To appear in the *Journal of Statistical Computation and Simulation*
Vol. 00, No. 00, 00000 0000, 1–6

Applying fully Bayesian spline smoothing to estimate yield curves

Jiayu Guo[a] and Xue Mi[b]

[a] *Department of Statistics Rutgers university* ; [b] *Department of Statistics Rutgers University*

(*December 2017*)

Abstract

Recent decades, splines smoothing regression has received more attention and recognition in contemporary non-parametric and semi-parametric regressions. In last century, Wahba (1978)[1] proved that smoothing spline estimate may be viewed as a Bayes estimate with respect to a certain prior on the class of possible response functions. In this work, spline smoothing regression will be stated from a Bayesian perspective. we will use real data to verify the conclusion which was carefully explained by Speckman and Sun(2003)[2], by showing the simulation on smoothing spline estimate that can match the marginal posterior very well.

Keywords: Bayesian Smoothing Spline, Partially Informative Normal Distribution, Monte Carlo Simulation, Ratio of Uniforms Methods Multivariate Normal

1. Introduction

Smoothing splines are always treated as tool for visualizing and analyzing noisy observational data. This method is superior to orthogonal polynomials because it makes no assumptions about the shape of the curve to be used for standardization. We choose smoothing spline as our final project topic because it is interesting to know about its principle both from the perspective of Frequentist and Bayesian. Bayesian analysis treats every parameter as random, assigns prior distribution before collecting the data, so as the posterior distribution is pretty much pragmatic by their incorporation multiple levels of randomness and resultant ability to combine information from a bunch of reasonable sources. Therefore, many authors have exploited the connections between smoothing splines and Bayesian estimation. However, interpretation of smoothing splines from Bayesian prospective is not given until 1970s by Kimeldorf (1971) and Wahba (1978)[3]. They demonstrated that spline smoothing with some certain forms can correspond to Bayesian estimates under a class of improper Gaussian prior distributions on function spaces. Speckman and Sun(2003) fully developed this approach by doing numerical simulations showing that multivariate smoothing splines outperform univariate smoothing splines. They extended earlier work in Sun et al. (1999)[4], stating sufficient conditions were derived for a class of proper priors on the variance components.

In the first part of this paper, we will briefly illustrates the similarities between the penalized likelihood for natural cubic spline smoothing regression and Bayes estimator

*Corresponding author. Email: jiayu.jenny67.guo@gmail.com

by doing formula derivation. The second part is to run Monte Carlo procedure and marginal posterior on the theoretical basis of the first part. By generating datasets from certain sine model and collecting treasury yield curve rates with different maturities(1 month,3 months, 6months, 1 year, 2 years, 3 years,5 years, 7 years,10 years, 20 years and 30 years). We will prove the smoothing parameter's MC simulation graph match its marginal posterior density curve very well, which was carefully explained by Sun(2003).

2. Methods

2.1. *The spline smoothing regression*

The basic idea of spline smoothing regression is the simplest nonparametric regression model with one independent variable t_i and response y_i opp;

$$y_i = f(t_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \delta_0), \qquad i = 1, \cdots, n \tag{1}$$

where a $\leq t_1 < \cdots < t_n \leq b$

Smoothing spline regression (SSR) model with one variable assumes that where $(y_1, \cdots, y_n)'$ are univariate responses; f is an unknown function of an independent variable t and ϵ_i are random errors with $\epsilon_i \overset{iid}{\sim} N(0, \delta_0)$. In order to produce a good fit to the data and avoid rapid local variation, a term called roughness penalty is introduced to measure such kind fluctuation. as shown below:

$$\sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \eta \int_a^b (f^{(p)}(t))^2 \mathrm{d}t \tag{2}$$

The smoothing parameter η controls trade-off between fidelity to the data and the smoothness of the estimation. The larger the value of η, the more the data will be smoothed to produce the curve estimate. we assume $P = 2$, known as cubic smoothing spline.

There is one specific form of cubic smoothing spline. Assume $a \leq t_1 < \ldots < t_n \leq b$, function g is called Natural cubic spline if its second and third derivatives are zero at a and b, so that g is linear on the two extreme intervals $[a, t_1]$and $[t_n, b]$. Based on this feature, we can solve the natural cubic spline smoothing regression, and take the first derivative on the penalized likelihood for natural cubic spline smoothing regression to minimize the least square term in the function (2), by set it to zero, we can get the result shown below and A_1 is a nonnegative definite symmetric matrix with rank n-2.

$$\hat{z} = (I + \eta A_1)^{-1} y \tag{3}$$

2.2. *Bayesian Inference for spline smoothing regression*

It has been shown from Sun(1999) that the priors appropriate for spline smoothing belong to a class we have termed partially informative normal. Let A be a symmetric positive semi-definite matrix having rank $n-p$ and variance δ_1 with the possibly improper density

$$p(z|\delta_1) \propto (\frac{1}{\delta_1})^{\frac{n-p}{2}} \exp(\frac{1}{2\delta_1} z' Az) \tag{4}$$

The model assumed by Sun(2003) are $y = z + \epsilon$, $z = T\theta + x$, $\varepsilon_i \sim N(0, \delta_0 W^{-1})$, $x \sim N_n(0, \delta_1 \sigma)$, $\theta \sim N_p(0, \delta_2 I_p)$, then $A_1 = \lim_{a \to \infty}(aTT' + \Sigma)^{-1}$

Based on the assumption of PIN model, the Bayesian Smoothing Spline model built by Monte Carlo Simulation

2.3. *Monte Carlo Simulation*

Monte Carlo simulations are used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. It is a technique used to understand the impact of risk and uncertainty in prediction and forecasting models.

(1) **using the Ratio-of-Uniforms method sample η from the density**
   - $p(\eta)\frac{\eta^{\frac{n-p}{2}}}{|W+\eta A|^{\frac{1}{2}}[\eta\overline{y}'W(W+\eta A)^{-1}A\overline{y}+SSE]^{\frac{N-p}{2}}}$

(2) **For give η, sample δ_0 from**
   - Inverse Gamma($\frac{N-p}{2}$, $\frac{1}{2}[\eta\overline{y}'W(W+\eta\mathrm{A})^{-1}A\overline{y}+SSE]$)

(3) **For given η and δ_0, sample z from**
   - N($(W+\eta\mathrm{A})^{-1}W\overline{y}$,$\delta_0$(W+$\eta$A)$^{-1}$))

Comparing the value of z in (3), the estimate value for $\widehat{z}$ in Bayesian smoothing spline is $(W+\eta\mathrm{A})^{-1}W\overline{y}$. So we assign $A = A_1$ in the Bayesian smoothing spline model.

2.4. *The Ratio-of-Uniforms method*

The most efficient algorithms for sampling from classical discrete distributions are based upon the acceptance/rejection principle. Since the density function of η is difficult to compute, sampling procedures can be established by using Ratio-of-Uniforms method. The methods is introduced by Kinderman and Monahan(1977)[5]. It is a popular transformation method that can be applied to generate nonuniform random variates, since it results in exact,efficient, fast, and easy-to-implement algorithms. The lemma for the method is defined as below:

lemma 1. Let h be a positive integrable function over χ a subset of $\Re$. Suppose that the variables(u, v) are uniformly distributed over

$$c_h = \{(u, v) : 0 < u \leq [h(\frac{v}{u})]^{\frac{1}{2}}\} \tag{5}$$

Then $x = \frac{v}{u}$ has density $\frac{h(x)}{\int h(x)\mathrm{d}x}$.

for sampling random points (u, v) uniformly distributed in C_h, the Accept-Reject algorithm from a two-dimensional rectangle $\tau = [0, a] \times [b^-, b^+]$is often used, where

$$a = \underset{x}{sup}\sqrt{h(x)}, \tag{6}$$

$$b^- = \underset{x}{inf}x\sqrt{h(x)}, \tag{7}$$

$$b^+ = \underset{x}{sup}x\sqrt{h(x)}, \tag{8}$$

That is $u \sim Unif(a, b)$ and $v \sim Unif(b^-, b^+)$

In this case $h(\eta)$ equals the following forms:

$$p(\eta)\frac{\eta^{\frac{n-p}{2}}}{|W+\eta A|^{\frac{1}{2}}[\eta\overline{y}'W(W+\eta A)^{-1}A\overline{y}+SSE]^{\frac{N-p}{2}}} \quad (9)$$

3. Results

By giving previous priors and using Monte Carlo Simulation which is described in section 2.2, 2.3. We simulated a sample of size 100 from the following model:

$$y = sin(0.2t_i) + \epsilon_i,\ z_i = sin(0.2t_i) \quad \epsilon_i \sim N(0, \delta_0),\ i = 1, \cdots, n \quad (10)$$

3.1. *Fitted Curve*

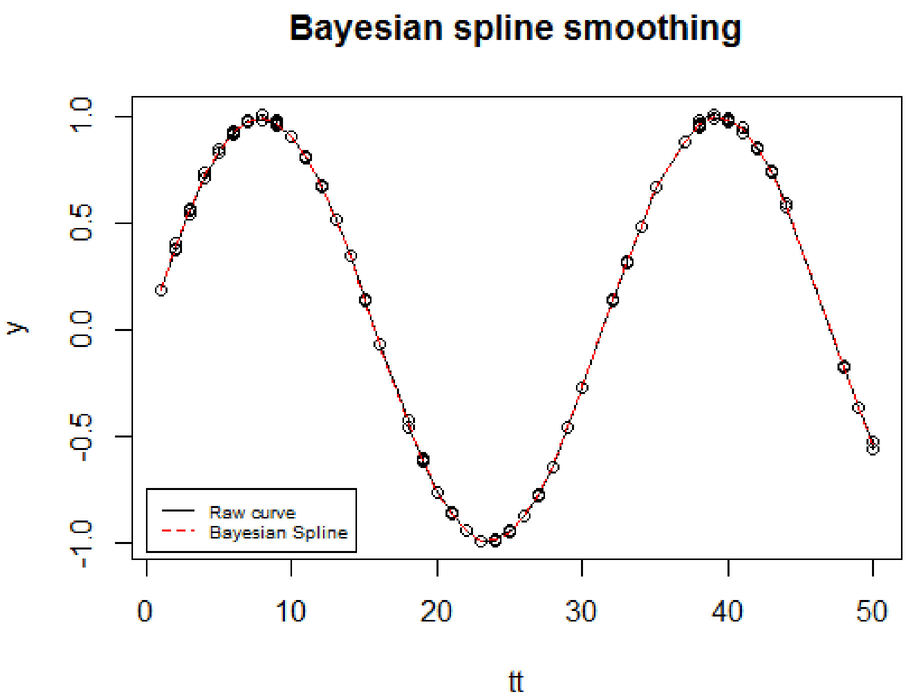

Figure 1. simulated data for $sinx$

The dash red line is simulated by Monte Carlo Methods described in 2.3. The curve drew by Bayesian spline and the raw curve are basically the same. tt is 100 different points random simulate from uniform distribution, y is the corresponding value of $sin(tt)$.

3.2. *Yield Rate*

We applied the Bayesian Smoothing Spline method into the real data. The data set of Daily Treasury Yield Curve Rates is downloaded from the website of U.S. Department of the Treasury on the day 04, Jan, 2016.

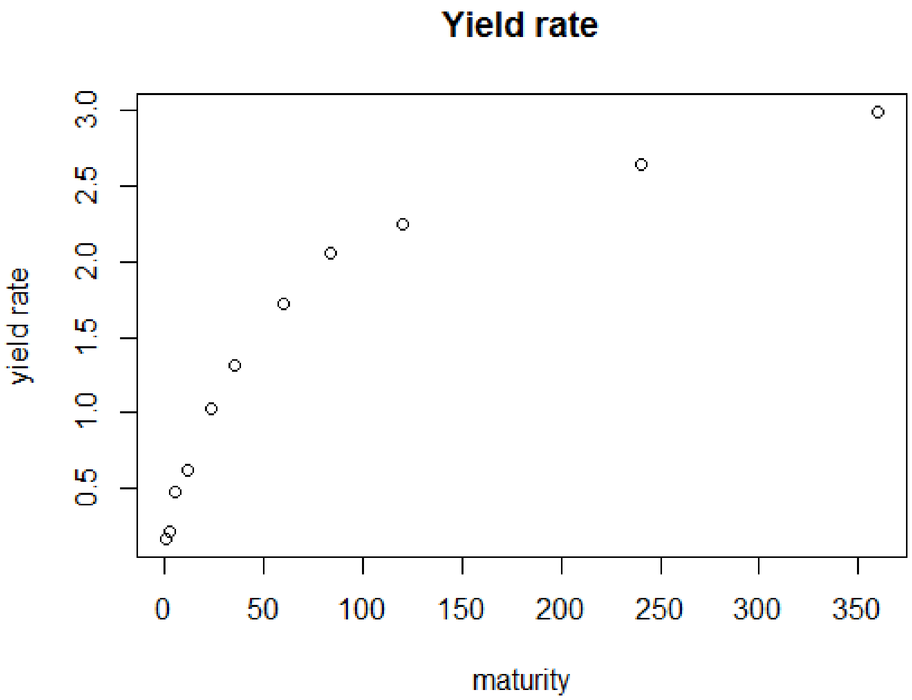

Figure 2. Relationship between maturity and yield rate

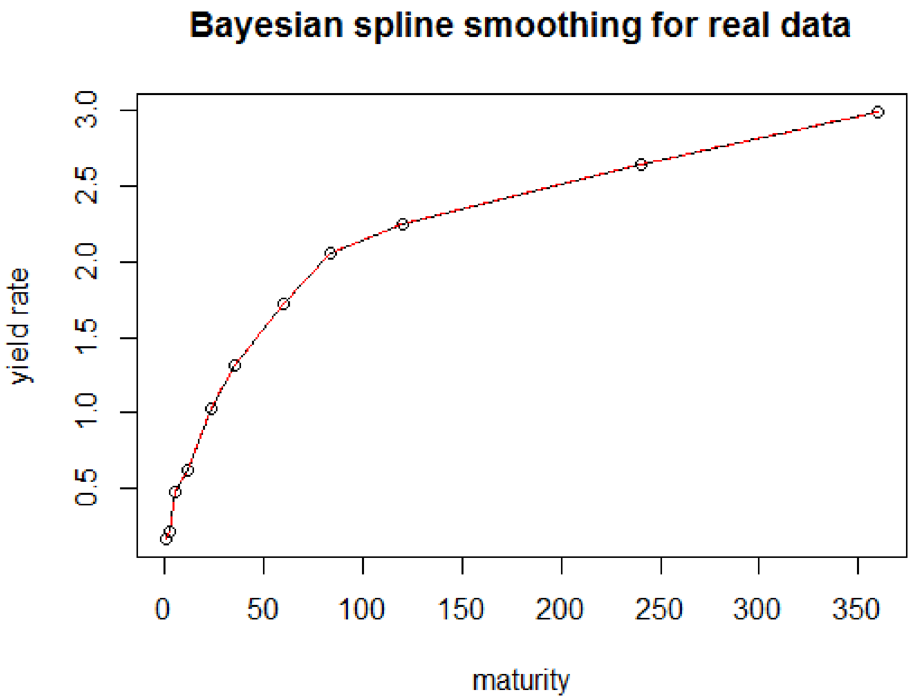

Figure 3. simulated data for real data

The relationship of Maturity and Yield rate showed below in Figure 2.
The smoothing spline curve drew by Bayesian smoothing spline(Figure 3) shows that

Bayesian Smoothing splines fit the simulated and real data very well.

4. Discussion

Through the first part of this work, it been proved that the connection between penalized likelihood for natural cubic spline smoothing regression and Bayes estimator from the theoretical perspective. The second part focuses on the application in the real data. We apply the daily treasury yield curve rates of Jan, 2016 to verify the relationship between Frequentist and Bayesian smoothing spline, resulting a consistent conclusion with the first part.

References

[1] Wahba, G., *Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression*, Journal of the Royal Statistical Society, Series B (Methodological) Vol. 40, No. 3 (1978), pp. 364-372

[2] Paul L. Speckman and Dongchu Sun, *Fully Bayesian Spline Smoothing and Intrinsic Autoregressive Priors*, Biometrika, Vol. 90, No. 2 (Jun., 2003), pp. 289-302

[3] Kimeldorf,G.and Wahba,G., *A correspondence between Bayesian estimation on stochastic processes and smoothing by splines*, Ann.Math.Statist., 41, pp. 495-502

[4] Sun, D., Tsutakawa, R. K., Speckman, P. L., *Posterior distribution of hierarchical models using CAR(1) distributions*, Biometrika, 86, pp. 341-350

[5] Authors:A. J. Kinderman, J. F. Monahan *Computer Generation of Random Variables Using the Ratio of Uniform Deviates*, ACM Transactions on Mathematical Software, Vol. 3 Issue 3,(Sept., 1977), pp. 257-260