* Multivariate Gaussian dist of $y = f(x) \rightarrow$ "prior" : mean = 0 , covariance = $k(x, x')$

# Gaussian Process

1. Regression

⇒ making inferences about the relationship between inputs and targets

( conditional distribution of the targets given the inputs. )

① Weight space view.

- standard linear model

: $f(x) = x^T w$ , $y = f(x) + \varepsilon$ where $\varepsilon \sim N(0, \sigma_n^2)$

⇒ likelihood ( $p(y \mid X, w)$ )

→ $p(y \mid X, w) = \prod_{i=1}^{n} p(y_i \mid x_i, w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y_i - x_i^T w)^2}{2\sigma_n^2}\right)$

$= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2} \lvert y - X^T w \rvert^2\right) = N(X^T w, \sigma_n^2 I)$

⇒ prior ( $p(w)$ )

→ $p(w) = N(0, \Sigma_p)$

⇒ posterior ( $p(w \mid y, X)$ )

→ $p(w \mid y, X) \propto p(y \mid X, w) \cdot p(w) = N(X^T w, \sigma_n^2 I) \cdot N(0, \Sigma_p)$

$\propto \exp\left(-\frac{1}{2\sigma_n^2}(y - X^T w)^T (y - X^T w)\right) \cdot \exp\left(-\frac{1}{2} w^T \Sigma_p^{-1} w\right)$

$\propto \exp\left(-\frac{1}{2}(w - \bar{w})^T \left(\frac{1}{\sigma_n^2} X X^T + \Sigma_p^{-1}\right)(w - \bar{w})\right)$

where $\bar{w} = \frac{1}{\sigma_n^2}\left(\frac{1}{\sigma_n^2} X X^T + \Sigma_p^{-1}\right)^{-1} X y$

→ $p(w \mid y, X) = N\left(\frac{1}{\sigma_n^2} A^{-1} X y, A^{-1}\right)$ where $A = \frac{1}{\sigma_n^2} X X^T + \Sigma_p^{-1}$ (D×D)

↳ MAP estimate ≈ penalized MLE (ridge regression)

⇒ predictive posterior ( $p(f_* \mid x_*, y, X)$ )

→ $p(f_* \mid x_*, y, X) = \int p(f_* \mid x_*, w) \cdot p(w \mid y, X)\, dw$

$= N\left(\frac{1}{\sigma_n^2} x_*^T A^{-1} X y, x_*^T A^{-1} x_*\right)$

- kernel trick

⇒ project the inputs into some high dimensional space using a set of basis functions

( $\phi : \mathbb{R}^D \rightarrow \mathbb{R}^N$ where $D < N$ and $f(x) = \phi(x)^T w$ )

⇒ predictive posterior ( $p(f_* \mid x_*, y, X)$ )

→ $p(f_* \mid x_*, y, X) = N\left(\frac{1}{\sigma_n^2} \phi(x_*)^T A^{-1} \phi(X) y, \phi(x_*)^T A^{-1} \phi(x_*)\right)$

where $A = \frac{1}{\sigma_n^2} \phi(X)\phi(X)^T + \Sigma_p^{-1}$ (N×N) → computationally expensive

→ $p(f_* \mid x_*, y, X) = N\Big(\phi(x_*)^T \Sigma_p \phi(X)\left(K + \sigma_n^2 I\right)^{-1} y,$

$\phi(x_*)^T \Sigma_p \phi(x_*) - \phi(x_*)^T \Sigma_p \phi(X)\left(K + \sigma_n^2 I\right)^{-1} \phi(X) \Sigma_p \phi(x_*)\Big)$

$= N\Big(k(x_*, X)\left(k(X, X) + \sigma_n^2 I\right)^{-1} y,$

$k(x_*, x_*) - k(x_*, X)\left(k(X, X) + \sigma_n^2 I\right)^{-1} k(X, x_*)\Big)$

where $K + \sigma_n^2 I = \phi(X)^T \Sigma_p \phi(X) + \sigma_n^2 I$ (n × n) → computationally relieved.

$k(x_i, x_j) = \phi(x_i)^T \Sigma_p \phi(x_j) = \left(\Sigma_p^{\frac{1}{2}} \phi(x_i)\right) \cdot \left(\Sigma_p^{\frac{1}{2}} \phi(x_j)\right) = \psi(x_i) \cdot \psi(x_j)$

↳ covariance function , kernel

② Function-space view

- Gaussian Process

: collection of random variable , any finite number of which have a joint Gaussian distribution

$\left(f(x) \sim GP(m(x), k(x, x'))\right.$

$\left.\text{where } m(x) = \mathbb{E}[f(x)],\ k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))^T]\right)$

⇒ marginalization property (consistency)

→ $(y_1, y_2) \sim N(\mu, \Sigma) \implies y_1 \sim N(\mu_1, \Sigma_{11})$ and $y_2 \sim N(\mu_2, \Sigma_{22})$

(examination of larger set of variables does not change the distribution of smaller set)

⇒ generating samples from GP(m, k)

1) Cholesky decomposition on covariance matrix : $K = LL^T$

( L is a lower triangle matrix )

2) Sample $u \sim N(0, I)$

3) Compute $x = m + L \cdot u$

( $\mathbb{E}[x] = 0$ , $Cov(x, x') = \mathbb{E}[Lu(Lu')^T] = L \cdot \mathbb{E}[uu^T] L^T = K$ )

- zero mean Gaussian Process

⇒ noise free observations

→ $\begin{bmatrix} f \\ f_* \end{bmatrix} \sim GP\left(0, \begin{bmatrix} k(X, X) & k(X, X_*) \\ k(X_*, X) & k(X_*, X_*) \end{bmatrix}\right)$

→ $p(f_* \mid X_*, f, X) = N\left(k(X_*, X)k(X, X)^{-1} f,\ k(X_*, X_*) - k(X_*, X)k(X, X)^{-1} k(X, X_*)\right)$

⇒ noisy observations

→ $\begin{bmatrix} y \\ f_* \end{bmatrix} \sim GP\left(0, \begin{bmatrix} k(X, X) + \sigma_n^2 I & k(X, X_*) \\ k(X_*, X) & k(X_*, X_*) \end{bmatrix}\right)$

→ $p(f_* \mid X_*, y, X) = N\Big(k(X_*, X)\left(k(X, X) + \sigma_n^2 I\right)^{-1} y,$

$k(X_*, X_*) - k(X_*, X)\left(k(X, X) + \sigma_n^2 I\right)^{-1} k(X, X_*)\Big)$

* Interpretation on mean prediction

1) linear combination of observation $y_i$ → weight ≈ (equivalent kernel / Nadaraya-Watson estimator)

2) linear combination of kernel $k(x_*, x_i)$ → representer theorem

* Interpretation on covariance prediction

1) do not depend on observation $y_i$

⇒ marginal likelihood ( $p(y \mid X)$ )

→ $p(y \mid X) = \int p(y \mid f, X) \cdot p(f \mid X)\, df$

$= \int N(f, \sigma_n^2 I) \cdot N(0, k(X, X))\, df$

$= N(0, k(X, X) + \sigma_n^2 I)$

- Squared Exponential kernel

⇒ $k(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}(x_p - x_q)^2\right) + \sigma_n^2 \delta_{pq}$

→ signal variance ($\sigma_f^2$)

→ length scale ($\ell$) → Median Heuristics : $\ell^2 = \frac{\text{median}}{2\log(n+1)}$ (distance among all pairs)

→ noise variance ($\sigma_n^2$)

- Decision theory for regression

: minimize the expected loss w.r.t. the model on what the truth might be.

⇒ $\tilde{R}_L(y_{guess} \mid x_*) = \int L(y_*, y_{guess}) \cdot p(y_* \mid x_*, y, X)\, dy_*$

⇒ $y_{optimal} \mid x_* = \underset{y_{guess}}{\arg\min}\ \tilde{R}_L(y_{guess} \mid x_*)$

$\left(L(y_*, y_{guess}) = \lvert y_{guess} - y_* \rvert \implies y_{optimal} \mid x^* = \text{median}\left(p(y_* \mid x_*, y, X)\right)\right)$

$\left(L(y_*, y_{guess}) = (y_{guess} - y_*)^2 \implies y_{optimal} \mid x^* = \text{mean}\left(p(y_* \mid x_*, y, X)\right)\right)$

$\left(L(y_*, y_{guess}),\ p(y_* \mid x_*, y, X) : \text{symmetric} \implies y_{optimal} \mid x^* = \text{mean}\left(p(y_* \mid x_*, y, X)\right)\right)$

- Non-zero mean Gaussian Process

⇒ with a fixed mean function ($m(\cdot)$)

→ $g(x) \sim GP(m(x), k(x, x'))$

→ $p(g_* \mid X_*, y, X) = N\left(m(X_*) + k(X_*, X)\left(k(X, X) + \sigma_n^2 I\right)^{-1}(y - m(X)),\ Cov(f_*)\right)$

⇒ with a few fixed basis functions ($h(\cdot)$)

: global linear model with residuals modelled by Gaussian Process

→ $g(x) = f(x) + h(x)^T \beta$ where $f(x) \sim GP(0, k(x, x'))$, $\beta \sim N(b, B)$

→ $g(x) \sim GP(h(x)^T b,\ k(x, x') + h(x)^T B\, h(x'))$

→ $p(g_* \mid X_*, y, X) = N\left(\mathbb{E}[f_*] + R^T \bar{\beta},\ Cov(f_*) + R^T (B^{-1} + h(X)\left(k(X, X) + \sigma_n^2 I\right)^{-1} h(X)^T)^{-1} R\right)$

where $\bar{\beta} = \left(B^{-1} + h(X)\left(k(X, X) + \sigma_n^2 I\right)^{-1} h(X)^T\right)^{-1}\left(h(X)\left(k(X, X) + \sigma_n^2 I\right)^{-1} h(X)^T\right)^{-1} R$

$R = h(X_*) - h(X)\left(k(X, X) + \sigma_n^2 I\right)^{-1} k(X, X_*)$

⇒ marginal likelihood ( $p(y \mid X)$ )

→ $p(y \mid X) = \int p(y \mid g, X) \cdot p(g \mid X)\, dg$

$= \int N(g, \sigma_n^2 I) \cdot N(h(X)^T b,\ k(X, X) + h(X)^T B\, h(X))\, dg$

$= N(h(X)^T b,\ k(X, X) + h(X)^T B\, h(X) + \sigma_n^2 I)$

# Sparse Gaussian Process

## 1. A unifying view of sparse approximate Gaussian Process Regression

: understand the theoretical foundations of the various approximations

⇒ every algorithm approximates the joint prior $p(f_*, f)$ by conditional independence

$$\left(p(f_*, f) \approx q(f_*, f) = \int q(f_*|u)\, q(f|u)\, p(u)\, du\right)$$

① GP

$$\rightarrow \begin{bmatrix} f \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K_{ff} & K_{f*} \\ K_{*f} & K_{**} \end{bmatrix}\right)$$

② Subset of Regressor (SoR)

$$\begin{cases} f_* = k_{*u} w_u, & p(w_u) = N(0, K_{uu}^{-1}) \\ u = k_{uu} w_u, & p(u) = N(0, K_{uu}) \end{cases}$$

$$\Rightarrow q(f|u) = N(k_{fu} \underbrace{k_{uu}^{-1} u}_{w_u}, 0) \quad , \quad q(f_*|u) = N(k_{*u} \underbrace{k_{uu}^{-1} u}_{w_u}, 0)$$

$$\Rightarrow q(f, f_*) = N\left(0, \begin{bmatrix} Q_{ff} & Q_{f*} \\ Q_{*f} & Q_{**} \end{bmatrix}\right)$$

$$\left(\because \mathrm{Cov}(f, f_*) = \mathbb{E}[(k_{fu}k_{uu}^{-1}u - 0)(k_{*u}k_{uu}^{-1}u - 0)^T] = \mathbb{E}[k_{fu}k_{uu}^{-1}u \cdot u^T k_{uu}^{-1}k_{uu}] = Q_{f*}\right)$$

② Deterministic Training Conditional (DTC) ⟺ PLV, PP

$$\Rightarrow p(y|f) \approx q(y|f) = N(k_{fu}k_{uu}^{-1}u, \sigma^2 I)$$

$$\Rightarrow q(f|u) = N(k_{fu}k_{uu}^{-1}u, 0) \quad , \quad q(f_*|u) = p(f_*|u)$$

$$\Rightarrow q(f, f_*) = N\left(0, \begin{bmatrix} Q_{ff} & Q_{f*} \\ Q_{*f} & K_{**} \end{bmatrix}\right) \rightarrow \text{violating consistency}$$

$$\left(\because \mathrm{Cov}(f_*, f_*) = \mathbb{E}[(f_* - 0)(f_* - 0)^T] = K_{**}\right)$$

③ Fully Independent Training Conditional (FITC) ⟺ SPGP

$$\Rightarrow p(y|f) \approx q(y|f) = N(k_{fu}k_{uu}^{-1}u, \mathrm{diag}[K_{ff} - Q_{ff}] + \sigma^2 I)$$

$$\Rightarrow q(f|u) = N(k_{fu}k_{uu}^{-1}u, \mathrm{diag}[K_{ff} - Q_{ff}]), \quad q(f_*|u) = p(f_*|u)$$

$$\Rightarrow q(f, f_*) = N\left(0, \begin{bmatrix} Q_{ff} + \mathrm{diag}[K_{ff} - Q_{ff}] & Q_{f*} \\ Q_{*f} & K_{**} \end{bmatrix}\right) \rightarrow \text{violating consistency}$$

$$\left(\because \mathrm{Cov}(f, f) = \mathbb{E}[(k_{fu}k_{uu}^{-1}u)(k_{fu}k_{uu}^{-1}u)^T] + \mathrm{diag}[K_{ff} - Q_{ff}]\right)$$

④ Fully Independent Conditional (FIC)

$$\Rightarrow p(y|f) \approx q(y|f) = N(k_{fu}k_{uu}^{-1}u, \mathrm{diag}[K_{ff} - Q_{ff}] + \sigma^2 I)$$

$$\Rightarrow \begin{cases} q(f|u) = N(k_{fu}k_{uu}^{-1}u, \mathrm{diag}[K_{ff} - Q_{ff}]) \\ q(f_*|u) = N(k_{*u}k_{uu}^{-1}u, \mathrm{diag}[K_{**} - Q_{**}]) \end{cases}$$

$$\Rightarrow q(f, f_*) = N\left(0, \begin{bmatrix} Q_{ff} + \mathrm{diag}[K_{ff} - Q_{ff}] & Q_{f*} \\ Q_{*f} & Q_{**} + \mathrm{diag}[K_{**} - Q_{**}] \end{bmatrix}\right)$$

$$\left(\because \mathrm{Cov}(f_*, f_*) = \mathbb{E}[(k_{*u}k_{uu}^{-1}u)(k_{*u}k_{uu}^{-1}u)^T] + \mathrm{diag}[K_{**} - Q_{**}]\right)$$

⑤ Partially Independent Training Conditional (PITC)

$$\Rightarrow p(y|f) \approx q(y|f) = N(k_{fu}k_{uu}^{-1}u, \mathrm{blockdiag}[K_{ff} - Q_{ff}] + \sigma^2 I)$$

$$\Rightarrow q(f|u) = N(k_{fu}k_{uu}^{-1}u, \mathrm{blockdiag}[K_{ff} - Q_{ff}]), \quad q(f_*|u) = p(f_*|u)$$

$$\Rightarrow q(f, f_*) = N\left(0, \begin{bmatrix} Q_{ff} + \mathrm{blockdiag}[K_{ff} - Q_{ff}] & Q_{f*} \\ Q_{*f} & K_{**} \end{bmatrix}\right) \rightarrow \text{violating consistency}$$

$$\left(\because \mathrm{Cov}(f, f) = \mathbb{E}[(k_{fu}k_{uu}^{-1}u)(k_{fu}k_{uu}^{-1}u)^T] + \mathrm{blockdiag}[K_{ff} - Q_{ff}]\right)$$

⑥ Partially Independent Conditional (PIC)

$$\Rightarrow p(y|f) \approx q(y|f) = N(k_{fu}k_{uu}^{-1}u, \mathrm{blockdiag}[K_{ff} - Q_{ff}] + \sigma^2 I)$$

$$\Rightarrow \begin{cases} q(f|u) = N(k_{fu}k_{uu}^{-1}u, \mathrm{blockdiag}[K_{ff} - Q_{ff}]) \\ q(f_*|u) = N(k_{*u}k_{uu}^{-1}u, \mathrm{blockdiag}[K_{**} - Q_{**}]) \end{cases}$$

$$\Rightarrow q(f, f_*) = N\left(0, \begin{bmatrix} Q_{ff} + \mathrm{blockdiag}[K_{ff} - Q_{ff}] & Q_{f*} \\ Q_{*f} & Q_{**} + \mathrm{blockdiag}[K_{**} - Q_{**}] \end{bmatrix}\right)$$

$$\left(\because \mathrm{Cov}(f_*, f_*) = \mathbb{E}[(k_{*u}k_{uu}^{-1}u)(k_{*u}k_{uu}^{-1}u)^T] + \mathrm{blockdiag}[K_{**} - Q_{**}]\right)$$

## 2. Variational Learning of Inducing Variables

: inducing variables and hyperparameters are jointly selected by maximizing the lower bound of the marginal likelihood to minimize the KL divergence between a variational GP and the true posterior GP

⇒ predictive distribution ($p(f_*|y)$)

$$\rightarrow p(f_*|y) = \int p(f_*|f)\, p(f|y)\, df$$

$$= \int p(f_*|f, u)\, p(f|y, u)\, p(u)\, df\, du \quad \text{where } p(u) = N(0, K_u)$$

$$= \int p(f_*|u)\, p(f|u)\, p(u)\, df\, du$$

$$= \int p(f_*|u)\, p(u)\, du = \int p(f_*, u|y)\, du$$

(u is assumed to be a sufficient statistic for f)

⇒ variational distribution ($q(f_*)$)

$$\rightarrow q(f_*) = \int p(f_*|u)\, p(f|u)\, \phi(u)\, df\, du \quad \text{where } \phi(u) = N(\mu, A)$$

$$= \int p(f_*|u)\, \phi(u)\, du = \int q(f_*, u)\, du$$

$$= N\left(K_{u*}^T K_u^{-1}\mu,\ K_{**} - K_{u*}^T K_u^{-1} K_{u*} + K_{u*}^T K_u^{-1} A K_u^{-1} K_{u*}\right)$$

⇒ Variational learning of z and kernel hyperparameters

$$\rightarrow \arg\min_{z,\mu,A} KL(q(f_*) \| p(f_*|y)) \Leftrightarrow \arg\min_{z,\mu,A} KL(q(f_*, u) \| p(f_*, u|y))$$

$$\rightarrow \log p(y_*|z_*)$$

$$\geq \int q(f_*, u) \log \frac{p(f_*, u, y_*)}{q(f_*, u)}\, df_*\, du$$

$$= \int p(f_*|u)\, \phi(u) \log \frac{p(y_*|f_*)\, p(u)}{\phi(u)}\, df_*\, du$$

$$= \int \phi(u) \left\{ \boxed{\int p(f_*|u) \log p(y_*|f_*)\, df_*} + \log \frac{p(u)}{\phi(u)} \right\} du$$

$$\Downarrow$$

$$\boxed{\begin{aligned} &\int p(f_*|u) \left\{ -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2} Tr\left[y_* y_*^T - 2 y_* f_*^T + f_* f_*^T\right] \right\} df_* \\ &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2} Tr\left[y_* y_*^T - 2 y_* \alpha^T K_{u*} + \alpha K_u \alpha^T + K_{**} - Q_{**}\right] \\ &= \log N(\alpha, \sigma^2 I) - \frac{1}{2\sigma^2} Tr[K_{**} - Q_{**}] \\ &\quad \text{where } \alpha = K_{uf}^T K_u^{-1} u, \quad Q_{**} = K_{uf}^T K_u^{-1} K_{uf} \end{aligned}}$$

$$= \int \phi(u) \log \frac{N(\alpha, \sigma^2 I)\, p(u)}{\phi(u)}\, du - \frac{1}{2\sigma^2} Tr[K_{**} - Q_{**}]$$

$$\geq \log \int N(\alpha, \sigma^2 I)\, p(u)\, du - \frac{1}{2\sigma^2} Tr[K_{**} - Q_{**}] \quad \left(\because \text{Jensen's inequality}\right)$$

$$= \log N(0, K_{uf}^T K_u^{-1} K_{uf} + \sigma^2 I) - \frac{1}{2\sigma^2} Tr[K_{**} - Q_{**}]$$

$$\begin{cases} \mu_{opt} = \sigma^{-2} K_u (K_u + \sigma^{-2} K_{uf} K_{uf}^T)^{-1} K_{uf} y \\ A_{opt} = K_u (K_u + \sigma^{-2} K_{uf} K_{uf}^T)^{-1} K_u \end{cases}$$

3. Sparse Orthogonal Variational Inference for Gaussian Process

: Introduce another set of inducing variables for the orthogonal complement, which is beneficial for the computational cost

⇒ Reinterpreting SVGP

→ $p(f|u) = \mathcal{N}\left(k_{uf}^T k_u^{-1} u,\ k_f - k_{uf}^T k_u^{-1} k_{uf}\right)$

→ $f = k_{uf}^T k_u^{-1} u + f_\perp$ where $f_\perp \sim \mathcal{N}(0,\ k_f - k_{uf}^T k_u^{-1} k_{uf})$

⇒ Orthogonal decomposition

→ $V = \left\{ \sum_{j=1}^{M} \alpha_j k(z_j, \cdot) \right\}$ : linear span of kernel basis function

→ $f = f_{||} + f_\perp$ where $f_{||} \in V$, $f_\perp \perp V$

$$\begin{cases} f_{||} = K_{uf}^T K_u^{-1} u \sim \mathcal{N}(0,\ K_{uf}^T K_u^{-1} K_{uf}) \\ f_\perp = f - f_{||} \sim \mathcal{N}(0,\ K_f - K_{uf}^T K_u^{-1} K_{uf}) \end{cases}$$

⇒ joint distribution ($p(y, u, f_\perp, v)$)

→ $p(y, u, f_\perp, v) = p(y | k_{uf}^T k_u^{-1} u + f_\perp)\, p(u)\, p(f_\perp|v)\, p(v)$ where $p(v) = \mathcal{N}(0, C_{vv})$

⇒ variational distribution ($q(u, f_\perp, v)$)

→ $q(u, f_\perp, v) = q(u)\, p(f_\perp|v)\, q(v)$ where $q(v) = \mathcal{N}(m_v, S_v)$

→ $q(f_\perp) = \int p(f_\perp|v)\, q(v)\, dv = \mathcal{N}(m_{f_\perp}, S_{f_\perp})$

$$\begin{cases} m_{f_\perp} = C_{vf}^T C_v^{-1} m_v \\ S_{f_\perp} = C_{ff} - C_{vf}^T C_v^{-1} (C_v - S_v) C_v^{-1} C_{vf} \end{cases} \quad \text{where } C_{ab} = k_{ab} - K_{ua}^T K_u^{-1} k_{ub}$$

⇒ variational learning of $z, \theta, m_v, S_v$ and kernel hyperparameters

→ $\log p(y|x)$

$$\geq \boxed{\mathbb{E}_{q(u) q(f_\perp)}\left[\log p(y | k_{uf}^T k_u^{-1} u + f_\perp)\right]} - KL\left[q(u) \| p(u)\right] - KL\left[q(v) \| p(v)\right]$$

$$\Downarrow$$

$$\begin{aligned} &= \mathbb{E}_{q(u) q(f_\perp)}\left[\log \mathcal{N}(k_{uf}^T k_u^{-1} u + f_\perp, \sigma^2 I)\right] \\ &= \mathbb{E}_{q(u) q(f_\perp)}\left[-\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - k_{uf}^T k_u^{-1} u - f_\perp)^T (y - k_{uf}^T k_u^{-1} u - f_\perp)\right] \\ &= \mathbb{E}_{q(u)}\left[\log \mathcal{N}(k_{uf}^T k_u^{-1} u + m_{f_\perp}, \sigma^2 I) - \frac{1}{2\sigma^2}\mathrm{Tr}(S_{f_\perp})\right] \end{aligned}$$

$$\geq \int \log \mathcal{N}(k_{uf}^T k_u^{-1} u + m_{f_\perp}, \sigma^2 I)\, p(u)\, du - \frac{1}{2\sigma^2}\mathrm{Tr}(S_{f_\perp}) - KL\left[q(v) \| p(v)\right]$$

$$= \log \mathcal{N}(m_{f_\perp}, k_{uf}^T k_u^{-1} k_{fu} + \sigma^2 I) - \frac{1}{2\sigma^2}\mathrm{Tr}(S_{f_\perp}) - KL\left[\mathcal{N}(m_v, S_v) \| \mathcal{N}(0, C_{vv})\right]$$

$$\begin{cases} m_v^{opt} = C_v \left[C_v + C_{vf}(k_{uf}^T k_u^{-1} k_{fu} + \sigma^2 I)^{-1} C_{vf}^T\right]^{-1} C_{vf} (k_{uf}^T k_u^{-1} k_{fu} + \sigma^2 I)^{-1} y \\ S_v^{opt} = C_v \left[C_v + \sigma^{-2} C_{vf} C_{vf}^T\right]^{-1} C_v \end{cases}$$

☆ If $m_v = 0$, $S_v = C_v$, then it reduces to SVGP

☆ If $S_v = C_v$, then it reduces to follows

: $\log \mathcal{N}(m_{f_\perp}, k_{uf}^T k_u^{-1} k_{fu} + \sigma^2 I) - \frac{1}{2\sigma^2}\mathrm{Tr}(C_f) - \frac{1}{2} m_v^T C_v^{-1} m_v$

# Gaussian Process Extrapolation

1. Gaussian Process kernels for pattern discovery and extrapolation
   : derive kernel by modeling a spectral density with a gaussian mixture