Stony Brook University

Explainable Learning with Gaussian Processes

Petar M. Djurić
Department of Electrical and Computer Engineering
Stony Brook University

Based on joint work with Kurt Butler and Guanchao Feng

Explainability in machine learning

Explainability in machine learning refers to the ability to make a model's decisions or predictions understandable to humans.

Why it matters?

- Trust and transparency
- Debugging and improvement:
- Accountability and ethics:
- Fairness and bias detection:

Types of explainability:

- Global explainability
- Local explainability

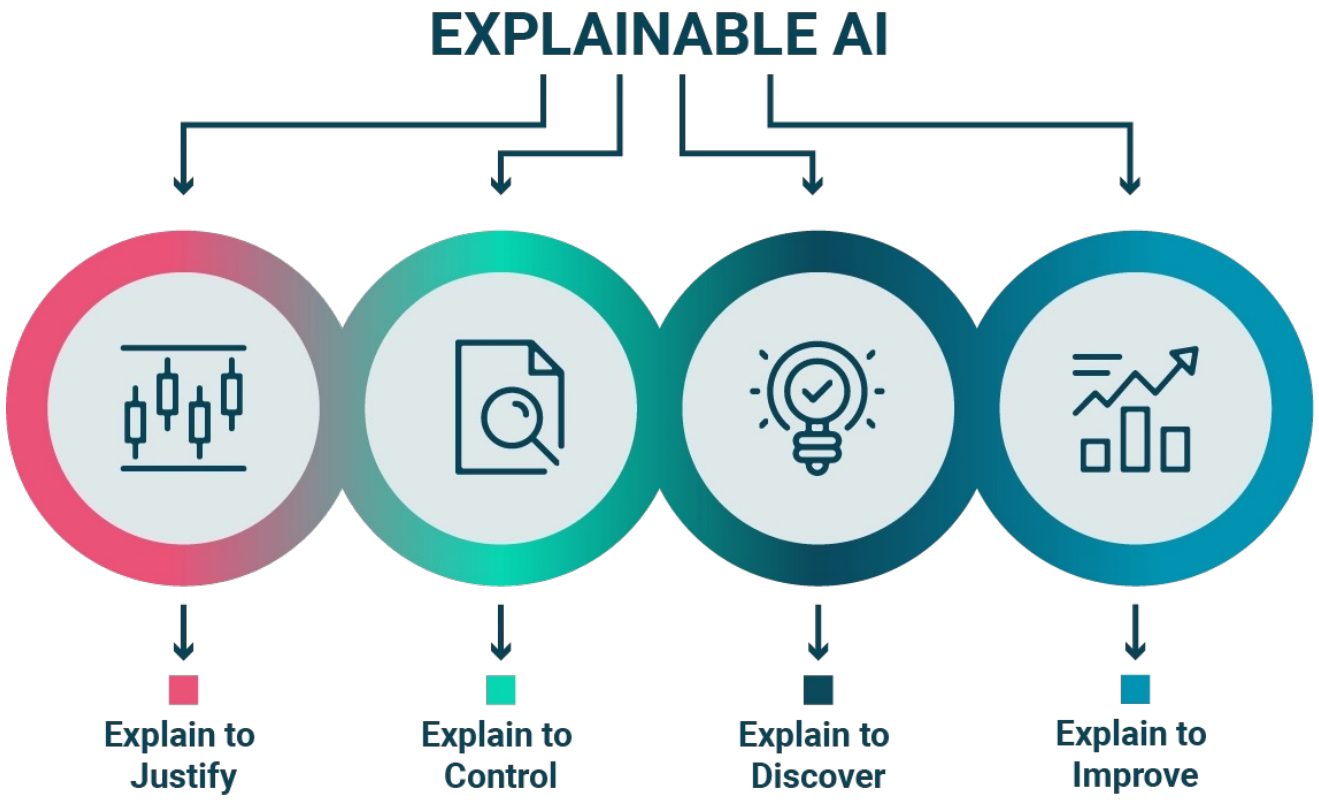


Image taken from Frontiere.io

The concept of attribution

Attribution refers to identifying the impact or contribution of input features on a model's output.

An example: *Medical diagnosis model*

Consider a model that predicts the likelihood of a disease based on **age, family history, blood pressure, and cholesterol levels**.

Suppose the model predicts a **70% probability** of the disease.
Attribution quantifies how much each feature contributed to the prediction. For instance:

- **High cholesterol** might contribute **30%**.
- **Age** could contribute **20%**.
- **Family history** could add **15%**.
- **Blood pressure** might add **5%**.

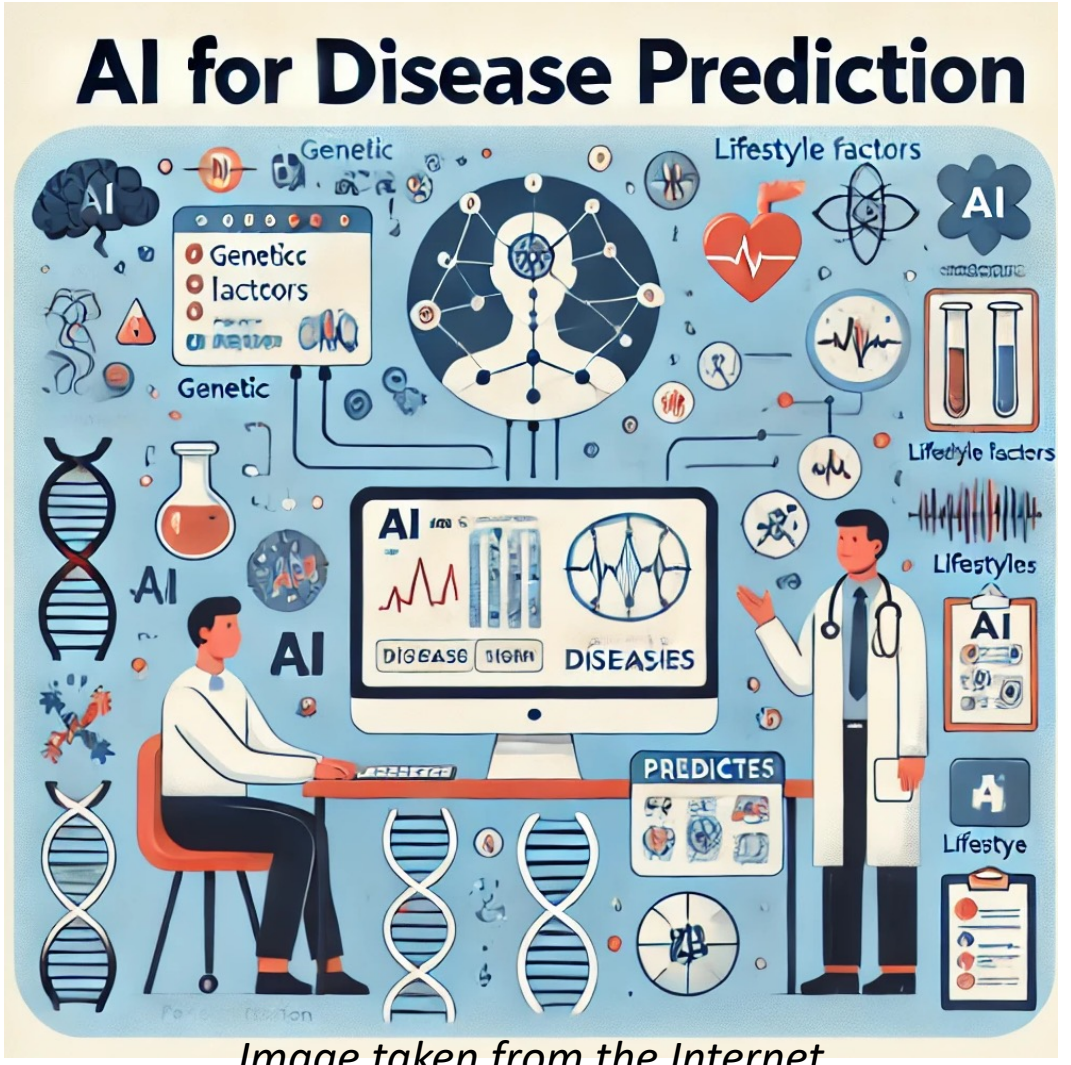


Image taken from the Internet.

Predicting house prices

Image taken from the Internet.

The model takes in features such as: **Square footage, Number of bedrooms, Location (e.g., distance to city center), Age of the house**, **Condition of the house.**

Now, suppose the model predicts that a specific house is worth **$500,000**.

Attribution answers the question: *How much did each feature contribute to this price prediction?*

Loan approval

Suppose a machine learning model predicts a 70% chance of loan approval for a customer based on:

- Income
- Credit score
- Employment status

An attribution method provides values for the contribution of each feature, e.g.,

- Income: +20% (pushed the prediction up by 20%)
- Credit Score: +30%
- Employment Status: +20%

Image taken from Forbes.com

Approaches

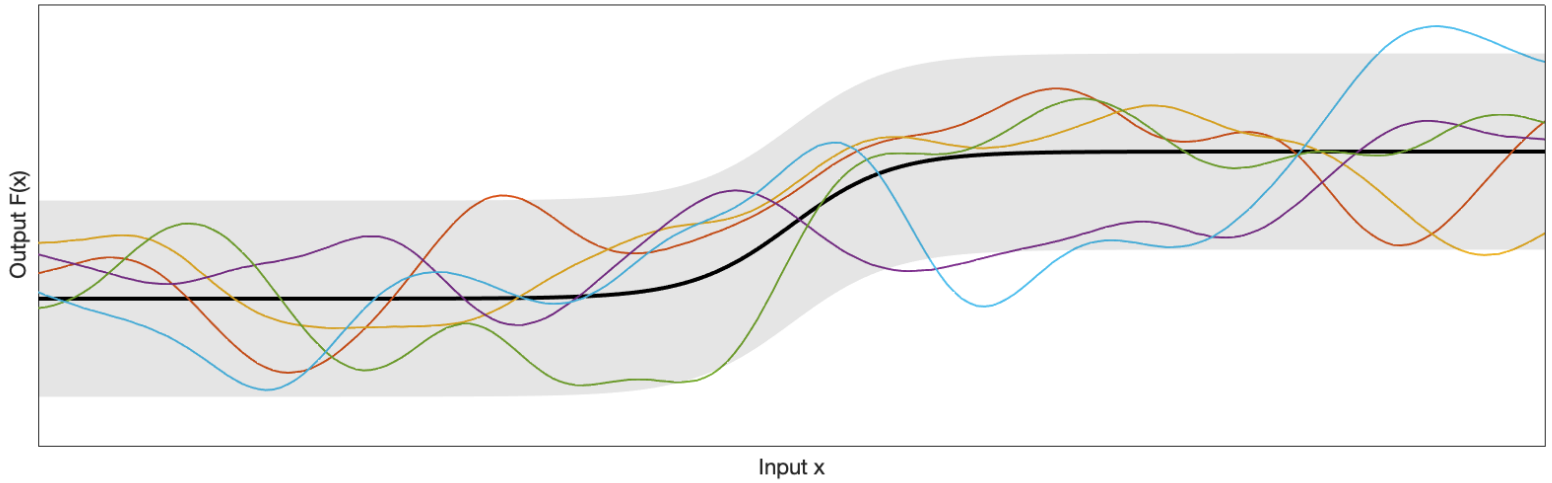
- Gradient based methods (saliency maps, integrated gradients)
- Perturbation-based methods (SHAP - SHapley Additive exPlanations)
- Surrogate models (decision trees)
- Attention mechanisms (transformers in NLP)
- Layer-wise relevance propagation
- Model-specific methods (tree SHAP)

Lloyd Shapley
1923-1926

Our workhorse: Gaussian processes

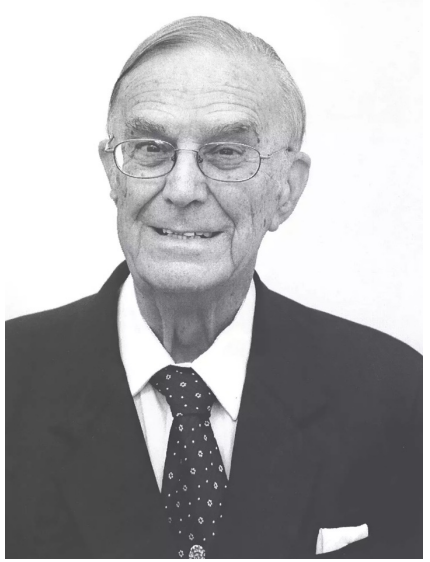
- A Gaussian process (GP) is a probability distribution over a space of functions.
  - A GP is specified by its mean and covariance functions, $m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')$.

$$F \sim \mathcal{GP}(m, k) \quad \Longleftrightarrow \quad \begin{aligned} \mathbb{E}(F(\mathbf{x})) &= m(\mathbf{x}) \\ \mathrm{Cov}(F(\mathbf{x}), F(\mathbf{x}')) &= k(\mathbf{x}, \mathbf{x}') \end{aligned}$$



A bit of history

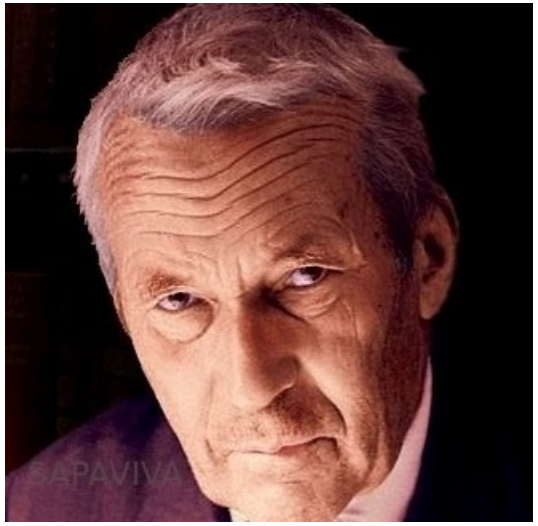
Danie Gerhardus Krige (26 August 1919 – 3 March 2013) - a South African statistician and mining engineer; was professor at the University of the Witwatersrand, Republic of South Africa. In the 1950s, sought a more accurate method to estimate ore grades by considering the spatial structure and the variability of samples.



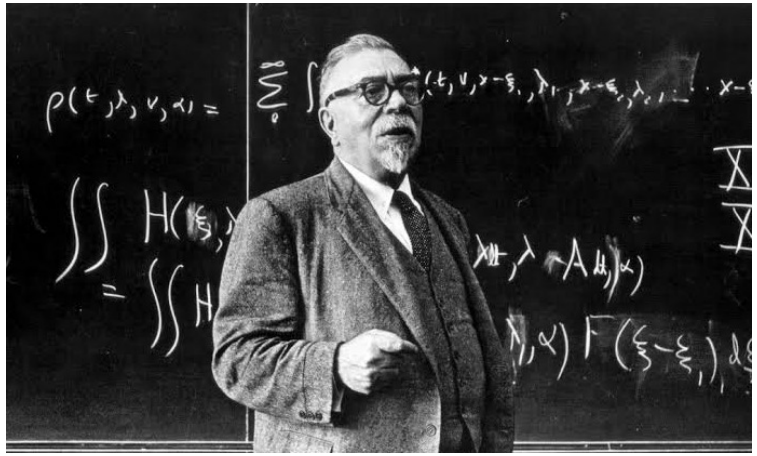
Georges François Paul Marie Matheron (2 December 1930 – 7 August 2000) was a French mathematician and civil engineer of mines. His PhD advisor was Paul Lévy. He established a rigorous mathematical framework for kriging, describing it as the **best linear unbiased estimator** for spatially correlated data. Published in **Traité** de **géostatistique appliquée,** Editions Technip 1962

A bit more history

1777-1855

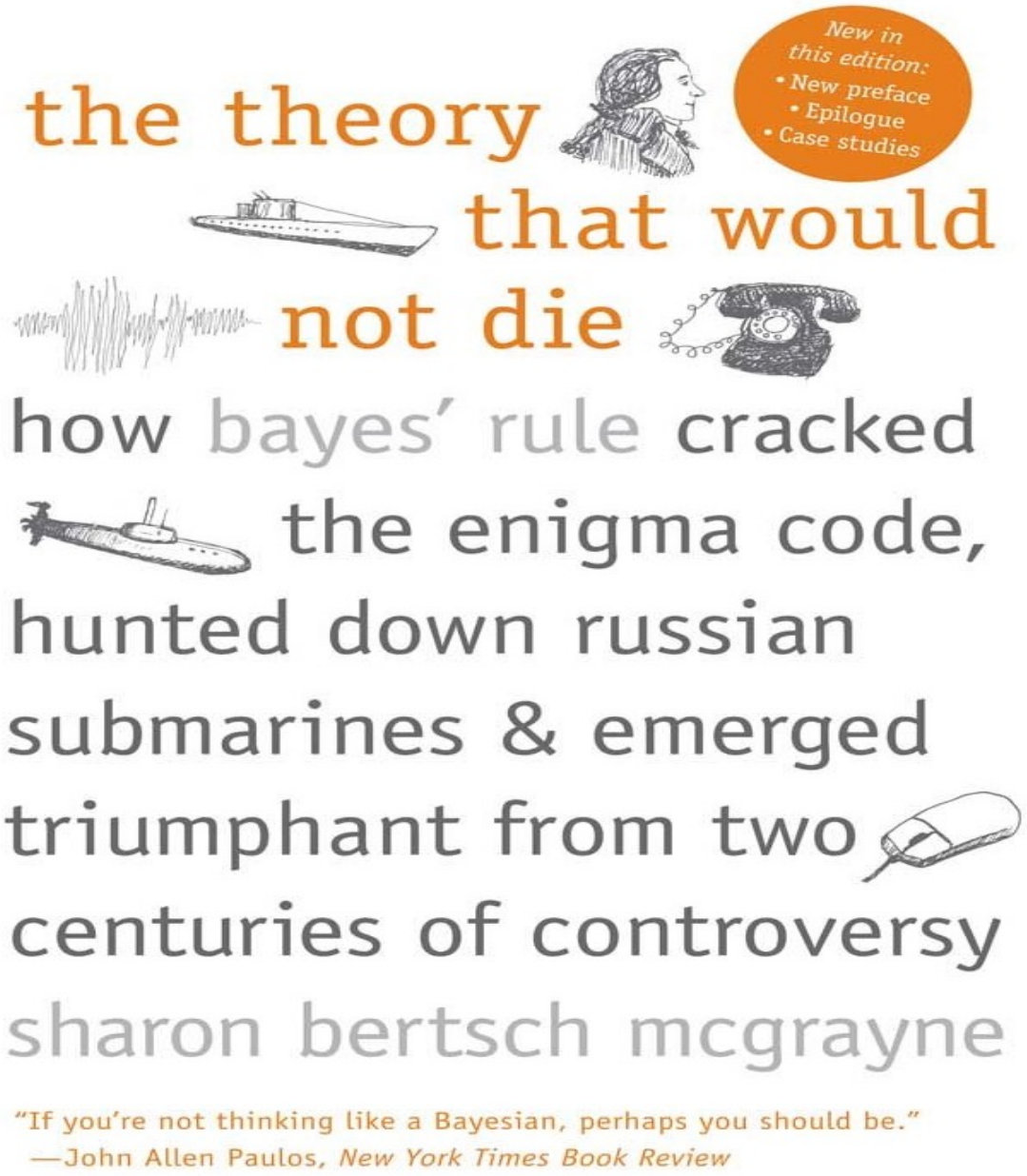


1903-1987



1894-1964

A recommendation for fn reading



Sharon Bertsch McGrayne is the author of highly-praised books about scientific discoveries and the scientists who make them.

Gaussian process regression

- Gaussian process regression (GPR) is a Bayesian approach to learning an unknown functional relationship.

Model $$y_n = F(\mathbf{x}_n) + \varepsilon_n$$
$$\varepsilon_n \sim \mathcal{N}(0, \sigma_n^2)$$

$$\mathcal{D} = \{(\mathbf{x}_n, y_n) | n = 1, \cdots, N\}$$

Likelihood $$\mathbf{y}|F \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad \mathbf{m} = \begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_N) \end{bmatrix}$$

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

Goal: Obtain a posterior distribution over possible functions

Gaussian Process Regression

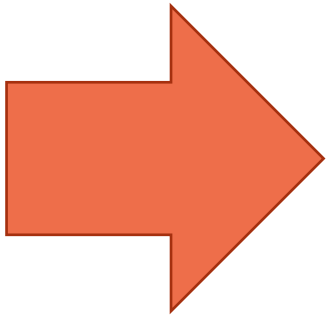
- Gaussian process regression (GPR) exploits GPs by recognizing that they are conjugate to Gaussian likelihoods.

Prior belief

$$F \sim \mathcal{GP}(m, k)$$

Likelihood

$$\mathbf{y}|F \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$



Posterior belief

$$F|\mathcal{D} \sim \mathcal{GP}(m_p, k_p)$$

$$m_p(\mathbf{x}) = m(\mathbf{x}) + \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{m})$$

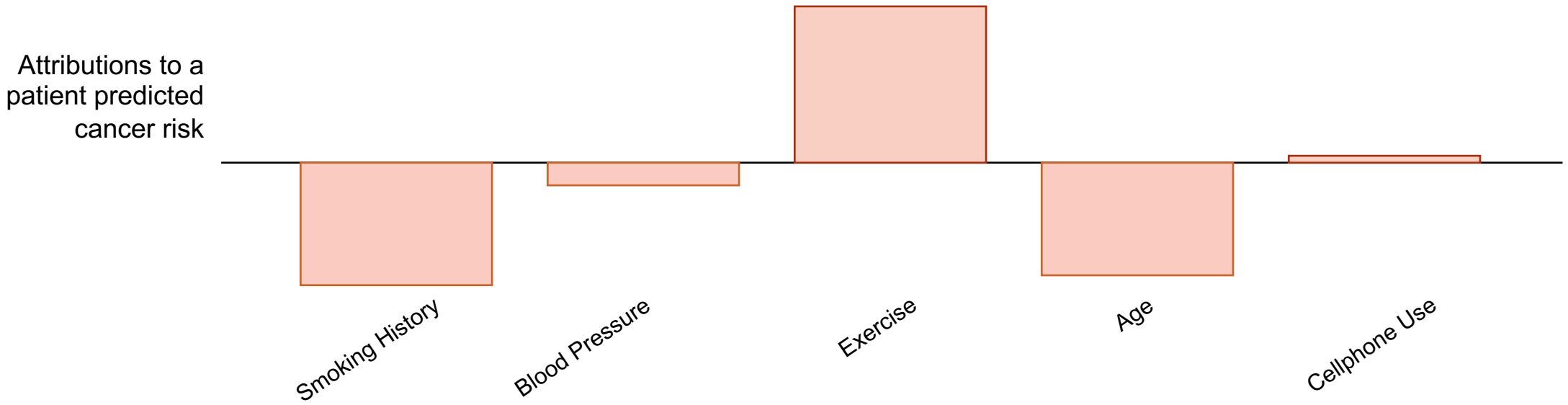
$$k_p(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}')$$

$$\mathbf{k}(\mathbf{x}) = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ k(\mathbf{x}, \mathbf{x}_N) \end{bmatrix}$$

Attribution Theory

- Feature attributions quantify how much each input variable contributed to the output result.

$$\underbrace{F(\mathbf{x})}_{\text{New prediction}} - \underbrace{F(\tilde{\mathbf{x}})}_{\text{Baseline point}} = \sum_{i=1}^{D} \underbrace{attr_i(\mathbf{x}|F)}_{\text{Contribution of feature } i}$$



Attribution Theory

Consider the Bayesian linear regression model

$$y|\boldsymbol{x}, \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2)$$

The predictions from the model are given by

$$F(\boldsymbol{x}) - F(\tilde{\boldsymbol{x}}) = \boldsymbol{w}^\top (\boldsymbol{x} - \tilde{\boldsymbol{x}}) = \sum_{i=1}^{D} w_i (x_i - \tilde{x}_i)$$

and the canonical choice for the attribution to the x_i feature is

$$attr_i(\boldsymbol{x}) = w_i (x_i - \tilde{x}_i)$$

Attribution Theory

Let

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Then we can show that

$$\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$$

where

$$\boldsymbol{\Sigma}' = \left(\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

$$\boldsymbol{\mu}' = \boldsymbol{\Sigma}'\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{y}\right)$$

And

$$attr_i(\boldsymbol{x})|\boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(\mu_i'(x_i - \tilde{x_i}), \Sigma_{ii}'(x_i - \tilde{x_i})^2)$$

Integrated Gradients

In this work, we use the Integrated Gradients definition of attributions:

$$attr_i(\mathbf{x}|F) = (x_i - \tilde{x}_i) \int_0^1 \frac{\partial F(\tilde{\mathbf{x}} + t(\mathbf{x} - \tilde{\mathbf{x}}))}{\partial x_i} dt$$

The **Fundamental Theorem of Line Integrals** states the following:

$$\int_\Gamma \nabla F \, d\boldsymbol{x} = F(\gamma(1)) - F(\gamma(0))$$

$$\nabla = \left[\frac{\partial}{\partial x_1} \ \ldots \ \frac{\partial}{\partial x_D}\right]$$

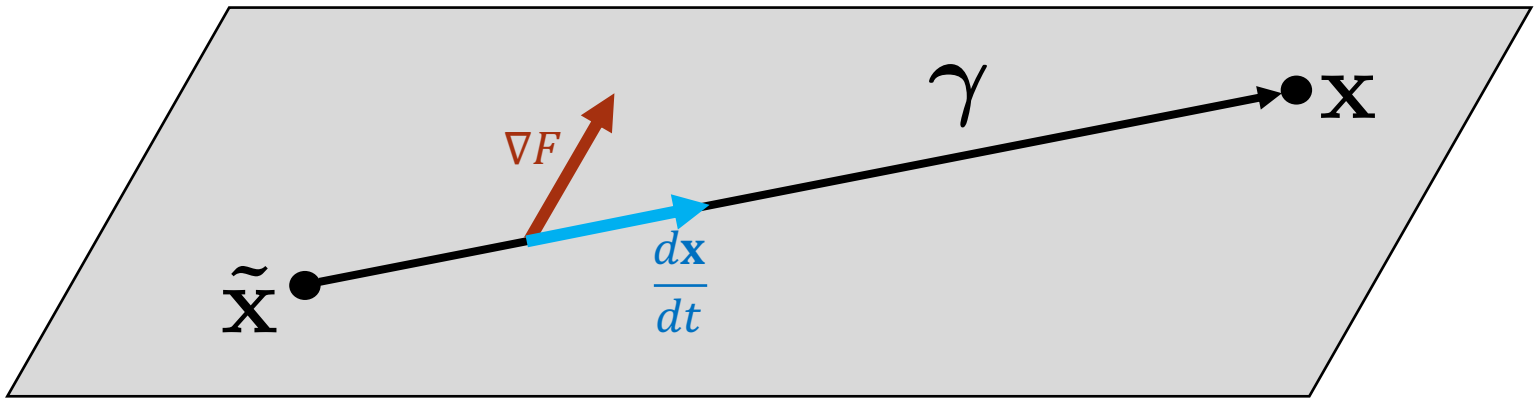
and Γ is a path in $\mathbb{R}^D$ parameterized by $\gamma(t)$, for $0 \leq t \leq 1$.

Geometric motivation for Integrated Gradients

- Let $\boldsymbol{\gamma}: [0,1] \rightarrow \mathbb{R}^D$ be a curve such that $\boldsymbol{\gamma}(t) = \tilde{\mathbf{x}} + t(\mathbf{x} - \tilde{\mathbf{x}})$.
- From the fundamental theorem of line integrals (Stokes' theorem):

$$F(\boldsymbol{\gamma}(1)) - F(\boldsymbol{\gamma}(0)) = \int_{\gamma} \nabla F \cdot d\mathbf{x}$$

$$F(\mathbf{x}) - F(\tilde{\mathbf{x}})$$



$$\int_{\Gamma} \nabla F \, d\boldsymbol{x} = \int_0^1 \sum_{i=1}^{D} \frac{\partial F(\boldsymbol{\gamma}(t)}{\partial \gamma_i} \frac{d\gamma_i}{dt} dt$$

$$= \sum_{i=1}^{D} (x_i - \tilde{x}_i) \int_0^1 \frac{\partial F(\boldsymbol{\gamma}(t))}{\partial \gamma_i} dt$$

$$= \sum_{i=1}^{D} attr_i(\boldsymbol{x}|F)$$

Attributions of GPs are GPs

Theorem: If F is distributed according to a GP,

$$F \sim \mathcal{GP}(m, k)$$

where $m \in C^1(\mathbb{R}^D)$ and $k \in C^2(\mathbb{R}^D \times \mathbb{R}^D)$, then

$$attr_i(\mathbf{x}|F) \sim \mathcal{GP}(\mu_i, \kappa_i)$$

where the mean and covariance functions are given by

$$\mu_i(\mathbf{x}) = (x_i - \tilde{x}_i) \int_0^1 \frac{\partial m(\tilde{\mathbf{x}} + t(\mathbf{x} - \tilde{\mathbf{x}}))}{\partial x_i} dt,$$

$$\kappa_i(\mathbf{x}, \mathbf{x}') = (x_i - \tilde{x}_i)(x_i' - \tilde{x}_i) \int_0^1 \int_0^1 \frac{\partial^2 k(\tilde{\mathbf{x}} + s(\mathbf{x} - \tilde{\mathbf{x}}), \tilde{\mathbf{x}} + t(\mathbf{x}' - \tilde{\mathbf{x}}))}{\partial x_i \partial x_i'} dt ds$$

Practical result

- The theorem shows that Integrated Gradients attributions preserve Gaussianity.
- Since we obtain a GP posterior from GPR, it follows that the attributions of functions modeled by GPRs are also GPs.
- Thus, we can sensibly discuss the distribution of $attr_i(\mathbf{x}|F)$ and its properties.

$$F|\mathcal{D} \sim \mathcal{GP}(m_p, k_p)$$

$$attr_1(\mathbf{x}|F)|D \sim \mathcal{GP}(\mu_1, \kappa_1)$$

$$attr_i(\mathbf{x}|F)|D \sim \mathcal{GP}(\mu_i, \kappa_i)$$

$$attr_D(\mathbf{x}|F)|D \sim \mathcal{GP}(\mu_D, \kappa_D)$$

Closed-Form Expressions for Attributions

- We can provide closed form expressions and code implementing the attributions of GPs with squared exponential type kernels.

$$k_{SE}(\mathbf{x}, \mathbf{x}') = \sigma_0^2 \exp\left(-\sum_{i=1}^{D} \frac{(x_i - x_i')^2}{2\ell^2}\right)$$

- After some derivations, the attributions have the following mean and variances,

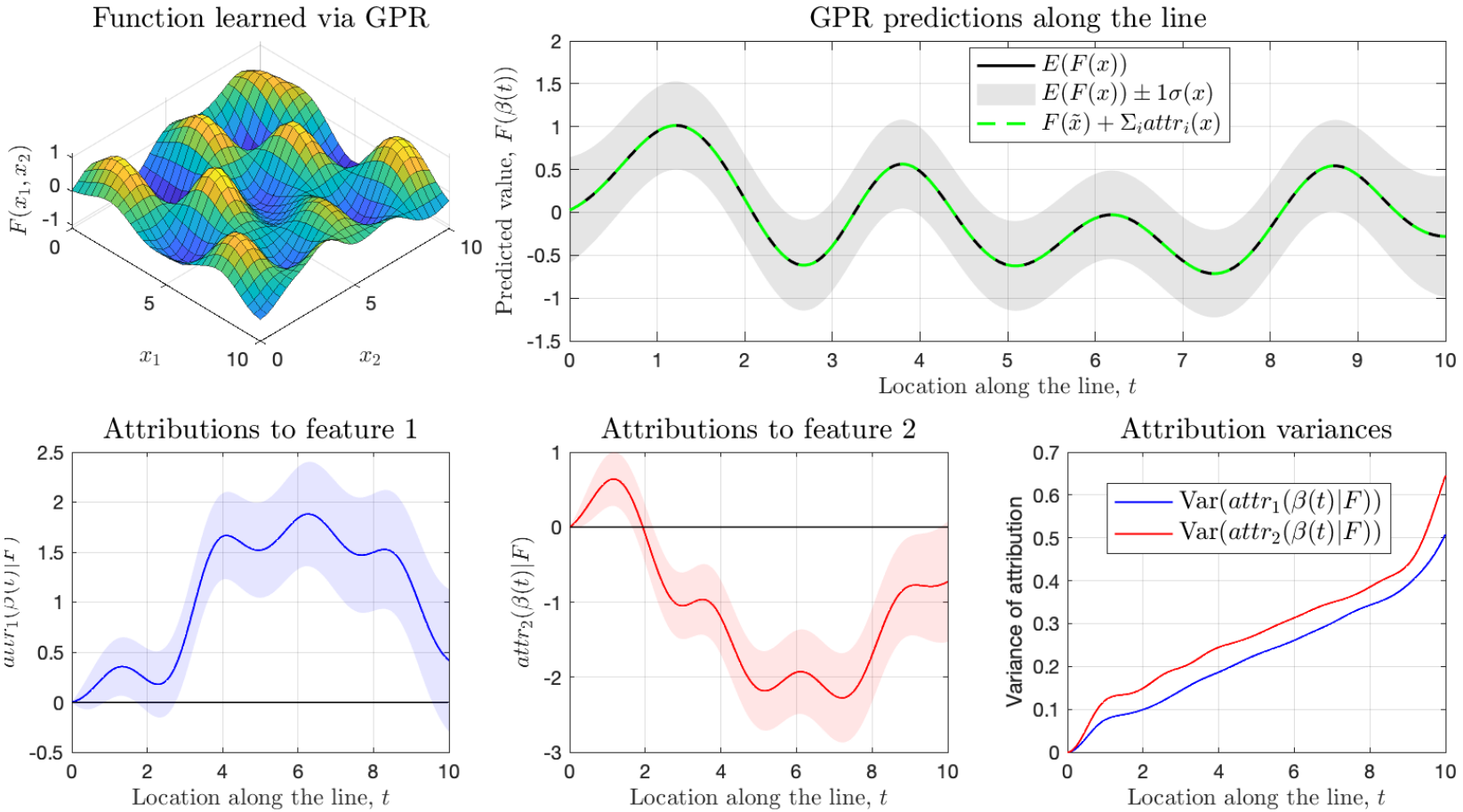
$$\mathbb{E}(attr_i(\mathbf{x})) = \sum_{n=1}^{N} \alpha_n A_{n,i}(\mathbf{x}),$$

$$\mathrm{Var}(attr_i(\mathbf{x})) = B_i(\mathbf{x}) - \sum_{n=1}^{N}\sum_{m=1}^{N} \Lambda_{n,m} A_{n,i}(\mathbf{x}) A_{m,i}(\mathbf{x}),$$

Experiment with Simulated Data

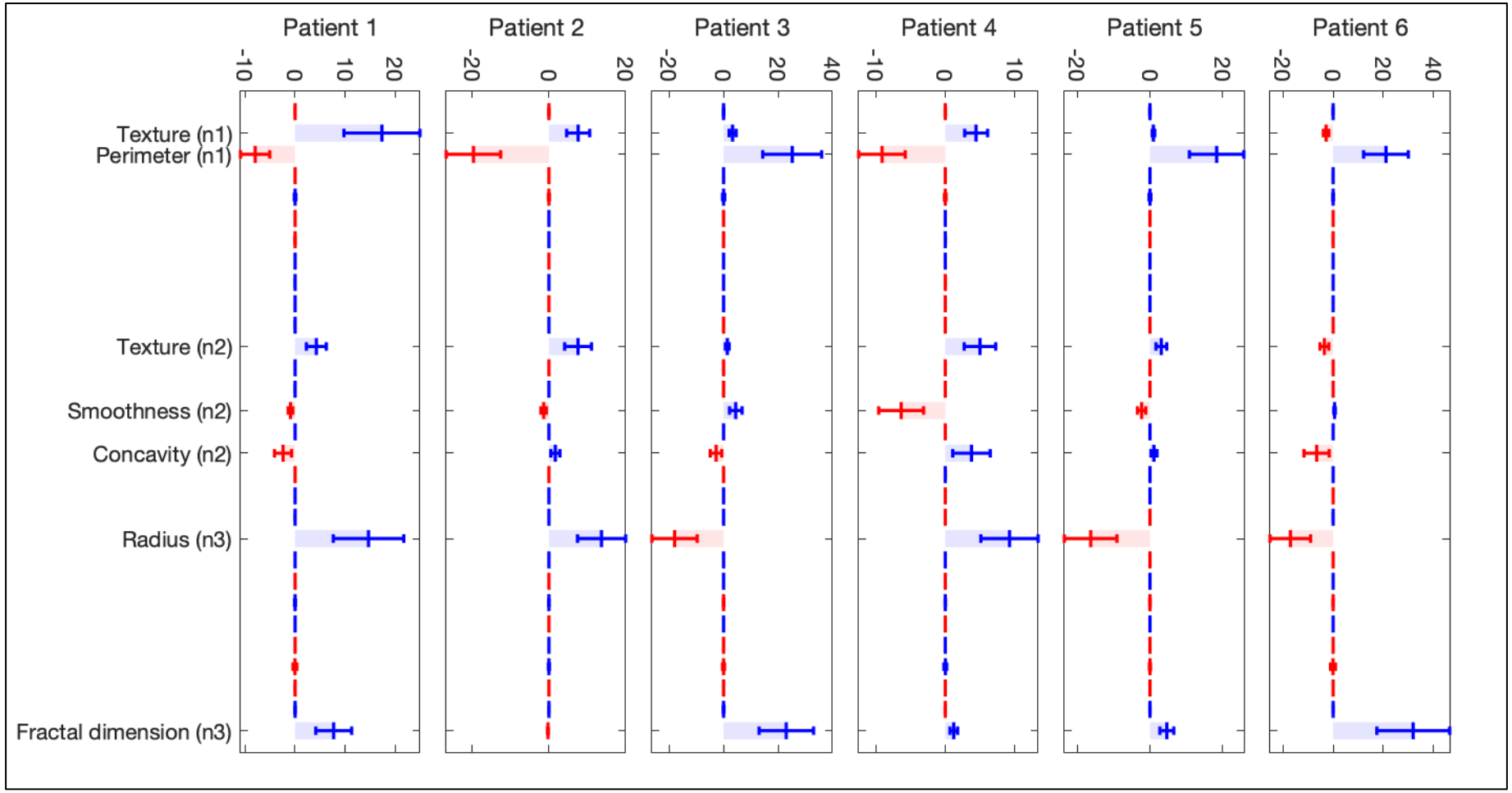
The data were generated according to $X_1 \sim \mathcal{U}(0,10), X_2 \sim \mathcal{U}(0,10)$,

$$N_Y \sim \mathcal{N}(0,1), \qquad y = \sin x_1 \sin 1.5x_2 + \frac{1}{2} n_y$$



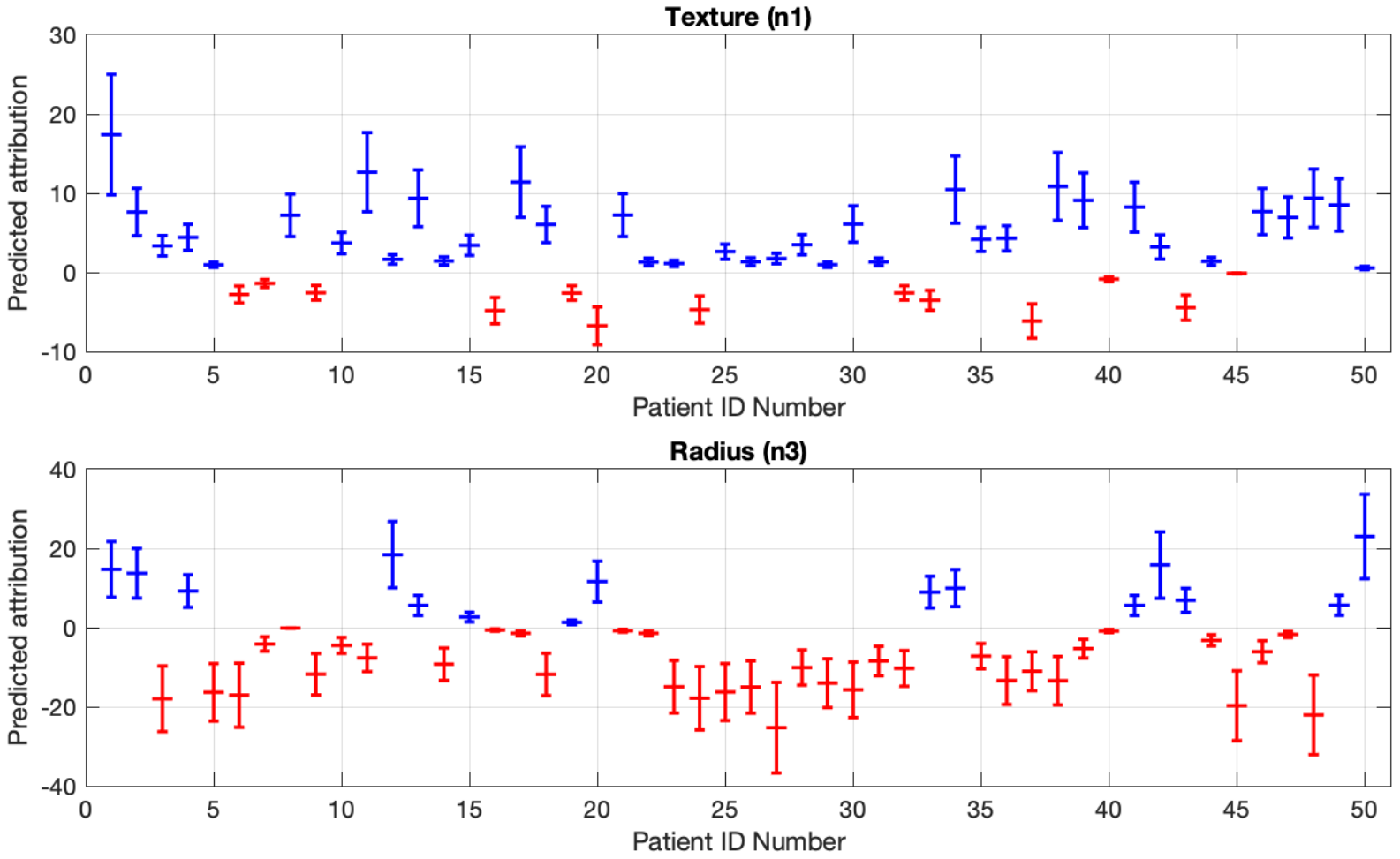
Breast Cancer Diagnosis

- GPR provides uncertainty in predictions
- More important features (of the 30 features) have more uncertainty.

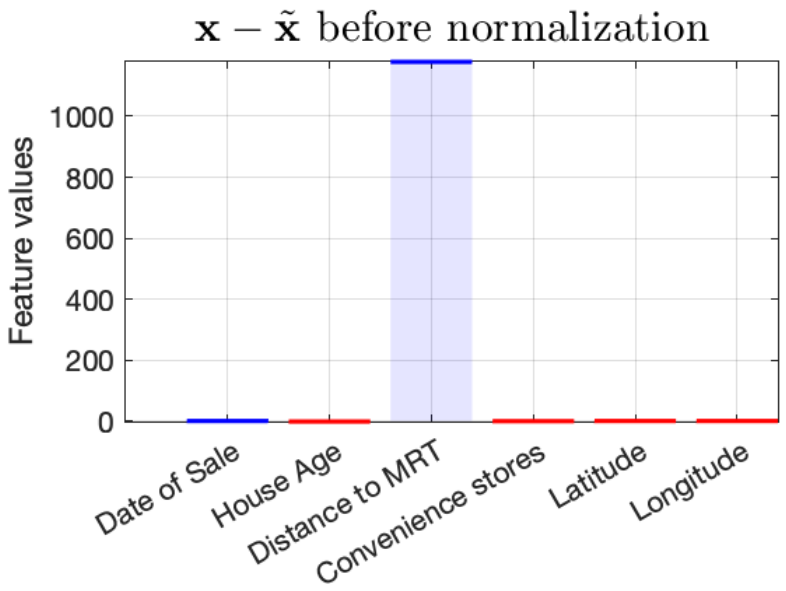


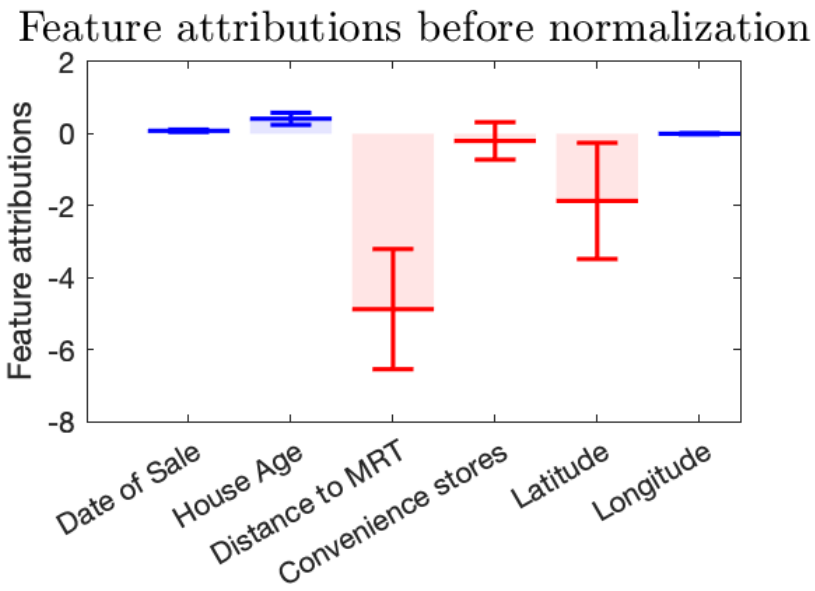
Breast Cancer Diagnosis (cont'd)

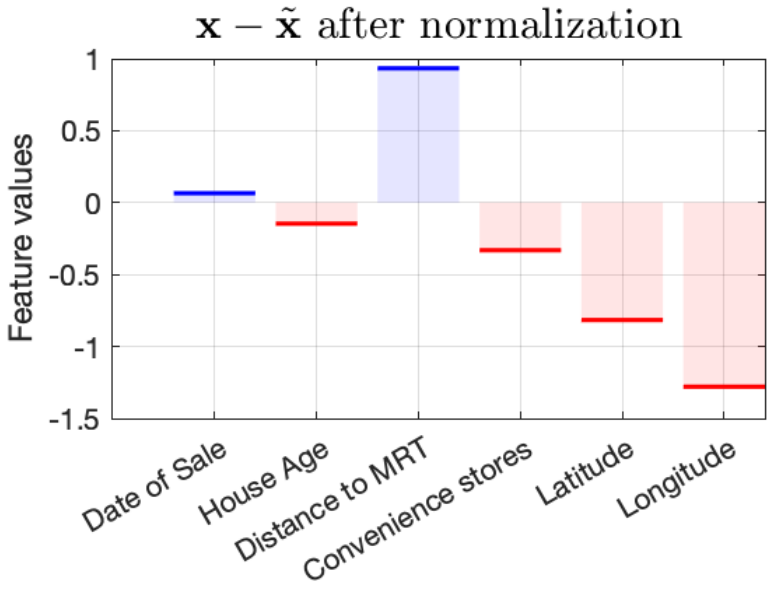
The variation of attribution of two features across patients

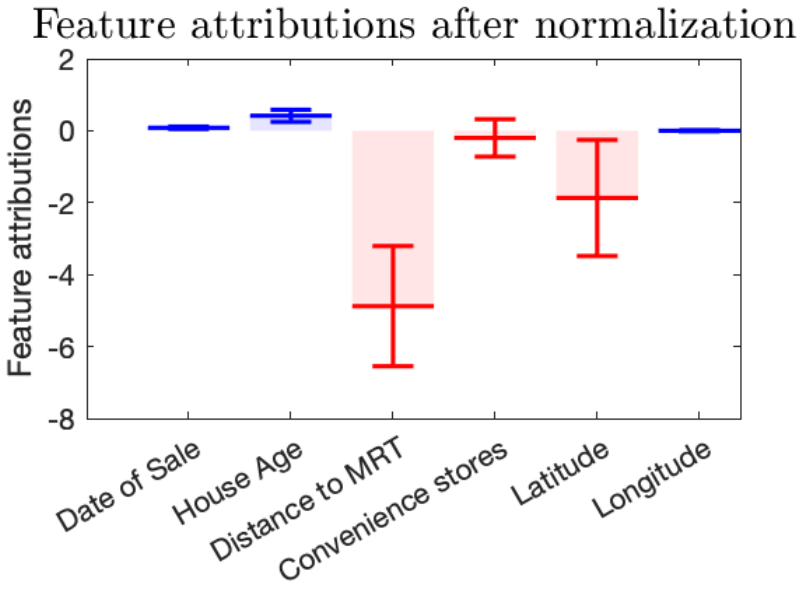


Predictions using the Taipei Housing Data









Random Feature GPs

- Random feature GPs use the spectral representation of the kernel function to approximate the kernel.

$$k(\mathbf{x}, \mathbf{x}') \approx \phi(\mathbf{x})^\top \phi(\mathbf{x}') \qquad \phi(\mathbf{x}) = \begin{bmatrix} \sin(\mathbf{v}_1^\top \mathbf{x}) \\ \cos(\mathbf{v}_1^\top \mathbf{x}) \\ \vdots \\ \sin(\mathbf{v}_M^\top \mathbf{x}) \\ \cos(\mathbf{v}_M^\top \mathbf{x}) \end{bmatrix}$$

  - The frequency vectors v_m are randomly sampled from the power spectral density of the kernel.
  - This representation yields equations which are more tractable during the training of a GPR model.

$$\begin{aligned} \mathbb{E}(F(\mathbf{x})|\mathcal{D}) &= m(\mathbf{x}) + \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{m}) \\ &\approx m(\mathbf{x}) + \phi(\mathbf{x})^\top \mathbf{\Phi} (\mathbf{\Phi}^\top \mathbf{\Phi} + \sigma_n^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{m}) \end{aligned}$$

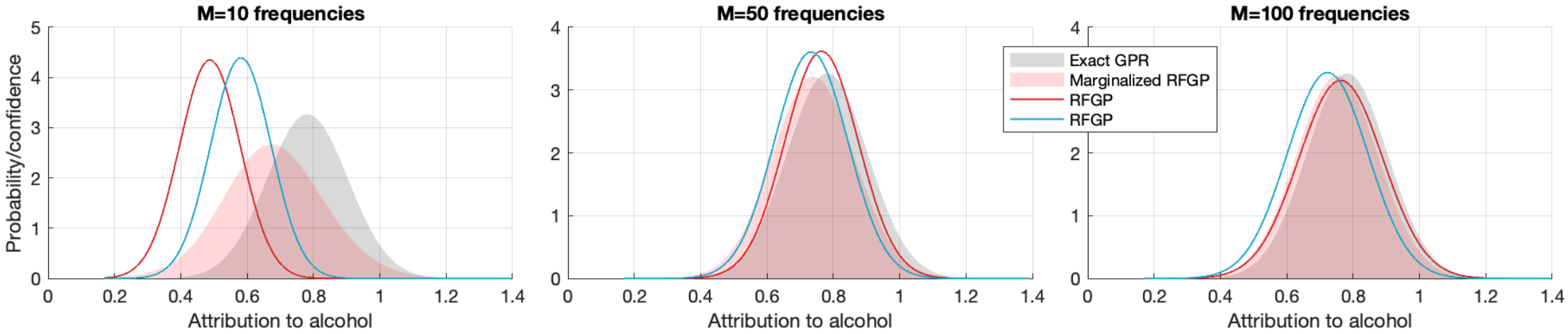
Random Feature GPs

- Random feature GPs admit simple to evaluate attributions.
- Since RFGPs approximate GPR, the RFGP attributions should approximate the GPR attributions.

$$\mathbb{E}(attr_i(\mathbf{x})) = (x_i - \tilde{x}_i)\boldsymbol{\zeta}(\mathbf{x})\mathbf{A}^{-1}\boldsymbol{\Phi}\mathbf{y},$$
$$\mathrm{Var}(attr_i(\mathbf{x})) = (x_i - \tilde{x}_i)^2\sigma_n^2\boldsymbol{\zeta}(\mathbf{x})^\top\mathbf{A}^{-1}\boldsymbol{\zeta}(\mathbf{x}),$$

$$\boldsymbol{\zeta}(\mathbf{x}) = \begin{bmatrix} \frac{v_{1,i}}{\mathbf{v}_1^\top(\mathbf{x}-\tilde{\mathbf{x}})}(\cos(\mathbf{v}_1^\top\mathbf{x}) - \cos(\mathbf{v}_1^\top\tilde{\mathbf{x}})) \\ \frac{v_{1,i}}{\mathbf{v}_1^\top(\mathbf{x}-\tilde{\mathbf{x}})}(\sin(\mathbf{v}_1^\top\mathbf{x}) - \sin(\mathbf{v}_1^\top\tilde{\mathbf{x}})) \\ \vdots \\ \frac{v_{M,i}}{\mathbf{v}_M^\top(\mathbf{x}-\tilde{\mathbf{x}})}(\cos(\mathbf{v}_M^\top\mathbf{x}) - \cos(\mathbf{v}_M^\top\tilde{\mathbf{x}})) \\ \frac{v_{M,i}}{\mathbf{v}_M^\top(\mathbf{x}-\tilde{\mathbf{x}})}(\sin(\mathbf{v}_M^\top\mathbf{x}) - \sin(\mathbf{v}_M^\top\tilde{\mathbf{x}})) \end{bmatrix} \in \mathbb{R}^{2M\times 1},$$

$$\mathbf{A} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \frac{M\sigma_n^2}{k(0,0)}\mathbf{I}_{2M}$$



Conclusions

- We developed a bridge between the theory of feature attribution and GPR.
- By representing attributions as distributions, we can quantify a model's confidence in its feature attributions.
- When the selected GPR model admits analytically tractable attributions, we mitigate the need for computationally expensive or otherwise inaccurate approximations.
- While non-parametric methods like GPR, at face value, offer a non-interpretable approach to ML, our study of feature attribution may provide more trust in GPR models.
- Furthermore, these results will extend the theoretical and practical scope of GPR models to more XAI domains.