Sparse Greedy Gaussian Process Regression

Han Chen

1 Maximize log posterior

Finite set inputs $X = \{x_1, \cdots, x_m\}$. $y(x) = t(x) + \xi$ where $\xi \sim \mathcal{N}(0, \sigma^2)$ and $(t_m) \sim \mathcal{N}(0, K)$.

Instead, assume y is generated by

$$y = K\alpha + \xi$$

where $\alpha \sim \mathcal{N}(0, K^{-1})$ and $\xi \sim \mathcal{N}(0, \sigma^2 \mathbf{1})$.
The posterior $p(\alpha|y, X)$ is proportional to

$$\Pi = \exp(-\frac{1}{2\sigma^2}\|y - K\alpha\|^2)\exp(-\frac{1}{2}\alpha^T K\alpha)$$

Let the maximizer be α_{opt}. Conditional expectation for $y(x)$ (new x) is

$$\mathbb{E}[y(x)|y, X] = k^T \alpha_{opt} \, ,$$

where $k = \big(k(x_1, x), \cdots, k(x_m, x)\big)$. We have

$$-\sigma^2 \log \Pi - \frac{1}{2}y^T y = -y^T K\alpha + \frac{1}{2}\alpha^T \big(\sigma^2 K + K^T K\big)\alpha$$

Therefore, α_{opt} minimizes $-\sigma^2 \log \Pi - \frac{1}{2}y^T y$.
Posterior mean $k^T(K + \sigma^2\mathbf{1})^{-1}y$, posterior variance $k(x, x) + \sigma^2 - k^T(K + \sigma^2\mathbf{1})^{-1}k$.
We have

$$\alpha_{opt} = (K + \sigma^2\mathbf{1})^{-1}y$$

2 Inequalities

For any positive semidefinite square matrix K, and vectors v, ξ, η, define

$$Q_v(\xi) \equiv -v^T K\xi + \frac{1}{2}\xi^T(\sigma^2 K + K^T K)\xi$$
$$Q_v^*(\eta) \equiv -v^T \eta + \frac{1}{2}\eta^T(\sigma^2\mathbf{1} + K)\eta$$

For all ξ, η, we have

$$Q_v(\xi) \geq Q_v^{\min} \geq -\frac{1}{2}\|v\|^2 - \sigma^2 Q_v^*(\eta)$$
$$Q_v^*(\eta) \geq Q_v^{*\ \min} \geq \sigma^{-2}\left(-\frac{1}{2}\|v\|^2 - Q_v(\xi)\right)$$

Equalities hold when $Q_v(\xi) = Q_v^{\min}$ and $Q_v(\eta) = Q_v^{*\ \min}$, that is $\alpha = \alpha_{opt}$ (notice $\xi, \eta = \alpha_{opt}$ minimizes both $Q_v(\xi)$ and $Q_v^*(\xi)$).

3 Error bounds

We have

$$\mathrm{Var}\big[y(x)\big|y, X\big] = k(x,x) + \sigma^2 + 2Q_k^{*\,min} \le k(x,x) + \sigma^2 + 2Q_k^*(\eta)$$

for any η, which gives an upper bound of the variance.
The lower bound of the variance is given by

$$\mathrm{Var}\big[y(x)\big|y, X\big] \ge k(x,x) + \sigma^2 + 2\sigma^{-2}\left(-\frac{1}{2}\|k\|^2 - Q_k(\xi)\right)$$

for any ξ.
Define "gap" to be

$$\frac{\text{Upper bound - lower bound}}{\text{Average variance reduction computed from upper/lower bound}}$$

We have

$$\mathrm{gap}(\xi,\eta) = \frac{2\Big(Q_k(\xi) + \sigma^2 Q_k^*(\eta) + \frac{1}{2}\|k\|^2\Big)}{-Q_k(\xi) + \sigma^2 Q_k^*(\eta) - \frac{1}{2}\|k\|^2}\,,$$

which is used as the stopping rule.

4 Model reduction

Define $P \in \mathbb{R}^{m\times n}$, $m \ge n$, with $P^T P = \mathbf{1}$. Let

$$\alpha_P \equiv P\beta\,,$$

where $\beta \in \mathbb{R}^n$. The minimizer of $Q_y(\alpha_P)$ and $Q_y^*(\alpha_P)$ is

$$\beta_{opt} = \Big(P^T(\sigma^2 K + K^T K)P\Big)^{-1} P^T K^T y = \Big(\sigma^2 P^T(KP) + (KP)^T(KP)\Big)^{-1}(KP)^T y$$

$$\beta^*_{opt} = (P^T KP + \sigma^2)^{-1} P^T y$$

If $m = n$ and P is full-rank, then $P\beta_{opt} = \alpha_{opt}$. Therefore,

$$Q_y(P\beta) = -y^T(KP)\beta + \frac{1}{2}\sigma^2\beta^T(P^T KP)\beta + \frac{1}{2}\beta^T(KP)^T(KP)\beta$$
$$Q_y^*(P\beta) = -y^T P\beta + \frac{1}{2}\sigma^2\beta^T\beta + \frac{1}{2}\beta(P^T KP)\beta$$

We choose P as a collection of unit vectors $\mathbf{e}_i$ where $(\mathbf{e}_i)_j = \delta_{ij}$. The statements hold when we replace y with k.

5 Algorithm

The paper here has many confusions, such as mixing Q_k with Q_y, β with β^*, k with y. Stop proceeding.

Data: $X = \{x_1, \cdots, x_m\}$, targets y, noise σ^2, precision ϵ
input : index sets $I, I^* = \{1, \cdots, m\}$, $S, S^* = \emptyset$
while $Q_k(P\beta_{opt}) + \sigma^2 Q_k^*(P^*\beta_{opt}^*) + \frac{1}{2}\|k\|^2 \leq \frac{\epsilon}{2}\Big(- Q_k(P\beta_{opt}) + \sigma^2 Q_k^*(P^*\beta_{opt}) - \frac{1}{2}\|k\|^2\Big)$ **do**
  Choose $M \subseteq I,\ M^* \subseteq I^*$
  Find $\arg\min_{i\in M} Q_k([P, e_i]\beta_{opt}^i)$ and $\arg\min_{i^*\in M^*} Q_k([P^*, e_i^*]\beta_{opt}^{*\,i})$
  Move i from I to S, move i^* from I^* to S^*.
  Set $P := [P, e_i]$, $P^* := [P^*, e_i^*]$.
end
output: S, β_{opt}, $Q_y^*(P^*\beta_{opt}^*)$.

6 Sparse likelihood approximation

Given samples $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, introduce latent variables u such that $P(y|u) = \mathcal{N}(y|u, \sigma^2)$. Denote the latent variables at the training points to be $u = \Big(u(x_1), \cdots, u(x_n)\Big)$, and the covariance of u to be $\mathbf{K} = (K(x_i, x_j))_{i,j} \in \mathbb{R}^{n,n}$. We have $P(u) = \mathcal{N}(u|\mathbf{0}, \mathbf{K})$.
To predict u_* at x_*, we have

$$\begin{aligned}\mu_* &= k_*^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}y\,, \quad k_* = \big(K(x_*, x_i)\big)_i \\ \sigma_*^2 &= K(x_*, x_*) - k_*^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}k_*\end{aligned}$$

However, this is costly to compute.

We replace the likelihood $\mathbb{P}(y|u)$ by (sparse likelihood)

$$Q(y|u_I) = \mathcal{N}\Big(y\Big|\mathbf{P}_I^T u_I, \sigma^2\mathbf{I}\Big),\ \mathbf{P}_I = \mathbf{K}_I^{-1}\mathbf{K}_{I,\cdot}\,,$$

where $P_I^T u_I = \mathbb{E}[u|u_I]$. Consider all distributions of the form $\propto \mathbb{P}(u)R(u_I)$, where $\mathbb{P}(u)$ indicates the prior of u. $R(u_I) = Q(y|u_I)$ minimizes the K-L divergence $\mathcal{D}\Big[\mathbb{P}(u)R(u_I)\Big\|\mathbb{P}[u|y]\Big]$.

Let

$$\begin{aligned}\mathbf{K}_I &= \mathbf{L}\mathbf{L}^T \qquad \text{(Cholesky)} \\ \mathbf{V} &= \mathbf{L}^{-1}\mathbf{K}_{I,\cdot} \\ \mathbf{M} &= \sigma^2\mathbf{I} + \mathbf{V}\mathbf{V}^T\end{aligned}$$

The approximate posterior of u_I can be obtained from the sparse likelihood:

$$\begin{aligned}Q(u_I|y) &= \mathbb{P}(u_I)Q(y|u_I) \\ &= \mathcal{N}(u_I|\mathbf{0}, \mathbf{L}\mathbf{L}^T)\cdot\mathcal{N}(y|K_{I,\cdot}^T K_I^{-1}u_I, \sigma^2\mathbf{I}) \\ &= \mathcal{N}\big(u_I\big|\mathbf{L}\mathbf{M}^{-1}\mathbf{V}y, \sigma^2\mathbf{L}\mathbf{M}^{-1}\mathbf{L}^T\big)\end{aligned}$$

To compute the approximate posterior of $u_* = u_*(x_*)$ at a new x_*, we define

$$\begin{aligned}\mathbf{M} &= \mathbf{L}_M\mathbf{L}_M^T \\ \beta &= \mathbf{L}_M^{-1}\mathbf{V}y \\ k_{I*} &= (K(x_i, x_*))_{i\in I} \\ l_* &= \mathbf{L}^{-1}k_{I*} \\ l_{M*} &= \mathbf{L}_M^{-1}l_*\end{aligned}$$

We have

$$\begin{aligned}Q(u_*|y) &= \int_{u_I}\mathbb{P}(u_*|u_I)Q(u_I|y)\,du_I \\ &= \int_{u_I}\mathcal{N}\big(u_*\big|k_{I*}^T\mathbf{K}_I^{-1}u_I,\ K(x_*, x_*) - k_{I*}^T\mathbf{K}_I^{-1}k_{I*}\big)\cdot\mathcal{N}\big(u_I\big|\mathbf{L}\mathbf{M}^{-1}\mathbf{V}y,\ \sigma^2\mathbf{L}\mathbf{M}^{-1}\mathbf{L}^T\big)\,du_I \\ &= \mathcal{N}\big(u_*\big|l_{M*}^T\beta,\ K(x_*, x_*) - \|l_*\|^2 + \sigma^2\|l_{M,*}\|^2\big)\end{aligned}$$

Notice the posterior mean

$$\mu(x_*) = k_{I*}^T \mathbf{L}^{-T} \mathbf{L}_M^{-T} \beta$$

and the posterior variance

$$\begin{aligned}
\sigma^2(x_*) &= K(x_*, x_*) - k_{I*}^T \mathbf{L}^{-T} \mathbf{L}^{-1} k_{I*} + \sigma^2 k_{I*}^T \mathbf{L}^{-T} \mathbf{M}^{-1} \mathbf{L}^{-1} k_{I*} \\
&= K(x_*, x_*) - k_{I*}^T \mathbf{L}^{-T} \mathbf{V} \mathbf{M}^{-1} \mathbf{V}^T \mathbf{L}^{-1} k_{I*} \qquad \text{(Woodbury identity)}
\end{aligned}$$

Therefore, we need to pre-compute $\mathbf{L}^{-T} \mathbf{L}_M^{-T} \beta$ and $\mathbf{L}^{-T} \mathbf{V} \mathbf{M}^{-1} \mathbf{V}^T \mathbf{L}^{-1}$.

7 Inclusion of a new point

Define

$$p = \operatorname{diag}(\mathbf{V}^T \mathbf{V}), \quad q = \operatorname{diag}(\mathbf{V}^T \mathbf{M}^{-1} \mathbf{V})$$

Let $\cdot'$ be the quantity associated with $\{I, i\}$ active set. We have

$$\mathbf{L}'_{d+1, \cdot \backslash d+1} = \left(\mathbf{L}^{-1} \mathbf{K}_{I,i}\right)^T \equiv v_i^T$$

$$\mathbf{L}'_{d+1,d+1} = \left(K(x_i, x_i) - v_i^T v_i\right)^{1/2}$$

$$\mathbf{V}'_{1\cdots d, \cdot} = \mathbf{V}$$

$$\mathbf{V}'_{d+1, \cdot} = \frac{1}{\mathbf{L}'_{d+1,d+1}} \left(\mathbf{K}_{\cdot, i} - \mathbf{V}^T v_i\right)$$

$$p' = p + \left(\left(\mathbf{V}'_{d+1,j}\right)^2\right)_j$$

8 Application to twin-model-GPO

Given samples $S = \{(x_1, \xi_1, \xi_{\tilde{\nabla}1}), \cdots, (x_n, \xi_n, \xi_{\tilde{\nabla}n})\}$, introduce latent variables $u = (\xi, \xi_\nabla) \in \mathbb{R}^{n(d+1)}$ such that

$$\mathbb{P}(\xi_{\tilde{\nabla}}) = \mathcal{N}\left(\xi_{\tilde{\nabla}} \middle| \xi_\nabla, \bar{\mathbf{G}}\right), \qquad \mathbb{P}(\xi, \xi_\nabla) = \mathcal{N}\left(\xi, \xi_\nabla \middle| 0, \begin{pmatrix} \mathbf{D} & \mathbf{H} \\ \mathbf{H}^T & \mathbf{E} \end{pmatrix} := \mathbf{K}\right)$$

where $\bar{\mathbf{G}} \in \mathbb{R}^{nd \times nd}$. Consider a subset of indices $I := \{1, \cdots, n\} \cup I_\nabla$, where $I_\nabla \subseteq T_\nabla := \{n+1, \cdots, n+nd\}$. Denote u_I, u_{I_∇} the subset of latent variables of u indexed by I, I_∇. We approximate likelihood $\mathbb{P}\left(\xi, \xi_{\tilde{\nabla}} \middle| u\right)$ by

$$Q\left(\xi, \xi_{\tilde{\nabla}} \middle| u_I\right) = \mathcal{N}\left(\xi_{\tilde{\nabla}} \middle| \mathbf{P}_{I_\nabla}^T u_{I_\nabla}, \bar{\mathbf{G}}\right),$$

where

$$\mathbf{P}_{I_\nabla} = \mathbf{K}_{I_\nabla}^{-1} \mathbf{K}_{I_\nabla, T_\nabla}$$

Define

$$\mathbf{P}_I = \begin{pmatrix} \mathbf{0}_{n \times nd} \\ \mathbf{P}_{I_\nabla} \end{pmatrix}$$

Notice

$$\mathbf{P}_I^T u_I = \mathbf{P}_{I_\nabla}^T u_{I_\nabla}$$

The approximate posterior of u_I is given by

$$Q(u_I | \xi, \xi_{\tilde{\nabla}}) = \delta(u_{\{1,\cdots,n\}}, \xi) \cdot \mathcal{N}\left(u_I \middle| (\mathbf{K}_I^{-1} + \mathbf{P}_I \bar{\mathbf{G}}^{-1} \mathbf{P}_I^T)^{-1} \mathbf{P}_I \bar{\mathbf{G}}^{-1} \xi_{\tilde{\nabla}},\ (\mathbf{K}_I^{-1} + \mathbf{P}_I \bar{\mathbf{G}}^{-1} \mathbf{P}_I^T)^{-1}\right).$$

To evalute the posterior mean and variance at new point x_*, define a length $\left|I\right|$ vector: k_{I*} which indicates the covariance between $\xi(x_*)$ and $u_I = \left(\xi, \xi_\nabla\right)$. We have

$$\begin{aligned}
Q(u_*|\xi, \xi_{\tilde{\nabla}}) &= \int_{u_I} \mathbb{P}(u_*|u_I) Q(u_I|\xi, \xi_{\tilde{\nabla}})\, du_I \\
&= \int_{u_I} \mathcal{N}\big(u_* \big| k_{I*}^T \mathbf{K}_I^{-1} u_I,\, K(x_*, x_*) - k_{I*}^T \mathbf{K}_I^{-1} k_{I*}\big) \cdot \\
&\quad \delta(u_{\{1,\cdots,n\}}, \xi) \cdot \mathcal{N}\Big(u_I \Big| (\mathbf{K}_I^{-1} + \mathbf{P}_I \bar{\mathbf{G}}^{-1} \mathbf{P}_I^T)^{-1} \mathbf{P}_I \bar{\mathbf{G}}^{-1} \xi_{\tilde{\nabla}}\,,\ (\mathbf{K}_I^{-1} + \mathbf{P}_I \bar{\mathbf{G}}^{-1} \mathbf{P}_I^T)^{-1}\Big)\, du_I
\end{aligned}$$

Define

$$\mathbf{S} = \left(\mathbf{0}_{|I_\nabla| \times n},\, \mathbf{I}_{|I_\nabla|}\right)$$

The posterior mean of $\xi(x_*)$ is

$$k_{I*}^T \mathbf{K}_I^{-1} \left(\xi^T\,,\ \left(\mathbf{S}(\mathbf{K}_I^{-1} + \mathbf{P}_I \bar{\mathbf{G}}^{-1} \mathbf{P}_I^T)^{-1} \mathbf{P}_I \bar{\mathbf{G}}^{-1} \xi_{\tilde{\nabla}}\right)^T\right)^T\,,$$

and the posterior variance is

$$K(x_*, x_*) - k_{I*}^T \mathbf{K}_I^{-1} k_{I*} + k_{I*}^T \mathbf{K}_I^{-1} \mathbf{S}^T (\mathbf{K}_I^{-1} + \mathbf{P}_I \bar{\mathbf{G}}^{-1} \mathbf{P}_I^T)^{-1} \mathbf{S} \mathbf{K}_I^{-T} k_{I*}$$

9 From Lehel Csato's thesis

$f_\mathbf{x}$ (GP) used as latent variable in a likelihood $P(y|\mathbf{x}, f_\mathbf{x})$. $K_0(\mathbf{x}, \mathbf{x}') = Cov(f_\mathbf{x}, f_{\mathbf{x}'})$. Training data $\{\mathbf{x}_n, y_n\}_{n=1}^N$. Assume $f_\mathbf{x}$ has zero mean. The kernel can be decomposed into dictionary $\phi_i(\mathbf{x})_{i=1}^\infty$. The dictionary defines the feature space $\mathcal{F}$.

$$< f(\mathbf{x}, f(\mathbf{x}')) >= K(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \sigma^2 \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')$$

If there are k elements in the dictionary, the feature space $\mathcal{F} = \mathbb{R}^k$. The projection function is

$$\phi(\mathbf{x}) = \left(\sigma_1 \phi_1(\mathbf{x}), \cdots\right)^T,$$

which is the image of $\mathbf{x}$ in the feature space.

$$K(\mathbf{x}, \mathbf{x}') = \phi_\mathbf{x}^T \phi_{\mathbf{x}'}$$