Inflammatix

A Practical Introduction to Gaussian Processes and their Biomedical Applications

Michael Mayhew, PhD
Senior Machine Learning Scientist

November 20, 2019
Cambridge, United Kingdom
Gaussian Processes Cambridge Meetup

DISCLAIMER (What is this talk?)

- This is a *summary* introduction to *applied* data science with Gaussian processes

- This is a talk delivered by someone learning Gaussian processes themselves

- Goals of this talk:
  - Give intuition on GPs
  - Leave pointers/references for relevant materials & resources
  - Showcase successful applications of GPs in biomedicine

Some useful prerequisites

- Regression
  - Simple linear
  - Generalized linear (e.g. logistic, probit)
  - Polynomial

- Multivariate normal theory

- Support vector machines and the ‘kernel trick’

What is a Gaussian Process?

- **A Gaussian process is a *distribution over functions***
  - ***Whether performing regression or classification, we're generally interested in learning functions (mappings) from some real-world inputs to some real-world outputs***

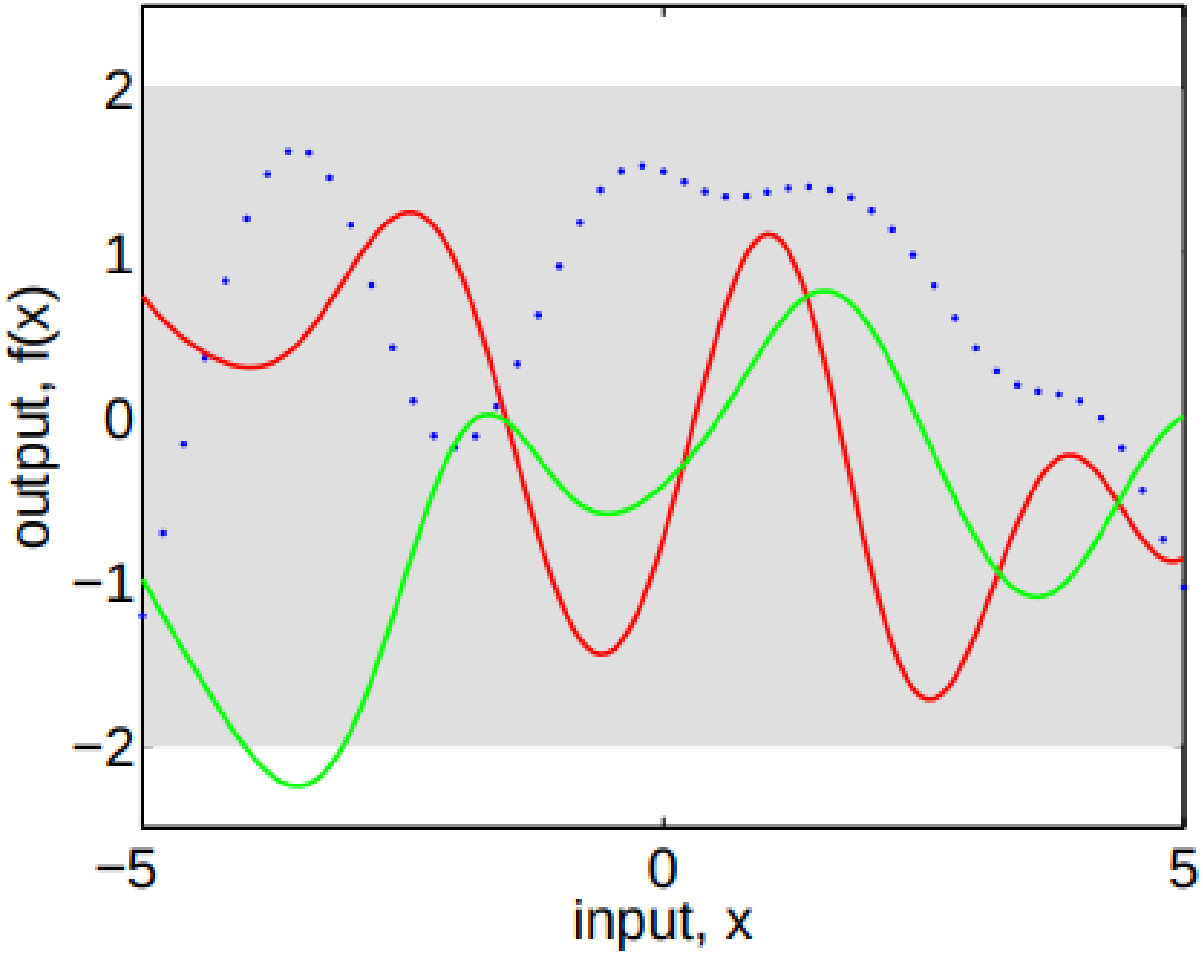
The mean function – in practice, often a constant function with value 0

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

x are your *inputs*

The covariance function – encodes assumptions about how variation in inputs affects variation in outputs

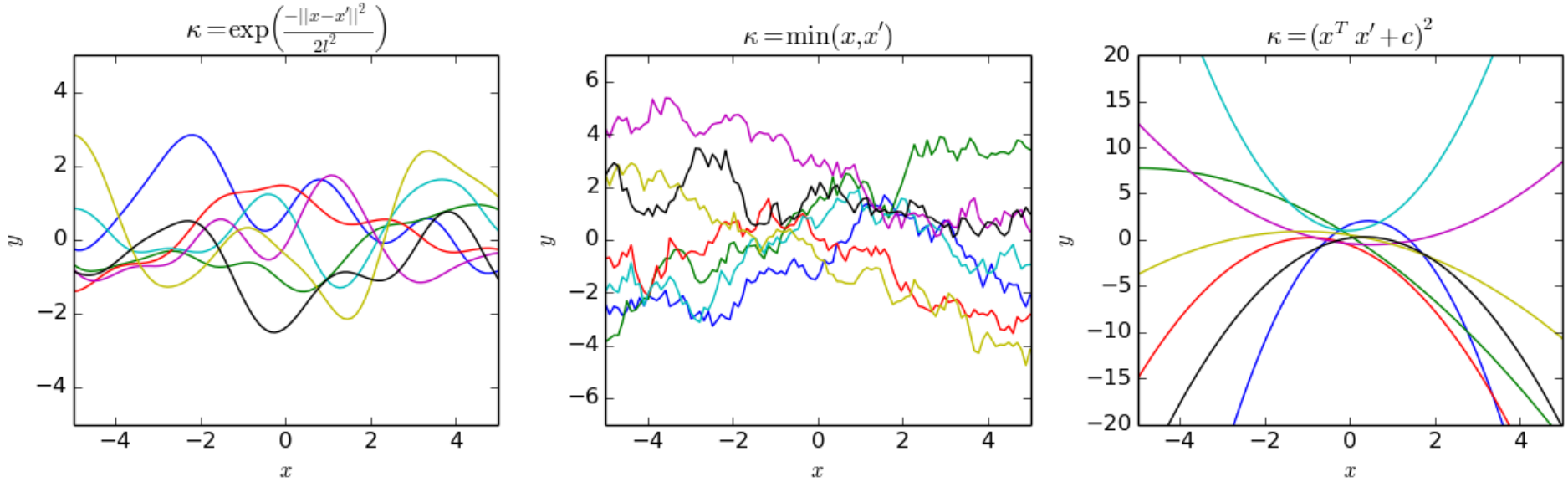
What is a Gaussian Process?



Rasmussen & Williams, *Gaussian Processes for Machine Learning (GPML)*

What is a Gaussian Process?

- **The covariance function, *k*, is the *<u>most important</u>* modeling choice for a GP**
  - ***Based on <u>similarity between input points</u>, encodes notions of:***
    - ***Smoothness/jaggedness of your function***
    - ***Structure/patterns in your function (e.g. periodicity and repetition)***



https://www.wikiwand.com/en/Gaussian_process

What is a Gaussian Process?

covariance function	expression
constant	σ_0^2
linear	$\sum_{d=1}^{D} \sigma_d^2 x_d x'_d$
polynomial	$(\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2)^p$
squared exponential	$\exp(-\frac{r^2}{2\ell^2})$
Matérn	$\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} r\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} r\right)$
exponential	$\exp(-\frac{r}{\ell})$
γ-exponential	$\exp\left(-(\frac{r}{\ell})^\gamma\right)$
rational quadratic	$(1 + \frac{r^2}{2\alpha\ell^2})^{-\alpha}$
neural network	$\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^\top \Sigma \tilde{\mathbf{x}}')}}\right)$

$$r = |x - x'|$$

Rasmussen & Williams, *Gaussian Processes for Machine Learning (GPML)*

What is a Gaussian Process?

Signal variance – (roughly) specifies the range over which the output varies

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}(x - x')^2 \right)$$

Length-scale – (roughly) the distance over which one moves in input space before seeing significant changes in the outputs

Rasmussen & Williams, *Gaussian Processes for Machine Learning (GPML)*

What is a Gaussian Process?

- A Gaussian process is (nearly) a multivariate normal distribution
  - More specifically, for finite observations (which is what we all observe anyway), a GP is equivalent to a multivariate normal distribution

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, K(X, X))$$

$$K(X, X) = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_N) \\ \dots & \dots & \dots & \dots \\ k(x_N, x_1) & k(x_N, x_2) & \dots & k(x_N, x_N) \end{pmatrix}$$

What is a Gaussian Process?

- **Example: Gaussian processes for regression**

Where are my model parameters?

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}})$$

$$\boxed{f(\mathbf{x}) = \boldsymbol{\beta}\mathbf{x}}$$

$$\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma_n^2 = f(\mathbf{x}) + \epsilon$$

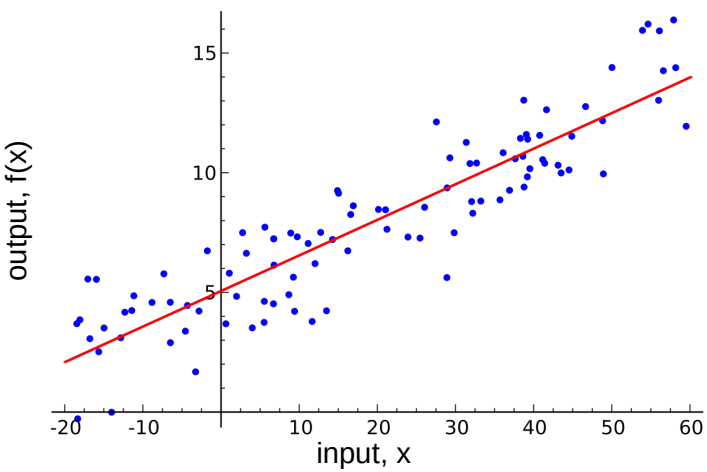
$$\epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

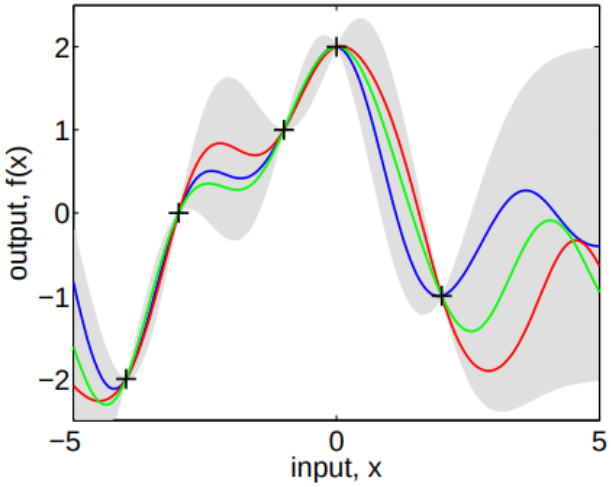
f is your 'parameter'

$$\boxed{f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))}$$

$$\mathbf{y}|\mathbf{x}, \sigma_n^2 = f(\mathbf{x}) + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2)$$





Rasmussen & Williams, *Gaussian Processes for Machine Learning (GPML)*

What is a Gaussian Process?

- **Example: Gaussian processes for classification**

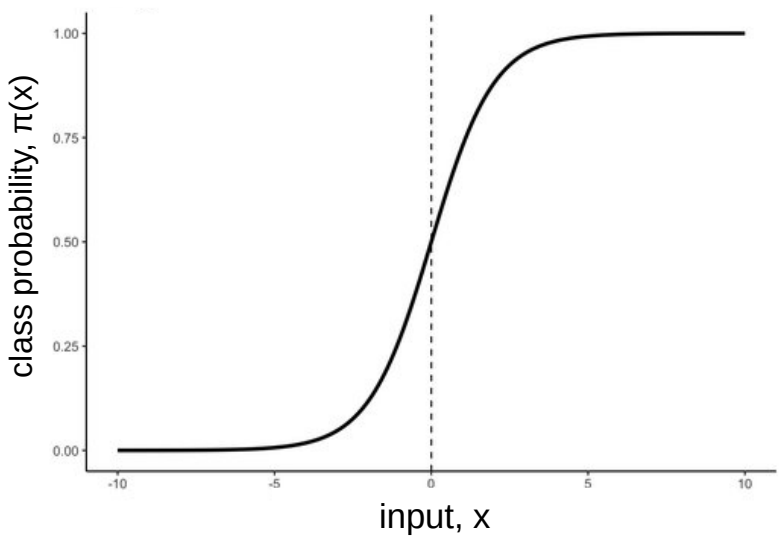
$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}})$$

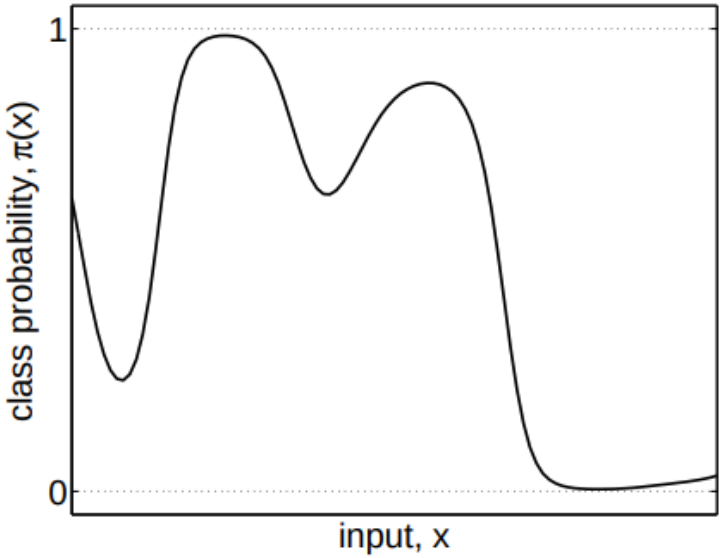
$$\log\left(\frac{p(y=1|\mathbf{x})}{1-p(y=1|\mathbf{x})}\right) = \boldsymbol{\beta}\mathbf{x}$$

$$\boxed{p(y=1|\mathbf{x}, \boldsymbol{\beta}) = \sigma(\boldsymbol{\beta}\mathbf{x})}$$

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

$$\boxed{p(y=1|\mathbf{x}, f) = \sigma(f(\mathbf{x}))}$$





Rasmussen & Williams, *Gaussian Processes for Machine Learning (GPML)*

Why Gaussian Processes?

1. GPs offer flexible nonparametric generalizations of familiar models for regression and classification
   Need a bit more than basic linear or logistic regression?

2. GPs are a useful building block for model development involving function learning
   We'll see biomedical examples in a moment!

3. GPs enable interpolation and (to an extent) extrapolation to unobserved points in the input space
   Useful for datasets with missingness/sparsity as well as for prediction

4. GPs commonly used to model objective functions in Bayesian Optimization (Mockus, 1974)
   Enables derivative-free global optimization of black-box functions (popular approach for hyperparameter optimization of neural networks)

Why not Gaussian Processes?

1. **Scalability**
   Without certain model adjustments/approximations (e.g. sparse GPs), fitting and prediction with GPs scale cubically with n

2. **Selection of the covariance function can be more art than science**
   When available covariance functions are inadequate, one may have to design one

3. **Limited adoption/cultural buy-in of GPs in the clinical world**
   Can make for a harder sell to superiors, teammates and collaborators

Gaussian Process Analysis Walk-Thru

How do I handle noisy outputs?

We can introduce noise models just as we would for, say, simple linear regression

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

$$\boxed{\begin{aligned} \mathbf{y} &= f(\mathbf{x}) + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma_n^2) \end{aligned}}$$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, K(X, X) + \sigma_n^2 I)$$

Gaussian Process Analysis Walk-Thru

How do I fit a Gaussian process model?

We determine the hyperparameter values that maximize log marginal likelihood of the training data

$$p_{\boldsymbol{\theta}}(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f}$$

$$\boxed{\boldsymbol{\theta} = \sigma_f^2, \ell, \sigma_n^2}$$

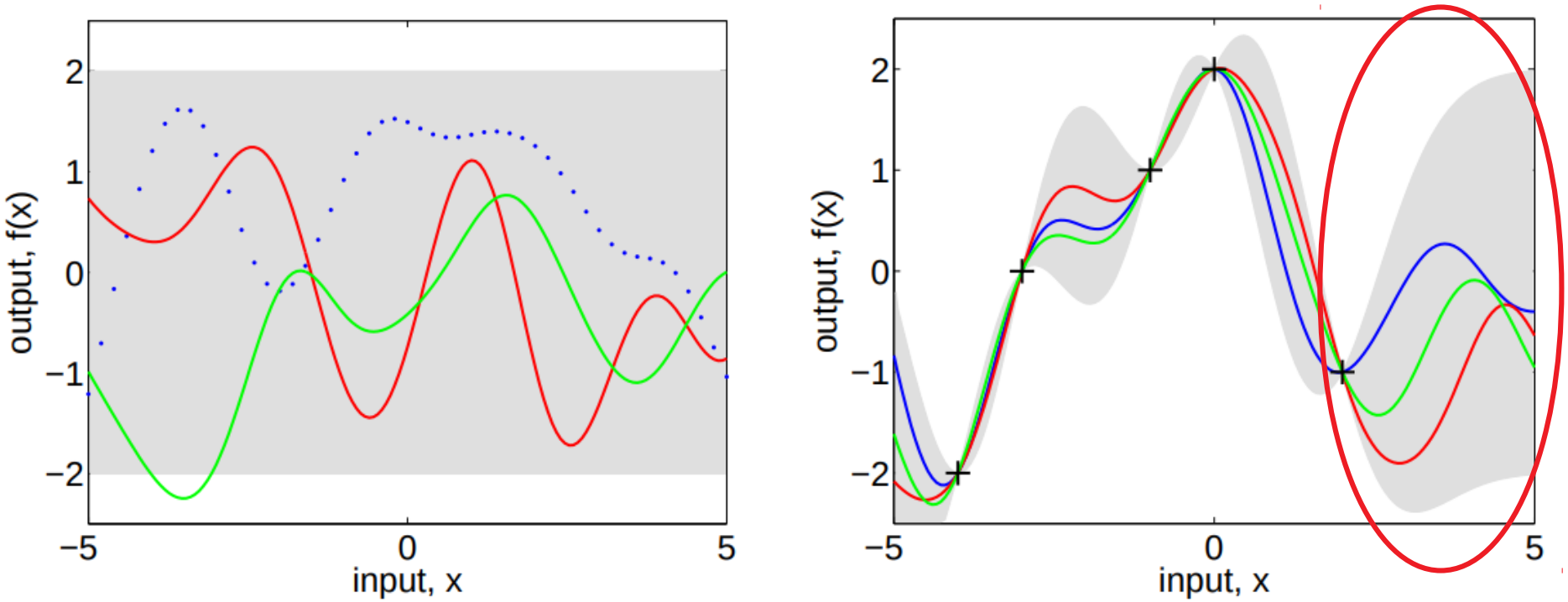
$$\log p_{\boldsymbol{\theta}}(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi$$

- **Alternatively, could select hyperparameter values that result in best performance on some held-out validation dataset or in cross-validation**

Gaussian Process Analysis Walk-Thru

How do I fit a Gaussian process model?

We determine the hyperparameter values that maximize log marginal likelihood of the training data



(a), prior (b), posterior

Rasmussen & Williams, *Gaussian Processes for Machine Learning (GPML)*

Gaussian Process Analysis Walk-Thru

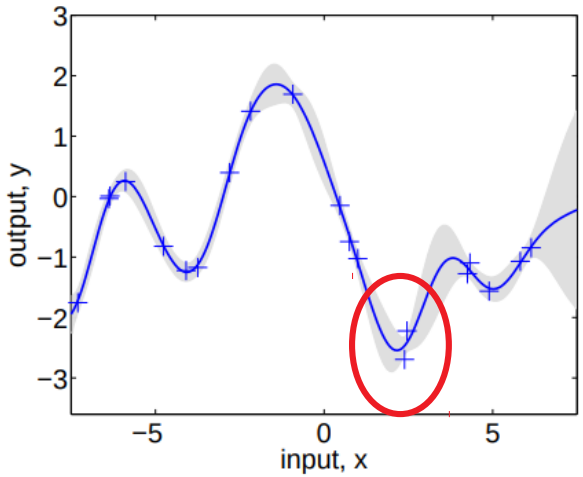
- **Data generated from GP with:**

$$\ell = 1$$

$$\sigma_f = 1$$

$$\sigma_n = 0.1$$

- **In parts (b) and (c), length-scale was fixed to shown value and signal variance and noise were estimated**



(a), $\ell = 1$

Rasmussen & Williams, *Gaussian Processes for Machine Learning (GPML)*

Gaussian Process Analysis Walk-Thru

How do I perform prediction?

For GP regression with Gaussian likelihood, we have the predictive distribution

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

$$\mathbb{E}[\mathbf{y}_*|X,\mathbf{y},X_*] = \mathbb{E}[\mathbf{f}_*|X,\mathbf{y},X_*] = K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}\mathbf{y}$$

$$\text{cov}(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}K(X,X_*)$$

$$\text{cov}(\mathbf{y}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}K(X,X_*) + \sigma_n^2 I_{d_{f_*}}$$

- **Situation is a little more complicated for classification (see Rasmussen & Williams *GPML* for more detail)**

Gaussian Process Analysis Walk-Thru

How do I handle multivariate inputs?

Covariance functions naturally handle vector-based (as well as scalar) inputs

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}(x - x')^2 \right)$$

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}') \right)$$

Gaussian Process Analysis Walk-Thru

How do I handle multi-typed inputs?

May require some covariance function engineering

- **For continuous input dimensions, plenty of kernels available**

- **For discrete input dimensions:**
  - **convert input dimension to 'one-hot' encoding**
  - **specify a squared-exponential kernel (with its own length-scale) for each dimension of the encoding**
  - **Take the product of squared-exponential kernels as the kernel for the discrete input dimension**

- **The covariance function for multi-typed inputs can then be the product of kernels for the different input dimensions**

- **Alternatively, valid kernels may exist that naturally accommodate multi-typed inputs**

David Duvenaud's Kernel Cookbook - https://www.cs.toronto.edu/~duvenaud/cookbook/

Gaussian Process Analysis Walk-Thru

How do I handle multivariate outputs?

Multi-task Gaussian processes

Inter-task covariance

$$\mathbf{K}_{MT}(\mathbf{X}_n, \mathbf{l}, \boldsymbol{\theta}_c, \boldsymbol{\theta}_t) = \mathbf{K}_c(\mathbf{l}, \boldsymbol{\theta}_c) \otimes \mathbf{K}_t(\mathbf{X}_n, \boldsymbol{\theta}_t)$$

Within-task covariance; what we've seen up to now

- **NOTE: This formulation also applies to multi-class classification**

Lecture by Mauricio Alvarez at GPSS 2017;
Bonilla, Chai & Williams, 2007; Ghassemi et al., *AAAI* 2015

Gaussian Process Analysis Walk-Thru

How do I perform feature selection?

Automatic relevance determination (ARD) covariance functions

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\Big(-\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^T M (\mathbf{x}_p - \mathbf{x}_q)\Big) + \sigma_n^2 \delta_{pq}$$

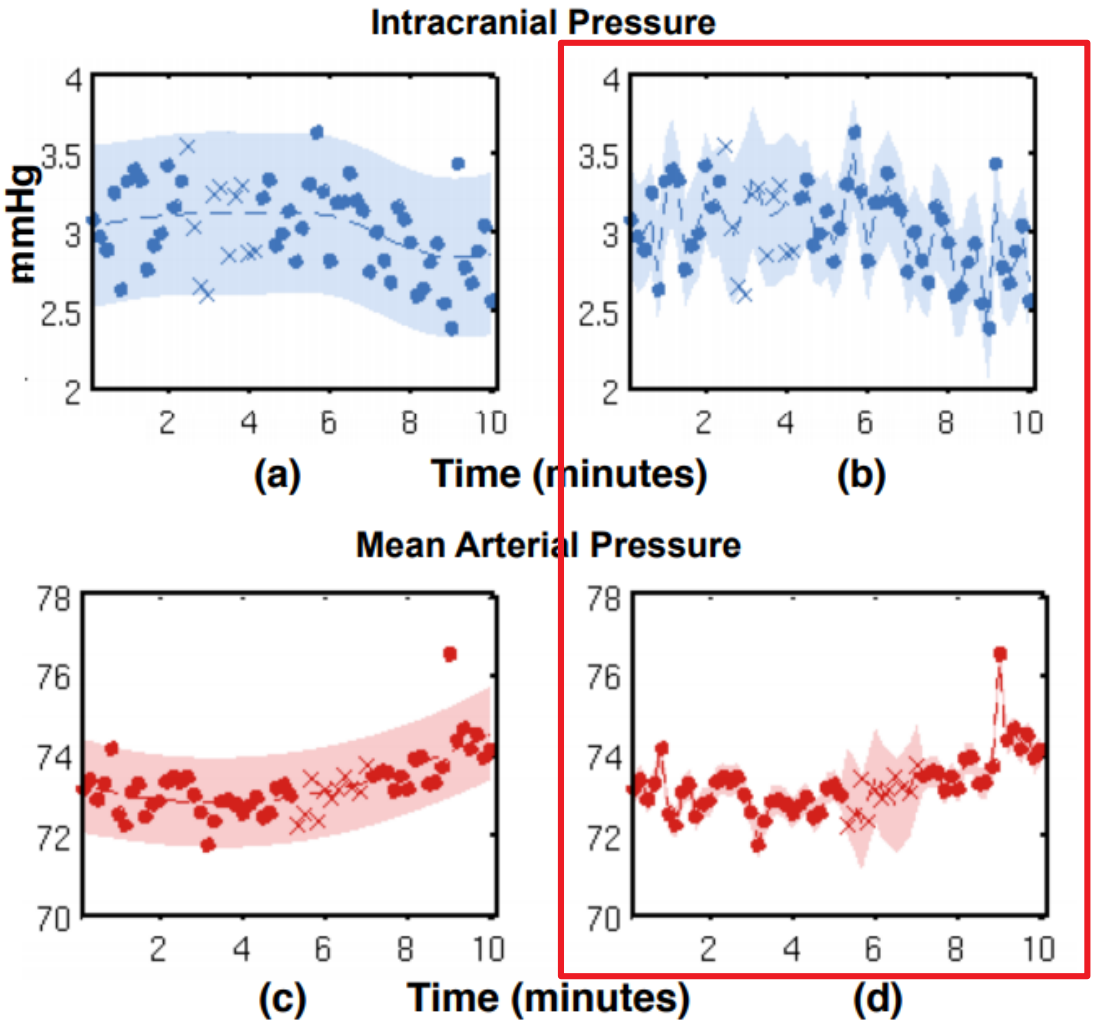
$$M = \mathrm{diag}(\boldsymbol{\ell})^{-2}$$

- **Vector, *l*, specifies length-scales for each input dimension**
- **Roughly speaking, the larger the length-scale, the greater the variation required in the corresponding input dimension for the output to change (and so, the less relevant that feature is to capturing the behavior of the function)**

Neal, “Bayesian Learning in Neural Networks” 1996; Williams & Rasmussen, 1996

Gaussian Processes in Biomedicine

Use case: Jointly model multiple patient time series from electronic health record

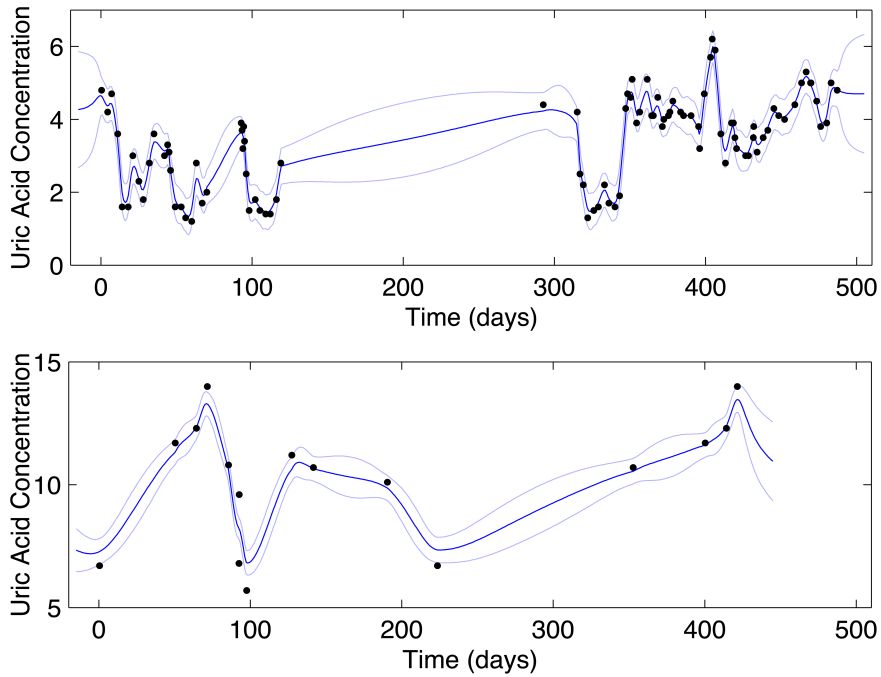


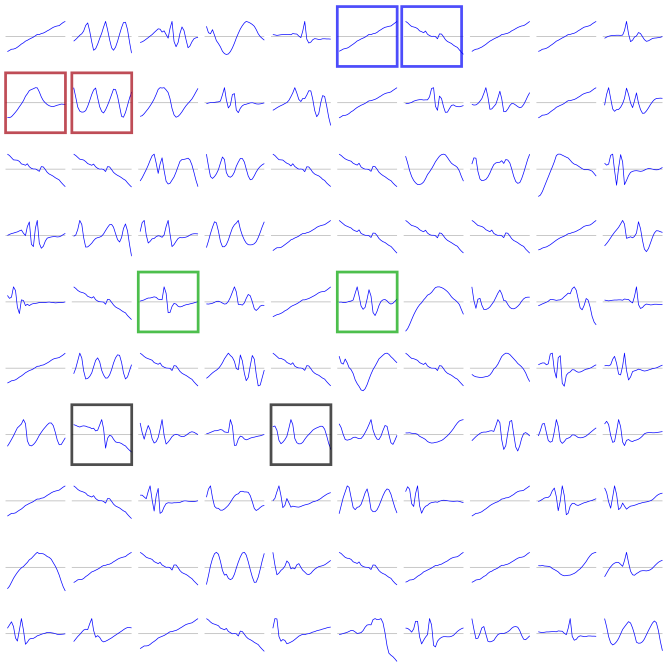
- Patient time series are generally:
  - Noisy
  - Irregularly sampled
  - Non-stationary (we haven't discussed this)
- Authors use multi-task GPs to jointly model patient time series of:
  - vital signs (n=35)
  - topic proportions derived from clinical notes (n=~7k)
- Hyperparameters of MTGPs with clinical notes were used as features to predict mortality

Ghassemi et al., *AAAI* 2015; Durichen et al., *IEEE TBME* 2015

Gaussian Processes in Biomedicine

Use case: Enable downstream unsupervised learning of temporal features of uric acid concentration related to gout and leukemia





- **Authors use GPs to fit patient (n=~4k) time series of uric acid concentration**
- **Samples of uric acid trajectories were sampled from the GP models and passed to a variational autoencoder to identify temporal features**
- **The features were used to predict gout or leukemia, and the features compared favorably to handcrafted (expert) features**

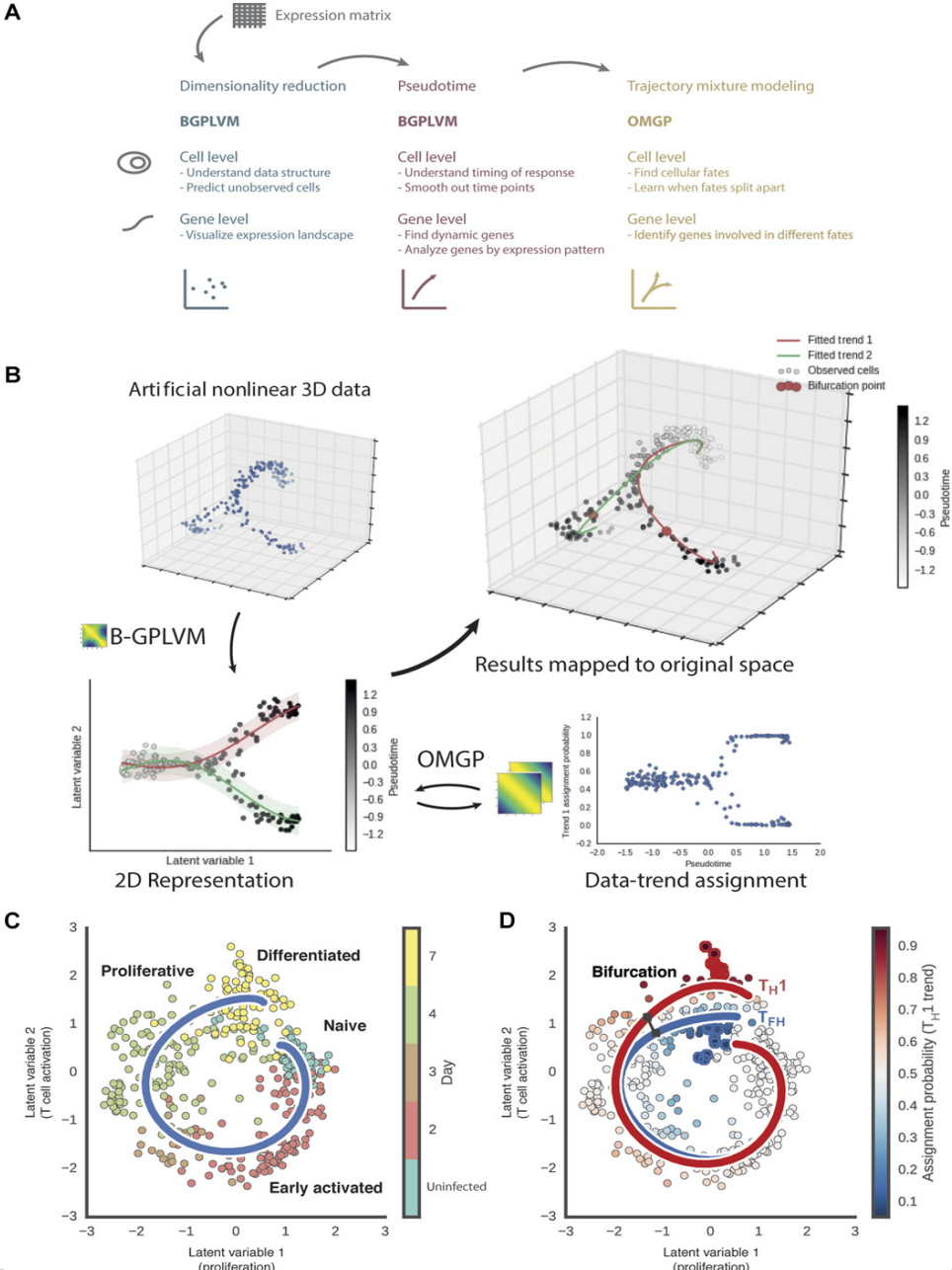
Lasko et al., *PloS ONE* 2013

Gaussian Processes in Biomedicine

Use case: Identify temporal trends associated with cell fate in malaria

- **Leverages two novel extensions of GPs**
  - ***GPLVMs (Gaussian process Latent Variable Models)* – inputs to be inferred and assumed to be lower dimensional than observed outputs**
  - ***OMGPs (Overlapping Mixtures of Gaussian processes)* – identify individual trajectories from mixed output observations**

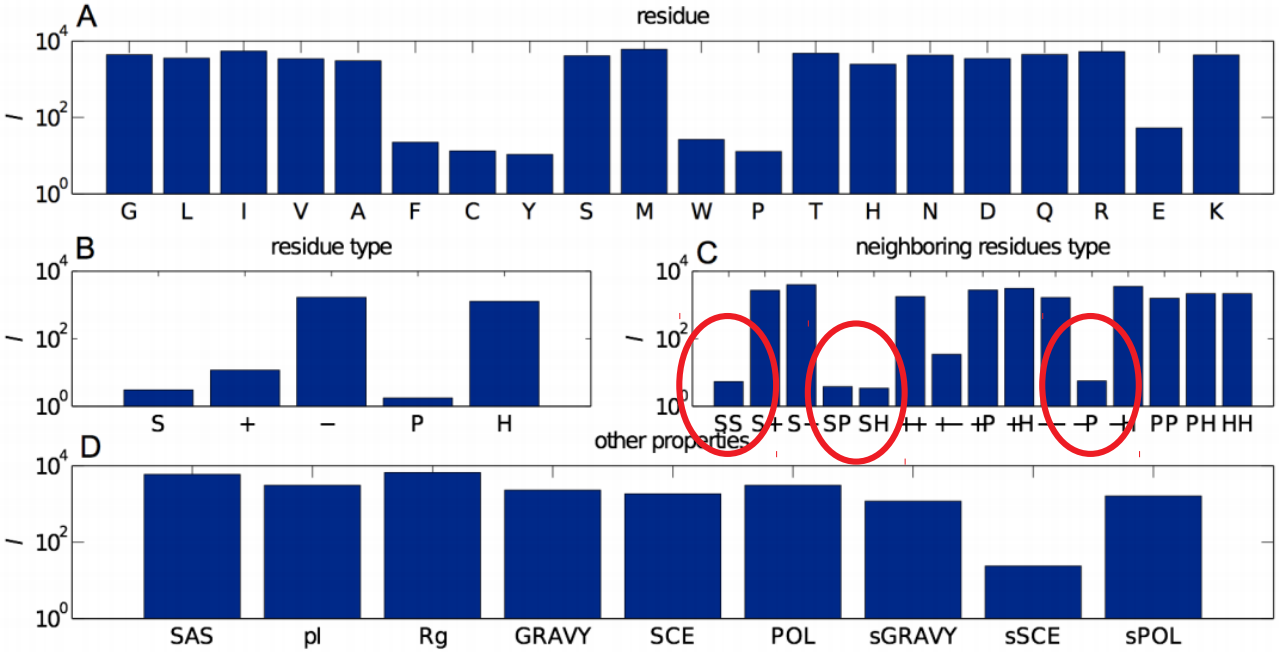
- **Modeling reveals gene expression patterns associated with helper T cell differentiation as bifurcation events in cell fate**



Lonnberg et al., *Sci. Immunology* 2017

Gaussian Processes in Biomedicine

Use case: Predict crystallization propensity of proteins



- **Authors performed GP regression to predict the propensity for crystallization of proteins (n=~100) based on a set of (largely) protein residue-based features (roughly 50)**

- **Used ARD covariance function to determine that certain neighboring residue types were more associated with crystallization propensity than others**

Fusco et al., *PLoS ONE* 2014

When Gaussian Processes?

Model	Type	Hyperparam. Dimensionality	Data Scale	Interpretability	Adoption/ Visibility
LASSO Logistic Regression	Parametric	Low	Low-High	High	High
Random Forests/GB Trees	Non-parametric	High	Medium-High	Medium	High
Neural Networks	Parametric	High	Medium-High	Low	High

When Gaussian Processes?

Model	Type	Hyperparam. Dimensionality	Data Scale	Interpretability	Adoption/ Visibility
LASSO Logistic Regression	Parametric	Low	Low-High	High	High
Random Forests/GB Trees	Non-parametric	High	Medium-High	Medium	High
Neural Networks	Parametric	High	Medium-High	Low	High
Gaussian Processes	Non-parametric	Low*	Low-Medium*	Medium	Medium*

Gaussian Process Resources

- Software Packages (non-exhaustive list)
  - GPy, GPflow, sklearn (Python)
  - GPML (MATLAB)
  - GPfit (R)
  - Bayesian Optimization: Spearmint, Dragonfly, HyperOpt, GpyOpt, BoTorch
- Books
  - *Gaussian Processes for Machine Learning (GPML)* by Rasmussen & Williams – start here!
- Lectures/Tutorials
  - Gaussian Process Summer School lectures (http://gpss.cc/gpss19/program)
- People!
  - Few places in the world with the amassed GP expertise of Cambridge

Thanks for your attention!
Any questions?

mmayhew@inflammatix.com
@DataForager
https://www.linkedin.com/in/michael-mayhew-40406468