Gaussian Processes

PETR CEZNER

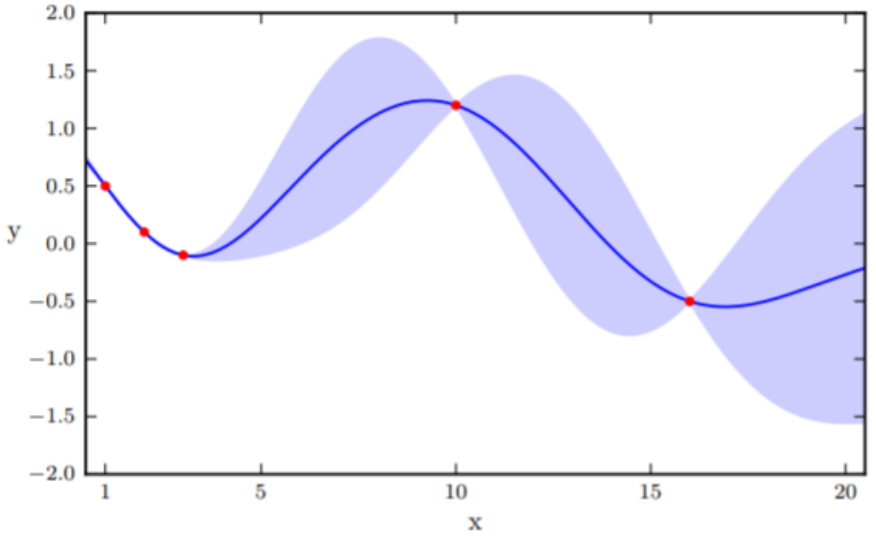
Outline

- Motivation
- Multivariate Gaussian distributions
- Gaussian Processes
  - Regression
  - Prediction

Motivation - Nonlinear regression

Yuge Shi

Motivation - Nonlinear regression



Yuge Shi

Multivariate Gaussian distributions

- Guassian distribution is a building block of Guassian Processes
- Multivariate Gaussian distribution is defined by: $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

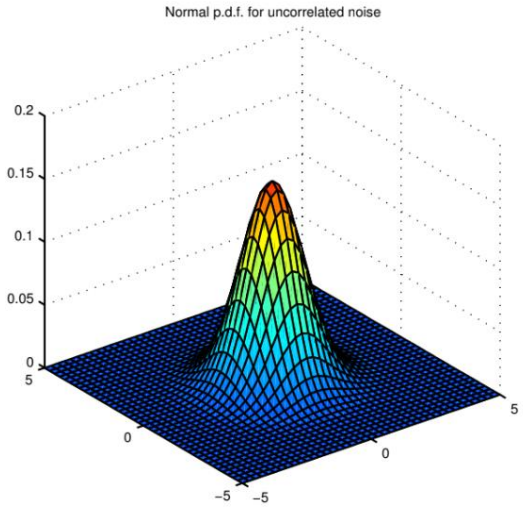
$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma)$$

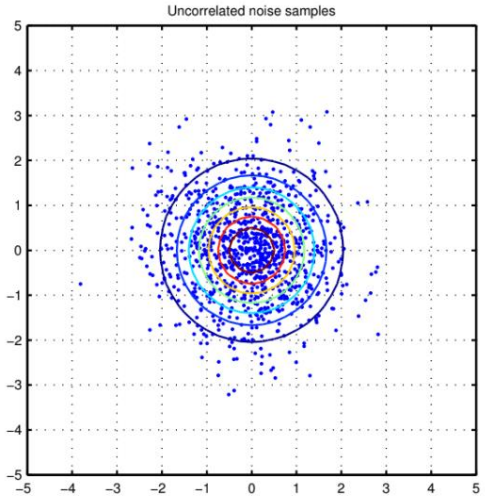
$$\Sigma = \mathrm{Cov}(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)^T\right]$$

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\lambda = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix} \qquad \boldsymbol{V} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$

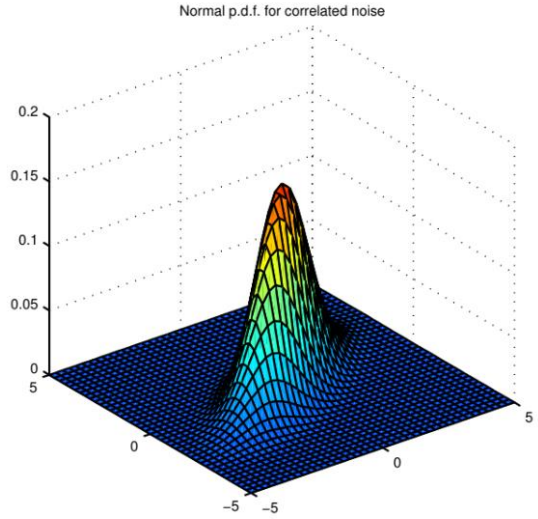


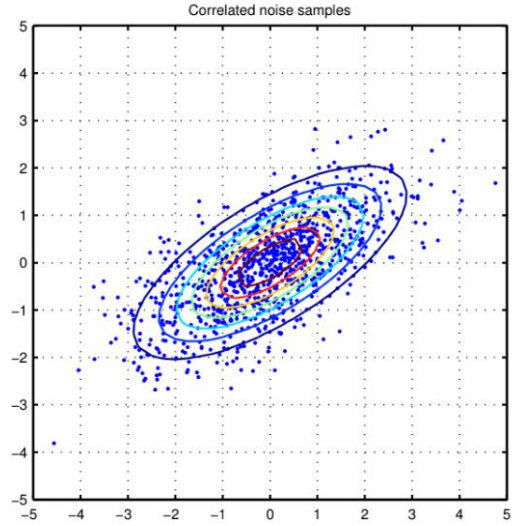


$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

$$\lambda = \begin{bmatrix} 2.6 \\ 0.4 \end{bmatrix} \qquad \boldsymbol{V} = \begin{bmatrix} 0.8 & 0.5 \\ 0.5 & -0.8 \end{bmatrix}$$





Multivariate Gaussian distributions - cont.

- Gaussian distribution has a nice property of being closed under conditioning and marginalization
- **Marginalization:**
  - We can extract partial information from multivariate probability distribution

$$P_{X,Y} = \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma) = \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} \, \Sigma_{XY} \\ \Sigma_{YX} \, \Sigma_{YY} \end{bmatrix} \right)$$

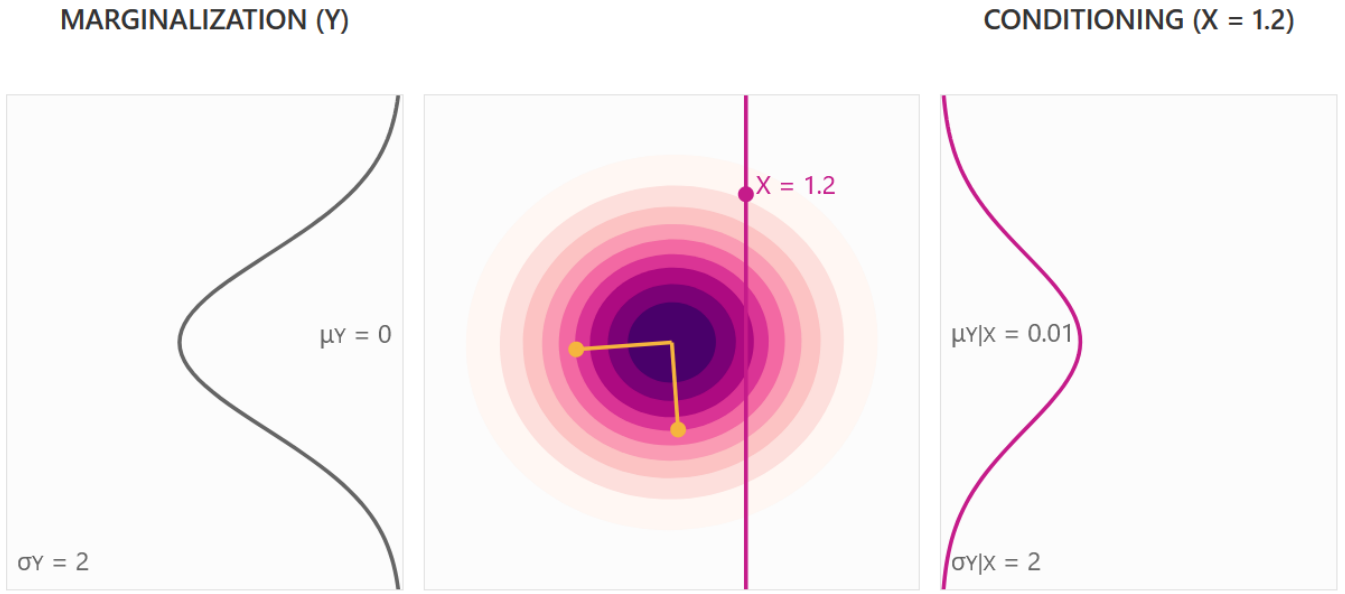
$$X \sim \mathcal{N}(\mu_X, \Sigma_{XX})$$
$$Y \sim \mathcal{N}(\mu_Y, \Sigma_{YY})$$

  - Each partition ***X*** and ***Y*** only depends on its corresponding entries in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

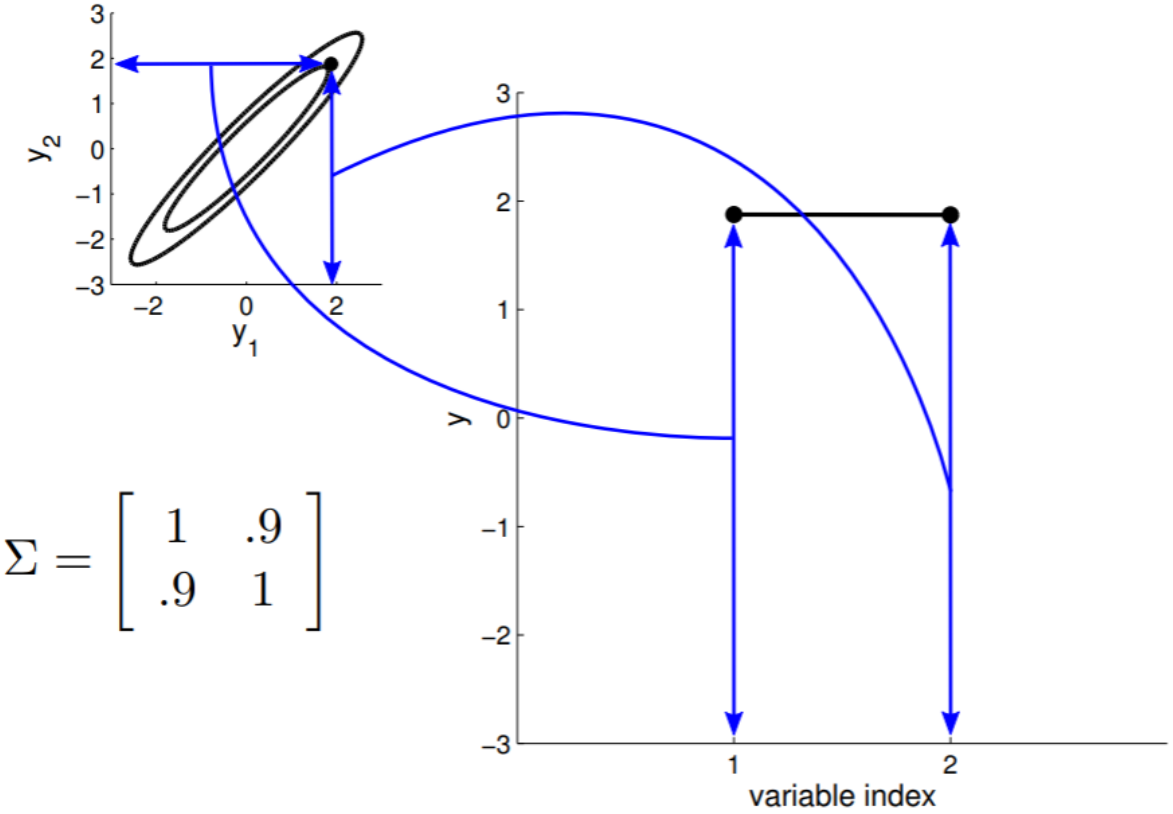
Conditioning

- It determinate the probability of one variable depending on another variable
- As with Marginalization conditioning leads to Gaussian distribution

$$X|Y \sim \mathcal{N}(\,\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y),\; \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\,)$$
$$Y|X \sim \mathcal{N}(\,\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X),\; \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\,)$$

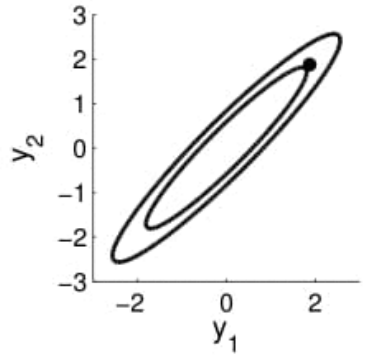


Görtler, et al

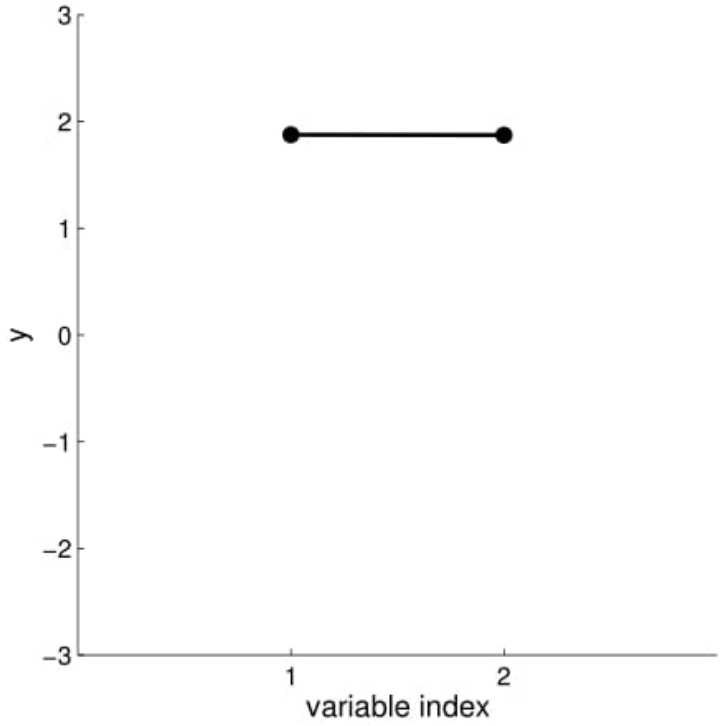


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

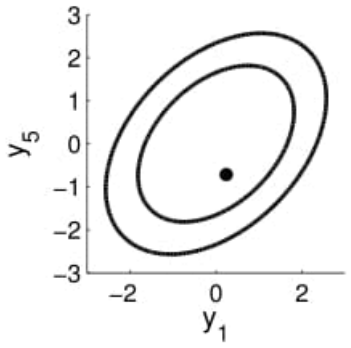
Yuge Shi



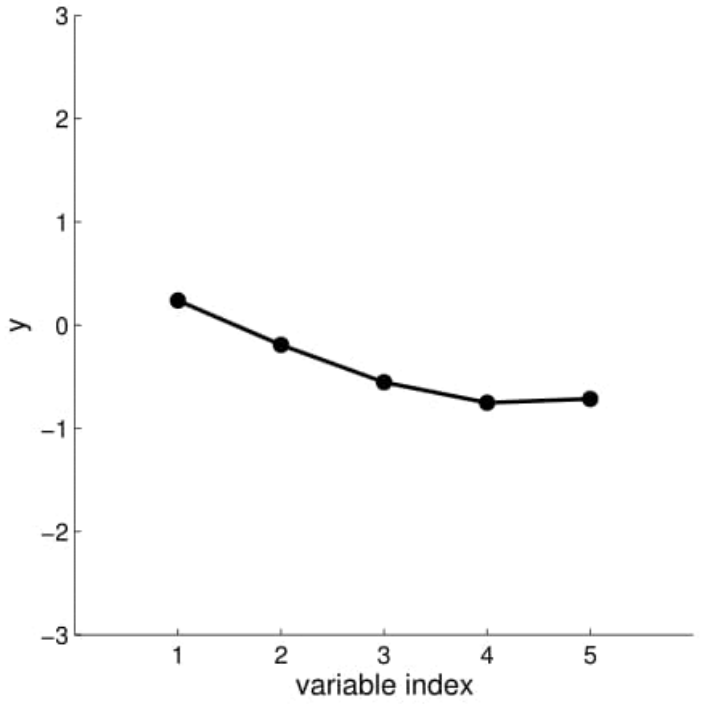
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



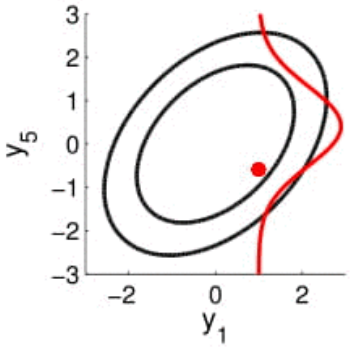
Yuge Shi



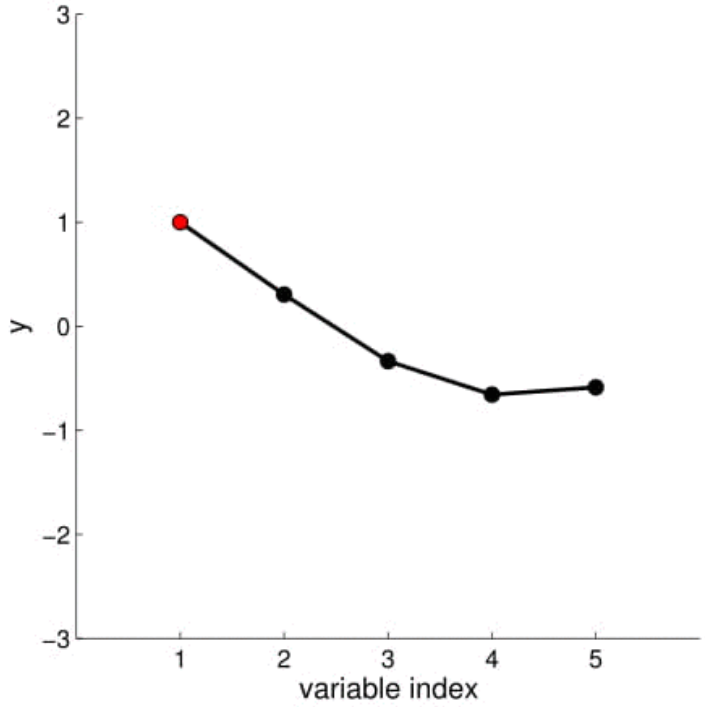
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$



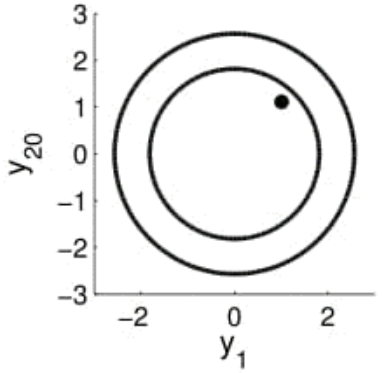
Yuge Shi



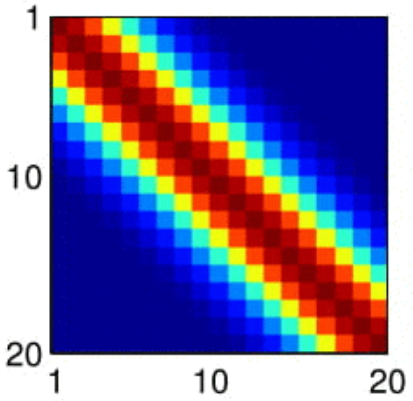
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

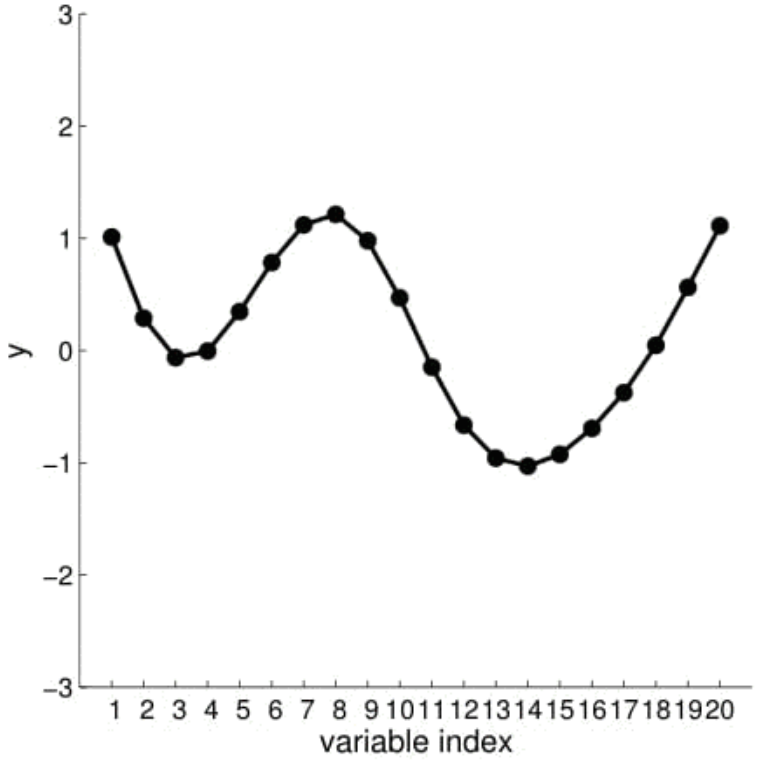


Yuge Shi

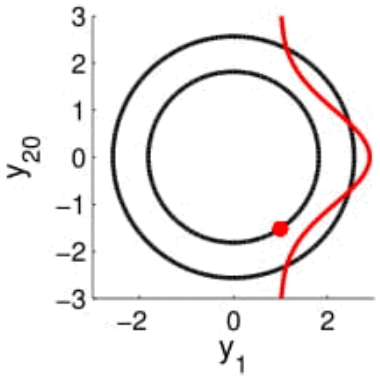


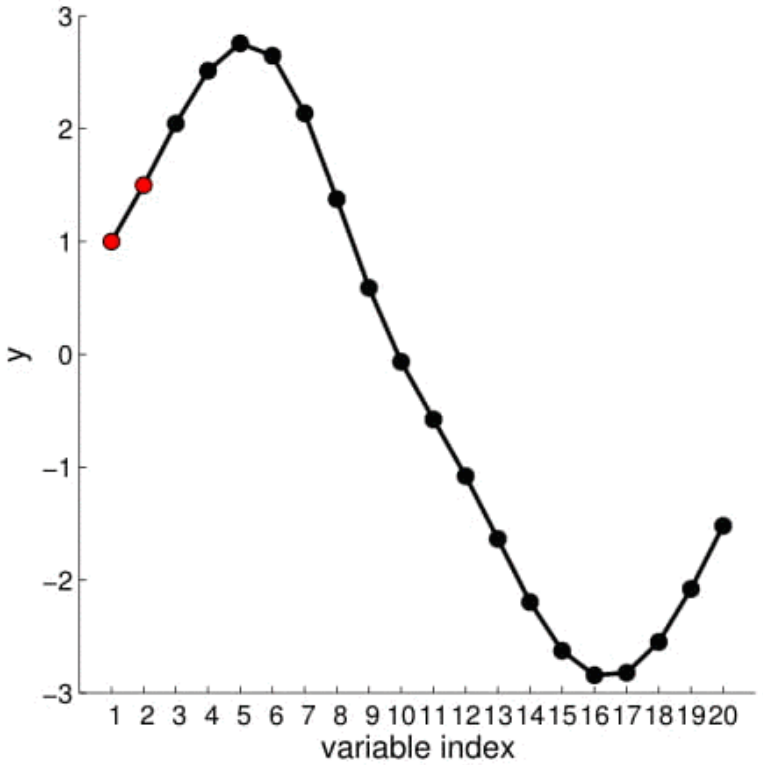
$\Sigma =$



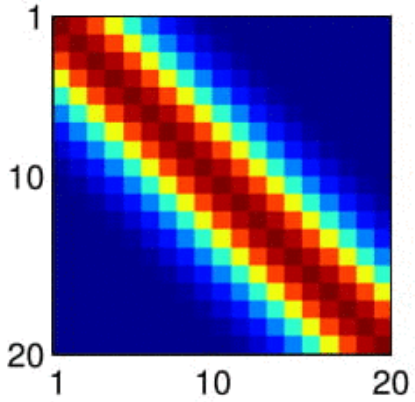


Yuge Shi



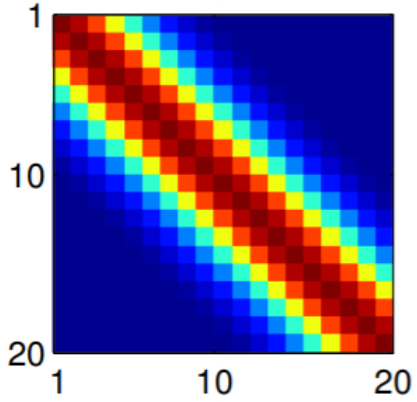


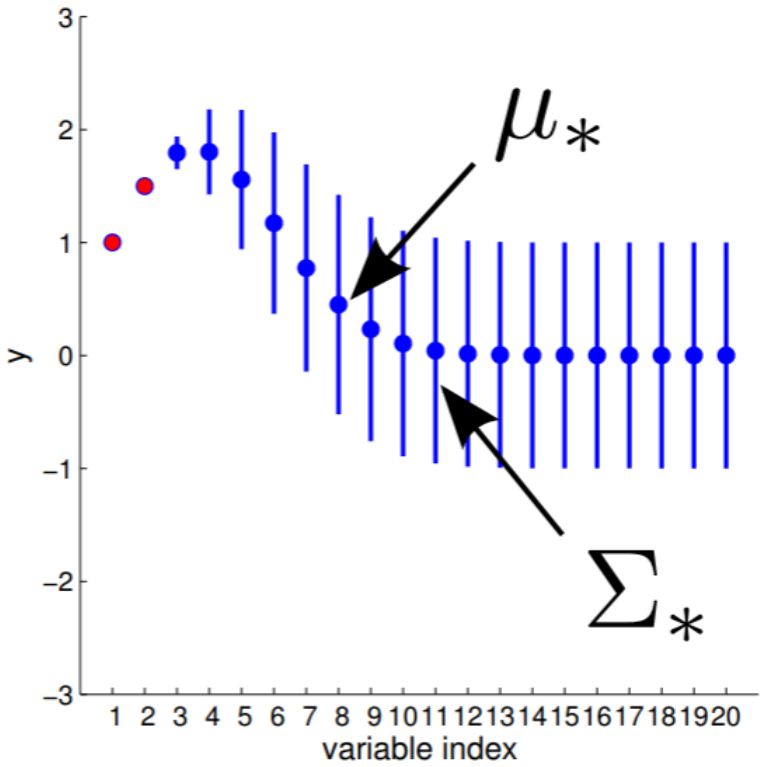
$$\Sigma =$$



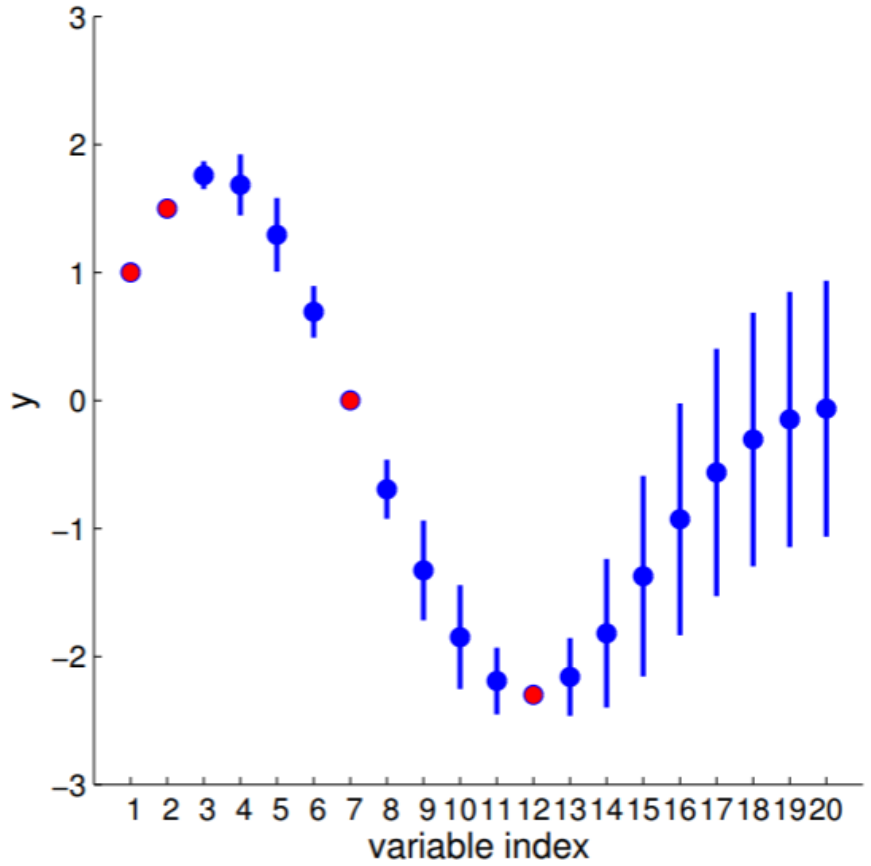
Yuge Shi

$\Sigma =$





Yuge Shi


3
2
1
0
−1
−2
−3
y
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
variable index

Yuge Shi

Intuition behind GP

- There is a problem with above mentioned approach for nonlinear regression
- The solution lies in how the covariance matrix is generated

$$\Sigma(\mathrm{x}_1, \mathrm{x}_2) = \mathrm{K}(\mathrm{x}_1, \mathrm{x}_2) + \mathrm{I}\sigma_{\mathrm{y}}^2$$

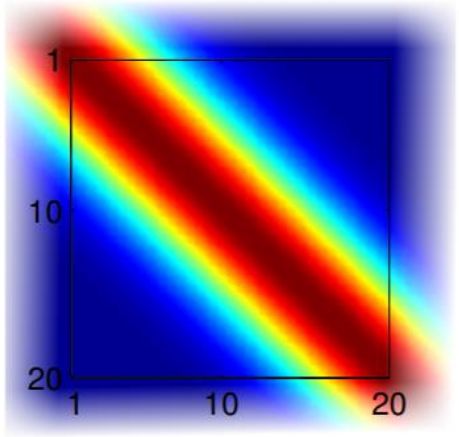
$$\mathrm{K}(\mathrm{x}_1, \mathrm{x}_2) = \sigma^2 \mathrm{e}^{-\frac{1}{2\mathrm{l}^2}(\mathrm{x}_1 - \mathrm{x}_2)^2}$$

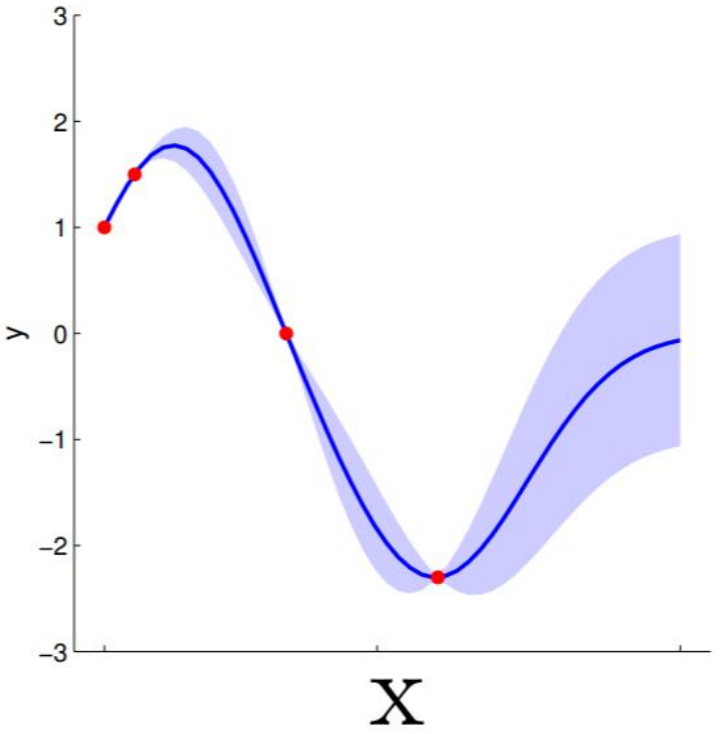
- The RBF kernel ensures the smoothness of the covariance matrix
- We can calculate convariance matrix for any real-values x_1 and x_2 by simply plugging them in
- The real-values x_s effectively result in an infinite-dimensional Gaussian defined by $\boldsymbol{\Sigma}$

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

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

$\Sigma =$





Yuge Shi

Gaussian Processes

- *"A Gaussian process is a collection of random variables, any finite number of which have consistent Gaussian distributions"*
- GP is defined by:
  - Mean function $m(x)$
  - Convariance function $K(x, x')$

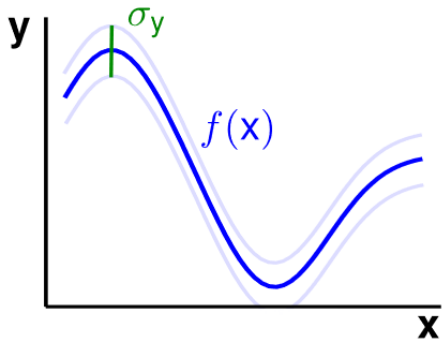
$$f(x) \sim \mathcal{GP}(m(x), K(x, x'))$$

Gaussian Processes are non-parametric

- Gaussian processes are non-parametric
  - non-parametric = model with infinite number of parameters

$$y(x) = f(x) + \epsilon\sigma_y$$

$$p(\epsilon) = \mathcal{N}(0, 1)$$



- We can place a Gaussian process prior over the nonlinear function (parametric function above is drawn from the Gaussian process)

$$p(f(x) \mid \theta) = \mathcal{GP}(0, K(x, x'))$$

$$K(x, x') = \sigma^2 \exp(-\frac{1}{2l^2}(x - x')^2)$$

- We can add Gaussian noise σ_y to the model, since the sum of Gaussian variables is also Guassian

$$p(f(x) \mid \theta) = \mathcal{GP}(0, K(x, x') + I\sigma_y^2)$$

RBF Kernel

Non-parametric (∞-parametric)

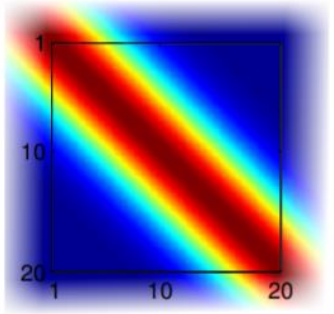
$p(\mathrm{y}|\theta) = \mathcal{N}(0, \Sigma)$

$\Sigma(\mathrm{x}_1, \mathrm{x}_2) = \mathrm{K}(\mathrm{x}_1, \mathrm{x}_2) + \mathrm{I}\sigma_\mathrm{y}^2$

$\mathrm{K}(\mathrm{x}_1, \mathrm{x}_2) = \sigma^2 \mathrm{e}^{-\frac{1}{2l^2}(\mathrm{x}_1 - \mathrm{x}_2)^2}$

vertical-scale horizontal-scale

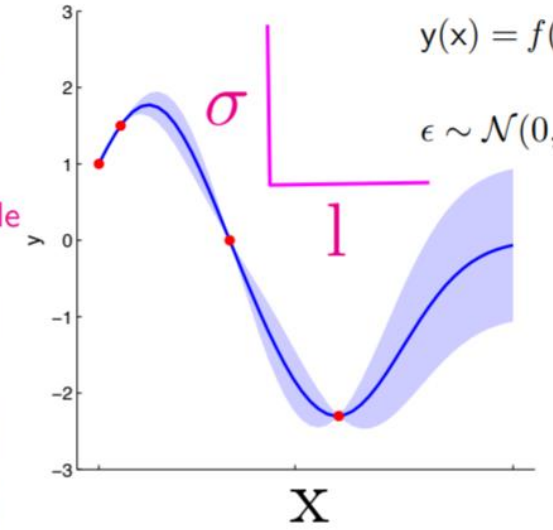
$\Sigma =$



Parametric model

$\mathrm{y}(\mathrm{x}) = f(\mathrm{x}; \theta) + \sigma_\mathrm{y}\epsilon$

$\epsilon \sim \mathcal{N}(0, 1)$

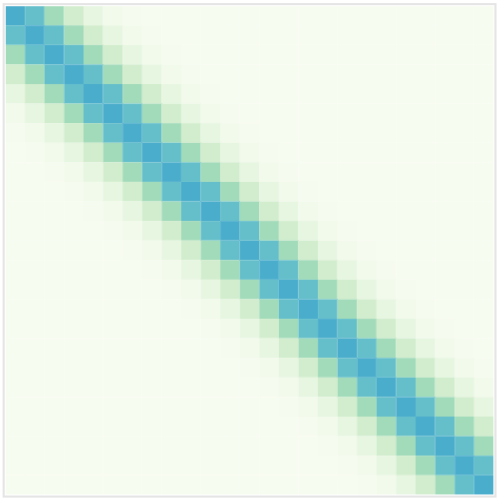


$$\arg\max_{l,\sigma^2} \log p(\mathbf{y}\,|\theta)$$

Yuge Shi

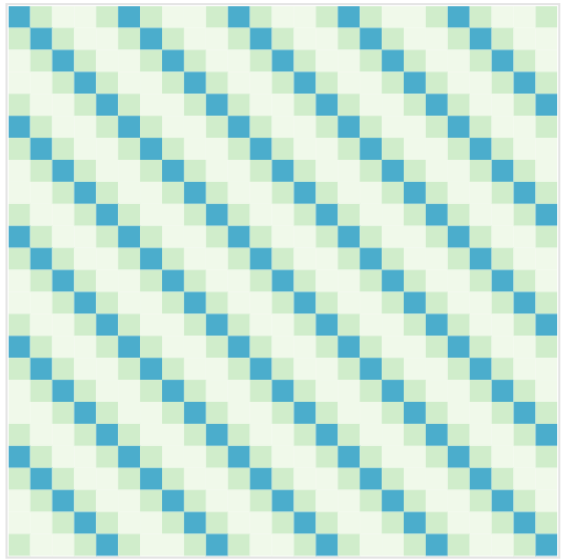
RBF KERNEL

$$\sigma^2 \exp\left(-\frac{||t-t'||^2}{2l^2}\right)$$



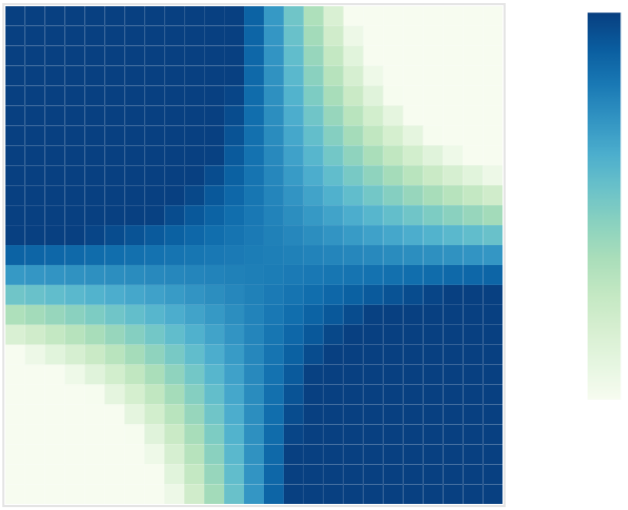
PERIODIC

$$\sigma^2 \exp\left(-\frac{2\sin^2(\pi|t-t'|/p)}{l^2}\right)$$



LINEAR

$$\sigma_b^2 + \sigma^2(t-c)(t'-c)$$



Görtler, et al

Computation

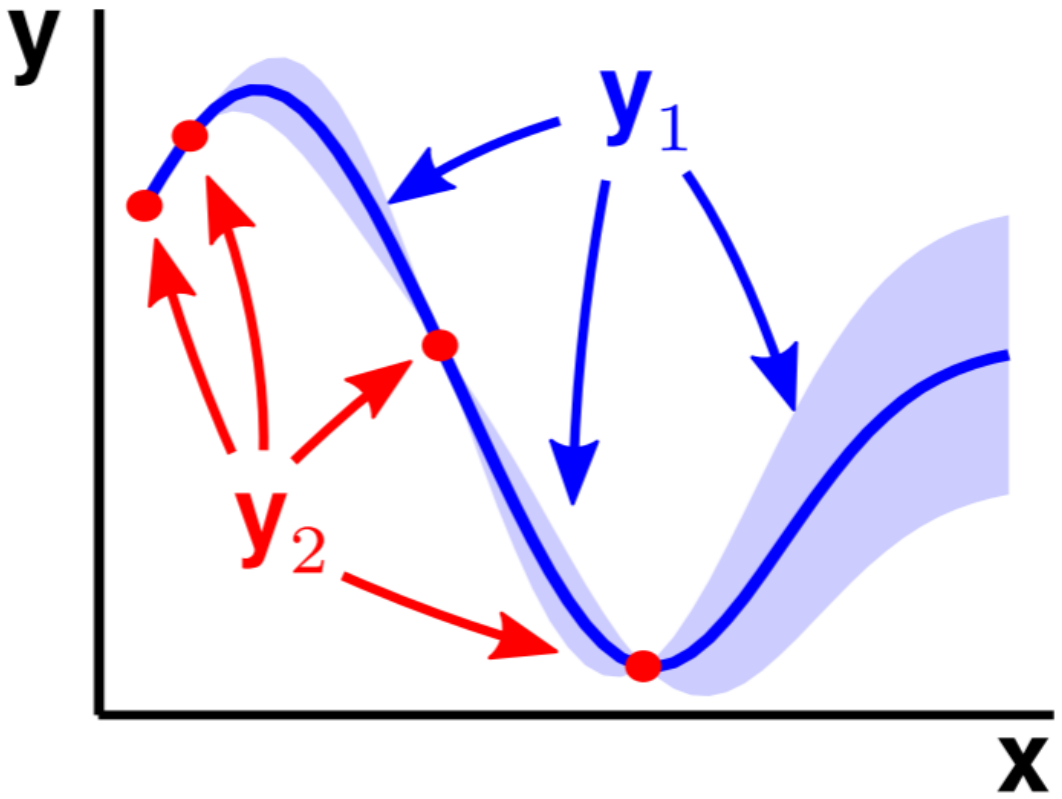
- How do we do computation with an infinite by infinite matrix?
- **Marginalization**

$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathsf{A} & \mathsf{B} \\ \mathsf{B}^\mathsf{T} & \mathsf{C} \end{bmatrix}\right)$$

- We can easily compute the probability of y_1 using the marginalisation property

$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathsf{A} & \mathsf{B} \\ \mathsf{B}^\mathsf{T} & \mathsf{C} \end{bmatrix}\right) \quad \Longrightarrow \quad p(\mathbf{y}_1) = \mathcal{N}(\mathbf{a}, \mathsf{A})$$

GP - Prediction



Yuge Shi

GP - Prediction

- To make predictions about y_1 given the observation y_2, we can use Bayes rules to calculate $p(y_1|y_2)$

$$p(\mathbf{y}_1|\mathbf{y}_2) = \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_2)}$$

- Because $p(y_1)$, $p(y_2)$ and $p(y_1, y_2)$ are all Gaussians, $p(y_1|y_2)$ is also Gaussian

$$p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\underbrace{\mathbf{a} + \mathrm{B}\mathrm{C}^{-1}(\mathbf{y}_2 - \mathbf{b})}_{\text{predictive mean } \mu}, \underbrace{\mathrm{A} - \mathrm{B}\mathrm{C}^{-1}\mathrm{B}^{\mathsf{T}}}_{\substack{\text{predictive} \\ \text{covariance } \Sigma}})$$

Yuge Shi

GP - Prediction

$$\implies p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\mathbf{a} + \mathrm{B}\mathrm{C}^{-1}(\mathbf{y}_2 - \mathbf{b}), \mathrm{A} - \mathrm{B}\mathrm{C}^{-1}\mathrm{B}^\intercal)$$

predictive mean

$$\mu_{\mathbf{y}_1|\mathbf{y}_2} = \mathbf{a} + \mathrm{B}\mathrm{C}^{-1}(\mathbf{y}_2 - \mathbf{b})$$
$$= \mathrm{B}\mathrm{C}^{-1}\mathbf{y}_2$$
$$= \mathrm{W}\mathbf{y}_2$$

linear in the data

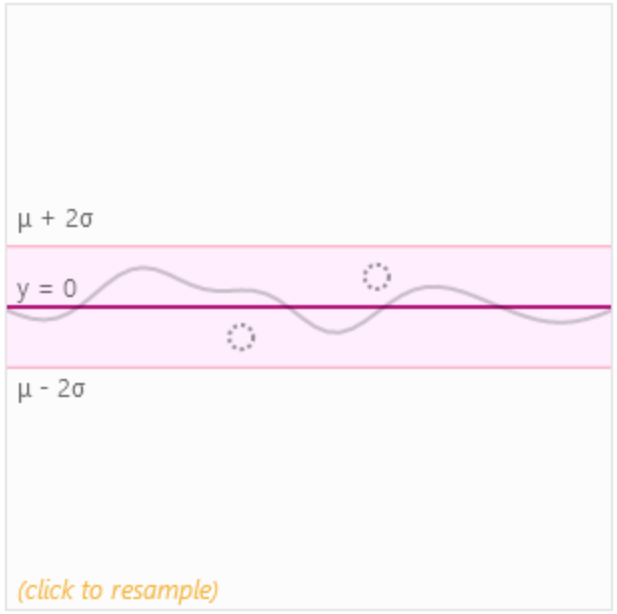
predictive covariance

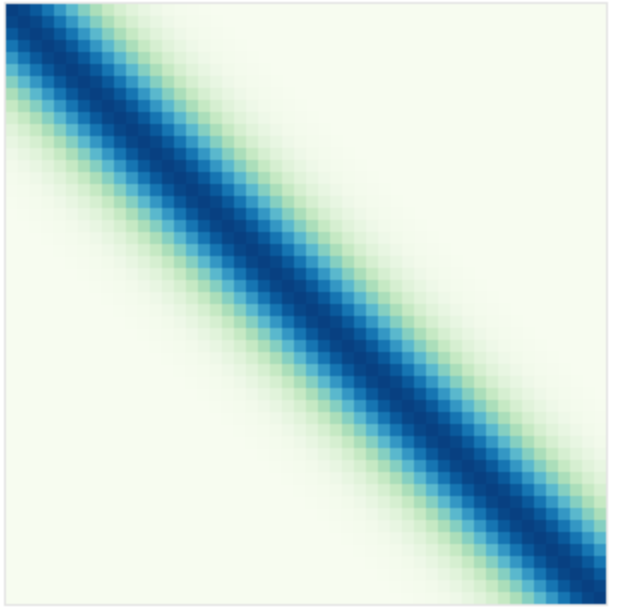
$$\Sigma_{\mathbf{y}_1|\mathbf{y}_2} = \mathrm{A} - \mathrm{B}\mathrm{C}^{-1}\mathrm{B}^\intercal$$

predictive uncertainty = prior uncertainty − reduction in uncertainty

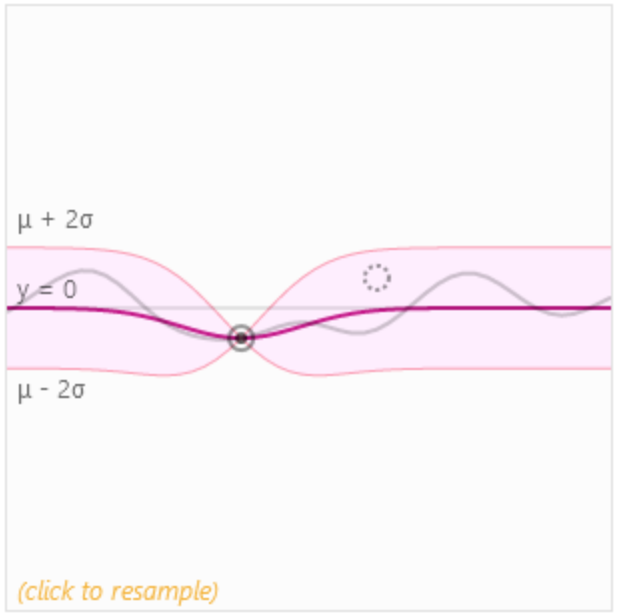
predictions more confident than prior

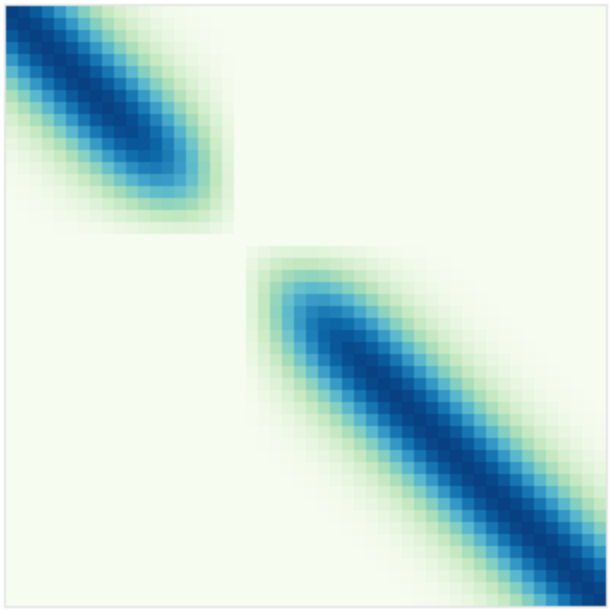
Yuge Shi



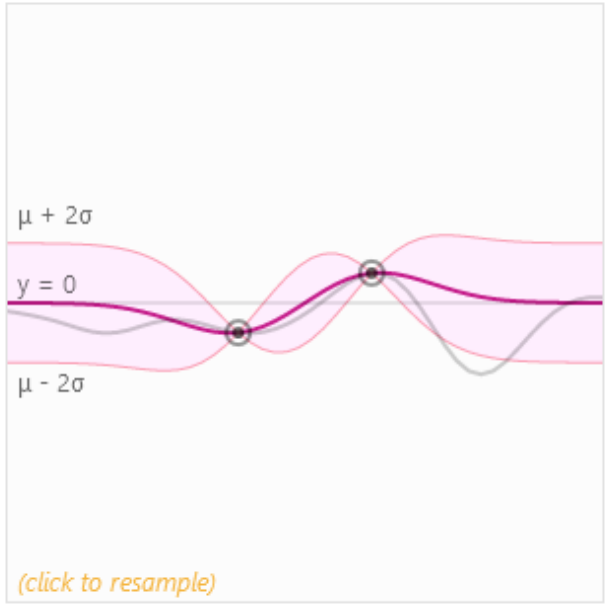


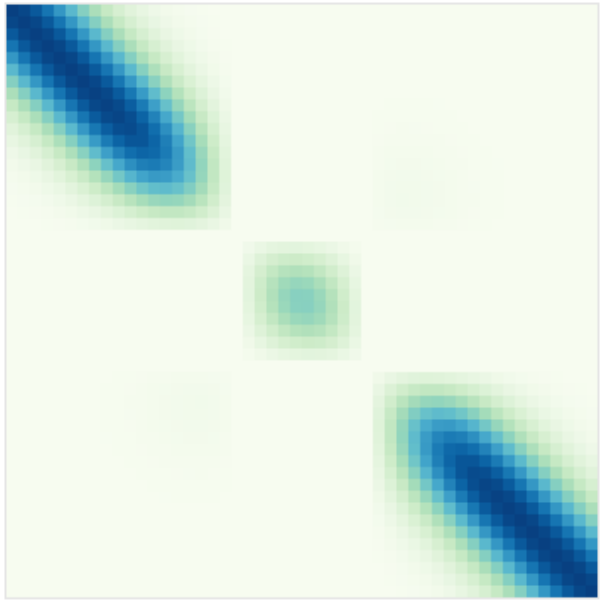
Görtler, et al





Görtler, et al





Görtler, et al

GP - Classification

- Covariance function (K) is called the latent function
- GP for classification also requires a *link* function
- The logistic function can be used

> *For the binary discriminative case one simple idea is to turn the output of a regression model into a class probability using a response function (the inverse of a link function), which "squashes" its argument, which can lie in the domain (−inf, inf), into the range [0, 1], guaranteeing a valid probabilistic interpretation.*
>
> Rasmussen, C. E.

- This is complicated, because now we're multiplying normal distribution with some non-normal distribution, so the resulting distribution is no longer normal. Due to that, marginalization (step in training on observations) is much harder to compute (approximations needed)

Sources

- Yuge Shi, *Gaussian Processes, not quite for dummies*, The Gradient, 2019
- Görtler, et al., *A Visual Exploration of Gaussian Processes*, Distill, 2019.
- Rasmussen, C. E., & Williams, C. K. (2008). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.
- Brownlee, J. (2020, August 02). Gaussian Processes for Classification With Python. Retrieved December 17, 2020, from *https://machinelearningmastery.com/gaussian-processes-for-classification-with-python/*
- Duvenaud, D. (n.d.). The Kernel Cookbook:. Retrieved December 17, 2020, from *https://www.cs.toronto.edu/~duvenaud/cookbook/*
- 1.7. Gaussian Processes¶. (n.d.). Retrieved December 17, 2020, from *https://scikit-learn.org/stable/modules/gaussian_process.html*