (23), pp. 5822–5829.
-  Ji, Z and H Ji (2016). "TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis". In: *Nucleic Acids Res* 44.13, e117–e117.

## References III

-  Klein, AM et al. (2015). "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells". In: *Cell* 161.5, pp. 1187–1201. DOI: [10.1016/j.cell.2015.04.044](https://doi.org/10.1016/j.cell.2015.04.044).
-  Moignard, Victoria et al. (2015). "Decoding the regulatory network of early blood development from single-cell gene expression measurements". In: *Nat Biotechnol* 33, 269 EP –. DOI: [dx.doi.org/10.1038/nbt.3154](http://dx.doi.org/10.1038/nbt.3154).
-  Murphy, KP (2012). *Machine learning: a probabilistic perspective*. Cambridge, MA: The MIT Press.
-  Reid, JE and L Wernisch (2016). "Pseudotime estimation: deconfounding single cell time series". In: *Bioinformatics* 32.19, pp. 2973–2980. DOI: [10.1093/bioinformatics/btw372](https://doi.org/10.1093/bioinformatics/btw372). URL: [+http://dx.doi.org/10.1093/bioinformatics/btw372](http://dx.doi.org/10.1093/bioinformatics/btw372).
-  Shalek, AK et al. (2014). "Single-cell RNA-seq reveals dynamic paracrine control of cellular variation". In: *Nature* 510, pp. 363–369.

## References IV

-  Strauss, Magdalena E et al. (2020). "GPpseudoClust: deconvolution of shared pseudo-profiles at single-cell resolution". In: *Bioinformatics* 36.5, pp. 1484–1491.
-  Strauß, Magdalena E, John E Reid and Lorenz Wernisch (2019). "GPpseudoRank: a permutation sampler for single cell orderings". In: *Bioinformatics* 35.4, pp. 611–618.
-  Stumpf, PS et al. (2017). "Stem Cell Differentiation as a Non-Markov Stochastic Process". In: *Cell Systems* 5.3, 268–282.e7.
-  Velten, Britta and Ricard Argelaguet (2020). *MOFA: Multi-Omics Factor Analysis (MOFA)*. R package version 1.4.0.
-  Virtanen, Seppo et al. (2012). "Bayesian Group Factor Analysis". In: ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. *Proceedings of Machine Learning Research*. La Palma, Canary Islands: PMLR, pp. 1269–1277. URL:  
<http://proceedings.mlr.press/v22/virtanen12.html>.



- Welch, JD et al. (2016). "SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data". In: *Genome Biol* 17.1, p. 106.

Work on Gaussian process modelling for pseudotime inference and clustering was joint with Lorenz Wernisch, John Reid and Paul Kirk as stated in the references, and supported by the Medical Research Council.

