Gaussian Processes and Bayesian Linear Regression

Yung-Kyun Noh

GP as an Infinite Dimensional Gaussians

Ex) $x_{t+1} = a x_t + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$

$E[\varepsilon_t] = 0$

$E[\varepsilon_t^2] = \sigma^2$

$E[\varepsilon_t \varepsilon_{t'}] = 0 \quad t \neq t'$

$$x_t = a x_{t-1} + \varepsilon_{t-1}$$
$$= a(a x_{t-2} + \varepsilon_{t-2}) + \varepsilon_{t-1}$$
$$= a^2 x_{t-2} + a\varepsilon_{t-2} + \varepsilon_{t-1}$$
$$= a^3 x_{t-3} + a^2 \varepsilon_{t-3} + a\varepsilon_{t-2} + \varepsilon_{t-1}$$
$$= \cdots$$
$$= \sum_{i=0}^{\infty} a^i \varepsilon_{t-1-i}$$

$$E[x_t] = \sum_{i=0}^{\infty} a^i E[\varepsilon_{t-1-i}] = 0$$

$$E[x_t x_{t+\Delta t}] = E\left(\sum_{i=0}^{\infty} a^i \varepsilon_{t-1-i}\right)\left(\sum_{j=0}^{\infty} a^j \varepsilon_{t+\Delta t-1-j}\right)$$

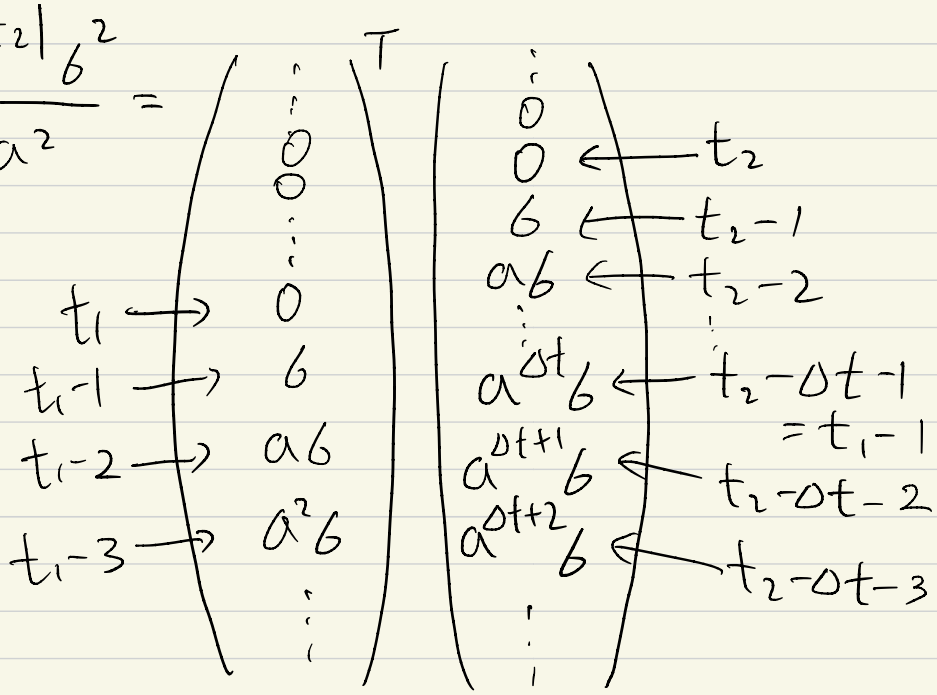
$t - 1 - i = t + \Delta t - 1 - j \Rightarrow j = i + \Delta t$

$$\left(\cdots \; a^2\varepsilon_{t-3} \;\; a\varepsilon_{t-2} \;\; \varepsilon_{t-1} \right.$$
$$\left. \cdots \; a^{\Delta t+2}\varepsilon_{t-3} \; a^{\Delta t+1}\varepsilon_{t-2} \; a^{\Delta t}\varepsilon_{t-1} \; \cdots \; a\varepsilon_{t+\Delta t-2} \;\; \varepsilon_{t+\Delta t-1} \right)$$

$$= \sum_{i=0}^{\infty} a^i a^{i+\Delta t} E[\varepsilon_{t-1-i}^2] = \sum_{i=0}^{\infty} a^{2i+\Delta t} \sigma^2 = \frac{a^{\Delta t}\sigma^2}{1-a^2}$$

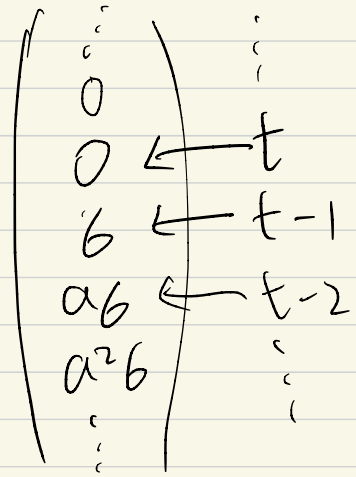
$$m(t) = 0, \quad k(t,t') = \frac{a^{|t-t'|}\,b^2}{1-a^2}$$

$$k(t_1,t_2) = \frac{a^{|t_1-t_2|}\,b^2}{1-a^2} =$$



$$= \phi(t_1)^T \phi(t_2)$$

$$\phi(t) \equiv$$



Covariance
= Inner product of two vectors in ϕ-space.

Note:
Inner product matrix is a p.d. matrix

$$K \equiv \begin{pmatrix} \phi(x_1)^T\phi(x_1) & \phi(x_1)^T\phi(x_2) & \cdots & \phi(x_1)^T\phi(x_N) \\ \vdots & & & \\ \phi(x_N)^T\phi(x_1) & \cdots & & \phi(x_N)^T\phi(x_N) \end{pmatrix}$$

For a nonzero $\vec{c} \in \mathbb{R}^N$,

$$\vec{c}^T K \vec{c} = \sum_{i,j=1}^{N} c_i K_{ij} c_j = \sum_{i,j=1}^{N} c_i \phi(x_i)^T \phi(x_j) c_j$$

$$= \left(\sum_{i=1}^{N} c_i \phi(x_i) \right)^T \left(\sum_{j=1}^{N} c_j \phi(x_j) \right) > 0$$

$\therefore$ Inner product matrix K is p.d.

$$E[y|x, \mathcal{D}] = x^T \left(XX^T + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} X y \quad \cdots (1)$$

$$\left(XX^T + \frac{\sigma^2}{\sigma_0^2} I\right) X = XX^T X + \frac{\sigma^2}{\sigma_0^2} X$$

$$= X\left(X^T X + \frac{\sigma^2}{\sigma_0^2} I\right)$$

$A'' \qquad\qquad = B$

$$AX = XB$$

$$A^{-1} A X B^{-1} = A^{-1} X B B^{-1}$$

$$XB^{-1} = A^{-1} X$$

$$\Rightarrow X\left(X^T X + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} = \left(XX^T + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} X$$

$$\therefore (1) = x^T X \left(X^T X + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} y$$

$$= k^T \left(K + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} y \quad \Leftarrow \text{Same as GP mean}$$

If $K_{ij} = x_i^T x_j$, $k_i = x^T x_i$.

When we consider GP with covariance

$$k(t_1, t_2) = \frac{a^{|t_1 - t_2|} b^2}{1 - a^2},$$ GP mean after

observation is the predictive mean of the linear Bayesian model in the $\phi(t)$ space, with $y = w^T \phi(x)$ and Gaussian prior on w.

Note: Inner product of two $\phi(x_1), \phi(x_2)$ is the covariance between $y(x_1)$ and $y(x_2)$,

$$k(x_1, x_2) = E[y(x_1) y(x_2)] - E[y(x_1)] E[y(x_2)]$$

Ex) $k(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$

$$\left(\exp(t) = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots \right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}\{x^2 + z^2 - 2x^T z\}\right)$$

$$= \frac{\exp\left(\frac{x^T z}{\sigma^2}\right)}{\exp\left(\frac{x^2}{\sigma^2}\right)^{\frac{1}{2}} \exp\left(\frac{z^2}{\sigma^2}\right)^{\frac{1}{2}}}$$

$$= \frac{1 + \frac{x^T z}{\sigma^2} + \frac{1}{2!}\left(\frac{x^T z}{\sigma^2}\right)^2 + \cdots}{\sqrt{1 + \frac{x^2}{\sigma^2} + \frac{1}{2!}\left(\frac{x^2}{\sigma^2}\right)^2 + \cdots}\sqrt{1 + \frac{z^2}{\sigma^2} + \frac{1}{2!}\left(\frac{z^2}{\sigma^2}\right)^2 + \cdots}}$$

$$= \frac{\psi(x)^T \psi(z)}{\|\psi(x)\| \, \|\psi(z)\|}$$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \quad \psi(x) = \begin{pmatrix} 1 \\ x_1/\sigma \\ x_2/\sigma \\ \vdots \\ x_D/\sigma \\ x_1^2/\sqrt{2}\sigma^2 \\ \vdots \\ x_D^2/\sqrt{2}\sigma^2 \\ x_1 x_2/\sigma^2 \\ x_1 x_3/\sigma^2 \\ \vdots \\ x_{D-1} x_D/\sigma^2 \\ x_1^3/\sqrt{3!}\,\sigma^3 \\ \vdots \\ x_D^3/\sqrt{3!}\,\sigma^2 \\ \vdots \end{pmatrix}$$

$$k(x, z) = \phi(x)^T \phi(z)$$

$$\phi(x) = \frac{\psi(x)}{\|\psi(x)\|}$$

$$(x^T z)^2 = (x_1 z_1 + \cdots + x_D z_D)^2$$
$$= x_1^2 z_1^2 + \cdots + x_D^2 z_D^2 + 2x_1 x_2 z_1 z_2 + \cdots + 2x_{D-1} x_D z_{D-1} z_D$$

$$(x^T z)^3 = (x_1 z_1 + \cdots + x_D z_D)^3$$
$$= x_1^3 z_1^3 + \cdots + x_D^3 z_D^3 + 3x_1^2 x_2 z_1^2 z_2 + \cdots + 6x_1 x_2 x_3 z_1 z_2 z_3 + \cdots + 6x_{D-2} x_{D-1} x_D z_{D-2} z_{D-1} z_D$$

$\sigma^2 \uparrow$ The element values decay fast

$\Rightarrow \phi(x)$ maps data to effectively low-dimensional space.

$\Rightarrow$ Flexibility $\downarrow$.

$\sigma^2 \downarrow$ The element values decay slow.

$\Rightarrow \phi(x)$ maps data to effectively high-dimensional space

$\Rightarrow$ Flexibility $\uparrow$

Easy to overfit.

Data away to the training data becomes orthogonal to the training data.

$k(x_i, x_j) = \phi^T(x_i)\phi(x_j) = 0$.

Bayesian linear predictive density:

$$E[y|x,\mathcal{Z}] = k^T\left(K + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} y$$

$$Var[y|x,\mathcal{Z}] = \sigma^2 + \sigma_0^2 k(x,x) - \sigma_0^2 k^T\left(K + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} k$$

GPs

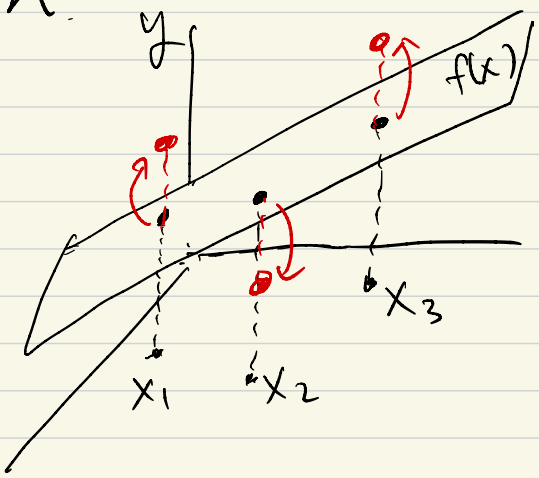
$$\mu(x|\mathcal{Z}) = k^T K^{-1} y \quad \leftarrow \mu \text{ at } x$$

$$k(x,x|\mathcal{Z}) = k(x,x) - k^T K^{-1} k$$

$$k_{GP}(x,z) = \sigma_0^2 k_{BL}(x,z) + \sigma^2 \delta_{x,z}$$

$$\delta_{x,z} = \begin{cases} 1 & \text{if } x = z \\ 0 & \text{otherwise} \end{cases}$$

Linear Regression.

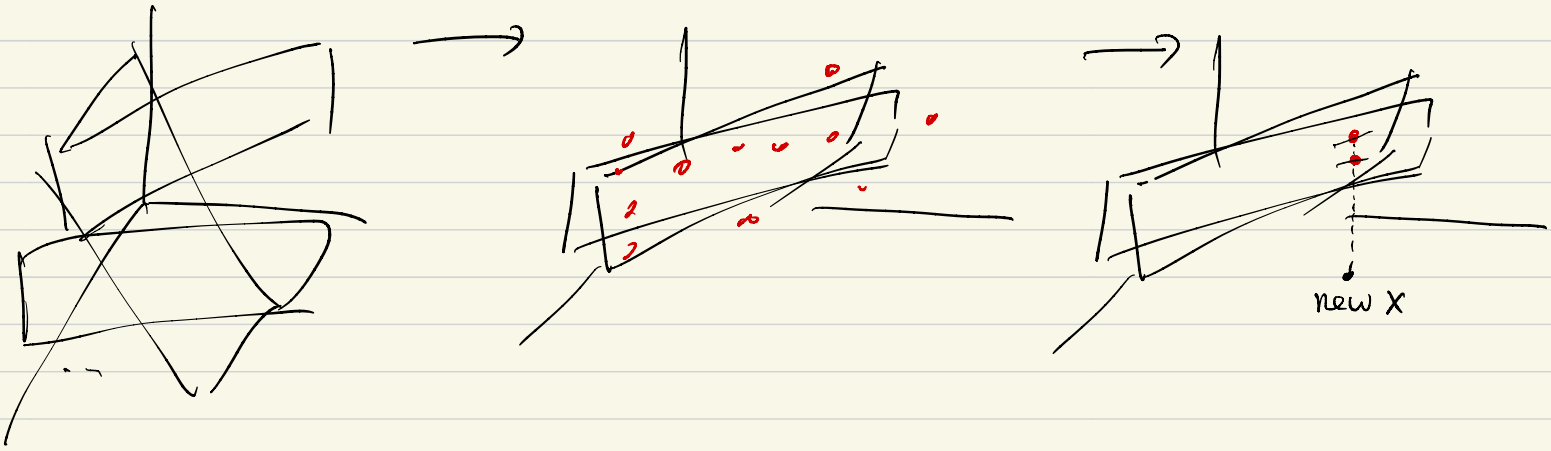


$$y = w^T x + \varepsilon$$

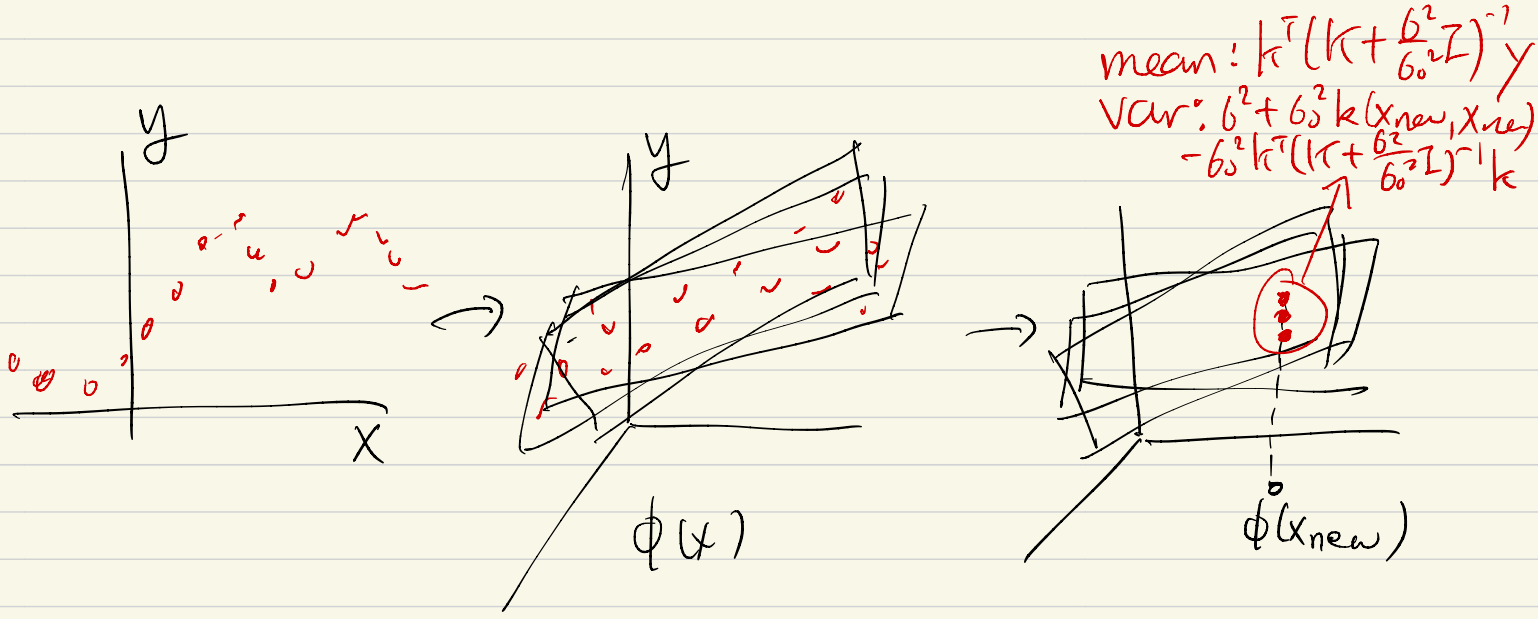
$$= f(x) + \varepsilon$$

Bayesian Linear Regression. (BLR)

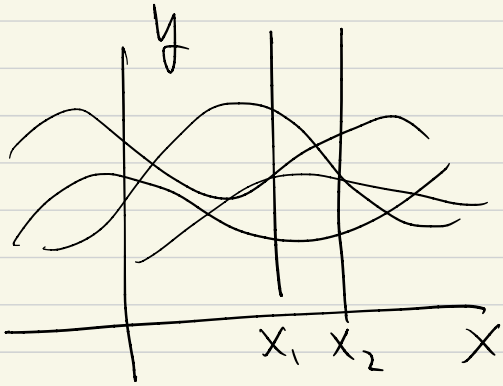
Prior for w → Data → Posterior for w → Prediction.



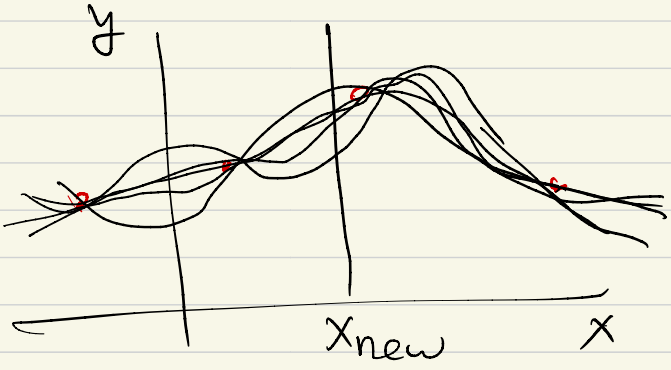
BLR with kernels.



GPs.



Data →



$\uparrow$ $\uparrow$

$\text{Cov}(y(x_1), y(x_2))$

$= k_{GP}(x_1, x_2)$

mean: $k_{GP}^T K_{GP}^{-1} y$

var: $k_{GP}(x_{new}, x_{new})$

$- k_{GP}^T K_{GP}^{-1} k_{GP}$

$k_{GP}(x_i, x_j)$

$= \sigma_0^2 k_{BLR}(x_i, x_j) + \sigma^2 \delta_{x_i, x_j}$