Exploratory data analysis

Daeyoung Lim[*]
Department of Statistics
Korea University

July 11, 2016

1 Pendulum Data

1.1 Data Description

The pendulum data obtained from the paper of David Nott consists of 9 covariates and one target variable. As described in the original paper, it is a simulated data of mechanical pendulum, covariates of which are different parameters of the system and the target variable is the angular velocity.

2 SSGP

For a short summary of Lazaro Gredilla's SSGP, the paper comes up with a decomposition of the function into

$$f\left(\mathbf{x}\right) = \sum_{r=1}^{m} a_r \cos\left(2\pi \mathbf{s}_r^\top \mathbf{x}\right) + b_r \sin\left(2\pi \mathbf{s}_r^\top \mathbf{x}\right) \tag{1}$$

where $a_r \sim \mathcal{N}\left(0, m^{-1}\sigma_0^2\right)$, $b_r \sim \mathcal{N}\left(0, m^{-1}\sigma_0^2\right)$. Then, by transforming $\mathbf{x}$ into

$$\phi\left(\mathbf{x}\right) = \begin{bmatrix} \cos\left(2\pi \mathbf{s}_1^\top \mathbf{x}\right) & \sin\left(2\pi \mathbf{s}_1^\top \mathbf{x}\right) & \cdots & \cos\left(2\pi \mathbf{s}_m^\top \mathbf{x}\right) & \sin\left(2\pi \mathbf{s}_m^\top \mathbf{x}\right) \end{bmatrix}, \tag{2}$$

the end result becomes similar to linear Gaussian process regression model:

$$\mathrm{E}\left(\mathbf{y}_*\right) = \phi\left(\mathbf{x}_*\right)^\top A^{-1}\mathbf{\Phi}\mathbf{y} \tag{3}$$

$$\mathrm{Var}\left(\mathbf{y}_*\right) = \sigma_n^2 + \sigma_n^2 \phi\left(\mathbf{x}_*\right)^\top A^{-1}\phi\left(\mathbf{x}_*\right). \tag{4}$$

The author suggests learning the parameters via optimizing the log-marginal likelihood:

$$\log p\left(\mathbf{y}|\theta\right) = -\frac{1}{2\sigma_n^2}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{\Phi}^\top A^{-1}\mathbf{\Phi}\mathbf{y}\right) - \frac{1}{2}\log|A| + m\log\frac{m\sigma_m^2}{\sigma_0^2} - \frac{n}{2}\log\left(2\pi\sigma_n^2\right) \tag{5}$$

by means of the conjugate gradient method.

2.1 Sidenote

The predictive mean of Gaussian process regression model is

$$p\left(f_*|\mathbf{x}_*, X, \mathbf{y}\right) = \int p\left(f_*|\mathbf{x}_*, \mathbf{w}\right) p\left(\mathbf{w}|X, \mathbf{y}\right)\, d\mathbf{w} = \int \mathbf{x}_*^\top \mathbf{w} p\left(\mathbf{w}|X, \mathbf{y}\right)\, d\mathbf{w} \tag{6}$$

$$= \mathcal{N}\left(\frac{1}{\sigma_n^2}\mathbf{x}_*^\top A^{-1} X\mathbf{y}, \mathbf{x}_*^\top A^{-1}\mathbf{x}_*\right). \tag{7}$$

[*] Prof. Taeryon Choi

The posterior distribution of the weights $\mathbf{w}$ is

$$p\left(\mathbf{w}|X,\mathbf{y}\right) \sim \mathcal{N}\left(\overline{\mathbf{w}} = \frac{1}{\sigma_n^2}A^{-1}X\mathbf{y}, A^{-1}\right) \tag{8}$$

where $A = \sigma_n^{-2}XX^\top + \Sigma_p^{-1}$. Therefore, this is the reason why we can plug in the test data $\widetilde{X}$ into the predictive distribution's $\mathbf{x}_*$ to get the fitted values of the unknown function, $\hat{f}$. (In the case of sparse spectrum decomposition, the design matrix is no longer X but rather Φ.)

3 VA for partially linear additive models

The paper suggests a model of the form

$$y_i = \mu + \sum_{j=1}^{p} f_j\left(x_{ij}\right) + \epsilon_i, \quad i = 1, \ldots, n. \tag{9}$$

- In the authors' exact words, *for simplicity of exposition*, they assumed all covariates to be continuous with support $[0,1]$.
- Discrete variables are put into the linear part.
- Impose $\mathrm{E}\left(f_j\left(x_j\right)\right) = 0$ constraint to achieve identiability.
- Transform the basis functions in such a way that they are orthogonal to the linear basis functions.
- Essentially VB for parameter estimation but offered some degree of freedom between Monte Carlo estimation and Laplace approximation for the intractable integration terms.

The real data analysis section of the original paper mentions that the crime data in R package {Ecdat} was used, which has 630 observations and 22 variables.

3.1 BPLAM code review

- Since one of the purposes of this paper was to propose a novel way of selecting variables using variational approximation, the code assumes there would be more than one covariate. The code crashes with only `x` of one column.
- The code still worked even though one of the covariates was outside of the range $[0,1]$.
- Functions are not self-contained!!!!! Who codes like this!?

4 Kneib VA

Model:

$$y_i = \mathbf{x}^\top\boldsymbol{\beta} + f_1\left(\mathrm{herdsize}_i\right) + f_2\left(\mathrm{capital}_i\right) + f_{\mathrm{spat}}\left(\mathrm{county}_i\right) + \epsilon_i. \tag{10}$$

Code doesn't work...

5 Datasets

5.1 Pole Telecomm and Elevators

According to Lazaro Gredilla, the data sets are taken from `http://www.liaad.up.pt/ ltorgo/Regression/Dat`

6 Transforming data with support $\mathcal{S}$

Normally, the data don't lie within the interval $[0,1]$. Therefore, we must transform them according to their support. One of the distributions with universal support is the Cauchy distribution. The pdf and cdf are

$$q\left(x\right)=\frac{1}{\pi\left(1+x^{2}\right)} \tag{11}$$

$$Q\left(x\right)=\frac{1}{\pi}\tan^{-1}\left(x\right)+\frac{1}{2}. \tag{12}$$

Now define $\varphi_0\left(x\right)=\sqrt{q\left(x\right)}$ and $\varphi_j\left(x\right)=\sqrt{2q\left(x\right)}\cos\left[\pi jQ\left(x\right)\right]$. Let's do one of the shape restriction models. Recall

$$\varphi^{a}_{j,k}\left(x\right)=\int_0^x\varphi_j\left(s\right)\overline{\varphi}_k\left(s\right)\,dx-\int_0^1\int_0^s\varphi_j\left(t\right)\overline{\varphi}_k\left(t\right)\,dt\,ds\ \text{ for } j,k\geq 0. \tag{13}$$

With the defined basis functions above,

$$\int_0^x\varphi_j\left(s\right)\overline{\varphi}_k\left(s\right)\,dx=\sqrt{2}\int_0^x\frac{1}{\pi\left(1+x^{2}\right)}\cos\left[j\tan^{-1}\left(s\right)+\frac{\pi j}{2}\right]\,ds \tag{14}$$

$$=\sqrt{2}\int_{Q(0)}^{Q(x)}\cos\left[\pi ju\right]\,du\quad\left(u=Q\left(x\right)\right) \tag{15}$$

$$=\frac{\sqrt{2}}{\pi j}\left\{\sin\left(j\tan^{-1}\left(x\right)+\frac{\pi j}{2}\right)-\sin\left(\frac{\pi j}{2}\right)\right\} \tag{16}$$

$$\int_0^1\int_0^s\varphi_j\left(t\right)\overline{\varphi}_k\left(t\right)\,dt\,ds=\int_0^1\int_{Q(0)}^{Q(s)}2\cos\left[\pi ju\right]\cos\left[\pi ku\right]\,du\,ds \tag{17}$$

$$=\int_0^1\frac{\sin\left(\pi\left(j-k\right)Q\left(s\right)\right)}{\pi\left(j-k\right)}+\frac{\sin\left(\pi\left(j+k\right)Q\left(s\right)\right)}{\pi\left(j+k\right)}\,ds \tag{18}$$

$$-\frac{\sin\left(\pi\left(j-k\right)/2\right)}{\pi\left(j-k\right)}-\frac{\sin\left(\pi\left(j+k\right)/2\right)}{\pi\left(j+k\right)} \tag{19}$$

Comparing with the one with support $[0,1]$ suggested in the original paper,

$$\int_0^x\varphi_j\left(s\right)\overline{\varphi}_k\left(s\right)\,ds=\frac{\sqrt{2}}{\pi j}\sin\left(\pi jx\right) \tag{20}$$

there is another term added and care must be taken in that we shouldn't simply apply the cdf to the data and do the rest equally. The math tells us it becomes different.

7 `ssgpr.R` manual

List of functions:

- `vbgpspectral`
- `minim`
- `ssgpr`

- `ssgpr_ui`
- `compareSSGPvsBSAR`

Core function is `compareSSGPvsBSAR` and all else can be ignored.

7.1 compareSSGPvsBSAR

Arguments:

- `data`
  - Choose one from `c('pendulum', 'elevators', 'kin', 'pol', 'pumadyn', 'simul')`.
  - All function similarly except for `'simul'` since it is the only option for simulation. All else are real data.
  - If `data` is set to other than `simul`, only the main title of the plot generated is affected.
  - **(Caution!)** If you choose one other than `simul` and set other path variable and filenames differently, then the function will not throw any error message, which could in the end be misleading!
- `fit`
  - Choose one from `c('training', 'test')`.
  - If `fit` is set to `'training'`, two of the subsequents will be ignored: `c('fileName_X_tst', 'fileName_T_tst')`. It also means one is free to not provide the argument which will then automatically be set to `NULL`.
  - If `fit` is set to `'test'`, the function will complain if you do not supply `'fileName_X_tst'` and `'fileName_T_tst'`.
- `path`
  - It is a string which is assigned the absolute/relative path to the directory that contains the desired data sets. All the files should be contained in the same directory.
  - The `path` variable should end with a forward slash `'/'`.
- `fileName_*_tr`, `fileName_*_tst`
  - String variables which are assigned the respective data set names.
  - The function will automatically concatenate `path` and `fileName_*_tr/fileName_*_tst` to load data sets.

Return values

- Most importantly, it generates a plot of the fitted/predicted values of SSGP and BSAR alongside the true values displayed in points.
- `res_SSGP`: Values returned from SSGP.
- `res_BSAR`: Estimated parameters returned from BSAR.
- `mu_SSGP`: Fitted value of SSGP if the argument `fit` is `'training'`. Predicted value if the argument `fit` is `'test'`.
- `mu_BSAR`: Fitted value of BSAR if the argument `fit` is `'training'`. Predicted value if the argument `fit` is `'test'`.
- `centred_SSGP`: Self-evident.

- `centred_BSAR`: Self-evident.

For an example code, refer to `demo.R`. If you pick `'simul'` for `data`, R will ask you to define a function that you want estimated. Use correct R syntax when defining your own function. For example, `f <- function(x) tan(x) - sin(x * pi)`.

8 Data set review

Note that, in BSAR, the variable that is not in the linear term is selected to have the smallest correlation coefficient with the target variable since it has to accommodate the nonlinear effects.

- As stated in the paper, SSGP has the tendency of overfitting the training data. In comparison, BSAR exhibits more stable performance.
- SSGP has a long execution time. BSAR is very fast.
- BSAR has lower NMSE in comparison to SSGP.