Derivative Estimation with Kernel

by Weihao LI

December 9, 2024

Outline

Problem setting

- Estimate $\boldsymbol{\theta}$, known f

$$\dot{x}(t) = \frac{dx(t)}{dt} = f(\boldsymbol{x}(t), \boldsymbol{\theta}, t), \quad t \in [0, T]$$

- Observation y with measurment error, $\{t_1, \cdots t_n\}$

$$y_i = x(t_i) + e_i, \quad i = 1, \ldots, n$$

Frequentist: two steps [LW08]

- Local polynomial to estimate derivative $\widehat{x}'(t_i)$
- Perform non-linear least square

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\widehat{x}'(t_i) - f\left\{\widehat{x}(t_i), \boldsymbol{\theta}\right\}\right]^2$$

Bayesian: MAGI [YWK21]

$$X(\cdot) \sim GP(\mu(\cdot), K(,))$$

- ODE constraint:

$$W_I = \max_{t \in I} \left| \dot{X}_d(t) - f(X(t), \theta, t)_d \right|$$

- Bayes rule:

$$\begin{aligned}
&p_{\boldsymbol{\Theta}, X(\boldsymbol{I}) | W_I, \boldsymbol{Y}(\boldsymbol{\tau})} \left(\boldsymbol{\theta}, \boldsymbol{x}(\boldsymbol{I}) \mid W_{\boldsymbol{I}} = 0, \boldsymbol{Y}(\boldsymbol{I}) = \boldsymbol{y}(\boldsymbol{I}) \right) \\
&\propto P \left(\boldsymbol{\Theta} = \boldsymbol{\theta}, \boldsymbol{X}(\boldsymbol{I}) = \boldsymbol{x}(\boldsymbol{I}), W_{\boldsymbol{I}} = 0, \boldsymbol{Y}(\boldsymbol{I}) = \boldsymbol{y}(\boldsymbol{I}) \right) \\
&= \pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \times P(\boldsymbol{X}(\boldsymbol{I}) = \boldsymbol{x}(\boldsymbol{I}) \mid \boldsymbol{\Theta} = \boldsymbol{\theta}) \\
&\quad \times P(\boldsymbol{Y}(\boldsymbol{I}) = \boldsymbol{y}(\boldsymbol{I}) \mid \boldsymbol{X}(\boldsymbol{I}) = \boldsymbol{x}(\boldsymbol{I}), \boldsymbol{\Theta} = \boldsymbol{\theta}) \\
&\quad P \left(W_{\boldsymbol{I}} = 0 \mid \boldsymbol{Y}(\boldsymbol{I}) = \boldsymbol{y}(\boldsymbol{I}), \boldsymbol{X}(\boldsymbol{I}) = \boldsymbol{x}(\boldsymbol{I}), \boldsymbol{\Theta} = \boldsymbol{\theta} \right)
\end{aligned}$$

MAGI

$$
\begin{aligned}
&P\left(W_l = 0 \mid \boldsymbol{Y}(l) = \boldsymbol{y}(l), \boldsymbol{X}(l) = \boldsymbol{x}(l), \boldsymbol{\Theta} = \theta\right) \\
&= P\left(\dot{\boldsymbol{X}}(l) - \mathbf{f}\left(\boldsymbol{x}(l), \theta, t_l\right) = \mathbf{0} \mid \boldsymbol{Y}(l) = \boldsymbol{y}(l), \boldsymbol{X}(l) = \boldsymbol{x}(l), \boldsymbol{\Theta} = \theta\right) \\
&= P\left(\dot{\boldsymbol{X}}(l) - \mathbf{f}\left(\boldsymbol{x}(l), \theta, t_l\right) = \mathbf{0} \mid \boldsymbol{X}(l) = \boldsymbol{x}(l)\right) \\
&= P\left(\dot{\boldsymbol{X}}(l) = \mathbf{f}\left(\boldsymbol{x}(l), \theta, t_l\right) \mid \boldsymbol{X}(l) = \boldsymbol{x}(l)\right),
\end{aligned}
$$

- $\dot{\boldsymbol{X}}(\cdot)$ is also a gaussian process

$$p_{\Theta,\boldsymbol{X}(\boldsymbol{I})|W_I,\boldsymbol{Y}(\tau)}\left(\boldsymbol{\theta},\boldsymbol{x}(\boldsymbol{I}) \mid W_I = 0, \boldsymbol{Y}(\boldsymbol{I}) = \boldsymbol{y}(\boldsymbol{I})\right)$$

$$\propto \pi_\theta(\theta) \exp\left\{ -\frac{1}{2}[\underbrace{|\boldsymbol{I}| \log(2\pi) + \log|C_d| + \|x(\boldsymbol{I}) - \mu(\boldsymbol{I})\|^2_{C^{-1}}}_{(1)}\right.$$

$$+ \underbrace{|\boldsymbol{I}| \log(2\pi) + \log|K| + \left\|\mathbf{f}_{\boldsymbol{I}}^{x,\theta} - \dot{\mu}(\boldsymbol{I}) - m\{x(\boldsymbol{I}) - \mu(\boldsymbol{I})\}\right\|^2_{K^{-1}}}_{(3)}$$

$$\left. + \underbrace{n \log\left(2\pi\sigma^2\right) + \|x(\boldsymbol{I}) - y(\boldsymbol{I})\|^2_{\sigma^{-2}}}_{(2)}]\right\}$$

$$\text{where} \begin{cases} C & = \mathcal{K}(\boldsymbol{I},\boldsymbol{I}) \\ m & = {}'\mathcal{K}(\boldsymbol{I},\boldsymbol{I})\mathcal{K}(\boldsymbol{I},\boldsymbol{I})^{-1} \\ K & = \mathcal{K}''(\boldsymbol{I},\boldsymbol{I}) - {}'\mathcal{K}(\boldsymbol{I},\boldsymbol{I})\mathcal{K}(\boldsymbol{I},\boldsymbol{I})^{-1}\mathcal{K}'(\boldsymbol{I},\boldsymbol{I}) \end{cases}$$

The difference operator m

- m perform numerical differentation: $m \cdot x(\boldsymbol{I}) \approx \dot{x}(\boldsymbol{I})$
- m perform derivative operation: $m \cdot g(\boldsymbol{I}) \approx \dot{g}(\boldsymbol{I})$ for any smooth function g, thus of independent interest
- m depend on kernel and design points

Take notes for consistency of kernel ridge regression

- is there an explicit solution for iterative optimization? (design by myself)
- Grace Wahba (Wahba, 1990) pioneered the study of nonparametric regression in reproducing kernel Hilbert spaces (RKHS) from the computational and statistical perspectives. One of the key aspects in that work is the role of the decay of eigenvalues of the kernel (at the population level) in rates of convergence. The analysis relies on explicit regularization (ridge parameter λ) for the bias-variance trade-off.
- more focus on risk bound?
- read this first:
- representer theorem can be generalized further, square error , glm likelihood ,etc (GP book p133 regularization)

KRR theory: wainwright

$$f^{\diamond} := \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2 + \lambda_n \|f\|_{\mathcal{H}}^2 \right\},$$

$$\omega^{\dagger} = \arg\min_{\omega \in \mathbb{R}^n} \left\{ \frac{1}{2} \omega^T K^2 \omega - \omega^T \frac{Ky}{\sqrt{n}} + \lambda_n \omega^T K \omega \right\},$$

$$f^{\diamond}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i^{\dagger} \mathcal{K}\left(\cdot, x_i\right).$$

- Kernel complexity measures and statistical guarantees $\widehat{\mathcal{R}}(\delta) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \min\left\{\delta^2, \widehat{\mu}_j\right\}}$ For a given kernel matrix and noise variance $\sigma > 0$, the critical radius is defined to be the smallest positive solution $\delta_n > 0$ to the inequality $\frac{\widehat{\mathcal{R}}(\delta)}{\delta} \leq \frac{\delta}{\sigma}$

$$\left\|f^{\diamond} - f^*\right\|_n^2 \leq c_u \left\{\lambda_n + \delta_n^2\right\} \quad \text{w.p} \geq 1 - c_1 e^{-c_2 n \delta_n^2}$$

If f^* is in $\mathcal{H}, \lambda \geq 2\delta_n^2$.

KRR theory: wainwright

- [ZDW13]: give bias and variance bound for KRR estimator

Consistency

- For gaussian process regression:[CS07] those work assume fourth order derivative exist in order to show derivative process is continuously differentiable, if we only need continuous, maybe second moment is enough?
- For frequentist regularized least square:[CO90] and [Wah90]

eigenvalue decay rate

- theorem 15 in this paper Under the above assumptions on K and Ω, the eigenvalues decay at least like

$$\sqrt{\lambda_{n+1}} < c_1 n^{-(\beta+d)/2d}$$

for a kernel with smoothness β, and at least like

$$\sqrt{\lambda_{n+1}} < c_2 \exp\left(-c_3 n^{1/d}\right)$$

for kernels with unlimited smoothness. The constants c_1, c_2, c_3 are independent of n, but dependent on K, Ω, and the space dimension.

- Hermite learning algorithm: no computational algorithm, perhaps only approximation
- Can truncated GP prior used to construct the derivative process?
- estimation accuracy($\boldsymbol{\theta}$) for MAGI is not that high compared with others, but with smaller RMSE

posterior consistency for Gaussian linear model with unknown variance

RKHS knowledge

- How $\mathcal{H}$ related with holder class, how relate with gaussian kernel,matern kernel; or how it related with eigendecay of kernel?
- how degree of freedom for matern kernel determine the size of Corresponding RKHS
- how to make sure ODE system generate the function belongs to RKHS/Holder class?

Eigendecay rate for specific kernel

- The rate of decay of the eigenvalues gives important information about the smoothness of the kernel.
- Gaussian kernel $K(x, x') = \exp\left(-\|x - x'\|_2^2\right)$

$$\mu_j \leq c_1 \exp\left(-c_2 j^2\right) \quad \text{for all } j = 1, 2, \ldots$$

Remark: minimax rate for this class is $\sigma^2 \frac{\sqrt{\log N}}{N}$, very small class.
how about the gaussian kernel with variable bandwidth

- First order Sobolev space: $K(x, x') = 1 + \min\{x, x'\}$, $\nu = 1$, Lipschitz functions with smoothness $\nu = 1$.

$$\mu_j \leq C j^{-2\nu} \quad \text{for all } j = 1, 2, \ldots$$

Remark: minimax rate for class with $\mu_j \leq C j^{-2\nu}$ for all $j = 1, 2, \ldots$ is $\left(\frac{\sigma^2}{N}\right)^{\frac{2\nu}{2\nu+1}}$

[TJW16] Matern kernel result

- Matern family:

$$\Phi(s,t;\nu,\phi) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}\phi\|s-t\|)^{\nu}K_{\nu}(2\sqrt{\nu}\phi\|s-t\|)$$

Linear space: $F_{\Phi}(\Omega) = \left\{\sum_{i=1}^{N}\beta_i\Phi(\cdot,x_i) : N\in\mathbb{N}, \beta_i\in\mathbf{R}, x_i\in\Omega\right\}$ and equip this space with the bilinear form

$$\left\langle \sum_{i=1}^{N}\beta_i\Phi(\cdot,x_i), \sum_{j=1}^{M}\gamma_j\Phi(\cdot,y_j)\right\rangle_{\Phi} := \sum_{i=1}^{N}\sum_{j=1}^{M}\beta_i\gamma_j\Phi(x_i,y_j)$$

Define the native space $\mathcal{N}_{\Phi}(\Omega)$ as the closure of $F_{\Phi}(\Omega)$ under the inner product $\langle\cdot,\cdot\rangle_{\Phi}$

- Corollary A.6: RKHS $\mathcal{N}_{\Phi}(\Omega)$ equals to the (fractional) Sobolev space $H^{\nu+d/2}(\Omega)$ and the two norms are equivalent for $\nu\geq 1$, where the Sobolev norm is defined by $\|f\|_{H^k(\Omega)} = \sqrt{\sum_{|\alpha|\leq k}\|D^{\alpha}f\|^2_{L_2(\Omega)}}$.

Matern kernel eigendecay rate

Eigenvalue decay for gaussian kernel with variable bandwidth

ODE

- Notation 1.3. Let

$$x_1' = f_1(t, x_1, x_2, \ldots, x_n)$$
$$x_2' = f_2(t, x_1, x_2, .., x_n)$$
$$\vdots$$
$$x_n' = f_n(t, x_1, x_2, \ldots, x_n)$$

be a system of differential equations. I will write this as $X' = F(t, X)$ where $X = (x_1, x_2, \cdots x_n)^\top$

Theorem 1.14. Given the initial value problem

$$X' = F(X), X(t_0) = X_0$$

for $X_0 \in \mathbb{R}^n$. Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is C^1. Then, there exists a unique solution to this initial value problem. That is, there exists an $a > 0$ and a solution $X : (t_0 - a, t_0 + a) \to \mathbb{R}^n$ of the differential equation such that $X(t_0) = X_0$.

Reference I

- Dennis D Cox and Finbarr O'Sullivan, *Asymptotic analysis of penalized likelihood and related estimators*, The Annals of Statistics (1990), 1676–1695.
- Taeryon Choi and Mark J Schervish, *On posterior consistency in nonparametric regression problems*, Journal of Multivariate Analysis **98** (2007), no. 10, 1969–1987.
- Hua Liang and Hulin Wu, *Parameter estimation for differential equation models using a framework of measurement error in regression models*, Journal of the American Statistical Association **103** (2008), no. 484, 1570–1583.
- Rui Tuo and CF Jeff Wu, *A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties*, SIAM/ASA Journal on Uncertainty Quantification **4** (2016), no. 1, 767–795.

Reference II

- Grace Wahba, *Spline models for observational data*, SIAM, 1990.
- Shihao Yang, Samuel WK Wong, and SC Kou, *Inference of dynamic systems from noisy and sparse data via manifold-constrained gaussian processes*, Proceedings of the National Academy of Sciences **118** (2021), no. 15, e2020397118.
- Yuchen Zhang, John Duchi, and Martin Wainwright, *Divide and conquer kernel ridge regression*, Conference on learning theory, PMLR, 2013, pp. 592–617.