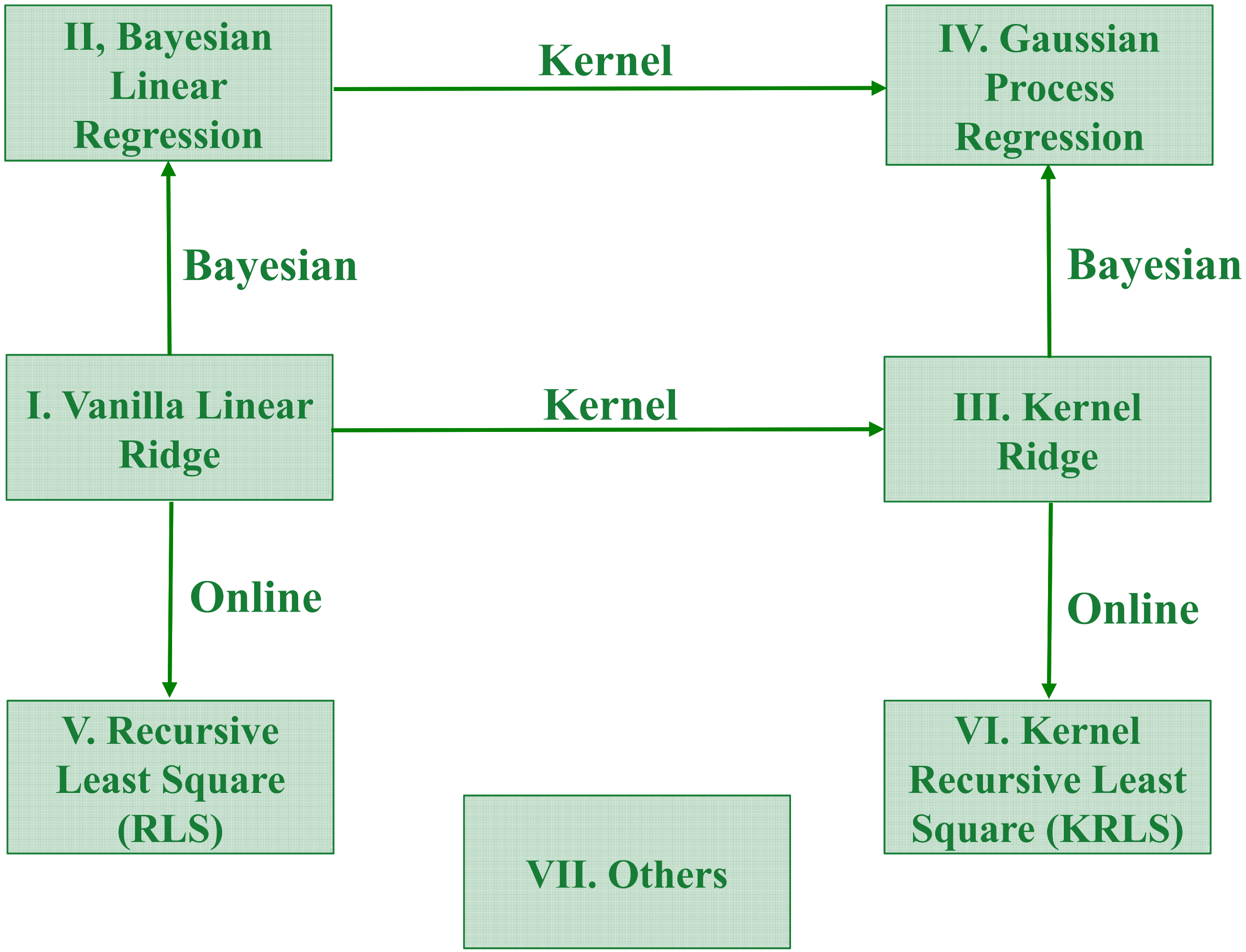
Advanced Linear Regression: Bayesian, Kernel and Online

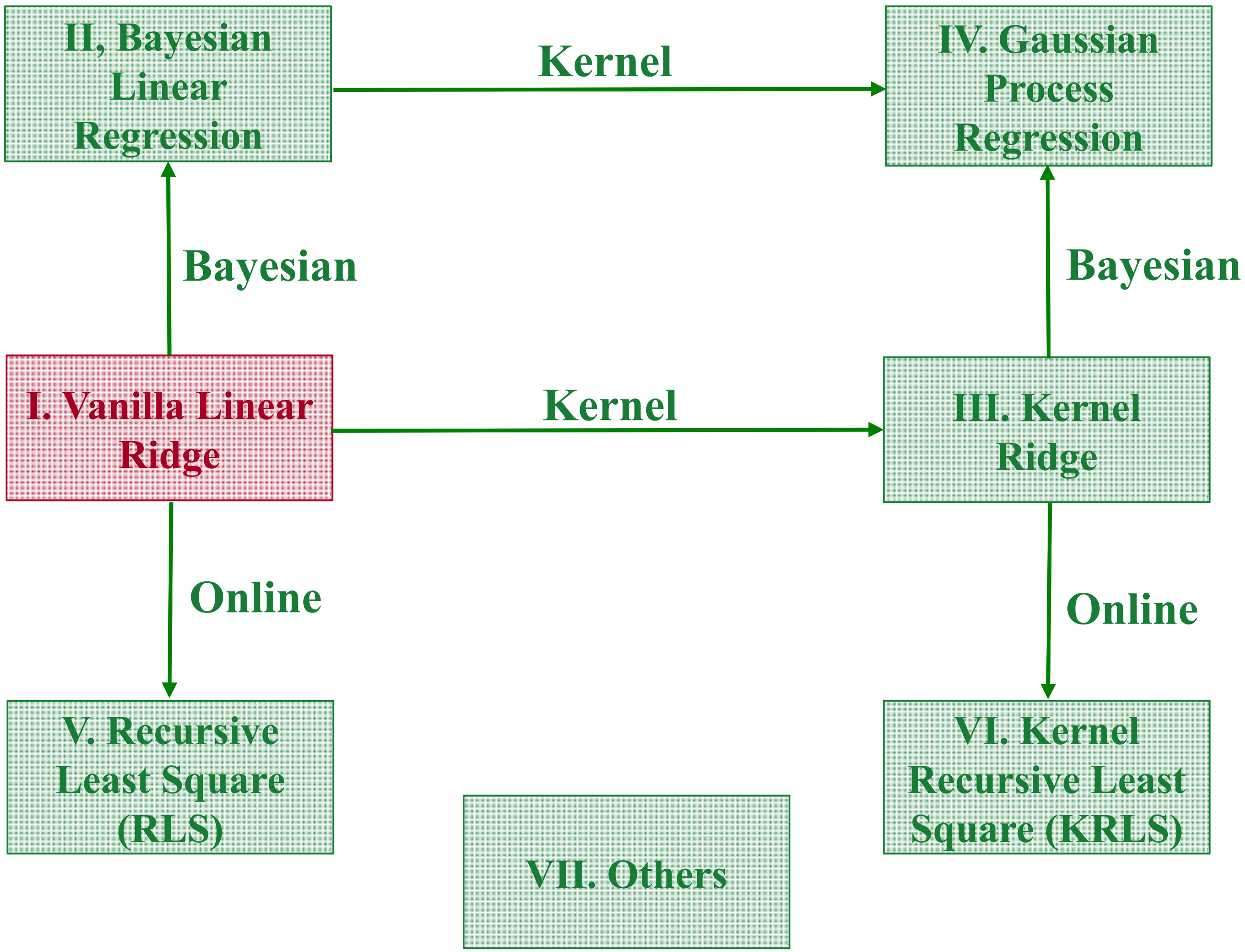
(Jan 11th, 2017)

YANG Jiancheng

Outline


II, Bayesian Linear Regression
Kernel
IV. Gaussian Process Regression
Bayesian
Bayesian
I. Vanilla Linear Ridge
Kernel
III. Kernel Ridge
Online
Online
V. Recursive Least Square (RLS)
VII. Others
VI. Kernel Recursive Least Square (KRLS)

Outline


II, Bayesian Linear Regression
Kernel
IV. Gaussian Process Regression
Bayesian
Bayesian
I. Vanilla Linear Ridge
Kernel
III. Kernel Ridge
Online
Online
V. Recursive Least Square (RLS)
VII. Others
VI. Kernel Recursive Least Square (KRLS)

• I. Vanilla Linear Ridge

- **Matrix Form**

$$\underset{n\times m}{Y} = \underset{n\times d}{X}\cdot \underset{d\times m}{W} + e$$

$d \to$ features

$m \to$ multiple output

$$\text{loss} = \|e\|^2 = \|Y - XW\|^2 = (Y-XW)^T(Y-XW)$$

$$= Y^TY - Y^TXW - W^TX^TY + W^TX^TXW$$

$$\frac{\partial}{\partial W}\text{loss} = -2X^TY + 2X^TXW = 0$$

$$\boxed{W = (X^TX)^{-1}X^TY = X^{+}Y}$$

$\to$ pseudo-inverse

• I. Vanilla Linear Ridge

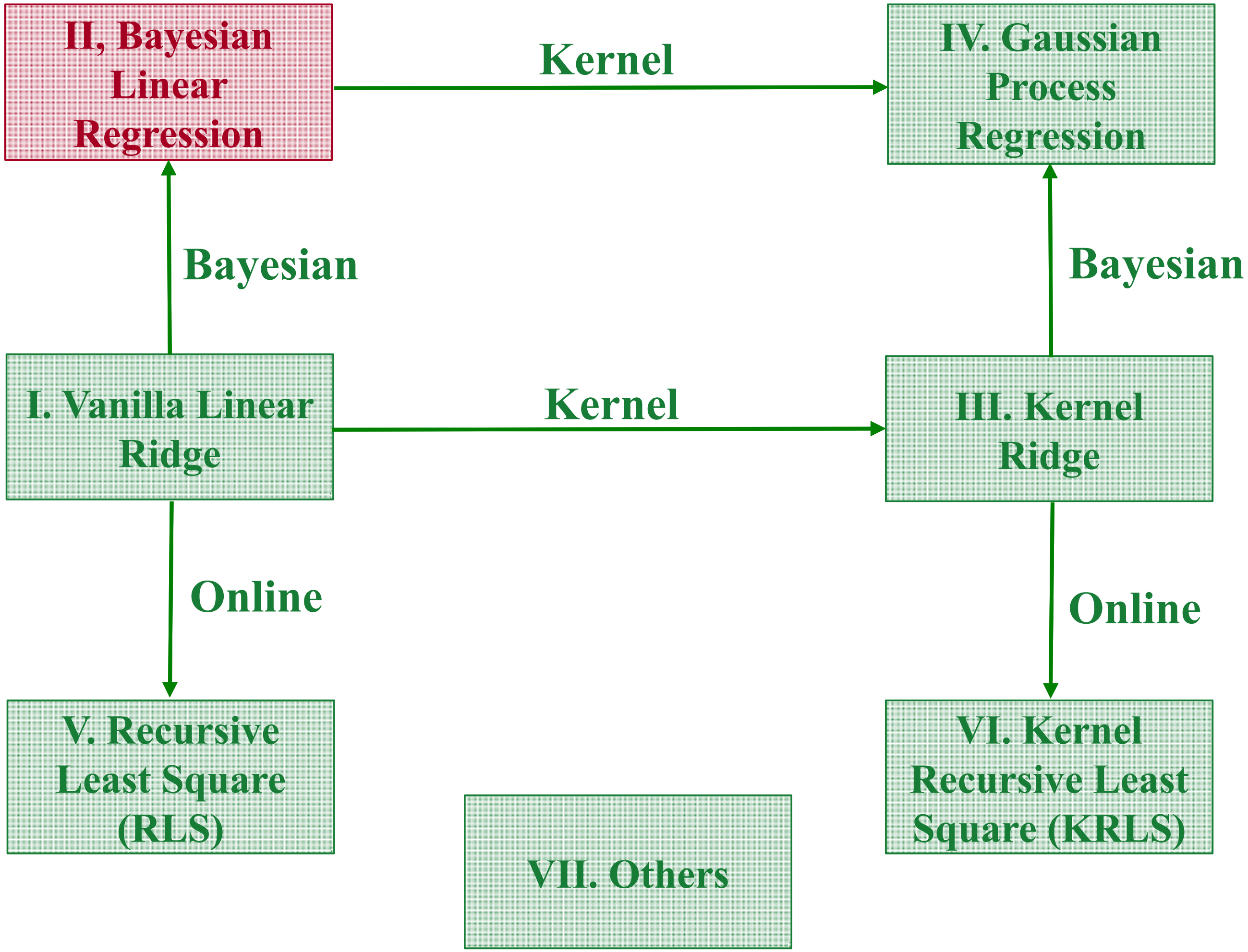
- **Ridge**

$$loss = \|y - Xw\|^2 + \lambda \|w\|^2$$

$$\frac{\partial}{\partial w} loss = -2X^T y + 2X^T X w + 2\lambda w = 0$$

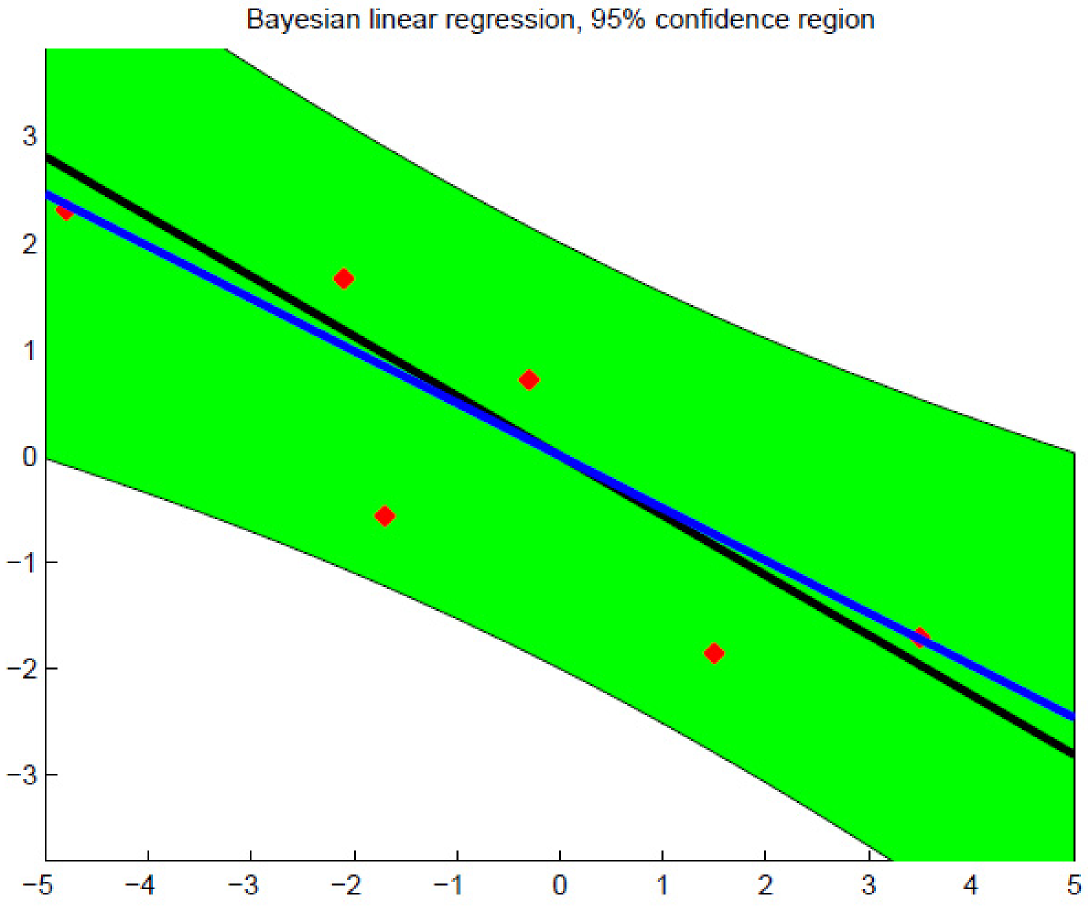
$$\boxed{w = (X^T X + \lambda I_d)^{-1} X^T y}$$

Outline


II, Bayesian Linear Regression
Kernel
IV. Gaussian Process Regression
Bayesian
Bayesian
I. Vanilla Linear Ridge
Kernel
III. Kernel Ridge
Online
Online
V. Recursive Least Square (RLS)
VII. Others
VI. Kernel Recursive Least Square (KRLS)

• II, Bayesian Linear Regression

- **Effect of Bayesian**



• II, Bayesian Linear Regression

- **Bayesian Interpretation**

For simplity, $m=1$ (single target)

$$\underset{n\times 1}{\underline{y}} = \underset{n\times d}{\underline{X}}\,\underset{d\times 1}{\underline{w}} + \underset{n\times 1}{\underline{\varepsilon}}, \quad \boxed{\varepsilon_i \sim N(0, \delta^2)}$$

$$\Rightarrow y \sim N(Xw, \delta^2 I_n)$$

Plus, we assume:

$$\boxed{w \sim N(0, \tau^2 I_d)}$$

• II, Bayesian Linear Regression

- **Posterior**

$$p(w \mid X, y) \propto P(X, y \mid w)\, P(w)$$

$$\propto \exp\Big(\underbrace{-\frac{(y-Xw)^T(y-Xw)}{2\delta^2} - \frac{w^T w}{2\tau^2}}_{\downarrow\ +\ \text{const}\quad (w-\mu)^T \Lambda (w-\mu)}\Big)$$

$$\Lambda_{d\times d} = \Big(\frac{1}{\delta^2} X^T X + \frac{1}{\tau^2} I_d\Big)$$

$$\mu = \frac{1}{\delta^2} \Lambda^{-1} X^T y = \Big(X^T X + \frac{\delta^2}{\tau^2} I\Big) X^T y$$

$$\boxed{P(w \mid D) \sim \mathcal{N}(\mu, \Lambda)}$$

MAP : $w = \mu$

$\Leftrightarrow$ ridge

• II, Bayesian Linear Regression

- **Predictive (1)**

$()_{d\times 1}$ $\rightarrow$ Scalar

Given a new sample (x_0, y_0)

We want

$$\underline{P(y_0 \mid x_0, D)}, \quad D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$$

Use Bayesian Average :

$$P(y_0 \mid x_0, D) = \int P(y_0, w \mid x_0, D)\, dw$$

$$= \int \underbrace{P(y_0 \mid x_0, D, w)}_{P(y_0 \mid x_0, w)} \underbrace{P(w \mid x_0, D)}_{P(w \mid D)}\, dw$$

$y_0 \perp D \mid w$ $\qquad$ $w \perp x_0 \mid D$

$$= \int N(w^T x_0, \delta^2)\, N(\mu, \Lambda^{-1})\, dw$$

• II, Bayesian Linear Regression

- **Predictive (2)**

$$p(y_0|x_0,D) \propto \int \exp\Big(\underbrace{\frac{(y_0-w^Tx_0)^2}{2\delta^2} - \frac{1}{2}(w-u)^T\Lambda(w-u)}_{(w-m)^TL(w-m)+\cdots}\Big)dw$$

Our goal:

$$\int g(y_0)\cdot N(w|\ldots)\,dw = g(y_0)$$

$$p(y_0|x_0,D) \propto \int \exp\Big(\frac{(w-m)^TL(w-m)}{2}\Big)\underbrace{\exp\Big(\frac{m^TLm}{2} - \frac{y_0^2}{2\delta^2}\Big)}_{g(y_0)}dw$$

$$\boxed{\begin{aligned} g(y_0) &= \exp\Big(-\frac{\lambda}{2}(y_0-\hat{y}_0)^2\Big) \\ \hat{y}_0 &= u^Tx_0 \\ \frac{1}{\lambda} &= \delta^2 + x_0^T\Lambda^{-1}x_0 \end{aligned}}$$

$$\boxed{\begin{aligned} L &= \frac{x_0x_0^T}{\delta^2} + \Lambda \\ m &= L^{-1}\Big(\frac{y_0x_0}{\delta^2} + \Lambda u\Big) \end{aligned}}$$

• II, Bayesian Linear Regression

- **Predictive (3)**

Same !

$$N(y_0 \mid x_0, D) \sim N(u^T x_0, \delta^2 + x_0^T \Lambda^{-1} x_0)$$

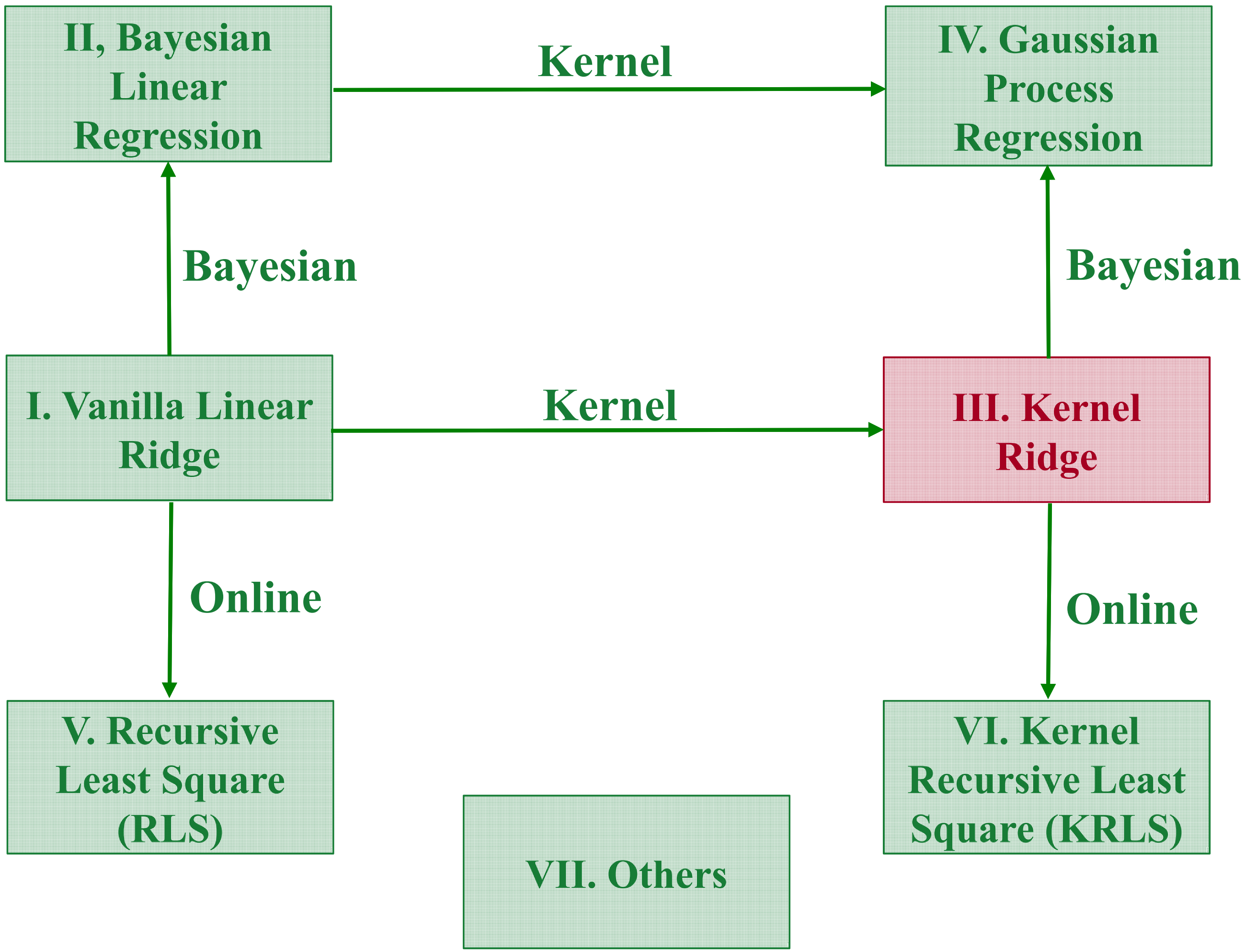
$$\varepsilon_i \sim N(0, \delta^2)$$

$$\omega \sim N(0, \tau^2 I_d)$$

$$u = (X^T X + \frac{\delta^2}{\tau^2} I) X^T y$$

$$\Lambda = \frac{1}{\delta^2} X^T X + \frac{1}{\tau^2} I_d$$

Outline


II, Bayesian Linear Regression
Kernel
IV. Gaussian Process Regression
Bayesian
Bayesian
I. Vanilla Linear Ridge
Kernel
III. Kernel Ridge
Online
Online
V. Recursive Least Square (RLS)
VII. Others
VI. Kernel Recursive Least Square (KRLS)

• III. Kernel Ridge

- **Dual Form**

$$(X^T X + \lambda I) w = X^T y$$

$$\lambda w = X^T (y - Xw)$$

$$w = X^T \cdot \boxed{\frac{y - Xw}{\lambda}} \stackrel{\text{def}}{=} w = X^T \alpha$$

$$\boxed{\begin{array}{l}\text{the optimal } w \text{ is} \\ \quad \text{the linear combination of } X\end{array}}$$

• III. Kernel Ridge

- **Kernelized Ridge**

$\overset{def}{=} K$

$\nearrow$ or: $K = \phi^T(X)\,\phi(X)$

$$\omega = X^T \alpha$$

$$y = Xw + e = \underbrace{X \cdot X^T}\alpha + \varepsilon$$

$$L_2 = \|\varepsilon\|_2 + \lambda \|w\|_2 = (y - K\alpha)^T (y - K\alpha) + \lambda \alpha^T \underbrace{XX^T}_{\to K} \alpha$$

$$L_2 = y^T y - \alpha^T K^T y - y^T K \alpha + \alpha^T K^T K \alpha + \lambda \alpha^T K \alpha$$

$$\frac{\partial}{\partial \alpha} L_2 = -2K^T y + 2K^T K \alpha + 2\lambda K^T = 0$$

$$\Rightarrow \boxed{\alpha = (K + \lambda I_n)^{-1} y}$$

• III. Kernel Ridge

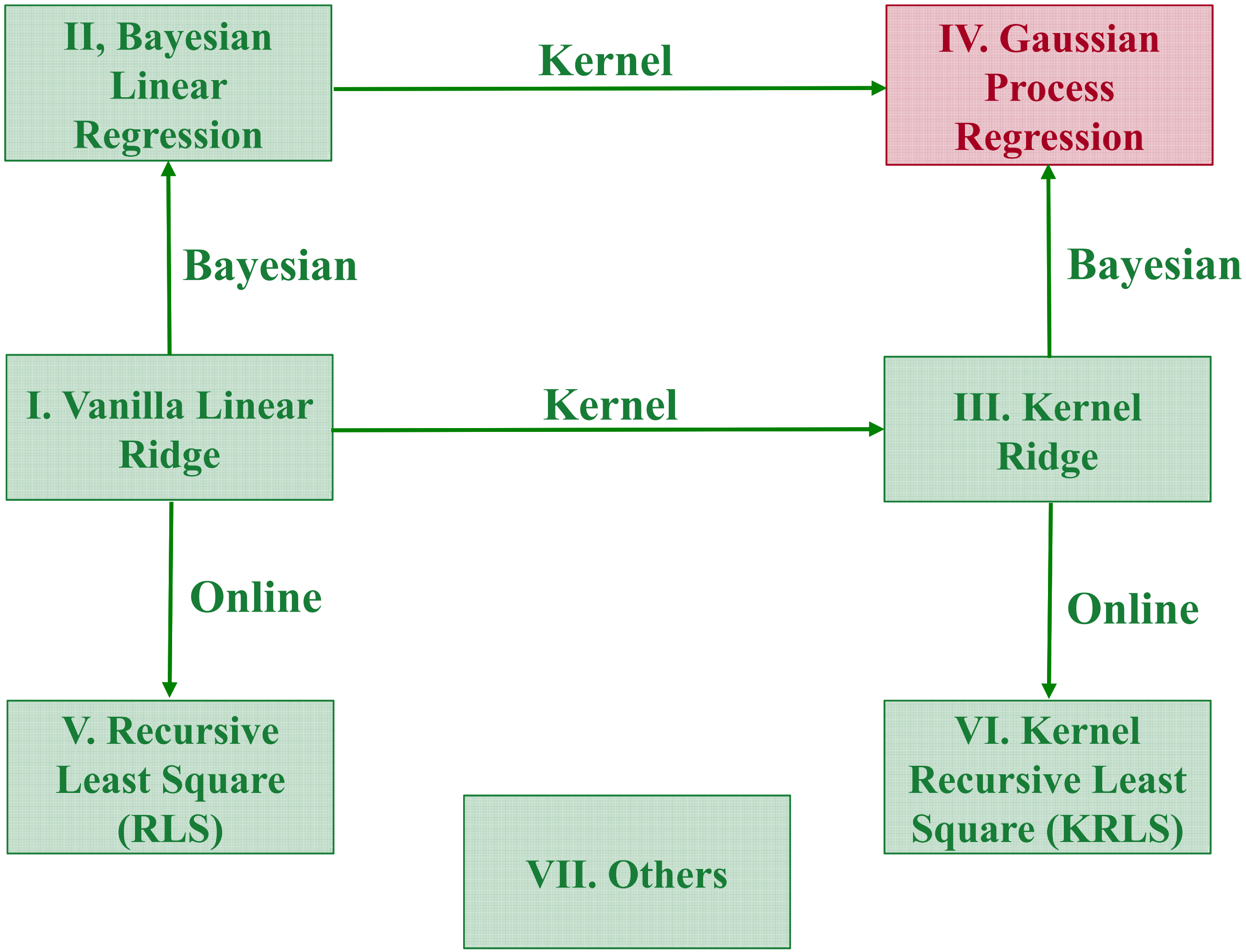
- **Big-O Analysis**

For a new entry x

$$\underset{1\times m}{\underline{y}} = \underset{1\times n}{\underline{K(x, X)}} \cdot \underset{n\times m}{\underline{\alpha}} + \varepsilon$$

	Linear Ridge	Kernel Ridge
Training Time	$O\left(d^3 + Nd^2\right)$	$O\left(N^3\right)$
Predicting Time	$O(d)$	$O(d)$
Model Size	$O(d)$	$O(d) + O(Nd) = O(Nd)$

Outline


II, Bayesian Linear Regression
Kernel
IV. Gaussian Process Regression
Bayesian
Bayesian
I. Vanilla Linear Ridge
Kernel
III. Kernel Ridge
Online
Online
V. Recursive Least Square (RLS)
VII. Others
VI. Kernel Recursive Least Square (KRLS)

• IV. Gaussian Process Regression

- **Idea**

a) Kernel as Covariance function

$$k(x, x') = \sigma_f^2 \exp\left[\frac{-(x-x')^2}{2l^2}\right] + \sigma_n^2 \delta(x, x')$$

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix}$$

$$K_* = \begin{bmatrix} k(x_*, x_1) & k(x_*, x_2) & \cdots & k(x_*, x_n) \end{bmatrix} \qquad K_{**} = k(x_*, x_*)$$

b) Sample from a multivariate Gaussian

$$\begin{bmatrix} \mathrm{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_*^{\mathrm{T}} \\ K_* & K_{**} \end{bmatrix}\right)$$

$$y_* | \mathrm{y} \sim \mathcal{N}(K_* K^{-1} \mathrm{y},\ K_{**} - K_* K^{-1} K_*^{\mathrm{T}})$$

c) Fit the **distribution of functions**

• IV. Gaussian Process Regression

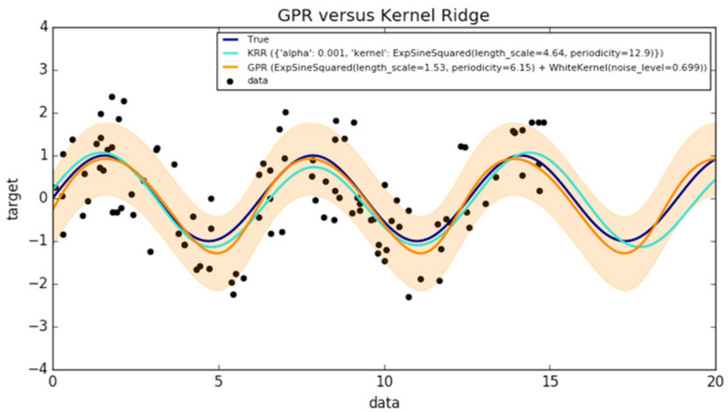
- **Hyperparameter as Optimization**

a) Use Conjugate Gradients or other optimizer to optimize the likelihood => **Bayesian Optimization**

$$\theta = \{l, \sigma_f, \sigma_n\}$$

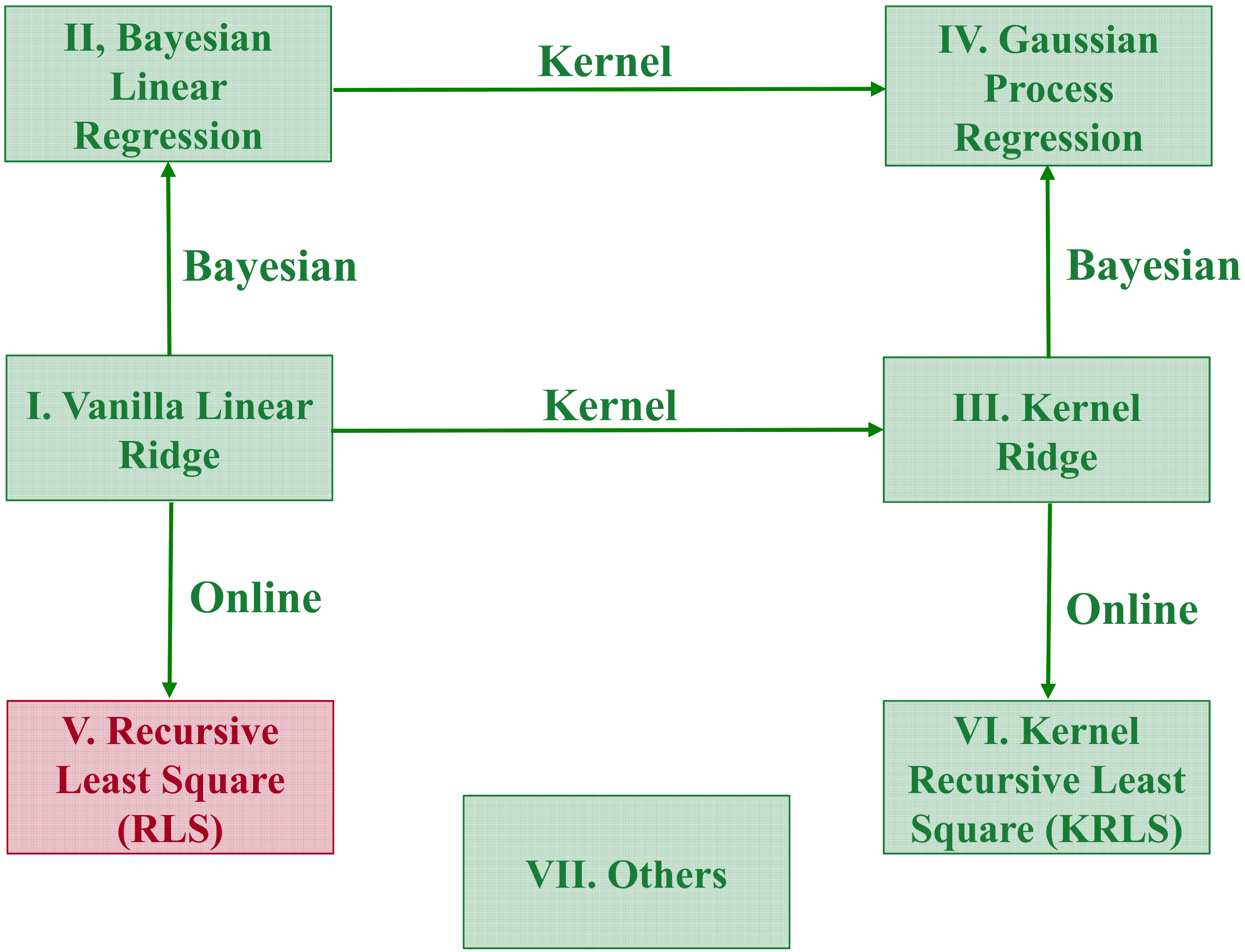
$$\log p(\mathrm{y}|\mathrm{x}, \theta) = -\frac{1}{2}\mathrm{y}^{\mathrm{T}}K^{-1}\mathrm{y} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi$$

b) GPR and Kernel Ridge



- *Time for KRR fitting: 10.243*
- *Time for GPR fitting: 0.223*
- *Time for KRR prediction: 0.083*
- *Time for GPR prediction: 0.091*
- *Time for GPR prediction with standard-deviation: 0.386*

Outline



II, Bayesian Linear Regression
Kernel
IV. Gaussian Process Regression
Bayesian
Bayesian
I. Vanilla Linear Ridge
Kernel
III. Kernel Ridge
Online
Online
V. Recursive Least Square (RLS)
VII. Others
VI. Kernel Recursive Least Square (KRLS)

• V. Recursive Least Square (RLS)

- **Formalization (1)**

$$Y_N = X_N \cdot w_N + \varepsilon$$

$$w_N = (\underbrace{X_N^T X_N + \lambda I_d}_{S_N})^{-1} X_N^T Y_N$$

$$S_N = X_N^T X_N + \lambda I_d$$

$$S_{N+1} = X_{N+1}^T X_{N+1} + \lambda I_d = [X_N^T \; x_{n+1}^T] \begin{bmatrix} X_N \\ x_{N+1} \end{bmatrix} + \lambda I_d$$

$$\Rightarrow \boxed{S_{N+1} = S_N + x_{N+1}^T x_{N+1}}$$

$$S_N w_N = X_N^T Y_N$$

$$S_{N+1} w_{N+1} = X_{N+1}^T Y_{N+1} = X_N^T Y_N + x_{n+1}^T y_{n+1}$$

$$\Rightarrow w_{N+1} = w_N + S_{N+1}^{-1} x_{n+1}^T (y_{n+1} - x_{n+1} \cdot w_n)$$

• V. Recursive Least Square (RLS)

- **Formalization (2)**

$$e_n = y_{n+1} - x_{n+1} w_n$$

$$S_{N+1} = S_N + x_{N+1}^T x_{N+1}$$

$$w_{N+1} = w_N + \underline{S_{N+1}^{-1}} x_{N+1}^T e_N$$

Can it be better ?

$$\boxed{\begin{array}{l} ① (A+B)^{-1} = A^{-1} - \frac{1}{1+g} A^{-1} B A^{-1} \\ \quad g = \mathrm{trace}(BA^{-1}) \\ ② \text{ trace trick} \\ \quad tr(uvw) = tr(wuv) = tr(vwu) \end{array}}$$

• V. Recursive Least Square (RLS)

- **Computation Optimization**

$$S_{N+1}^{-1} = (S_N + x_{N+1}^T x_{N+1})^{-1}$$

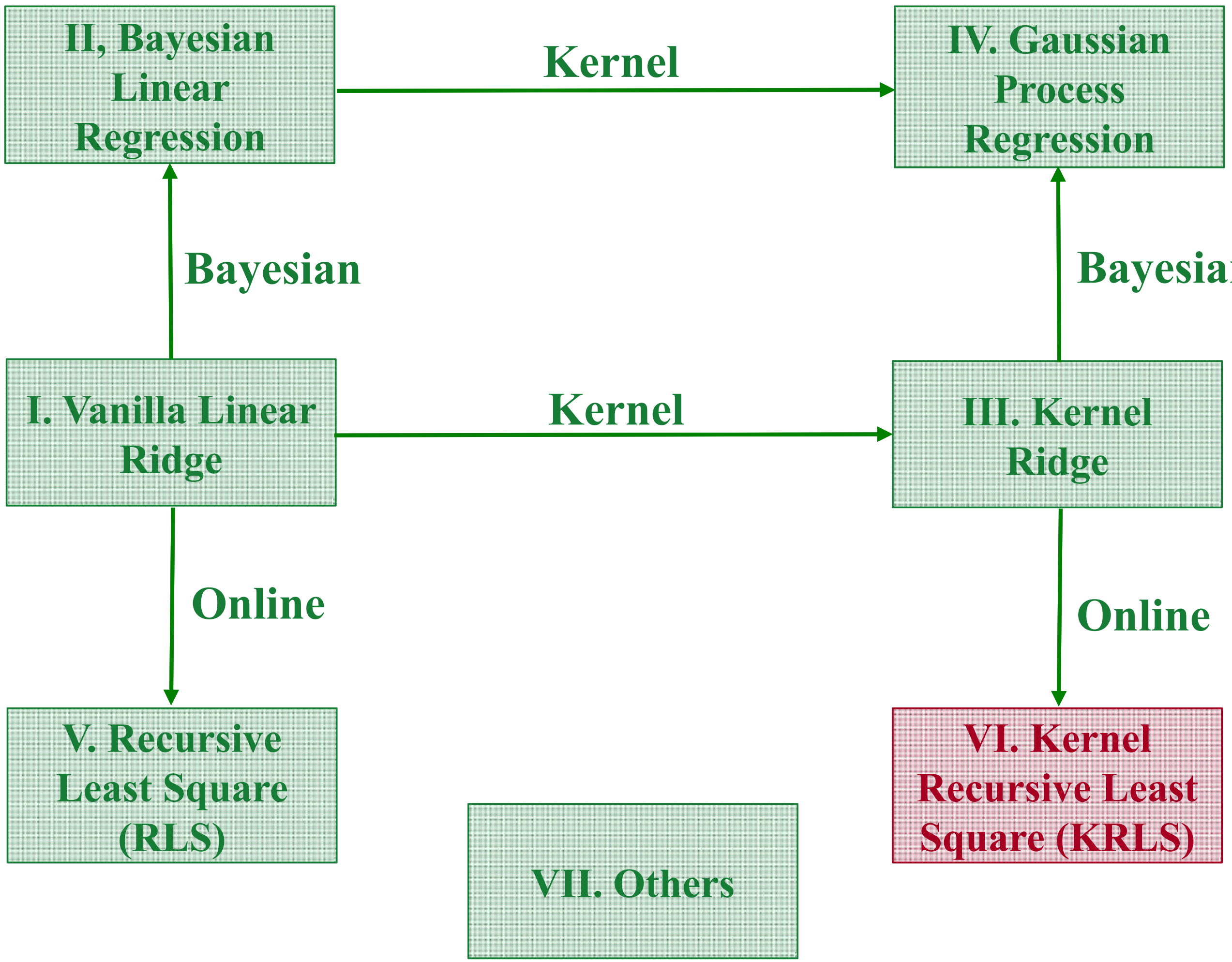
$$= S_N^{-1} - \frac{1}{1+\underline{tr(x_{N+1}^T x_{N+1} S_N^{-1})}} S_N^{-1} x_{N+1}^T \cdot x_{N+1} S_N^{-1}$$

$O(d^3)$ to $O(d^2)$

$$tr(x_{N+1}^T x_{N+1} S_N^{-1}) = tr(x_{N+1} S_N^{-1} x_{N+1}^T)$$

$$= x_{N+1} S_N^{-1} x_{N+1}^T$$ scalar

Outline


II, Bayesian Linear Regression
Kernel
IV. Gaussian Process Regression
Bayesian
Bayesian
I. Vanilla Linear Ridge
Kernel
III. Kernel Ridge
Online
Online
V. Recursive Least Square (RLS)
VII. Others
VI. Kernel Recursive Least Square (KRLS)

• VI. Kernel Recursive Least Square (KRLS)

- **Infinite Dictionary**

- **The core step is to get the inverse of new kernel $O(N^2)$**

$$\dot{\mathbf{K}} = \mathbf{K} + c\mathbf{I} \qquad \dot{\mathbf{K}}_n = \begin{bmatrix} \dot{\mathbf{K}}_{n-1} & \mathbf{k}_n \\ \mathbf{k}_n^\top & k_{nn} + c \end{bmatrix}$$

$$\mathbf{a}_n = \dot{\mathbf{K}}_{n-1}^{-1}\mathbf{k}_n \qquad \gamma_n = k_{nn} + c - \mathbf{k}_n^\top \mathbf{a}_n$$

$$\dot{\mathbf{K}}_n^{-1} = \frac{1}{\gamma_n}\begin{bmatrix} \gamma_n \dot{\mathbf{K}}_{n-1}^{-1} + \mathbf{a}_n \mathbf{a}_n^\top & -\mathbf{a}_n \\ -\mathbf{a}_n & 1 \end{bmatrix}$$

- **The new instance parameters will be easy to get $O(N^2)$**

$$\boldsymbol{\alpha}_{n-1} = \dot{\mathbf{K}}_{n-1}^{-1}\mathbf{y}_{n-1} \qquad \hat{y}_n = \mathbf{k}_n^\top \boldsymbol{\alpha}_{n-1} \qquad e_n = y_n - \hat{y}_n$$

$$\boldsymbol{\alpha}_n = \begin{bmatrix} \boldsymbol{\alpha}_{n-1} - \mathbf{a}_n e_n / \gamma_n \\ e_n / \gamma_n \end{bmatrix}$$

- **There are many criteria for growing and pruning dictionary**

• VI. Kernel Recursive Least Square (KRLS)

- **Dictionary Update**

criterion	type	complexity
all	growing	—
coherence	growing	$\mathcal{O}(m)$
ALD	growing	$\mathcal{O}(m^2)$
oldest	pruning	—
least weight	pruning	$\mathcal{O}(m)$
least a posteriori SE	pruning	$\mathcal{O}(m^2)$

• VI. Kernel Recursive Least Square (KRLS)

- Sliding-Window

- Very simple update to the dictionary: remove the oldest

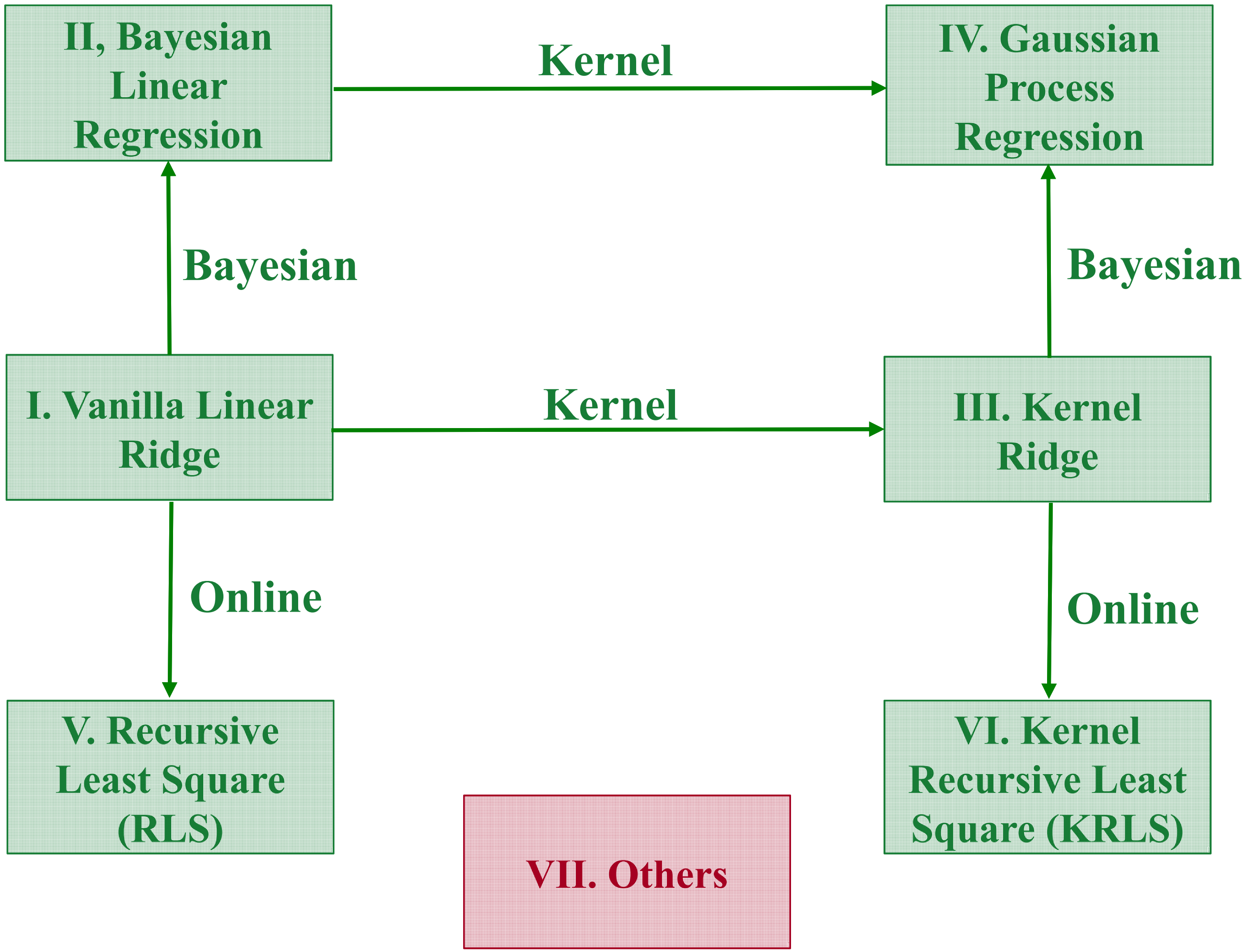
- Suitable for Time Series

- Inverse Kernel $\dot{\mathrm{K}}_{n-1} = \begin{bmatrix} a & \mathrm{b}^T \\ \mathrm{b} & \mathrm{D} \end{bmatrix}$ $\dot{\mathrm{K}}_{n-1}^{-1} = \begin{bmatrix} e & \mathrm{f}^T \\ \mathrm{f} & \mathrm{G} \end{bmatrix}$

$$\mathrm{D}^{-1} = \mathrm{G} - \mathrm{f}\mathrm{f}^T/e.$$

- Remove the oldest instance, and its instance parameter

Outline



II, Bayesian Linear Regression
Kernel
IV. Gaussian Process Regression
Bayesian
Bayesian
I. Vanilla Linear Ridge
Kernel
III. Kernel Ridge
Online
Online
V. Recursive Least Square (RLS)
VII. Others
VI. Kernel Recursive Least Square (KRLS)

• VII. Others

- **Kalman Filter for Online Linear Regression**

θ_t is the $\boxed{\text{state}}$, y_t is the $\boxed{\text{observation}}$.

$$\begin{cases} \theta_t = a_t \theta_{t-1} + w_t \\ y_t = F_t^T \theta_t + v_t \end{cases}$$

Online linear regression?

$\theta_t \leftarrow \beta_t$, $y_t = \beta_t^T x_t + \varepsilon$

$$\begin{cases} \beta_t = \overbrace{\boxed{I}}^{a_t} \cdot \beta_{t-1} + w_t \\ y_t = \underbrace{\boxed{x_t}}_{F_t^T} \cdot \beta_t + v_t \end{cases}$$

Bibliography

- Online regression with kernel
- Gaussian Processes: A Quick Introduction
- CS229 Lecture note of Gaussian Process
- Bayesian Linear Regression YouTube
- Recursive Least Square (RLS) Otexts
- Dynamic Hedge Ratio Between ETF Pairs Using the Kalman Filter

Thanks for listening!