Statistica Sinica **34** (2024), 1585-1602
doi:https://doi.org/10.5705/ss.202022.0297

FUNCTIONAL HORSESHOE SMOOTHING FOR FUNCTIONAL TREND ESTIMATION

Tomoya Wakayama* and Shonosuke Sugasawa

The University of Tokyo

Abstract: As a result of developments in instruments and computers, functional observations are becoming increasingly prevalent. However, few existing methodologies can flexibly estimate the underlying trends with valid uncertainty quantification for a sequence of functional data (e.g., functional time series). In this work, we develop a locally adaptive smoothing method, called functional horseshoe smoothing, by introducing a shrinkage prior to the general order of differences of functional variables. This allows us to capture abrupt changes by making the most of the shrinkage capability, and to assess uncertainty by using a Bayesian inference. The fully Bayesian framework allows us to select the number of basis functions using the posterior predictive loss. We provide theoretical properties of the model, which support the shrinkage ability. Furthermore, by taking advantage of the nature of functional data, the proposed method can handle heterogeneously observed data without data augmentation. Simulation studies and a real-data analysis demonstrate that the proposed method has desirable properties.

Key words and phrases: Functional time series, MCMC, shrinkage prior, tail robustness, trend filtering.

1. Introduction

The recent development of measuring instruments and computers has made it possible to obtain high-dimensional data in various fields. However, analyzing such data using a classical multivariate analysis requires a huge number of parameters making it difficult to extract valuable information from the data. A promising methodology to solve these problems is functional data analysis (FDA), which treats and analyzes high-dimensional data as a curve (function). Functional versions for various branches of statistics have been provided; see Ramsay and Silverman (2005), Kokoszka and Reimherr (2017), and Horváth and Kokoszka (2012).

The traditional FDA approach for independent functional data has recently been extended to time series. In fact, for functional time series data, the standard stationary model for multivariate data has been extended (e.g., Besse, Cardot and Stephenson (2000); Klepsch and Klüppelberg (2017); Klepsch, Klüppelberg and Wei (2017); Hörmann, Horváth and Reeder (2013); Gao, Shang and Yang (2019);

*Corresponding author.

Hörmann, Kidziński and Hallin (2015); Martínez-Hernández and Genton (2021)), and its theoretical properties have been studied extensively (e.g., Bosq (2000); Aue and Klepsch (2017); Spangenberg (2013); Aue, Horváth and F. Pellatt (2017); Kühnert (2020); Cerovecki et al. (2019)). However, in actual data, such as GDP data, the assumption of stationarity is often not satisfied because the expected value varies across periods. There are few cases in which the trend can be analyzed appropriately using existing methods.

The stationarity of conventional FDA methods means they cannot capture rapid changes in trend estimation. Wakayama and Sugasawa (2021) solve this difficulty by developing a new type of lasso and proposing functional trend filtering. This new method can capture local changes, while removing observation errors in the data. In other words, the method can clearly identify when structural changes occur in time series data. In order for the inferred results to be used for decision-making, it is essential to evaluate the interpretability of the model and the uncertainty of the estimation. However, few methods for uncertainty evaluation in functional time series analysis have been developed (Petris (2013); Canale and Ruggiero (2016)).

In this work, we propose an approach in a Bayesian framework, hoping to assess the uncertainty and estimate the trend accurately and flexibly, as seen in univariate models (Faulkner and Minin (2018)). In the context of FDA, a shrinkage prior on the functional space is introduced by Shin, Bhattacharya and Johnson (2020). Using a similar idea, we construct a locally adaptive smoother for functional data via the shrinkage prior. Because the priors in the model can be represented as a scale mixture of normals, the model is easy to implement using the Gibbs sampler, and minor extensions make it possible to analyze heterogeneously observed data. Furthermore, in the proposed Bayesian approach, we select the number of basis functions, often done by cross-validation or information-based criteria (e.g., Yao, Müller and Wang (2005); Aue and Klepsch (2017); Tang, Wang and Zhang (2020)), by adopting a posterior predictive loss Gelfand and Ghosh (1998).

We also discuss the theoretical justification for this approach. The essence of this method is that it removes noise, while leaving the change points large. The property of the prior that keeps the signal from shrinking is called "tail robustness". For time series data, analyzing tail robustness is complicated, but here we have shown the properties proposed by Okano et al. (2022). This argument is not limited to functional data, but also justifies a locally adaptive method for finite-dimensional data (Faulkner and Minin (2018); Kakikawa, Shimamura and Kawano (2022)).

The remainder of the paper is structured as follows. Section 2.1 introduces the setting and model for the trend estimation. Section 2.2 gives the posterior computation algorithm. In Section 2.3, we present the way to select the number of basis functions. Section 3 discusses the theoretical properties of the proposed prior and its posterior distribution. In Section 4, we investigate the performance of the proposed method for homogeneously observed data and for heterogeneously observed data. We apply our method to a real data set in Section 5. The contribution of the article is discussed in Section 6. All proofs and detailed posterior densities are given in the online Supplementary Material.

2. Functional Horseshoe Smoothing

2.1. Settings and models

Let $Y_1(\cdot), \ldots, Y_T(\cdot)$ be observed functional data on $\mathcal{S} \subset \mathbb{R}$, ordered as $t = 1, \ldots, T$. Suppose that we are interested in the mean function $Z_t(\cdot) \equiv \mathrm{E}[Y_t(\cdot)]$, which may change smoothly or abruptly over t. To estimate Z_t, we employ the following measurement error model:

$$Y_t(s) = Z_t(s) + \varepsilon_t(s), \quad \varepsilon_t(s) \sim N(0, \sigma^2), \quad t = 1, \ldots, T, \quad s \in \mathcal{S},$$

where $\varepsilon_t(s)$ are error terms, independent over different values of t and s, and σ^2 is an unknown variance. Such measurement models are adopted in the context of Bayesian modeling of functional data (Yang et al. (2016, 2017)).

Let $\phi_1(\cdot), \ldots, \phi_L(\cdot)$ be basis functions on $\mathcal{S}$ (e.g., B-spline function) common over t. We model $Z_t(\cdot)$ as

$$Z_t(\cdot) = \sum_{\ell=1}^{L} b_{t\ell}\phi_\ell(\cdot), \quad t = 1, \ldots, T,$$

where $\boldsymbol{b}_t = (b_{t1}, \ldots, b_{tL})^\top$ is a vector of coefficients, and L is the number of basis functions. Thus, the heterogeneity of the mean function $Z_t(s)$ is characterized by the heterogeneous coefficients, $\boldsymbol{b}_t$. The choice of L controls the smoothness of the estimates of $Z_t(\cdot)$; later we discuss a data-dependent selection of L.

Let $Y_t(s_{t1}), \ldots, Y_t(s_{tn_t})$ be discrete observations, where $s_{t1}, \ldots, s_{tn_t}$ are observation points, and n_t is the number of discrete observations. Note that we allow the number of sampling points and the sampled locations to be heterogeneous over t. Under the settings described above, the model for $\boldsymbol{y}_t = (Y_t(s_{t1}), \ldots, Y_t(s_{tn_t}))^\top$ is

$$\boldsymbol{y}_t = \Phi_t\boldsymbol{b}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \sigma^2 I_{n_t}), \quad t = 1, \ldots, T,$$

where Φ_t is an $n_t \times L$ matrix in which the (i, ℓ)-element is $\phi_\ell(s_{ti})$.

Now, we consider prior distributions on $\boldsymbol{b}_t$. Here, $\otimes$ denotes the Kroneker product. Let Δ_k be the kth order forward difference operators, defined as

$$\Delta_k = \begin{cases} D^{(0)} & \text{for } k = 0, \\ D^{(k)}\Delta_{k-1} & \text{for } k \geq 1, \end{cases}$$

where $D^{(k)}$ is the following $(T-k-1)L \times (T-k)L$ matrix:

$$D^{(k)} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix} \otimes I_L.$$

We then define $(\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_{T-k-1})^\top = \Delta_k(\boldsymbol{b}_1, \dots, \boldsymbol{b}_T)^\top$, and consider a model for this. Although $\boldsymbol{\delta}_t$ depends on k, we write $\boldsymbol{\delta}_t$ rather than $\boldsymbol{\delta}_t^{(k)}$, for simplicity. For example, if $k = 0$, each $\boldsymbol{\delta}_t$ can be written as $\boldsymbol{\delta}_t = \boldsymbol{b}_t - \boldsymbol{b}_{t+1}$, for $t = 1, \dots, T-1$. Hence, if the vector $\boldsymbol{\delta}_t$ is shrunk toward the origin, the two adjacent coefficient vectors $\boldsymbol{b}_{t+1}$ and $\boldsymbol{b}_t$ are identical, leading to the same mean functions for $Z_t(\cdot)$ and $Z_{t+1}(\cdot)$. To encourage such structures, we introduce shrinkage priors on $\boldsymbol{\delta}_t$, which has a large mass around the origin, whereas the tail of the prior is sufficiently large to allow possible abrupt changes over t. Note that under general k, shrinking $\boldsymbol{\delta}_t$ toward the origin can be regarded as smoothing the $k+1$th order derivatives of $Z_t(\cdot)$ with respect to t. We introduce the following hierarchical prior for $\boldsymbol{\delta}_t$:

$$\boldsymbol{\delta}_t | \lambda_t, \tau, \sigma \sim N(\mathbf{0}, \sigma^2\tau^2\lambda_t^2(\Phi_t^\top \Phi_t)^{-1}), \quad \lambda_t \sim C^+(0,1), \quad \tau \sim C^+(0,1), \qquad (2.1)$$

where $C^+(0,1)$ denotes the standard half-Cauchy distribution. The conditional prior of $\boldsymbol{\delta}_t$ has the form of a g-prior (Zellner (1986)), and Shin, Bhattacharya and Johnson (2020) use a hierarchical prior similar to (2.1) in the context of linear regression models. Here, λ_t is a local shrinkage parameter that controls the amount of shrinkage, and λ_t is common to all components of $\boldsymbol{\delta}_t$, so that the vector $\boldsymbol{\delta}_t$ can be simultaneously shrunk toward the origin. Using the half-Cauchy for the local parameter λ_t leads to a multivariate horseshoe-like prior for $\boldsymbol{\delta}_t$. The prior distribution has favorable properties, given in Section 3, and we make the most of them to handle sparsity.

2.2. Posterior computation algorithm

The joint posterior distribution is given by

$$\pi(\sigma^2)\pi(\tau^2)\prod_{t=1}^{T} p(\boldsymbol{y}_t; \Phi_t \boldsymbol{b}_t, \sigma^2 I_n) \prod_{t=1}^{T-k-1} p(\boldsymbol{\delta}_t; \mathbf{0}, \sigma^2\tau^2\lambda_t^2(\Phi_t^\top \Phi_t)^{-1})\pi(\lambda_t),$$

where $\pi(\sigma^2)$, $\pi(\tau^2)$, and $\pi(\lambda_t)$ are prior distributions for σ^2, τ^2, and λ_t^2, respectively. The priors of τ^2 and λ_t^2 are defined in (2.1), and we use the conjugate prior $\sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma)$, which denotes an inverse gamma distribution with shape parameter a_σ and scale parameter b_σ. Using the data augmentation technique of the horseshoe prior (e.g., Makalic and Schmidt (2015)), we can sample from the joint posterior using a simple Gibbs sampling, described as follows:

- Sampling from σ^2: The full conditional distribution of σ^2 is $\mathrm{IG}(\widetilde{a}_\sigma, \widetilde{b}_\sigma)$, where

$$\widetilde{a}_\sigma = a_\sigma + \frac{1}{2}L(T-k-1) + \frac{1}{2}nT,$$
$$\widetilde{b}_\sigma = b_\sigma + \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t - \Phi_t\boldsymbol{b}_t)^\top(\boldsymbol{y}_t - \Phi_t\boldsymbol{b}_t) + \frac{1}{2\tau^2}\sum_{t=1}^{T-k-1}\frac{\boldsymbol{\delta}_t^\top\Phi_t^\top\Phi_t\boldsymbol{\delta}_t}{\lambda_t^2}.$$

- Sampling from τ^2: The full conditional distribution of τ^2 is

$$\mathrm{IG}\left(\frac{L(T-k-1)+1}{2}, \frac{1}{\xi} + \sum_{t=1}^{T-k-1}\frac{\boldsymbol{\delta}_t^\top\Phi_t^\top\Phi_t\boldsymbol{\delta}_t}{2\sigma^2\lambda_t^2}\right),$$

  where ξ is an auxiliary variable with full conditional distribution IG(1, $1 + 1/\tau^2$).

- Sampling from λ_t^2: The full conditional distribution of λ_t^2 is

$$\mathrm{IG}\left(\frac{L+1}{2}, \frac{1}{\nu_t} + \frac{\boldsymbol{\delta}_t^\top\Phi_t^\top\Phi_t\boldsymbol{\delta}_t}{2\tau^2\sigma^2}\right),$$

  where ν_t is an auxiliary variable with full conditional distribution IG(1, $1 + 1/\lambda_t^2$).

- Sampling from $\boldsymbol{b}_t$: The full conditional distribution of $\boldsymbol{b}_t$ is of the form $N(\boldsymbol{\mu}_t, c_t(\Phi_t^\top\Phi_t)^{-1})$, where the specific forms of $\boldsymbol{\mu}_t$ and c_t are dependent on k, the order of difference. Detailed expressions under $k = 0$ and $k = 1$ are provided in the Supplementary Material.

As shown above, the full conditional distributions are all familiar forms, allowing us to compute the posterior computation efficiently. Given the posterior samples of $\boldsymbol{b}_t$, we can generate posterior samples of $Z_t(s)$ at an arbitrary location $s \in \mathcal{S}$, which gives a point estimate (e.g., posterior mean) and an interval estimation (e.g., 95% credible interval).

2.3. Selection of the number of basis functions

In practice, specifying the number of basis functions, L, is an important task. If L is smaller than necessary, the basis function approximation gives over-smoothed results. On the other hand, the estimation results can be inefficient if L is larger than necessary. We suggest adopting a model selection criterion to select L in a data-dependent manner. Here we use the posterior predictive loss (PPL) proposed by Gelfand and Ghosh (1998).

To clarify the number of bases used in the estimation, we write $\boldsymbol{b}_t(L)$ and $\Phi_t(L)$, rather than $\boldsymbol{b}_t$ and Φ_t, respectively. Given $\boldsymbol{b}_t(L)$, the conditional distribution of $\boldsymbol{y}_t$ is $N\left(\Phi_t(L)\boldsymbol{b}_t(L), \sigma^2 I_n\right)$. We then define the PPL as

$$\begin{aligned}\mathrm{PPL}(L) = \frac{T}{T+1}\sum_{t=1}^{T}\left\{\boldsymbol{y}_t - \Phi_t(L)\mathrm{E}_p[\boldsymbol{b}_t(L)]\right\}^\top\left\{\boldsymbol{y}_t - \Phi_t(L)\mathrm{E}_p[\boldsymbol{b}_t(L)]\right\}\\ + nT\mathrm{E}_p[\sigma^2] + \sum_{t=1}^{T}\mathrm{tr}(\Phi_t(L)\mathrm{Cov}_p(\boldsymbol{b}_t(L))\Phi_t(L)^\top),\end{aligned}$$

where E_p and Cov_p are the expectation and covariance with respect to the posterior distribution. We choose the number of basis functions by minimizing the criterion $\mathrm{PPL}(L)$. The order of differences, k, can also be selected using the PPL.

2.4. Extension to irregular grids

We here consider an extension to the proposed smoothing techniques under irregularly spaced functional data. Let $Z_{t_1}(\cdot), Z_{t_2}(\cdot), \ldots, Z_{t_n}(\cdot)$ be a sequence of functional random variables indexed by t_i. This situation requires that distance information be incorporated into the model. Using the same basis expansion in Section 2.1, we introduce the following prior:

$$\boldsymbol{b}_{t_i+h} - \boldsymbol{b}_{t_i}|\lambda_{t_i}, \tau, \sigma \sim N(\mathbf{0}, h\sigma^2\tau^2\lambda_{t_i}^2(\Phi_{t_i}^\top\Phi_{t_i})^{-1}), \quad h \geq 0, \tag{2.2}$$

where we use the same prior for λ_t^2. Although the prior formulation (2.2) corresponds to the extension under the first-order difference, second-order cases can be extended to the irregular grids, following Lindgren and Rue (2008).

3. Theoretical properties of the model

This section presents the theoretical properties of the prior distribution and its periphery. The proofs are provided in the Supplementary Material.

In Section 2.1, we formulated the prior as (2.1). Here we investigate this in greater detail. The marginal prior of λ_t is

$$\pi(\boldsymbol{\delta}_t \mid \tau, \sigma) \propto \int_0^\infty \frac{1}{\lambda_t^L(1+\lambda_t^2)}\exp\left\{-\frac{1}{2\sigma^2\tau^2\lambda_t^2}\boldsymbol{\delta}_t^\top\Phi_t^\top\Phi_t\boldsymbol{\delta}_t\right\}d\lambda_t.$$

Then, notable properties of the marginal prior are given by the following proposition.

Proposition 1.

(i) $\pi(\boldsymbol{\delta}_t \mid \tau, \sigma) \to \infty$ *as* $\boldsymbol{\delta}_t \to 0$.

(ii) $\pi(\boldsymbol{\delta}_t \mid \tau, \sigma) = O(\|\Phi_t\boldsymbol{\delta}_t\|_2^{-L-1})$

Here, (i) implies that, for given τ, σ^2, and L, the density diverges at the origin $\boldsymbol{\delta}_t = \mathbf{0}$, like the original horseshoe prior (Carvalho, Polson and Scott (2009, 2010)). Conspicuously, this property strongly shrinks trivial noise toward zero at the posterior inference. On the other hand, (*ii*) suggests that the tail decay of the marginal prior is slow. The random variables from the prior are expected to take large values with greater probability, owing to the heavy tail. These critical features of the prior distribution contribute to handling sparsity.

Next, we consider the posterior mean deduced from the prior. For simplicity, we focus on $k = 0$. In this case, the model can be rewritten as

$$\boldsymbol{z}_t \equiv \boldsymbol{y}_{t+1} - \boldsymbol{y}_t \sim N(\Phi_t \boldsymbol{\delta}_t, 2\sigma^2 I_n), \quad \boldsymbol{\delta}_t \equiv \boldsymbol{b}_{t+1} - \boldsymbol{b}_t \sim N(\mathbf{0}, \sigma^2\lambda_t^2\tau^2(\Phi_t^\top\Phi_t)^{-1}),$$

so that the model is defined for the observed value of difference $\boldsymbol{z}_t$.

Proposition 2. *The posterior mean of the model is weakly tail robust, that is,*

$$\frac{|\mathrm{E}[\Phi_{t^*}\boldsymbol{\delta}_{t^*}|\boldsymbol{z}] - \boldsymbol{z}_{t^*}|}{\|\boldsymbol{z}_{t^*}\|_2} \to \mathbf{0} \quad \text{as} \quad \|\boldsymbol{z}_{t^*}\|_2 \to \infty \quad \text{for any} \quad t^* \in \{1, \ldots, T-1\}.$$

This claim implies that the difference between the posterior expectation and the original observation is relatively subtle when $\|\boldsymbol{z}_{t^*}\|_2$ is large. This property is weaker than the tail robustness (Carvalho, Polson and Scott (2010)). In our setting, the dependencies between the data make it challenging to analyze the tail robustness. Nevertheless, weak tail robustness still holds, implying that the signal is preserved in the posterior analysis without shrinkage. This property is derived from the fact that the prior has considerable mass on the tail. Using $C^+(0,1)$ for λ_t is motivated by this argument.

The tail robust-related properties of time series have not been determined for ordinary multivariate analysis or in functional data. The result of this theorem also applies to an ordinary multivariate analysis if we ignore Φ_t, which is important in the context of shrinkage estimation.

4. Simulation Studies

4.1. Simulation settings

We evaluate the performance of the proposed and existing methods using simulation studies. For $t = 1, \ldots, T(= 50)$ and the domain $\mathcal{X} = [1, n]$, with $n = 120$, we have the following four scenarios as the true trend function $\beta_t(x)$:

(1) Constant: $\beta_t(x) = f_1(x)$,

(2) Smooth: $beta_t(x) = f_1(x)\sin((t+x)/5)$,

(3) Piecewise constant: $\beta_t(x) = \sum_{i=1}^5 f_i(x)\mathbb{I}_{\{10(i-1)<t\leq 10i\}}$,

(4) Varying smoothness: $\beta_t(x) = f_1(x) + 20\{\sin((4t/n)-2) + 2\exp(-30((4t/n)-2)^2))\}$,

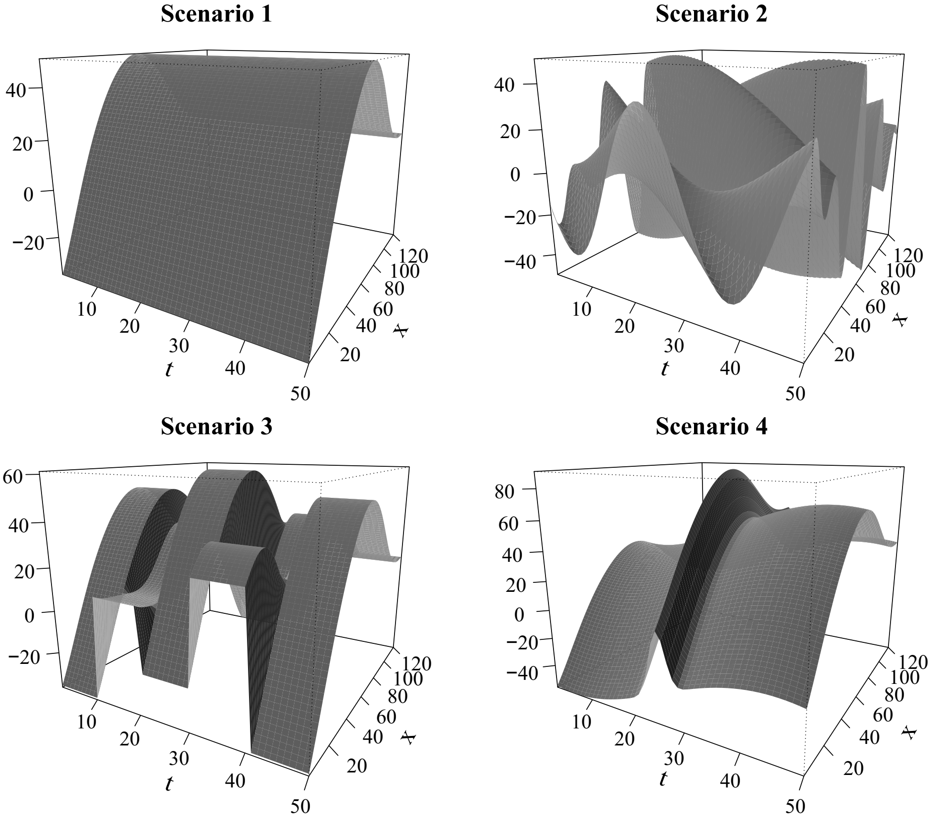

Figure 1. Each surface represents a three-dimensional plot of the true trend.

where f_i $(i = 1, \ldots, 5)$ is a sample path of the Gaussian process associated with the radial basis function (RBF) kernel $k_i(x_1, x_2) = \theta_i^2 \exp(-\|x_1 - x_2\|^2/2\theta_i^2)$, with a hyperparameter θ_i. We set $\theta_i = 30, 20, 35, 25, 30$, for $i = 1, \ldots, 5$.

The observed data are generated by adding noise from $N(0, 5^2)$ to the trend functions at equally spaced $H = 120$ points of x, namely, $x \in \{1, \ldots, H\}$. Figure 1 shows how the trends change over time.

In scenario 1, we investigate whether the proposed methods discover that the trend is constant over time, even in the presence of noise. Scenario 2 checks the ability of the methods to extract a continuous curve from noisy data. Scenario 3 reveals the ability of the methods to detect abrupt changes between intermittent straight horizons, that is, discontinuity points. In scenario 4, we examine whether the methods can capture periods when the smoothness of the trend changes significantly.

4.2. Homogeneously observed data

We first use the full data set generated using the method presented in the previous section, which we call *homogeneously observed data*. For the simulated data, we apply the following methods:

- FHS: functional horseshoe smoothing.
- FLS: an artificial alternative method using a Laplace-like prior, that is, $\lambda_t^2 \sim \text{Exp}(1)$.
- B-spline: a curve-fitting method using a B-spline function of order four and uniform knots. This method is implemented using "fda.usc" package (Febrero-Bande and de la Fuente (2012)). Note that the obtained basis functions are used in FHS and FLS.
- FTF: functional trend filtering developed by Wakayama and Sugasawa (2021).
- BART: Bayesian additive regression trees, as developed by Chipman, George and McCulloch (2010).

The motivation for using the FLS method is to address the importance of the half-Cauchy prior for the local parameter λ_t, as discussed in Section 2. The purpose of using BART is to compare FHS with existing flexible methods. In fact, this time-dependent functional analysis can be reframed as a bivariate (t and x are explanatory variables) regression problem, to which BART can be applied. In addition, to compare FHS with a locally adaptive frequentist method, we implement FTF. The comparison with the B-spline is to determine how much better the FHS and FLS are when compared with the case without smoothing.

For the Bayesian methods, we use 3,000 posterior draws, after discarding 3,000 burn-in samples. For the FHS and FLS, we select the optimal number of the basis functions, L, and the order of difference, k, using the PPL criterion from among candidates $L \in \{5, 9, 13, 17, 21, 25\}$ and $k \in \{0, 1\}$, respectively.

To evaluate the point estimates, we adopt the following criterion:

- Mean absolute deviation (MAD): difference between the posterior medians and the true values, defined as

$$\text{MAD} = \frac{1}{HT} \sum_{x=1}^{H} \sum_{t=1}^{T} \big| \widehat{\beta}_t(x) - \beta_t(x) \big|.$$

Moreover, we use the following two criteria to evaluate the 95% credible intervals obtained using the Bayesian methods:

- Mean credible interval width (MCIW): the width of intervals, defined as

$$\text{MCIW} = \frac{1}{HT} \sum_{x=1}^{H} \sum_{t=1}^{T} \widehat{\beta}_t^{97.5}(x) - \widehat{\beta}_t^{2.5}(x),$$

Table 1. The averaged values of MAD (mean absolute deviation), MCIW (mean credible interval width), CP (coverage probability of credible interval), and number L of bases for FHS, FLS, Spline1 (B-spline estimator with the same basis as FHS), Spline2 (B-spline estimator with the same basis as FLS), BART (Bayesian additive regression trees) and FTF (functional trend filtering) for scenario 1 and scenario 2.

Scenario	Method	MAD	MCIW	CP(%)	L
1	**FHS**	0.538	3.321	98.1	17.9
	FLS	0.673	4.374	99.0	24.9
	Spline1	1.495	-	-	17.9
	Spline2	1.127	-	-	24.9
	BART	0.656	3.187	93.4	-
	FTF	0.490	-	-	-
2	**FHS**	0.961	5.214	96.6	23.4
	FLS	0.940	4.590	94.1	24.9
	Spline1	1.739	-	-	23.4
	Spline2	1.333	-	-	24.9
	BART	2.539	9.122	83.0	-
	FTF	1.700	-	-	-

where $\widehat{\beta}_t^{97.5}(x)$ and $\widehat{\beta}_t^{2.5}(x)$ correspond to the 97.5 and 2.5 percentiles, respectively, of the posterior distribution for $\beta_t(x)$.

- Coverage probability (CP): the coverage accuracy of the credible interval, defined as

$$\mathrm{CP}(\%) = \frac{100}{HT}\sum_{x=1}^{H}\sum_{t=1}^{T}\mathbb{I}_{\{\widehat{\beta}_t^{97.5}(x)>\beta_t(x)>\widehat{\beta}_t^{2.5}(x)\}}.$$

We repeated the simulations 150 times; the averages across the simulations are presented in Tables 1 and 2.

Overall, the results in the tables show that FHS outperforms the other method. Specifically, we have the following results:

Scenario 1: In general, frequentist shrinkage methods have a stronger ability to shrink estimators to zero than Bayesian methods do, and hence FTF yields the best results. However, the difference between it and FHS is subtle, and FHS can also shrink the estimators significantly.

Scenario 2: The generated functional data have a complex and ever-changing signal, but FHS and FLS capture it well, owing to their flexibility. Here, we found that the proposed methods fitted successfully for smoothly transitioning functional time series data. However, when abrupt changes in the data do not exist, the curve is estimated better by FLS.

Scenario 3: Here, the distinction between FHS and the other methods becomes evident. FHS estimates the trend almost horizontally, where the amount of change is zero, and made large changes only at the discontinuous points. In contrast, FLS was less able to make sparse estimates, resulting in a gently curved estimate. This suggests that choosing an appropriate prior is essential.

Scenario 4: The BART and FHS results eclipse those of the other methods. In fact, FHS is inferior to BART, but still guarantees the flexibility to capture abrupt changes.

Thus, FHS performs favorably, and any slight differences between it and other methods should not compromise its usefulness. With respect to CP, FHS always achieves a value close to 95, indicating that it provides a better inference than other Bayesian methods do in terms of uncertainty evaluation. Moreover, the stability of the estimation accuracy and coverage of FHS compared with that of BART, a mere nonparametric method, indicates that FHS is more successful in analyzing it as functional time series data.

Next, we investigate how the credible intervals change with the sample size. We change the number H of data at each time from 120 to 60, and compare the results with those of the original settings with respect to MCIW and CP. From Table 3, obtaining additional data narrows the range of CI even though CP remains almost the same, which is consistent with the fact that uncertainty decreases with more data. This suggests that a trend estimation of functional data should consider the number of functions and the number of observation points for each function.

To investigate whether the choice of a basis is meaningful, that is, the advantage of selecting a basis rather than preparing many basis functions, we fix the number of basis functions at 25 and implement our method. The results are shown in Table 4, indicating that the accuracy of the point estimation improves by adaptively choosing the number of basis functions using PPL. Furthermore, the performance of the interval estimation also improves, because MCIW gets smaller when we select the number of basis functions, while preserving the CP values.

4.3. Heterogeneously observed data

Here, We examine cases in which 5% and 10% of the data are omitted at random. We report the results in Table 5, based on 3,000 MCMC iterations, obtained after a burn-in period of 3,000 iterations. As the percentage of omitted data increases, the data become more unequally spaced, and the estimation becomes less precise, although it is still able to capture the trend accurately. The wide MCIW implies an increase in uncertainty due to the omitted data. In addition, it should be challenging to estimate the mean function when some points

Table 2. The averaged values of MAD (mean absolute deviation), MCIW (mean credible interval width), CP (coverage probability of credible interval), and number L of bases for FHS, FLS, Spline1 (B-spline estimator with the same basis as FHS), Spline2 (B-spline estimator with the same basis as FLS), BART (Bayesian additive regression trees), and FTF (functional trend filtering) for scenario 3 and scenario 4.

Scenario	Method	MAD	MCIW	CP(%)	L
3	**FHS**	0.713	4.350	98.2	21.5
	FLS	3.701	6.307	59.2	9.24
	Spline1	1.560	-	-	21.5
	Spline2	4.045	-	-	9.24
	BART	1.217	5.061	89.1	-
	FTF	2.245	-	-	-
4	**FHS**	0.814	3.988	93.7	19.0
	FLS	1.523	4.276	80.7	10.1
	Spline1	1.542	-	-	19.0
	Spline2	1.868	-	-	10.1
	BART	0.711	3.210	92.6	-
	FTF	1.213	-	-	-

Table 3. MCIW (mean credible interval width), MASVD (mean absolute sequentially variational deviation), and CP (coverage probability of credible interval) of FHS with a horseshoe prior for $H = 120$ and $H = 60$.

Scenario	H	MCIW	CP(%)
1	120	3.321	98.1
	60	4.879	98.5
2	120	5.214	96.6
	60	6.513	96.4
3	120	4.350	98.2
	60	6.234	97.9
4	120	3.988	93.7
	60	5.389	94.0

are missing, but Figure 2 suggests that the estimator detects the trend. Thus, heterogeneity in both the number of sampling points and the sampled locations does not make implementation challenging, nor does it have a significant negative effect on accuracy.

5. Empirical Application

Many studies have demonstrated the applicability and performance of functional time series analysis methods using age-specific fertility data (e.g., Hyndman and Ullah (2007); Wakayama and Sugasawa (2021)). This section presents an empirical application of the proposed method using annual age-specific

Table 4. MAD (mean absolute deviation), MCIW (mean credible interval width), and CP (coverage probability of credible interval) for FHS (the estimator using the horseshoe-like prior) with different number of bases for each scenario.

Scenario	Method	L	MAD	MCIW	CP(%)
1	selected	17.9	0.538	3.321	98.1
	fixed	25.0	0.571	3.658	98.6
2	selected	23.4	0.961	5.214	96.6
	fixed	25.0	0.981	5.362	96.7
3	selected	21.5	0.713	4.350	98.2
	fixed	25.0	0.749	4.572	98.2
4	selected	19.0	0.814	3.988	93.7
	fixed	25.0	0.896	4.361	93.5

Table 5. MAD (mean absolute deviation), MCIW (mean credible interval width), and CP (coverage probability of credible interval) for FHS (the estimator based on the horseshoe-like prior) for each scenario under heterogeneously observed points.

Scenario	omitted rate	MAD	MCIW	CP
1	0%	0.538	3.321	98.1
	5%	0.674	4.335	98.5
	10%	0.727	4.622	98.5
2	0%	0.961	5.214	96.6
	5%	1.017	5.599	96.8
	10%	1.066	5.822	96.8
3	0%	0.713	4.350	98.2
	5%	0.821	5.056	98.2
	10%	0.862	5.277	98.0
4	0%	0.814	3.988	93.7
	5%	0.889	4.784	96.4
	10%	0.889	4.806	96.4

Australian fertility rates, obtained from the Australian Bureau of Statistics, defined as the number of births per 1,000 female residents. These data cover the age group 15 to 49 and the period 1921 to 2015. Then, we then consider that there are 95 functions with the domain $[15, 49]$. Our interest is the transition of the functions over time.

We apply FHS and its Laplace prior version (FLS). The numbers of bases for FHS and FLS are 27 and 20, respectively, chosen using PPL from $\{5, 6, \ldots, 30\}$. Furthermore, the difference order k is selected as one using PPL. The observation is shown in the upper left part of Figure 3, and its surface is rugged. FLS smoothed the surface and largely removed the noise. This is also the case for FHS, where the surface is smooth and the denoising effect is confirmed. The difference is that FHS left the sharp edges intact, whereas FLS erased the sharp

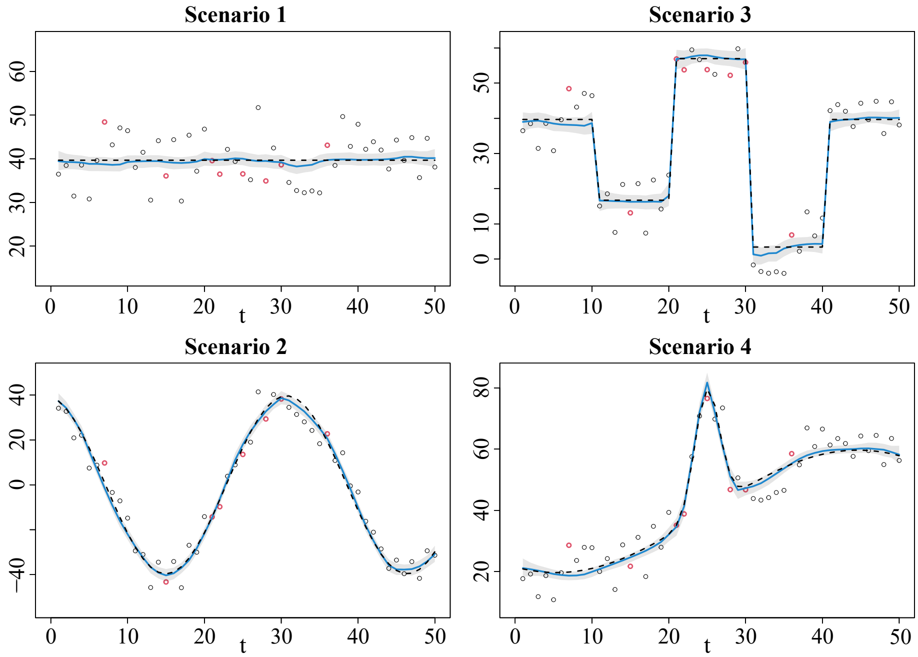

Figure 2. Each panel is a cross-section at $x = 40$. The dotted line is the true trend, the blue line is the estimated trend, the gray area shows the 95% credible intervals at $x = 40$, and red dots indicate omitted measurements. The data acquisition interval is uneven, and some values are missing.

edges implying that the latter reduces the signal and the noise, and hides the change points. Not doing so is a strength of FHS.

We next focus on the CI. Figure 4 shows the difference between the 97.5 percentile and the 2.5 percentile for each year. This is a three-dimensional representation of the size of the 95% credible region. Here, we find that FHS has a smaller credible area than FLS, and thus regard FHS as a more plausible model.

6. Conclusion

We have presented a Bayesian nonparametric smoothing method for functional time series data. This enables a locally adaptive estimation by exploiting the sparsity from the shrinkage prior distributions. The result of our simulation studies and empirical applications suggest that the proposed method performs well, especially with a horseshoe-like prior, even with the presence of sharp changes.

Moreover, we have elucidated the theoretical properties of the proposed method. We discussed two significant issues with the shrinkage prior distribution. The first is the spike at the origin of the marginal prior, and the second is

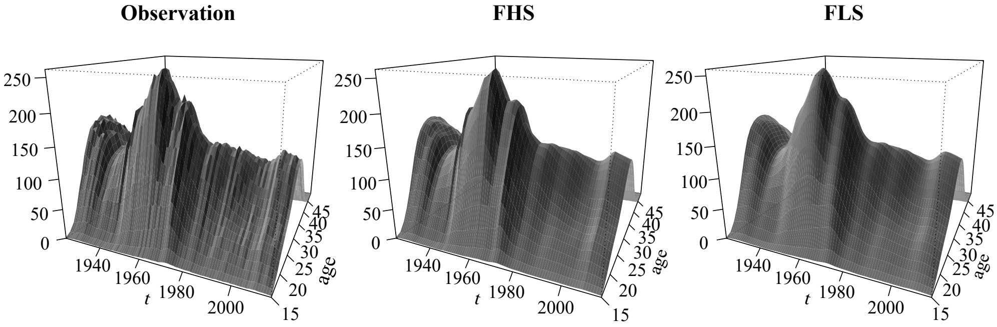

Figure 3. The number of births per 1,000 female residents by age in each year in Australia. The left is the observed quantity, and the middle (right) is the smoothed surface using FHS (FLS).

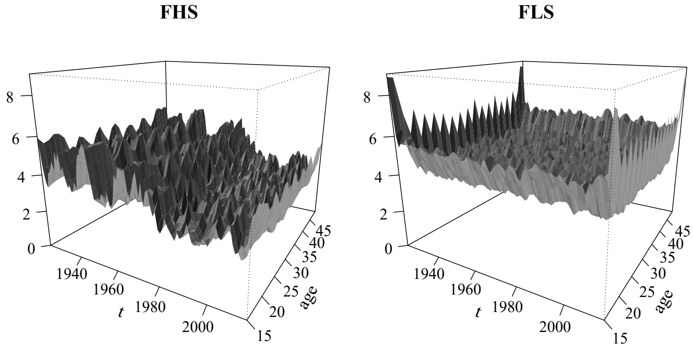

Figure 4. The left (right) figure shows the credible interval width of FHS (FLS) in three dimensions.

the thickness of its tail. Because of this, we expect the estimator to be very good at eliminating noise and detecting abrupt changes simultaneously. We further checked the latter by proving the weak tail robustness of the posterior expectation. These are the theoretical reasons why using a horseshoe-like prior is favorable.

FHS has two advantages over FTF (Wakayama and Sugasawa (2021)). One is that selecting the parameters is easy. In the optimization approach, we need to select the tuning parameter (penalty parameter) using K-fold cross-validation, and the computational complexity increases with the number of folds and the number of candidate parameters. The time and effort required to select the three parameters k (the order of difference), L (the number of basis functions), and λ (the penalty parameter) is enormous. In our Bayesian approach, penalty parameters such as local parameters and global parameter are selected automatically using MCMC. The parameter selection is easier than that of the frequentist method, because it does not require selecting penalty parameters, and the PPL (criterion of model) can be calculated. In addition, Bayesian models can estimate a trend more accurately than existing methods do including FTF, especially when flexible shrinkage is required. This is because, as noted by Polson and Scott (2011) and Carvalho, Polson and Scott (2010), by using local and global parameters, the horseshoe prior can shrink each part of the estimation to a different degree. Hence, FHS allows the smooth and sharp parts of the estimation to coexist.

Furthermore, FHS can deal with heterogeneously observed data. From simulation studies, FHS can capture the trend accurately, even when some data are randomly omitted. There are two critical reasons why trends can be estimated accurately without completing missing values. The first is that even though there is little information at each time (some data are missing), accumulating information from all functions yields, a lot of information. The second is that when estimating trends, one can borrow information from the adjacent time. This is a key advantage of functional data analysis.

Our model is also useful for estimating the varying-coefficient functional linear model (VCFLM) (Matsui (2022); Wu, Fan and Müller (2010); Cardot and Sarda (2008)). The VCFLM is a combination of a scalar-on-function model and a varying-coefficient model, where the coefficients are functions that depend on exogenous variables. By expanding the predictor function and the coefficient function in an orthonormal basis, and introducing our prior into the coefficients of the basis expansion, the VCFLM can be analyzed within the FHS framework.

Supplementary Material

The Supplementary Material contains detailed forms of full conditional distributions of our method and proofs of propositions.

Acknowledgments

We thank the two anonymous referees for their helpful comments. This research was supported, in part, by JSPS KAKENHI Grant Numbers 18H03628, 20H00080, and 21H00699 from the Japan Society for the Promotion of Science.

References

Aue, A., Horváth, L. and F. Pellatt, D. (2017). Functional generalized autoregressive conditional heteroskedasticity. *Journal of Time Series Analysis* **38**, 3–21.

Aue, A. and Klepsch, J. (2017). Estimating functional time series by moving average model fitting. *arXiv:1701.00770.*

Besse, P. C., Cardot, H. and Stephenson, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics* **27**, 673–687.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications.* Springer Science & Business Media.

Canale, A. and Ruggiero, M. (2016). Bayesian nonparametric forecasting of monotonic functional time series. *Electronic Journal of Statistics* **10**, 3265–3286.

Cardot, H. and Sarda, P. (2008). Varying-coefficient functional linear regression models. *Communications in Statistics—Theory and Methods* **37**, 3186–3203.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, 73–80. PMLR.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

Cerovecki, C., Francq, C., Hörmann, S. and Zakoian, J.-M. (2019). Functional GARCH models: The quasi-likelihood approach and its applications. *Journal of Econometrics* **209**, 353–375.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics* **4**, 266–298.

Faulkner, J. R. and Minin, V. N. (2018). Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Analysis* **13**, 225–252.

Febrero-Bande, M. and de la Fuente, M. O. (2012). Statistical computing in functional data analysis: The R package fda. usc. *Journal of statistical Software* **51**, 1–28.

Gao, Y., Shang, H. L. and Yang, Y. (2019). High-dimensional functional time series forecasting: An application to age-specific mortality rates. *Journal of Multivariate Analysis* **170**, 232–243.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

Hörmann, S., Horváth, L. and Reeder, R. (2013). A functional version of the ARCH model. *Econometric Theory* **29**, 267–288.

Hörmann, S., Kidziński, Ł. and Hallin, M. (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)* **77**, 319–348.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications.* Springer Science & Business Media.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis* **51**, 4942–4956.

Kakikawa, Y., Shimamura, K. and Kawano, S. (2022). Bayesian fused Lasso modeling via horseshoe prior. *arXiv:2201.08053.*

Klepsch, J. and Klüppelberg, C. (2017). An innovations algorithm for the prediction of functional linear processes. *Journal of Multivariate Analysis* **155**, 252–271.

Klepsch, J., Klüppelberg, C. and Wei, T. (2017). Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics* **1**, 128–149.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis.* CRC Press.

Kühnert, S. (2020). Functional ARCH and GARCH models: A yule-walker approach. *Electronic Journal of Statistics* **14**, 4321–4360.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics* **35**, 691–700.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters* **23**, 179–182.

Martínez-Hernández, I. and Genton, M. G. (2021). Nonparametric trend estimation in functional time series with application to annual mortality rates. *Biometrics* **77**, 866–878.

Matsui, H. (2022). Truncated estimation for varying-coefficient functional linear model. *arXiv:2203.10268.*

Okano, R., Hamura, Y., Irie, K. and Sugasawa, S. (2022). Locally adaptive Bayesian isotonic regression using half shrinkage priors. *arXiv:2208.05121.*

Petris, G. (2013). A Bayesian framework for functional time series analysis. *arXiv:1311.0098.*

Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics* **9**, 501–538. Oxford University Press, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis.* 2nd Edition. Springer Science+Business Media, Inc., New York.

Shin, M., Bhattacharya, A. and Johnson, V. E. (2020). Functional horseshoe priors for subspace shrinkage. *Journal of the American Statistical Association* **115**, 1784–1797.

Spangenberg, F. (2013). Strictly stationary solutions of ARMA equations in banach spaces. *Journal of Multivariate Analysis* **121**, 127–138.

Tang, C., Wang, T. and Zhang, P. (2020). Functional data analysis: An application to covid-19 data in the united states. *arXiv:2009.08363.*

Wakayama, T. and Sugasawa, S. (2021). Trend filtering for functional data. *arXiv:2104.02456.*

Wu, Y., Fan, J. and Müller, H.-G. (2010). Varying-coefficient functional linear regression. *Bernoulli* **16**, 730–758.

Yang, J., Cox, D. D., Lee, J. S., Ren, P. and Choi, T. (2017). Efficient Bayesian hierarchical functional data analysis with basis function approximations using Gaussian–Wishart processes. *Biometrics* **73**, 1082–1091.

Yang, J., Zhu, H., Choi, T. and Cox, D. D. (2016). Smoothing and mean–covariance estimation of functional data with a Bayesian hierarchical model. *Bayesian Analysis* **11**, 649–670.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical Association* **100**, 577–590.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*, 233–243. Elsevier Science Publishers, New York.

Tomoya Wakayama

Department of Statistics, The University of Tokyo, Tokyo 113-0033, Japan.

E-mail: tomow.9.12@gmail.com

Shonosuke Sugasawa

Center for Spatial Information Science, The University of Tokyo, Kashiwa, Chiba 277-8568, Japan.

E-mail: sugasawa@csis.u-tokyo.ac.jp

(Received August 2022; accepted November 2022)