Viterbi Decoder and Regression

Anders Hansson and Martin Andersen

Linköping University and Technical Unversity of Denmark

April 18, 2024

Viterbi Decoder

Consider a signal $u \in \{0,1\}^{\mathbf{Z}_+}$ that is coded into $y \in \{0,1\}^{\mathbf{Z}_+}$ using a convolution

$$y_k = \sum_{i=1}^{n} c_i u_{k-i}$$

where $c_i \in \{0,1\}$ for $i \in \mathbf{N}_n$ describes the code. Summations are carried out modulo two.

Let $x_k \in \{0,1\}^n$ be such that

$$x_{k+1} = Ax_k + Bu_k$$
$$y_k = Cx_k$$

where A is a lower shift matrix with ones on the first sub-diagonal, where $B = e_1$, where $C = \begin{bmatrix} c_1 & \cdots & c_n \end{bmatrix}$, and where $x_0 = 0$.

Received signal is $r_k = y_k + e_k$ where e_k is a realization of a sequence of independent normally distributed random variables with zero mean and unit variance.

Maximum Likelihood Problem

for estimating u_k for $0 \leq k \leq N-1$ equivalent to optimal control problem

$$\text{minimize} \sum_{k=0}^{N} (r_k - Cx_k)^2$$

$$\text{subject to } x_{k+1} = Ax_k + Bu_k, \quad 0 \leq k \leq N-1$$

with variables $(u_0, x_1, \ldots, u_{N-1}, x_N)$, where $x_0 = 0$.

Introduce for $0 \leq k \leq N-1$ the functions
$f_k : \{0,1\}^n \times \{0,1\} \times \{0,1\}^n \to \mathbf{R}$ as

$$f_k(x, u, x^+) = (r_k - Cx)^2 + I(x, u, x^+)$$

where $I : \{0,1\}^n \times \{0,1\} \times \{0,1\}^n \to \mathbf{R}$ is the indicator function for the set

$$\{(x, u, x^+) \in \{0,1\}^n \times \{0,1\} \times \{0,1\}^n \mid x^+ = Ax + Bu\}$$

Also let $\phi : \{0,1\}^n \to \mathbf{R}$ be defined as $\phi(x) = (y_N - Cx)^2$.

Equivalent Formulation

$$\text{minimize } \phi(x_N) + \sum_{k=0}^{N-1} f_k(x_k, u_k, x_{k+1})$$

with variables $(u_0, x_1, \ldots, u_{N-1}, x_N)$, where $x_0 = 0$.

This is a partially separable optimization problem.

Dynamic Programming Formulation

Let $V_k : \{0,1\}^n \to \mathbf{R}$ be defined as

$$V_{k+1}(x^+) = \min_{u,x} \left\{ V_k(x) + f_k(x,u,x^+) \right\}$$

for $k = 0, \ldots, N-1$, where $V_0(x) = 0$. Also define $\mu_{k+1} : \{0,1\}^n \to \{0,1\} \times \{0,1\}^n$ as the minimizing argument in the above minimization, i.e.

$$\mu_{k+1}(x^+) = \operatorname*{argmin}_{x,u} \left\{ V_k(x) + f_k(x,u,x^+) \right\}$$

The function $V_N(x) + (r_N - Cx)^2$ is then finally minimized with respect to x to obtain the optimal x_N.

After this has been done all optimal variables can be recovered from the recursion

$$(u_{k-1}, x_{k-1}) = \mu_k(x_k), \quad 1 \leq k \leq N$$

Detailed Analysis

The minimization in each step in the recursion can be written as

$$\begin{aligned} &\text{minimize } V_k(x) + (r_k - Cx)^2 \\ &\text{subject to } Ax + Bu = x^+ \end{aligned}$$

with variables (u, x). Consider $n = 3$. Then the constraints are

$$\begin{aligned} u &= x_1^+ \\ x_1 &= x_2^+ \\ x_2 &= x_3^+ \end{aligned}$$

and hence only free variable to optimize over is x_3. Therefore optimization problem is

$$\text{minimize}_{x_3} \ (r_k - c_1 x_2^+ - c_2 x_3^+ - c_3 x_3)^2 + V_k((x_2^+, x_3^+, x_3))$$

where minimizing argument x_3 will be a function of x_2^+ and x_3^+. Also notice that V_{k+1} will only be a function of x_2^+ and x_3^+.

Detailed Analysis ctd.

The minimizing argument is hence

$$\mu_k(x) = \begin{bmatrix} x_1^+ \\ x_2^+ \\ x_3^+ \\ x_2(x_2^+, x_3^+) \end{bmatrix}$$

Notice that only the last component is non-trivial, and can be computed by enumeration of all values of x_2^+ and x_3^+. Need tables for V_k and μ_k that have 2^{n-1} entries.

In practical use of Viterbi decoding the value of N is not fixed but it is increasing and represents the current time. The decoding is done with some fixed dealy d measured from N, i.e. it is u_{N-d} that is estimated.

Need to store one table for V_N and d tables for μ_k, where $N - d + 1 \leq k \leq N$. In case $d2^{n-1}$ is large this could be costly. Approximations possible

Regression

Fit polynomial function $y : \mathbf{R} \to \mathbf{R}$ where

$$y(x) = a_1 + a_2 x + \cdots a_m x^{m-1}$$

to pairs (x_k, y_k) for $i \in \mathbf{N}_N$.

Define $a = (a_1, \ldots, a_m) \in \mathbf{R}^m$ and $\beta : \mathbf{R} \to \mathbf{R}^m$ by $\beta(x) = (1, x, \cdots, x^{m-1})$. Then

$$y(x) = a^T \beta(x) \tag{1}$$

called a *linear regression* model.

Fit by solving

$$\text{minimize} \frac{1}{2} \sum_{k=1}^{N} \left(a^T \beta(x_k) - y_k \right)^2 \tag{2}$$

with variable a. This is a linear Least Squares (LS) problem

Generalization

General functions $\varphi_j : \mathbf{R} \to \mathbf{R}$, $j \in \mathbf{N}_m$ and let

$$y(x) = \sum_{j=1}^{m} a_j \varphi_j(x)$$

Often $\varphi_1(x) = 1$, $\forall x$. With $\beta(x) = (\varphi_1(x), \ldots, \varphi_m(x))$.

$$y(x) = a^T \beta(x)$$

Vector-Valued Regression Model

Let $y : \mathbf{R}^n \to \mathbf{R}^p$ be given by

$$y(x) = A\beta(x)$$

where $A \in \mathbf{R}^{p \times m}$, $\beta : \mathbf{R}^n \to \mathbf{R}^m$ with $\beta(x) = (\varphi_1(x), \ldots, \varphi_m(x))$ and $\varphi_j : \mathbf{R}^n \to \mathbf{R}$, $j \in \mathbf{N}_m$. Important special case: $\beta(x) = x$.

LS criterion is the sum of the LS criteria for each row in regression model, or

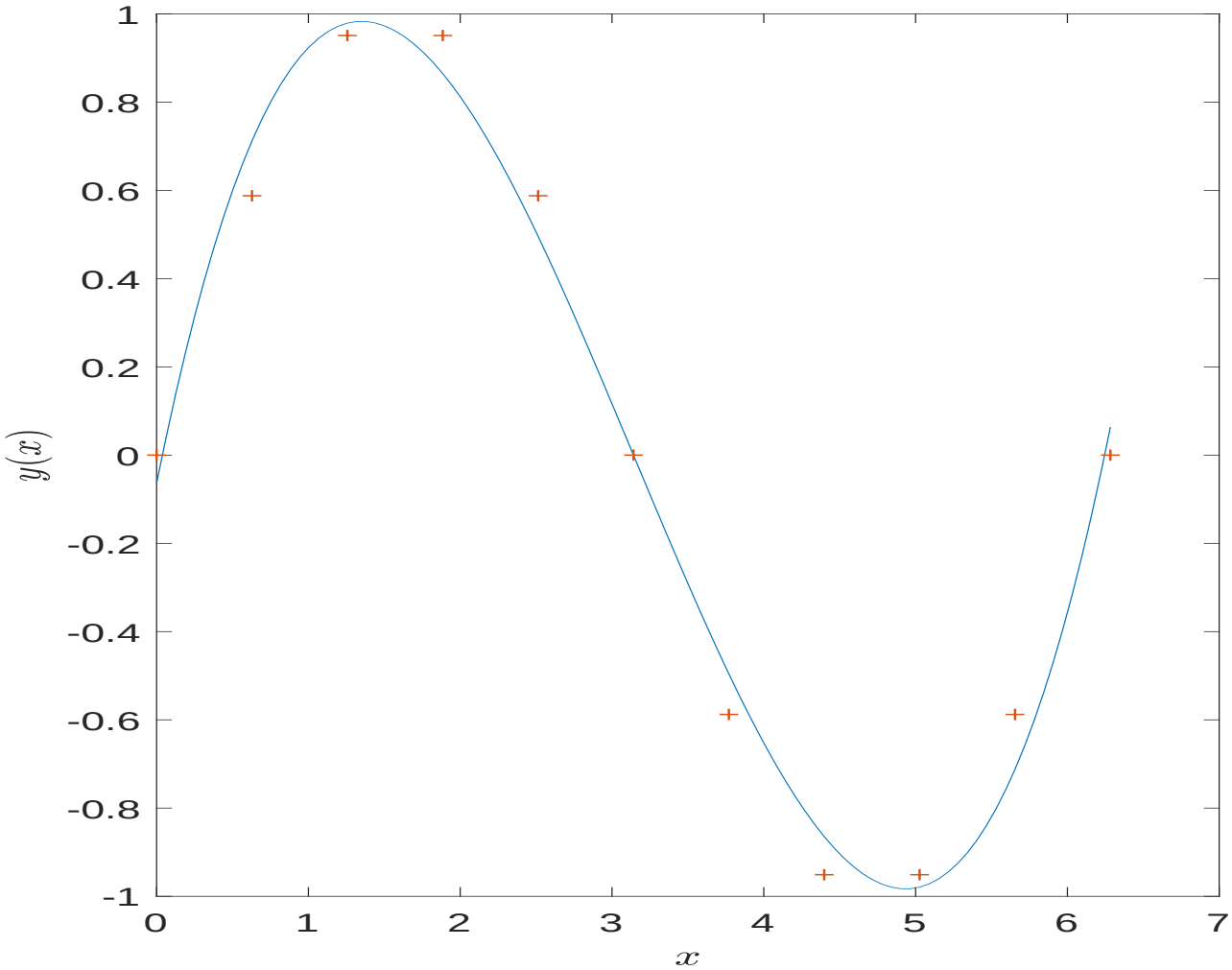
$$\text{minimize} \frac{1}{2} \sum_{k=1}^{N} (y_k - A\beta(x_k))^T (y_k - A\beta(x_k))$$

with variable A. The solution to this problem is closely related to the stochastic averaging approximation.

$y(x)$ often called a *predictor*.

Example

Given pairs of data (x_k, y_k) that satisfy $y_k = \sin x_k$ and we want to find a linear regression model (1) that solves (2) for $\beta(x) = (1, x, x^2, x^3, x^4)$.



Statistical Interpretation

Let E_k be independent with pdf $f_{E_k} : \mathbf{R} \to \mathbf{R}_+$ where

$$f_{E_k}(e_k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2} e_k^2}$$

Let e_k be the outcomes of E_k and

$$y_k = a^T \beta(x_k) + e_k$$

with $a \in \mathbf{R}^m$ and $\beta : \mathbf{R}^n \to \mathbf{R}^m$ model the observations y_k, $k \in \mathbf{N}_N$.

The pdf of observation $f_{Y_k} : \mathbf{R} \to \mathbf{R}_+$ is

$$f_{Y_k}(y_k) = f_{E_k}(y_k - a^T \beta(x_k))$$

Maximum Likelihood Estimation

Likelihood function $\ell : \mathbf{R}^N \times \mathbf{R}^m \to \mathbf{R}_+$

$$\ell\left(y_, \ldots y_N; a\right) = \prod_{k=1}^{N} f_{Y_k}(y_k) = \prod_{k=1}^{N} f_{E_k}\left(y_k - a^T \beta(x_k)\right)$$

Then

$$-\ln \ell\left(y_1, \ldots y_N; a\right) = \sum_{k=1}^{N} \frac{1}{2\sigma^2}\left(y_k - a^T \beta(x_k)\right)^2 + N \ln\left(\sqrt{2\pi}\sigma\right)$$

Minimizing a same as LS solution.

Maximum A Posteriori (MAP) Estimation

Assume that also a is the outcome of a random variable A with some pdf $f_A : \mathbf{R}^m \to \mathbf{R}_+$ independent of E_k.

Joint distribution $f : \mathbf{R}^N \times \mathbf{R}^m \to \mathbf{R}_+$ of the observations $y = (y_1, \ldots, y_N)$ and parameter a:

$$f(y; a) = f_A(a) \prod_{k=1}^{N} f_{E_k}\left(y_k - a^T \beta(x_k)\right)$$

which is proportional to distribution of A conditioned on the observations $Y = y$.

$$\text{minimize} \sum_{k=1}^{N} \frac{1}{2\sigma^2}\left(y_k - a^T \beta(x_k)\right)^2 + N \ln\left(\sqrt{2\pi}\sigma\right) - \ln f_A(a)$$

with varible a.

Regularization

Different distributions f_A result in different regularizations.

Normal distribution with zero mean and covariance Σ:

$$-\ln f_A(a) \sim a^T \Sigma^{-1} a$$

Called *Ridge regression* or *Tikhonov regularization*.

Components A_i of A have double-sided exponential distribution:

$$f_A(a) = \prod_{i=1}^{m} \frac{1}{2\lambda} e^{-|a_i|/\lambda} \Rightarrow -\ln q(a) \sim \sum_{i=1}^{m} |a_i|$$

Called *Lasso regularization.*

Any regularization function $h_i(a_i)$ can be translated into a pdf for A_i by the formula $f_{A_i}(a_i) = e^{-h_i(a_i)}/\mu$, where $\mu > 0$ is a normalization constant.

Constrained Regression

Consider

$$f_{A_i}(a_i) = \begin{cases} \frac{1}{2\lambda}, & a_i \in [-\lambda, \lambda] \\ 0 & a_i \notin [-\lambda, \lambda] \end{cases}$$

$$\Longrightarrow$$

$$-\ln f_{A_i}(a_i) = \begin{cases} \ln(2\lambda), & a_i \in [-\lambda, \lambda] \\ +\infty & a_i \notin [-\lambda, \lambda] \end{cases}$$

Hence the terms $-\ln q_i(a_i)$ in the objective function of the MAP problem can be replaced with the constraint $a_i \in [-\lambda, \lambda]$ for $i \in \mathbf{N}_m$ or equivalently $\|a\|_\infty \leq \lambda$.

Hilbert Spaces

For $a = (a_1, a_2, \ldots)$, where $a_i \in \mathbf{R}$, $i \in \mathbf{N}$ we say that $a \in \ell_2$ if $\sum_{i=1}^{\infty} a_i^2 < \infty$.

Inner product: $\langle \cdot, \cdot \rangle_{\ell_2} : \ell_2 \times \ell_2 \to \mathbf{R}$ by $\langle a, b \rangle_{\ell_2} = \sum_{i=1}^{\infty} a_i b_i$.

Corresponding norm: $\| \cdot \|_{\ell_2} \to \mathbf{R}_+$ is defined as $\|a\|_{\ell_2}^2 = \langle a, a \rangle_{\ell_2}$.

For $f : \mathbf{R}^n \to \mathbf{R}$ we say that $f \in L_2$ if

$$\int_{\mathbf{R}^n} f(x)^2 dx < \infty$$

Inner product: $\langle \cdot, \cdot \rangle_{L_2} : L_2 \times L_2 \to \mathbf{R}$ defined as

$$\langle f, g \rangle_{L_2} = \int_{\mathbf{R}^n} f(x) g(x) dx$$

Corresponding norm: $\| \cdot \|_{L_2} : L_2 \to \mathbf{R}_+$ is defined as $\|f\|_{L_2}^2 = \langle f, f \rangle_{L_2}$.

Regression in Hilbert Spaces

Let $\beta(x) = (\varphi_1(x), \varphi_2(x), \ldots)$, where $\varphi_i : \mathbf{R}^n \to \mathbf{R}$, $i \in \mathbf{N}$.

Assume $\varphi_i \in L_2$, $i \in \mathbf{N}$:

$$\begin{aligned} \langle \varphi_i, \varphi_j \rangle_{L_2} &= 0, \qquad i \neq j \\ \langle \varphi_i, \varphi_j \rangle_{L_2} &= 1, \qquad i = j \end{aligned}$$

Let $f(x) = \sum_{i=1}^{\infty} a_i \varphi_i(x)$ be a regressor and consider

$$\text{minimize} \frac{1}{2} \sum_{k=1}^{N} (y_k - f(x_k))^2 + \frac{\nu}{2} \|f\|_{L_2}^2$$

with variable a.

Equivalent Optimization Problem

A sum $f(x) = \sum_{i=1}^{\infty} a_i \varphi_i(x)$ is convergent and belongs to L_2 for orthonormal φ_i if and only if $a \in \ell_2$. Then $a_i = \langle f, \varphi_i \rangle_{L_2}$ and hence $\|f\|_{L_2} = \|a\|_{\ell_2}$.

We reformulate this problem as a constrained problem with variables (a, e):

$$\text{minimize } \frac{1}{2} \sum_{k=1}^{N} e_k^2 + \frac{\nu}{2} \|a\|_{\ell_2}^2$$

$$\text{subject to } e_k = y_k - \sum_{i=1}^{\infty} a_i, \varphi_i(x), \quad k \in \mathbf{N}_K$$

where $e = (e_1, \ldots, e_N)$.

Vector Notation

To ease notation:

$$\begin{aligned} &\text{minimize } \frac{1}{2} e^T e + \frac{\nu}{2} a^T a \\ &\text{subject to } e = y - \Phi a \end{aligned}$$

with variables (a, e), where Φ is infinite-dimensional matrix:

$$\Phi = \begin{bmatrix} \beta^T(x_1) \\ \vdots \\ \beta^T(x_N) \end{bmatrix}$$

Lagrangian

Lagrangian $L : \mathbf{R}^N \times \ell_2 \times \mathbf{R}^N \to \mathbf{R}$ defined by

$$L(e, a, \lambda) = \frac{1}{2} e^T e + \frac{\nu}{2} a^T a + \lambda^T (e - y + \Phi a)$$

Completing the squares:

$$L(e, a, \lambda) = \frac{1}{2}(e + \lambda)^T (e + \lambda) + \frac{\nu}{2} \left(a + \frac{1}{\nu} \Phi^T \lambda \right)^T \left(a + \frac{1}{\nu} \Phi^T \lambda \right)$$
$$- \frac{1}{2} \lambda^T \lambda - \frac{1}{2\nu} \lambda^T \Phi \Phi^T \lambda - \lambda^T y$$

Minimized for

$$e = -\lambda; \qquad a = -\frac{1}{\nu} \Phi^T \lambda$$

Kernel Trick

Lagrange dual function $g : \mathbf{R}^N \to \mathbf{R}$ defined by

$$g(\lambda) = -\frac{1}{2}\lambda^T\lambda - \frac{1}{2\nu}\lambda^T\Phi\Phi^T\lambda - \lambda^T y$$

Kernel function $K : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ defined by
$K(x, \bar{x}) = \beta(x)^T\beta(\bar{x}) = \sum_{i=1}^{\infty} \varphi_i(x)\varphi_i(\bar{x})$

Define $\mathcal{K} \in \mathbf{S}^N$ with elements $\mathcal{K}_{i,j} = K(x_i, x_j)$. Then dual optimization problem

$$\text{minimize}\frac{1}{2}\lambda^T\left(I + \frac{1}{\nu}\mathcal{K}\right)\lambda + \lambda^T y$$

with variable λ, with solution

$$\lambda = -\left(I + \frac{1}{\nu}\mathcal{K}\right)^{-1} y$$

We may also write

$$f(x) = a^T\beta(x) = -\frac{1}{\nu}\lambda^T\Phi\beta(x) = -\frac{1}{\nu}\sum_{k=1}^{N}\lambda_k K(x_k, x)$$

Kernel Functions

Positive semidefinite kernel: for any $N \in \mathbf{N}$, and any $c_i \in \mathbf{R}$, and $x_i \in \mathbf{R}^n$, $i \in \mathbf{N}_N$ it should hold that

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K(x_i, x_j) \geq 0$$

Mercer's theorem: If K is a positive definite kernel, then there are orthonormal φ_i that defines it. (could be finitely or infinitely many)

dth degree polynomials: $K(x, \bar{x}) = \left(1 + x^T \bar{x}\right)^d$

Radial basis functions: functions that only depend on $\|x - \bar{x}\|_2$

Gaussian radial basis kernel

$$K(x, \bar{x}) = \exp\left(-\frac{1}{2\sigma^2}\|x - \bar{x}\|_2^2\right)$$

where $\sigma \in \mathbf{R}_{++}$.

Gaussian Linear Regression

Linear regression model:

$$Y_k = A^T x_k + E_k, \qquad k \in \mathbf{N}_N$$

where E_k, Y_k random variables and A is random vector with same dimension as x_k.

For the outcomes (y_k, e_k, a) of an experiment we have

$$y_k = a^T x_k + e_k, \quad k \in \mathbf{N}_N$$

Joint Gaussian

Assume A is Gaussian with zero mean and covariance $\Sigma_a \in \mathbf{S}^n_{++}$, and $E = (E_1, \ldots, E_N)$ Gaussian with zero mean and covariance $\Sigma_e \in \mathbf{S}^N_{++}$, where E_k and A independent. With

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix}$$

we have

$$Y = XA + E$$

Hence (Y, A) is Gaussian with zero mean and covariance

$$\begin{bmatrix} R & X\Sigma_a \\ \Sigma_a X^T & \Sigma_a \end{bmatrix}$$

where $R = \Sigma_e + X\Sigma_a X^T$.

Conditional pdf

Let $f : \mathbf{R}^N \times \mathbf{R}^n \to \mathbf{R}_+$ be joint pdf for (Y, A),
$f_{A|Y} : \mathbf{R}^N \times \mathbf{R}^n \to \mathbf{R}_+$ conditional pdf for A given Y, and
$f_Y : \mathbf{R}^N \to \mathbf{R}_+$ marginal pdf for Y.

With $\Sigma_{a|y} = \Sigma_a - \Sigma_a X^T R^{-1} X \Sigma_a$ we have
$f(y, a) = f_Y(y) f_{A|Y}(a|y)$, where

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^N \det R}} e^{-\frac{1}{2} y^T R^{-1} y}$$

$$f_{A|Y}(a|y) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma_{a|y}}} e^{-\frac{1}{2}(a-\mu_{a|y})^T \Sigma_{a|y}^{-1} (a-\mu_{a|y})}$$

where $\mu_{a|y} = \Sigma_a X^T R^{-1} y$ is the conditional mean for A given $Y = y$, c.f. predictor for unsupervised learning.

Predictor

Conditional pdf maximized by $a = \mu_{a|y}$.

Prediction with a new value x given by

$$\mu_{a|y}^T x = x^T \mu_{a|y} = x^T \Sigma_a X^T \left(\Sigma_e + X \Sigma_a X^T \right)^{-1} y$$

Kernel Function Interpretation

Notice that Σ_a only appears in terms of expressions of the form $z^T \Sigma_a \bar{z}$ for $z \in \mathbf{R}^n$ and $\bar{z} \in \mathbf{R}^n$.

Can instead write predictor in terms of kernel function $K : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}_+$ defined by $K(z, \bar{z}) = z^T \Sigma_a \bar{z}$.

Let $\mathcal{K} \in \mathbf{S}_+^N$: $\mathcal{K}_{i,j} = K(x_i, x_j)$, and let $\eta_k = K(x_k, x)$, $k \in \mathbf{N}_N$. Then

$$\mu_{a|y}^T x = \sum_{k=1}^{N} \eta_k \left(\Sigma_e + \mathcal{K}\right)^{-1} y$$

Hilbert Space Regression Interpretation

If $\Sigma_e = \nu I$ and if φ_i orthonormal such that resulting kernel function is $K(z, \bar{z}) = z^T \Sigma_a \bar{z}$, then above predictor is same as for Hilbert space regression.

$K(z, \bar{x})$ is covariance between $A^T z$ and $A^T \bar{z}$.

Gaussian Processes

Generalization of regressor:

$$Y_k = \mathcal{A}(x_k) + E_k$$

with $\mathcal{A} : \mathbf{R}^n \to \mathbf{R}$, where we specify the covariance between $\mathcal{A}(z)$ and $\mathcal{A}(\bar{z})$ for any $z \in \mathbf{R}^n$ and $\bar{z} \in \mathbf{R}^n$ by specifying the kernel function.

Still assume that the joint distribution is Gaussian. This defines a zero mean real-valued *Gaussian random process* on $\mathbf{R}^n$.

Covariance Functions

Any of the previous kernel functions may be used.

The Gaussian radial basis function is called the squared exponential covariance function.

Another common covariance function:

$$K(z, \bar{z}) = \exp\left(\frac{1}{\sigma}\|z - \bar{z}\|_2\right)$$

which defines the *Ornstein–Uhlenbeck* process, where $\sigma \in \mathbf{R}_{++}$.

Properties of Random Processes

When covariance function only dependes on $z - \bar{z}$, the process is *stationary*.

When it only depends on $\|z - \bar{z}\|_2$ is also *isotropic*.

Stationary and isotropic together is called *homogeneous*.

We specifically realize that infinite dimensional regression in Hilbert spaces can be equivalently expressed as MAP estimation for Gaussian processes.

Example

Again consider data from a sinusoidal, but collected from three periods of the sinusoidal. We take $\Sigma_e = 0$, since we have no measurement errors of the sinusoidal. We use the squared exponential covariance function with parameter $\sigma = 1$.