STAT 625 - Advanced Bayesian Inference

Lecture 1

Meng Li

Department of Statistics

Bayes Nonparametrics

- Nonparametric statistical models are increasingly replacing parametric models, to overcome the latter's inflexibility to address a wide variety of data.
- A nonparametric model involves at least one infinite-dimensional parameter (such as a function or measure) and hence may also be referred to as an "infinite-dimensional model".
- Keeping it aside to specify a prior distribution, the Bayesian approach is extremely straightforward, in principle.
- The full inference is based on the posterior distribution only.

Linear regression

- Linear regression:

  $$y = f(\boldsymbol{x}) + \varepsilon$$
  $$f(\boldsymbol{x}) = \boldsymbol{x}^T \cdot \boldsymbol{\beta},$$

  where $\boldsymbol{x} \in \mathbb{R}^p$ and $\varepsilon \sim N(0, \sigma^2)$.
- Linear regression can capture non-linear shapes via basis functions, i.e.,
  - $$f(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_N(\boldsymbol{x}))^T \cdot \boldsymbol{\beta}.$$
  - Popular basis systems:
    - wavelets
    - trigonometric functions
    - polynomials
    - splines, etc.
  - Is this a nonparametric model?

Bayesian linear regression

- Model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the design matrix $\boldsymbol{X}$ is $n \times p$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$.
- Suppose the variance σ^2 is known.
- Prior: $\boldsymbol{\beta} \sim N(0, \Sigma_p)$
- Then the posterior of $\boldsymbol{\beta}$ is

$$\begin{aligned}\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{y} &\sim N\left(\bar{\boldsymbol{\beta}}, \boldsymbol{A}^{-1}\right)\\ \boldsymbol{A} &= \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \Sigma_p^{-1}\\ \bar{\boldsymbol{\beta}} &= \sigma^{-2}\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y} = \left(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\Sigma_p^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

- Predictive density for the mean $f(x_*)$ at a new location x_* is

$$f(\boldsymbol{x}_*) | \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y} \sim N\left(\boldsymbol{x}_*^T\bar{\boldsymbol{\beta}}, \boldsymbol{x}_*^T\boldsymbol{A}^{-1}\boldsymbol{x}_*\right)$$

- Predictive density for the response at a new location x_* is

$$\boldsymbol{y}_* | \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y} \sim N\left(\boldsymbol{x}_*^T\bar{\boldsymbol{\beta}}, \boldsymbol{x}_*^T\boldsymbol{A}^{-1}\boldsymbol{x}_* + \sigma^2\right)$$

Nonparametric regression

- **Nonparametric regression**: using infinitely many parameters characterizing the regression function $f(\cdot)$ evaluated at all possible predictor values $\boldsymbol{x}$.

- **Weight space view**
  - Restrict attention to a grid of $\boldsymbol{x}$-values: $\boldsymbol{x}_1, \boldsymbol{x}_2, .., \boldsymbol{x}_n$.
  - Put a joint prior on the n function values: $f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), ..., f(\boldsymbol{x}_n)$.

- **Function space view**
  - Treat f as an unknown function.
  - Put a prior over a set of functions.
- Kolmogorov's existence theorem for stochastic processes equates the two views.
  - Just make sure that the set of finite-dimensional distributions are consistent: symmetric to permutation and marginalization.

Gaussian process regression

- Weight-space view. GP assumes

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\boldsymbol{m}, \boldsymbol{K})$$

- But how do we specify the $k \times k$ **covariance matrix $\boldsymbol{K}$**?

$$Cov(f(x_p), f(x_q))$$

  - An example of covariance function:

$$Cov(f(x_p), f(x_q)) = k(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(x_p - x_q)^2\right)$$

  - Nearby x's have highly correlated function ordinates $f(x)$.
  - We can compute $Cov(f(x_p), f(x_q))$ for *any* x_p and x_q.
  - Extension to multiple covariates: $(x_p - x_q)$ replaced by $\|\boldsymbol{x}_p - \boldsymbol{x}_q\|$.

Gaussian process regression, cont.

Definition

A **Gaussian process** (**GP**) is a stochastic process $W = (W_t : t \in T)$ indexed by an arbitrary set T such that the vector $(W_{t_1}, \ldots, W_{t_k})$ possesses a multivariate normal distribution, for every $t_1, \ldots, t_k \in T$ and $k \in \mathbb{N}$.

- Therefore, a Gaussian process is a collection of random variables, any finite number of which have a multivariate Gaussian distribution.
- A GP is a probability distribution of functions. No need for a grid!
- A GP is completely specified by a **mean** and a **covariance kernel**

$$m(x) = \mathrm{E}[f(x)]$$

$$k(x,x') = E\left[(f(x) - m(x))\left(f(x') - m(x')\right)\right]$$

for any two inputs x and x'.

- A **Gaussian process** is denoted by

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

- The mean function $m(\cdot)$ is an arbitrary function from T to $\mathbb{R}$.
  - It is often taken equal to zero as a prior; a shift to a nonzero mean can also be incorporated in the model.
- The covariance kernel is a bilinear, symmetric nonnegative-definite function from $T \times T$ to $\mathbb{R}$.
- There exists a Gaussian process for any mean function and covariance kernel.
- From a Bayesian point of view, $f(x) \sim GP$ describes **prior beliefs** about the unknown $f(\cdot)$.

- Example (squared exponential GP):

$$m(x) = 0, \qquad k(x,x') = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x-x'}{\ell}\right)^2\right)$$

  - Here $\ell > 0$ is the length scale parameter controlling smoothness.
    - Larger ℓ gives more smoothness in $f(x)$.
  - σ_f^2 controls the magnitude.
- Simulate draw from $f(x) \sim GP(m(x), k(x,x'))$ over a grid $\boldsymbol{x}_* = (x_1, ..., x_n)$ by using that

$$f(\boldsymbol{x}_*) \sim N(m(\boldsymbol{x}_*), K(\boldsymbol{x}_*, \boldsymbol{x}_*))$$

- Note that the **kernel** $k(x,x')$ produces a **covariance matrix** $K(\boldsymbol{x}_*, \boldsymbol{x}_*)$ when evaluated at the vector $\boldsymbol{x}_*$.

Simulating a GP

- The joint way: Choose a grid $x_1, ..., x_k$. Simulate the k-vector

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\boldsymbol{m}, \boldsymbol{K})$$

- The conditional decomposition:

$$\begin{aligned} p(f(x_1), f(x_2), ..., f(x_k)) &= p(f(x_1))\, p(f(x_2)|f(x_1)) \cdots \\ &\times p(f(x_k)|f(x_1), ..., f(x_{k-1})) \end{aligned}$$

The posterior for a GPR

- **Model**: $y_i = f(x_i) + \varepsilon_i, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$
- **Prior**: $f(x) \sim GP(0, k(x, x'))$.
- Data: $\boldsymbol{x} = (x_1, ..., x_n)^T$ and $\boldsymbol{y} = (y_1, ..., y_n)^T$.
- Goal: the posterior of $f(\cdot)$ over a grid of x-values: $\boldsymbol{f}_* = \boldsymbol{f}(\boldsymbol{x}_*)$.
- Intermediate step: joint distribution of $\boldsymbol{y}$ and $\boldsymbol{f}_*$

$$\left(\begin{array}{c} \boldsymbol{y} \\ \boldsymbol{f}_* \end{array} \right) \sim N \left\{ \left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left[\begin{array}{cc} K(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 I & K(\boldsymbol{x}, \boldsymbol{x}_*) \\ K(\boldsymbol{x}_*, \boldsymbol{x}) & K(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{array} \right] \right\}$$

- The **posterior**

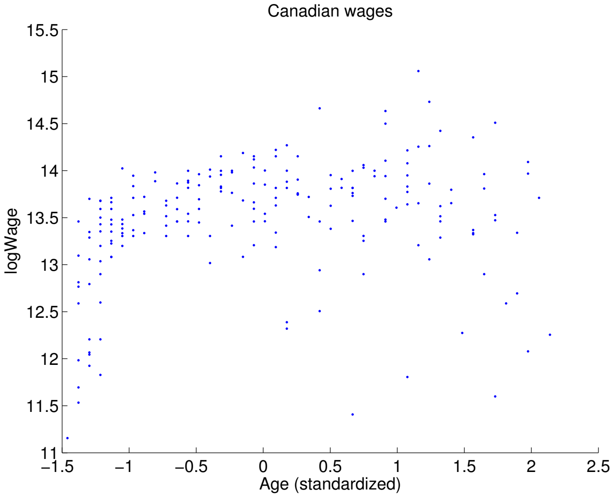
$$\boldsymbol{f}_* | \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}_* \sim N\left(\bar{\boldsymbol{f}}_*, \text{cov}(\boldsymbol{f}_*)\right)$$

$$\bar{\boldsymbol{f}}_* = K(\boldsymbol{x}_*, \boldsymbol{x}) \left[K(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 I\right]^{-1} \boldsymbol{y}$$

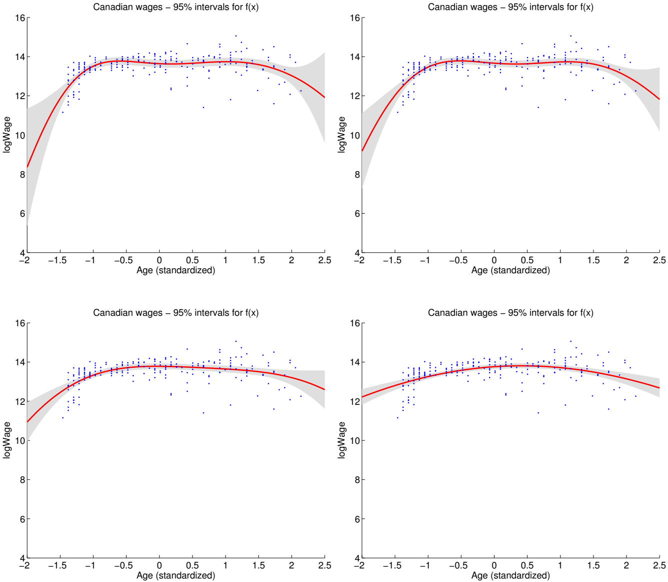
$$\text{cov}(\boldsymbol{f}_*) = K(\boldsymbol{x}_*, \boldsymbol{x}_*) - K(\boldsymbol{x}_*, \boldsymbol{x}) \left[K(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 I\right]^{-1} K(\boldsymbol{x}, \boldsymbol{x}_*)$$

- **Computational complexity**: $O(n^3)$ for matrix inversion. It needs to be repeated at each MCMC step if we change hyperparameters. Hence, the computation becomes challenging for large n or large p.

Example - Canadian wages



Posterior of f - $\ell = 0.2, 0.5, 1, 2$



Prediction and Decision

- Predicting a new set of y-values $\boldsymbol{y}_* = f(\boldsymbol{x}_*) + \varepsilon$ is easy

$$\boldsymbol{y}_* | \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}_* \sim N\left(\bar{\boldsymbol{f}}_*, \mathrm{cov}(\boldsymbol{f}_*) + \sigma^2 I\right)$$

- Choosing a point prediction $\boldsymbol{y}_{guess}$ by maximizing expected utility

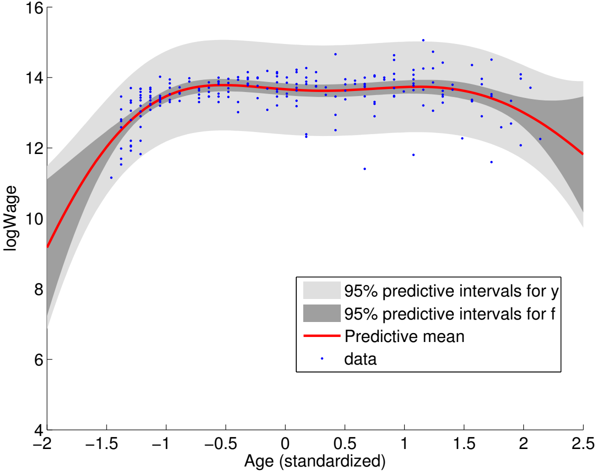
$$\bar{\mathscr{U}}(\boldsymbol{y}_{guess} | \boldsymbol{x}_*) = \int \mathscr{U}(\boldsymbol{y}_*, \boldsymbol{y}_{guess}) p(\boldsymbol{y}_* | \boldsymbol{x}_*, \boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{y}_*$$

- Have to make a decision $a \in \mathscr{A}$ whose consequences (utility) depends on the uncertain $\boldsymbol{f}_*$ (or $\boldsymbol{y}_*$)? Just maximize expected utility

$$\bar{\mathscr{U}}(a) = \int \mathscr{U}(a, \boldsymbol{f}_*) p(\boldsymbol{f}_* | \boldsymbol{x}_*, \boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{f}_*$$

where $\mathscr{U}(a, \boldsymbol{f}_*)$ is the utility of action $a \in \mathscr{A}$ if $\boldsymbol{f}_*$ turns out to be the “true state of the world”.

Canadian wages - prediction with $\ell = 0.5$



Stationary processes and smoothness

- A stochastic process (field) $\{f(\boldsymbol{x}), x \in \mathbb{R}^p\}$ is **weakly stationary** if $E(f(\boldsymbol{x})) = \mu$ and its covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$ is a function of $\boldsymbol{t} = \boldsymbol{x} - \boldsymbol{x}'$

$$k(\boldsymbol{x}, \boldsymbol{x}') = Cov\left[f(\boldsymbol{x}), f(\boldsymbol{x}')\right] = k(\boldsymbol{t}).$$

- The covariance function is **isotropic** if it only depends on the distance $t = \|\boldsymbol{x} - \boldsymbol{x}\|$ (invariant to directions)

$$k(\boldsymbol{x}, \boldsymbol{x}') = Cov\left[f(\boldsymbol{x}), f(\boldsymbol{x}')\right] = k(t).$$

- The **smoothness** of a stationary process is determined by the smoothness of the covariance function.
- A stationary (isotropic) process is **continuous in quadratic mean**

$$E\left(|f(\boldsymbol{x} + t) - f(\boldsymbol{x})|^2\right) \to 0 \text{ as } t \to 0$$

iff $k(t)$ is continuous at $t = 0$.

Commonly used covariance kernels

- Let $r = \|x - x'\|$. All kernels can be scaled by $\sigma_f > 0$.
- **Squared exponential** (**SE**) ($\ell > 0$)

$$K_{SE}(r) = \exp\left(-\frac{r^2}{2\ell^2}\right)$$

  - Infinitely mean square differentiable. Very smooth.
- **Matérn** ($\ell > 0,\ \nu > 0$)

$$K_{Matern}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu r}}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu r}}{\ell}\right)$$

  - Here $\Gamma(\cdot)$ is the Gamma function, and K_ν is the modified Bessel function of the second kind.
  - As $\nu \to \infty$, Matérn's kernel approaches SE kernel. Very rough.

Commonly used covariance kernels, cont.

- **γ-exponential** ($\ell > 0$, $0 < \gamma \leq 2$)

$$K_\gamma(r) = \exp\left[-\left(\frac{r}{\ell}\right)^\gamma\right]$$

  - Mean square differentiable only when $\gamma = 2$ (SE).
- **Rational quadratic** ($\ell > 0$, $\alpha > 0$)

$$K_{RQ}(r) = \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha}$$

  - Scale mixture of SE covariance functions with different length scales.
  - $K_{RQ}(r)$ approaches the SE kernel as $\alpha \to \infty$.

More on kernels

- Anisotropic version of isotropic kernels by setting $r^2(\boldsymbol{x},\boldsymbol{x}') = (\boldsymbol{x}-\boldsymbol{x}')^T \boldsymbol{M}(\boldsymbol{x}-\boldsymbol{x}')$ where $\boldsymbol{M}$ is positive definite.
- **Automatic Relevance Determination** (ARD): $\boldsymbol{M} = Diag(\ell_1^{-2},...,\ell_p^{-2})$ is diagonal with different length scales.
- **Factor kernels**: $M = \Lambda\Lambda^T + \Psi$, where Λ is $p \times k$ for low rank k.
- Kernels are often combined into **composite kernels**. The sum of kernels is a kernel. The product of kernels is a kernel.
- Kernels can be used for non-vectorial inputs by defining distance functions between objects (e.g., words). String kernels for text analysis. Fisher kernels.

Hyperparameters

- The kernel can depend on hyperparameters θ. Example: SE kernel $[\theta = (\sigma_f, \ell)^T]$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{\ell^2}\right)$$

- We have two strategies for unknown hyperparameters.
- The first strategy proceeds with computing the posterior

$$p(\theta|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\boldsymbol{X}, \theta)p(\theta|\boldsymbol{X}).$$

- We need to compute

$$p(\boldsymbol{y}|\boldsymbol{X}, \theta) = \int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{f}, \theta)p(\boldsymbol{f}|\boldsymbol{X}, \theta)d\boldsymbol{f}$$

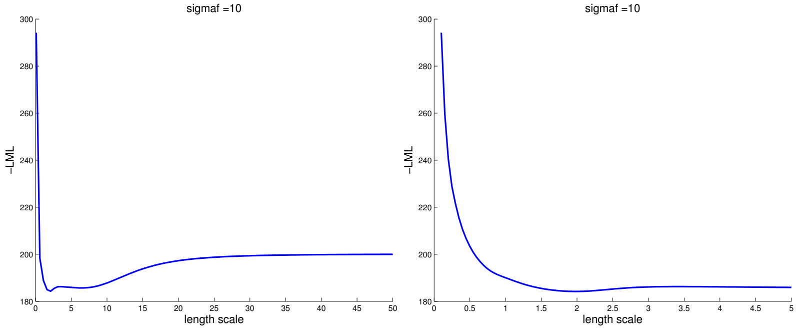
where $\boldsymbol{f} = f(\boldsymbol{X})$ is a vector with function values in the training data.

- For Gaussian process regression, the marginal likelihood of data is analytically available: [since $\boldsymbol{y}|\boldsymbol{X},\theta \sim N(0, K+\sigma^2 I)$]

$$\log p(\boldsymbol{y}|\boldsymbol{X},\theta) = -\frac{1}{2}\boldsymbol{y}^T\left(K+\sigma^2 I\right)^{-1}\boldsymbol{y} - \frac{1}{2}\log\left|K+\sigma^2 I\right| - \frac{n}{2}\log(2\pi)$$

- We may choose θ by maximizing $\log p(\boldsymbol{y}|\boldsymbol{X},\theta)$ (maximum marginal likelihood estimate, or MMLE; sometimes it is called Type-2 MLE).
- The second strategy: A fully Bayesian approach would use a prior on θ.

Canadian wages - determination of ℓ


sigmaf =10
−LML
length scale
sigmaf =10
−LML
length scale