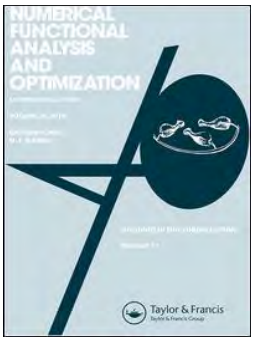

Numerical Functional Analysis and Optimization

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lnfa20

Estimating the Nonparametric Regression Function by Using Padé Approximation Based on Total Least Squares

Syed Ejaz Ahmed , Dursun Aydin & Ersin Yilmaz

NUMERICAL FUNCTIONAL ANALYSIS AND OPTIMIZATION
https://doi.org/10.1080/01630563.2020.1794891

Taylor & Francis
Taylor & Francis Group

Estimating the Nonparametric Regression Function by Using Padé Approximation Based on Total Least Squares

Syed Ejaz Ahmed[a], Dursun Aydin[b], and Ersin Yilmaz[b]

[a]Department of Mathematics & Statistics, Brock University, St. Catharines, Canada; [b]Department of Statistics, Mugla Sitki Kocman University, Mugla, Turkey

ABSTRACT

In this paper, we propose a Padé-type approximation based on truncated total least squares ($P-TTLS$) and compare it with three commonly used smoothing methods: Penalized spline, Kernel smoothing and smoothing spline methods that have become very powerful smoothing techniques in the nonparametric regression setting. We consider the nonparametric regression model, $y_i = g(x_i) + \varepsilon_i$, and discuss how to estimate smooth regression function g where we are unsure of the underlying functional form of g. The Padé approximation provides a linear model with multi-collinearities and errors in all its variables. The $P-TTLS$ method is primarily designed to address these issues, especially for solving error-contaminated systems and ill-conditioned problems. To demonstrate the ability of the method, we conduct Monte Carlo simulations under different conditions and employ a real data example. The outcomes of the experiments show that the fitted curve solved by $P-TTLS$ is superior to and more stable than the benchmarked penalized spline ($B-PS$), Kernel smoothing (KS) and smoothing spline (SS) techniques.

ARTICLE HISTORY
Received 14 August 2019
Revised 8 July 2020
Accepted 8 July 2020

KEYWORDS
Alternative smoothing method; nonparametric regression; Padé approximation; splines; truncated total least squares

1. Introduction

Suppose that we have a response variable $y = (y_{1,}, _{,}y_n)'$ produced by the model

$$y_i = \mathrm{g}(x_i) + \varepsilon_i, a = x_1 < \cdots < x_n = b \tag{1}$$

where" g" is an unknown smooth function, x_i represents the values of covariate, ε_i represents the independent random error terms with zero mean and common variance σ^2, and the symbol $(.)'$ indicates a transposed vector. In this case we focus on a single covariate x_i, however, this model can be extended to include more than one.

CONTACT Ersin Yilmaz yilmazersin13@hotmail.com Department of Statistics, Mugla Sitki Kocman University, 48000 Mugla, Turkey.

The relationships between the variables stated in equation (1) can be modeled by splines over small ranges of the values of covariate x_i. Note also that these splines are widely used in situations where the researcher knows that nonlinear effects occur in the real response function. In the literature, different methods based on splines have been proposed, including penalized splines, additive splines, partial splines, tensor product splines, and thin plate splines. In addition to smoothing using splines, there are other families of smoothing methods such as kernel, wavelet smoothers, and orthogonal series approximations. There has been extensive literature written on the topic of nonparametric regression: for example, general methods [13; Schimek 2000; Hastie et al. 2001], spline smoothing [20, 32], kernel smoothing (Nadaraya 1964; Watson 1964) and local polynomial smoothing (Fan and Gijbels 1996).

The aim of this study is to estimate the unknown regression function using a Padé approximation expressed as the ratio of two polynomial functions. Such an approximation also provides an alternative estimation procedure for nonparametric regression models based on the different smoothing methods covered in the previous paragraph. One of the strong motivations for studying the Padé approach is that it offers a concrete problem that represents functions efficiently by easily computed expression. It should be also noted that the Padé approximation has a numerical approach that works directly on data; there is no need to convert to a Fourier (or other) domain. Furthermore, the problem is converted to linear least squares fits by removing the denominator of a rational function. Finally, the method provides direct control of the coefficients in the rational approximation, permitting for restrictions to be placed on certain terms as motivated by the physics of the model being studied.

For these reasons, it is a very useful task to attempt using the Padé method on a data set that belongs to a nonparametric regression setting. There is has been extensive literature written on the general methods of the Padé approximation, including [3, 4, 6, 7, 10, 19, 37]. In addition, the connections between rational functions and splines can be found in Petrushev and Popov (1987). Petrushev and Papov also demonstrate that the rational functions are not worse than splines as tool approximation. Moreover, the use of rational functions introduces a nonlinear problem and requires an iterative procedure rather than a direct procedure, as in linear least squares problems. Our point of view is that splines are the well-known nonlinear tool for approximation and therefore it is very useful to investigate the connections between rational and spline approximations of functions.

As indicated in the above, we mainly consider a Padé approach to find a better approximation of the unknown regression function $g(x)$ expressed in the model (1). The method used in initial computations leads to an ill-conditioned problem. This is only related to the formulation of this procedure, as shown in the equation (3). Moreover, in this system, both the input data matrix and the response observations are contaminated by error and noise.

For these cases, as a solution technique, called total least squares (TLS) method, is devised by Golub and Van Loan (1980). Note also that in statistics literature this technique is sometimes known as an errors-invariables (EIV) modeling or orthogonal regression. An extension covering the randomized truncated TLS with the known or estimated rank as the regularization parameter are recently introduced by [38] for the large-scale and ill-conditioned cases. It should be noted that the literature on TLS and their extensions focuses mostly on solution and numerical algorithms [see Van Huffel and Vandewalle 1991; Cheng and Van Ness 2000; 36, 38 among many others]. It is worthwhile to note that here, although the TLS method is suggested as a basic approximation way of overdetermined system 2, it cannot to solve the ill-conditioned problem caused by the relationship of rational function terms.

To overcome multi-collinearity problem and to get a stable estimation of Padé coefficients, we used a truncated total least squares (*TTLS*) method. For convenience, this estimation procedure is hereafter referred to as the "$P-TTLS$" method, as indicated in our abstract. It can also be viewed as a new technique of least squares from minimization problem with regularization. We also compare the performance of the $P-TTLS$ method with that of a p^{th}-degree penalized spline with truncated polynomial basis, as a benchmark method (i.e., $B-PS$), and *KS* and *SS* methods. To our knowledge, such a study has not yet carried out for this purpose. However, many authors have used the Padé approximation (or rational approximation) in numerical modeling. Ref. [39] used the Padé approximations for identification of air bubble volume. These authors used this approximation for the problem of estimation of microstructural parameters in finely-structured heterogeneous mixtures in the following year, [40]. In the study conducted by [41], the Padé approximation is considered as a numerical inversion method for the estimation of the quality Q factor and phase velocity in linear, viscoelastic, and isotropic media using the reconstruction of relaxation spectrum. Also recently, [1] studied the selection of optimum truncation parameter for estimation of the nonparametric regression model based on Padé approximation.

Our paper is organized as follows: In section 2, necessary fundamental information are given for the proposed $P-TTLS$ method and the nonparametric regression model. In Section 3, the Padé approximation is introduced and the solution to the total least squares problem and a regularized solution are given. Also, $P-TTLS$ estimator is obtained and an algorithm for that is provided. In section 4, the penalized spline is discussed. Section 5 reviews the smoothing parameter selection criteria. Statistical properties of the coefficients' estimates are defined in section 6. Simulation experiments are carried out in section 7, and regularization methods are applied to four different real data sets in section 8. Finally, the conclusions and recommendations are presented in the last section.

2. Preliminaries

A basic task of the Padé approximation is to determine an estimate of the unknown coefficients vector β from particular measurements of the variables. However, an approximation of this type gives rise an overdetermined system of equations just like in (13). Conceptually, this system is convenient to re-express in the equivalent form

$$\mathbf{X}\beta \approx \mathbf{y} \tag{2}$$

where $\mathbf{X} \in R^{n\times\mathbf{m}}(n>m)$ is data matrix, $\mathbf{y} \in \mathrm{R}^{n\times 1}$ is the response vector, and $\beta \in \mathrm{R}^{m\times 1}$ is the Padé coefficients vector to be estimated, as expressed in (13). Note also that $\mathbf{X}$ and $\mathbf{y}$ are known and assumed to be error contaminated. This problem is referred to as the total least squares (TLS) problem [see Ref. 17].

In this section, the main goal is to focus on a regularization (or a stabilization) technique in order to solve the ill-posed problem (14). However, before exploring these issues, we must first discuss singular value decomposition (SVD) and its variants that from a basis for the TLS problems. We first discuss singular value decomposition (SVD) and its variants that form a basis for the TLS problems.

Theorem 2.1. *(SVD). If* $\mathbf{X}$ *is an* $n\times m$ *matrix, then there exist orthogonal matrices* $\tilde{\mathbf{U}} = (\tilde{u}_1, \ldots, \tilde{u}_n) \in R^{n\times n}$ *and* $\tilde{\mathbf{V}} = (\tilde{v}_1, \ldots, \tilde{v}_n) \in R^{m\times m}$ *such that*

$$\tilde{U}X\tilde{V} = \tilde{\Sigma} = \mathrm{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_k) \in R^{n\times m}, k = \min(n, m), \tag{3}$$

where $\tilde{\mathbf{U}}'\tilde{\mathbf{U}} = \tilde{\mathbf{V}}'\tilde{\mathbf{V}} = \mathbf{I}$ *and* $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_k \geq 0$

Proof. See ref. [18].

Note that the positive diagonal entries in $\tilde{\Sigma}$ are called singular values of $\mathbf{X}$. The singular values are the square roots of the eigenvalues of the square matrices $\mathbf{X}'\mathbf{X}$ or $\mathbf{X}\mathbf{X}'$. The number of these singular values is also equal to the rank of $\mathbf{X}$. If the $rank(\mathbf{X}) = r$, we have $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \ldots \geq \tilde{\sigma}_k \geq 0$, and if $r \leq min(nim)$ then $\tilde{\sigma}_{r+1} = \cdots = \tilde{\sigma}_k = 0$. This means that SVD allows us to define a cutoff point r for a given an $n\times m$ matrix X such that

$$\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_r > \tilde{\sigma}_{r+1} = \cdots = \tilde{\sigma}_k = 0, k = \min(n, m)$$

In this case,

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \tag{4}$$

where $\tilde{\Sigma}_r = \mathrm{diag}(\tilde{\sigma}_1, ..., \tilde{\sigma}_r)$. It can easily be seen that the diagonal matrix $\tilde{\Sigma}$ has k entries on the diagonal, but the $(k-r)$ of these entries equal zero. If the matrix $\mathbf{X}$ has an additional $(k-r)$ zero singular values, then this matrix is not full-rank. Also, if the rank of diagonal matrix $\tilde{\Sigma}$ equals the number of nonzero diagonal elements, then, $\mathrm{rank}(\mathbf{x}) = \mathrm{rank}(\tilde{\mathbf{\Sigma}}) = r$.

Theorem 2.2. *The SVD of the $n \times (m+1)$ augmented matrix $[\mathbf{X}, \mathbf{y}]$ can be defined by*

$$[\mathbf{X}, \mathbf{y}] = \mathbf{U}\Sigma\mathbf{V}' \tag{5}$$

where $\mathbf{U}$ is an $(n \times n)$ orthogonal matrix and satisfies $\mathbf{U}'\mathbf{U} = \mathbf{I}$, $\mathbf{V}$ is an $(m+1) \times (m+1)$ orthogonal matrix and satisfies $\mathbf{V}'\mathbf{V} = \mathbf{I}$, and Σ is an $n \times (m+1)$ diagonal matrix with nonnegative entries such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{m+1} >= 0$.

Proof. See proof of theorem 2.1.

Corollary 2.1. *If $[\mathbf{X}, \mathbf{y}]$ is an $n \times (m+1)$ augmented matrix and $r = rank\{[\mathbf{X}, \mathbf{y}]\}$, then*

$$[\mathbf{X}, \mathbf{y}] = \mathbf{U}diag(\sigma_1, \sigma_2, ..., \sigma_r)\mathbf{V}' = \sum_{i=1}^{r} \sigma_i u_i v_i' \tag{6}$$

where diag(.) is the diagonal matrix, and $\mathbf{U}$ and $\mathbf{V}$ are the orthogonal matrices, as defined in Theorem 2.2. Moreover, the column vectors of $\mathbf{U} = (u_1, , u_n)$ are called the left singular vectors (or unitary) and the vectors of $\mathbf{V} = (v_1, , v_n)$ are called the right singular vectors.

Proof. Suppose $[\mathbf{X}, \mathbf{y}] = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}'$ is the SVD of $[\mathbf{X}, \mathbf{y}]$ and let $r = rank\{[\mathbf{X}, \mathbf{y}]\}$ Considering equation 8, we have

$$\begin{aligned} [\mathbf{X}, \mathbf{y}] &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}' = \mathbf{U}\begin{bmatrix} \mathbf{\Sigma}_{r\times r} & 0 \\ 0 & 0 \end{bmatrix}\mathbf{V}' \\ &= (u_1, ..., u_n)\begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \ddots & 0 & \dots & 0 \\ \vdots & 0 & \sigma_r & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}\begin{pmatrix} v_1' \\ \vdots \\ v_{m+1}' \end{pmatrix} = \sigma_1\mathbf{u}_1\mathbf{v}_1' + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r' \end{aligned}$$

completing the proof of Corollary 2.1.

Equation (6) is a dyadic form of the SVD. This form plays an important role in applications where a matrix is approximated with a lower rank. It should also be noted that we need a suitable matrix norm in order to measure the size of the errors in the matrix of the linear system given in the equation (2). In this context, the Frobenius norm is the commonly used matrix norm in TLS problems. For a matrix $X = (x_{ij})$ it is defined as

$$||\mathbf{X}||_F = \sqrt{\sum_{i,j} x_{ij}^2} = \sqrt{\mathrm{tr}(\mathbf{X}'\mathbf{X}} = \sqrt{\sum_{i=1}^{k} \tilde{\sigma}_i^2} \tag{7}$$

Furthermore, in this kind of application, the following Eckart-Young-Mirsky theorem provides a convenient solution to the problem of approximating a matrix by another of lower rank.

Theorem 2.3. ***(Eckart-Young-Mirsky matrix approximation theorem)****. Let the SVD of* $\mathbf{X} \in R^{n\times m}$ *be given by* $\mathbf{X} = \sum_{i=1}^{r} \tilde{\sigma}_i \tilde{u}_i v_i'$ *with* $r = rank(\mathbf{X})$*. If* $h<r$ *and* $\mathbf{x}_h = \sum_{i=1}^{h} \tilde{\sigma} u_i v_i'$*, then*

$$\min_{\mathrm{rank}(Z)=h} ||\mathbf{X}-\mathbf{Z}||_2 = ||\mathbf{X} - \mathbf{X}_h||_2 = \tilde{\sigma}_{h+1} \tag{8}$$

and

$$\min_{\mathrm{rank}(Z)=h} ||\mathbf{X}-\mathbf{Z}||_p = ||\mathbf{X} - \mathbf{X}_h||_F = \left(\sum_{i=h+1}^{k} \tilde{\sigma}_i^2 \right)^{1/2} \quad \text{where } k = \min(n, m) \tag{9}$$

Proof. See refs. [12], [16] and [24].

3. Padé-type approximation

Consider the following approximation problem. The key idea is to approximate the unknown regression function $g(x)$ in *refeq11* by the function of the form $g_{[p,q]}(x) = A(x)/B(x)$ where

$$\begin{aligned} A(x) &= a_0 + a_1x + a_2x^2 + \cdots + a_px^p \\ B(x) &= b_0 + b_1x + b_2x^2 + \cdots + b_qx^q \end{aligned} \tag{10}$$

From (10), it is clear that an approximation of function $g(x)$ is also a rational function approximation, $g_{[p,q]}$ (x) of order $(p + q + 1)$. In this sense, there are $(p + q + 1)$ coefficients to be estimated. It should also be noted that $g(x)$ is a continuous function defined by the intervals $[a, b]$. A main problem here is to estimate, for fixed degree p and q, the coefficients of the real polynomials $A(x)$ and $B(x)$ so that the absolute error of this approximation, $\left|\mathrm{g}(x) - g_{[p,q]}(x)\right| \leq \varepsilon$, is the smallest possible.

If function $g(.)$ is given by measured data pairs (x_i, y_i) then the Padé approximation can be made through the following methods: For each of the data points and the continuous function $g(.)$, and the integers p and q, the Padé approximation can be written as

$$y_i = g(x_i) \cong \frac{a_0 + a_2 x_i + a_2 x_i^2 + \cdots + a_p x_i^p}{1 + b_1 x_i + b_2 x_i^3 + \cdots + b_q x_i^q} = g_{[p,q]}(x_i), 1 \leq i \leq n \qquad (11)$$

where $a_j (j = 0, 1, ..., p)$ and $b_k (k = 0, 1, ..., q)$ are the unknown coefficients to be estimated from the data. It must be noted that the constant coefficient $b_0 = 1$ in the denominator that allows us to determine the non-zero poles of $g_{[p,q]}(x_i)$. The most useful of the Padé approximations are those with order of the numerator equal to, or one greater than, the degree of denominator.

The equation (3.2) above equivalently can be rewritten as,

$$\begin{aligned}&\left[a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_p x_i^p - b_1 y_i x_i - b_2 y_i x_i^2 - \cdots - b_q y_i x_i^q\right] \\ &\quad \cong y_i, 1 \leq i \leq n\end{aligned} \qquad (12)$$

It is clear that equation (12) produces a system of linear equations in terms of n observations. The mentioned system can be expressed using matrix and vector notation as

$$\{\mathbf{g} = \mathbf{X}\beta\} = \begin{bmatrix} 1 & x_1 & \ldots & x_1^p & -y_1 x_1 & -y_1 x_1^2 & \ldots & -y_1 x_1^q \\ 1 & x_2 & \ldots & x_2^p & -y_2 x_2 & -y_2 x_2^2 & \ldots & -y_2 x_2^q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & x_n & \ldots & x_n^p & -y_n x_n & -y_n x_n^2 & \ldots & -y_n x_n^q \end{bmatrix}_{(mx1)} \begin{bmatrix} a_0 \\ \vdots \\ a_p \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$$

$$= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n, x_1)} \cong \mathbf{y} \qquad (13)$$

Note that to uniquely determine the coefficients $\beta = (a_j, b_k)'$, the sample size should be at least as large as the number of coefficients, i.e., $n > m$ (where $m = p + q + 1$). From the above matrix $\mathbf{X}$ we can see that the natural polynomials x_i^j's $(j = 0, 1, ..., p)$ are nearly linearly dependent, which leads to $\mathbf{X}'\mathbf{X}$ being close to singular. This case implies that the matrix $\mathbf{X}$

has an ill-conditioned problem, so that the solution is exceedingly sensitive to perturbations. As mentioned earlier, the main consideration in this article is to estimate the nonparametric regression function **g** using the Padé approach. A slight advantage is that by its nature, a Padé-type approximant to function $\mathbf{g}(.)$ leads to the linear equation systems $\mathbf{g} = \mathbf{X}\beta$, as expressed in (13). However, it should be emphasized that the singularity of $\mathbf{X}'\mathbf{X}$ matrix causes this system to become unsatisfactory. In this case, the regression function $\mathbf{g}(.)$ is said to be suffering from the problem of multicollinearity.

To obtain a stable solution, one needs a regularization method that replaces the above problem with a well-posed problem. For these purposes, the most commonly used regularization methods are discussed in the Appendix, but there are limitations on their uses. For example, these regularization methods assume that the errors are confined to the right-hand side (response vector **y**) of the equation $\mathbf{X}\beta \approx \mathbf{y}$. Unfortunately, in our study, we also need to consider the errors (or perturbations) occurring in the data matrix **X** with ill-conditioned problems. Therefore there is a need for using a regularization method that takes the errors on both sides of equation (13) into account. To overcome this problem, this paper applies the TTLS problem introduced by ref. [17], in the field of computational mathematics.

3.1. Solution of padé coefficients based on the TLS method

Assume an over-determined system $\mathbf{X}\beta \approx \mathbf{y}$ in which both the matrix **X** and the observation vector **y** are subject to errors. The basic objective of the TLS method is to find a solution to the unknown Padé coefficients vector β that minimizes the Frobenius norm of the errors

$$\min_{[\hat{\mathbf{X}},\hat{\mathbf{y}}] \in R^{n\times(m+1)}} ||[\mathbf{X},\mathbf{y}] - [\hat{\mathbf{X}},\hat{\mathbf{y}}]||_F \text{ subject to } \hat{\mathbf{y}} \in R(\hat{\mathbf{X}}) \tag{14}$$

where $\hat{\mathbf{y}}$ is the smallest possible perturbation of **y** that lies in the range (column space) $R(\mathbf{X})$ of **X**, and $\hat{\mathbf{X}}$ is the subject to noise part of **X**. In contrast to this, the ordinary least squares method requires that $\mathbf{X} = \hat{\mathbf{X}}$, and minimizes the 2-norm of the residual vector $(\mathbf{y} - \hat{\mathbf{y}})$.

It should be noted here that we seek $\hat{\mathbf{X}}$ and $\hat{\mathbf{y}}$ such that equation (14) is as small as possible. $\hat{\mathbf{X}}\beta = \hat{\mathbf{y}}$ is thus a TLS solution to the equation (2), assuming that the random variables $\hat{\mathbf{y}}$ and $\hat{\mathbf{X}}$ are measured with errors, so we only observe

$$\hat{\mathbf{y}} = \mathbf{y} + \varepsilon_y \text{ and } \hat{\mathbf{X}} = \mathbf{X} + \varepsilon_x \tag{15}$$

where ε_y and ε_x satisfy

$$\min_{[\varepsilon_y,\varepsilon_x]\in R^{n\times(m+1)},\beta\in R^m} \|\varepsilon_x,\varepsilon_y\|_F \text{ subject to } \mathbf{y}+\varepsilon_y=(\mathbf{X}+\varepsilon_x)\beta \tag{16}$$

[see ref. 16]. Here ε_y is the error of response vector $\mathbf{y}$ and ε_x is the error of the matrix $\mathbf{X}$. One may notice that (14) and (16) are equal, and once a minimizing $[\hat{\varepsilon}_y,\hat{\varepsilon}_x]$ is found, then any vector β satisfying $\mathbf{y}+\hat{\varepsilon}_y=(\mathbf{X}+\hat{\varepsilon}_x)\beta$ is stated as a solution of Padé coefficients vector β based on TLS. The following theorem provides a basic solution for such problems:

Theorem 3.1. ***Solution of the basic TLS problem***. *Let (5) be the SVD of $[\mathbf{X},\mathbf{y}]$ with unitary matrices* $[\mathbf{U},\ \mathbf{V}]$ *and rectangular diagonal matrix $\boldsymbol{\Sigma}$. Here, $\tilde{\sigma}_m$ is the smallest singular value of $\mathbf{X}$. If $\tilde{\sigma}_m>\sigma_{m+1}$, then*

$$[\hat{\mathbf{x}},\hat{\mathbf{y}}]=\left[\mathbf{X}+\hat{\varepsilon}_x,\mathbf{y}+\hat{\varepsilon}_y\right]=\mathbf{U}\hat{\Sigma}\mathbf{V}' \text{ with } \hat{\Sigma}=\operatorname{diag}(\sigma_1,\ldots,\sigma_m,0) \tag{17}$$

with the corresponding TLS correction matrix

$$\left[\hat{\varepsilon}_x,\hat{\varepsilon}_y\right]=[\mathbf{X},\mathbf{y}]-\left[\hat{\mathbf{X}},\hat{\mathbf{y}}\right]=\tilde{\sigma}_{m+1}u_{m+1}v'_{m+1} \tag{18}$$

solving the TLS problem

$$\beta_{P-TLS}=-\frac{1}{v_{m+1,m+1}}\left[v_{1,m+1},\ldots,v_{m,m+1}\right]' \tag{19}$$

exists and is the unique solution $\hat{\mathbf{X}}\beta=\hat{\mathbf{y}}$

Proof. (See Van Huffel and Vandewalle (1991), for a complete proof). Rewrite equation (13) in the form

$$[\mathbf{X},\mathbf{y}]\left[\beta',-1\right]\approx 0 \tag{20}$$

If $\sigma_{m+1}\neq 0,\operatorname{rank}\{[\mathbf{X},\mathbf{y}]\}=m+1$ then there is no nonzero vector in the null space of the matrix [**X**, **y**]. In order to find an appropriate vector, the rank of matrix [**X**, **y**] should be reduced to m. Using the Eckart-Young-Mirsky Theorem (4), the best *m* – *rank* TLS approximation $[\hat{\mathbf{X}},\hat{\mathbf{y}}]$ of [**X**, **y**], which minimizes the deviations in variance, is obtained by setting the smallest singular value of [**X**, **y**] to zero, $\sigma_{m+1}=0$. The minimal correction is then

$$\sigma_{m+1}=\min_{rank[\hat{\mathbf{x}},\hat{y}]=m}\left\|[\mathbf{X},\mathbf{y}]-\left[\hat{\mathbf{X}},\hat{\mathbf{y}}\right]\right\|_F=\|\hat{\varepsilon}_x,\hat{\varepsilon}_y\|_F \text{ where } \left[\hat{\mathbf{X}},\hat{\mathbf{y}}\right]=\sum_{i=1}^{m}\sigma_i u_i v'_i \tag{21}$$

and is attained for TLS correction matrix (18) which has rank one. Now, $\mathbf{v}_{(m+1)}$ is a vector in the null space of $[\hat{\mathbf{X}},\hat{\mathbf{y}}]$, and the approximate equation $[\hat{\mathbf{X}},\hat{\mathbf{y}}]\left[\beta',-1\right]=0$ is compatible for some β. The TLS solution is then obtained by normalizing $\mathbf{v}_{(m+1)}$ until last element is -1.

Remark 3.1. A s discussed in Van Huffel and Vandewalle (1991),

$$\tilde{\sigma}_m > \sigma_{m+1} \leftrightarrow \sigma_m > \sigma_{m+1} \text{ and } v_{m+1,m+1} \neq 0$$

Remark 3.2. If $\sigma_{m+1} = 0$ and $\text{rank}\{[\mathbf{X}, \mathbf{y}]\} = m \rightarrow$ equation (20) is compatible and any approximation is not required to obtain the exact solution stated in (19).

Remark 3.3. If $v_{m+1,m+1} \neq 0$ the TLS problems (14) and (16) are solvable and are therefore generic.

According to Remark 3.3, it is important to note that the fitted values (indexed by $\hat{\mathbf{y}}_{PTLS}$ are obtained in the following way ($\hat{\beta}_{PTLS}$ denotes the Padé coefficients estimated by TLS):

$$\hat{\mathbf{y}}_{P-TLS} = \hat{\mathbf{X}}\hat{\beta}_{P-TLS} = -\frac{1}{v_{m+1,m+1}}\hat{\mathbf{X}}[v_{1,m+1}, \ldots, v_{m,m+1}]' \in R(\hat{\mathbf{X}}) \tag{22}$$

where $\mathbf{v}_{(m+1,m+1)}$ is the $(m+1)^{th}$ component of the last column vector $\mathbf{v}_{(m+1)}$ of V that belongs to the null space of $[\hat{\mathbf{X}}, \hat{\mathbf{y}}]$. Hence $\hat{\beta}_{P-TLS}$ provides a solution to problem (14). After reconstruction of the coefficient vector $\hat{\beta}_{P-TLS}$ of the rational function approximation, the fitted values expressed in (22) will also give Padé-type approximation of function g(x).

Theorem 3.2. ***(Closed-form TLS solution).*** *Let (3) and (5), be the SVD of* $\mathbf{X}$ *and* $[\mathbf{X}, \mathbf{y}]$*, respectively. If* $\tilde{\sigma}_m > \sigma_{m+1}$*, then*

$$\hat{\beta}_{P-TLS} = \left(\mathbf{X}'\mathbf{X} - \tilde{\sigma}^2_{m+1}\mathbf{I}_m\right)^{-1}\mathbf{X}'\mathbf{y} \tag{23}$$

Proof. For proof, See ref. [35].

Our focus in this study is on very ill-conditioned problems. Since the data matrix $\mathbf{X}$ is ill-conditioned, the augmented matrix $[\mathbf{X}, \mathbf{y}]$ is also ill-conditioned as a direct result. In this case, the TLS problem is unstable whenever $\tilde{\sigma}_m$ is close to $\tilde{\sigma}_{m+1}$. Hence the TLS estimators given in the equations (19) and (23) often yield unstable solutions, and a regularization approach is necessary to stabilize them.

3.2. Estimation of padé coefficients by P-TTLS method

As noted in Section 3.1, the TTLS method is commonly used to solve linear ill-conditioned problems in the presence of measurement errors. The most important idea underlying the TTLS method is that one neglects the smaller singular values of the augmented matrix $[\mathbf{X}, \mathbf{y}]$ [see 15, 30, for more thorough discussions]. Roughly speaking, the TTLS method aims to reduce the contribution of errors by cutting off a certain number of singular values in the SVD of the augmented data matrix. The mechanism of the TTLS method is a similar to a truncated SVD, a generalization of the ordinary least squares method for nearly rank-deficient problems (see Appendix for more detail). To obtain a stable solution vector β of Padé coefficients for (2), the ideas of describing TTLS method are expressed with the following algorithm:

3.2. Algorithm TTLS

1. Compute the SVD of the augmented matrix $[\mathbf{X}, \mathbf{y}]$, as described in the equation (5):
   $$[\mathbf{X}, \mathbf{y}] = \mathbf{U}\Sigma\mathbf{V}' \text{ with diagonal elements } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{m+1} \geq 0$$
2. Select an appropriate truncation (or regularization) parameter $t \leq m = rank[\mathbf{X}, \mathbf{y}]$ (i.e., the number of maintained the singular values of the matrix $[\mathbf{X},\mathbf{y}]$).
3. Block-partition the $(m+1) \times (m+1)$ matrix $\mathbf{V}$ such that
   $$V = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}, \text{ where } \mathbf{V}_{11} \in R^{m\times t} \text{ and}$$
   $$\left\{\mathbf{V}_{22} \equiv \left[v_{m+1,t+1}, \ldots, v_{m+1,m+1}\right] \neq 0\right\} \in R^{1\times(m+1-t)}$$
4. Compute the P-TTLS solution as
   $$\hat{\beta}^t_{P-TTLS} = -\mathbf{V}_{12}(\mathbf{V}_{22})^+ = -\mathbf{V}_{12}\frac{\mathbf{V}'_{22}}{\|\mathrm{V}_{22}\|_2^2} \tag{24}$$

where $\mathbf{V}_{22})^+$ denotes the pseudoinverse of $\mathbf{V}_{22}$ and $\hat{\beta}^t_{P-TTLS}$ shows the estimates of Padé coefficients.

If $t = m$, then it is clear that the equation (19) has been obtained. The 2-norm of (24) and the equivalent Frobenius norm of TLS residuals are defined as, respectively:

$$\|\hat{\beta}^t_{P-TTLS}\|_2 = \sqrt{\|\mathbf{V}_{22}\|_2^{-2}-1} \text{ and } \|[\mathrm{X}, \mathbf{y}]-\left[\hat{\mathbf{X}}, \hat{\mathbf{y}}\right]\|_F = \sqrt{\sigma^2_{t+1} + \cdots + \sigma^2_{m+1}} \tag{25}$$

As stated in [15], the aforementioned equations denote that $\|\hat{\beta}^t_{P-TTLS}\|_2$ increases with t, while the norm of the TLS residual decreases with t. Note also that the small singular values of $[\mathbf{X}, \mathbf{y}]$ are neglected with the help of a chosen regularization parameter t. For these reasons, it is very important to select an appropriate regularization (or truncated) parameter t. In the context of this paper, the generalized cross validation criterion (*GCV*) is used to find this parameter.

Alternatively, analogous to the case presented in equation (24), the solution vector β^t_{P-TTLS} may be easily expressed. Firstly, we define $\mathbf{P}_t$ to be the orthogonal projection onto a t dimensional subspace of $[\mathbf{X}, \mathbf{y}]$, given by

$$\mathbf{P}_t = \mathbf{U}_t\mathbf{U}_t' = (u_1, ..., u_n)(u_1, ..., u_n)'$$

where $\mathbf{U}_t$ denotes the first t columns of the orthogonal matrix $\mathbf{U}$ defined in (5). In this case, for the reduced rank system $\mathbf{P}_t[\mathbf{X}, \mathbf{y}]$ we can obtain a useful alternative P – $TTLS$ solution minimizes the $||[\mathbf{X}, \mathbf{y}] - [\hat{\mathbf{X}}, \hat{\mathbf{y}}]||_F$ such that

$$\hat{\beta}^t_{P-TTLS} = (\mathbf{P}_t\mathbf{X})^+\mathbf{P}_t\mathbf{y} \tag{26}$$

Note that arguments similar to those used in solution (22) show that the fitted (or predicted) values at training inputs are

$$\hat{\mathbf{y}}_{P-TTLS} = \mathbf{X}\hat{\beta}^t_{P-TTLS} = \mathbf{X}(\mathbf{P}_t\mathbf{X})^+\mathbf{P}_t\mathbf{y} = \mathbf{H}_t\mathbf{y} \tag{27}$$

where $\mathbf{H}_t = \mathbf{X}(\mathbf{P}_t\mathbf{X})^+\mathbf{P}_t$ is called "the hat matrix" because it transforms real response vector $\mathbf{y}$ into the fitted observations vector $\hat{\mathbf{y}}$. The matrix $\mathbf{H}_t$ computes the orthogonal projection, and hence it is also known as a projection matrix. From equation (27) above, it should be emphasized that the Padé coefficients estimated by the TTLS method are linear smoothers, and therefore the vector of fitted values (27) can be written as

$$\hat{\mathbf{y}}^t_{P-TTLS} = \left(\hat{g}^t_{P-TTLS}(x_1), ..., \hat{g}^t_{P-TTLS}(x_n)\right)' = \hat{\mathbf{g}}^t_{P-TTLS} = \mathbf{H}_t\mathbf{y} \tag{28}$$

For computational and conceptual simplicity, an alternative formulation of the solution (24) can be expressed in terms of filter factor, as in the TSVD solution (A.5) given in the Appendix. The main idea is to write a expression for the $\hat{\beta}^t_{P-TTLS}$ based on the SVD of $\mathbf{X}$, rather than of the SVD of $[\mathbf{X}, \mathbf{y}]$. The filter factor formulation used for TTLS solution can be given by

$$\hat{\beta}^t_{P-TTLS} = \sum_{i=1}^{m} f_i \frac{u_i\mathbf{y}}{\tilde{\sigma}_i} v_i, f_i = \sum_{j=t+1}^{m} \frac{v^2_{m+1.j}}{\|\mathbf{V}_{22}\|^2}\left(\frac{\gamma_i^2}{\sigma_i^2 - \sigma_j^2}\right) = \sum_{j=1}^{t} \frac{v^2_{m+1,j}}{\|\mathbf{V}_{22}\|^2}\left(\frac{\tilde{\sigma}_i^2}{\tilde{\sigma}_i^2 - \sigma_j^2}\right) \tag{29}$$

where $\tilde{\sigma}_i \neq \sigma_j$, f_i's are the filter factor values, the numbers $\tilde{\sigma}_i$ are the nonzero singular values of $\mathbf{X}$ while the quantities σ_j's are the nonzero singular values of $[\mathbf{X}, \mathbf{y}]$, and $\mathbf{v}_{m+1,j}$ indicates $(m+1, j)^{th}$ element of $\mathbf{V}_{22}$ defined in the third step of algorithm TTLS [see 15, for more information].

It follows from equation (29) that it is provided a valid expression for $\hat{\beta}^t_{P-TTLS}$ with truncated parameter t. Moreover, the filter factor formulation shows that the SVD elements of $\hat{\beta}^t_{P-TTLS}$ associated with the smallest singular values σ_i are indeed filtered out.

3.3. Multi-collinearity problem and P-TTLS solution

The $P - TTLS$ method is a generalized version of the ordinary TTLS, and it is motivated by linear approximation problem

$$\mathbf{X}\beta \approx \mathbf{y}, \mathbf{X} \in R^{n\times \mathbf{m}}(n>m), \mathbf{y} \in \mathbf{R}^{n\times 1}, \beta \in \mathbf{R}^{m\times 1} \tag{30}$$

where the matrix $\mathbf{X}$ and the vector $\mathbf{y}$ have errors, as discussed earlier. Note that the linear system (30) is also known as the EIV model in the statistical literature and there are many approaches that are closely associated with the solution of this system. The main difficulty here is that the system (30) is ill-conditioned, here matrix $\mathbf{X}$ is often numerically rank deficient and contains small singular values. In such cases, the ordinary TLS or similar techniques can give an unstable solution due to the dominance of errors in the data. In this sense, regularization techniques based on truncation for TLS problems are considered to obtain a stable solution. Note that truncation techniques used to reduce the dimension of the linear equations system (30) can also be considered as a type of regularization.

From Algorithm TTLS, we see that the $P - TTLS$ can be viewed as an extended version of TTLS method used for the regularization of ill-conditioned linear systems (30). Notice that the SVD of the matrix $\mathbf{X}$ gives us some additional insight into the nature of an ill-conditioned regression problem. For example, small singular values of $\mathbf{X}$ identify multi-collinearities, as mentioned in the previous paragraph. The basic idea here is to limit the contribution of noise or errors by cutting a certain number of terms in an expansion such as SVD. In other words, in the $P - TTLS$ method the key idea is to truncate the small singular values of $(\mathbf{X}, \mathbf{y})$, by setting these values to zero. This really means that the augmented matrix $(\mathbf{X}, \mathbf{y})$ is reduced to a rank $t \leq m$ matrix and the minimum norm solution of the truncated problem $\mathbf{X}_t\beta \approx \mathbf{y}_t$ is obtained by

$$\hat{\beta}^t_{P-TTLS} = -\mathbf{V}^t_{12}\frac{\left(\mathbf{V}^t_{22}\right)'}{\|\mathbf{V}^t_{22}\|^2_2} \tag{31}$$

called $P-TTLS$ solution in nonparametric regression, as defined in (24) or (26).

As defined in step 3 of Algorithm TTLS, $\mathbf{V}^t_{22} \neq 0$ is a required condition. Also, the vector $\mathbf{V}^t_{22}$ is zero in nongeneric cases (see Van Huffel and Vandewalle 1991 for more details), but some elements of this vector can be almost 0 in close-to-nongeneric cases, that can happen in ill-conditioned problems. As we indicated earlier, the truncation parameter t can always be reduced sufficiently so that the vector $\mathbf{V}^t_{22}$ has large enough norm. For these purposes, a truncation level t must be carefully selected. Not that this task is carried out by GCV criterion defined in section 5.

In the light of the detailed information above, it can be said that the modified truncated TLS solution (or $P - TTLS$) stabilizes the TLS solution in nonparametric regression models when multi-collinearities are exist in the data matrix $[\mathbf{X}, \mathbf{y}]$.

4. Penalized spline method

The unknown function $g(x)$ can be well approximated by a p^{th} degree regression spline model with a truncated power basis

$$g(x;\beta) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k}(x - \kappa_k)_+^p \tag{32}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_{p+K})'$ is a vector of unknown coefficients to be estimate, $p \geq 1$ is an integer $(x - \kappa_k)_+^p = (x - \kappa_k)^p$, if $x > \kappa_k$ otherwise zero, and $\{\kappa_1, \ldots, \kappa_K\}$ is a set of fixed knots

$$\{\min(x) \leq \kappa_1 \leq \cdots \leq \kappa_K \leq \max(x)\}$$

[see ref. 28].

Using the above truncated polynomial, it follows that model (1) can be re-written as

$$y_i = \left(g(x;\beta) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k}(x - K_k)_+^p \right) + \varepsilon_i, 1 \leq i \leq n \tag{33}$$

where ε_i represents the random error terms with zero mean and variance σ^2. In matrix and vector form, model (11) can be stated as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{34}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & \ldots & x_1^p & (x_1 - k_1)_+^p & \ldots & (x_1 - k_K)_+^p \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_2 & \ldots & x_n^p & (x_1 - k_1)_+^p & \ldots & (x_n - k_K)_+^p \end{bmatrix} \text{ and}$$

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{(p+1)\times(p+1)} & \mathbf{0}_{(p+1)\times 1} \\ \mathbf{0}_{K\times(p+1)} & \mathbf{I}_{K\times K} \end{bmatrix}$$

and $\mathbf{y}$ is a vector, as defined in (1). Then, the penalized spline (PS) estimates of the vector β are obtained by minimizing the penalized least squares problem

$$\min\left\{\sum_{i=1}^{n}(y_i-g(x_i))^2+\lambda\sum_{k=1}^{K}\beta_{p+k}^2=||\mathbf{y}-\mathbf{X}\beta||^2+\lambda\beta'\mathbf{D}\beta\right\} \tag{35}$$

Here, the regularization parameter $\lambda>0$ controls the weight given to minimization of the penalty term $\beta'\mathbf{D}\beta$, which is also known as the regularization term. In general, large values of λ produce smoother estimators, while smaller values produce wigglier estimators. As can be seen here, the parameter λ plays a key role in estimating the parameters of the nonparametric model (1).

The *PS* estimates are obtained by setting derivatives with respect to β equal to zero, giving the form

$$\hat{\beta}_{B-PS}^{\lambda}=(\mathbf{X}'\mathbf{X}+\lambda\mathbf{D})^{-1}\mathbf{X}'\mathbf{y} \tag{36}$$

where $\hat{\beta}_{B-PS}^{\lambda}$ indicates the coefficients of the truncated power basis function estimated using the penalized spline method. The resulting estimated $\hat{\beta}_{B-PS}^{\lambda}$ can be used to provide the corresponding fitted values (indexed by $\hat{\mathbf{y}}_{B-PS}^{\lambda}$ for $g(x)$:

$$\hat{\mathbf{g}}_{B-PS}^{\lambda}=\left(\hat{g}_{B-PS}^{\lambda}(x_1),\ldots,\hat{g}_{B-PS}^{\lambda}(x_n)\right)'=\mathbf{X}\beta_{B-PS}^{\lambda}=\mathbf{S}_{\lambda}\mathbf{y}=\hat{\mathbf{y}}_{B-PS}^{\lambda} \tag{37}$$

where $\mathbf{S}_{\lambda}=\mathbf{X}(\mathbf{X}'\mathbf{X}+\lambda\mathbf{D})^{-1}\mathbf{X}'$ is a well-known as smoothing matrix for penalized splines. As shown in above equation, the quality of the fitted values depends on choices of parameter λ and $\{\kappa_1,\ldots,\kappa_K\}$, the number of knots. The following paragraph expresses an algorithm for choosing knots. See [2], for a detailed discussion on different knot selection algorithms.

Full-search algorithm: Assume that we have a sequence of candidate values of

$$K=\left\{(\kappa_1,\ldots,\kappa_K)=(5,10,20,40,80,120)\right\}$$

for sample size $n\geq 120$. Moreover, suppose that $\lambda=(\lambda_1,\ldots,\lambda_6)$ is a vector of the regularization parameters. In this case, we use generalized cross validation (GCV) as the regularization parameter selection criterion (or λ) in the knot selection procedure. For $\kappa_j, j=1,\ldots,6$ the algorithm works as follows:

1. The penalized spline fits are performed using the smoothing parameter λ_j, which is chosen by GCV for the knots $\kappa_j, j=1,\ldots,6$.
2. For $j=1,,6$ the value of κ_j that minimizes the $GCV(\lambda_j)$ criterion is selected.

It should be noted here that we use the GCV criterion in the full search algorithm; however, any selection criteria (cross-validation and Akaike information criterion, risk estimation criterion, and so on) can be used in the algorithm. Refs. [28] and [27], provide more detailed information regarding knot selection methods.

5. Selecting the regularization parameter

For the evaluation of the nonparametric regression model, we have to select a regularization parameter. A good regularization parameter should yield a fair balance between the perturbation error and the regularization error in the regularized solution. This section is devoted to choosing good regularization parameter levels for $B-PS$ and $P-TTLS$ methods. Although there are many selection criteria, we have focused on the GCV criterion, which is widely used in the literature. We have modified the GCV criterion, which is used mainly for $B-PS$, to suit the $P-TTLS$ environment. Note that, bandwidth and smoothing parrameters for KS and SS methods are selected by GCV criterion, as in $B-PS$ method.

GCV Criterion for B – PS: GCV is a modified form of the ordinary cross-validation (CV) model, which is a traditional method for choosing the regularization parameter. The GCV score can be computed from the ordinary residuals by dividing by the factors $1-(\mathbf{S}_\lambda)_{ii}$. The GCV score is defined by

$$GCV(\lambda) = \frac{n^{-1}\sum_{i=1}^{n}\left[y_i - \hat{g}^{\lambda}_{B-PS}(x_i)\right]^2}{\left[1-n^{-1}\mathrm{tr}(\mathbf{S}_\lambda)\right]^2} = \frac{n^{-1}\|(\mathbf{I}-\mathbf{S}_\lambda)\|^2}{\left[n^{-1}\mathrm{tr}(\mathbf{I}-\mathbf{S}_\lambda)\right]^2} \tag{38}$$

where $\mathbf{S}_\lambda$ is a smoother matrix, as defined in (37), and $\mathrm{tr}(\mathbf{A})$ denotes the trace of a matrix $\mathbf{A}$ [11, 32].

The value of λ that minimizes the equation (38) is selected as a regularization parameter. The key idea here is to select an appropriate estimate of" g" from the set of corresponding penalized spline estimates $\mathbf{H} = \left\{\hat{\mathbf{g}}^{\lambda_1}_{B-PS}, \ldots, \hat{\mathbf{g}}^{\lambda_k}_{B-PS}\right\}$ for a set of pre-given positive regularization parameters, $\lambda_1 < \lambda_2 < \cdots < \lambda_k$. For example, if we let $\hat{\mathbf{g}}^{\lambda_1}_{B-PS}$ be the minimizer of

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda\sum_{k=1}^{K}\beta^2_{p+k} \tag{39}$$

then the GCV estimate of parameter λ is also an estimate of the minimizer of the mean square error (MSE) function

$$\mathrm{MSE}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{g}^{\lambda}_{B-PS}(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{g}^{\lambda_1}_{B-PS}(x_i)\right)^2 \tag{40}$$

GCV Criterion for P – TTLS: We know from system (7) that the matrix $\mathbf{X}$ is subject to noise. Accordingly, there can be important changes in the norm of residuals located at numerator of (38). The norm square of residuals in the case of TTLS are defined by $\|\mathbf{X}\hat{\beta}^{t}_{P-TTLS} - \mathbf{y}\|^2$. This provides a similar equation to (38) for choosing the regularization parameter that minimizes the following modified GCV criterion

$$GCV(t) = \|\mathbf{X}\hat{\beta}^{t}_{P-TTLS} - \mathbf{y}\|^2 / \left(n - p^{eff}_t\right)^2 \tag{41}$$

where p^{eff}_t is the effective number of parameters, defined as sum of the TTLS filter factors stated in (29), and it can also be given as

$$p^{eff}_t = \sum_{i=1}^{m} f_i = \sum_{i=1}^{m} \sum_{j=t+1}^{m+1} \frac{v^2_{m+1,j}}{\|V_{22}\|^2} \left(\frac{\tilde{\sigma}^2_i}{\sigma^2_i - \sigma^2_j} \right) \tag{42}$$

The definition of an effective parameter (or numerical rank) expressed here is related to the optimum accuracy that can be found in the solution with the given data. It should be emphasized that the values of t for which p^{eff}_t becomes less than n are taken in account. In practice, the value of t that minimizes the GCV criterion (41) is chosen as a truncation parameter. Notice that we only use the values of t in the interval $\left\{[1, t_{\max}] | t_{\max} \le m\right\}$, such that $p^{eff}_1 \le p^{eff}_2 \le \cdots \le p^{eff}_{t_{\max}} \le m$.

6. Statistical properties of the coefficient estimates

The statistical properties of the estimates are related to the Padé-type approximation of function $g(x)$ to be estimated. To evaluate the statistical properties of the TTLS problem, one needs an appropriate model. In statistics literature, the most suitable models are referred to as errors in variable models (*EIV*), which are characterized by the fact that true values of observed variables consist of some unknown true values plus measurements error. The TTLS technique is especially effective in these models with only collinear and measurement error, as stated in linear relationships of the form (13). The relationship between unknown true variables $\mathbf{y}$ and $\mathbf{X}$ has the form

$$\mathbf{y} + \varepsilon_y = (\mathbf{X} + \varepsilon_x)\beta \tag{43}$$

where ε_x and ε_y are considered to be random error components. It is also assumed that $E(\varepsilon_x) = E(\varepsilon_y) = 0$ and their variances are $Var(\varepsilon_x) = \sigma^2_x$ and $Var(\varepsilon_y) = \sigma^2_y$, respectively. Moreover, the error components are mutually uncorrelated.

Note that the estimates $\hat{\beta}^{t}_{P-TTLS}$ of the vector β in (6.1) are described as a linear smoother, as discussed in section 3.3. The algebraic properties of the vector $\hat{\beta}^{t}_{P-TTLS}$ can be expressed as follows:

1. Assume a truncation parameter of $t \in \left\{[1, t_{\max}] | t_{\max} \le m\right\}$, where $m = \text{rank}\{[\mathbf{X}, \mathbf{y}]\}$, as stated in the previous section. If $t = m$, then the expected value and variance-covariance matrix of the $\hat{\beta}^{t}_{P-TTLS}$ are described, respectively, as:

$$E\left(\hat{\beta}^{t}_{P-TTLS}\right) = E\left((\mathbf{P}_t\mathbf{X})^{+}\mathbf{P}_t\mathbf{y}\right) = E\left((\mathbf{P}_t\mathbf{X})^{+}P_t(\mathbf{X}\beta+\varepsilon)\right) = \beta \tag{44}$$

and

$$\mathrm{Var}\left(\hat{\beta}^{t}_{P-TTLS}\right) = \mathrm{Var}\left((\mathbf{P}_t\mathbf{X})^{+}\mathbf{P}_t\mathbf{y}\right) = \sigma^2\left(\mathbf{P}'_t\left((\mathbf{P}_t\mathbf{X})^{+}\right)^2\mathbf{P}_t\right) \tag{45}$$

2. If $t<m$, then $E(\hat{\beta}^{t}_{P-TTLS}) \neq \beta$, but $\hat{\mathbf{y}}^{t}_{P-TTLS} = \hat{\mathbf{g}}^{t}_{P-TTLS} = \mathbf{H}_t\mathbf{y}$ is approximately the same for all such $\hat{\beta}^{t}_{P-TTLS}$.

As seen from equation (45), the variance-covariance matrix is not practical because they depend on the unknown σ^2. One can see that the estimate of σ^2 is needed to construct the aforementioned variance-covariance matrices. The variance of error is estimated by the residual sum of squares (RSS):

$$\begin{aligned} RSS &= \left(\mathbf{y}-\hat{\mathbf{y}}^{t}_{P-TTLS}\right)'\left(\mathbf{y}-\hat{\mathbf{y}}^{t}_{P-TTLS}\right) = \left(\mathbf{y}-\hat{\mathbf{g}}^{t}_{P-TTLS}\right)'\left(\mathbf{y}-\hat{\mathbf{g}}^{t}_{P-TTLS}\right) \\ &= (\mathbf{y}-\mathbf{H}_t\mathbf{y})'(\mathbf{y}-\mathbf{H}_t\mathbf{y}) = \|(\mathbf{I}-\mathbf{H}_t)\mathbf{y}\|_2^2 \end{aligned} \tag{46}$$

where $\mathbf{H}_t$ is the projection matrix stated in (28). Therefore, like ordinary least squares regression, estimation of the error variance can be defined by

$$\hat{\sigma}^2 = \frac{RSS}{tr(\mathbf{I}-\mathbf{H}_t)^2} = \frac{\|(\mathbf{I}-\mathbf{H}_t)\mathbf{y}\|_2^2}{n-m} \tag{47}$$

where $\mathrm{tr}(\mathbf{I}-\mathbf{H}_t)^2 = n-\mathrm{tr}(2\mathbf{H}_t-\mathbf{H}'_t\mathbf{H}_t) = (n-m)$ denotes the residual degrees of freedom. From equation (47), one can see that the degrees of freedom for RSS is also the number of total observations minus total number of the parameters in the model.

6.1. Measuring the risk and error

Generally, the expected loss of fitted values is measured by risk (i.e., the bias-variance decomposition). Our task is now to approximate the risk in the nonparametric regression model. Such approximations have the advantage of being simpler to optimize than the practical selection of truncation parameter t. For convenience, we will work with the mean square errors (MSEs) and therefore compare the accuracy of the $P-TTLS$ and $B-PS$ solution with respect to their squared bias and variance. If we take the expected value of RSS expressed in (46), the MSE is obtained as

$$\begin{aligned} E(RSS) &= E\left(\|(\mathbf{I}-\mathbf{H}_t)\mathbf{y}\|_2^2\right) = E\left(\|\mathbf{y}(\mathbf{I}-\mathbf{H}_t)(\mathbf{I}-\mathbf{H}_t)\mathbf{y}\|^2\right) \\ &= \sigma^2\mathrm{tr}(\mathbf{I}-\mathbf{H}_t)^2 + E\left(\mathbf{y}'\right)(\mathbf{I}-\mathbf{H}_t)'(\mathbf{I}-\mathbf{H}_t)E(\mathbf{y}) \\ &= \sigma^2\left[n-\mathrm{tr}\left(2\mathbf{H}_t-\mathbf{H}_t^2\right)\right] + \hat{\mathbf{g}}^{t}_{P-TTLS}(\mathbf{I}-\mathbf{H}_t)^2\hat{\mathbf{g}}^{t}_{P-TTLS} = MSE \end{aligned} \tag{48}$$

where the first term measures the variance and the second term measures bias, respectively. The error variance σ^2 in (48) is usually unknown. In practice, it can be used as $\hat{\sigma}^2$, as given in (47), instead of σ^2. To show whether $\hat{\sigma}^2$ is an unbiased or biased estimator, $E(\hat{\sigma}^2)$ is found as

$$E(\hat{\sigma}^2) = \frac{1}{n-m} E\left(\|(\mathbf{I} - \mathbf{H}_t)\mathbf{y}\|_2^2 \right) = \frac{1}{n-m} E(RSS)$$

The expected value of $E(RSS)$ implies that the estimator of σ^2 in equation (47) has a positive bias. However, it should be noted that (47) yields asymptotically negligible bias. Considering this fact, it is noteworthy that σ^2 is equivalent to the MSE, which is a widely used criterion for measuring the quality of estimation (see Speckman 1988).

Error analysis is an important subject due to perturbations for the solutions of an overdetermined system (13). The objective is to obtain solutions that are accurate or with small errors. To measure approximation (or truncation) errors between real observations and their fitted values, we consider the relative error (RE) criterion defined by

$$RE = \frac{||\hat{\mathbf{y}} - \mathbf{y}||_{\mathrm{F}}}{||\mathbf{y}||_F} = \frac{\|\mathbf{H}_{\mathbf{t}}\mathbf{y} - \mathbf{y}\|_{\mathrm{F}}}{||\mathbf{y}||_F} = \frac{||\hat{\mathbf{g}}^t_{P-TTLS}}{||\mathbf{y}||_F} \tag{49}$$

As shown in (49), this criterion is based on Frobenius sum of relative errors. We are also aware that many previous researchers have considered this criterion. See [9, 23, 30], for more detailed discussions.

7. Simulation study

This section explores a simulation study that illustrates the effectiveness and performances of the Padé-type approximation based on the TTLS method. The simulation experiments are designed to study the effects of three experimental factors: noise levels, degree of spatial variation, and variance function. The factor levels are changed six times to indicate the effects of these factors on the quality of the estimates. We also compare the fits from $P{-}TTLS$ to the fits computed by B – PS (considered as benchmark method), the KS and the SS methods.

The specification of the simulation setup is designed in the following framework:

- We seek to approximate the unknown regression function g(.) by a rational function of the form

$$g_{[p,q]}(x_i) = \frac{A(x_i)}{B(x_i)} = \frac{a_j x_i^J}{1 + b_k x_i^k}, j = 0, 1, \ldots, p, k = 1, \ldots, q \text{ and } p \leq q \tag{50}$$

The degrees of p and q have been suitably chosen so as to minimize the Frobenius norm of the errors stated in (25).

- In total, three sets of numerical experiments are performed. For each set of experiments, only one of the three experimental factors has been changed, while the remaining two have been left unaffected. Within each set of experiments, the factor levels are changed six times $(i.e., c = 1, 2, 3, 4, 5, 6)$; consequently, there are 18 different configurations altogether, which are used to detect the effects of varying the values of the experimental factors.
- To see the performance of the small, medium, and large samples of the estimates, we generated four different simulation data sets with sample sizes $n = 60, 120, 200,$ and 400. The number of replications was 1000 for each of the 72 numerical experiments.
- In order to obtain appropriate estimates of the parameters expressed in the equations (13) and (35), we determined the optimum regularization parameters (i.e., the truncated parameter for $P-TTLS$ and smoothing parameter for B – PS) that minimize the GCV selection criterion.

For completeness, the data derivation procedures from the equation (50) are given in detail sections 7.1-7.3. Furthermore, in this simulation, we obtained 1000 estimates of function "g" for two methods. As mentioned in previous sections, the estimated MSE values are computed for corresponding "g" functions. Here we use the MSE formula

$$\text{MSE}(\hat{\text{g}}, \text{g}) = \frac{1}{1000}\sum_{j=1}^{1000}\sum_{i=1}^{n}\left\{\hat{\text{g}}(x_{ij}) - \text{g}(x_i)\right\}^2 \tag{51}$$

where $\hat{\text{g}}(x_{ij})$ shows the estimated value at i.th point of the function g(.) in j.th iterations.

The outcomes from the simulation experiments are summarized in the following figures and tables. As indicated before, in this simulation study, because 72 different experiments are made in total, it is not possible to adequately display all the numerical results here. Therefore, only some different configurations are given in the following figures for different samples sized n. The results of the simulation experiments are reported in the following sections under separate headings for each experimental factor.

7.1. Noise level factor

For the first simulation experiment, six true regression functions that include different noise levels are considered. To assess the performance of

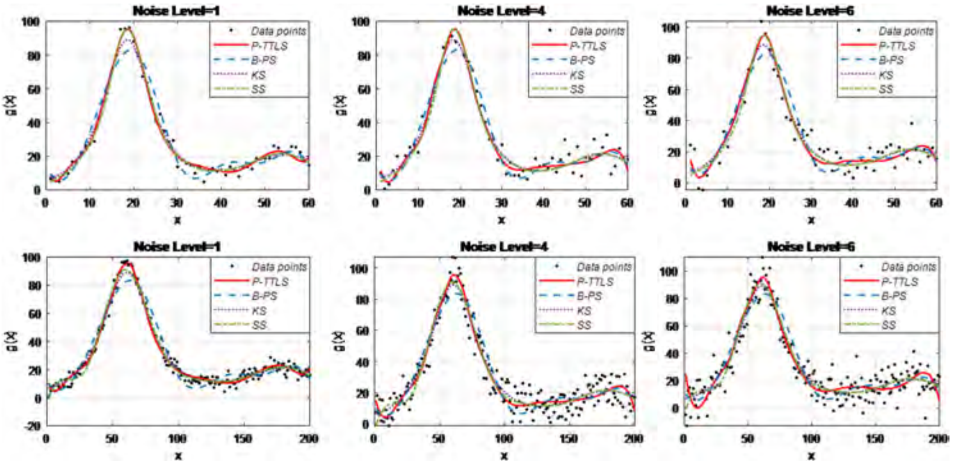

Figure 1. Results for varying the noise level factor. Panels display the fitted curves from the Padé based on TTLS (*P – TTLS* in the legend), penalized spline (*B – PS* in the legend), Kernel smoothing (*KS* in the legend) and Smoothing spline (*SS* in the legend) together with real data observations. The top panel compares a *P – TTLS* fit to a *B – PS* fit, *KS* and *SS* fits for the sample size of $n = 60$. The bottom panel does the same for a sample size of $n = 200$.

our estimation procedures, the data sets $\{(x_i, y_i), i = 1, \ldots, n\}$ are generated from the model

$$y_i = g(x_i) + \sigma_j \varepsilon_i, \sigma_j = \left(0.2 + 0.07(j–0.1)^2\right), j = 1, \ldots, 6 \text{ and } \varepsilon_i \sim NIID(0, 1) \tag{52}$$

with $x_i = (i–0.5)/n$. The true regression function expressed above is defined as

$$g(x_i) = \frac{1}{\left[(x_i - 0.3)^2 + 0.01\right]} + \frac{1}{\left[(x_i - 0.9)^2 + 0.04\right]} –6$$

The main goal is to find the best approximation to the true function $g(x_i)$. In the light of the information presented in section 3, examination of the fits from the Padé based on the TTLS method (*P – TTLS*) show that it is quite reasonable for the true function, when compared to the more traditional penalized spline (*B – PS*), Kernel smoothing (*KS*) and Smoothing spline (*SS*) methods, which is considered here as benchmark.

In Figure 1, each panel presents a single realization of simulated data, hence four different fitted curves. As illustrated in Figure 1, Padé provides a better approximation than *B – PS*, especially for the low noise levels. This idea is also supported by the MSEs given in Table 1. As seen in Figure 1, the curves obtained from the *KS* and *SS* methods appear close to *P – TTLS*, but the differences between them can be seen more clearly in Table 1. It is important that the *P – TTLS* technique has a better and closer fit than the commonly used *B – PS*, *KS* and *SS* smoothing techniques. Although Padé

Table 1. MSEs obtained by the methods based on varying noise levels.

n	60				120			
Noise level	P-TTLS	B-PS	SS	KS	P-TTLS	B-PS	SS	KS
1	**0.239**	0.365	0.257	0.441	0.197	0.351	0.249	0.374
2	**0.281**	0.354	0.342	0.495	**0.230**	0.365	0.271	0.448
3	0.374	**0.373**	0.503	0.555	**0.357**	0.405	0.331	0.520
4	**0.464**	0.484	0.558	0.606	0.465	0.456	**0.423**	0.576
5	0.542	**0.519**	0.619	0.604	0.539	**0.521**	0.588	0.605
6	0.623	**0.602**	0.678	0.715	0.624	**0.608**	0.638	0.683
n	200				400			
Noise level	P-TTLS	B-PS	SS	KS	P-TTLS	B-PS	SS	KS
1	**0.137**	0.300	0.312	0.250	**0.090**	0.262	0.209	0.203
2	**0.180**	0.323	0.352	0.280	**0.102**	0.278	0.214	0.169
3	**0.280**	0.344	0.440	0.359	0.230	0.296	0.253	**0.211**
4	0.405	**0.401**	0.466	0.434	0.323	0.345	0.312	**0.288**
5	**0.450**	0.534	0.493	0.477	**0.375**	0.456	0.416	0.376
6	0.575	0.567	0.561	**0.527**	0.493	0.492	0.505	0.527

makes good estimates for low and medium noise levels, it begins to give slightly fluctuating curves when the noise level is high. Of course, the solutions corresponding to Figure 1 are obtained by using the simulated data samples with two different sized $n = 60$ and $n = 200$ under three noise levels. Further details on other possible combinations in this regard are provided in Figure 2 and Table 1.

The 3 D diagrams of MSE versus varying noise levels are plotted for the estimates obtained by both approaches in Figure 2. In addition, a 3 D plot displaying the two methods together is given in the bottom panel of Figure 2. As can be seen in this figure, the $P - TTLS$ diagram is below the $B - PS$ diagram for small noise levels, which means that the $P - TTLS$ has smaller MSE values, but at higher noise levels it passes $B - PS$. This graph gives a visual way of understanding the behavior of MSE values under increasing noise levels and sample sizes. From this point of view, the performances of both approaches in simulated data sets appear quite similar, and both perform well. However, we get better and more stable estimates from the $P - TTLS$ than when using $B - PS$, especially for low noise levels.

According to Figure 2, as expected, MSEs decrease when the sample size is larger and the noise level is reduced. Correspondingly, when we look carefully at the MSE axis of both 3 D plots, we realize that $P - TTLS$ has a lower minimum and higher maximum than $B - PS$. This means that $P - TTLS$ is better for lower noise levels and worse for higher noise levels than the benchmark $B - PS$. In addition, 3 D diagrams are given in Figure B1 and Appendix B for comparison of $P - TTLS$ with KS and SS methods. According to Figure B1, as in the previous figure, $P - TTLS$ can be said to perform well at lower noise levels, but both the KS and SS methods follow a more stable process in terms of total noise level. However, as one can see in Table 1, $P - TTLS$ still has satisfying results in estimating the nonparametric regression model.

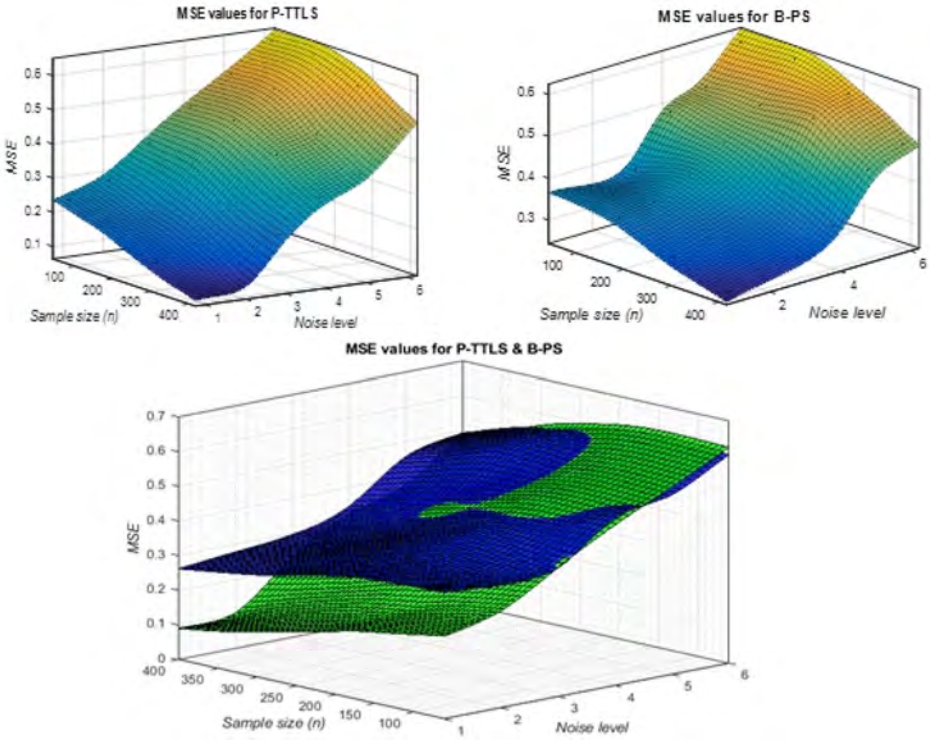

Figure 2. Diagram of MSE versus noise levels for different sample sizes. The left panel shows the MSE for the fits from Padé using the TTLS method, while the right panel denotes the MSE for the fits obtained by $B-PS$.

7.2. Variance function factor

For the second simulation experiment, we considered a regression problem with non-constant variance. Six real regression function are used, as in the varying noise level experiment. The data set of observed pairs of values, $\{(x_i, y_i), i = 1, \dots, n\}$, is constructed by

$$y_i = g(x_i) + \sqrt{v_c(x_i)}\varepsilon_i, c = 1, \dots, 6, \varepsilon_i \sim N(0, 0.5)$$

where $v_c(.)$ is the variance function defined as $v_c(x) = [0.15(1 + 0.4(2c-7)(x-0.5))]$ and x_i values are drawn from uniformly distributed on the interval $(0, 1)$. Finally, the generic form of regression function g is defined as

$$g(x_i) = \left[1.5e^{-\frac{u_{1i}^2}{2}}\sqrt{2\pi}\right] - \left[e^{-\frac{u_{2i}^2}{2}}/\sqrt{2\pi}\right] \text{ with } u_{1i} = \frac{(x_i - 0.35)}{0.15} \text{ and } u_{2i} = (x_i - 0.8)/0.04$$

The outcomes from the simulation experiments are summarized in the following figures and tables. Four fitted curves from methods under non-constant variances, are plotted in Figure 3 alongside one typical simulated data

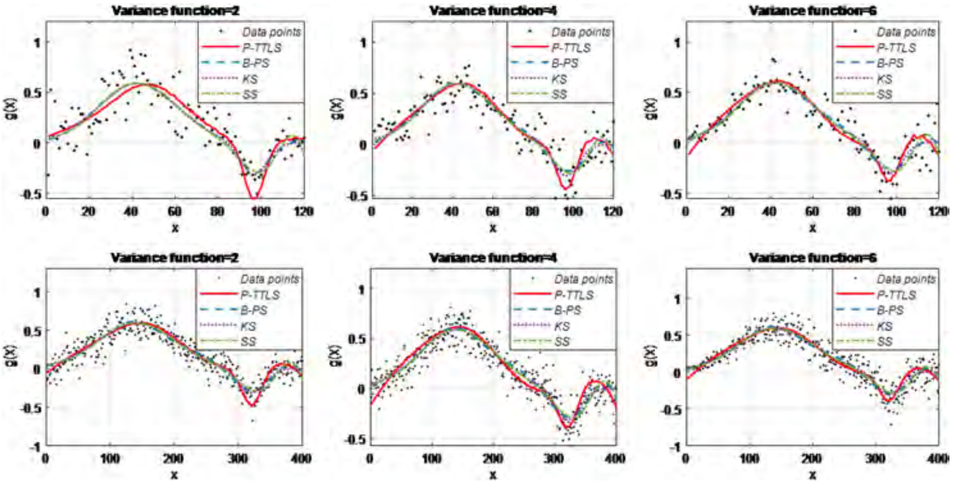

Figure 3. Results for varying variance functions. Like in Figure 1, each panel compares the fitted curves obtained by four approaches. The top panel compares a $P - TTLS$ fit to $B - PS$, KS and SS fits for $n = 120$. The bottom panel compares the same fits for a sample size of $n = 400$.

set. Three-dimensional plots of the MSEs for each method, together with their different variance functions and sample sizes, are illustrated in Figure 4 like the previous simulation experiment. The MSE results obtained by each of the four methods are shown in Table 2 for different combinations.

It can be seen from Figure 3 that both methods deliver reasonable results under the non-constant variance function. However, the Padé approximation based on TTLS, taking into account the rational fraction, performed better (see the bottom panel of Figure 4). In addition, it is clear from Figure 3 and Table 2 that in most cases, fits computed by $P - TTLS$ are better than the other three methods in terms of performance indicators (MSEs) for non-constant variance functions under different sized samples. Note also that $P - TTLS$ is superior to others especially for sample of size $n = 60$. In this sense, it can be said that $P - TTLS$ provides more satisfactory results for different variance functions and sample sizes compared to $B - PS$, KS and SS (see Table 2 and Figures 3,4 and Figure B2).

7.3. Spatial variation factor

For the third numerical simulation experiment, we generated six regression functions with different degrees of spatial variation. Our data set consists of n pairs $\{(x_i, y_i), i = 1, \ldots, n\}$ constructed by

$$y_i = g_c(x_i) + \sigma\varepsilon_i, c = 1, \ldots, 6, \varepsilon_i \sim N(0, 1)$$

where values of x_i are generated as in section 7.2, $\sigma = 0.1,$ and the regression function g is indexed by a single parameter c, and defined in the following way:

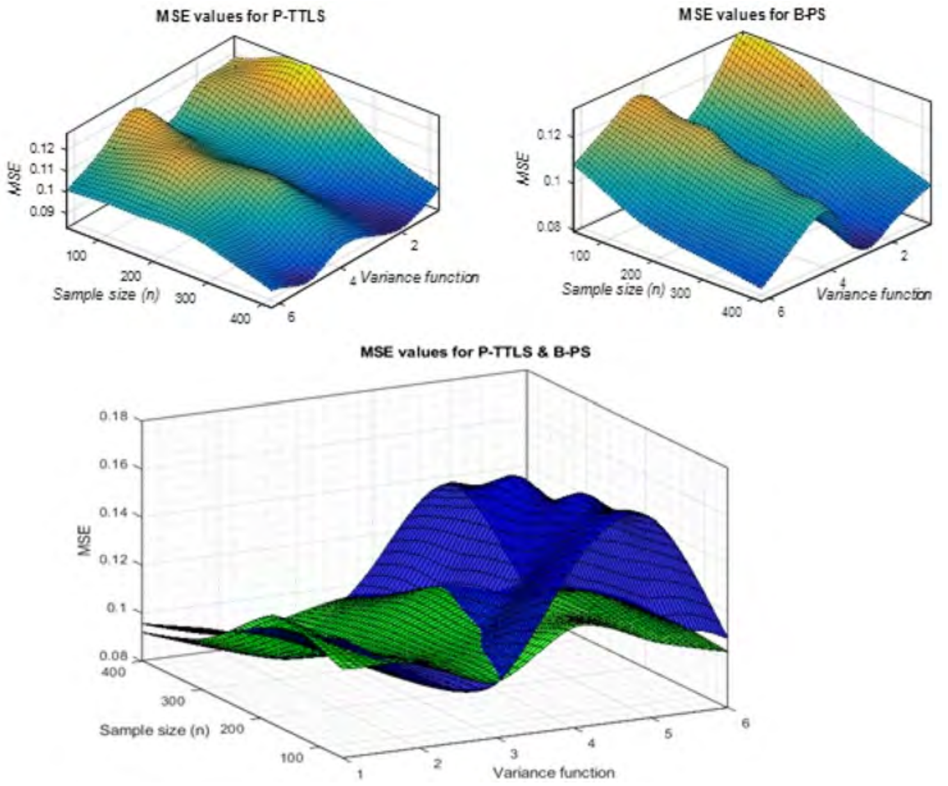

Figure 4. Like Figure 2, the left panel shows the MSE for the fits from the *P* – *TTLS* method and the right panel denotes the MSE for the fits obtained by *B* – *PS*, but for variance functions.

Table 2. MSEs from the methods based on varying variance functions.

n	60				120			
Var. func.	P-TTLS	B-PS	SS	KS	P-TTLS	B-PS	SS	KS
1	**0.117**	0.130	0.130	0.133	0.126	**0.121**	0.163	0.160
2	0.116	**0.110**	0.148	0.147	0.105	**0.102**	0.147	0.141
3	**0.103**	0.103	0.133	0.134	0.102	**0.096**	0.106	0.105
4	0.122	0.125	0.140	0.140	**0.112**	0.121	0.145	0.145
5	**0.112**	0.119	0.152	0.146	0.119	**0.110**	0.126	0.125
6	**0.103**	0.109	0.107	0.103	0.105	**0.097**	0.097	0.118
n	200				400			
Var. func.	P-TTLS	B-PS	SS	KS	P-TTLS	B-PS	SS	KS
1	**0.103**	0.105	0.109	0.108	0.092	0.096	**0.085**	**0.085**
2	**0.090**	0.099	0.104	0.091	**0.085**	0.089	0.091	0.093
3	0.099	**0.087**	0.095	0.095	0.091	**0.082**	0.095	0.096
4	0.112	0.112	0.103	**0.102**	**0.095**	0.107	0.124	0.121
5	**0.102**	0.103	**0.102**	**0.102**	**0.086**	0.100	0.107	0.105
6	0.102	**0.092**	0.102	0.102	0.092	**0.088**	0.089	**0.088**

$$g_c(x_i) = \sqrt{x_i(1-x_i)}\sin\left[\frac{2\pi\left(1+2^{(9-4c)/5}\right)}{x_i+2^{(9-4c)/5}}\right] \tag{53}$$

In a similar manner as in the previous simulation experiments, the fitted curves obtained by the four methods, together with a typical simulated data

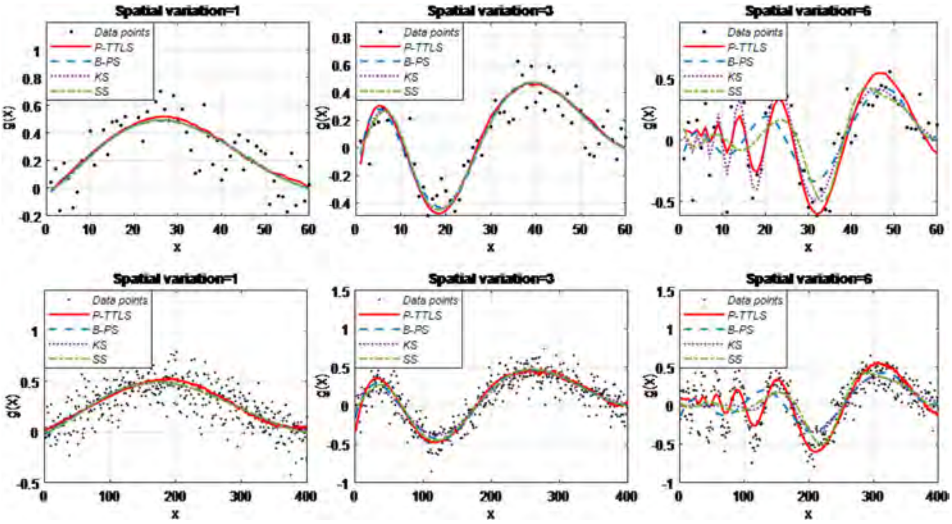

Figure 5. Results for varying spatial variation functions. Like in Figures 1 and 3, the top panel compares the fits from four methods for $n=60$, while the bottom panel compares the same fits for sample size of $n = 400$..

set, are illustrated in Figure 5. For four methods, the MSEs versus different spatial variation functions and sample sizes are displayed in Figure 6. The performance of the methods, the MSEs, are also stated in Table 3.

As can be seen in Figure 5, the $P - TTLS$ and other three methods produce approximately the same fitted values for the first three levels of spatial variation. However, when levels of the spatial variation function are large, the curves fitted by $P - TTLS$ are stable and remain close to real observations, especially for the variation level of 6. The validity of this case can also be confirmed by looking at the MSEs given in Table 3. As seen in this table, the KS method gives the second best performance for a high level of spatial variation. Figure 6 compares the performances of the fitted values from the two methods $P - TTLS$ and $B - PS$. In Figure 6, one can see that MSEs from $B - PS$ are much larger than the MSEs obtained by $P - TTLS$, yielding significant Wilcoxon test rankings with the rejection of the null hypothesis of equality of median of the MSEs (see also Table 4). In addition, 3 D diagrams for $P - TTLS$, SS and KS methods are displayed in Figure B3 given in Appendix B. As can be seen from these Figures, the $P - TTLS$ method appears to have a stationary way from low to high-level spatial variations in terms of MSE scores. However, KS and SS methods show relatively hard transitions between factor levels. From these results, we conclude that the Padé approach (i.e., $P - TTLS$) also works remarkably well in the context of spatial variation problems, as in many applications.

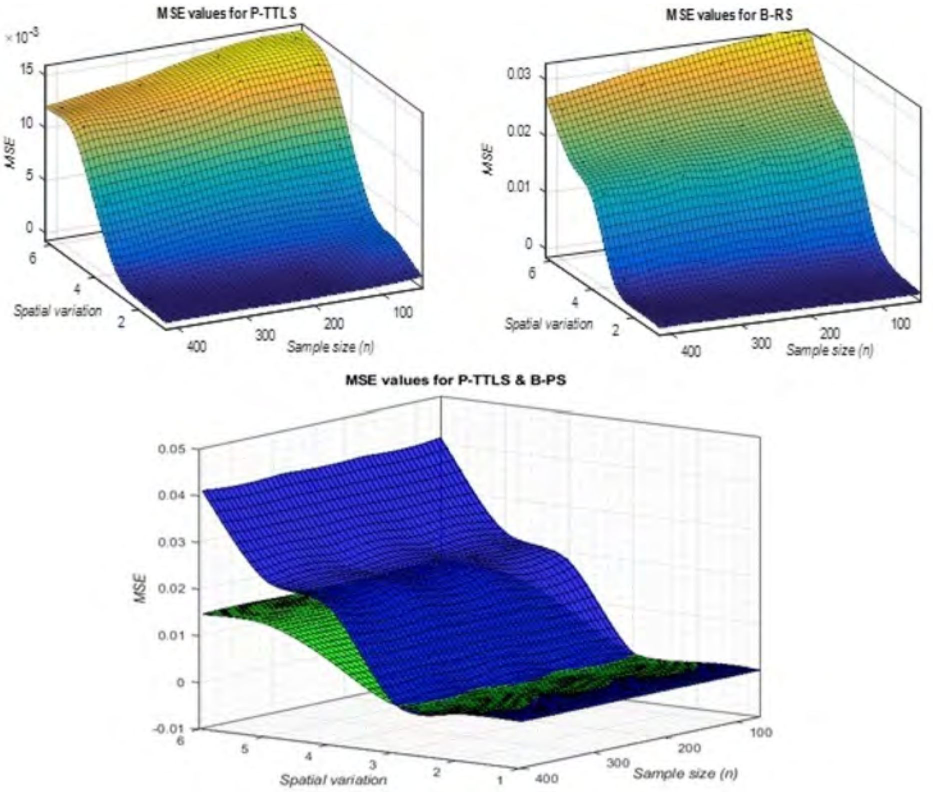

Figure 6. Like Figures 2 and 4, the left panel shows the MSE for the fits from the $P-TTLS$ method and the right panel denotes the MSE for the fits obtained by PS, but for different spatial variation functions.

Table 3. MSEs obtained from methods under spatial variation functions.

n	60				120			
Spatial var.	P-TTLS	B-PS	SS	KS	P-TTLS	B-PS	SS	KS
1	0.001	**0.000**	0.001	0.001	**0.000**	**0.000**	**0.000**	**0.000**
2	0.002	**0.000**	**0.000**	0.001	**0.000**	**0.000**	0.001	0.001
3	**0.003**	0.005	**0.001**	**0.001**	**0.003**	0.004	0.001	0.001
4	**0.008**	0.018	0.009	**0.006**	0.008	0.017	0.007	**0.006**
5	0.014	0.023	0.033	**0.013**	**0.014**	0.021	0.019	0.016
6	0.015	0.031	0.034	0.023	**0.015**	0.030	0.032	0.020
n	200				400			
Spatial var.	P-TTLS	B-PS	SS	KS	P-TTLS	B-PS	SS	KS
1	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
2	**0.000**	**0.000**	0.001	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
3	0.002	0.004	**0.001**	0.003	0.001	0.002	**0.000**	0.001
4	**0.007**	0.017	0.009	0.009	**0.006**	0.015	**0.005**	**0.005**
5	0.012	0.020	0.014	**0.011**	**0.012**	0.018	0.014	**0.012**
6	**0.013**	0.028	0.027	0.016	**0.012**	0.025	0.022	0.014

7.4. Performance comparisons

As indicated in section 7, we use the MSEs to evaluate the quality of the fitted values. In this context, we use paired Wilcoxon tests to determine

Table 4. Averaged Wilcoxon test rankings related to the regularization methods.

Sample Sizes	n = 60	n = 120	n = 200	n = 400	Method
Overall	1.500*	1.667*	1.667*	1.167*	P-TTLS
	2.167	2.167	2.167	3.333	B-PS
	3.333	3.167	3.167	3.167	SS
	3.000	3.000	3.000	2.333	KS

whether the difference between the median of the MSEs obtained from each of the methods is statistically significant at a significance level of 5%. The methods are also ranked in the following manner: If the median MSE value of a method is significantly less than other, it is assigned a rank of 1, and a rank of 2 otherwise. Methods with non-significantly different median values share the same averaged rank. The resulting Wilcoxon test rankings are illustrated in Figure 7, and the averaged rankings values are conveyed in Table 4. As we can see from Figure 7, there is a contraction in the range of estimates as the sample sizes increase from 60 to 400. In general, we see that the superior performance of the $P - TTLS$ here may be due to the fact that the Padé approximation provides more optimum estimators to the parameters being estimated than the $B - PS$, SS and KS estimators, especially in the spatial variation and heterogeneous variance factors.

Note that Table 4 is constructed based on the rankings of the median values of the MSEs given in Figure 7. According to Table 4, for all samples, the $P - TTLS$ method has had a good empirical performance for variance function and spatial variation factors. Furthermore, $B - PS$ has shared the better performance after proposed $P - TTLS$ method for sample of sizes $n = 60$, 120 and 200. Note also that KS has had a good performance after $P - TTLS$ especially for big sized samples (i.e., $n = 400$). For a detailed discussion of this issue, box plots of MSE values obtained from each method under each factor are also displayed in Figure 7. These results show that $P - TTLS$ has a good ability to estimate the response variable under spatial variation factors. However, it cannot be said that $P - TTLS$ method shows the same success under the variance factors and varying noise levels. What we are seeing here is that the $B - PS$, KS and SS methods provide good estimates under data sets with noise level and variance factors. It should also be emphasized that KS and SS techniques are not as successful as the $B - PS$ method. One of the most important reasons for this case is that the $B - PS$ has a knot selection procedure. As we described in section 4, the $B - PS$ uses a full search algorithm for estimating the parameters of the regression model. This algorithm considers the knot points from 5 to $(n-1)$ and thus determines the optimal number of knots according to minimum error [See Aydin and Yilmaz 2017, 27]. As a result, the $B - PS$ method fails to cope with the fluctuations in the data under spatial

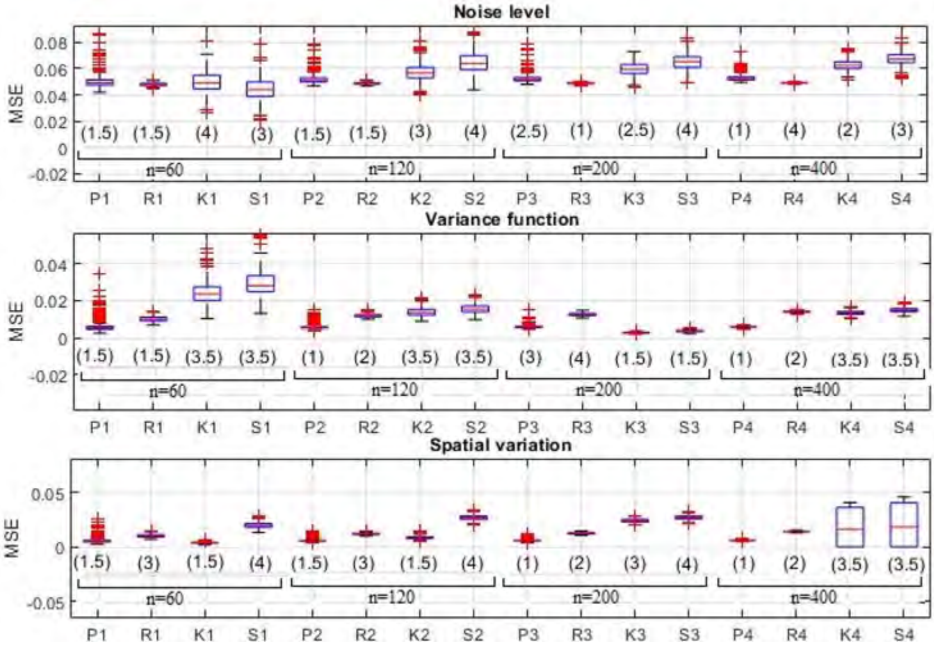

Figure 7. Each panel shows the boxplots of the MSEs for fitted curves. The numbers below the boxplots are the paired Wilcoxon test rankings. For each sample size P1, P2, and P3, the boxplots of the replications of the MSEs when fits are constructed by *P – TTLS* under varying noise levels in the uppermost panel. Similarly, R1, R2, and R3 represent the boxplots of the MSEs determined by *B – PS*. In a similar fashion K1, K2, and K3 denote the boxplots of the MSEs computed by *KS*, while S1, S2 and S3 indicate the boxplots of the MSEs from *SS* method. The remaining panels use the same labeling system as the first, but for variance function (middle panel) and spatial variation (bottom panel).

variation factor because it takes fixed knot points into account. This method is successful for the noise level factor due to the same reasons.

We now proceed to evaluate the discrepancy between real and fitted values. A useful technique, which takes measurement errors into account, is known as the relative error of approximation. The key idea is to understand the reasons for the discrepancy of the *P – TTLS*, *B – PS*, *KS* and *SS* solutions. When these solutions are investigated, we see that the choice of the regularization parameters plays an important role. To obtain an optimum solution for each method under the aforementioned factors, we generated 1000 simulated data sets, and for each of the 1000 fits, we calculated the regularization parameter by GCV criterion. For comparison, for each fit, we also computed the optimum regularization parameters (i.e., a truncation level for *P – TTLS* and a smoothing parameter for *B – PS*, *SS* and *KS*) that minimize the relative errors, as given in (41 and 38).

$$RE = ||\hat{\mathbf{y}} - \mathbf{y}||_F / ||\mathbf{y}||_F$$

Table 5. The average relative errors.

	Noise level				Variance function				Spatial variation			
Sample size	P-TTLS	B-PS	SS	KS	P-TTLS	B-PS	SS	KS	P-TTLS	B-PS	SS	KS
60	**0.829**	0.959	1.035	1.319	**0.885**	0.927	1.127	1.114	**0.720**	2.279	1.319	0.807
120	0.845	**0.776**	0.906	1.337	0.922	**0.877**	1.119	1.144	**0.761**	2.090	1.024	0.815
200	**0.779**	**0.756**	1.245	1.034	1.008	**0.985**	1.025	0.991	**0.787**	1.949	0.981	1.059
400	**0.787**	1.377	1.140	1.005	**0.944**	0.987	1.049	1.044	0.801	1.977	1.063	**0.784**

are averaged over 1000 simulated data sets affected by different experimental factors. For each of the two methods, the averaged *RE* values are given in Table 5, while their column graphs are displayed in Figure 8.

In this part of the analysis, the magnitude of the absolute error ratios in terms of the simulated data measurements is determined. From Table 5, we see that when regularization methods, $P - TTLS$, $B - PS$, KS and SS, are applied to these data sets, the regularized solutions are obtained with relatively small averaged values of the relative errors by means of the GCV criterion. However, it should be noted that $P - TTLS$ performs substantially well in spatial variation and variance function factors, as mentioned earlier. In variance function factor, all methods give similar results but $P - TTLS$ has the smaller (total and relative) error(s). This idea is also clearly supported by the two column (or bar) graphs illustrated in Figure 8.

Finally, we developed a web application by using $R - shiny$ software. The key idea is to compare the proposed $P - TTLS$ method with the widely used methods, such as Kernel smoothing and smoothing spline, in terms of MSE scores under the same simulation setup. See, https://ey13.shinyapps.io/pttls_comparison/ for more detailed information.

8. Real data example

8.1. Fuel consumption data

In this section of the study, we introduce a real data example to see how the proposed method performs. We applied the $P - TTLS$ method to fuel consumption data collected by Carnegie Mellon University Statistics library in 1983 and it is used by Quinlan (1993) for estimating fuel consumption. This data has totally 398 observations and 9 variables. Note that the data set is collected to investigate 8 factors such as number of cylinders, engine displacement, engine horsepower, vehicle weight, acceleration, model year, origin of car and vehicle name (string) on the city-cycle fuel consumption in miles per gallon. In this study, since a nonparametric regression model with a single explanatory variable is considered, only two variables belonging to this dataset are used; city-cycle fuel consumption (*fc*) as a response variable and displacement (*disp*) as a nonparametric explanatory variable. The main goal in this example is to clarify the relationship between these two variables using the model

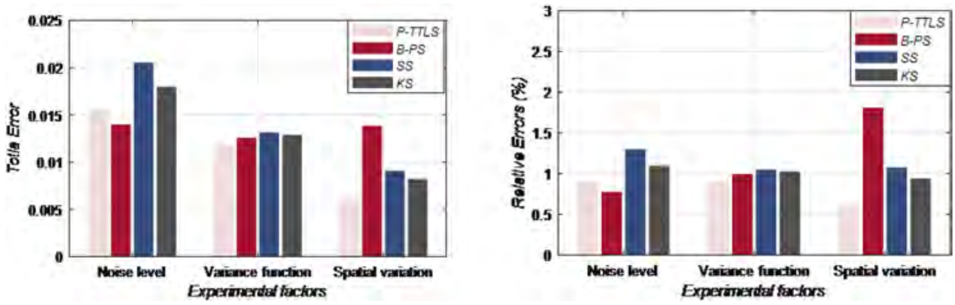

Figure 8. These column charts present the relative errors computed by each of the regularization methods for all sample sizes under different factors. The left chart denotes the means of the relative errors, while the right displays the percentage errors found by multiplying the relative error by 100%.

$$y_i = fc_i = \mathrm{g}(disp_i) + \varepsilon_i, i = 1, \ldots, 398 \tag{54}$$

In the light of the ideas stated in section 3, a linear system of equations is constructed, as in (13). To obtain the estimates of Padé coefficients in the equations, we used the equalities (24) or (26) as the regularization solutions and correspondingly the value of the truncation level is selected with the GCV criterion, as stated in (41). The value of the truncation parameter in this example is $t = 8$. The results from the regularization method, $P - TTLS$, are compared to those from $B - PS$ defined in (36) based on a smoothing parameter. The value of this parameter, $\lambda = 0.001$, is also chosen by GCV criterion, but for penalized splines, as described in (35).

We begin by defining the existence of the nonlinear relationship between variables disp and fc. In this regard, we used the following F-test statistics described in [22], for testing null hypotheses: $H_0 : E(y_i) = \mu(linearfunction)$ against the alternative $H_1 : E(y_i) = g(x_i)(smoothfunction)$

$$F_{df_1 - df_0, n - df_1} = \frac{\left(\sum_i^n \hat{\varepsilon}_i^2 - \sum_i^n \hat{v}_i^2\right) / (df_1 - df_0)}{\sum_i^n \hat{v}_i^2 / (n - df_1)} \tag{55}$$

where $\hat{\varepsilon}_i = (y_i - x_i' \hat{\beta}_{OLS})$ and $\hat{v}_i = (\hat{\mathbf{g}}_{P-TTLS}(x_i) - x_i' \hat{\beta}_{OLS})$. Here, $\hat{\beta}_{OLS}$ is the estimates of parameters from the ordinary least squares, $\hat{\mathbf{g}}_{P-TTLS}$ is the fitted values, as defined in (28), $df_1 = \mathrm{tr}(2\mathbf{H}_{P-TTLS} - \mathbf{H}_{P-TTLS}\mathbf{H}'_{P-TTLS})$ where $\mathbf{H}_{P-TTLS}$ denotes the hat matrix for the $P - TTLS$, as expressed in (27), and df_1 is the number of the parameters solved by the OLS method. It should also be noted that we used only the $P - TTLS$ method to get the above F-test statistics. However, the $B - PS$ method could be used instead.

Using the F-statistics (55) we obtain $F_{7,390} = 45.2425$ with 7 and 390 degrees of freedom. Comparing $F = 45.2425$ with critical value $F_{(0.05;7,390)} = 3.08$, we reject the null hypothesis, showing that the linear

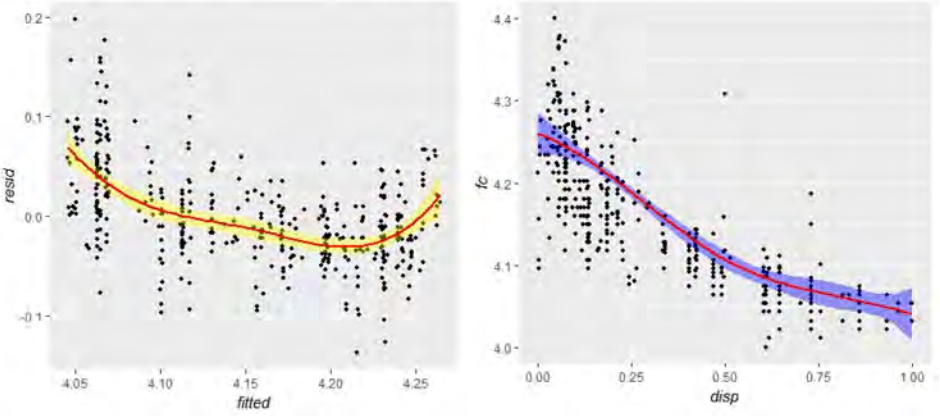

Figure 9. The left panel shows the scatter plot of residuals versus fitted values solved by the *P* – *TTLS* method, while the right panel displays the scatter plot of *B* – *PS* fits against disp. In these panels, the yellow and blue shaded regions denote the 95% confidence intervals form the *P* – *TTLS* and *B* – *PS* fits, respectively.

function is appropriate. This result is also supported graphically through the two different scatter plots displayed in Figure 9. It follows that the relationship between *disp* and *fc* may be nonlinear, especially when we added the fitted values with 95% confidence intervals indicated by shaded regions.

The remaining outcomes from the *P* – *TTLS* and *B* – *PS* methods applied to the fuel consumption data are summarized in Figure 10 and Table 6. Figure 10 compares the fits from the *P* – *TTLS* and *B* – *PS* methods on the fuel consumption data collected by Carnegie Mellon University in 1983. Note that the horizontal axis indicates the scaled values of the explanatory variable. The main purpose of using scaled values in this graphic is to get a better visualization of the data. However, all other calculations are based on real observations. Besides the graphical result given above, the MSE values of the fitted values solved by *P* – *TTLS*, *B* – *PS*, *KS* and *SS* are 0.2163, 0.2109, 0.2916 and 0.5493 respectively. Their performance is almost identical and all three are optimal, except for *SS*. Apparently, the *SS* has not had a good performance in this dataset. In our view, the proposed *P* – *TTLS* approach to estimation of a nonparametric regression model has produced satisfying conclusions for the real data example, as well as the simulated data sets under different varying factors.

The outcomes so far demonstrate that the *B* – *PS* methods provides a better performance than *P* – *TTLS*, *KS* and *SS* in terms of varying noise level factors. When real data is examined in detail, the typical behavior of a data set with high noise level factors is shown to be similar to that of the simulation examples. From the simulation experiments, we know that in general, the *P* – *TTLS* method performs worse than *B* – *PS* when it comes to this factor (noise levels). Therefore, it can be said that the actual data

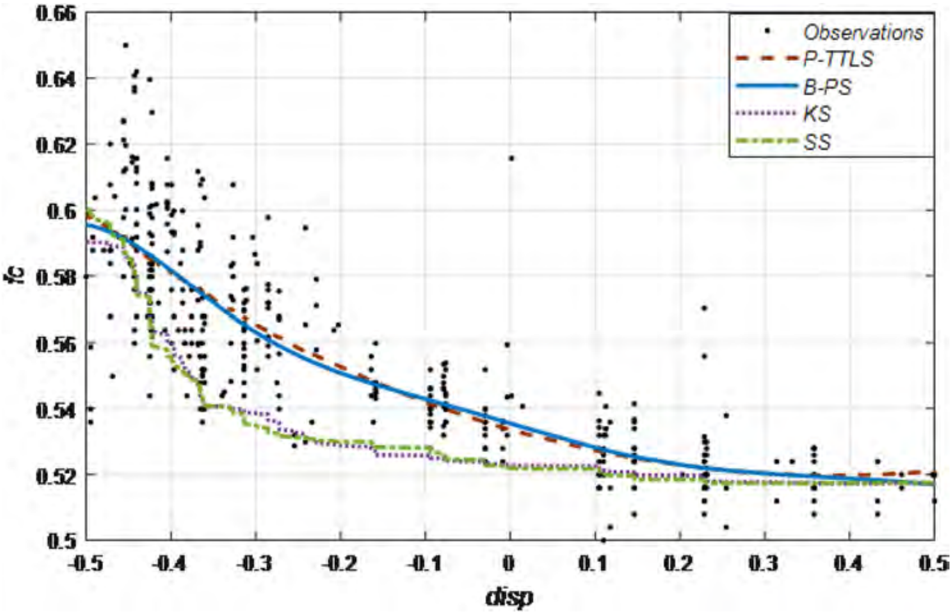

Figure 10. The fitted values solved using $P - TTLS$ and B-PS for the model on a scatterplot of fuel consumption data. The $B - PS$ fit is given by solid line (in blue), while $P - TTLS$ fit is represented by the dashed line (in red), KS is denoted by dotted line (in purple) and SS is represented by the dashed line (in green).

Table 6. Values of MSE. variance and relative errors for the fuel consumption data.

Methods	MSE	Variance	Relative Errors
B-PS	**0.2109**	**0.0292**	**0.5205**
P-TTLS	0.2163	0.0302	0.5402
KS	0.2916	0.0368	0.8154
SS	0.5493	0.0502	1.7569

works in harmony with the simulation. From (7.3), it is seen that noise level factor is considered as a multiplier of the variance of the error terms. Technically, this means that the aforementioned factor can also be interpreted as a uniform amplifier of noisiness of the simulated data. In this context, the variance of the measurement error (the noise) is found as 2.97 in the fuel consumption data used as a real data example. It is seen that this value is approximately equivalent to the 6th noise level (i.e., 2.63) given in the simulation study. As can be seen from simulation studies, it is clear that $P - TTLS$ is superior especially in low noise levels. The outcomes from fuel consumption data show that the recommended $P - TTLS$ and the widely used $B - PS$ method have almost the same performance in terms of evaluation criteria for high noise levels. Note also that $P - TTLS$ and $B - PS$ methods give better scores than KS and SS techniques. This case proves that the proposed $P - TTLS$ method is as successful as the $B - PS$. In addition, we calculated some further results found by the two methods to evaluate their performance comparatively, as reported in Table 6.

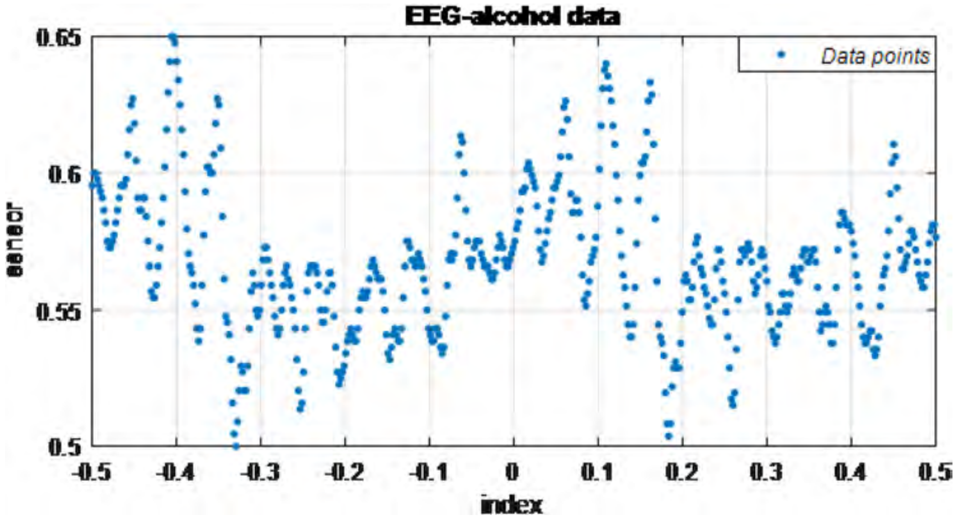

Figure 11. Scatter plot of EEG-alcohol data.

According to Table 6, our conclusion is that the performance measurements in value are fairly close to each other, except for *SS*. However, the findings here give strong evidence in favor of the proposed *P* – *TTLS* and *B* – *PS* method.

8.2. Eeg-alcohol data

The EEG-alcohol data is used to see the success of *P* – *TTLS* applied to a data with spatial variation, which has a type of nonlinear structure, as in the simulation experiments. This data set is obtained from a study conducted at the Neurodynamic Laboratory in State University of New York Health science Center. Note also that data can be accessed via the link https://archive.ics.uci.edu/ml/datasets/eeg+database. (Also, see Zhang et al. 1995 for detailed process of data collection). The dataset has four main features, such as sensor location, sensor value, subject identifier and sample index. In this study, sensor value ($sensor_i$) is determined as a response variable, sample index ($index_i$) is considered as a nonparametric covariate. We can emphasize here that $index_i$ has a uniform structure that is similar to covariate produced in the simulation study. From information given above, the nonparametric regression model for this example is specified by

$$y_i = \text{sensor}_i = \text{g}(\text{index}_i) + \varepsilon_i, i = 1, \ldots, 500 \tag{56}$$

As seen from inspection of Figure 11, one can see that there is a nonlinear and nonparametric relationship between two variables.

Estimates of unknown Padé coefficients are obtained similar to Section 8.1. Note also that the truncation level and smoothing parameters that

Table 7. Values of MSE, variance and relative errors for the EEG-alcohol data.

Methods	MSE	Variance	Relative Errors
B-PS	0.4703	0.0471	0.8562
P-TTLS	**0.0743**	**0.0007**	**0.1953**
KS	0.1576	0.0157	0.6945
SS	0.4895	0.0490	2.6837

minimize the GCV criteria defined in section 5 are selected and they are found to be $t=16$ and $\lambda=0.00019$ for $P-TTLS$ and $B-PS$ methods, respectively. Using these parameters, the comparative outcomes of the nonparametric model (56) are summarized in the following Table 7 and Figure 12 for the EEG-alchol data set. In this sense, Figure 12 depicts the fitted curves obtained using both methods. As mentioned above, since this data set has a high spatial variation, it can be said that based on $P-TTLS$ method, the nonparametric model fits much better than the nonparametric model considered $B-PS$, KS and SS, as in simulation experiments. On the other hand, to evaluate the performance of each method, we compute the MSE values, variances, and relative errors obtained from two different estimators. For example, if MSE value of one estimator is smaller than the other then, this indicates the superiority of this method over the other estimator, as discussed earlier. These results are displayed in Table 7.

8.3. Age-income data

In this section, "Age-Income" data is used to show the performances of the proposed $P-TTLS$, $B-PS$, KS and SS techniques on a real data set with different local variances. Furthermore, this data set can be considered as the verification of the results of the variance function factor included in the simulation study. This real data set has 205 pairs observations on Canadian workers from a 1971 Canadian Census [34]. One can access the dataset by using the "*semipar*" package in the open source *R*-software. The data set has two variables: age (age_i) as a non-parametric predictor and logarithm of income $log(income_i))$ as the response variable. Thus, the nonparametric regression model for age-income data is provided by

$$y_i = \log(\text{income}_i) = \text{g}(\text{age}) + \varepsilon_i, i = 1, \ldots, 205 \qquad (57)$$

Similar to Sections 8.1 and 8.2, performance scores and curves fitted by $P-TTLS$, $B-PS$, KS and SS techniques are obtained and these comparative results are summarized in the following Table 8 and Figure 13.

Figure 13 shows the curves fitted by four methods considered in this study. As can be seen in scattered observations, there is an obvious change in variance in the data, especially after age more than forty years. According to the contoured curves and MSE scores in Table 8, it can be said that $B-PS$ is a good representation method for this data. However, it

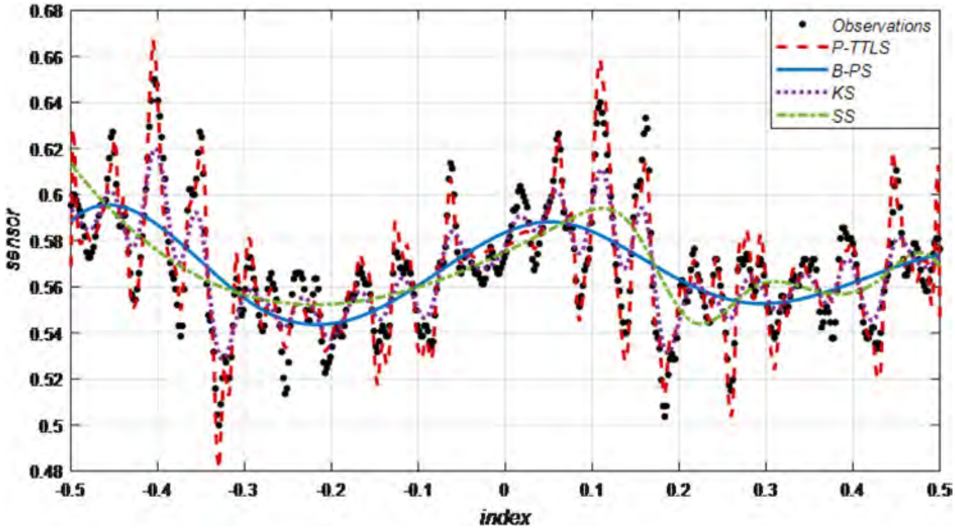

Figure 12. The fitted values solved using $P - TTLS$, $B - PS$, KS and SS for the model (56) on a scatterplot of EEG-alcohol data.

can be said that there is no strongly advantageous method among the methods in terms of superiority. One of the most important indicators of this statement is that the performance values of each method come close to each other. In this context, it is clear that $P - TTLS$ can easily find a place in the literature as a nonparametric smoothing technique among other popular methods.

In this real data example, although $P - TTLS$ does not produce an expected result, most of the methods in the real world may behave differently from theory. However, it can be clearly understood that when the Figure 8 is examined, the results are really close to each other in the variance function factor. Therefore, it would not be wrong for this real data sample to be consistent with the simulation study in this respect. These ideas are supported by the MSE values, the variances and the relative errors of methods given in Table 8.

9. Conclusions and recommendations

In this paper, in order to estimate an unknown smooth function in a nonparametric regression environment, we proposed a Padé-type approximation based on the TTLS technique ($P - TTLS$), compared with the benchmarked penalized spline method ($B - PS$), Kernel smoothing (KS) and smoothing spline (SS). For a comprehensive understanding of approximation theory expressed here, we carried out simulation experiments under different factors. Also, three real data examples are presented in Section 8. Note that each real data study corresponds to one of the experimental

Table 8. Outcomes for age-income data.

Methods	MSE	Variance	Relative Errors
B-PS	**0.0175**	**0.0017**	**0.8337**
P-TTLS	0.0224	0.0022	1.1605
KS	0.0205	0.0021	1.0338
SS	0.0202	0.0020	1.0138

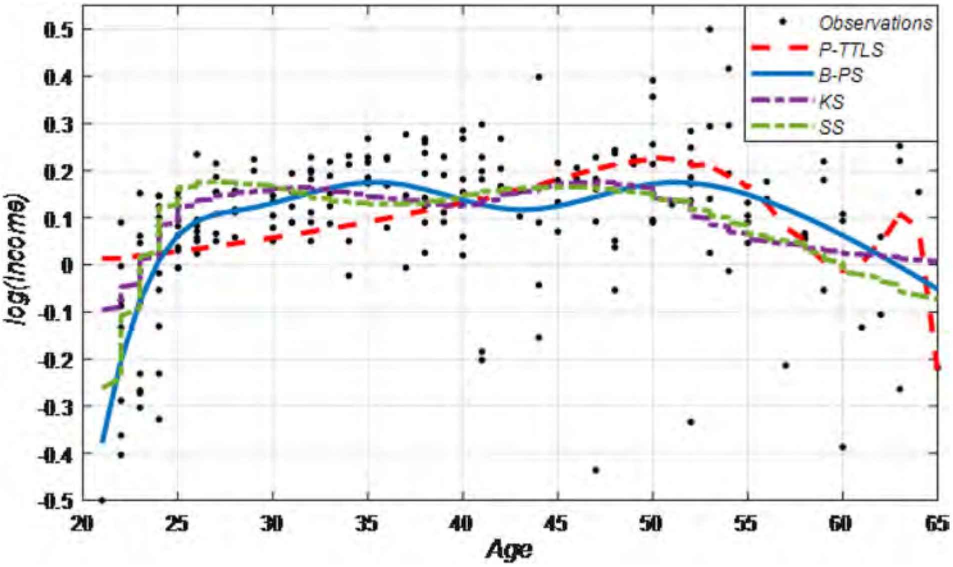

Figure 13. The fitted curves obtained by $P - TTLS$, $B - PS$, KS and SS for the model (57) on a scatterplot of Age-Income data.

factors used in the simulation study. The outcomes from the simulation and real data examples are satisfactory and also demonstrate that $P - TTLS$ is both useful and a feasible method in the estimation procedure of the nonparametric regression function. The empirical results confirmed that the regularization methods have similar performance measurements, based on the noise level factor. However, the $P - TTLS$ shows superior performance when compared to other three widely used smoothing techniques under the varying variance function and spatial variation factors. In summation, based on the numerical simulation experiments and real data results, the following suggestions and conclusions should be considered:

- We see that although $P - TTLS$ provides better fits of data sets based on lower noise levels, it exhibits poor performance for the data sets under high noise levels. We also see that proposed $P - TTLS$ solutions can compete with those obtained with traditional $B - PS$, KS and SS methods.
- Interestingly, the $P - TTLS$ performs much better than the other three in all simulation scenarios based on varying spatial variation factors.

We therefore recommend the $P - TTLS$ as a good approximation method for spatial variation data. It should also be noted that KS is the second method that works good under this factor (See Figures 7 and 8).

- In general, under the heterogeneous variance factors, the $P - TTLS$ method gives more satisfactory fits compared to the benchmarked $B - PS$, KS and SS methods. However, it should be emphasized that results of the four methods for the variance function factor are really close to each other (see Figures 8 and 13). Therefore, it can be concluded that all of these methods give satisfactory outcomes. It is the great opportunity for proposed $P - TTLS$ and it shows that it deserves a place among other popular smoothing techniques.
- Despite the noisy structure of real data, $P - TTLS$ method gives convincing results. However, it should be stressed that $B - PS$ works better in terms of MSE and relative errors for real data (see Table 6).
- For a nonparametric regression function with the spatial variation or variance factor, the $P - TTLS$ approximation method seems to be superior (see Figure 7 and Table 4 for sample sizes of $n = 120$ and 400). In most simulation scenarios, the proposed $P - TTLS$ method gave very effective solutions (see Table 5 and Figure 8).
- In three real data examples, results are obtained in harmony with the simulation study. However, in Section 8.3, although $P - TTLS$ has not given the best score as in simulation, still, the four methods shows really close results which also supports the simulation study.

Finally, the overall results of two numerical studies demonstrated that the proposed $P - TTLS$ method provides feasible estimates for the nonparametric regression model. Furthermore, as the sample size n increases, the range of estimates decreases. The estimates from medium and large samples are more stable than those from small samples (see Figure 7).

References

[1] Aydin, D., Yilmaz, E. (2017). Modified spline regression based on randomly right-censored data: a comparison study. *Commun. Stat. –Simul. Comput.* 47(9): 2587–2611. DOI: 10.1080/03610918.2017.1353615.

[2] Aydin, D., Yilmaz, E. (2019). Truncation level selection in nonparametric regression using Padé approximation. *Commun. Stat. –Simul. Comput.*:1–20. DOI: 10.1080/03610918.2019.1565586.

[3] Baker, G. (1975). *Essentials of Padé Approximants*. 1st ed. New York, NY: Academic Press.

[4] Baker, G. A., Jr., Graves-Morris, P. (1996). *Padé Approximations*. 2nd ed. New York: Cambridge University Press.

[5] Björck, A. (1996). *Numerical Methods for Least Squares Problems*. Philadelphia: Society for Industrial and Applied Mathematics.

[6] Bonizzoni, F., Nobile, F., Perugia, I., Pradovera, D. (2018). Least-squares Padé approximation of parametric and stochastic Helmholtz maps. Adv. Comput. Math. 46(3):46.

[7] Brezinski, C. (1980). *Padé-Type Approximation and General Orthogonal Polynomials*. Basel: Birkhäuser.

[8] Brezinski, C., Rodriguez, G., Seatzu, S. (2009). Error estimates for the regularization of least squares problems. *Numer. Algor*. 51(1):61–76. DOI: 10.1007/s11075-008-9243-2.

[9] Chen, K., Shaojun, G., Lin, Y., Zhiliang, Y. (2010). Least absolute relative error estimation. *J. Am. Stat. Assoc*. 105(491):1104–1112. DOI: 10.1198/jasa.2010.tm09307.

[10] Cheney, E. W. (1996). *Introduction to Approximation Theory*. 2nd ed. New York, NY: Chelsea Publishing Company.

[11] Cheng, C.-L., Van Ness, J.W. (2000). Statistical regression with measurement error. London: Wiley.

[12] Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math*. 31(4):377–403. DOI: 10.1007/BF01404567.

[13] Eckart, C., Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*. 1(3):211–218. DOI: 10.1007/BF02288367.

[14] Eubank, R. (1999). *Nonparametric Regression and Spline Smoothing*. New York, NY: Marcel Dekker.

[15] Fan, J., Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

[16] Fierro, R. D., Golub, G. H., Hansen, P. C., O'Leary, D. P. (1997). Regularization by truncated total least squares. *SIAM J. Sci. Comput*. 18(4):1223–1241. DOI: 10.1137/S1064827594263837.

[17] Golub, G.H., Hoffman, A., Stewart, G.W. (1987). A generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra Appl*. 88–89(89):317–327. DOI: 10.1016/0024-3795(87)90114-5.

[18] Golub, G.H., Van Loan, C.F. (1980). An analysis of the total least squares problem. *SIAM J. Numer. Anal*. 17(6):883–893. DOI: 10.1137/0717073.

[19] Golub, G.H., Van Loan, C.F. (2013). *Matrix Computations*. Baltimore, MD: Johns Hopkins University Press.

[20] Gonnet, P., Pachon, R., Trefethen, L.N. (2011). Robust rational interpolation and least squares. 38:146–167.

[21] Green, P.J., Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Model*. Florida: Chapman & Hall.

[22] Hansen, P.C. (1998). *Rank-Deficient and Discrete ill-Posed Problems*, Philadelphia: SIAM.

[23] Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

[24] Hastie, T., Tibshirani, R., Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer-Verlag.

[25] Khoshgoftaar, T. M., Bhattacharyya, B.B., Richardson, G.D. (1992). Predicting software errors, during development, using nonlinear regression models: a comparative study. *IEEE Trans. Rel*. 41(3):390–395. DOI: 10.1109/24.159804.

[26] Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *Q J. Math*. 11(1):50–59. DOI: 10.1093/qmath/11.1.50.

[27] Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl*. 9(1):141–142. DOI: 10.1137/1109020.

[28] Petrushev, P.P., Popov, V.A. (1987). *Rational Approximation of Real Functions*. New York, NY: Cambridge University Press.

[29] Quinlan, R. (1993). Combining instance-based and model-based learning. In Proceedings on the Tenth International Conference of Machine Learning, 236–243, University of Massachusetts, Amherst, Morgan Kaufmann.

[30] Ruppert, D., Wand, M.P., Carroll, R.J. (2003). *Semiparametric Regression*. New York, NY, Cambridge University Press.

[31] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graphical Stat.* 11(4):735–757. DOI: 10.1198/106186002853.

[32] Schimek, G. M. (2000). *Smoothing and Regression: Approaches, Computation, and Application*. Hoboken, NJ: John Wiley & Sons, Inc, pp. 19–39.

[33] Sima, D.M., Huffel, S. (2007). Level choice in truncated total least squares. *Comput. Stat. Data Anal.* 52(2):1103–1118. DOI: 10.1016/j.csda.2007.05.015.

[34] Speckman, P. (1988). Kernel smoothing in partial linear models. J. R. Statist. Soc. B. 50(3):413–436.

[35] Tikhonov, A.N. (1963). Solution of incorrectly formulated problems and the regularization method. *Sov. Math. Dokl.* 4:1035–1038.

[36] Wahba, G. (1990). *Spline Model for Observational Data*. Philadelphia, PA: SIAM.

[37] Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Ser. A.* 26(15):175–184.

[38] Ullah, A. (1985). Specification analysis of econometric models. *J. Quant. Econ.* 2: 187–209.

[39] Van Huffel, S., Vandewalle, J., Staar, J. (1991). The total linear least squares problem: Properties, applications and generalization. Internal Report. K. U. Leuven, Belgium: ESAT Lab., Department of Electrical Engineering.

[40] Van Huffel, S., Vandewalle J. (1991). The total least squares problem: computational aspects and analysis. Society for Industrial and Applied Mathematics (SIAM).

[41] Wei, Y., Xie, P., Zhang, L. (2016). Tikhonov regularization and randomized GSVD. *SIAM J. Matrix Anal. Appl.* 37(2):649–675. DOI: 10.1137/15M1030200.

[42] Wuytack, L. (1979). *Padé Approximation and Its Applications*. Vol. 765 of Lectures Notes in Mathematics. Berlin: Springer.

[43] Xie, P., Xiang, H., Wei, Y. (2019). Randomized algorithms for total least squares problems. *Numer. Linear Algebra Appl.* 26(1):e2219. DOI: 10.1002/nla.2219.

[44] Zhang, D., Cherkaev, E. (2008). Padé approximations for identification of air bubble volume from temperature or frequency dependent permittivity of a two-component mixture. *Inverse Prob. Sci. Eng.* 16(4):425–445. DOI: 10.1080/17415970701529213.

[45] Zhang, D., Cherkaev, E. (2009). Reconstruction of spectral function from effective permittivity of a composite material using rational function approximations. *Comput. Phys.* 228(15):5390–5409. VolDOI: 10.1016/j.jcp.2009.04.014.

[46] Zhang, D., Lamoureux, M.P., Margrave, G.F., Cherkaev, E. (2011). Rational approximation for estimation of quality q factor and phase velocity in linear, viscoelastic, isotropic media. *Comput. Geosci.* 15(1):117–133. DOI: 10.1007/s10596-010-9201-7.

[47] Zhang, X.L., Begleiter, H., Porjesz, B., Wang, W., Litke, A. (1995). Event related potentials during object recognition tasks. Brain Research Bulletin. 38(6):531–538.

Appendix A

Consider a model described by (1). Algebraic calculations show that $y_i = g(x_i)$ gives rise an over-determined linear system, $\mathbf{X}\beta \approx \mathbf{y}$. We want to find a vector β such that $\mathbf{X}\hat{\beta}$ is the best approximation to $\mathbf{y}$. The most popular method is ***least squares***, in which we choose the vector $\hat{\beta}$ that minimizes the least squares problem

$$\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - g(x_i))^2 = ||\mathbf{y} - \mathbf{X}\beta||_2^2\right\} \tag{A.1}$$

The SVD can be used to compute this minimum norm solution. If the SVD of matrix $\mathbf{X}$ is rewritten as

$$\mathbf{X} = \sum_{i=1}^{k}\tilde{\sigma}_i\tilde{u}_i\tilde{v}_i' \text{ where } k = \min(n, m) = \operatorname{rank}(\mathbf{X}) = m$$

where $\{\tilde{u}_i \in R^n, \tilde{v}_i \in R^m, \tilde{\sigma}_i \in R^+\}$ are the left and right singular vectors, and singular values of the data matrix $\mathbf{X}$, as defined in (2), respectively, then the minimum norm least squares solution is given by

$$\hat{\beta}_{LS} = \sum_{i=1}^{k}\frac{\tilde{u}_i\mathbf{y}}{\tilde{\sigma}_i}\tilde{v}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{A.2}$$

It is clear that the existence of small singular values $\tilde{\sigma}$ means ill-conditioned data matrix $\mathbf{X}$. Note also that if the data matrix ill-conditioned, the variance is still high. In this case, if the estimator $\hat{\beta}_{LS}$ is allowed to be biased, the variance can be substantially reduced. One way to deal with this problem is to compute the truncated SVD (TSVD) that gives the stable solutions.

The main objective of TSVD method is to replace the ill-conditioned matrix $\mathbf{X}$ with the best rank-r matrix $\mathbf{X}_r$. To perform this procedure, we remove the small singular values of $\mathbf{X}$, by setting the value of those below a given threshold to zero. If $\mathbf{X} \in \mathrm{R}^{n\times m}$ is replaced by $\mathbf{X}_r$, then we select a new vector β to the minimize the least squares problem

$$\min_{\beta}\left\{||\mathbf{y} - \mathbf{X}_r\beta||_2^2\right\}, \mathbf{X}_r = \sum_{i=1}^{r}\tilde{\sigma}_i\tilde{u}_i\tilde{v}_i' \tag{A.3}$$

As expressed in the above, the minimum-norm solution to (A.3) is obtained by truncating the singular value for $r<k = \operatorname{rank}(\mathbf{X}) \leq m$. In this case, the TSVD solution of (A.3) can be described in terms of the SVD as

$$\hat{\beta}_{TSVD} = \sum_{i=1}^{r}\frac{\tilde{u}_i'\mathbf{y}}{\tilde{\sigma}_i}\tilde{v}_i \tag{A.4}$$

where $r = rank(\mathbf{X}_r)$ is equivalent to the regularization (or truncation) parameter. It should be noted here that the singular values are ordered as$\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_r > \tilde{\sigma}_{r+1} = \cdots \tilde{\sigma}_k = 0$ and the small nonzero singular values $(\tilde{\sigma}_{r+1}, ..., \tilde{\sigma}_k)$ are changed by exact zeros. The idea, then, is to choose the desired parameter r that filters the elements of the solution corresponding to the smallest values of $\tilde{\sigma}_i$. For our purposes, the SVD solution in (A.4) has a simple interpretation in terms of filter factors. If one set

$$f_1 = \cdots = f_r = 1 \text{ and } f_{r+1} = \cdots f_k = 0$$

then the TSVD solution based on truncated parameter r is obtained by

$$\hat{\beta}_{TSVD(f)} = \sum_{i=1}^{k} f_i \frac{\tilde{u}_i'\mathbf{y}}{\tilde{\sigma}_i}\tilde{v}_i \text{ where } k = \min(n, m) \le m \tag{A.5}$$

Hence, it follows from (A.5) that TSVD solution is also a filtered version of the least squares solution in (A.2).

In equation (A.4), the solution $\hat{\beta}_{TSVD}$ is referred to as the regularized SVD (or truncated SVD, the TSVD). An alternative approach to the TSVD method is ***Tikhonov regularization***, introduced by [31]. This regularization provides a solution to the overdetermined system by minimizing the penalized least squares problem

$$\min_{\beta}\left\{\|\begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{L}\end{bmatrix}\beta - \begin{bmatrix}\mathbf{y}\\ 0\end{bmatrix}\|_2^2\right\} = \min_{\beta}\left\{\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda\|\mathbf{L}\beta\|_2^2\right\} \tag{A.6}$$

where $\mathbf{L}$ shows a regularization matrix, and the scalar λ is known as the regularization parameter to be selected. When $\lambda>0$ this problem is always of full column rank and has a unique solution [5]. The matrix $\mathbf{L}$ is commonly selected to be the identity matrix $\mathbf{I}$; however, if the solution β has particular known properties, then we may set the $\mathbf{L}$ matrix as the first or second derivative operator [see 8, 21].

The problem (A.6), also called a damped least squares (DLS), is equivalent to the least squares normal equations

$$\left(\frac{\mathbf{X}}{\sqrt{\lambda}\mathbf{L}}\right)'\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{L}\end{pmatrix}\beta = \begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{L}\end{pmatrix}'\begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix} \tag{A.7}$$

The equation (A.7) can be equivalently rewritten as

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{L}'\mathbf{L})\beta = \mathbf{X}'\mathbf{y} \tag{A.8}$$

In order to analyze how $\mathbf{X}$ and $\mathbf{L}$ interact in the Tikhonov problem, it would be useful to convert (A.8) into an equivalent diagonal problem. If $\mathbf{L} = \mathbf{I}$, then the SVD overcomes this task for the ridge regression problem. For the Tikhonov problem, a generalized version of the SVD that diagonalizes both $\mathbf{X}$ and $\mathbf{L}$ is discussed by [18].

For each value of λ, when $\mathbf{L} = \mathbf{I}$ it is easy prove that the Tikhonov solution based on SVD of the (A.8) is given by

$$\hat{\beta}_{TTK} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{L}'\mathbf{L})^{-1}\mathbf{X}'\mathbf{y} = \sum_{i=1}^{k} f_i \frac{\tilde{u}_i'\mathbf{y}}{\tilde{\sigma}_i}\tilde{v}_i, f_i = \frac{\tilde{\sigma}_i^2}{\tilde{\sigma}_i^2 + \lambda^2} \tag{A.9}$$

As expressed before, the quantities f_i are commonly known as filter factors. It is also noted that as long as $f_i \cong 1, \hat{\beta}_{TTK}$ will approximately equal the $\hat{\beta}_{LS}$ and $\hat{\beta}_{TVSD}$ for $k = r$, respectively. Specifically, standard Tikhonov solution defined in (A.9) is also known as Ridge regression solution.

Appendix B

In order to see comparison of P – $TTLS$, KS and SS methods, Figures B1–B3 are given below.

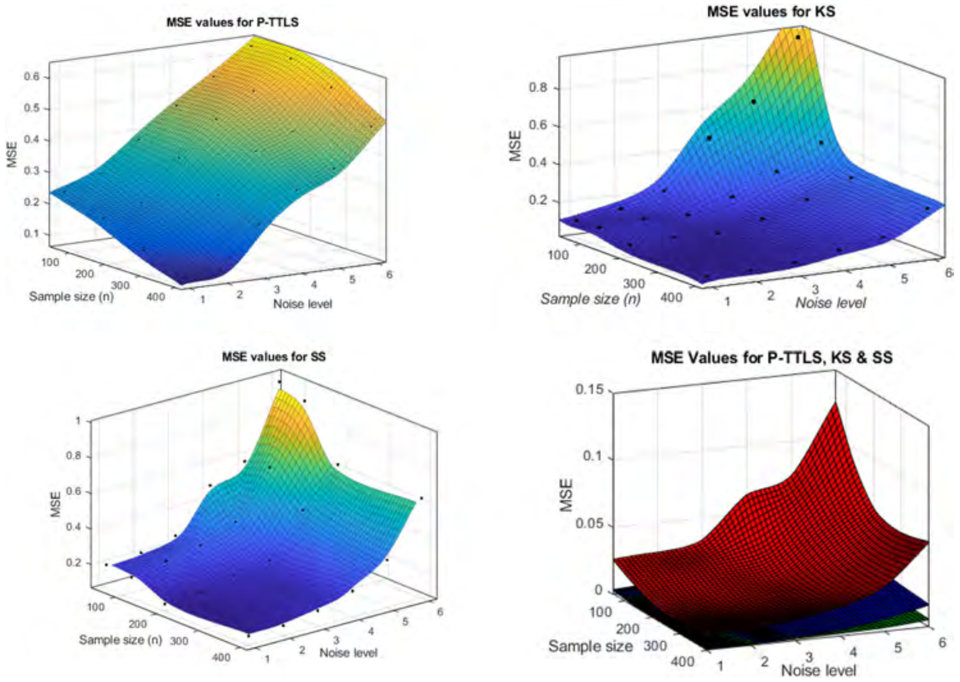

Figure B1. 3 D plots for noise level factor to see performances of $P - TTLS$, KS and SS. Right of the bottom panel, 3 D diagram is given for three methods $P - TTLS$ (blue), KS (green) and SS (red).

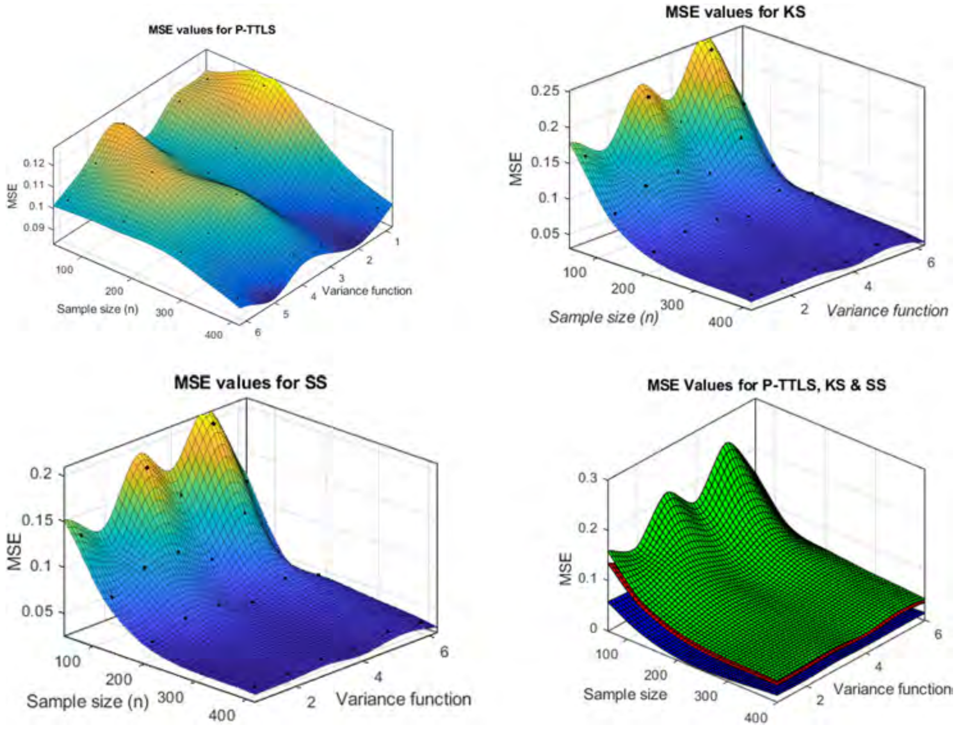

Figure B2. 3 D plots for variance function factor to see performances of $P - TTLS$, KS and SS.

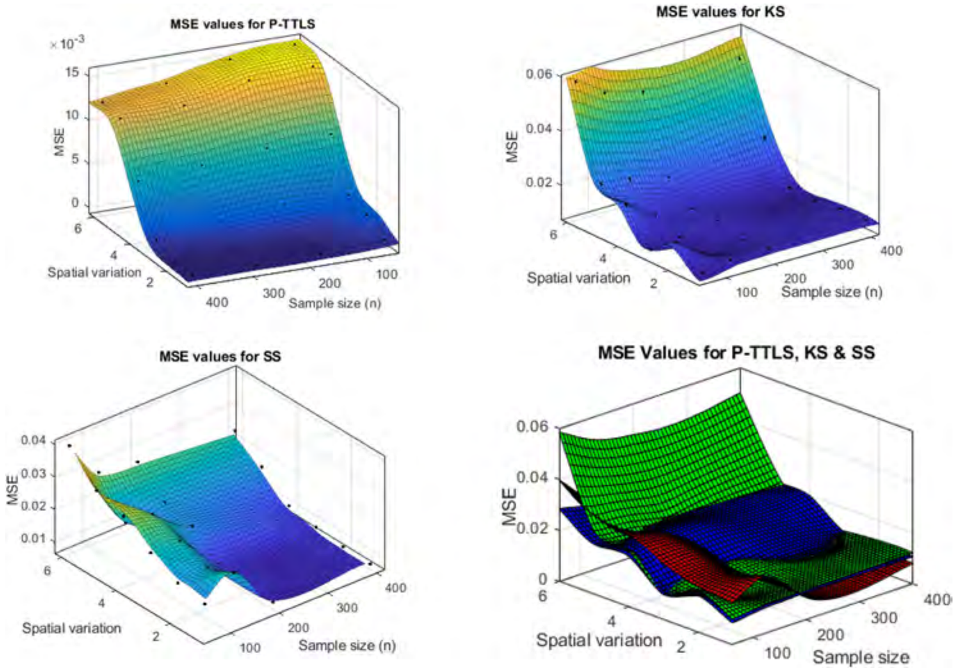

Figure B3. 3 D plots for spatial variation factor to see performances of *P* – *TTLS*, *KS* and *SS*.