Bayesian Regression and Gaussian Processes

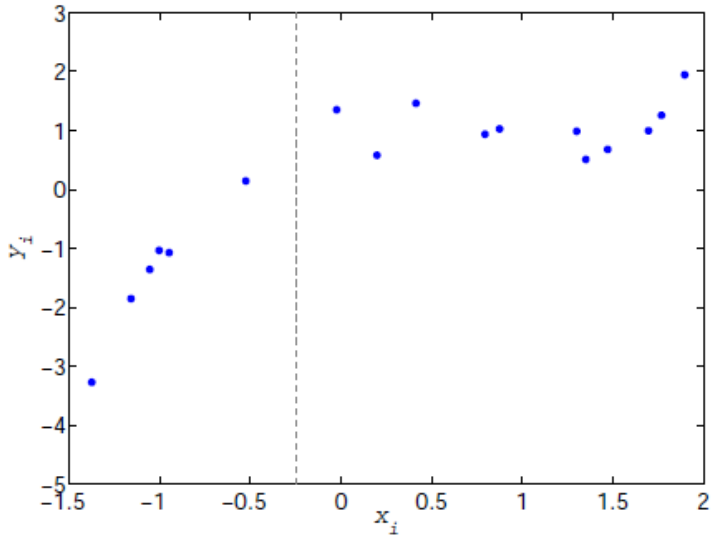
Sara Wade
University of Cambridge

Charles University
8-19 April 2013, Prague

References

- Rasmussen, C.E. and Williams, C.K.I. (2006). Gaussian process for machine learning. *MIT press*.
- Rasmussen, C.E. and Ghahramani, Z. (2013). Machine learning course. `http://mlg.eng.cam.ac.uk/teaching/4f13/1213`

How do we fit this dataset?



- Data: input $\mathbf{x} = (x_1, \ldots, x_n)^T$ for $x_i \in \mathcal{X}$ and output $\mathbf{y} = (y_1, \ldots, y_n)^T$ for $y_i \in \mathcal{Y}$, where y_i may be **continuous** (regression problem) or discrete (classification problem).
- Goal: to predict y_* for a new input x_* given the data $(\mathbf{x}, \mathbf{y})$.
- Model: the input-output mapping $f : \mathcal{X} \to \mathcal{Y}$ is contaminated by noise,

$$Y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathsf{N}(0, \sigma^2).$$

Assumptions on $f(x)$

- Classic approaches: represent $f(x)$ as linear combination of basis functions $\{\phi_m(x)\}$,

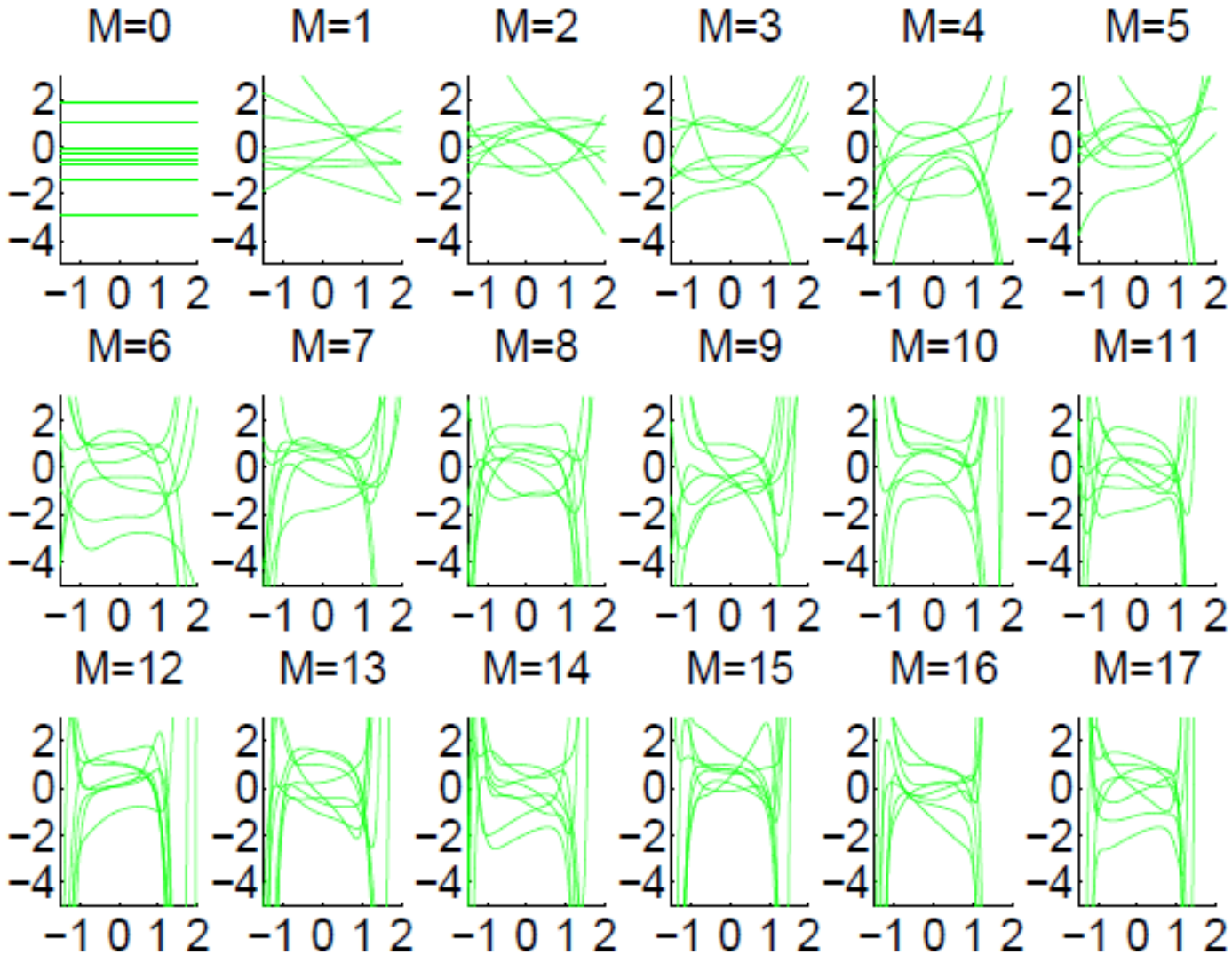
$$f(x) = \sum_{m=1}^{M} w_m \phi_m(x).$$

For example, the class of polynomials $\phi_m(x) = x^m$ and

$$f(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M.$$

- Our model parameters are the weights $w = (w_1, \ldots, w_M)^T$ and the structure of the model is defined by M and $\{\phi_m\}$.

Examples of polynomials as M and w vary



Bayesian prior (M fixed)

- What values of w do we believe are probable?

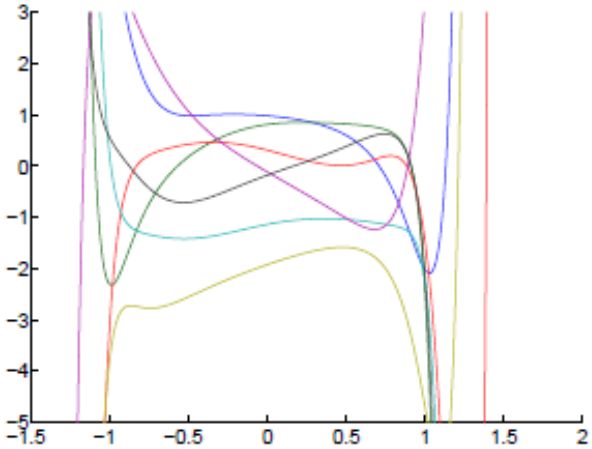
$\rightarrow$ This is expressed through the prior on w, ex.

$$p(w) = (2\pi)^{-M/2}|\Sigma_0|^{-1/2}\exp\left(-\frac{1}{2}(w-\mu_0)^T\Sigma_0^{-1}(w-\mu_0)\right),$$

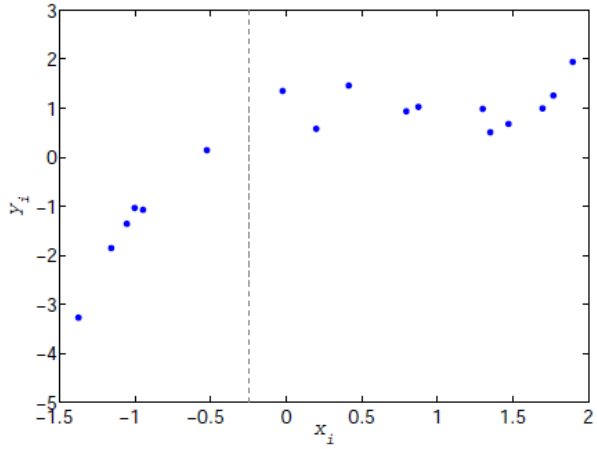
i.e. $w \sim \mathsf{N}_M(\mu_0, \Sigma_0)$.

- The normal prior is specified by
  1) prior guess $\mu_0 = \mathsf{E}[w]$ and
  2) variability around μ_0, $\Sigma_0 = \mathsf{E}[(w-\mu_0)^2]$.

Prior Samples



(a) Prior Samples

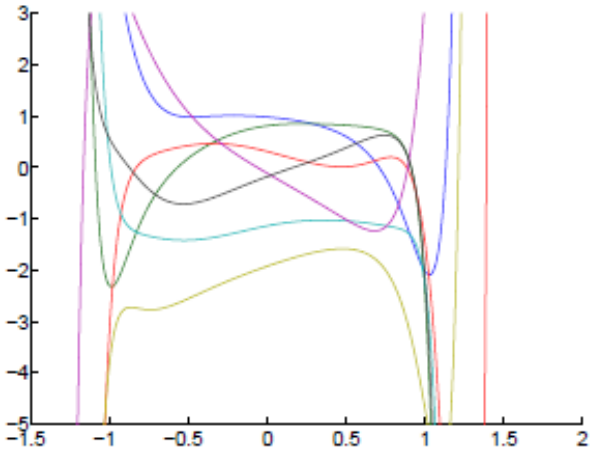


(b) Data

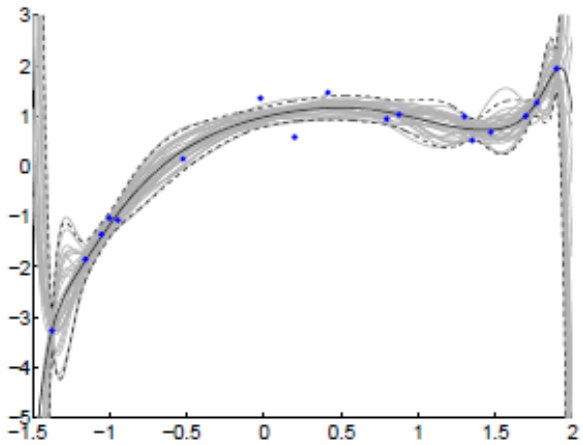
How do we sample from the prior?
Imagine $\phi_m(x) = x^m$, $M = 17$ and select prior parameters μ_0 and Σ_0.

- Define a grid of x values.
- Compute $\phi_m(x)$ for $m = 0, \ldots, M$ and x in the grid.
- Sample $w \sim \mathsf{N}_M(\mu_0, \Sigma_0)$.
- Compute $f(x) = \sum_{m=0}^{M} w_m \phi_m(x)$ for x in the grid.

Posterior



(c) Prior Samples



(d) Posterior Samples

$$\text{Posterior: } p(w|\mathbf{y},\mathbf{x}) = \frac{p(w)p(\mathbf{y}|w,\mathbf{x})}{p(\mathbf{y}|\mathbf{x})}.$$

- The prior density is adjusted by the likelihood $p(\mathbf{y}|w,\mathbf{x})$, which measures the closeness between the data and the function defined by w.
- We don't pick a single w but many, weighted by the posterior density.

Posterior

The posterior is easily computed if:

- the likelihood is Gaussian, $Y_i|w, x_i \overset{ind}{\sim} \mathsf{N}(f(x_i), \sigma^2)$,
- the prior of w is Gaussian, $w \sim \mathsf{N}_M(\mu_0, \Sigma_0)$,

then, the posterior is Gaussian

$$w|\mathbf{y}, \mathbf{x} \sim \mathsf{N}_M(\hat{w}, \hat{\Sigma}),$$

where

$$\hat{\Sigma} = (\Sigma_0^{-1} + \sigma^{-2}\Phi^T\Phi)^{-1},$$
$$\hat{w} = \hat{\Sigma}(\Sigma_0^{-1}\mu_0 + \sigma^{-2}\Phi^T y),$$

and Φ is the n by M matrix with elements $\phi_m(x_i)$.

Note: the Gaussian prior is the *conjugate* prior.

Posterior derivation

$$p(w|\mathbf{y},\mathbf{x}) \propto p(w)p(\mathbf{y}|\mathbf{x},w)$$

$$\propto \exp\left(-\frac{1}{2}(w-\mu_0)^T\Sigma_0^{-1}(w-\mu_0)\right)\exp\left(-\frac{1}{2}\sigma^{-2}\underbrace{(\mathbf{y}-\Phi w)^T(\mathbf{y}-\Phi w)}_{\mathbf{y}^T\mathbf{y}+w^T\Phi^T\Phi w-2w^T\Phi^T\mathbf{y}}\right)$$

$$\propto \exp\left(-\frac{1}{2}(w^T\Sigma_0^{-1}w + w^T(\sigma^{-2}\Phi^T\Phi)w - 2w^T(\Sigma_0^{-1}\mu_0) - 2w^T(\sigma^{-2}\Phi^T\mathbf{y})\right)$$

$$\propto \exp\left(-\frac{1}{2}\underbrace{(w^T(\Sigma_0^{-1}+\sigma^{-2}\Phi^T\Phi)w - 2w^T(\Sigma_0^{-1}\mu_0+\sigma^{-2}\Phi^T\mathbf{y})}_{\text{complete the square}}\right)$$

$$\propto \exp\left(-\frac{1}{2}(w-\hat{w})^T\hat{\Sigma}^{-1}(w-\hat{w})\right)$$

$$\Rightarrow w|\mathbf{y},\mathbf{x} \sim \mathsf{N}_M(\hat{w},\hat{\Sigma}).$$

Predictive distribution

- The regression function at a new value of the input x_* is $f(x_*) = w^T \phi(x_*)$, where $\phi(x_*) = (\phi_1(x_*), \ldots, \phi_M(x_*))^T$.
- From properties of the Gaussian distribution:

$$f(x_*)|\mathbf{y}, \mathbf{x}, x_* \sim \mathsf{N}(\hat{w}^T \phi(x_*), \phi(x_*)^T \hat{\Sigma} \phi(x_*)).$$

- The predictive density at y_* is

$$p(y_*|\mathbf{y}, \mathbf{x}, x_*) = \int p(y_*|w, x_*) p(w|\mathbf{y}, \mathbf{x}) dw.$$

$$\Rightarrow Y_*|\mathbf{y}, \mathbf{x}, x_* \sim \mathsf{N}(\hat{w}^T \phi(x_*), \phi(x_*)^T \hat{\Sigma} \phi(x_*) + \sigma^2).$$

- We average the prediction arising from each w with its posterior density.

Point estimation

How do we summarize the posterior or predictive?

$\rightarrow$ Define an appropriate loss function and find the estimator that minimizes the posterior expected loss.

Ex. let $f_* = f(x_*)$ and define $L(f_*, \hat{f}_*) = (f_* - \hat{f}_*)^2$. Then our point estimate $\hat{f}_*$ is

$$\begin{aligned}
\hat{f}_* &= \arg\min_{\tilde{f}_*} \mathsf{E}[(f_* - \tilde{f}_*)^2 | \mathbf{y}, \mathbf{x}] \\
&= \arg\min_{\tilde{f}_*} \mathsf{E}[(f_* - \hat{w}^T\phi(x_*) + \hat{w}^T\phi(x_*) - \tilde{f}_*)^2 | \mathbf{y}, \mathbf{x}] \\
&= \arg\min_{\tilde{f}_*} \mathsf{E}[(f_* - \hat{w}^T\phi(x_*))^2 | \mathbf{y}, \mathbf{x}] + \mathsf{E}[(\hat{w}^T\phi(x_*) - \tilde{f}_*)^2 | \mathbf{y}, \mathbf{x}] \\
&\qquad + 2\mathsf{E}[(f_* - \hat{w}^T\phi(x_*))(\hat{w}^T\phi(x_*) - \tilde{f}_*) | \mathbf{y}, \mathbf{x}] \\
&\Rightarrow \hat{f}_* = \hat{w}^T\phi(x_*).
\end{aligned}$$

Other examples include:

- $L(f_*, \hat{f}_*) = |f_* - \hat{f}_*| \rightarrow \hat{f}_*$ is the posterior median.
- $L(f_*, \hat{f}_*) = 1_{f_* \neq \hat{f}_*} \rightarrow \hat{f}_*$ is the posterior mode (MAPE).

Connections with penalized regression

Notice that:

$$\log(p(w|\mathbf{y},\mathbf{x})) \propto -\frac{1}{2}\sigma^{-2}(\mathbf{y}-\Phi w)^T(\mathbf{y}-\Phi w) - \frac{1}{2}(w-\mu_0)^T\Sigma_0^{-1}(w-\mu_0).$$

The MAPE corresponds to the estimator in penalized regression.

The prior corresponds to the penalization term (e.g. normal $\Leftrightarrow$ ridge, laplace $\Leftrightarrow$ lasso)

Gaussian Process

Model: $Y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$.
$\rightarrow$ with a *Gaussian process* prior, we can specify a prior directly on f.

A Gaussian process prior is a generalization of a multivariate Gaussian distribution on a random vector to an infinite collection of random variables.

Definition

A **Gaussian process** (GP) is an infinite collection of random variables, where any finite number have Gaussian distribution with consistent parameters.

Gaussian process

- A Gaussian distribution is fully specified by a mean vector μ and covariance matrix Σ;

$$(f_1, \dots f_n)^T \sim \mathsf{N}_n(\mu, \Sigma), \quad \text{indexed by } i = 1, \dots, n.$$

- A Gaussian process is fully specified by a mean function $\mu(x)$ and symmetric positive-semidefinite covariance function $k(x, x')$; for any $x_1, \dots, x_n$

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathsf{N}_n \left(\begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \right),$$

indexed by $x \in \mathcal{X}$, denoted by $f(x) \sim \mathsf{GP}(m(x), k(x, x'))$.

Properties of multivariate Gaussian

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \sim \mathsf{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{1\,1} & \Sigma_{1\,2} \\ \Sigma_{2\,1} & \Sigma_{2\,2} \end{bmatrix} \right)$$

1. Marginalization property: $f_i \sim \mathsf{N}(\mu_i, \Sigma_i)$.
2. Conditional property: $f_2|f_1 \sim \mathsf{N}(\beta_0 + \beta_1 f_1, \Sigma_{2|1})$,

$$\text{where} \quad \beta_0 = \mu_2 - \Sigma_{2\,1}\Sigma_{1\,1}^{-1}\mu_1; \quad \beta_1 = \Sigma_{2\,1}\Sigma_{1\,1}^{-1},$$

$$\text{and} \quad \Sigma_{2|1} = \Sigma_{2\,2} - \Sigma_{2\,1}\Sigma_{1\,1}^{-1}\Sigma_{1\,2}.$$

Existence of Gaussian process

Kolmogorov extension theorem

For any $x_1, \ldots, x_n$, $x_i \in \mathcal{X}$ and $n \in \mathbb{N}$, let $P_{x_1,\ldots,x_n}$ be a collection of probability measures on $\mathbb{R}^n$. If $(P_{x_1,\ldots,x_n})$ satisfy

1. for any permutation π and measurable sets $A_i \subseteq \mathbb{R}$,

$$P_{x_1,\ldots,x_n}(A_1 \times \ldots \times A_n) = P_{x_{\pi(1)},\ldots,x_{\pi(n)}}(A_{\pi(1)} \times \ldots \times A_{\pi(n)});$$

2. for any measurable sets $A_i \subseteq \mathbb{R}$,

$$P_{x_1,\ldots,x_{n-1}}(A_1 \times \ldots \times A_{n-1}) = P_{x_1,\ldots,x_n}(A_1 \times \ldots \times A_{n-1} \times \mathbb{R});$$

then there exists a stochastic process $(f(x))_{x\in\mathcal{X}}$ taking values in $\mathbb{R}^{\mathcal{X}}$ with marginals $P_{x_1,\ldots,x_n}$.

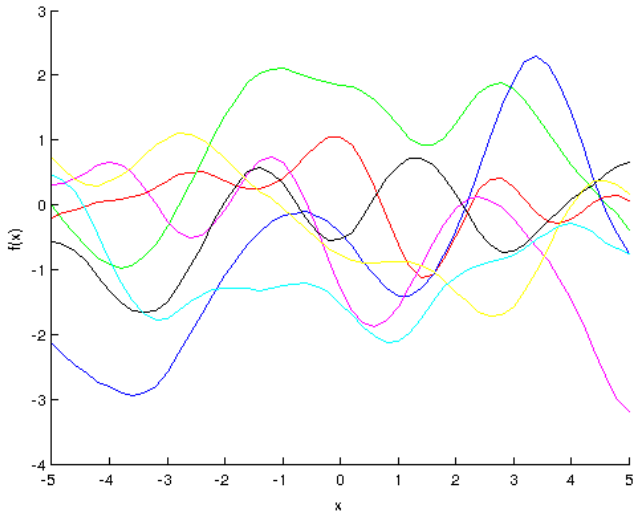
$\rightarrow$ Existence of a Gaussian process is obtained from Kolmogorov extension theorem and the marginalization property.

Prior Samples from a GP

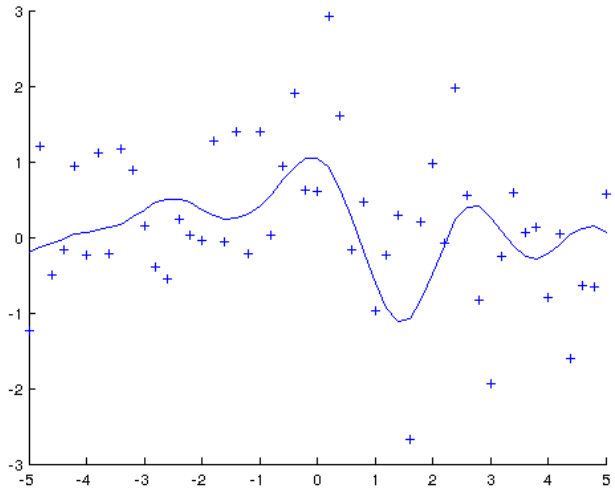
How do we sample from the prior?

- Specify input $\mathbf{x}$ (ex. grid).
- Compute $K(\mathbf{x}, \mathbf{x})$ and $m(\mathbf{x})$.
- Set $f(\mathbf{x}) = \text{chol}(K(\mathbf{x}, \mathbf{x}))^T \mathbf{z} + m(\mathbf{x})$, where $Z_i \overset{iid}{\sim} \text{N}(0, 1)$.

Ex. $m(x) = 0$ and $k(x, x') = \exp(-\frac{1}{2}(x - x')^2)$.



(e) Prior samples



(f) Prior sample with data

Posterior and Predictive

1. Gaussian likelihood: $Y_i|x_i, f \overset{ind}{\sim} \mathsf{N}(f(x_i), \sigma^2)$.
2. Zero-mean Gaussian process prior: $f(x) \sim \mathsf{GP}(0, k(x, x'))$.

$\rightarrow$ leads to a Gaussian process posterior,

$$f(x)|\mathbf{x}, \mathbf{y} \sim \mathsf{GP}(\hat{m}(x), \hat{k}(x, x')),$$

$$\text{where} \quad \hat{m}(x) = K(x, \mathbf{x}) \left(K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right)^{-1} \mathbf{y},$$

$$\text{and} \quad \hat{k}(x, x') = k(x, x') - K(x, \mathbf{x}) \left(K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right)^{-1} K(\mathbf{x}, x').$$

$\rightarrow$ leads to a Gaussian predictive distribution,

$$y_*|x_*\mathbf{x}, \mathbf{y} \sim \mathsf{N}(\hat{m}(x_*), \hat{k}(x_*, x_*) + \sigma^2),$$

Posterior and Predictive

Derivation: since $Y_i = f(x_i) + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} \mathsf{N}(0, \sigma^2)$, it follows that for any $\mathbf{x}_* = (x_{*1}, \ldots, x_{*k})^T$ with $k \in \mathbb{N}$,

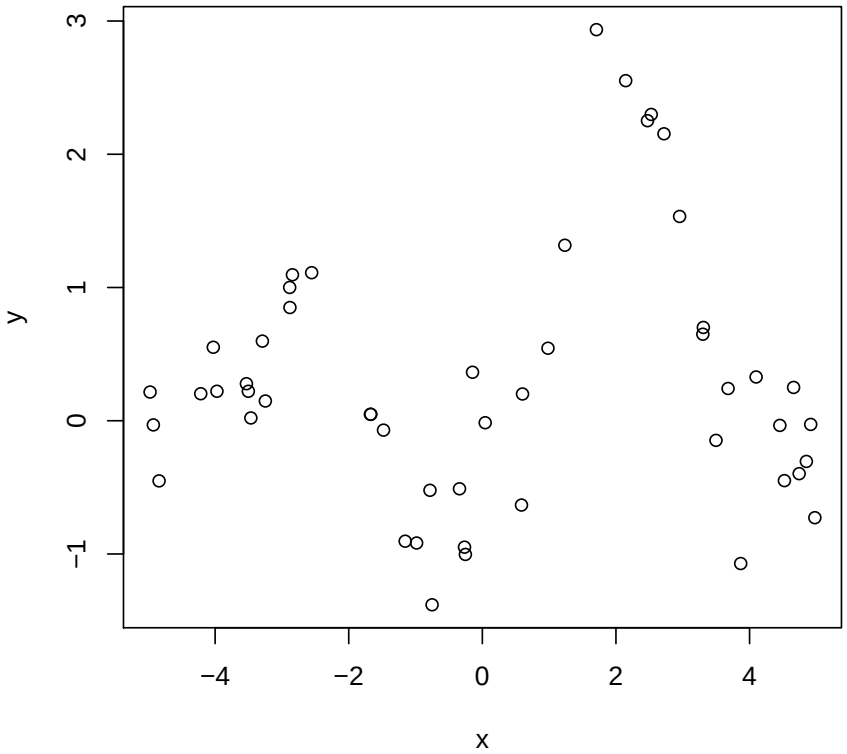
$$\begin{bmatrix} \mathbf{y} \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathsf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma^2 I & K(\mathbf{x}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{x}) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right).$$

From the conditional property of the multivariate Gaussian, the posterior distribution is obtained.
The predictive distribution is easily found since

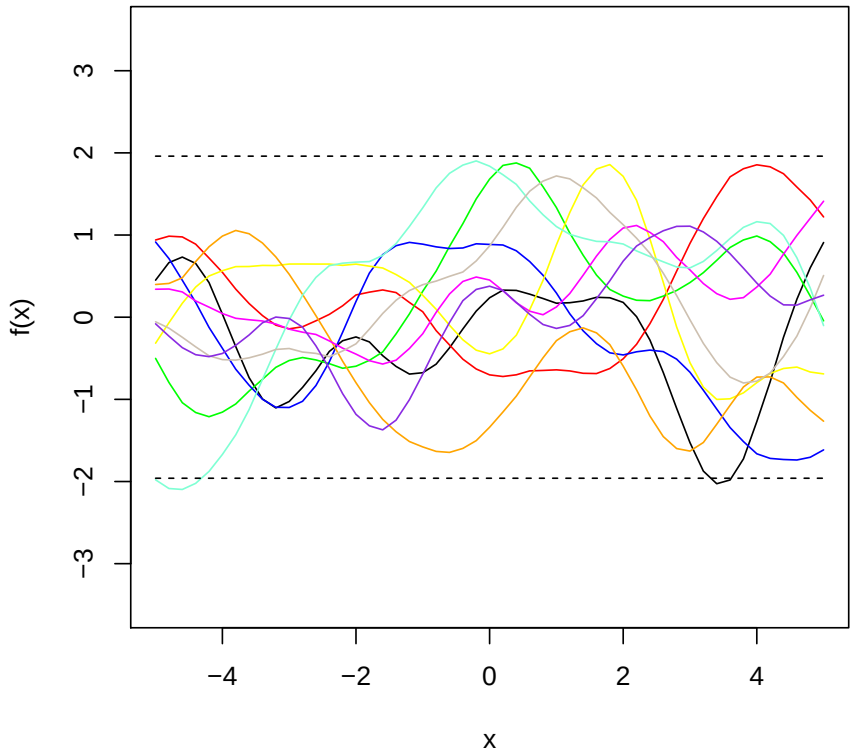
$$p(y_*|x_*, \mathbf{y}, \mathbf{x}) = \int \underbrace{p(y_*|f(x_*))}_{\mathsf{N}(y_*|f(x_*),\sigma^2)} \underbrace{p(f(x_*)|\mathbf{y}, \mathbf{x})}_{\mathsf{N}(f(x_*)|\hat{m}(x_*),\hat{k}(x_*,x_*))} df(x_*).$$

GP example

Ex. $m(x) = 0$, $k(x, x') = \exp(-\frac{1}{2}(x - x')^2)$, $\sigma^2 = 0.1$.



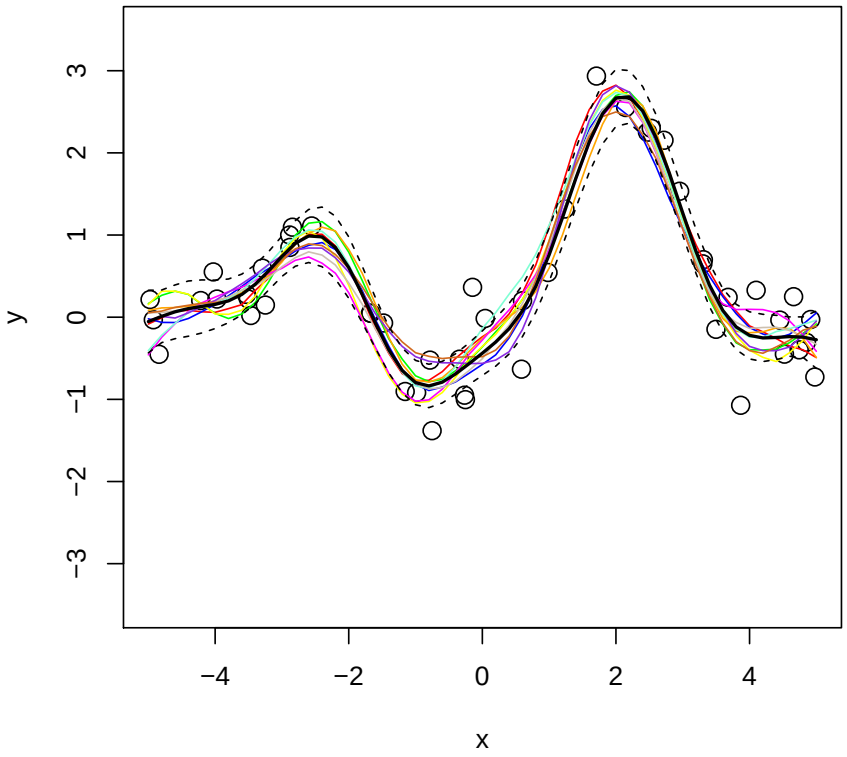
(g) Data



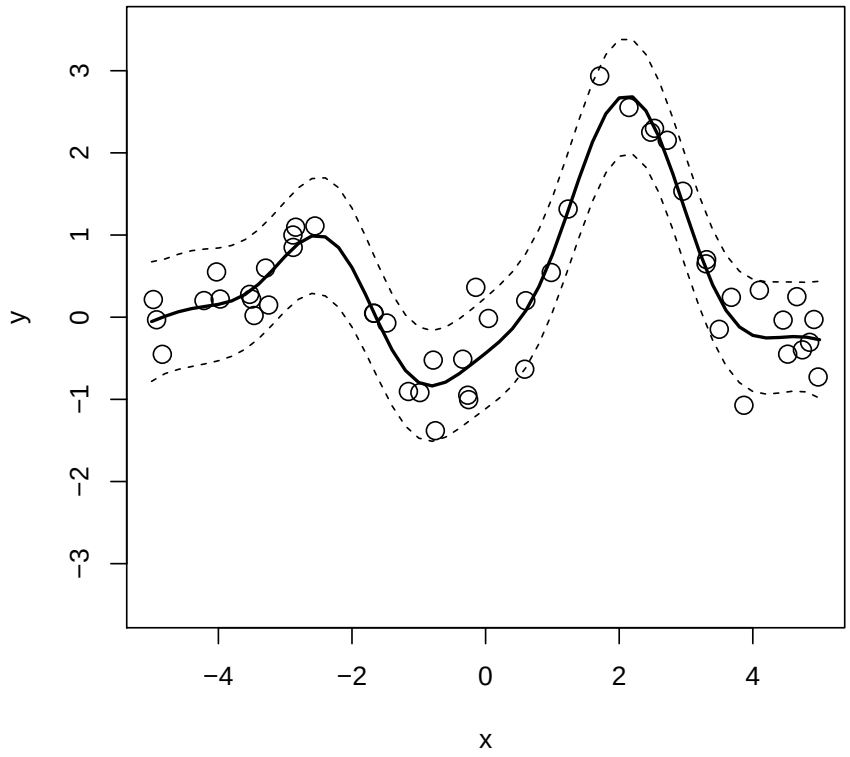
(h) Prior Samples

GP example

Ex. $m(x) = 0$, $k(x, x') = \exp(-\frac{1}{2}(x - x')^2)$, $\sigma^2 = 0.1$.



(i) Posterior Samples



(j) Predictive

Interpretation

Posterior:

$$f(x)|\mathbf{x}, \mathbf{y} \sim \mathrm{GP}(\hat{m}(x), \hat{k}(x, x')),$$

$$\text{where} \quad \hat{m}(x) = K(x, \mathbf{x}) \left(K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right)^{-1} \mathbf{y},$$

$$\text{and} \quad \hat{k}(x, x') = k(x, x') - K(x, \mathbf{x}) \left(K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right)^{-1} K(\mathbf{x}, x').$$

Posterior mean is linear in two ways:

$$\hat{m}(x) = \sum_{i}^{n} \beta_i y_i = \sum_{i=1}^{n} \alpha_i k(x, x_i).$$

Posterior variance is the difference between terms:

$$\hat{k}(x, x) = k(x, x) - K(x, \mathbf{x}) \left(K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right)^{-1} K(\mathbf{x}, x),$$

the prior variance from which a positive term telling us how much the data x has explained is subtracted.

Connections with finite linear models

Finite linear models with a **Gaussian prior** on the weights:

$$f(x) = \sum_{m=1}^{M} w_m \phi_m(x), \quad w \sim \mathsf{N}(0, \Sigma_0).$$

$\rightarrow$ for any $x_1, \ldots, x_n$, the joint distribution of $(f(x_1), \ldots, f(x_n))$ is multivariate Gaussian, i.e $f(x) \sim$ GP.

Mean function:

$$m(x) = \mathsf{E}_w[f(x)] = \sum_{m=1}^{M} \underbrace{\mathsf{E}_w[w_m]}_{0} \phi_m(x) = 0.$$

Connections with finite linear models

Covariance function:

$$k(x, x') = \mathsf{E}_w[f(x)f(x')] - \underbrace{\mathsf{E}_w[f(x)]\mathsf{E}_w[f(x')]}_{0}$$

$$= \mathsf{E}_w\left[\sum_{m=1}^{M}\sum_{m'=1}^{M} w_m w_{m'} \phi_m(x)\phi_{m'}(x')\right]$$

$$= \sum_{m=1}^{M}\sum_{m'=1}^{M} \underbrace{\mathsf{E}_w[w_m w_{m'}]}_{\Sigma_{0\,m,m'}} \phi_m(x)\phi_{m'}(x') = \phi(x)^T \Sigma_0 \phi(x').$$

In summary, the finite linear model with a Gaussian prior on the weights:

$$f(x) = \sum_{m=1}^{M} w_m \phi_m(x), \quad w \sim \mathsf{N}(0, \Sigma_0),$$

corresponds to a Gaussian process prior where

$$f(x) \sim \mathsf{GP}(0, k(x, x')), \quad k(x, x') = \phi(x)^T \Sigma_0 \phi(x').$$

Connections with infinite linear models

Consider finite linear models with

- Basis functions: $\phi_m(x) = \exp(-(x - \frac{m}{M})^2) \rightarrow$ uniformly placed Gaussian-shaped basis functions.
- Weights: Gaussian prior $w_m \overset{iid}{\sim} \mathsf{N}(0, 1)$.

Consider the limiting class of functions:

$$f(x) = \lim_{M \to \infty} \frac{1}{M} \sum_m w_m \phi_m(x) \quad = \int_{-\infty}^{\infty} w(u) \exp(-(x-u)^2) du,$$

where $w(u) \overset{iid}{\sim} \mathsf{N}(0, 1)$.

The **mean function** is:

$$m(x) = E_w[f(x)] = \int_{-\infty}^{\infty} \underbrace{E_w[w(u)]}_{0} \exp(-(x-u)^2) du = 0.$$

Connections with infinite linear models

The **covariance function** is:

$$k(x, x') = E_w[f(x)f(x')] = \int_{-\infty}^{\infty} \exp(-(x-u)^2) \exp(-(x'-u)^2) du$$

$$= \int_{-\infty}^{\infty} \exp\left(-2\left(u - \frac{x+x'}{2}\right)^2 + \left(\frac{x+x'}{2}\right)^2 - x^2 - x'^2\right) du$$

$$\propto \exp(-\frac{1}{2}(x-x')^2).$$

$\rightarrow$ a Gaussian process with squared exponential covariance function is equivalent to a regression model with infinitely many Gaussian-shaped basis functions placed everywhere.

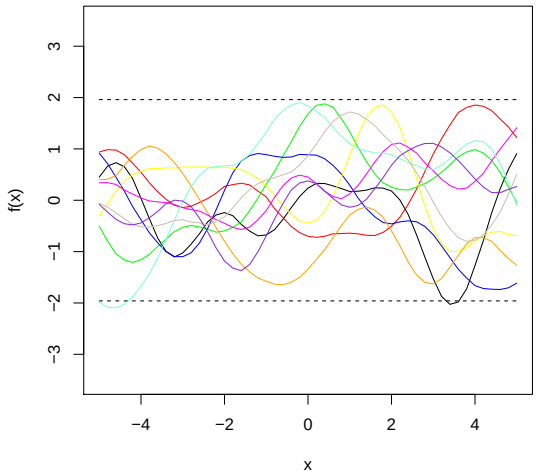
Indeed, for every positive definite covariance function, there exists a (infinite) basis function expansion (*Mercer's Theorem*, see Rasmussen and Williams (2006) pg. 96)

Squared exponential covariance function

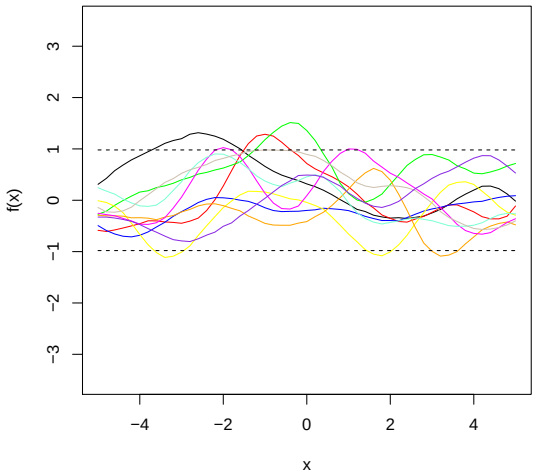
The squared exponential covariance function,

$$k_{\mathsf{SE}}(x, x') = v_0 \exp(-\frac{1}{2l^2}(x - x')^2),$$

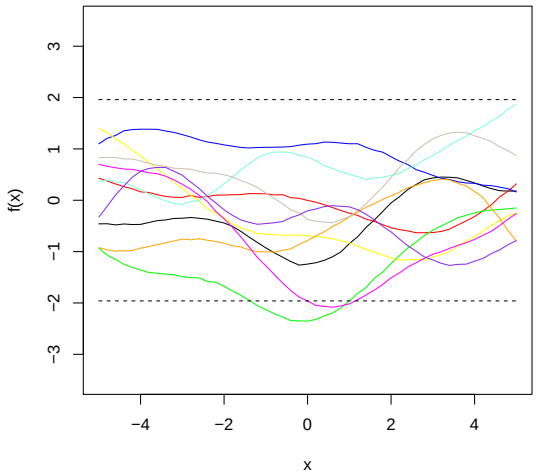
is stationary (a function of $(x - x')$) and infinitely differentiable (smooth realizations), with parameters v_0, which controls the pointwise prior variability, and l, the length scale.



(k) $v_0 = 1$, $l = 1$



(l) $v_0 = 1/4$, $l = 1$



(m) $v_0 = 1$, $l = 2$

Rational quadratic covariance function

The rational quadratic covariance function,

$$k_{\mathsf{RQ}}(x, x') = \left(1 + \frac{(x - x')^2}{2\alpha l^2}\right)^{-\alpha},$$

with parameters $\alpha > 0$, $l > 0$ can be viewed as a mixture of squared exponential covariance functions with different length scales.

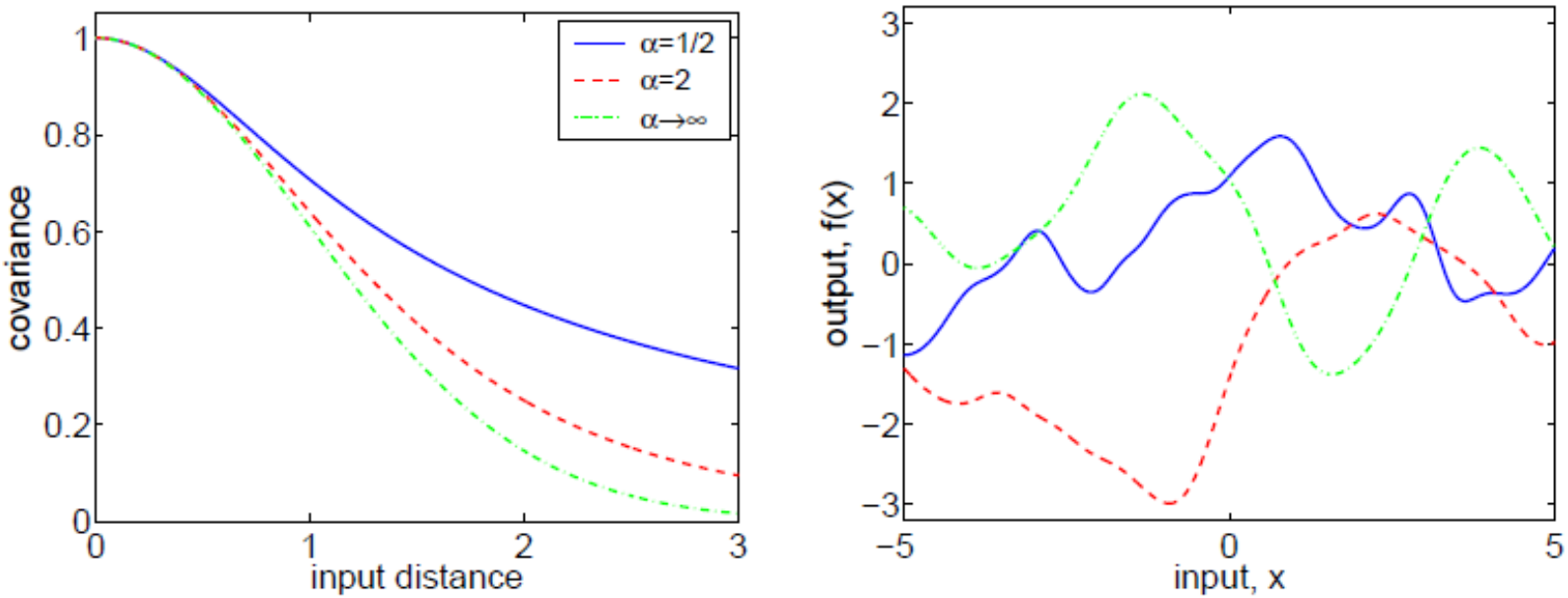
Let $\tau = l^{-2}$, $d = (x - x')$ and assume $\tau \sim \mathsf{Gam}(\alpha, \alpha/\beta)$,

$$\begin{aligned} k_{\mathsf{RQ}}(d) &= \int k_{\mathsf{SE}}(d|\tau) p(\tau|\alpha, \beta) d\tau \\ &\propto \int \exp\left(-\frac{1}{2}\tau d^2\right) \tau^{\alpha-1} \exp\left(-\frac{\alpha}{\beta}\tau\right) d\tau \\ &\propto \left(1 + \frac{(x - x')^2}{2\alpha l^2}\right)^{-\alpha}, \end{aligned}$$

where we set $\beta = l^{-2}$.

Rational quadratic covariance function

Covariance function and prior samples when $l = 1$.



As $\alpha \to \infty$, the RQ covariance function coverges to SE.

Matern covariance function

The Matern class of covariance functions is

$$k_{\mathrm{M}}(x, x') = \frac{2^{1-v}}{\Gamma(v)} \left(\frac{\sqrt{2v}}{l} |x - x'| \right)^{v} K_v \left(\frac{\sqrt{2v}}{l} |x - x'| \right),$$

where K_v is the modified Bessel function of the second kind of order v and l is the length scale.

Sample functions are $\lfloor v \rfloor$ times differentiable; the parameter v controls the smoothness.

Matern covariance function

Special cases:

- $v = 1/2$: Laplacian covariance function, Brownian motion (Ornstein-Uhlenbeck),

$$k(x, x') = \exp\left(-\frac{1}{l}|x - x'|\right).$$

- $v = 3/2$: (once differentiable)

$$k(x, x') = \left(1 + \frac{\sqrt{3}}{l}|x - x'|\right)\exp\left(-\frac{\sqrt{3}}{l}|x - x'|\right).$$

- $v = 5/2$, (twice differentiable)

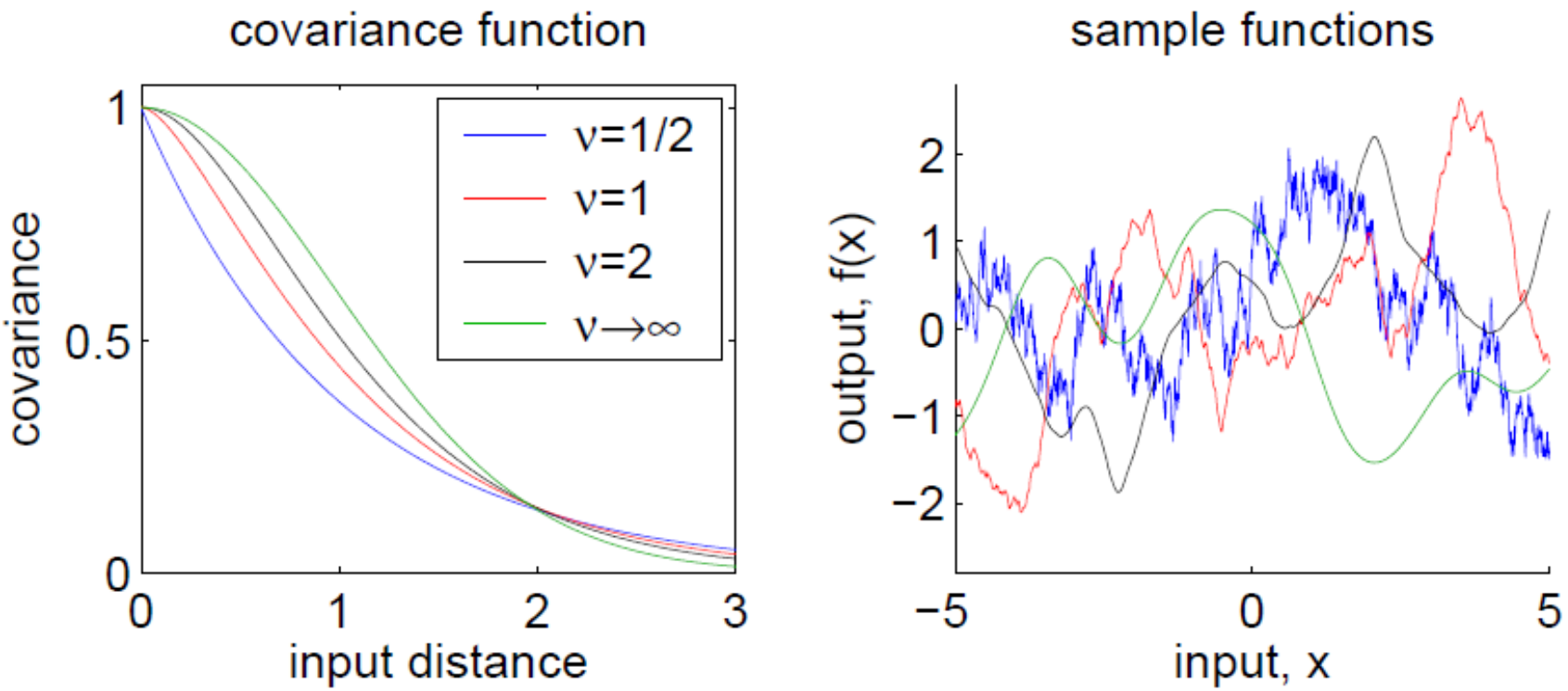
$$k(x, x') = \left(1 + \frac{\sqrt{5}}{l}|x - x'| + \frac{5}{3l^2}(x - x')^2\right)\exp\left(-\frac{\sqrt{5}}{l}|x - x'|\right).$$

- $v \to \infty$, (infinitely differentiable)

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

Rational quadratic covariance function

Covariance function and prior samples when $l = 1$.



As $v \to \infty$, the Matern covariance function converges to SE.

Periodic covariance functions

A prior over periodic functions can be obtained by 1) mapping the input to $u = (\sin(x), \cos(x))^T$ and 2) measuring distances in the u-space. For example, combined with the SE covariance function, we get $k(x, x') = v_0 \exp(-\frac{2}{l^2} \sin^2(\pi(x - x')))$.

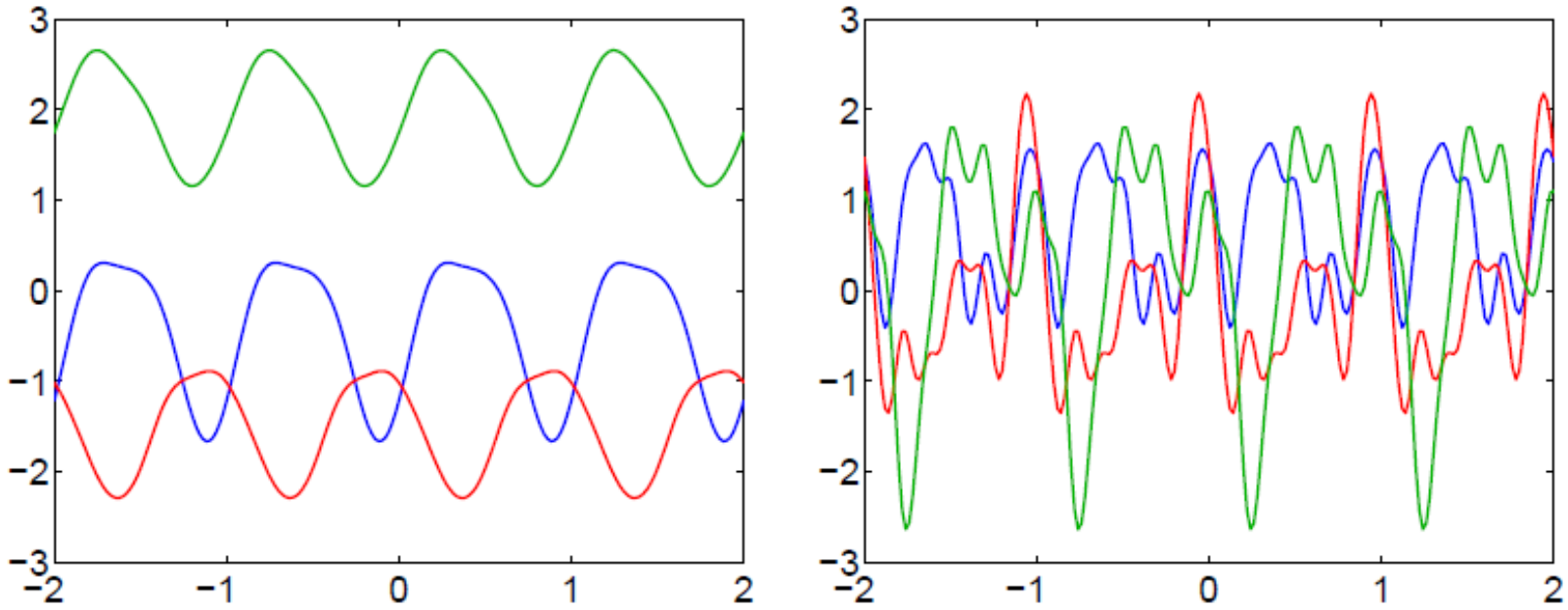


Figure: with $l > 1$ (left) and $l < 1$ (right)

Multivariate Extension

Multivariate extensions may be obtained by setting $d^2(x, x') = (x - x')^T M(x - x')$ for some positive semidefinite matrix M. If M is diagonal, this corresponds to different length scales on each dimension.

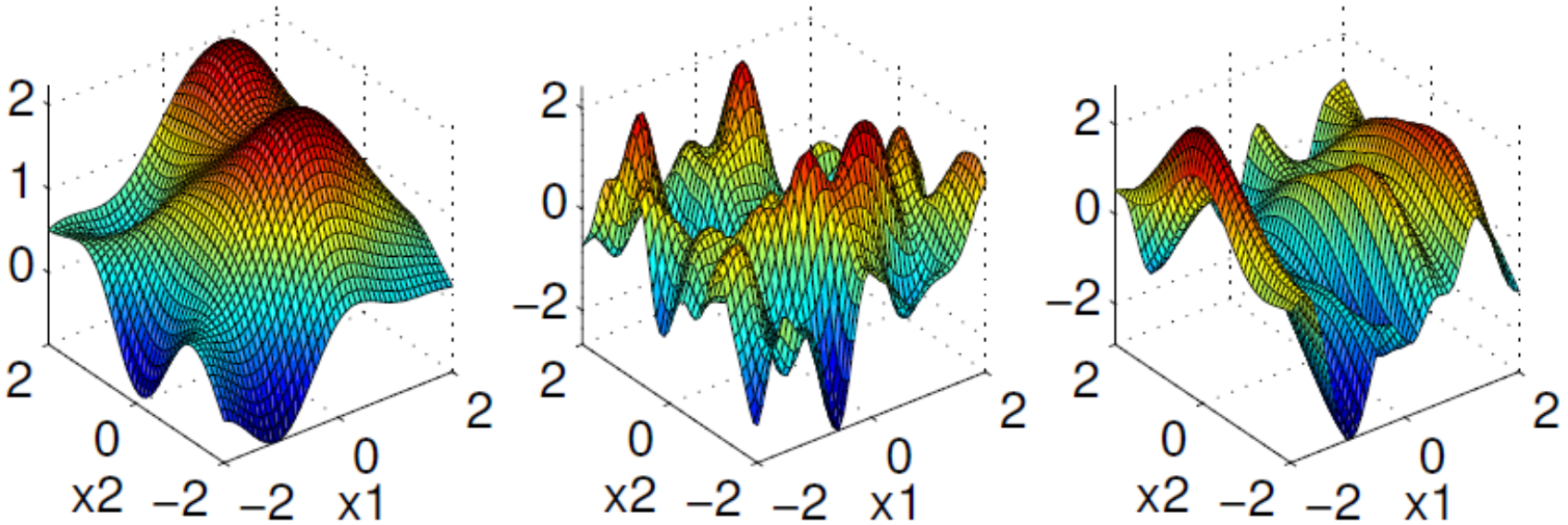


Figure: SE with a) $l_1 = 1$ and $l_2 = 1$; b) $l_1 = 0.32$ and $l_2 = 0.32$; c) $l_1 = 0.32$ and $l_2 = 1$

Inference

Two key elements to define with Gaussian processes:

- covariance function
- hyperparameters θ; such as the noise variance σ^2 and parameters of covariance functions (ex. length scale l).

The form of the covariance function is chosen by the researcher and the hyperparameters may be found by optimizing the marginal likelihood:

$$\log(p(\mathbf{y}|\mathbf{x},\theta)) \propto -\frac{1}{2}\mathbf{y}^T(K+\sigma^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K+\sigma^2 I|.$$

References

- Rasmussen, C.E. and Williams, C.K.I. (2006). Gaussian process for machine learning. *MIT press*.
- Rasmussen, C.E. and Ghahramani, Z. (2013). Machine learning course. `http://mlg.eng.cam.ac.uk/teaching/4f13/1213`