Applications of RNA Velocity and Gaussian Processes for scRNA-seq Data

Author: Inna Ivanova
StudentID: 2267477i

Contents

- Single Cell RNA sequencing
- RNA Velocity
- Gaussian Processes
- Experiments
- Results
- Conclusion

Single Cell RNA Sequencing

- A technology that provides the expression profiles of individual cells
- In-depth sequencing of individual mRNA molecules
  - Analysis of the gene expression
  - Cell lineage
  - Gene profiling
- Different protocols for sequencing: well and drop based
  - Smart-seq2, 10X Genomics, inDrop
  - Different sequencing depth

How Single Cell RNA Sequencing is done?

- Isolate cells of interest

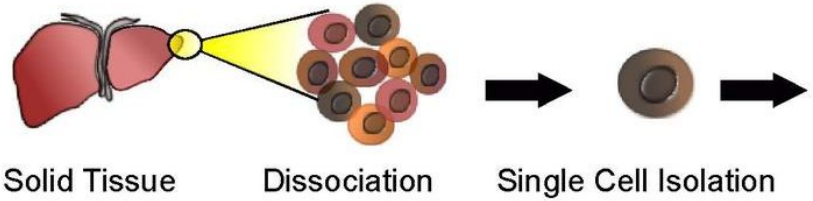


Image adapted from Wikipedia.

How Single Cell RNA Sequencing is done?

- Extract individual RNA molecules, reverse transcriptomics to construct cDNA

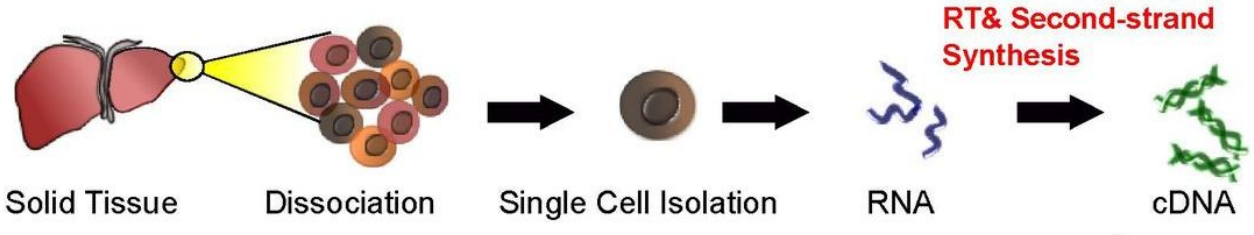


Image adapted from Wikipedia.

How Single Cell RNA Sequencing is done?

- PCR Amplification on the cDNA followed by sequencing and gene alignment

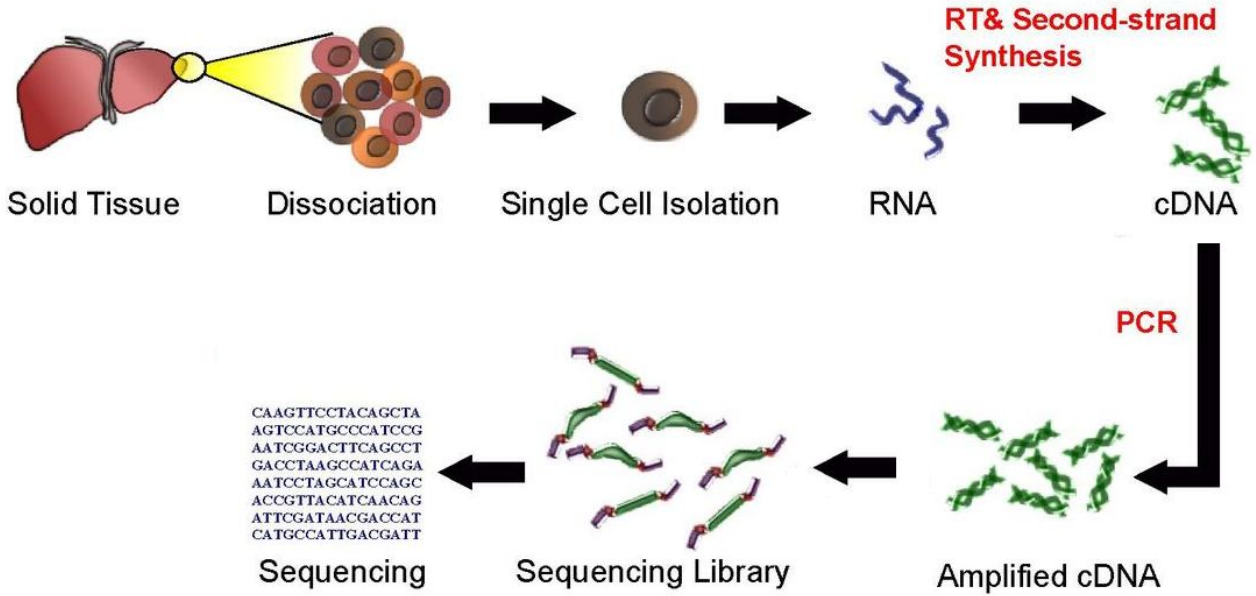


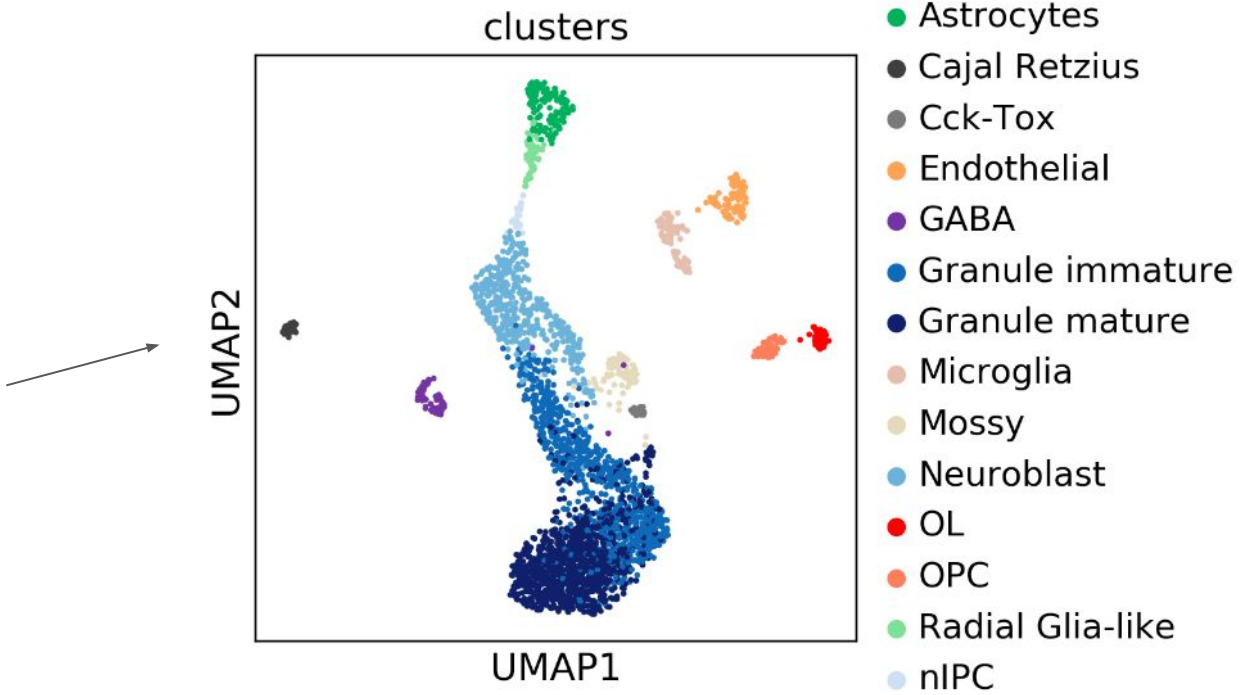
Image adapted from Wikipedia.

The result from scRNAseq: Count Matrix

index Cell_ID	Lypla1	Tcea1	Atp6v1h	Rb1cc1	St18	Pcmtd1	Rrs1	Adhfe1	Sgk3	Mcmdc2	...	Tlr7	Prps2
0	0.0	0.0	0.0	2.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
5	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
6	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

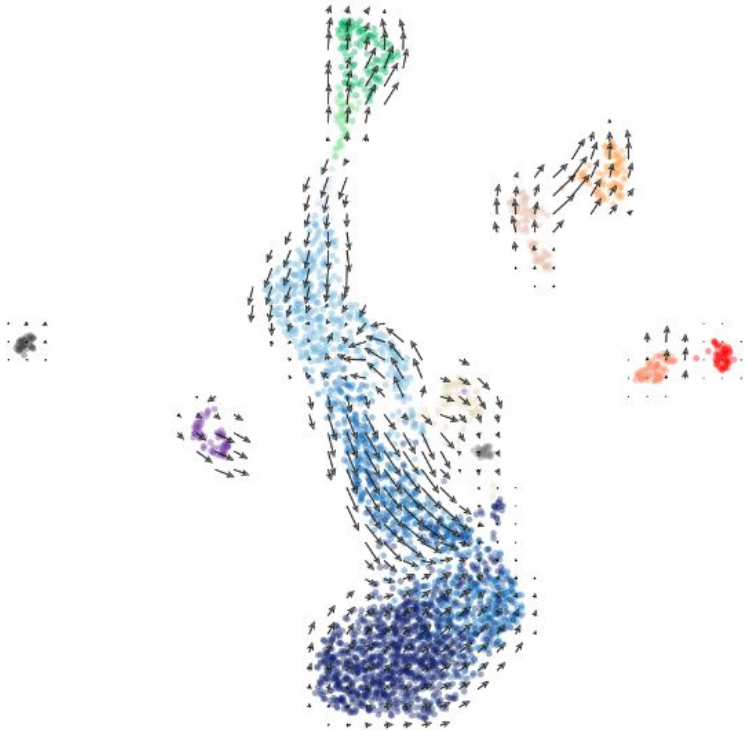
scRNA-seq Count Matrix Processing

index Cell_ID	Lypla1	Tcea1	Atp6v1h	Rb1cc1	St18	Pcmtd1
0	0.0	0.0	0.0	2.0	0.0	1.0
1	0.0	1.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	1.0
4	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	1.0	0.0	0.0
6	0.0	0.0	0.0	1.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0



RNA Velocity

- High dimensional vector
- Obtained from the differential equation describing the dynamics of spliced (mature) and unspliced (immature) RNA cell abundance
- Can be mapped into the low dimensional mapping
- Allows future predictions of the cell differential pathways

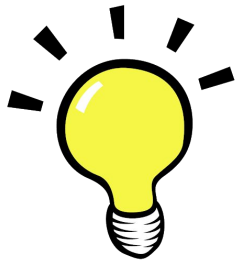


Problem Scope

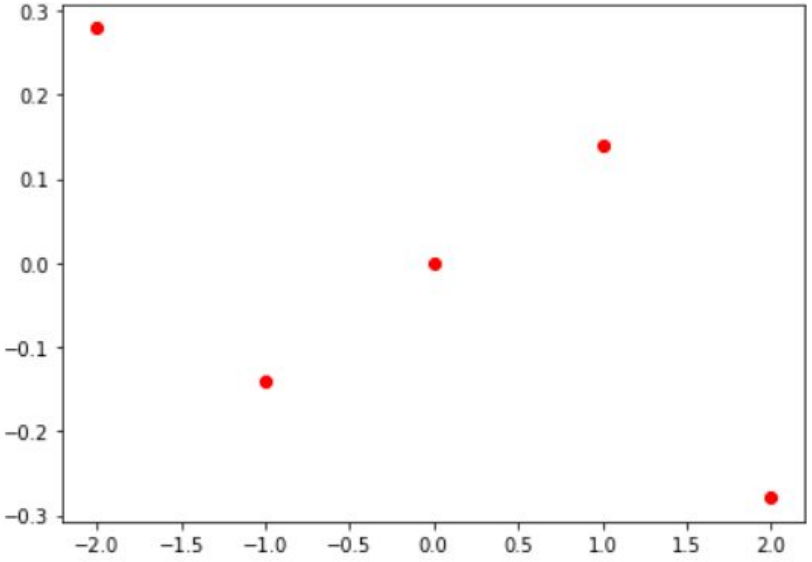
- A way to identify smoothly varying genes in the low dimension space
  - They could potentially identify interesting properties of the biological system
  - Unravel hidden structures
  - EL Low (2019) suggests that smoothly varying genes have been identified as possibly relating to consequence of differentiating events
- A way to rank genes by their smoothness
- Do this in an unsupervized way

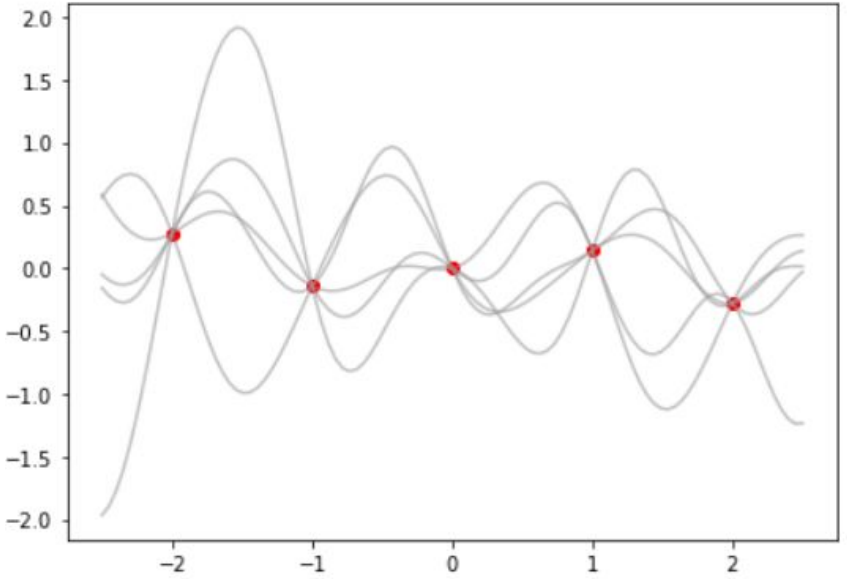
Gaussian Process

- Models a system of observations
- Distribution over all possible functions constrained at specific points
  - Bayesian approach
  - The mean is the prediction
  - Analytical approach of obtaining the mean and the covariance matrix
- Analytically computing the marginal likelihood
  - Used as a criteria for how good a fit the data is to the model
  - Used in the optimization of the hyperparameters
- Different kernel functions defining the covariance matrix
  - Describe different smoothness

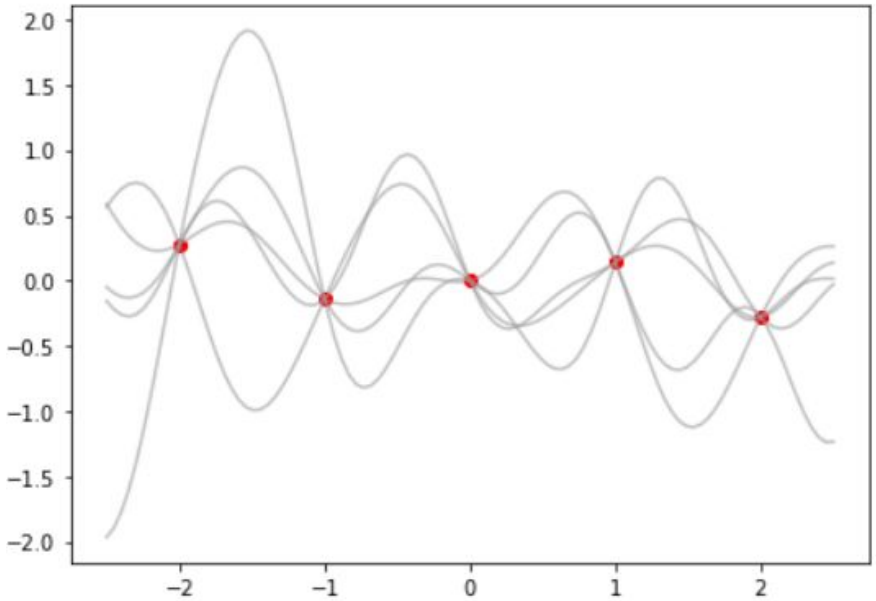


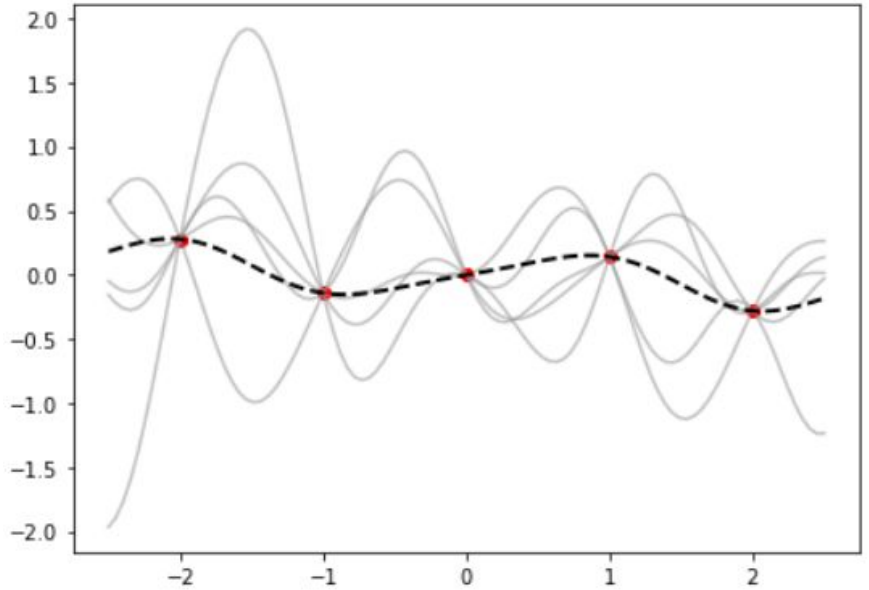
Gaussian Process



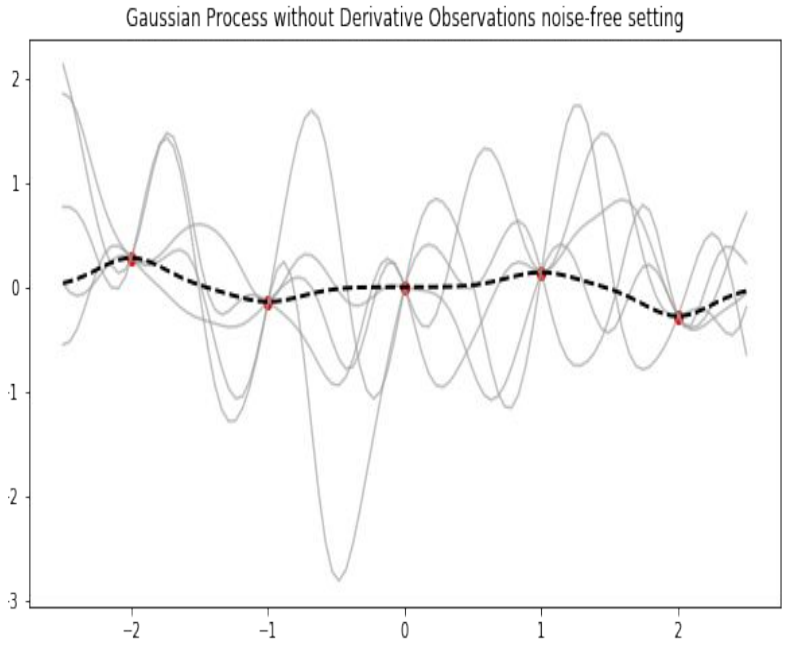


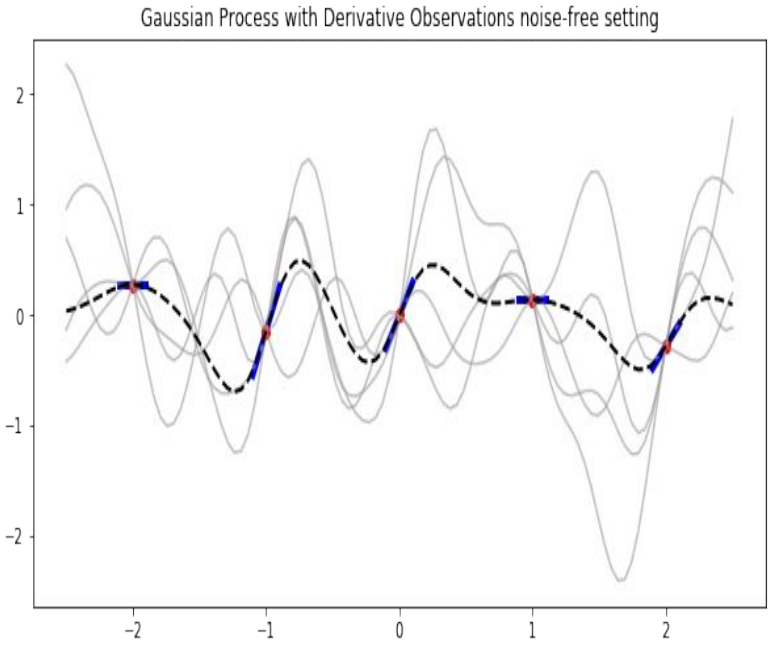
Gaussian Process





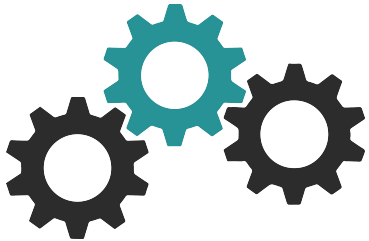
Gaussian Process with Derivative Observations





Experiment Set-up

1. Use the low dimensional mapping of the count matrix (via t-SNE, UMAP, PCA)
2. Map the resulting points and their gene expression levels for each gene to the GP model with and without RNA Velocity
3. Calculate the marginal log likelihood of each gene
4. Rank the genes by their obtained likelihood (smoothness)
5. Examine gene expression functions, rankings and mean squared errors obtained by the two models



Results

Baseline Gaussian Process Model			
Smoothest Genes		Least Smooth Genes	
1.*CALD1*	2.*TNFRSF11B1*	1.*KLHL4*	2.*LDLRAD4*
3.*SERPINE1*	4.*CCDC80*	3.*RAB27B*	4.*LINC01139*
5.*PXDC1*	6.*DDAH1*	5.*DNM1*	6.*SEPT6*
7.*WWTR1*	8.*EXT1*	7.*KRT8*	8.*DEPTOR*
9.*FGF2*	10.*DLC1*	9.*AC083967.1*	10.*RTKN2*

Table 6.1: *Top 10 smoothest and least smooth genes after running the Baseline Gaussian Model on 40% of the Human Saphenous Vein dataset.*

Gaussian Process Model with RNA Velocity Observations			
Smoothest Genes		Least Smooth Genes	
1.*CALD1*	2.*SERPINE1*	1.*SEMA3D*	2.*AL355607.2*
3.*DDAH1*	4.*TNFRSF11B*	3.*VIPR1*	4.*LINC01239*
5.*CCDC80*	6.*FGF2*	5.*HAPLN1*	6.*DRAXIN*
7.*PXDC1*	8.*FST*	7.*PKIB*	8.*EDN1*
9.*DLC1*	10.*EXT1*	9.*PRG4*	10.*SNTB1*

Table 6.3: *Top 10 smoothest and least smooth genes after running the Gaussian Model with RNA Velocity Observations on 20% of the Human Saphenous Vein dataset.*

Results

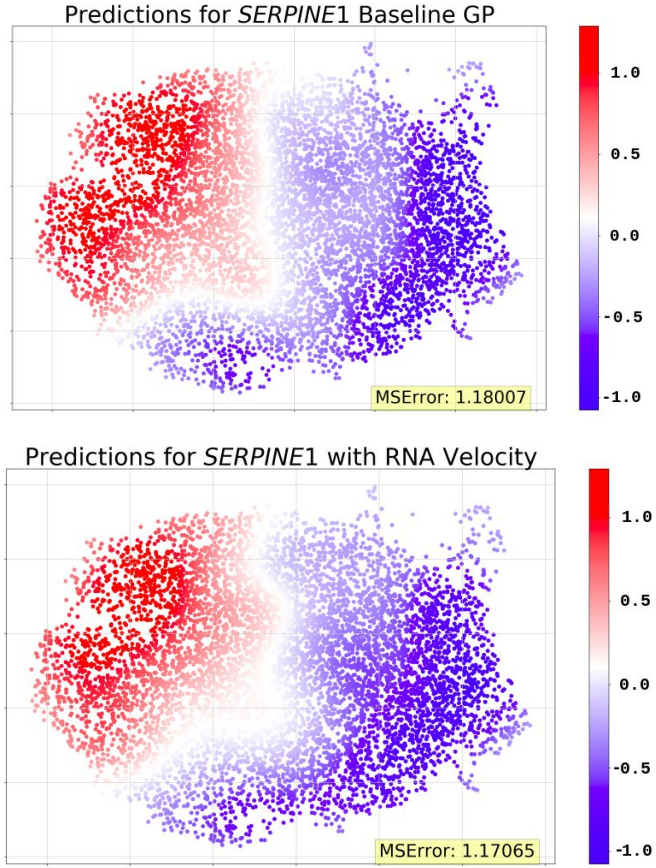
Baseline Gaussian Process Model			
Smoothest Genes		Least Smooth Genes	
1.*CALD1*	2.*TNFRSF11B1*	1.*KLHL4*	2.*LDLRAD4*
3.*SERPINE1*	4.*CCDC80*	3.*RAB27B*	4.*LINC01139*
5.*PXDC1*	6.*DDAH1*	5.*DNM1*	6.*SEPT6*
7.*WWTR1*	8.*EXT1*	7.*KRT8*	8.*DEPTOR*
9.*FGF2*	10.*DLC1*	9.*AC083967.1*	10.*RTKN2*

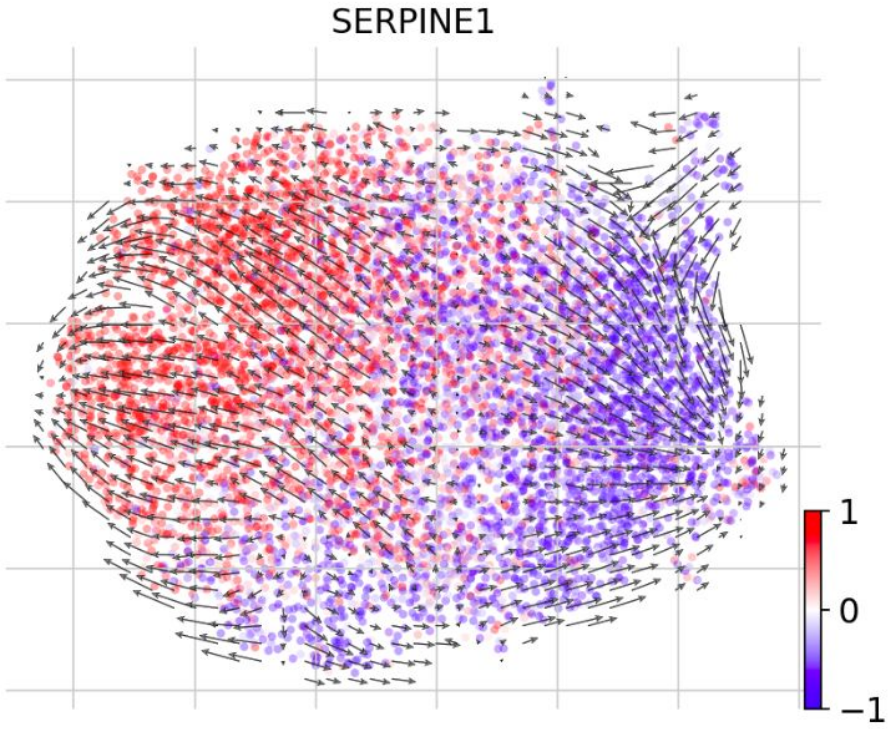
Table 6.1: *Top 10 smoothest and least smooth genes after running the Baseline Gaussian Model on 40% of the Human Saphenous Vein dataset.*

Gaussian Process Model with RNA Velocity Observations			
Smoothest Genes		Least Smooth Genes	
1.*CALD1*	2.*SERPINE1*	1.*SEMA3D*	2.*AL355607.2*
3.*DDAH1*	4.*TNFRSF11B*	3.*VIPR1*	4.*LINC01239*
5.*CCDC80*	6.*FGF2*	5.*HAPLN1*	6.*DRAXIN*
7.*PXDC1*	8.*FST*	7.*PKIB*	8.*EDN1*
9.*DLC1*	10.*EXT1*	9.*PRG4*	10.*SNTB1*

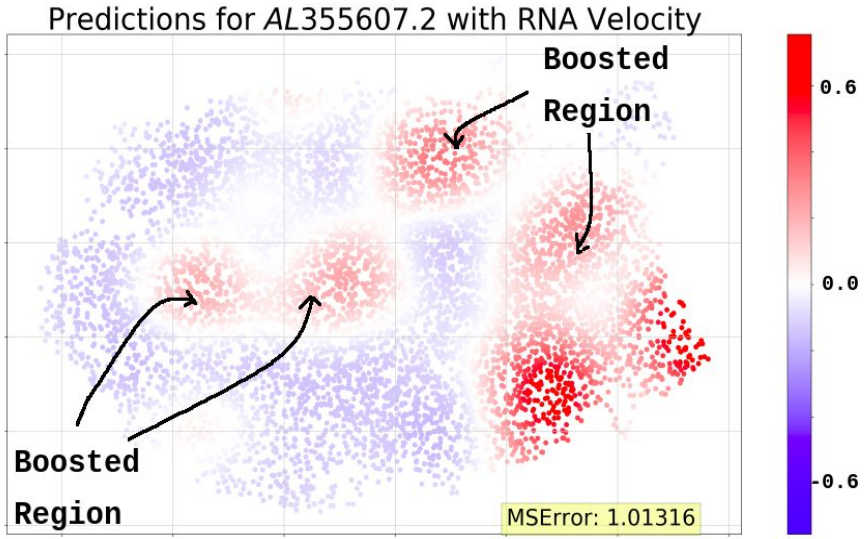
Table 6.3: *Top 10 smoothest and least smooth genes after running the Gaussian Model with RNA Velocity Observations on 20% of the Human Saphenous Vein dataset.*

Gene Expression Function

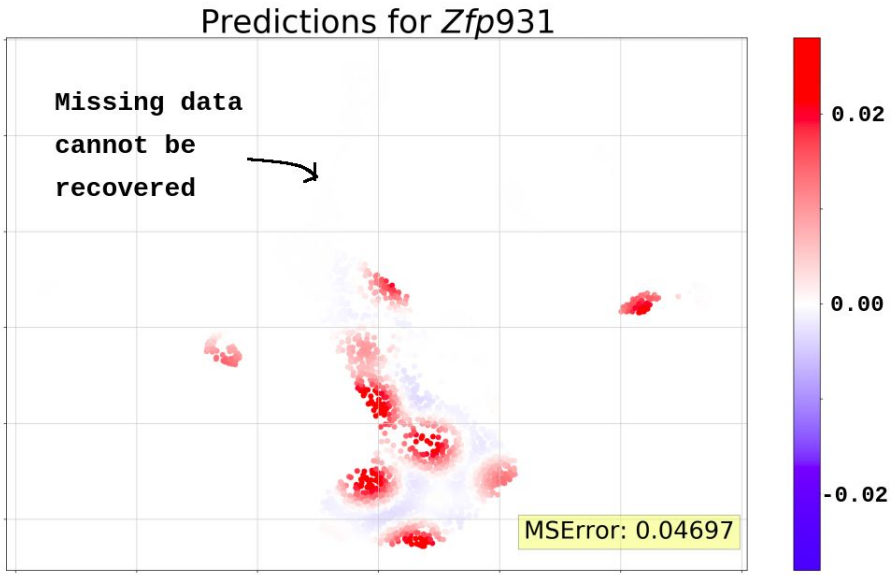


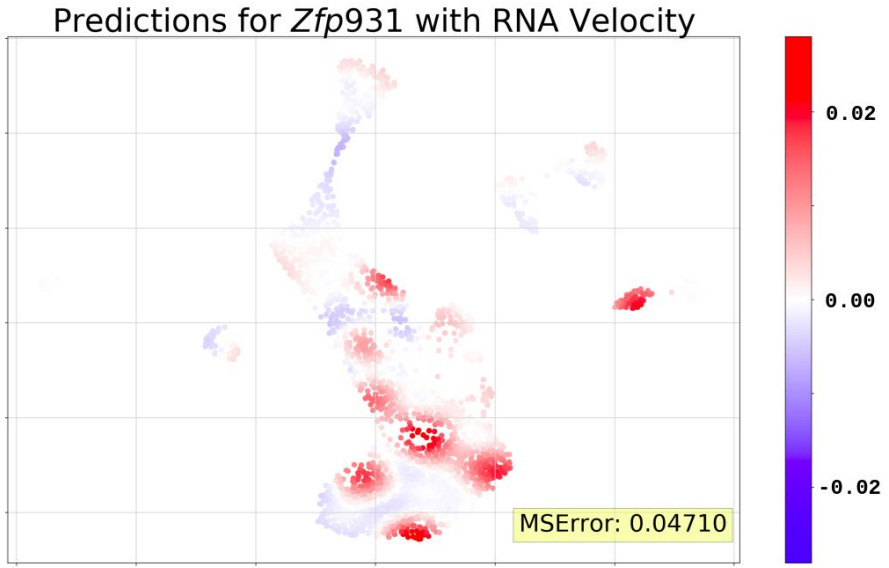


Effect of adding RNA Velocity



Effect of adding RNA Velocity on Very Sparse Data





Conclusion

- Gaussian Processes are very good method for identifying smooth genes
  - Fast computations
  - Allows only a subset of the datapoints to be used
  - Various kernel functions that allow different smoothness levels
  - Analytically obtained marginal likelihood which allows
    - Unsupervized ranking of genes
    - Fast optimization of hyperparameters
- Adding RNA Velocity results in generally better results
  - RNA velocity needs a scaling constant for some genes
  - Potentially, can be used as a reconstruction technique
  - Nevertheless, it can add noise where data is missing
- Further work
  - Better initial assumptions for the distribution of data
  - Scaling RNA Velocity constant