Variable selection consistency of Gaussian process regression

Sheng Jiang [1]

Department of Statistics
University of California Santa Cruz

July 1, 2022

(joint work with Surya Tokdar)

Regression problem setup

- We observe n pairs of (X_i, Y_i):

$$\text{response } Y_i \in \mathbb{R}, \text{ explanatory } X_i \in \mathbb{R}^d.$$

- In many modern datasets,
  1. d can be large or larger than n.
     - e.g., DNA micro-arrays
  2. The dependence of Y on X can be non-linear.
     - e.g., the motorcycle accidents data (mcycle (MASS) in R)

Regression problem setup

- We observe n pairs of (X_i, Y_i):

$$\text{response } Y_i \in \mathbb{R}, \text{ explanatory } X_i \in \mathbb{R}^d.$$

- In many modern datasets,
  1. d can be large or larger than n.
     - e.g., DNA micro-arrays
  2. The dependence of Y on X can be non-linear.
     - e.g., the motorcycle accidents data (mcycle (MASS) in R)

Regression problem setup

- We observe n pairs of (X_i, Y_i):

$$\text{response } Y_i \in \mathbb{R}, \text{ explanatory } X_i \in \mathbb{R}^d.$$

- In many modern datasets,
  1. d can be large or larger than n.
     - e.g., DNA micro-arrays
  2. The dependence of Y on X can be non-linear.
     - e.g., the motorcycle accidents data (mcycle (MASS) in R)

Non-parametric regression

- General formulation:

$$Y_i = f(X_i) + \epsilon_i, \qquad \epsilon_i | X_i \overset{iid}{\sim} N(0, \sigma^2),$$

  where
  - f is some unknown function to be estimated
  - Y_i may depend on some of the regressors. (some regressors can be redundant)
- A few options:
  - Kernel smoothing, local polynomials, splines, tree methods, random forests,...
  - Gaussian process (GP) regression

Non-parametric regression

- General formulation:

$$Y_i = f(X_i) + \epsilon_i, \qquad \epsilon_i | X_i \overset{iid}{\sim} N(0, \sigma^2),$$

  where
  - f is some unknown function to be estimated
  - Y_i may depend on some of the regressors. (some regressors can be redundant)
- A few options:
  - Kernel smoothing, local polynomials, splines, tree methods, random forests,...
  - Gaussian process (GP) regression

Non-parametric regression

- General formulation:

$$Y_i = f(X_i) + \epsilon_i, \qquad \epsilon_i | X_i \overset{iid}{\sim} N(0, \sigma^2),$$

  where
  - f is some unknown function to be estimated
  - Y_i may depend on some of the regressors. (some regressors can be redundant)
- A few options:
  - Kernel smoothing, local polynomials, splines, tree methods, random forests,...
  - Gaussian process (GP) regression

Gaussian process regression

- GP prior on $f(\cdot)$: for any test point $X_* \in \mathbb{R}^d$,

$$\begin{bmatrix} f(X_1) \\ f(X_2) \\ \vdots \\ f(X_n) \\ f(X_*) \end{bmatrix} \sim N_{n+1}\left(\mu, \begin{bmatrix} K(X_1,X_1) & K(X_1,X_2) & \dots & K(X_1,X_*) \\ K(X_2,X_1) & K(X_2,X_2) & \dots & K(X_2,X_*) \\ \vdots & \vdots & \ddots & \vdots \\ K(X_*,X_1) & K(X_*,X_2) & \dots & K(X_*,X_*) \end{bmatrix}\right)$$

  where $K(\cdot,\cdot)$ covariance function/kernel controls sample paths' smoothness.
- For the ease of notation, we write $f \sim W$ with W being some GP.
- With ind. Gaussian noise, we know the distribution of

$$f(X_*) | \{(X_i, Y_i)\}_{i=1}^n, \text{ for every } X_* \in \mathbb{R}^d,$$

  and

$$Y_* | \{(X_i, Y_i)\}_{i=1}^n.$$

(See more details in, e.g., Rasmussen and Williams (2006))

Gaussian process regression

- GP prior on $f(\cdot)$: for any test point $X_* \in \mathbb{R}^d$,

$$\begin{bmatrix} f(X_1) \\ f(X_2) \\ \vdots \\ f(X_n) \\ f(X_*) \end{bmatrix} \sim N_{n+1}\left(\mu, \begin{bmatrix} K(X_1,X_1) & K(X_1,X_2) & \dots & K(X_1,X_*) \\ K(X_2,X_1) & K(X_2,X_2) & \dots & K(X_2,X_*) \\ \vdots & \vdots & \ddots & \vdots \\ K(X_*,X_1) & K(X_*,X_2) & \dots & K(X_*,X_*) \end{bmatrix}\right)$$

  where $K(\cdot,\cdot)$ covariance function/kernel controls sample paths' smoothness.
- For the ease of notation, we write $f \sim W$ with W being some GP.
- With ind. Gaussian noise, we know the distribution of

$$f(X_*)|\{(X_i,Y_i)\}_{i=1}^n, \text{ for every } X_* \in \mathbb{R}^d,$$

  and

$$Y_*|\{(X_i,Y_i)\}_{i=1}^n.$$

(See more details in, e.g., Rasmussen and Williams (2006))

Gaussian process regression

- GP prior on $f(\cdot)$: for any test point $X_* \in \mathbb{R}^d$,

$$\begin{bmatrix} f(X_1) \\ f(X_2) \\ \vdots \\ f(X_n) \\ f(X_*) \end{bmatrix} \sim N_{n+1}\left(\mu, \begin{bmatrix} K(X_1,X_1) & K(X_1,X_2) & \dots & K(X_1,X_*) \\ K(X_2,X_1) & K(X_2,X_2) & \dots & K(X_2,X_*) \\ \vdots & \vdots & \ddots & \vdots \\ K(X_*,X_1) & K(X_*,X_2) & \dots & K(X_*,X_*) \end{bmatrix}\right)$$

  where $K(\cdot,\cdot)$ covariance function/kernel controls sample paths' smoothness.
- For the ease of notation, we write $f \sim W$ with W being some GP.
- With ind. Gaussian noise, we know the distribution of

$$f(X_*)|\{(X_i,Y_i)\}_{i=1}^n, \text{ for every } X_* \in \mathbb{R}^d,$$

  and

$$Y_*|\{(X_i,Y_i)\}_{i=1}^n.$$

(See more details in, e.g., Rasmussen and Williams (2006))

Rescaled SE GP

- Rescaled Gaussian process (GP) prior

$$f|A \sim W_{At}, \qquad A^d \approx Gamma$$

- Squared Exponential (SE) covariance kernel $k(\cdot,\cdot)$:

$$k(s,t) := \mathbb{E}[W_s W_t] = e^{-||s-t||^2}, \text{ for } s,t \in \mathbb{R}^d.$$

- Posterior contraction rates in L_2 are near minimax-optimal and adaptive to
  - unknown smoothness with covariate dimension d fixed,

$$n^{-\beta/(2\beta+d)} \log^{\kappa} n.$$

    (van der Vaart and van Zanten, 2009)
  - both unknown smoothness and true sparsity $d_0 \le d$,

$$n^{-\beta/(2\beta+d_0)} \log^{\kappa} n.$$

    (Tokdar, 2011; Bhattacharya et al., 2014; Yang and Tokdar, 2015)

Rescaled SE GP

- Rescaled Gaussian process (GP) prior

$$f|A \sim W_{At}, \qquad A^d \approx Gamma$$

- Squared Exponential (SE) covariance kernel $k(\cdot,\cdot)$:

$$k(s,t) := \mathbb{E}[W_s W_t] = e^{-||s-t||^2}, \text{ for } s,t \in \mathbb{R}^d.$$

- Posterior contraction rates in L_2 are near minimax-optimal and adaptive to
  - unknown smoothness with covariate dimension d fixed,

$$n^{-\beta/(2\beta+d)} \log^\kappa n.$$

(van der Vaart and van Zanten, 2009)

  - both unknown smoothness and true sparsity $d_0 \leq d$,

$$n^{-\beta/(2\beta+d_0)} \log^\kappa n.$$

(Tokdar, 2011; Bhattacharya et al., 2014; Yang and Tokdar, 2015)

Rescaled SE GP

- Rescaled Gaussian process (GP) prior

$$f|A \sim W_{At}, \qquad A^d \approx Gamma$$

- Squared Exponential (SE) covariance kernel $k(\cdot,\cdot)$:

$$k(s,t) := \mathbb{E}[W_s W_t] = e^{-||s-t||^2}, \text{ for } s,t \in \mathbb{R}^d.$$

- Posterior contraction rates in L_2 are near minimax-optimal and adaptive to
  - unknown smoothness with covariate dimension d fixed,

$$n^{-\beta/(2\beta+d)} \log^\kappa n.$$

(van der Vaart and van Zanten, 2009)

  - both unknown smoothness and true sparsity $d_0 \leq d$,

$$n^{-\beta/(2\beta+d_0)} \log^\kappa n.$$

(Tokdar, 2011; Bhattacharya et al., 2014; Yang and Tokdar, 2015)

Rescaled SE GP

- Rescaled Gaussian process (GP) prior

$$f|A \sim W_{At}, \qquad A^d \approx Gamma$$

- Squared Exponential (SE) covariance kernel $k(\cdot, \cdot)$:

$$k(s,t) := \mathbb{E}[W_s W_t] = e^{-||s-t||^2}, \text{ for } s, t \in \mathbb{R}^d.$$

- Posterior contraction rates in L_2 are near minimax-optimal and adaptive to
  - unknown smoothness with covariate dimension d fixed,

$$n^{-\beta/(2\beta+d)} \log^\kappa n.$$

(van der Vaart and van Zanten, 2009)

  - both unknown smoothness and true sparsity $d_0 \leq d$,

$$n^{-\beta/(2\beta+d_0)} \log^\kappa n.$$

(Tokdar, 2011; Bhattacharya et al., 2014; Yang and Tokdar, 2015)

Rescaled SE GP

- Rescaled Gaussian process (GP) prior

$$f|A \sim W_{At}, \qquad A^d \approx Gamma$$

- Squared Exponential (SE) covariance kernel $k(\cdot,\cdot)$:

$$k(s,t) := \mathbb{E}[W_s W_t] = e^{-||s-t||^2}, \text{ for } s,t \in \mathbb{R}^d.$$

- Posterior contraction rates in L_2 are near minimax-optimal and adaptive to
  - unknown smoothness with covariate dimension d fixed,

$$n^{-\beta/(2\beta+d)} \log^{\kappa} n.$$

(van der Vaart and van Zanten, 2009)

  - both unknown smoothness and true sparsity $d_0 \leq d$,

$$n^{-\beta/(2\beta+d_0)} \log^{\kappa} n.$$

(Tokdar, 2011; Bhattacharya et al., 2014; Yang and Tokdar, 2015)

The variable selection consistency challenge

- When $d_0 \leq d$,

which regressors are relevant?

- Can we equip GP regression with variable selection?
- How accurate is the selection?
  - A more refined question than estimation accuracy.

(Rasmussen and Williams, 2006; Savitsky et al., 2011; Tokdar, 2011)

The variable selection consistency challenge

- When $d_0 \leq d$,

  which regressors are relevant?

- Can we equip GP regression with variable selection?
- How accurate is the selection?
  - A more refined question than estimation accuracy.

(Rasmussen and Williams, 2006; Savitsky et al., 2011; Tokdar, 2011)

The variable selection consistency challenge

- When $d_0 \leq d$,

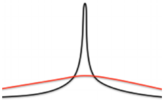
  which regressors are relevant?
- Can we equip GP regression with variable selection?
- How accurate is the selection?
  - A more refined question than estimation accuracy.

(Rasmussen and Williams, 2006; Savitsky et al., 2011; Tokdar, 2011)

Bayesian sparsity priors

- The spike-and-slab prior: a mixture of two components

$$\pi(\theta) = (1 - w)\psi_0(\theta) + w\psi_1(\theta).$$



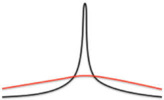
  - Continuous version, aka soft spike-and-slab
  - Discrete version, aka hard spike-and-slab
- Continuous shrinkage priors
  - Exponential scale mixture of Normal $\rightarrow$ Laplace, aka, Bayesian lasso.
  - Half-Cauchy scale mixture of Normal, aka, Horseshoe prior.
  - ...

Handbook of Bayesian VS(Tadesse and Vannucci, 2021)

Bayesian sparsity priors

- The spike-and-slab prior: a mixture of two components

$$\pi(\theta) = (1 - w)\psi_0(\theta) + w\psi_1(\theta).$$



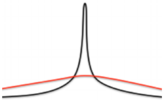
  - Continuous version, aka soft spike-and-slab
  - Discrete version, aka hard spike-and-slab
- Continuous shrinkage priors
  - Exponential scale mixture of Normal $\rightarrow$ Laplace, aka, Bayesian lasso.
  - Half-Cauchy scale mixture of Normal, aka, Horseshoe prior.
  - ...

Handbook of Bayesian VS (Tadesse and Vannucci, 2021)

Bayesian sparsity priors

- The spike-and-slab prior: a mixture of two components

$$\pi(\theta) = (1 - w)\psi_0(\theta) + w\psi_1(\theta).$$



  - Continuous version, aka soft spike-and-slab
  - Discrete version, aka hard spike-and-slab
- Continuous shrinkage priors
  - Exponential scale mixture of Normal $\rightarrow$ Laplace, aka, Bayesian lasso.
  - Half-Cauchy scale mixture of Normal, aka, Horseshoe prior.
  - ...

Handbook of Bayesian VS(Tadesse and Vannucci, 2021)

Gaussian process priors with stochastic variable selection

- Introduce model index parameter Γ supported on $\{0,1\}^d$
  - $\Gamma_i = 1$ means X_i is included; $\Gamma_i = 0$ means X_i is not included
- Put a prior on Γ,
  - e.g., independent Bernoulli which works in fixed dimension case.
  - In high dimensions, we need to carefully penalize large models
- Given $\gamma \in \{0,1\}^d$, put a $|\gamma|$-dimensional rescaled GP over $X_{[\gamma]} \equiv \{X_i : \gamma_i = 1\}$.
- Essentially, hard spike-and-slab. Alternative methods can be developed.

Gaussian process priors with stochastic variable selection

- Introduce model index parameter Γ supported on $\{0,1\}^d$
  - $\Gamma_i = 1$ means X_i is included; $\Gamma_i = 0$ means X_i is not included
- Put a prior on Γ,
  - e.g., independent Bernoulli which works in fixed dimension case.
  - In high dimensions, we need to carefully penalize large models
- Given $\gamma \in \{0,1\}^d$, put a $|\gamma|$-dimensional rescaled GP over $X_{[\gamma]} \equiv \{X_i : \gamma_i = 1\}$.
- Essentially, hard spike-and-slab. Alternative methods can be developed.

Gaussian process priors with stochastic variable selection

- Introduce model index parameter Γ supported on $\{0,1\}^d$
  - $\Gamma_i = 1$ means X_i is included; $\Gamma_i = 0$ means X_i is not included
- Put a prior on Γ,
  - e.g., independent Bernoulli which works in fixed dimension case.
  - In high dimensions, we need to carefully penalize large models
- Given $\gamma \in \{0,1\}^d$, put a $|\gamma|$-dimensional rescaled GP over $X_{[\gamma]} \equiv \{X_i : \gamma_i = 1\}$.
- Essentially, hard spike-and-slab. Alternative methods can be developed.

Gaussian process priors with stochastic variable selection

- Introduce model index parameter Γ supported on $\{0,1\}^d$
  - $\Gamma_i = 1$ means X_i is included; $\Gamma_i = 0$ means X_i is not included
- Put a prior on Γ,
  - e.g., independent Bernoulli which works in fixed dimension case.
  - In high dimensions, we need to carefully penalize large models
- Given $\gamma \in \{0,1\}^d$, put a $|\gamma|$-dimensional rescaled GP over $X_{[\gamma]} \equiv \{X_i : \gamma_i = 1\}$.
- Essentially, hard spike-and-slab. Alternative methods can be developed.

VS consistency of nonparametric regression

- In high dimensions, with d_0 being fixed, VS consistency is achievable if

$$\limsup_{n\to\infty} d_0 \log(d)/n < c_*$$

(Comminges and Dalalyan, 2012)

- Estimation accuracy?
- How about GP VS?

VS consistency of nonparametric regression

- In high dimensions, with d_0 being fixed, VS consistency is achievable if

$$\limsup_{n\to\infty} d_0 \log(d)/n < c_*$$

(Comminges and Dalalyan, 2012)

- Estimation accuracy?
- How about GP VS?

VS consistency of nonparametric regression

- In high dimensions, with d_0 being fixed, VS consistency is achievable if

$$\limsup_{n \to \infty} d_0 \log(d)/n < c_*$$

(Comminges and Dalalyan, 2012)

- Estimation accuracy?
- How about GP VS?

Variable selection consistency

- In the Bayesian setting, posterior prob. on wrong models goes to 0 in prob..

$$\lim_{n\to\infty} \mathbb{P}_0 \left[\Pi\left(\Gamma \neq \gamma_0 | \mathcal{D}_n\right)\right] = 0$$

- Variable selection via the lens of posterior concentration

$$\Pi\left(\Gamma \neq \gamma_0, ||f - f_0||_2 \leq M\varepsilon_n | \mathcal{D}_n\right)$$

where $\varepsilon_n \asymp n^{-\beta/(2\beta+d_0)} \log^\kappa n$.

- By Schwartz method,

$\Pi\left(||f - f_0||_2 \geq M\varepsilon_n | \mathcal{D}_n\right) \to 0.$

(Ghosal and van der Vaart, 2017)

Variable selection consistency

- In the Bayesian setting, posterior prob. on wrong models goes to 0 in prob..

$$\lim_{n\to\infty} \mathbb{P}_0 \left[\Pi \left(\Gamma \neq \gamma_0 | \mathcal{D}_n\right)\right] = 0$$

- Variable selection via the lens of posterior concentration

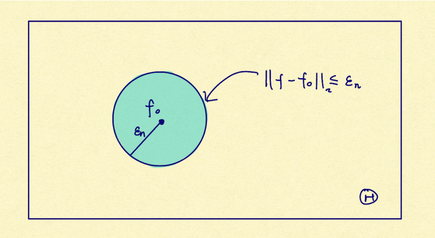
$$\Pi \left(\Gamma \neq \gamma_0, ||f - f_0||_2 \leq M\varepsilon_n | \mathcal{D}_n\right)$$

where $\varepsilon_n \asymp n^{-\beta/(2\beta+d_0)} \log^{\kappa} n$.

- By Schwartz method,

$$\Pi \left(||f - f_0||_2 \geq M\varepsilon_n | \mathcal{D}_n\right) \to 0.$$

(Ghosal and van der Vaart, 2017)



Variable selection consistency

- In the Bayesian setting, posterior prob. on wrong models goes to 0 in prob..

$$\lim_{n\to\infty} \mathbb{P}_0 \left[\Pi\left(\Gamma \neq \gamma_0 | \mathcal{D}_n\right)\right] = 0$$

- Variable selection via the lens of posterior concentration

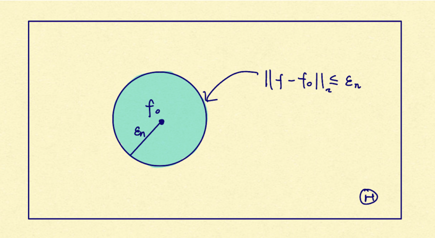
$$\Pi\left(\Gamma \neq \gamma_0, ||f - f_0||_2 \leq M\varepsilon_n | \mathcal{D}_n\right)$$

where $\varepsilon_n \asymp n^{-\beta/(2\beta+d_0)} \log^\kappa n$.

- By Schwartz method,

$$\Pi\left(||f - f_0||_2 \geq M\varepsilon_n | \mathcal{D}_n\right) \to 0.$$

(Ghosal and van der Vaart, 2017)



VS consistency via posterior contraction rates

- The model space $\{\Gamma \neq \gamma_0\}$ decomposes into two parts:

false negative (FN) models	false positive (FP) models
exclude relevant regressors	include all relevant ones + irrelevant ones

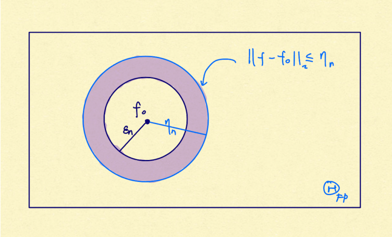
VS consistency via posterior contraction rates

- The model space $\{\Gamma \neq \gamma_0\}$ decomposes into two parts:

false negative (FN) models	false positive (FP) models
exclude relevant regressors	include all relevant ones + irrelevant ones
$\Pi\left(\Gamma \in FN, \|\|f - f_0\|\|_2 \leq M\varepsilon_n \| \mathcal{D}_n\right)$	
beta-min condition	

- The model space $\{\Gamma \neq \gamma_0\}$ decomposes into two parts:

false negative (FN) models	false positive (FP) models
exclude relevant regressors	include all relevant ones + irrelevant ones
$\Pi\left(\Gamma \in FN, \|\|f - f_0\|\|_2 \leq M\varepsilon_n \| \mathcal{D}_n\right)$	$\Pi\left(\Gamma \in FP, \|\|f - f_0\|\|_2 \leq M\varepsilon_n \| \mathcal{D}_n\right)$
beta-min condition	posterior contraction rates slowdown



Polynomial slowdown in posterior contraction rates

- Posterior contraction rates of FP models are slower by n^{κ}.
- The prior mass comparison argument in Castillo (2008).
  - If
    $$\frac{\Pi(E_n)}{\Pi\left(B_{KL}(f_0, \eta_n)\right)} \leq e^{-2n\eta_n^2}$$
    where η_n satisfies $\eta_n \to 0$ and $n\eta_n^2 \to \infty$. Then,
    $$\mathbb{P}_0\left[\Pi(E_n|\mathcal{D}_n)\right] \to 0$$
  - Let $E_n = \{f : \Gamma \in FP(\gamma_0), ||f - f_0||_{L_2(Q_n)} \leq \varepsilon_n\}$
- We need upper bounds of the concentration prob. at every rescaling level.
  - Cf. lower bounds of the concentration prob. at some rescaling levels.
    (van der Vaart and van Zanten, 2009; Bhattacharya et al., 2014; Tokdar, 2011; Yang and Tokdar, 2015)

Polynomial slowdown in posterior contraction rates

- Posterior contraction rates of FP models are slower by n^{κ}.
- The prior mass comparison argument in Castillo (2008).
  - If
    $$\frac{\Pi(E_n)}{\Pi\left(B_{KL}(f_0, \eta_n)\right)} \leq e^{-2n\eta_n^2}$$
    where η_n satisfies $\eta_n \to 0$ and $n\eta_n^2 \to \infty$. Then,
    $$\mathbb{P}_0\left[\Pi(E_n|\mathcal{D}_n)\right] \to 0$$
  - Let $E_n = \{f : \Gamma \in FP(\gamma_0), ||f - f_0||_{L_2(Q_n)} \leq \varepsilon_n\}$
- We need upper bounds of the concentration prob. at every rescaling level.
  - Cf. lower bounds of the concentration prob. at some rescaling levels.

    (van der Vaart and van Zanten, 2009; Bhattacharya et al., 2014; Tokdar, 2011; Yang and Tokdar, 2015)

Polynomial slowdown in posterior contraction rates

- Posterior contraction rates of FP models are slower by n^{κ}.
- The prior mass comparison argument in Castillo (2008).
  - If
    $$\frac{\Pi(E_n)}{\Pi\left(B_{KL}(f_0, \eta_n)\right)} \leq e^{-2n\eta_n^2}$$
    where η_n satisfies $\eta_n \to 0$ and $n\eta_n^2 \to \infty$. Then,
    $$\mathbb{P}_0\left[\Pi(E_n|\mathcal{D}_n)\right] \to 0$$
  - Let $E_n = \{f : \Gamma \in FP(\gamma_0), ||f - f_0||_{L_2(Q_n)} \leq \varepsilon_n\}$
- We need upper bounds of the concentration prob. at every rescaling level.
  - Cf. lower bounds of the concentration prob. at some rescaling levels.
    (van der Vaart and van Zanten, 2009; Bhattacharya et al., 2014; Tokdar, 2011; Yang and Tokdar, 2015)

Small ball probability

- For a GP W on some Banach space $\mathbb{B}$, (shifted) small ball probability:

$$\Pi\left(||W - w_0||_{\mathbb{B}} < \varepsilon\right)$$

- SBP exponent is equivalent to concentration function:

$$\phi_{w_0}(\varepsilon) \leq -\log\Pi\left(||W - w_0||_{\mathbb{B}} < \varepsilon\right) \leq \phi_{w_0}(\varepsilon/2)$$

(See more details in van der Vaart and van Zanten (2008); Kuelbs and Li (1993); Li and Linde (1999))

- Concentration function:

$$\phi_{w_0}(\varepsilon) = \underbrace{\frac{1}{2}\inf_{h\in\mathcal{H}:||h-w_0||_{\mathbb{B}}<\varepsilon}||h||^2_{\mathcal{H}}}_{\text{decentering}} \underbrace{-\log\Pi\left(||W||_{\mathbb{B}} \leq \varepsilon\right)}_{\text{centered}},$$

where $\mathcal{H}$ is the RKHS associated with W.

- In our case, GP is $W^{A,\Gamma}$ with SE kernel, and $\mathbb{B}$ is chosen to be $L_2(Q_n)$.

Small ball probability

- For a GP W on some Banach space $\mathbb{B}$, (shifted) small ball probability:

$$\Pi\left(||W - w_0||_{\mathbb{B}} < \varepsilon\right)$$

- SBP exponent is equivalent to concentration function:

$$\phi_{w_0}(\varepsilon) \leq -\log\Pi\left(||W - w_0||_{\mathbb{B}} < \varepsilon\right) \leq \phi_{w_0}(\varepsilon/2)$$

(See more details in van der Vaart and van Zanten (2008); Kuelbs and Li (1993); Li and Linde (1999))

- Concentration function:

$$\phi_{w_0}(\varepsilon) = \underbrace{\frac{1}{2}\inf_{h\in\mathcal{H}:||h-w_0||_{\mathbb{B}}<\varepsilon}||h||_{\mathcal{H}}^2}_{\text{decentering}} \underbrace{-\log\Pi\left(||W||_{\mathbb{B}} \leq \varepsilon\right)}_{\text{centered}},$$

where $\mathcal{H}$ is the RKHS associated with W.

- In our case, GP is $W^{A,\Gamma}$ with SE kernel, and $\mathbb{B}$ is chosen to be $L_2(Q_n)$.

Small ball probability

- For a GP W on some Banach space $\mathbb{B}$, (shifted) small ball probability:

$$\Pi\left(||W - w_0||_{\mathbb{B}} < \varepsilon\right)$$

- SBP exponent is equivalent to concentration function:

$$\phi_{w_0}\left(\varepsilon\right) \leq -\log\Pi\left(||W - w_0||_{\mathbb{B}} < \varepsilon\right) \leq \phi_{w_0}\left(\varepsilon/2\right)$$

(See more details in van der Vaart and van Zanten (2008); Kuelbs and Li (1993); Li and Linde (1999))

- Concentration function:

$$\phi_{w_0}\left(\varepsilon\right) = \underbrace{\frac{1}{2}\inf_{h\in\mathcal{H}:||h-w_0||_{\mathbb{B}}<\varepsilon}||h||^2_{\mathcal{H}}}_{\text{decentering}} \underbrace{-\log\Pi\left(||W||_{\mathbb{B}} \leq \varepsilon\right)}_{\text{centered}},$$

where $\mathcal{H}$ is the RKHS associated with W.

- In our case, GP is $W^{A,\Gamma}$ with SE kernel, and $\mathbb{B}$ is chosen to be $L_2(Q_n)$.

Small ball probability

- For a GP W on some Banach space $\mathbb{B}$, (shifted) small ball probability:

$$\Pi\left(||W - w_0||_{\mathbb{B}} < \varepsilon\right)$$

- SBP exponent is equivalent to concentration function:

$$\phi_{w_0}(\varepsilon) \leq -\log\Pi\left(||W - w_0||_{\mathbb{B}} < \varepsilon\right) \leq \phi_{w_0}(\varepsilon/2)$$

(See more details in van der Vaart and van Zanten (2008); Kuelbs and Li (1993); Li and Linde (1999))

- Concentration function:

$$\phi_{w_0}(\varepsilon) = \underbrace{\frac{1}{2}\inf_{h\in\mathcal{H}:||h-w_0||_{\mathbb{B}}<\varepsilon}||h||_{\mathcal{H}}^2}_{\text{decentering}} \underbrace{-\log\Pi\left(||W||_{\mathbb{B}} \leq \varepsilon\right)}_{\text{centered}},$$

where $\mathcal{H}$ is the RKHS associated with W.

- In our case, GP is $W^{A,\Gamma}$ with SE kernel, and $\mathbb{B}$ is chosen to be $L_2(Q_n)$.

The decentering	The centered
$\inf_{h \in \mathcal{H}: \lvert\lvert h - w_0 \rvert\rvert_2 < \varepsilon} \lvert\lvert h \rvert\rvert_{\mathcal{H}}^2$	$-\log\Pi\left(\lvert\lvert W \rvert\rvert_{L_2(Q_n)} \leq \varepsilon\right)$

Technical developments: SBP estimates

The decentering	The centered
$\inf_{h \in \mathcal{H}: \lVert h - w_0 \rVert_2 < \varepsilon} \lVert h \rVert_{\mathcal{H}}^2$	$-\log \Pi \left(\lVert W \rVert_{L_2(Q_n)} \leq \varepsilon \right)$
function approximation	

(van der Vaart and van Zanten, 2011)

Technical developments: SBP estimates

The decentering	The centered
$\inf_{h \in \mathcal{H}: \|\|h - w_0\|\|_2 < \varepsilon} \|\|h\|\|^2_{\mathcal{H}}$	$-\log \Pi \left(\|\|W\|\|_{L_2(Q_n)} \leq \varepsilon \right)$
function approximation	the metric entropy method
(van der Vaart and van Zanten, 2011)	(Kuelbs and Li, 1993; Li and Linde, 1999)
	Metric entropy of the RKHS unit ball
	Series representation of SEGP

Details: series representation

- SE kernel's series expansion
  - Univariate case under Gaussian design:

$$K_{a,1}(s,t) = e^{-a^2(s-t)^2} = \sum_{j=0}^{\infty} \lambda_j \varphi_j(s) \overline{\varphi_j(t)},$$

(See Chapter 4.3 of Rasmussen and Williams (2006) for more details.)

  - For model γ,

$$K_{a,\gamma}(s,t) = \prod_{\{i:\gamma_i=1\}} \left(\sum_{j=0}^{\infty} \lambda_j^{(i)} \varphi_j^{(i)}(s_i) \overline{\varphi_j^{(i)}(t_i)} \right) \equiv \sum_{k=0}^{\infty} \mu_k^{(\gamma)} \psi_k^{(\gamma)}(s) \overline{\psi_k^{(\gamma)}(t)}$$

- Series representation of $W^{a,\gamma}$, aka, Karhunen-Loève expansion: for $Z_j \overset{iid}{\sim} N(0,1)$, $j = 0, 1, \ldots$,

$$W_t^{a,\gamma} = \sum_{j=0}^{\infty} Z_j \sqrt{\mu_j^{(\gamma)}} \psi_j^{(\gamma)}(t).$$

  - $W^{a,\gamma} \in L_2(Q_n)$, a.s..
  - RKHS unit ball $\mathcal{H}_1^{a,\gamma}$ of $W^{a,\gamma}$:

$$\left\{ \{\theta_j\}_{j=1}^{\infty} : \sum_{j=1}^{\infty} \theta_j^2 / \mu_j^{(\gamma)} \le 1 \right\} \subseteq \ell^2(\mathbb{N}),$$

whose metric entropy can be sharply bounded. (See example 5.12 of Wainwright (2019))

Details: series representation

- SE kernel's series expansion
  - Univariate case under Gaussian design:

$$K_{a,1}(s,t) = e^{-a^2(s-t)^2} = \sum_{j=0}^{\infty} \lambda_j \varphi_j(s) \overline{\varphi_j(t)},$$

(See Chapter 4.3 of Rasmussen and Williams (2006) for more details.)

  - For model γ,

$$K_{a,\gamma}(s,t) = \prod_{\{i:\gamma_i=1\}} \left(\sum_{j=0}^{\infty} \lambda_j^{(i)} \varphi_j^{(i)}(s_i) \overline{\varphi_j^{(i)}(t_i)} \right) \equiv \sum_{k=0}^{\infty} \mu_k^{(\gamma)} \psi_k^{(\gamma)}(s) \overline{\psi_k^{(\gamma)}(t)}$$

- Series representation of $W^{a,\gamma}$, aka, Karhunen-Loève expansion: for $Z_j \overset{iid}{\sim} N(0,1)$, $j = 0, 1, \ldots,$

$$W_t^{a,\gamma} = \sum_{j=0}^{\infty} Z_j \sqrt{\mu_j^{(\gamma)}} \psi_j^{(\gamma)}(t).$$

  - $W^{a,\gamma} \in L_2(Q_n)$, $a.s.$.
  - RKHS unit ball $\mathcal{H}_1^{a,\gamma}$ of $W^{a,\gamma}$:

$$\left\{ \{\theta_j\}_{j=1}^{\infty} : \sum_{j=1}^{\infty} \theta_j^2 / \mu_j^{(\gamma)} \leq 1 \right\} \subseteq \ell^2(\mathbb{N}),$$

  whose metric entropy can be sharply bounded. (See example 5.12 of Wainwright (2019))

Discussions

- Gaussian random design
  - necessary for Karhunen-Loève expansion of the SE covariance kernel
- Limited smoothness of f_0: f_0 is about $\beta-$smooth.
  - Cf. the self-similarity assumption, minimal "signal strength"
  - VS consistency can be shown for a wider class of functions. If f_0 is ∞ smooth, consider

$$\tilde{Y}_i = Y_i + g(Z_i) = f_0(X_i) + g(Z_i) + \epsilon_i$$

    where g has finite smoothness and Z_i is generated. Regress $\tilde{Y}_i$ on (X_i, Z_i).

Discussions

- Gaussian random design
  - necessary for Karhunen-Loève expansion of the SE covariance kernel
- Limited smoothness of f_0: f_0 is about $\beta-$smooth.
  - Cf. the self-similarity assumption, minimal "signal strength"
  - VS consistency can be shown for a wider class of functions.
    If f_0 is ∞ smooth, consider

$$\tilde{Y}_i = Y_i + g(Z_i) = f_0(X_i) + g(Z_i) + \epsilon_i$$

    where g has finite smoothness and Z_i is generated.
    Regress $\tilde{Y}_i$ on (X_i, Z_i).

More discussions

- VS consistency v.s. estimation optimality
  - Design dimension growth rate: $\log d_n \precsim n^{d_0/(2\beta+d_0)}$
  - Consistent VS is possible for $\log d_n = O(n)$. (Comminges and Dalalyan, 2012)
- Continuous shrinkage approach to GP VS?
- See more details and rigorous statements in the paper available at

  https://arxiv.org/pdf/1912.05738.pdf

More discussions

- VS consistency v.s. estimation optimality
  - Design dimension growth rate: $\log d_n \precsim n^{d_0/(2\beta+d_0)}$
  - Consistent VS is possible for $\log d_n = O(n)$. (Comminges and Dalalyan, 2012)
- Continuous shrinkage approach to GP VS?
- See more details and rigorous statements in the paper available at

  `https://arxiv.org/pdf/1912.05738.pdf`

Posterior contraction rates

- Schwartz method to establish as $n \to \infty$, for every large M,

$$\mathbb{P}_0[\Pi(f : \rho(f, f_0) \geq M\varepsilon_n)] \to 0.$$

  - Existence of exponentially powerful test on certain sieve $\mathcal{F}_n$
    - If ρ is dominated by Hellinger distance $\to$ prior mass condition:

$$\Pi_n(f \in B_n(f_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$$

    where $B_n(g, \epsilon) = \{f : K(\mathbb{P}_g^1, \mathbb{P}_f^1) \leq \epsilon^2, V(\mathbb{P}_g^1, \mathbb{P}_f^1) \leq \epsilon^2\}$,
    K is KL divergence; V is KL variation.
  - The complexity of the sieve is under control: $\log N(\varepsilon_n, \mathcal{F}_n, \rho) \leq n\varepsilon_n^2$
  - The sieve is essentially the parameter space: $\Pi(\mathcal{F}_n) \geq 1 - e^{-Cn\varepsilon_n^2}$
- In our context, $\rho(\cdot, \cdot)$ is $L_2(Q_n)$ distance.
  - Prior mass condition w.r.t.

$$B_n(g, \epsilon) = \{f \in L_2(Q_n) : \|f - g\|_{L_2(Q_n)} \leq \epsilon\}.$$

  - Construct a sieve $\mathbb{B}_n \subset L_2(Q_n)$

(Ghosal and van der Vaart, 2017)

Posterior contraction rates

- Schwartz method to establish as $n \to \infty$, for every large M,

$$\mathbb{P}_0[\Pi(f : \rho(f, f_0) \geq M\varepsilon_n)] \to 0.$$

  - Existence of exponentially powerful test on certain sieve $\mathcal{F}_n$
    - If ρ is dominated by Hellinger distance $\to$ prior mass condition:

$$\Pi_n(f \in B_n(f_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$$

    where $B_n(g, \epsilon) = \{f : K(\mathbb{P}_g^1, \mathbb{P}_f^1) \leq \epsilon^2, V(\mathbb{P}_g^1, \mathbb{P}_f^1) \leq \epsilon^2\}$,
    K is KL divergence; V is KL variation.
  - The complexity of the sieve is under control: $\log N(\varepsilon_n, \mathcal{F}_n, \rho) \leq n\varepsilon_n^2$
  - The sieve is essentially the parameter space: $\Pi(\mathcal{F}_n) \geq 1 - e^{-Cn\varepsilon_n^2}$
- In our context, $\rho(\cdot, \cdot)$ is $L_2(Q_n)$ distance.
  - Prior mass condition w.r.t.

$$B_n(g, \epsilon) = \{f \in \mathcal{L}_2(Q_n) : \|f - g\|_{L_2(Q_n)} \leq \epsilon\}.$$

  - Construct a sieve $\mathbb{B}_n \subset L_2(Q_n)$

(Ghosal and van der Vaart, 2017)

References

Bhattacharya, A., D. Pati, and D. Dunson (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Annals of statistics 42*(1), 352.

Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics 2*, 1281–1299.

Comminges, L. and A. S. Dalalyan (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics 40*(5), 2667–2696.

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*, Volume 44. Cambridge University Press.

Kuelbs, J. and W. V. Li (1993). Metric entropy and the small ball problem for Gaussian measures. *Journal of Functional Analysis 116*(1), 133–157.

Li, W. V. and W. Linde (1999). Approximation, metric entropy and small ball estimates for Gaussian measures. *The Annals of Probability 27*(3), 1556–1578.

Rasmussen, C. E. and C. K. Williams (2006). *Gaussian processes for machine learning*, Volume 1. MIT press Cambridge.

Savitsky, T. D., M. Vannucci, and N. Sha (2011). Variable selection for nonparametric gaussian process priors: Models and computational strategies. *Statistical science 26*(1), 130–149.

Tadesse, M. G. and M. Vannucci (2021). Handbook of bayesian variable selection.

Tokdar, S. T. (2011). Dimension adaptability of Gaussian process models with variable selection and projection. *arXiv preprint arXiv:1112.0716*.

van der Vaart, A. and H. van Zanten (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research 12*(Jun), 2095–2119.

van der Vaart, A. W. and J. H. van Zanten (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pp. 200–222. Institute of Mathematical Statistics.

van der Vaart, A. W. and J. H. van Zanten (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 2655–2675.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press.

Yang, Y. and S. T. Tokdar (2015). Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics 43*(2), 652–674.