AIMS Mathematics, 7(10): 19202–19220.
DOI: 10.3934/math.20221054
Received: 26 March 2022
Revised: 18 August 2022
Accepted: 26 August 2022
Published: 30 August 2022

http://www.aimspress.com/journal/Math

Research article

On smoothing of data using Sobolev polynomials

Rolly Czar Joseph Castillo and Renier Mendoza*

Institute of Mathematics, University of the Philippines Diliman, Quezon City, Philippines

* **Correspondence:** Email: rmendoza@math.upd.edu.ph.

Abstract: Data smoothing is a method that involves finding a sequence of values that exhibits the trend of a given set of data. This technique has useful applications in dealing with time series data with underlying fluctuations or seasonality and is commonly carried out by solving a minimization problem with a discrete solution that takes into account data fidelity and smoothness. In this paper, we propose a method to obtain the smooth approximation of data by solving a minimization problem in a function space. The existence of the unique minimizer is shown. Using polynomial basis functions, the problem is projected to a finite dimension. Unlike the standard discrete approach, the complexity of our method does not depend on the number of data points. Since the calculated smooth data is represented by a polynomial, additional information about the behavior of the data, such as rate of change, extreme values, concavity, etc., can be drawn. Furthermore, interpolation and extrapolation are straightforward. We demonstrate our proposed method in obtaining smooth mortality rates for the Philippines, analyzing the underlying trend in COVID-19 datasets, and handling incomplete and high-frequency data.

Keywords: data smoothing; Whittaker-Henderson method; Sobolev polynomials; high-frequency data; approximation; generalized cross validation score
Mathematics Subject Classification: 65K10, 90C23, 35A15

1. Introduction

Data smoothing is a method commonly used to obtain a smooth approximation of a crude data set. If $\{f_1, f_2, \ldots, f_n\}$ is a sequence of n data points, the goal of data smoothing is to find a corresponding sequence of graduated data points $\{u_1, u_2, \ldots, u_n\}$ that provides a better representation of the underlying unknown true values [7, 33]. Data smoothing is also referred to as data graduation in actuarial science and is commonly used in actuarial studies to compute smooth mortality rates from crude data.

Goodness-of-fit and smoothness are the two most important criteria in data smoothing [7]. The Whittaker-Henderson method (WHM), a common non-parametric technique often presented as an

alternative to the moving average method [18, 20] in data smoothing, generates the smoothed data points while balancing both criteria. In WHM, the sequence of smoothed data points is the minimizer $u = \{u_i\}_{i=1}^n$ of the function

$$Q(u) = \sum_{i=1}^{n} \omega_i (u_i - f_i)^2 + \lambda \sum_{i=1}^{n} (\Delta u_i)^2, \tag{1.1}$$

where

$$\begin{aligned} \omega_i &: \text{positive weight vector associated with the data,} \\ \Delta u_i &: \text{Newton's advancing operator defined by } \Delta u_i := u_{i+1} - u_i, \\ \lambda &: \text{a positive smoothing parameter.} \end{aligned}$$

The first term is a measure of fidelity to the original data, while the second term is a measure of smoothness. The weights ω_i's are pre-determined and data-dependent. Note that for high values of λ, the smoothing is favored, while lower values result in smoothed data that is closer to actual values. A WHM where the order of the difference operator Δ is set to 2 is known as the Hodrick-Prescott (HP) filter [15].

Several techniques for data graduation can be found in [12, 16, 33, 43, 44]. These methods treat data graduation as a minimization problem in $\mathbb{R}^n$. In [26], the WHM in (1.1) is generalized as a minimization problem in the function space $L^2(\Omega)$. By treating the data $\{u_1, u_2, \ldots, u_n\}$ as a piecewise linear function, the data graduation problem in (1.1) is solved by minimizing the functional $\tilde{J}(u) : L^2(\Omega) \longrightarrow \mathbb{R}$ given by

$$\tilde{J}(u) = \frac{1}{2} \int_\Omega \omega (u - f)^2 dx + \frac{\lambda}{2} \int_\Omega |\mathbf{D}u|^2 dx, \tag{1.2}$$

where f is an interpolation of the data points $\{f_1, f_2, \ldots, f_n\}$, ω is an interpolation of the weights $\{\omega_1, \omega_2, \ldots, \omega_n\}$, and $\mathbf{D}$ is the derivative operator.

The functional (1.2) can be viewed as the infinite-dimensional generalization of (1.1). Note that the second term in (1.2) is a penalty term for data smoothing. This technique of adding a term for smoothing is called *regularization*, which has gained popularity in various applications [2, 30, 31, 34, 39]. In this work, we modify the smoothing term of the functional $\tilde{J}$ in (1.2) and consider

$$J(u) = \frac{1}{2} \int_\Omega \omega (u - f)^2 dx + \frac{\lambda}{2} \int_\Omega |\mathbf{D}^m u|^2 dx, \tag{1.3}$$

where m is the order of the differential operator $\mathbf{D}$. By doing this, the solution gains more regularity. To minimize J in (1.3), we project the problem to a finite-dimensional polynomial space. Hence, the solution that we obtain is a polynomial. This approach is motivated by [23], where it is argued that the WHM is useful in obtaining smooth data points when these graduated data points approximate a polynomial. One advantage of having a polynomial solution is that it is easier to analyze the solution. With a polynomial solution, one can easily obtain extreme values, the relevant points at which the graduated data are increasing or decreasing, and concavity of the smooth data. Furthermore, having a continuous solution instead of discrete points makes interpolation and extrapolation much easier.

Our proposed method can be used to understand time series trends by analyzing the resulting polynomial solution. In [40], the use of polynomials is demonstrated in the analysis of COVID-19

cases. By partitioning the daily new COVID-19 cases into sections and fitting these sections into either a logarithmic, exponential, or linear function, the authors characterized the trend of the disease and identified periods when the incidence of the disease is rapidly or slowly increasing. However, it is not clear how the data set is partitioned. Identifying all possible partitions of data points and evaluating the fit of these points to each of the three functions may be computationally expensive. With polynomials, partitioning can be set at inflection points.

The function Q in (1.1) uses the Newton's advancing operator Δu_i, which assumes that data points of independent variables must be evenly spaced. Hence, the WHM cannot be used for data with missing terms. Our proposed scheme does not have this limitation because one can still compute the polynomial interpolation of the data, regardless if it is not evenly spaced or some of its terms are missing. Another disadvantage of WHM is that it solves a linear system whose dimension is equal to the number of variables in (1.1). Hence, smoothing high frequency data using WHM requires solving a high-dimensional linear system. In our proposed method, the number of variables depends only on the degree of the polynomial. Thus, regardless of having incomplete or high frequency data, the resulting minimization problem has the same dimension.

This paper is organized as follows: Section 2 discusses the derivation of the minimizer of (1.3) in a finite-dimensional polynomial space. Section 3 presents our proposed algorithm. In Section 4, we present the applications of our method on mortality rates and COVID-19 data. We also apply our method to data with missing terms and high frequency data. Finally, we give our conclusions and recommendations in Section 5.

2. Theoretical framework

Because we expect the minimizer of (1.3) to be at least m-times differentiable, we consider the Sobolev space $H^m(\Omega)$ [9]. We present a first-order optimality condition for (1.3). We show that solving this optimality condition is equivalent to minimizing J in (1.3).

Theorem 1. *Suppose f and ω are sufficiently smooth functions. Because ω is an interpolation of the weights, then we can assume that $0 < \omega(x) \leq \bar{\omega}$, $\forall x \in \Omega$, for some $\bar{\omega} > 0$. Then, $u \in H^m(\Omega)$ is a minimizer of* (1.3)*, if and only if, u satisfies*

$$\int_\Omega \omega(uv)dx + \lambda \int_\Omega \boldsymbol{D}^{(m)}u \cdot \boldsymbol{D}^{(m)}v dx = \int_\Omega fv dx, \tag{2.1}$$

for all $v \in H^m(\Omega)$ with $m \in \mathbb{N}$.

Proof. Let $v \in H^m(\Omega)$ with $m \in \mathbb{N}$. Define $r : \mathbb{R} \to \mathbb{R}$ by $r(t) := J(u + tv) - J(u)$. Suppose u is the minimizer of (1.3). Then $J(u) \leq J(u + tv)$ for any $t \in \mathbb{R}$. Hence, $r(t) \geq 0$, $\forall t \in \mathbb{R}$ and $r(t) = 0$ at $t = 0$. By the definition of J, we can simplify $r(t)$ as

$$r(t) = t\left[\int_\Omega \omega\,(uv - fv)\,dx + \lambda \int_\Omega \mathbf{D}^{(m)}u \cdot \mathbf{D}^{(m)}v dx\right] + \frac{t^2}{2}\int_\Omega \omega v^2 + |\mathbf{D}^m v|^2 dx.$$

Hence,

$$r'(t) = \left[\int_\Omega \omega\,(uv - fv)\,dx + \lambda \int_\Omega \mathbf{D}^{(m)}u \cdot \mathbf{D}^{(m)}v dx\right] + t\int_\Omega \omega v^2 + \lambda|\mathbf{D}^m v|^2 dx.$$

Since r is optimal at $t = 0$, then

$$0 = r'(0) = \left[\int_\Omega \omega\,(uv - fv)\,dx + \lambda \int_\Omega \mathbf{D}^{(m)}u \cdot \mathbf{D}^{(m)}v dx\right].$$

Rearranging the above equation gives us (2.1).

Now, suppose u satisfies (2.1). Let $z \in H^m(\Omega)$ and define $v := z - u$. Then,

$$J(z) - J(u) = \left[\int_\Omega \omega\,(uv - fv)\,dx + \lambda \int_\Omega \mathbf{D}^{(m)}u \cdot \mathbf{D}^{(m)}v dx\right] + \frac{1}{2}\int_\Omega \omega v^2 + \lambda|\mathbf{D}^m v|^2 dx.$$

From the assumption, the first two terms of the above equation is 0. Thus,

$$J(z) - J(u) = \frac{1}{2}\int_\Omega \omega v^2 + \lambda|\mathbf{D}^m v|^2 dx \geq 0.$$

Therefore, for any $z \in H^m(\Omega)$, $J(z) \geq J(u)$, which means that u is a minimizer of J. □

Theorem 1 tells us that it is sufficient to solve (2.1) to minimize J in (1.3). To show that (2.1) has a unique minimizer, we first need the following results.

Lemma 1. *Let $\Omega \subset \mathbb{R}^d$ be bounded with Lipschitz boundary. Suppose g satisfies the following conditions:*

(1) $g : H^m(\Omega) \to [0, +\infty)$ is a seminorm.
(2) There exists a positive constant C such that $0 \leq g(v) \leq C\|v\|_{H^m(\Omega)}$, for all $v \in H^m(\Omega)$.
(3) If $v \in \mathcal{P}_{k-1} := \{$space of polynomials of degree $k-1\}$ and $g(v) = 0$, then $v \equiv 0$. Then the norm

$$\|u\|_{H^m(\Omega)} := \left(\sum_{|\alpha\|\leq m} \|\boldsymbol{D}^\alpha u\|_{L^2(\Omega)}\right)^{1/2} \tag{2.2}$$

is equivalent to the norm

$$\|u\|'_{H^m(\Omega)} := \left(g^2(u) + \sum_{|\alpha|=m}\int_\Omega |\boldsymbol{D}^m u|^2 dx\right)^{1/2} \tag{2.3}$$

Proof. This is a special case of the result proven in [38], where the equivalence was shown on a system of functionals and on a more general Sobolev space $W^{m,p}(\Omega)$. □

Lemma 2. *For $\Omega \subset \mathbb{R}$, we define*

$$|u|'_{H^m(\Omega)} := \left(\int_\Omega \omega u^2 + \lambda|\boldsymbol{D}^m u|^2 dx\right)^{1/2} \tag{2.4}$$

Then the norm in (2.4) *is equivalent to the norm in* (2.2)*, or equivalently, $\exists \rho_1, \rho_2 > 0$ such that*

$$\rho_1\|u\|_{H^m(\Omega)} \leq |u|'_{H^m(\Omega)} \leq \rho_2\|u\|_{H^m(\Omega)}. \tag{2.5}$$

Proof. The norm in (2.4) is equivalent to

$$|u|'_{H^m(\Omega)} := \left(\lambda \int_\Omega \frac{\omega}{\lambda} u^2 + |\mathbf{D}^m u|^2 dx\right)^{1/2}.$$

We can set

$$g(u) := \left(\int_\Omega \frac{\omega}{\lambda} u^2 dx\right)^{1/2}.$$

Note that g is a weighted $L^2(\Omega)$-norm, and hence the first condition of Lemma 1 is satisfied. Because $\omega \leq \bar{\omega}$, for all $x \in \Omega$ and from the definition of norm in (2.2), we get

$$g(v) \leq \sqrt{\frac{\bar{\omega}}{\lambda}} \|v\|_{L^2(\Omega} \leq \sqrt{\frac{\bar{\omega}}{\lambda}} \|v\|_{H^m(\Omega)}.$$

Therefore, the second condition of Lemma 1 is satisfied. The last condition easily follows because $\dfrac{\omega(x)}{\lambda} > 0$ for all $x \in \Omega$. Therefore, by Lemma 1, our assertion holds. □

Theorem 2. *The variational formulation in* (2.1) *has a unique solution in* $H^m(\Omega)$.

Proof. We define

$$a(u, v) := \int_\Omega \omega(uv) dx + \lambda \int_\Omega \mathbf{D}^{(m)} u \cdot \mathbf{D}^{(m)} v dx$$

and

$$b(v) := \int_\Omega f v dx.$$

We use Lax-Milgram's Lemma to prove the existence of the unique solution. For the discussion of this lemma, we refer the readers to [9]. To do this, we need to show that a is bilinear, bounded, and coercive, and b is linear and bounded. The bilinearity and linearity of a and b, respectively, follows directly from their respective definitions.

We now show that a is bounded, that is, $\exists C_1 > 0$ such that $|a(u, v)| \leq C_1 \|u\|_{H^m(\Omega)} \|v\|_{H^m(\Omega)}$. We use triangle inequality, Cauchy-Schwarz inequality, and the definition of the norm in (2.2). Thus,

$$\begin{aligned} |a(u, v)| &\leq \left|\int_\Omega \omega(uv) dx\right| + \lambda \left|\int_\Omega \mathbf{D}^{(m)} u \cdot \mathbf{D}^{(m)} v dx\right| \\ &\leq \bar{\omega} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \lambda \|\mathbf{D}^{(m)} u\|_{L^2(\Omega)} \|\mathbf{D}^{(m)} v\|_{L^2(\Omega)} \\ &\leq \underbrace{\max(\bar{\omega}, \lambda)}_{C_1} \|u\|_{H^m(\Omega)} \|v\|_{H^m(\Omega)}. \end{aligned}$$

We use (2.5) from Lemma 2 to show that a is coercive, that is, $\exists C_2 > 0$ such that $|a(u, u)| \geq C_2 \|u\|^2_{H^m(\Omega)}$ for all $u \in H^m(\Omega)$. Indeed,

$$|a(u, u)| = \int_\Omega \omega u^2 + \lambda |\mathbf{D}^m u|^2 dx$$

$$= \left[|u|'_{H^m(\Omega)}\right]^2$$
$$\geq \underbrace{\rho_1}_{C_2} \|u\|^2_{H^m(\Omega)}.$$

Finally, we show that b is bounded, that is, $\exists C_3 > 0$ such that $|b(v)| \leq C_3\|v\|_{L^2(\Omega)}$. Using Cauchy-Schwarz inequality, we get

$$|b(v)| \leq \underbrace{\|f\|_{L^2(\Omega)}}_{C_3}.\|v\|_{L^2(\Omega)}.$$

□

We have shown in Theorem 1 that we can minimize J in (1.3) by solving an equivalent variational formulation, which we have shown in Theorem 2 to have a unique solution in $H^m(\Omega)$. Because of the equivalence of $\|\cdot\|_{H^m(\Omega)}$ and $|\cdot|'_{H^m(\Omega)}$, we can use the following inner product for $H^m(\Omega)$:

$$\langle u, v\rangle_H := \int_\Omega \omega(uv)dx + \lambda \int_\Omega \mathbf{D}^{(m)}u \cdot \mathbf{D}^{(m)}vdx, \tag{2.6}$$

where ω is a continuous function in Ω, and $\lambda > 0$. With the above inner product, we can rewrite the optimality condition in (2.1) as

$$\langle u, v\rangle_H = \langle f, v\rangle_{L^2(\Omega)} \quad \forall v \in H^m(\Omega).$$

In this work, we pose the above variational formulation in a finite-dimensional subspace S of $H^m(\Omega)$, that is, we solve

$$\langle u, v\rangle_H = \langle f, v\rangle_{L^2(\Omega)} \quad \forall v \in S.$$

We consider the subspace S which is spanned by a set of orthonormal polynomials $\{p_1, p_2, \ldots, p_l,\}$ for some $l \in \mathbb{N}$. We refer to $\{p_1, p_2, \ldots, p_l\}$ as *Sobolev polynomials* because they are constructed using the inner product in Eq (2.6), which involves derivatives, such that each polynomial in this subspace are also in the Sobolev space $W^{m,p}(\Omega)$. We were motivated to use Sobolev polynomials because they are the best polynomial approximation to a function f with respect to the $L^2(\Omega)$ norm [28]. A comprehensive review of the history and recent development in the study of Sobolev polynomials can be found in [27].

Since $S = \text{span}\{p_1, p_2, \ldots, p_l\}$, it is sufficient to find $u \in S$ such that

$$\langle u, p_j\rangle_H = \langle f, p_j\rangle_{L^2(\Omega)} \quad \forall j \in \{1, 2, \ldots, l\}.$$

If $u \in S$, then $u = \sum_{i=1}^{l} u_i p_i$ and so, we solve

$$\langle \sum_{i=1}^{n} u_i p_i, p_j\rangle_H = \langle f, p_j\rangle_{L^2(\Omega)} \quad \forall j \in \{1, 2, \ldots, l\}.$$

Equivalently,

$$\sum_{i=1}^{n} u_i \langle p_i, p_j\rangle_H = \langle f, p_j\rangle_{L^2(\Omega)} \quad \forall j \in \{1, 2, \ldots, l\}.$$

Because $\langle p_i, p_j \rangle_H = \delta_{ij}$, we obtain

$$u_i = \langle \omega f, p_i \rangle_{L^2(\Omega)} \quad \forall i \in \{1, 2, \ldots, l\}. \tag{2.7}$$

Therefore, the solution $u \in S$ of

$$\langle u, v \rangle_H = \langle f, v \rangle_{L^2(\Omega} \quad \forall v \in S$$

is given by

$$u = \sum_{i=1}^{l} p_i \langle \omega f, p_i \rangle_{L^2(\Omega)}. \tag{2.8}$$

We construct the orthonormal polynomials using Gram-Schmidt orthonormalization process [24]. We will also use polynomial interpolations for ω and f so that the solution u is a linear combination of polynomials. The Gram-Schmidt process is started using Chebyshev polynomials as the initial basis functions [36].

3. Proposed algorithm

To compute u in (2.8), we first obtain the polynomial interpolations of ω and f. Then we get a set of orthonormal polynomials from an independent set of polynomials. In our case, we use the Chebyshev polynomials $\{c_1, c_2, \ldots, c_l\}$, for some user-defined degree $l - 1$. Then, using Gram-Schmidt orthonormalization process, we construct a set of orthonormal polynomials $\{p_1, p_2, \ldots, p_l\}$. The order of differentiation m is also user-defined. The construction of the Sobolev polynomials is summarized in Algorithm 1.

Algorithm 1 Creating a set of orthonormal Sobolev polynomials from Chebyshev polynomials

1: *Input:* Set the desired degree of polynomial $l - 1$, and the endpoints a, b of the interval of interest according to the crude data points.
2: Construct a set of basis Chebyshev polynomials of dimension l over the interval (a, b).
3: Apply the Gram-Schmidt orthogonalization procedure that uses the inner product in Eq (2.6) on the set of Chebyshev polynomials from step 2.
4: *Output:* The Gram-Schmidt orthogonalization procedure that uses the specified inner product in step 3 results in a set of orthonormal Sobolev polynomials $\{p_1, p_2, \ldots, p_l\}$.

The user can also specify the value for the smoothing parameter λ. However, by default, the algorithm uses the approach presented in [12, 42]. The smoothing parameter λ is computed as the minimizer of the generalized cross validation (GCV) score, which is expressed as

$$GCV(\lambda) = \frac{n \sum_{i=1}^{n} (\hat{f}_i - f_i)^2}{\left(n - \sum_{i=1}^{n} (1 + \lambda \gamma_i^2)^{-1} \right)^2}, \tag{3.1}$$

where

$$\hat{f} = \left(I_n + \lambda D^T D\right)^{-1} f,$$

n is the number of data points, $\gamma_i' s$ are the eigenvalues of $D^T D$, and D is a tridiagonal matrix with entries given by

$$D_{i,i-1} = \frac{2}{h_{i-1}(h_{i-1} + h_i)}, \quad D_{i,i} = \frac{-2}{h_{i-1}h_i}, \quad D_{i-1,i} = \frac{2}{h_i(h_{i-1} + h_i)},$$

with h_i representing the step between $\hat{f}_i$ and $\hat{f}_{i+1}$.

For a detailed discussion on how to use GCV to identify the smoothing parameter, we refer the readers to [11, 12, 14, 41, 42]. To compute the GCV value numerically, we used the Matlab code provided in the Appendix of [12]. The *GCV* function in (3.1) may have multiple local minima so we have to use a global minimizer. In this study, we use Genetic Algorithm (GA) which has growing applications in science and engineering because of its capability to estimate the global minimum and its non-reliance on the derivative of the objective function [21, 22, 37]. We utilize the Matlab built-in function `ga`, which only requires the upper and lower bounds for the smoothing parameter. Although GA can converge to the global minimum, it can still sometimes get stuck at a local minimum. To guarantee global convergence, we run `ga` 10 times and store the best solution λ_{GA}. To further improve accuracy, we hybridize GA with `fmincon`, a Matlab built-in code for interior point algorithm [6], which is a local search technique. We use λ_{GA} as the initial guess of `fmincon`. We denote $\lambda^\star$ as the minimizer of the hybrid GA-interior point method. Upon obtaining $\lambda^\star$, we can use the orthogonal polynomial calculated in Algorithm 1 and obtain the smooth polynomial approximation u in (2.8). We summarize the proposed method in Algorithm 2.

Algorithm 2 Data graduation using Sobolev polynomials

1: *Input:* The crude data $\{f_i\}_{i=1}^n$.
2: Set the value for l (dimension of the polynomial space) and m (order of differentiation).
3: Determine the polynomial interpolations, ω and f, of the weights and the data, respectively.
4: Set the bounds for the smoothing parameter $[\lambda_{\min}, \lambda_{\max}]$.
5: Estimate the minimizer the function *GCV* in (3.1) over the interval $[\lambda_{\min}, \lambda_{\max}]$ using the Matlab built-in program `ga` 10 times. Set λ_{GA} as the minimizer among the 10 solutions with the least GCV score.
6: To get a more accurate global minimizer $\lambda^\star$, implement the interior point method to the *GCV* in (3.1) over the interval $[\lambda_{\min}, \lambda_{\max}]$ using the Matlab built-in pogram `fmincon`.
7: Using the smoothing parameter $\lambda^\star$, obtain a set of orthonormal Sobolev polynomial functions $\{p_1, p_2, \ldots, p_l\}$ according to Algorithm 1.
8: *Output:* The smooth approximation of the data as a polynomial function $u = \sum_{i=1}^{l} p_i \langle \omega f, p_i \rangle_{L^2(\Omega)}$.

Given a data $\{f_1, f_2, \ldots, f_n\}$, a unique smooth approximation u is calculated using (2.8). Note that the orthogonal polynomials p_i do not rely on the data. This means that if the data has a particular error structure, uncertainty might occur in the coefficient terms u_i in (2.7). In this study, we rely on a general bootstrap approach presented in [8, 10] to quantify the uncertainty of the coefficients and construct confidence intervals. The step-by-step procedure is stated in Algorithm 3.

Algorithm 3 Uncertainty quantification

1: *Input:* The coefficients u_i calculated from (2.7).
2: Generate M simulated data $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n$ by adding noise assuming an error structure to the calculated Sobolev polynomial u at points $x_1, x_2, \ldots, x_n$.
3: Recalculate the coefficients $\hat{u}_i$ from (2.7) given the simulated data $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n$.
4: From the recalculated M set of values of $\hat{u}_i$, characterize the distribution of the coefficients and calculate the confidence interval.
5: *Output:* Histograms to display the empirical distributions of the coefficients and the corresponding confidence intervals.

4. Results and discussion

In this section, we test the proposed algorithm to determine the graduated values of different datasets. All numerical simulations were performed in Matlab 2021a on a computer with Intel(R) Core(TM) i7-8550U CPU clocked at 1.80GHz and 1.99 GHz with 8 GB of RAM that runs Windows 10 OS. By default, we set the order of derivative to 10 and the degree of the polynomial to 8. In all simulations, the bounds for the smoothing parameter is set to $[\lambda_{\min}, \lambda_{\max}] := [0.1, 5]$.

For the first application, we test our method to calculate the graduated value of male mortality rates. We obtained the crude mortality rates from the 2017 Philippine Intercompany Mortality Study of the Actuarial Society of the Philippines [1]. The illustration of our proposed scheme is shown in the left panel of Figure 1. The red curve represents the linearly interpolated data and the black dashed curve shows our proposed method. The smooth approximation using Whittaker-Henderson graduation is shown in green. It can be seen how our proposed method obtained similar results as the Whittaker-Henderson method. Both methods produced mortality rates that corrected the fluctuations observed in the crude data. Moreover, plots produced from both methods also continuously increase across all ages. While graduated data produced by the discrete Whittaker-Henderson method may appear smooth because of interpolation, the graduated data produced by our method is guaranteed to be smooth because it is represented by a polynomial. Note that the crude data for the mortality rates for ages 80 to 85 are not available. The Whittaker-Henderson method cannot be used to obtain the graduated values outside the range of the data set. In [1], extrapolation was used to calculate the graduated values for ages 80–85. The authors used the Gompertz-Makeham model (blue curve) to do this. An advantage of our proposed method is that extrapolation is straightforward. One simply needs to evaluate the polynomial at the points outside the range of the data set. The computed graduated value using the Gompertz-Makeham model and our method are both shown in Figure 1. The right panel shows the plot of the GCV function in (3.1) for $\lambda \in [0.1, 5]$. Observe that the plot (blue curve) is nonlinear and multi-modal which justifies the use of the hybrid GA and interior point method as a global minimization algorithm. The red dot shows the global minimize ($\lambda^\star = 0.5285$), which we set as the smoothing parameter for this problem. Figure 2 presents our uncertainty analysis for this problem using Algorithm 3. For this simulation, we assume a Gamma error structure [5]. The uncertainty in our calculated coefficient values translates into the confidence bounds (red dashed lines) around the smooth approximation of the data. Since the degree of the polynomial is 8, the dimension of the subspace S is 9, which means that we calculate 9 coefficients u_i's. The histograms and confidence intervals are

presented in the bottom panels of Figure 2. Observe that all the computed coefficients fall within their corresponding 95% confidence intervals.

In the proposed algorithm, the user can choose the degree of polynomial. The left panel of Figure 3 demonstrates our next example to determine the graduation of male mortality rates for ages 0–30 using various degrees of polynomial while fixing the order of derivative to 10. For this example, we calculated $\lambda^\star = 1.3798$. As presented, the graduation of mortality rates can produce a range of polynomial functions that practitioners can choose from depending on their purpose. However, it is also possible to set the degree of polynomial based on the fit of the smoothing curve. One approach is to iterate over a set of positive integers and select the degree that produces the polynomial that has least distance to the piecewise linear interpolation of data. The right panel of Figure 3 demonstrates this and shows that a polynomial with degree 8 ($l = 9$) has the best fit. This polynomial is shown as the red curve in the left panel.

We also applied our proposed method to obtain a smooth time series data for daily new COVID-19 cases in Germany [17]. This dataset spans new COVID-19 cases from 1 September 2020 to 31 May 2021. Here, $\lambda^\star = 0.7227$. The left panel in Figure 4 presents the crude daily new COVID-19 cases in Germany as blue line, and the graduated data as red line. As shown, the plot of the graduated data points exhibits peaks and troughs that follow the crude data. The underlying trend in the data can also be observed after smoothing.

The right panel in Figure 4 compares the results of smoothing using different methods, implemented using the Matlab built-in function `smoothdata`. Simple moving average, gaussian or kernel smoothing, and smoothing using locally weighted scatter plot smoothing (LOWESS) all use a window size of 7 because this allows coverage of both the incubation period and the time from the first appearance of symptoms to diagnosis [19].

Daily new COVID-19 cases can be viewed as the change in the total number of cases for every small change in time. Hence, by integrating the interpolating polynomial, the cumulative number of COVID-19 cases can be approximated. As an illustration, we integrate the polynomial obtained in Figure 4 to compute the cumulative number of cases. In Figure 5 (left panel), we illustrate how the integral of the polynomial closely approximates the actual cumulative data. The resulting approximation of the cumulative data is also a polynomial, which makes finding the inflection points and concavity of the curve easier. These changes in concavity are shown in Figure 5 (right panel). The segments of the plot in red indicate a deceleration in the number of cases while the blue segments indicate acceleration. Plots like this are useful in evaluating the effectiveness of policies in containing the pandemic, or in identifying events that may have contributed to a more rapid spread of the disease.

We also applied our method to obtain the trend in the total COVID-19 deaths in China from 23 January 2020 to 22 February 2020 [29]. For this example, we computed $\lambda^\star = 0.8018$. It was shown in [3] that the growth in the number of deaths related to COVID-19 in the early stages of the pandemic follows a quadratic trend. By setting the degree of polynomial to 2, we obtain the graduation in Figure 6, which captures the trend of the data of COVID-19 deaths.

One of the important applications of our proposed method is in handling time series data with missing terms. Figure 7 demonstrates this in the case of missing time series data points for Schistosomiasis cases in years 2009–2011 and 2013 [35]. Here, $\lambda^\star = 1.5577$. As shown in the left panel, our method produced graduated data points that are also consistent with the general trend in the dataset. Note that both WHM and moving average technique require that the time series data

be evenly spaced. We also applied other methods in extrapolating the missing data points. The right panel of Figure 7 compares the results of our smoothing algorithm when paired with either linear, spline, piecewise cubic hermite interpolating polynomial (PCHIP), and the modified Akima piecewise cubic hermite interpolation (MAkima) to fill the missing data. We used the Matlab built-in function `fillmissing` using different methods in our numerical experiments. Our method was applied after extrapolating the data. For this example, smoothing using the points generated by the piecewise linear interpolation produced the least L^2-norm error. The user can choose which approach they prefer when filling in missing data. This is a pre-processing step that should be implemented before applying our method to data with missing values.

Our method offers a useful alternative in smoothing high-frequency datasets. WHM entails high computational cost in smoothing large datasets because it requires solving a linear system whose dimension is equal to the number of data points. On the contrary, our method is more suitable because the number of unknowns depends only on the degree of the polynomial used in the graduation. We demonstrate this in Figure 8 using temperature data from [32] collected every 5 minutes from 24 April 2019 to 1 October 2020. The data set considered includes 140 thousand data points. For this example, we calculated $\lambda^\star = 0.9586$. Applying WHM to this dataset is not possible because our computer can only handle a linear system of a dimension of at most 30 thousand. Moreover, observe that there are missing data points shown in Figure 8 as a break in the blue line. Based on the results of Figure 7, we used linear interpolation to fill the missing data. Our method was able to provide predicted graduated values for the missing sequence in this high-frequency dataset. Moreover, the smooth approximation also exhibits the trend in the data.

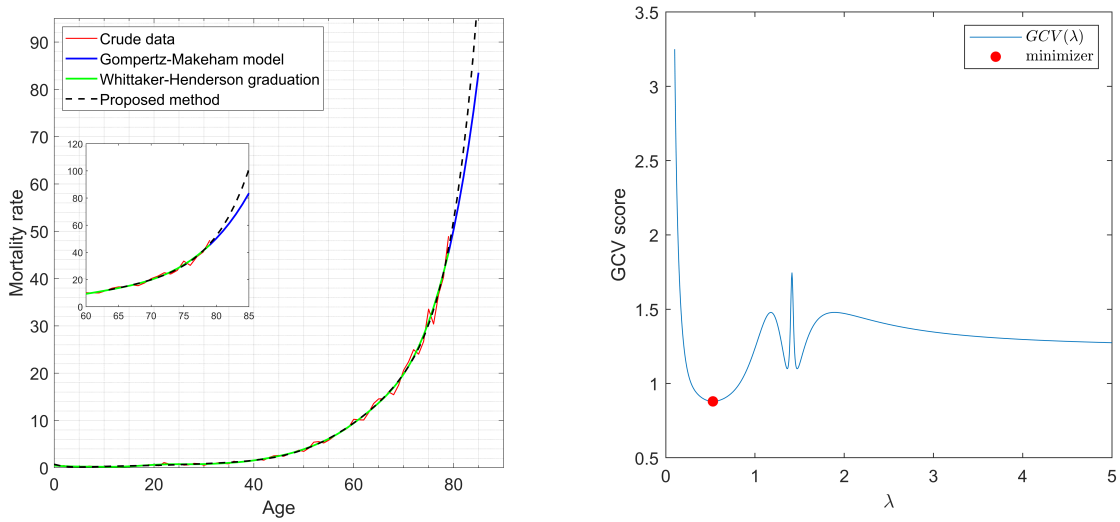

Figure 1. Graduation of male mortality rates. The black dotted line in the left panel represents the graduated rates using our method. The minimization of the *GCV* function is presented on the right panel.

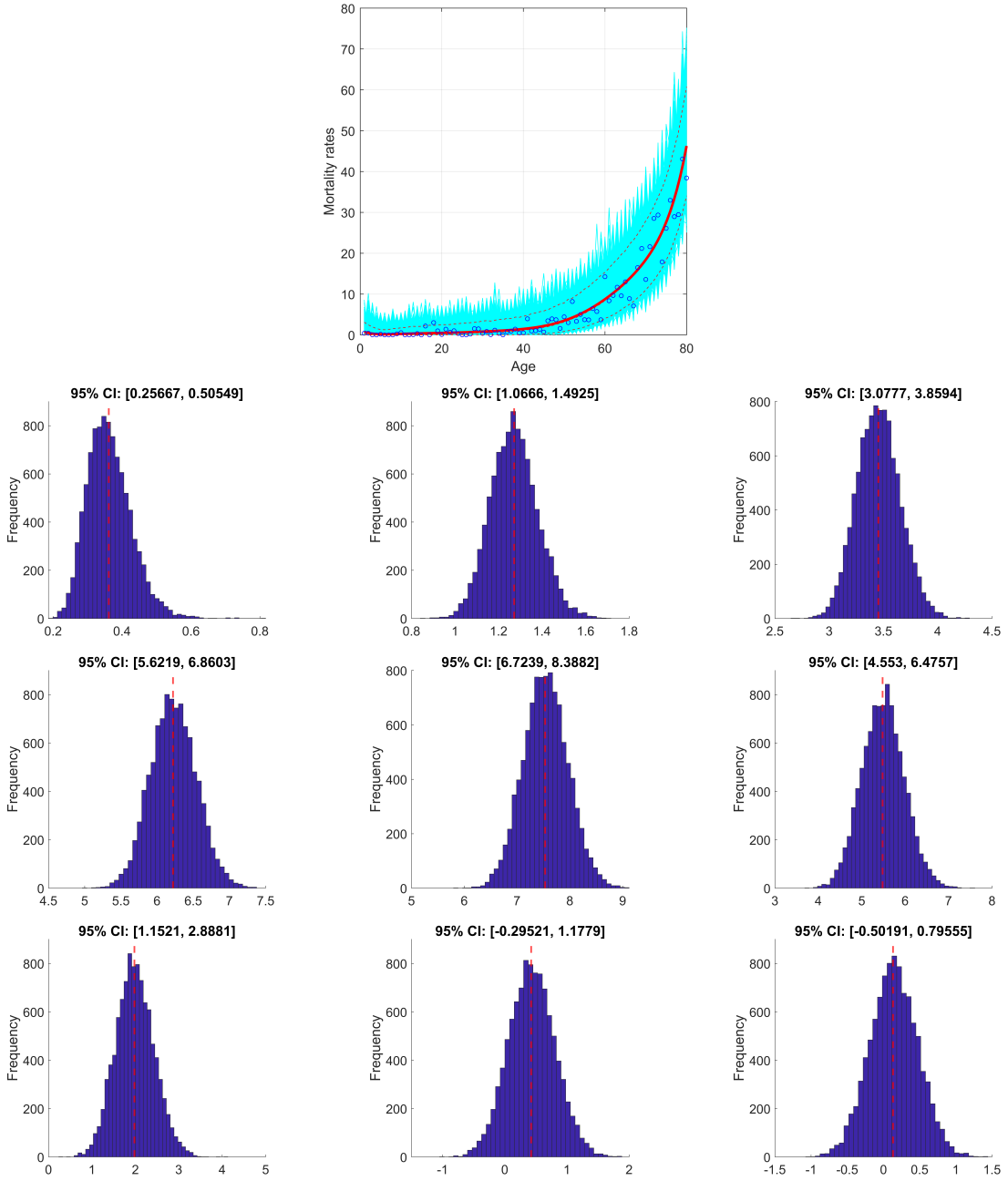

Figure 2. Graduation of male mortality rates (red solid line) with the quantified uncertainty (uppermost panel). The blue circles are the mortality data. The cyan lines are the 10000 realizations of the mortality data assuming Gamma error structure. The red dashed lines illustrate the 95% confidence bands around the smooth approximation. The bottom panels show the histogram that shows the empirical distribution of the coefficients u_i.

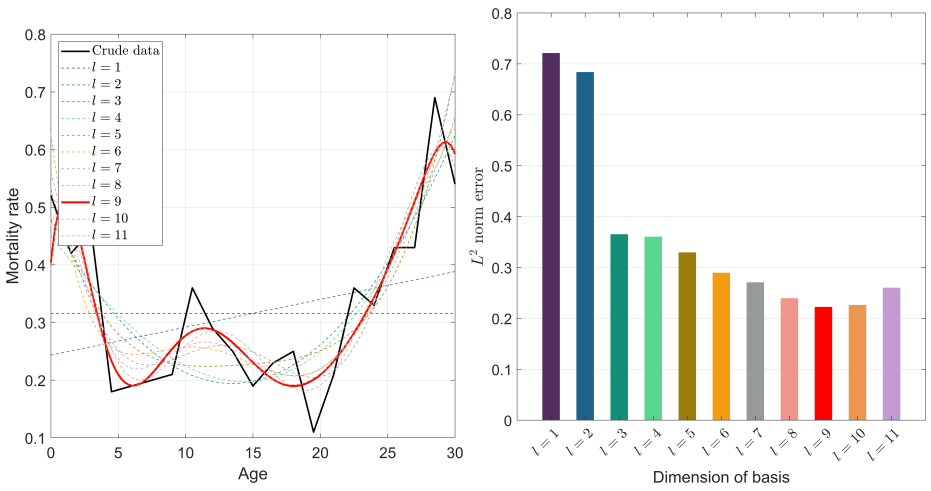

Figure 3. Graduation of male mortality rates for age group 0–30. The left panel presents the smoothing of data using different dimensions of basis (or $l - 1$ degree of polynomial), while the right panel presents the L^2-norm error corresponding to each dimension of basis. The polynomial with degree equal to 8, which is shown as red curve in the left panel produced the best fit according to its L^2-norm error, shown as red bar in the right panel.

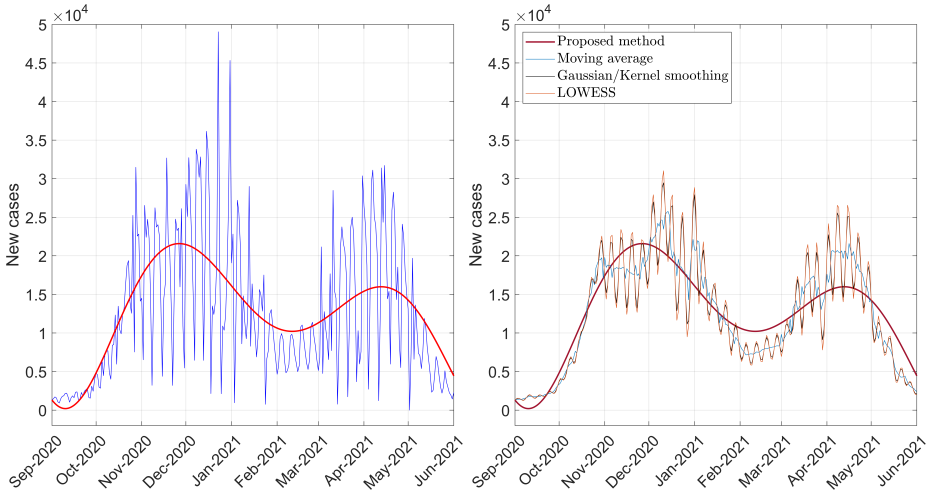

Figure 4. Data smoothing of daily new COVID-19 cases in Germany, September 2020–May 2021. The red line in the left panel represents the graduated new cases. The right panel compares our proposed method with other smoothing techniques.

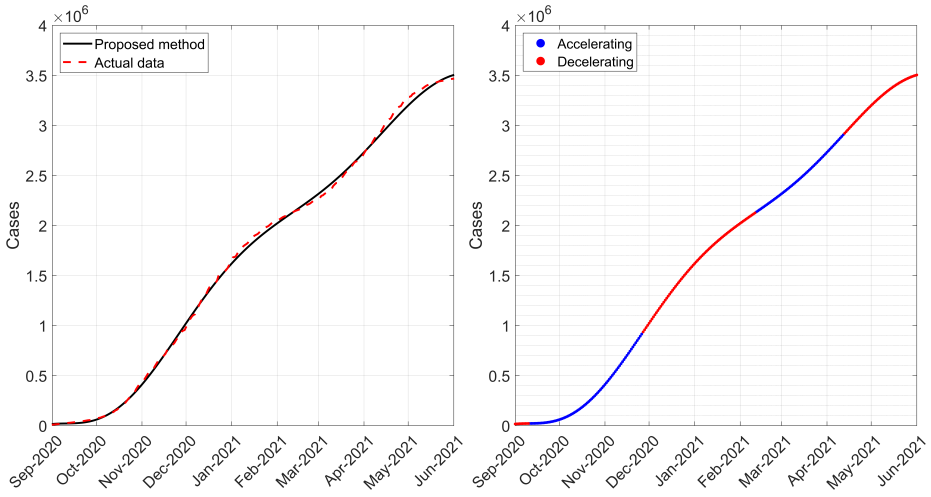

Figure 5. Cumulative COVID-19 cases in Germany, September 2020–May 2021. The left panel compares the cumulative COVID-18 data with the proposed method. The right panel visualizes the time intervals when cases are increasing and decreasing based on the concavity of the calculated Sobolev polynomial.

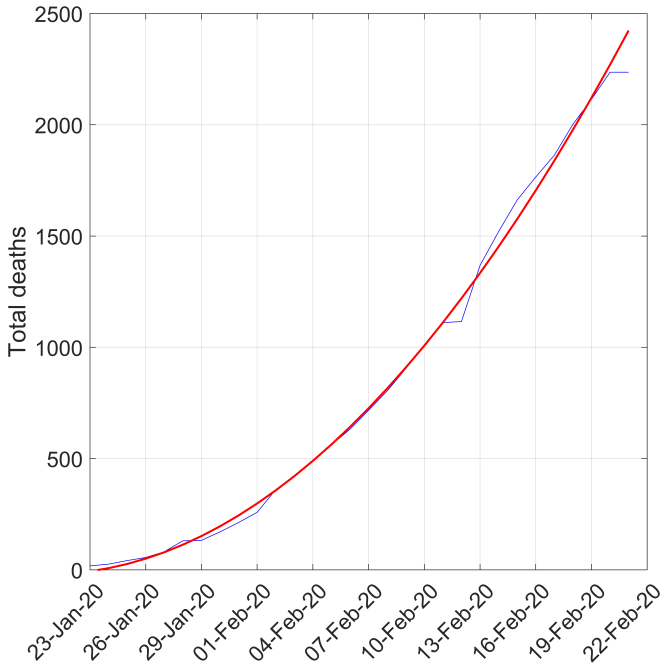

Figure 6. Graduation of total COVID-19 deaths in China, 23 January–22 February 2020. The red line represents the graduated total number of deaths due to COVID-19, assuming that the data follow a quadratic trend [3].

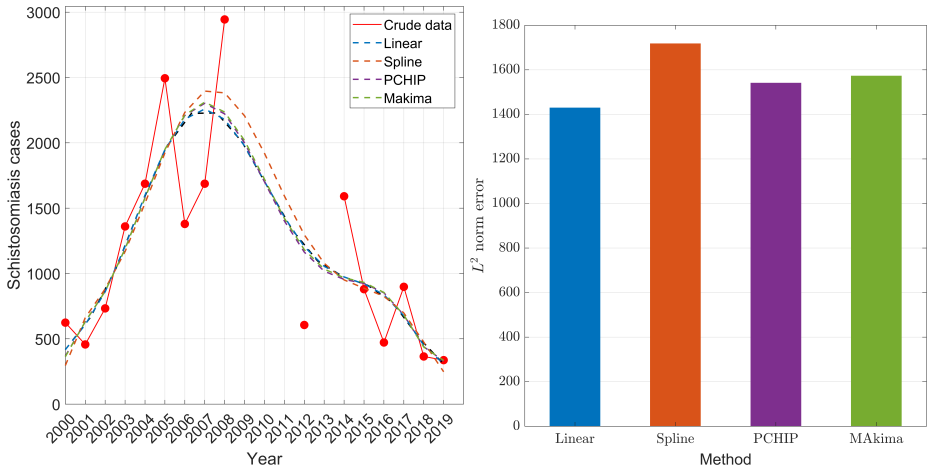

Figure 7. Schistosomiasis cases, 2000–2019. The left panel presents the smooth data using different techniques to fill the missing data while the right panel shows the corresponding L^2-norm error.

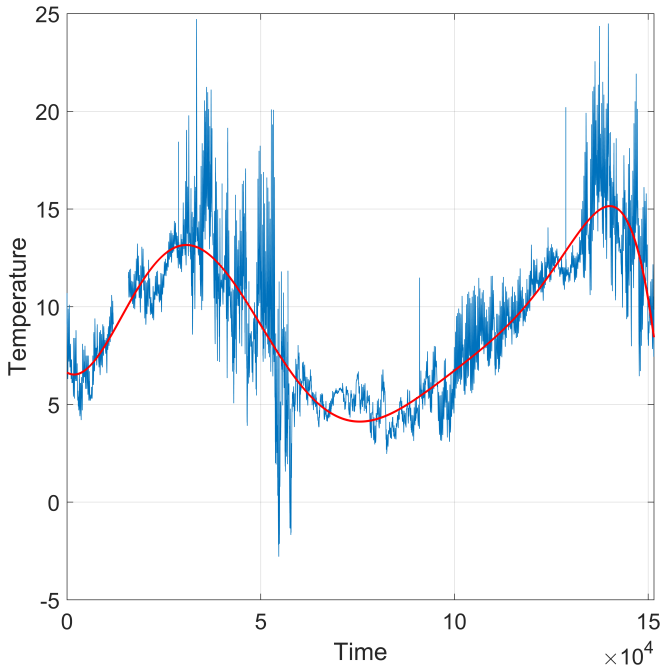

Figure 8. Graduation of water temperature, 24 April 2019–1 October 2020. Blue line represents the crude data. Red line represents the smoothed data.

5. Conclusions

In this paper, we presented an alternative method in data graduation that uses Sobolev polynomials. We have proven that the resulting minimization problem has a unique solution in a suitable function space. Furthermore, we formulated an approach using the Gram-Schmidt orthogonalization process to find the solution in an approximate polynomial space. We applied this method in obtaining the graduated values of male mortality rates in the Philippines, and COVID-19 data. We also demonstrated

the usefulness of our method in addressing missing data points and smoothing high frequency data. Because the obtained results are polynomials, interpolation and extrapolation become straightforward. Furthermore, inflection points and concavity of the the graduated values are easier to identify. The regularity of polynomials makes the analysis simpler.

Our study has some limitations that can be explored in future research. First, although one can calculate the smoothing parameter by minimizing the GCV score, the order of the derivative and the degree of polynomial are entirely user-defined. If the user does not want to use the default values, the user can fix the order of the derivative and choose the degree of the polynomial that will yield the least L^2-error. This can be time-consuming especially if the user sets the order of the derivative to a high value. As a future study, we can explore how both of these relevant parameters can be chosen based solely on data. Second, our study is posed in one-dimensional time series data. For future work, one can extend our method to solve multi-dimensional data smoothing problems. For example, one can consider mortality rates that depend on age and policy duration. Third, one can also consider using other regularization terms (e.g., total variation) and other non-polynomial basis functions. Fourth, in the case of high-frequency data, one can cut the computational cost if the interval is subdivided into segments to reduce the data size [15]. Once the data is segmented, the smooth approximation can be implemented using parallel computing. This means that the solution in each segment represents a spline of the entire smooth approximation. However, this will require continuity conditions between segments, which needs a rigorous theoretical analysis. Fifth, another area that can be explored is the relation of our method with space-state models. Since we are treating the smoothing as a minimization in an infinite dimensional Lebesgue space, an extensive study on how to formulate the corresponding difference equations needs to be carried out. These are exciting research directions but demand a thorough investigation and would require other numerical optimization algorithms.

Acknowledgments

This research was supported by a grant from the Computational Research Laboratory of the Institute of Mathematics, University of the Philippines Diliman.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. *Philippine intercompany mortality study 2017*, Actuarial Society of the Philippines, 2017. Available from: http://www.actuary.org.ph/wp-content/uploads/2017/05/2017-PICM-Study-Final-Report-18May2017.pdf.
2. L. Ambrosio, V. M. Tortelli, Approximation of functional depending on jumps by elliptic functional via t-convergence, *Commun. Pure Appl. Math.*, **43** (1990), 999–1036. https://doi.org/10.1002/cpa.3160430805
3. A. Brandenburg, Piecewise quadratic growth during the 2019 novel coronavirus epidemic, *Infect. Dis. Model.*, **5** (2020), 681–690. https://doi.org/10.1016/j.idm.2020.08.014

4. R. J. Brooks, M. Stone, F. Y. Chan, L. K. Chan, Cross-validatory graduation, *Insur. Math. Econ.*, **7** (1988), 59–66. https://doi.org/10.1016/0167-6687(88)90097-2
5. M. A. Buford, W. L. Hafley, Probability distributions as models for mortality, *Forest Sci.*, **31** (1985), 331–341.
6. R. H. Byrd, J. C. Gilbert, J. Nocedal, A trust region method based on interior point techniques for nonlinear programming, *Math. Program.*, **89** (2000), 149–185. https://doi.org/10.1007/PL00011391
7. F. Y. Chan, L. K. Chan, E. R. Mead, Properties and modifications of Whittaker-Henderson graduation, *Scand. Actuar. J.*, **1982** (1982), 57–61. https://doi.org/10.1080/03461238.1982.10405433
8. G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, *Infect. Dis. Model.*, **2** (2017), 379–398.
9. D. Cioranescu, P. Donato, M. P. Roque, *An introduction to second order partial differential equations: Classical and variational solutions*, World Scientific, Singapore, 2018.
10. B. Efron, R. J. Tibshirani, *An introduction to the bootstrap*, CRC Press, 1992.
11. P. H. Eilers, A perfect smoother, *Anal. Chemis.*, **75** (2003), 3631–3636. https://doi.org/10.1021/ac034173t
12. D. Garcia, Robust smoothing of gridded data in one and higher dimensions with missing values, *Comput. Stat. Data Anal.*, **54** (2010), 1167–1178. https://doi.org/10.1016/j.csda.2009.09.020
13. L. Grafakos, *Classical Fourier analysis*, Springer, New York, 2008.
14. P. Graven, Smoothing noisy data with spline function: Estimating the correct degree of smoothing by the method of Generalized Cross-Validaton, *Numer. Math.*, **31** (1978), 377–403.
15. V. Guerrero, E. Silva, Smoothing a time series by segments of the data range, *Commun. Stat.-Theor. M.*, **44** (2015), 4568–4585. https://doi.org/10.1080/03610926.2014.901372
16. V. Guerrero, Estimating trends with percentage of smoothness chosen by the user, *Int. Stat. Rev.*, **76** (2008), 182–202. https://doi.org/10.1111/j.1751-5823.2008.00047.x
17. R. Hannah, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, et al., *Coronavirus pandemic (COVID-19)*, Our World in Data, 2020. Available from: `https://ourworldindata.org/coronavirus`.
18. S. Hansun, *A new approach of moving average method in time series analysis*, 2013 conference on new media studies (CoNMedia), IEEE, 2013, 1–4.
19. Y. He, X. Wang, H. He, J. Zhai, B. Wang, Moving average based index for judging the peak of COVID-19 epidemic, *Int. J. Environ. Res. Pub. He.*, **17** (2021), 5288. https://doi.org/10.3390/ijerph17155288
20. R. J. Hyndman, *Moving averages*, International Encyclopedia of Statistical Science, Springer, Berlin, Heidelberg, 2011, 866–896. https://doi.org/10.1007/978-3-642-04898-2 https://doi.org/10.1007/978-3-642-04898-2_380
21. C. U. Jamilla, R. G. Mendoza, V. M. P. Mendoza, Parameter estimation in neutral delay differential equations using genetic algorithm with multi-parent crossover, *IEEE Access*, **9** (2021), 131348–131364. https://doi.org/10.1109/ACCESS.2021.3113677

22. S. Katoch, S. S. Chauhan, V. Kumar, A review on genetic algorithm: Past, present, and future, *Multimed. Tools Appl.*, **9** (2021), 8091–8126.

23. F. Knorr, Multidimensional Whittaker-Henderson graduation, *Trans. Soc. Actuar.*, **36** (1984), 213–255.

24. D. C. Lay, *Linear algebra and its applications*, 5 Eds., Pearson, Boston, 2016.

25. F. Macaulay, *The Whittaker-Henderson method of graduation*, The smooting of time series, National Bureau of Economic Research, New York, 1931, 89–99.

26. J. L. Manejero, R. Mendoza, Variational approach to data graduation, *Philipp. J. Sci.*, **149** (2020), 431–449.

27. F. Marcellan, Y. Xu, On Sobolev orthogonal polynomials, *Expo. Math.*, **33** (2015), 308–352. https://doi.org/10.1016/j.exmath.2014.10.002

28. F. Marcellan, M. Alfaro, M. L. Rezola, Orthogonal polynomials on Sobolev spaces: Old and new directions, *J. Comput. Appl. Math.*, **48** (1993), 113–131. https://doi.org/10.1016/0377-0427(93)90318-6

29. E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, et al., A global database of COVID-19 vaccinations, *Nat. Hum. Behav.*, **5** (2021), 947–953. https://doi.org/10.1038/s41562-021-01122-8 https://doi.org/10.1101/2021.03.22.21254100

30. R. Mendoza, S. Keeling, A two-phase segmentation approach to the impedance tomography problem, *Inverse Probl.*, **33** (2016), 015001. https://doi.org/10.1088/0266-5611/33/1/015001

31. D.B. Mumford, J. Shah, Optimal approximations by piecewise smooth functions and associated variational problems, *Commun. Pure Appl. Math.*, **32** (1989), 577–685. https://doi.org/10.1002/cpa.3160420503

32. E. Nixdorf, M. Hannemann, M. Kreck, A. Schoßland, Hydrological records in 5 min resolution of tributaries in the Mueglitz River Basin, Germany, *Data Brief*, 2021. https://doi.org/10.1594/PANGAEA.927729

33. A. Nocon, W. Scott, An extension of the Whittaker-Henderson method of graduation, *Scand. Actuar. J.*, **2012** (2012), 70–79. https://doi.org/10.1080/03461238.2010.534257

34. K. R. O. Recio, R. G. Mendoza, Three-step approach to edge detection of texts, *Philipp. J. Sci.*, **148** (2019). 193–211.

35. J. B. E. Riñon, R. Mendoza, A. A. de los Reyes V, V. Y. Belizario Jr., V. M. P. Mendoza, Management and control of schistosomiasis in Agusan del Sur, Philippines: A modeling study, *Research Square*, 2020.

36. T. J. Rivlin, *Chebyshev polynomials*, Courier Dover Publications, 2020.

37. S. Sharma, V. Kumar, Application of genetic algorithms in healthcare: A review, *Next Gener. Healthc. Inform.*, 2021, 75–86. https://doi.org/10.1007/978-981-19-2416-3_5

38. W. I. Smirnow, *Lehrgang der höheren Mathematik: Teil V*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1967.

39. A. N. Tikhonov, V. Y. Arsenin, *Solutions of ill-posed problems*, Wiley, New York, 1977.

40. L. Tribe, R. Smith, Modeling global outbreaks and proliferation of COVID-19, *SIAM News*, 2020.

41. G. Wahba, *Spline models for observational data*, Society for Industrial and Applied Mathematics, 1990.

42. H. Weinert, Efficient computation for Whittaker-Henderson smoothing, *Comput. Stat. Data An.*, **52** (2007), 959–974. https://doi.org/10.1016/j.csda.2006.11.038

43. H. Yamada, A note on Whittaker-Henderson graduation: Bisymmetry of the smoother matrix, *Commun. Stat.-Theor. M.*, **49** (2020), 1629–1634. https://doi.org/10.1080/03610926.2018.1563183

44. H. Yamada, F. T. Jahra, Explicit formulas for the smoother weights of Whittaker-Henderson graduation of order 1, *Commun. Stat.-Theor. M.*, **48** (2018), 3153–3161. https://doi.org/10.1080/03610926.2018.1476713

AIMS Press