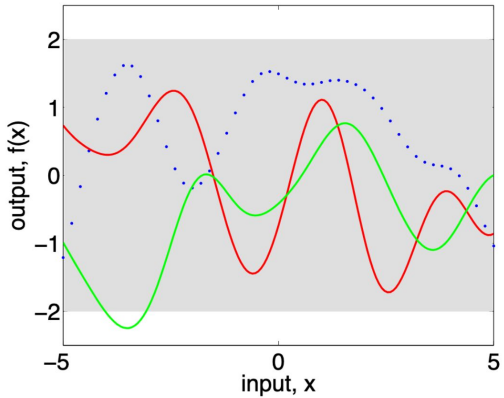
Gaussian Processes

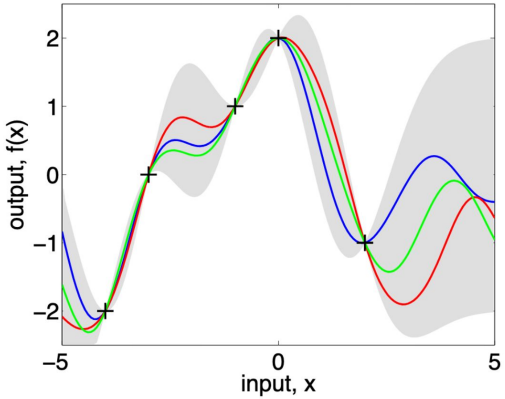
Recap Friday
12th March 2021

Introduction

- Most common application: Regression (Intuition)
- Idea of GPs: parametrize the space of all possible functions
- Bayesian/probabilistic approach to regression



(a), prior



(b), posterior

Recap: Multivariate Gaussian

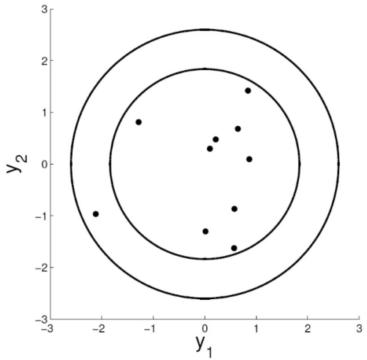
$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma)$$

mean (n-dim vector, for simplicity often =0 today)

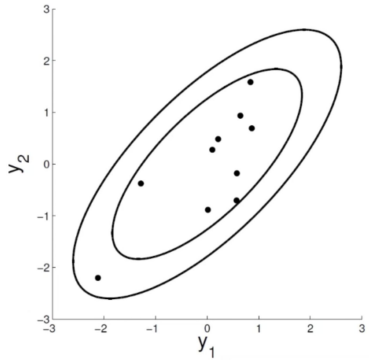
Covariance (nxn matrix, symmetric & positive definite)

$$\Sigma = \mathrm{Cov}(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)^T\right]$$

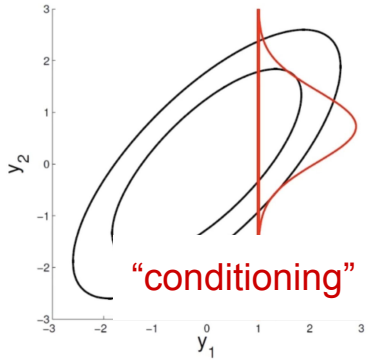
$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\intercal\Sigma^{-1}\mathbf{y}\right)$ $\quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\intercal\Sigma^{-1}\mathbf{y}\right)$ $\quad \Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$



$p(\mathrm{y}_2|\mathrm{y}_1, \Sigma) \propto \exp\left(-\frac{1}{2}(\mathrm{y}_2 - \mu_*)\Sigma_*{}^{-1}(\mathrm{y}_2 - \mu_*)\right)$



Theory

Video Richard Turner (Imperial)

https://www.youtube.com/watch?v=92-98SYOdlY

- Intro to high-dim Gaussians and GPs at 08:30-16:00 & 19:00-24:00
- How to predict with GPs at 36:31-40:00

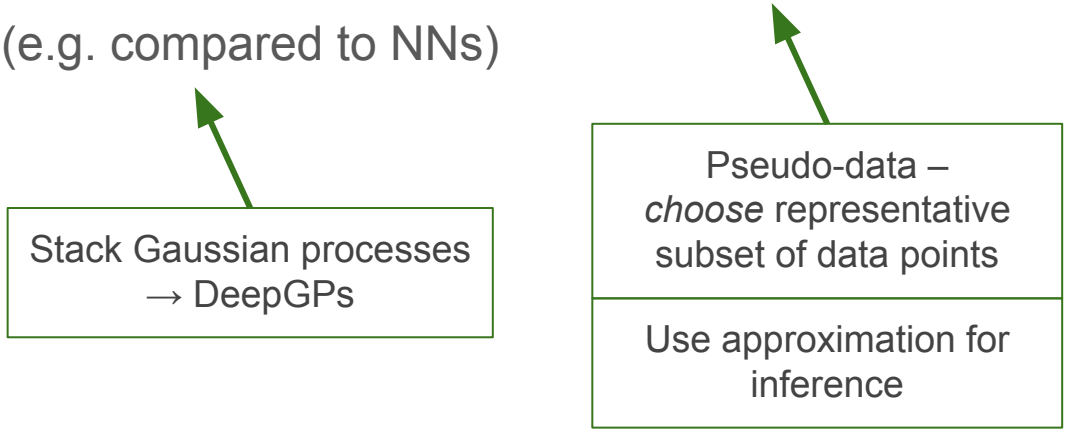
Key points

(+) Non-parametric model (capacity is not bound by #parameters)

(+) Probabilistic interpretation (e.g. predictive uncertainty)

(–) Data inefficient (need to compute matrix inverse over all data points)

(–) Not hierarchical (e.g. compared to NNs)



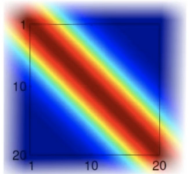
Different kernel functions

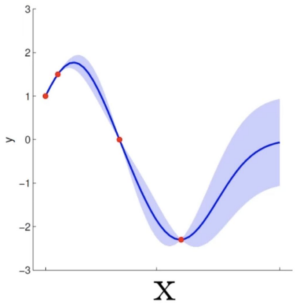
$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$

$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$

RDF kernel

$\Sigma =$





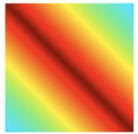
$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2}|x_1 - x_2|\right)$

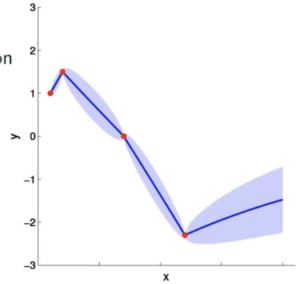
Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$

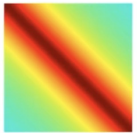


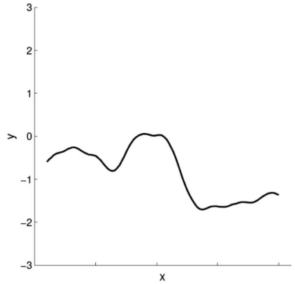


$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2}|x_1 - x_2|\right)^{-\alpha}$

Rational Quadratic

$\Sigma =$



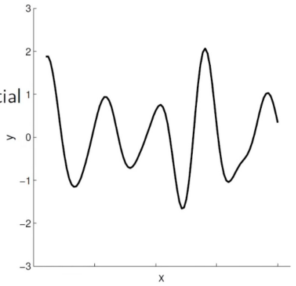


$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$

Periodic

sinusoid × squared exponential

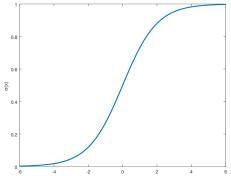
$\Sigma =$



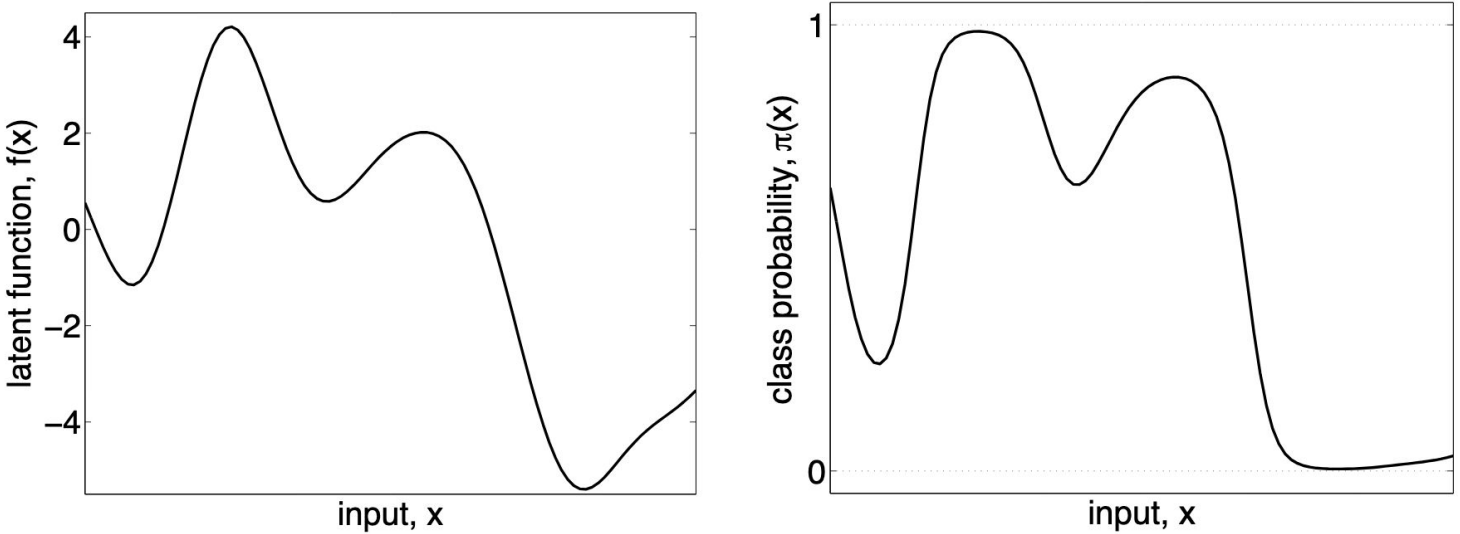
- Other kernels: e.g. linear kernel
- We can combine different kernels, e.g. by multiplication

Applications

1. Regression
2. Classification

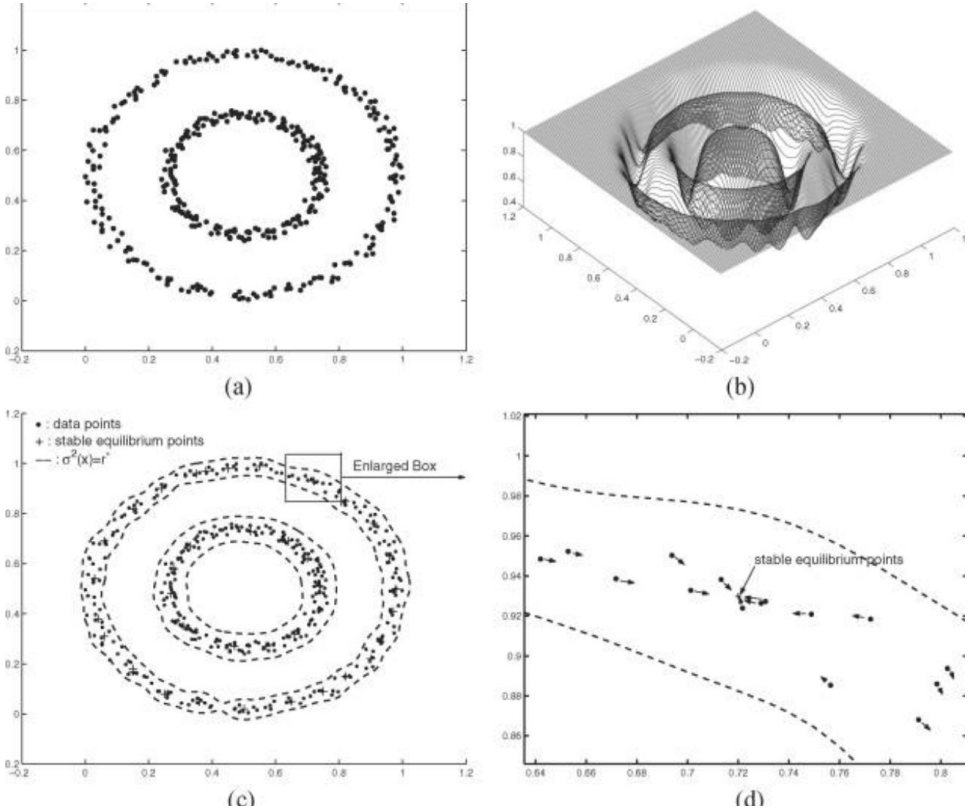


Logistic function



Applications

1. Regression
2. Classification
3. Clustering

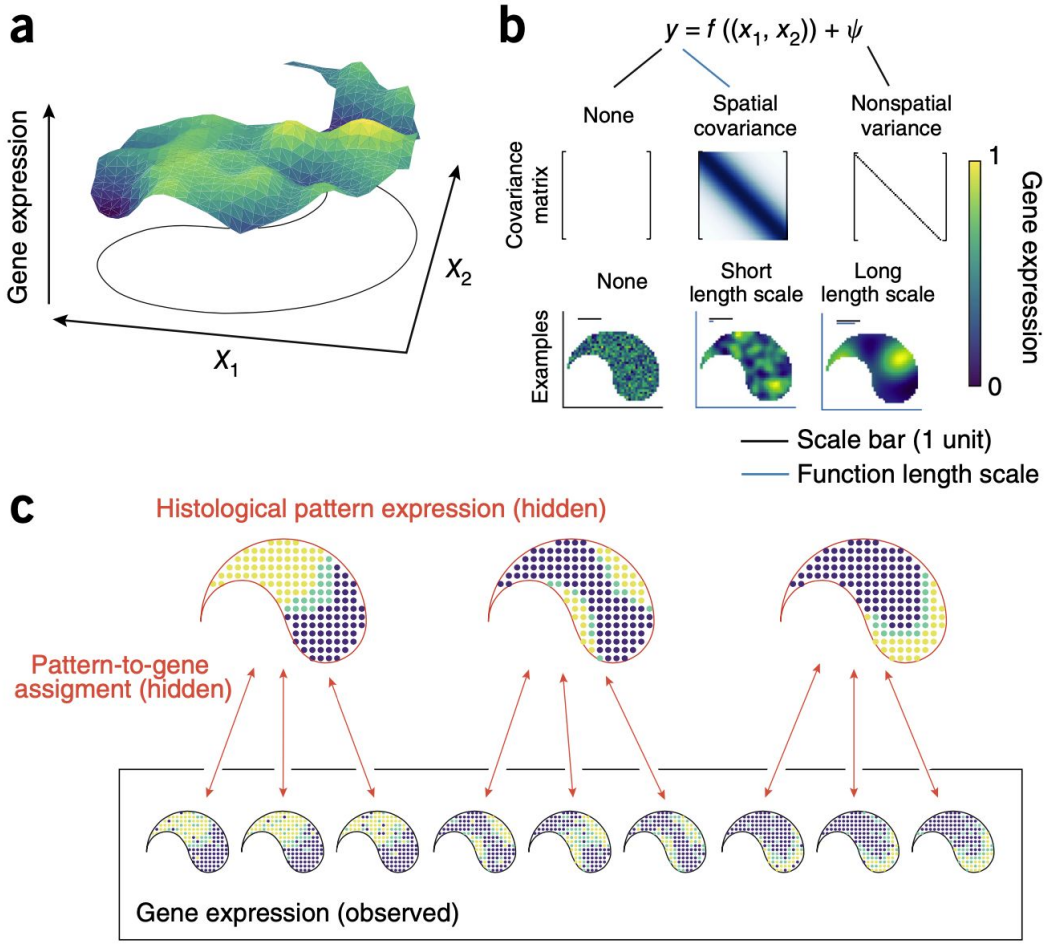


Application: Spatial transcriptomics

SpatialDE: identification of spatially variable genes

Valentine Svensson[1,2], Sarah A Teichmann[1,3] & Oliver Stegle[2,4]

Technological advances have made it possible to measure spatially resolved gene expression at high throughput. However, methods to analyze these data are not established. Here we describe SpatialDE, a statistical test to identify genes with spatial patterns of expression variation from multiplexed imaging or spatial RNA-sequencing data. SpatialDE also implements 'automatic expression histology', a spatial gene-clustering approach that enables expression-based tissue histology.



SpatialDE model. SpatialDE models gene expression profiles $y = (y_1, \ldots, y_N)$ for a given gene across spatial coordinates $X = (x_1, \ldots, x_N)$, using a multivariate normal model of the form

$$P(y \mid \mu, \sigma_s^2, \delta, \Sigma) = N(y \mid \mu \cdot 1, \sigma_s^2 \cdot (\Sigma + \delta \cdot I)) \quad (1)$$

The fixed effect $\mu_g \cdot 1$ accounts for the mean expression level, and Σ denotes a spatial covariance matrix defined on the basis of the input coordinates of pairs of cells. SpatialDE uses the so-called squared exponential covariance function to define Σ:

$$\Sigma_{i,j} = k(x_i, x_j) = \exp\left(- \frac{|x_i - x_j|^2}{2 \cdot l^2} \right) \quad (2)$$

whereby the covariance between pairs of cells i and j is modeled to decay exponentially with the squared distance between them. The hyperparameter l, also known as the characteristic length scale, determines how rapidly the covariance decays as a function of distance[21].

The second covariance term $\delta \cdot I$ accounts for independent non-spatial variation in gene expression, where the ratio $1/(1 + \delta)$ can be interpreted as the fraction of expression variance attributable to spatial effects. Model parameters are fit by maximizing the marginal log likelihood (LL),

$$LL = -\frac{1}{2} \cdot N \cdot \log(2\pi) - \frac{1}{2} \cdot \log(|\sigma_s^2 \cdot (\Sigma + \delta \cdot I)|) - \frac{1}{2} \cdot (y - \mu \cdot 1)^T (\sigma_s^2 \cdot (\Sigma + \delta \cdot I))^{-1} (y - \mu \cdot 1) \quad (3)$$

Statistical significance. To estimate statistical significance, we compared the model likelihood of the fitted SpatialDE model with the likelihood of a model that corresponds to the null hypothesis of no spatial covariance,

$$P(y \mid \mu, \sigma^2) = N(\mu \cdot 1, \sigma^2 \cdot I) \quad (4)$$

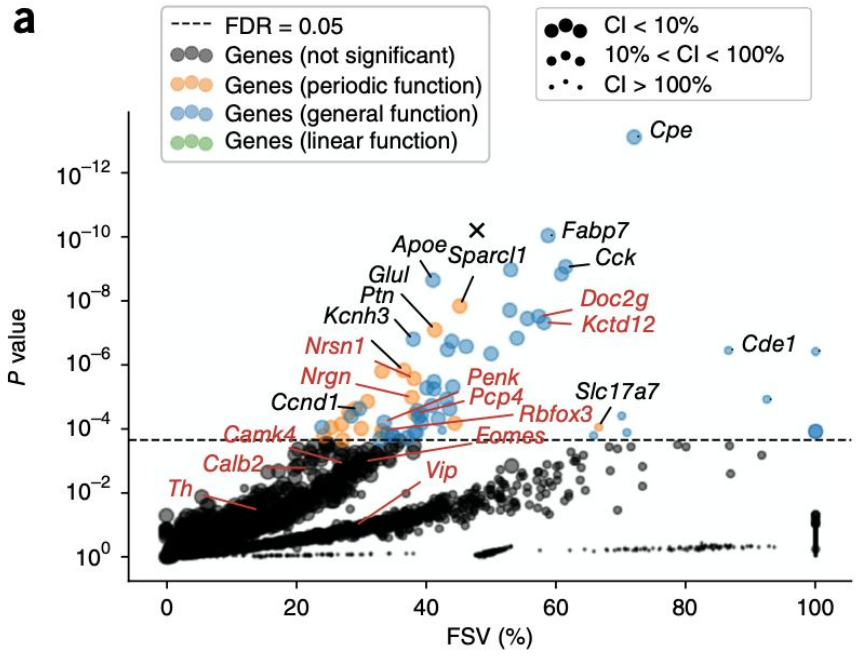
We then estimated P values analytically on the basis of the χ^2 distribution transformation with one degree of freedom. Unless stated otherwise, we used the Q-value method[23] to adjust for multiple testing, thereby controlling the FDR.

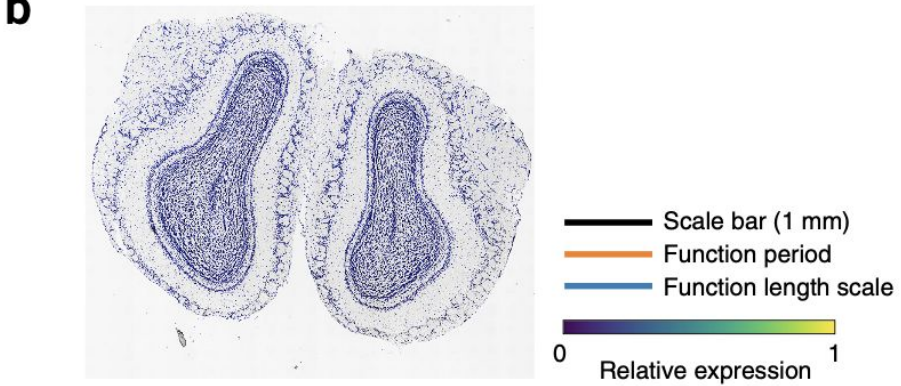
Model selection. After significance testing, the spatial covariance patterns identified can be further investigated through comparisons of models with alternative covariance functions.

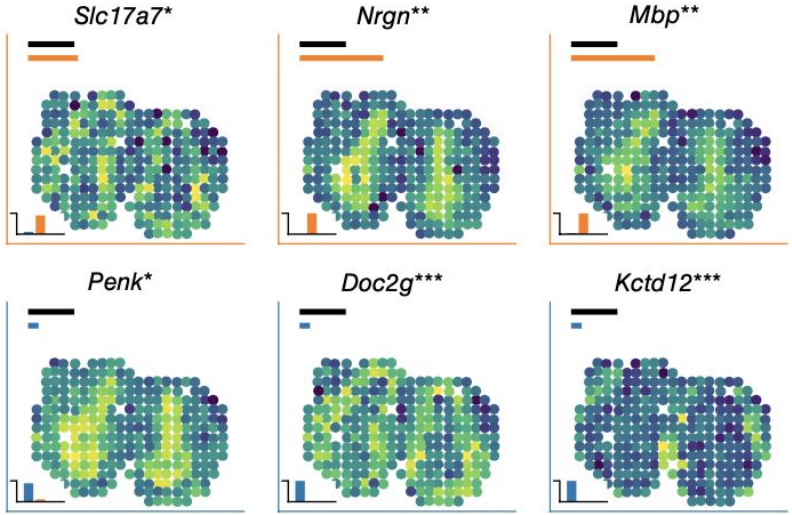
In addition to the squared exponential covariance (equation (2)), SpatialDE implements covariance functions that assume linear trends and periodic patterns of gene expression variation (**Supplementary Fig. 1**), which are compared using the Bayesian information criterion (BIC):

$$\mathrm{BIC} = \log(N) \cdot M - 2 \cdot \mathrm{LL}$$

Here M is the number of hyperparameters of a given model, N is the number of samples, and LL (equation (3)) is the log marginal likelihood of the data. Guidance on how to interpret these inferences and alternative functional forms is available in **Supplementary Note 1**.


a
FDR = 0.05
Genes (not significant)
Genes (periodic function)
Genes (general function)
Genes (linear function)
CI < 10%
10% < CI < 100%
CI > 100%
P value
FSV (%)
Cpe
Fabp7
Apoe
Sparcl1
Cck
Glul
Ptn
Doc2g
Kcnh3
Kctd12
Cde1
Nrsn1
Nrgn
Penk
Slc17a7
Ccnd1
Pcp4
Rbfox3
Camk4
Eomes
Calb2
Vip
Th


b
Scale bar (1 mm)
Function period
Function length scale
0
1
Relative expression


c
Slc17a7*
Nrgn**
Mbp**
Penk*
Doc2g***
Kctd12***

Relations to other models in ML

1. Close correspondence between SVM solution and MAP estimate of GP classification
    a. SVMs: hinge loss gives rise to sparse solutions
    b. GPs have probabilistic interpretation
2. Neural Networks
    a. Infinitely wide NNs converge to a Gaussian process
    b. Gaussian processes can be used in a stacked manner similar to NNs

Links to Implementations

https://blog.dominodatalab.com/fitting-gaussian-process-models-python/

https://nbviewer.jupyter.org/github/adamian/adamian.github.io/blob/master/talks/Brown2016.ipynb

Python library GPy https://github.com/SheffieldML/notebook/tree/master/GPy

Further Readings / References

- A whole book on **Gaussian Processes in Machine Learning (Rasmussen and Williams, 2006)** http://www.gaussianprocess.org/gpml/
- One chapter in David MacKay's book: http://www.inference.org.uk/mackay/itprnn/ps/534.548.pdf
- https://katbailey.github.io/post/gaussian-processes-for-dummies/